SpudDB: A database for accessing potato genomic data

John P. Hamilton[1,2], Julia Brose[2], and C. Robin Buell[1,2,3,4,*]
[1] Department of Crop & Soil Sciences, University of Georgia, Athens, GA, 30602, USA
[2] Center for Applied Genetic Technologies, University of Georgia, Athens, GA, 30602, USA
[3] Institute of Plant Breeding, Genetics & Genomics, University of Georgia, Athens, GA, 30602, USA
[4] The Plant Center, University of Georgia, Athens, GA, 30602, USA

*Corresponding author: Center for Applied Genetic Technologies, 111 Riverbend Rd, University of Georgia, Athens, GA 30602; +1-706-432-4021; Robin.Buell@uga.edu

Keywords: Potato, diversity, co-expression, annotation

Abstract

Potato is a key food crop with a complex, polyploid genome. Advancements in sequencing technologies coupled with improvements in genome assembly algorithms have enabled generation of phased, chromosome-scale genome assemblies for cultivated tetraploid potato. The SpudDB database houses potato genome sequence and annotation, with the doubled monoploid DM 1-3 516 R44 (hereafter DM) genome serving as the reference genome and haplotype. Diverse annotation data types for DM genes are provided through a suite of Gene Report Pages including gene expression profiles across 438 potato samples. To further annotate potato genes based on expression, 65 gene co-expression modules were constructed that permit identification of tightly co-regulated genes within DM across development and responses to wounding, abiotic stress, and biotic stress. Genome browser views of DM and 28 other potato genomes are provided along with a download page for genome sequence and annotation. To link syntenic genes within and between haplotypes, syntelogs were identified across 25 cultivated potato genomes. Through access to potato genome sequences and associated annotations, SpudDB can enable potato biologists, geneticists and breeders to continue to improve this key food crop.

Introduction

Solanum Sect Petota contains approximately 100 species including *Solanum tuberosum* L. (cultivated potato) (Spooner, 2009). Potatoes were domesticated from wild potato species approximately 8,000-10,000 years ago in the Andes (Spooner et al., 2005) and have since spread throughout the world serving as a critical food crop. Potato tubers are modified underground stem structures and are clonally derived. While tubers serve as a mode of asexual reproduction and as a mechanism to overwinter and evade predation, the reliance on clonal propagation results in high genetic load due to the lack of a meiotic sieve to remove deleterious and dysfunctional alleles.

Most cultivated potato cultivars are autotetraploids ($2n = 4x = 48$) and due to its complex genome, the first potato genome sequenced was that of *S. tuberosum* Group Phureja DM 1-3 516 R44 (hereafter DM), a doubled monoploid derived from a diploid clone via anther culture (Paz & Veilleux, 1999). Its homozygosity permitted assembly in 2011 of the DM genome via *de novo* assembly of short-read sequences prior to development of third generation long read sequencing platforms (Potato Genome Sequencing Consortium et al., 2011). Access to the DM reference genome sequence, albeit a single haplotype, permitted an explosion of genome-enabled discoveries in potato including development of SNP-chip genotyping arrays (Felcher et al., 2012; Vos et al., 2015), assessment of potato structural genome diversity (Hardigan et al., 2015, 2016), understanding genes underlying domestication (Hardigan et al., 2017), discovery of genes associated with agronomic traits (e.g., (Sharma et al., 2018; Klaassen et al., 2019; Khlestkin et al., 2020; Prodhomme et al., 2020)), and furthering our knowledge of potato biology (e.g. (Peterson et al., 2016; Ye et al., 2018; Enciso-Rodriguez et al., 2019; Laimbeer et al., 2020; Zhang et al., 2020; Eggers et al., 2021; Ma et al., 2021; Ramírez Gonzales et al., 2021)). With respect to the genome landscape of cultivated tetraploids, whole genome resequencing coupled with alignments to the DM reference genome revealed a high degree of heterozygosity coupled with rampant structural variation attributable to mutation and wild species introgressions (Hardigan et al., 2017).

In the last 15 years, the advances in sequencing technologies, enhanced genome assembly algorithms, and increased computing capacities have resulted in generation of multiple potato genome sequences, including heterozygous diploid genomes ($2n = 2x = 24$) as well as phased, tetraploid genomes. Some of the first cultivated potato genomes available subsequent to DM were diploid or dihaploid ($2n = 2x = 24$) genomes derived from cultivated tetraploids (van Lieshout et al., 2020; Zhou et al., 2020; Jayakody et al., 2023). With access to long-read sequencing platforms and improved algorithms that can phase genome assemblies in the last few years, phased tetraploid genomes are now feasible (Hoopes et al., 2022; Sun et al., 2022; Bao et al., 2023) confirming earlier estimates of genome heterogeneity, allelic diversity, structural variation, wild species introgressions, and a high degree of dysfunctional and deleterious alleles that were derived from short-read sequencing alignments to the DM reference genome. While most potato genome sequences have been generated from cultivated potato clones, a subset of wild potato species (Leisner et al., 2018; Yan et al., 2021; Tang et al., 2022; Feng et al., 2024; Hosaka et al., 2024) have been sequenced. Central to these emerging genome sequences was updating the DM reference genome to a chromosome-scale using long-read sequences coupled with Hi-C data that was re-annotated using RNA-sequencing (RNA-seq) and full-length cDNA sequences, greatly improving the quality of the genome sequence and the gene annotation (Pham et al., 2020).

Genome and genetic data for potato are housed in a limited number of databases. The PoMaMo database included molecular maps, genome sequences, and suite tools (Meyer et al., 2005); however, it is no longer available. A subset of potato genomic data is available on the Solanaceae Genomics

Network (Fernandez-Pozo et al., 2015), yet this database is highly focused on tomato. Potato genome sequences are also available at the SpudDB (Hirsch et al., 2014) which was created for breeders to mine genotype and phenotype diversity data primarily derived from North American cultivated potato as part of the USDA-funded SolCAP project (Douches et al., 2014). With the recent availability of extensive chromosome-scale genome assemblies, we have updated SpudDB with new content, features, and access tools.

Overview and Navigating SpudDB

SpudDB provides access to potato genome sequences via genome browsers, search tools, download pages, and diverse annotation data types for the DM v6.1 reference genome. The home page of SpudDB (https://spuddb.uga.edu/index.shtml) highlights recent updates, a summary of the content of the database including links to literature associated with large datasets in SpudDB, and a quick search tool for DM v6.1 genes either by gene identifier or keyword. The menu provides access to the JBrowse2 genome browsers (Diesh et al., 2023), database search tools, dataset download pages, and the results of various analyses such as gene expression. Archived updates of SpudDB are available on the What's New page (https://spuddb.uga.edu/new.shtml).

To facilitate access to potato genome sequences, we have deployed a JBrowse2 genome browser (Diesh et al., 2023) hosting 29 total genomes including the DM v6.1 reference genome (Pham et al., 2020), phased diploid breeding line RH 89-039-16 (Zhou et al., 2020), phased tetraploid genomes (Hoopes et al., 2022; Sun et al., 2022; Bao et al., 2023), *S. chacoense* M6--a source of self-compatibility (Jansky et al., 2011; Eggers et al., 2021; Ma et al., 2021), and *S. candolleanum*, the progenitor of cultivated potato (Spooner et al., 2005). A suite of search tools for the DM v6.1 gene annotation is located on the Search Tool Page (https://spuddb.uga.edu/integrated_searches.shtml) including a BLAST (v2.2.26) (Altschul et al., 1997) search tool, functional annotation keyword search tool, InterPro identifier and key word search tool, Gene Ontology identifier and key word search tool, Pfam accession search tool, and a sequence identifier search tool. On the top menu bar is the Analyses tab which provides links to DM specific analyses such as gene expression, gene co-expression, and potato syntelogs. A Contact tab is also present to permit users to send feedback. A Download tab on the top menu bar contains links to webpages that describe available genome datasets including:

- DM v6.1 genome assembly, annotation, gene expression matrix, and variant calls
- Eight phased tetraploid genome assemblies and annotation (cv. Altus, Atlantic, Avenger, Castle Russet, Colomba, Cooperation-88, Otava, Spunta)
- Twenty phased dihaploid genome assemblies and annotation from the Potato 2.0 project
- Diploid RH89-039-16 (v3) genome assembly and annotation
- Doubled monoploid DM1S1 genome assembly, annotation, and variant calls
- Updated *S. chacoense* M6 (v5.0) genome assembly and annotation
- *S. candolleanum* (v1.0) genome assembly and annotation
- Archived *S. chacoense* M6 (v4.1) genome assembly and annotation
- Archived DM (PGSC v4.03/v4.04) genome assembly and annotation
- Tomato (*Solanum lycopersicum*) M82 (SollycM82_v1) genome assembly and annotation

In addition, links to the genome browser for each of these assemblies are provided via its download webpage.

DM v6.1 as the reference genome for potato

The DM v6.1 reference genome serves as the foundation for the potato community. Not only is it a high-quality, chromosome-scale genome assembly (Pham et al., 2020), it represents a single haplotype that serves to link alternative haplotypes present in highly heterozygous, phased diploid and tetraploid genome assemblies thereby linking alleles and syntelogs. The 741 Mb DM v6.1 genome was scaffolded into 12 chromosomes and annotated using extensive transcript evidence, including full-length cDNAs, resulting in 40,652 working protein-coding genes encoding 52,953 gene models and 32,917 high-confidence protein-coding genes encoding 44,851 gene models (Pham et al., 2020). The DM genes have been annotated for a suite of annotation data types to aid in understanding gene function which are available on individual Gene Report Pages for each gene. These include putative functional description assigned through BLAST searches against the *Arabidopsis thaliana* proteome, Swiss-Prot plant proteins, and the Pfam database as well as gene expression abundances. Further annotations include Gene Ontology classifications, BLAST searches against UniRef database, gene co-expression module assignment, and syntelogs across cultivated potato genomes.

The major update to SpudDB was expansion of gene expression profiles to include additional RNA-sequencing (RNA-seq) datasets from a broader group of developmental stages, tissues, and treatments. To obtain relevant expression datasets, the National Center for Biotechnology Information Sequence Read Archive (Sayers et al., 2022) was queried for *S. tuberosum*. Initial filtering for paired end RNA-seq datasets resulted in 4,571 Sequence Read Archive accessions. These were then filtered for sequencing platform requiring the Illumina platform, RNA-seq library, minimum of 20 million reads, paired end sequences, and informative sample description. A subset of 456 accessions were downloaded and quality checked using FastQC (v0.12.1; (Wingett SW, 2018)) and MultiQC (Ewels et al., 2016) using default parameters and were then classified based on organ and treatment/conditions [Organ: fruit, flower, leaf, root, seedling, stem, tuber; Treatment/Conditions: abiotic stress, biotic stress, development, photoperiod, wounding] based on the BioProject and BioSample description. Expression abundances were calculated using kallisto quant (Bray et al., 2016) with the parameter -t 8 and represented as transcripts per million (TPMs). All of the RNA-seq samples downloaded for the gene expression analysis were clustered to identify mis-labeled samples and 18 accessions were removed based on their PCC or PCA clustering generated using the R commands: prcomp (R Core Team, 2023) with default parameters and cor with method option set to pearson, respectively, with aberrant tissue types. The final RNA-seq dataset has 438 samples. Of the 40,652 DM genes, 39,651 are expressed at $\geq$ 0 TPM in at least one sample. Expression abundances are available for the entire DM genome via the Gene Expression page (https://spuddb.uga.edu/expression.shtml) or individually for each gene via the Gene Report Page.

Gene co-expression

Gene co-expression network modules were generated from all representative working gene models using Simple Tidy GeneCoEx with default parameters (Li et al., 2023). Co-expression modules were built using all 438 RNA-seq libraries (156 samples after replicate averaging) representing a diverse set of tissues and conditions/treatments including a tuber developmental series, abiotic stress, biotic stress, and a set of photoperiod conditions (Fig. 1a). To build the correlation network edge table, only edges with r > 0.8 were used, which corresponded to the top 1% of all edges. A network object was then constructed using the 'graph_from_data_frame()' function of igraph (Csárdi et al.) with option directed set to F. Graph based clustering was performed using the Leiden algorithm (implemented as the 'cluster_leiden()' function in R as part of the igraph package(Csárdi et al.) with a resolution parameter of 4 and objective_function parameter set to modularity. Of the 40,652 DM genes, 36,025 were placed into co-expression modules based on their expression pattern generating 65 modules containing between 5 and 2,943 genes.

The co-expression modules can then be used to identify genes with expression patterns associated with specific tissues or treatments. Tuber bulking relies on the accumulation of amylopectin in the amyloplasts that is catalyzed by starch synthases (Nazarian-Firouzabadi & Visser, 2017). Starch synthase V (Soltu.DM.02G027020.1) was previously identified as vital to tuber bulking (Li et al., 2024). In the co-expression analyses, starch synthase V is a member of Module 6 which has peak expression in the tuber short day time course sample collected at 3pm. The expression of starch synthase V was plotted in red along with the other genes in module 6 showing high expression during later stages of tuber development and the time course data sets (Fig. 1b). This expression profile is expected based on the activity of starch synthase V in amylopectin accumulation.

Co-expression modules can also be mined to identify additional genes involved in biological processes. For example, Module 11 has peak expression in wounded tubers after 3 days and generally high expression in all the wounded tubers after 1 day (Fig. 1a). Of the highly expressed genes, there was one MYB transcription factor. MYB transcription factors are known to be involved in wound healing through their regulation of suberin biosynthesis (Han et al., 2024). This uncharacterized MYB transcription factor (Soltu.DM.04G025530.1) exhibits an expression pattern with high expression after 14 days of wounding and could play an important role in wound healing in potato (Fig. 1c). Co-expression module membership for the entire DM genome can be obtained via the Gene Co-expression page (https://spuddb.uga.edu/coexpression.shtml) and the Gene Report page for each individual gene.

Potato genome sequences and linking across haplotypes

To facilitate traversing between alleles within and between genome assemblies of cultivated potato, we determined syntenic relationships between 25 cultivated potato genomes. Using DM as the reference genome, the representative gene models from four phased tetraploid genomes (Hoopes et al., 2022; Sun et al., 2022; Bao et al., 2023) and 20 phased dihaploid genomes from the Potato 2.0 project (https://potatov2.github.io/) were input into GENESPACE (Lovell et al., 2022) and syntelogs for each DM gene were identified. To account for the ploidy of the phased genomes, the ploidy parameter of the 'init_genespace' function was set to "1" for the DM v6.1 genome, "2" for the 20 dihaploid genomes, and "4" for the four tetraploid genomes. Syntelogs for each gene are viewable on the Gene Report Page and as a track on the DM v6.1 JBrowse.

DM v6.1 JBrowse

The DM v6.1 reference genome is available as a JBrowse2 instance (Diesh et al., 2023)(Fig. 2). Tracks available include reference sequence, loci and gene models with separate tracks for representative high confidence gene models, high confidence gene models, and working gene models. Gene expression data is available as RNA-seq coverage tracks of the 438 RNA-seq samples that were generated using HISAT2 (Kim et al., 2019) that are grouped based on classification. Syntelogs from GENESPACE (v1.3.1; (Lovell et al., 2022) are provided as well. For variant data, SNPs from genotyping-by-sequencing using a set of 57,054 SureSelect baits (Uitdewilligen et al., 2013) and the SolCAP SNP project that utilized RNA-seq and draft genome sequence to develop an Affymetrix SNP array platform (Hamilton et al., 2011) are provided to link positions in the DM genome with widely used genetic markers which are in use in community-based genotyping platforms. Individual JBrowse instances are also available for 28 other potato genomes [Tetraploid cultivars: Atlantic, Castle Russet, Cooperation-88, Otava; Diploid/Dihaploids: RH and 20 dihaploids from the Potato 2.0 project (https://potatov2.github.io/), Doubled Monoploid: DM1S1; Wild species: *S. chacoense* M6, *S. candolleanum*] and available via the top menu bar or via links on their individual Download page.

SpudDB Gene Report Pages

For a biologist, access to an array of annotation data types can facilitate understanding gene function. For each DM gene model, a Gene Report Page is available either through a search via gene identifier, key word identifier, or from a locus or gene model link in the JBrowse. The SpudDB Gene Report page (Fig. 3) has a summary of each gene model including putative function description, locus name, and alternative splice form (gene model). Basic metrics for each gene model include chromosome or scaffold location, coordinates on the DM v6.1 genome for the mRNA (predicted transcript = gene model), coding sequence length, and predicted amino acid length. The sequences of the genomic sequence, transcript sequence, coding sequence, and predicted protein sequence are provided in FASTA format. A link to the DM v6.1 JBrowse genome browser for each gene model is also included in the Gene Report Page. Gene ontology classifications are also provided as are searches against UniRef100. To facilitate development of targeted simple molecular markers, putative Simple Sequence Repeats (SSRs) for the locus are available with their coordinates.

A table of gene expression abundances for each gene model is provided along with its run identifier from National Center for Biotechnology Information Sequence Read Archive, short description of the sample, classification of the study, and expression abundance in TPMs. Co-expression module assignment is also listed with the co-expression module membership and the module peak expression assignment. To facilitate traversing from DM to other potato genomes, syntelogs from the potato syntelog analysis are listed for each gene.

Improvements in architecture

SpudDB has undergone a number of back-end improvements and enhancements since the last release to support future updates and to continue to provide useful tools to the user community. The entire SpudDB website has been converted to use HTTPS for increased compatibility and security. The search tools and Gene Report pages have been migrated from a PostgreSQL instance to SQLite for increased performance and reliability. The number of genome browsers on SpudDB had grown from the original MySQL based Gbrowse1 for the legacy DM annotation to include a number of JBrowse1 and JBrowse2 instances for the new and updated genomes added to SpudDB. All of these have been replaced by a single, unified JBrowse2 instance which is easier to maintain and provides an enhanced user experience.

Future Directions

As a substantial number of potato researchers are geneticists and breeders, SpudDB serves a key function in provision of genomic data not only from the DM reference genome but also from new emerging genome assemblies. We anticipate that more potato genome sequences and annotation will continue to be generated and made available to the public in the coming years. The back-end improvements to SpudDB, especially the use of JBrowse 2, will enable streamlined addition of new genomes to SpudDB. These assemblies, and importantly, the alleles in these genomes can be linked through synteny with new GENESPACE builds with the addition of new genomes to SpudDB. Continued development of new functional annotation datatypes will facilitate data-mining the potato genome and can readily be added to the Gene Report page for each gene. We also anticipate that development of a pan-genome for potato that captures novel structural variation in cultivated potato will provide new resources for potato researchers.

Data availability

All data are freely available at SpudDB (https://spuddb.uga.edu) for searching and download. We have also deposited the new gene expression, gene co-expression, and syntelog datasets in Figshare under https://doi.org/10.6084/m9.figshare.27471918.v1.

Funding

We gratefully acknowledge funding to C.R.B. from the National Science Foundation (IOS-2140176), U.S. Department of Agriculture (2019-51181-30021), Georgia Research Alliance, Georgia Seed Development, and the University of Georgia.

Conflicts of interest

The authors declare no conflict of interest.

REFERENCES

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

Bao, Z., Li, C., Li, G., Wang, P., Peng, Z., Cheng, L., Li, H., Zhang, Z., Li, Y., Huang, W., Ye, M., Dong, D., Cheng, Z., VanderZaag, P., Jacobsen, E., Bachem, C.W.B., Dong, S., Zhang, C., Huang, S., & Zhou, Q. (2023). Genome architecture and tetrasomic inheritance of autotetraploid potato. *Molecular Plant*, *16*, 1866

Bray, N.L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*, 525–527

Csárdi, G., Nepusz, T., Traag, V., Horvát, S., & Zanini, F. igraph: Network Analysis and Visualization in R. 2023. *URL https://CRAN. R-project. org*,

Diesh, C., Stevens, G.J., Xie, P., De Jesus Martinez, T., Hershberg, E.A., Leung, A., Guo, E., Dider, S., Zhang, J., Bridge, C., Hogue, G., Duncan, A., Morgan, M., Flores, T., Bimber, B.N., Haw, R., Cain, S., Buels, R.M., Stein, L.D., & Holmes, I.H. (2023). JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biology*, *24*, 74

Douches, D., Hirsch, C.N., & Manrique-Carpintero, N.C. (2014). The contribution of the Solanaceae coordinated agricultural project to potato breeding. *Potato Research*, *57*, 215-224

Eggers, E.-J., van der Burgt, A., van Heusden, S.A.W., de Vries, M.E., Visser, R.G.F., Bachem, C.W.B., & Lindhout, P. (2021). Neofunctionalisation of the Sli gene leads to self-compatibility and facilitates precision breeding in potato. *Nature Communications*, *12*, 4141

Enciso-Rodriguez, F., Manrique-Carpintero, C., N., Nadakuduti, S., S., Buell, R., C., Zarka, D., Douches, & . D. (2019). Overcoming Self-Incompatibility in Diploid Potato Using CRISPR-Cas9. *Frontiers in Plant Science*, *10*, 376

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*, 3047–3048

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

Feng, Y., Zhou, J., Li, D., Wang, Z., Peng, C., & Zhu, G. (2024). The haplotype-resolved T2T genome assembly of the wild potato species *Solanum commersonii* provides molecular insights into its freezing tolerance. *Plant Communications*, 100980

Fernandez-Pozo, N., Menda, N., Edwards, J.D., Saha, S., Tecle, I.Y., Strickler, S.R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., Yan, A., & Mueller, L.A. (2015). The Sol Genomics Network (SGN)--from genotype to phenotype to breeding. *Nucleic Acids Research*, *43*, D1036-41

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

Han, Y., Yang, R., Zhang, X., Wang, Q., Wang, Y., Li, Y., Prusky, D., & Bi, Y. (2024). MYB24, MYB144, and MYB168 positively regulate suberin biosynthesis at potato tuber wounds during healing. *The Plant Journal*, *119*, 1239–1257

Hardigan, M.A., Bamberg, J., Buell, C.R., & Douches, D.S. (2015). Taxonomy and Genetic Differentiation among Wild and Cultivated Germplasm of *Solanum* sect. Petota. *The plant genome*, *8*, eplantgenome2014.06.0025

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

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., Veilleux, R.E., & Buell, C.R. (2017). Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proceedings of the National Academy of Sciences*, *114*, E9999–E10008

Hirsch, C.D., Hamilton, J.P., Childs, K.L., Cepela, J., Crisovan, E., Vaillancourt, B., Hirsch, C.N., Habermann, M., Neal, B., & Buell, C.R. (2014). Spud DB: A resource for mining sequences, genotypes, and phenotypes to accelerate potato breeding. *The Plant Genome*, *7*. https://doi.org/10.3835/plantgenome2013.12.0042

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., & Finkers, R. (2022). Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. *Molecular plant*, *15*, 520–536

Hosaka, A.J., Sanetomo, R., & Hosaka, K. (2024). A de novo genome assembly of *Solanum bulbocastanum* Dun., a Mexican diploid species reproductively isolated from the A-genome species, including cultivated potatoes. *G3*, *14*, jkae080

Jansky, H., S., Chung, S., Y., Kittipadukal, & P. (2011). M6: A Diploid Potato Inbred Line for Use in Breeding and Genetics Research. *Journal of Plant Registrations*, *8*, 195–199

Jayakody, T.B., Hamilton, J.P., Jensen, J., Sikora, S., Wood, J.C., Douches, D.S., & Buell, C.R. (2023). Genome Report: Genome sequence of 1S1, a transformable and highly regenerable diploid potato for use as a model for gene editing and genetic engineering. *G3*, *13*, jkad036

Khlestkin, V.K., Erst, T.V., Rozanova, I.V., Efimov, V.M., & Khlestkina, E.K. (2020). Genetic loci determining potato starch yield and granule morphology revealed by genome-wide association study (GWAS). *PeerJ*, *8*, e10286

Kim, D., Paggi, J.M., Park, C., Bennett, C., & Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*, 907–915

Klaassen, M.T., Willemsen, J.H., Vos, P.G., Visser, R.G.F., van Eck, H.J., Maliepaard, C., & Trindade, L.M. (2019). Genome-wide association analysis in tetraploid potato reveals four QTLs for protein content. *Molecular Breeding*, *39*, 151

Laimbeer, F.P.E., Bargmann, B.O.R., Holt, S.H., Pratt, T., Peterson, B., Doulis, A.G., Buell, C.R., & Veilleux, R.E. (2020). Characterization of the F Locus Responsible for Floral Anthocyanin Production in Potato. *G3*, *10*, 3871–3879

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., & Buell, C.R. (2018). Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species *Solanum chacoense*, reveals residual heterozygosity. *The Plant Journal 94*, 562–570

Li, C., Deans, N.C., & Buell, C.R. (2023). “Simple Tidy GeneCoEx”: A gene co-expression analysis workflow powered by tidyverse and graph-based clustering in R. *The Plant Genome*, *16*, e20323

Li, H., Brouwer, M., Pup, E.D., van Lieshout, N., Finkers, R., Bachem, C.W.B., & Visser, R.G.F. (2024). Allelic variation in the autotetraploid potato: genes involved in starch and steroidal glycoalkaloid metabolism as a case study. *BMC Genomics*, *25*, 274

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., & Finkers, R. (2020). Solyntus, the New Highly Contiguous Reference Genome for Potato (Solanum tuberosum). *G3* , *10*, 3489–3495

Lovell, J.T., Sreedasyam, A., Schranz, M.E., Wilson, M., Carlson, J.W., Harkess, A., Emms, D., Goodstein, D.M., & Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife*, *11*, eLife.78526

Ma, L., Zhang, C., Zhang, B., Tang, F., Li, F., Liao, Q., Tang, D., Peng, Z., Jia, Y., Gao, M., Guo, H., Zhang, J., Luo, X., Yang, H., Gao, D., Lucas, W.J., Li, C., Huang, S., & Shang, Y. (2021). A nonS-locus F-box gene breaks self-incompatibility in diploid potatoes. *Nature Communications*, *12*, 4142

Meyer, S., Nagel, A., & Gebhardt, C. (2005). PoMaMo--a comprehensive database for potato genome data. *Nucleic Acids Research*, *33*, D666-70

Nazarian-Firouzabadi, F., & Visser, R.G.F. (2017). Potato starch synthases: Functions and relationships. *Biochemistry and Biophysics Reports*, *10*, 7–16

Paz, M.M., & Veilleux, R.E. (1999). Influence of culture medium and *in vitro* conditions on shoot regeneration in *Solanum phureja* monoploids and fertility of regenerated doubled monoploids. *Plant Breeding*, *118*, 53–57

Peterson, B.A., Holt, S.H., Laimbeer, F.P.E., Doulis, A.G., Coombs, J., Douches, D.S., Hardigan, M.A., Buell, C.R., & Veilleux, R.E. (2016). Self-Fertility in a Cultivated Diploid Potato Population Examined with the Infinium 8303 Potato Single-Nucleotide Polymorphism Array. *The plant genome*, *9*. https://doi.org/10.3835/plantgenome2016.01.0003

Pham, G.M., Hamilton, J.P., Wood, J.C., Burke, J.T., Zhao, H., Vaillancourt, B., Ou, S., Jiang, J., & 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

Prodhomme, C., Vos, P.G., Paulo, M.J., Tammes, J.E., Visser, R.G.F., Vossen, J.H., & van Eck, H.J. (2020). Distribution of P1(D1) wart disease resistance in potato germplasm and GWAS identification of haplotype-specific SNP markers. *TAG. Theoretical and Applied Genetics*, *133*, 1859–1871

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/.

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

Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B.W., Pruitt, K.D., & Sherry, S.T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*, D20–D26

Sharma, S.K., MacKenzie, K., McLean, K., Dale, F., Daniels, S., & Bryan, G.J. (2018). Linkage Disequilibrium and Evaluation of Genome-Wide Association Mapping Models in Tetraploid Potato. *G3*, *8*, 3185–3202

Spooner, D.M. (2009). DNA barcoding will frequently fail in complicated groups: An example in wild potatoes. *American Journal of Botany*, *96*, 1177–1189

Spooner, D.M., McLean, K., Ramsay, G., Waugh, R., & Bryan, G.J. (2005). A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 14694–14699

Sun, H., Jiao, W.-B., Krause, K., Campoy, J.A., Goel, M., Folz-Donahue, K., Kukat, C., Huettel, B., & Schneeberger, K. (2022). Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nature Genetics*, *54*, 342–348

Tang, D., Jia, Y., Zhang, J., Li, H., Cheng, L., Wang, P., Bao, Z., Liu, Z., Feng, S., Zhu, X., Li, D., Zhu, G., Wang, H., Zhou, Y., Zhou, Y., Bryan, G.J., Buell, C.R., Zhang, C., & Huang, S. (2022). Genome evolution and diversity of wild and cultivated potatoes. *Nature*, *606*, 535–541

Uitdewilligen, J.G., Wolters, A.M., D'Hoop B, B., Borm, T.J., Visser, R.G., & van Eck, H.J. (2013). A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PloS One*, *8*, e62355

Vos, P.G., Uitdewilligen, J.G.A.M.L., Voorrips, R.E., Visser, R.G.F., & van Eck, H.J. (2015). Development and analysis of a 20K SNP array for potato (*Solanum tuberosum*): an insight into the breeding history. *TAG. Theoretical and Applied Genetics*, *128*, 2387–2401

Wingett SW, A.S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, *7*, 1338

Yan, L., Zhang, Y., Cai, G., Qing, Y., Song, J., Wang, H., Tan, X., Liu, C., Yang, M., Fang, Z., & Lai, X. (2021). Genome assembly of primitive cultivated potato *Solanum stenotomum* provides insights into potato evolution. *G3*, *11*, jkab262

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

Zhang, X., Campbell, R., Ducreux, L.J.M., Morris, J., Hedley, P.E., Mellado-Ortega, E., Roberts, A.G., Stephens, J., Bryan, G.J., Torrance, L., Chapman, S.N., Prat, S., & Taylor, M.A. (2020). TERMINAL FLOWER-1/CENTRORADIALIS inhibits tuberisation via protein interaction with the tuberigen activation complex. *The Plant Journal*, *103*, 2263–2278

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., Zhang, C., & Huang, S. (2020). Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nature Genetics*, *52*, 1018–1023

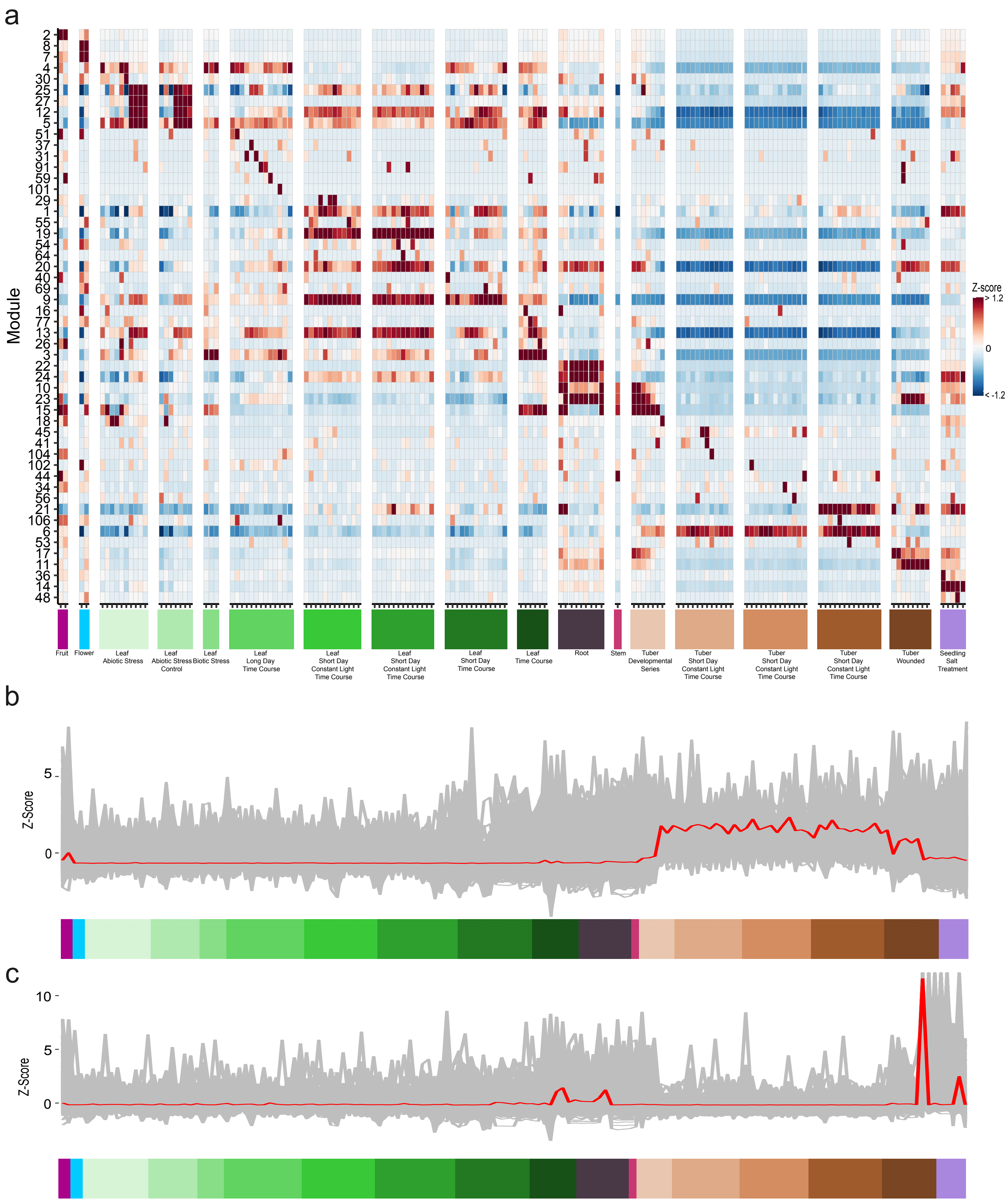
Figures

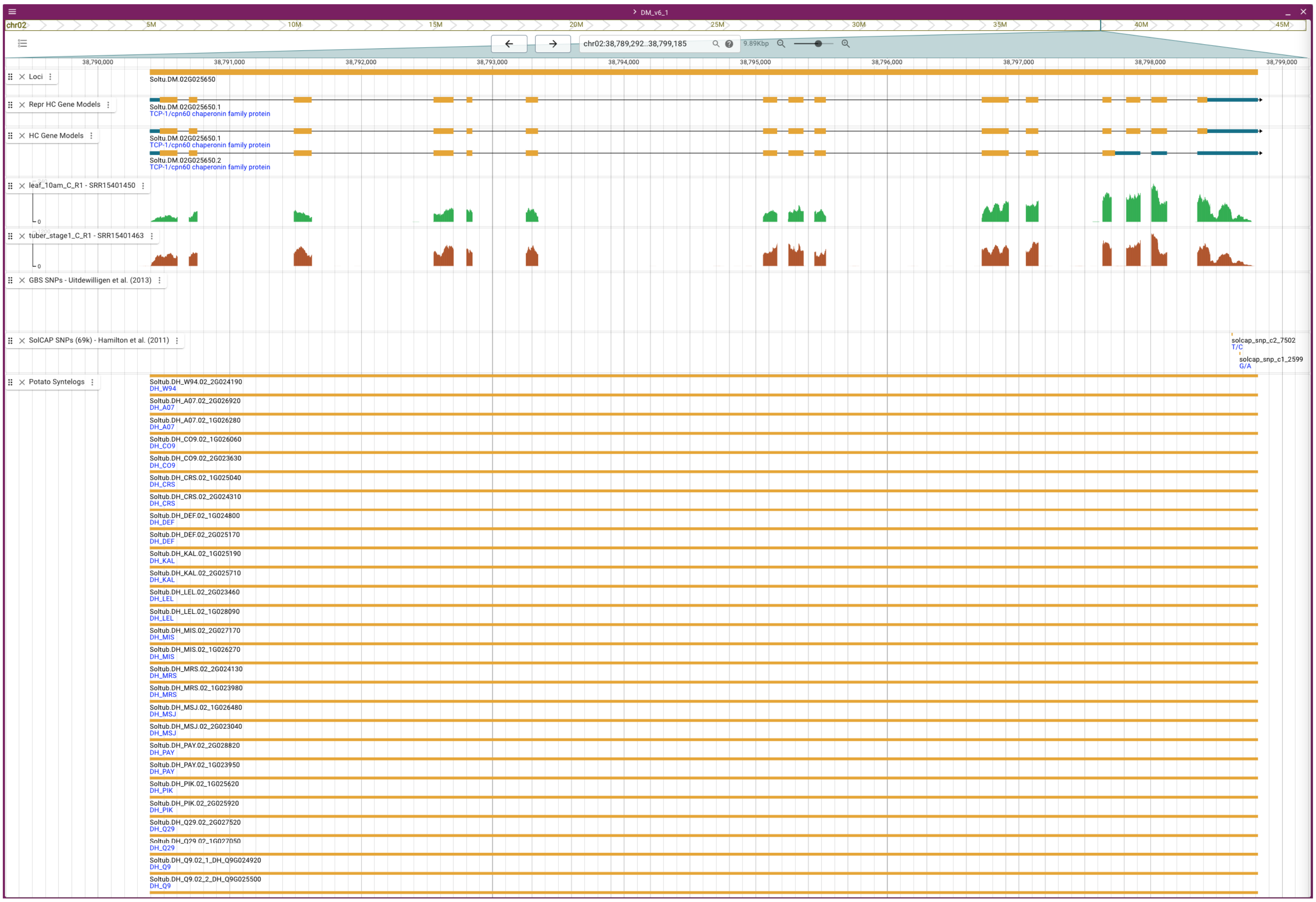
Fig. 1. Gene Co-expression Module Expression. a) Module expression of 438 RNA-sequencing libraries representing 18 conditions/treatments over seven tissue types. b) Expression of starch synthase V (Soltu.DM.02G027020). Z-score expression of all genes Module 6 is plotted in grey with starch synthase V in red. c) Expression of a MYB transcription factor (Soltu.DM.04G025530). Z-score expression of all genes in Module 11 is plotted in grey with the MYB transcription factor in red.

Fig. 2. SpudDB genome browser. Exemplar JBrowse2 screenshot of the DM v6.1 browser featuring Soltu.DM.02G025650 which encodes a TCP-1/cpn60 chaperonin family protein. The locus, representative high confidence gene model (Soltu.DM.02G025650.1), all high confidence gene models, two RNA-sequencing read alignments (leaf, 10 AM, Rep1 and tuber, stage 1, Rep1), variants from genotyping-by-sequencing and the SolCAP SNP array, and syntelogs within cultivated potato identified by GENESPACE are shown.

Fig. 3. Gene Report Page features in SpudDB. a) General information regarding Soltu.DM.02G025650.1 including link to the gene model in the genome browser, putative function, locus name, alternative splice form, and gene attributes (chromosome, coordinates, CDS length, and amino acid length). b) Gene ontology information. c) BlastP search of UniRef100 showing accession identifier, percent similarity, percent coverage, description, and p-value. d) Pfam and InterPro matches including accession identifier, method, name, match positions, and e-value. e) Potato syntelogs identified through GENESPACE. F) Coexpression module assignment including peak expression within the module. g) RNA-seq gene expression values in transcripts per million with National Center for

Biotechnology Information Sequence Read Archive accession identifier and sample description.


a
Module
2
8
7
4
30
25
27
12
5
51
37
31
91
59
101
29
1
55
19
54
64
20
40
69
9
16
77
13
26
3
22
24
10
23
15
18
45
41
104
102
44
34
56
21
106
6
53
17
11
36
14
48
Z-score
> 1.2
0
< -1.2
Fruit
Flower
Leaf Abiotic Stress
Leaf Abiotic Stress Control
Leaf Biotic Stress
Leaf Long Day Time Course
Leaf Short Day Constant Light Time Course
Leaf Short Day Constant Light Time Course
Leaf Short Day Time Course
Leaf Time Course
Root
Stem
Tuber Developmental Series
Tuber Short Day Constant Light Time Course
Tuber Short Day Constant Light Time Course
Tuber Short Day Constant Light Time Course
Tuber Wounded
Seedling Salt Treatment
b
Z-Score
5
0
c
Z-Score
10
5
0



DM_v6_1
chr02
chr02:38,789,292..38,799,185
9.89Kbp
Loci
Soltu.DM.02G025650
Repr HC Gene Models
Soltu.DM.02G025650.1
TCP-1/cpn60 chaperonin family protein
HC Gene Models
Soltu.DM.02G025650.1
TCP-1/cpn60 chaperonin family protein
Soltu.DM.02G025650.2
TCP-1/cpn60 chaperonin family protein
leaf_10am_C_R1 - SRR15401450
tuber_stage1_C_R1 - SRR15401463
GBS SNPs - Uitdewilligen et al. (2013)
SolCAP SNPs (69k) - Hamilton et al. (2011)
solcap_snp_c2_7502
T/C
solcap_snp_c1_2599
G/A
Potato Syntelogs
Soltub.DH_W94.02_2G024190
DH_W94
Soltub.DH_A07.02_2G026920
DH_A07
Soltub.DH_A07.02_1G026280
DH_A07
Soltub.DH_C09.02_1G026060
DH_C09
Soltub.DH_C09.02_2G023630
DH_C09
Soltub.DH_CRS.02_1G025040
DH_CRS
Soltub.DH_CRS.02_2G024310
DH_CRS
Soltub.DH_DEF.02_1G024800
DH_DEF
Soltub.DH_DEF.02_2G025170
DH_DEF
Soltub.DH_KAL.02_1G025190
DH_KAL
Soltub.DH_KAL.02_2G025710
DH_KAL
Soltub.DH_LEL.02_2G023460
DH_LEL
Soltub.DH_LEL.02_1G028090
DH_LEL
Soltub.DH_MIS.02_2G027170
DH_MIS
Soltub.DH_MIS.02_1G026270
DH_MIS
Soltub.DH_MRS.02_2G024130
DH_MRS
Soltub.DH_MRS.02_1G023980
DH_MRS
Soltub.DH_MSJ.02_1G026480
DH_MSJ
Soltub.DH_MSJ.02_2G023040
DH_MSJ
Soltub.DH_PAY.02_2G028820
DH_PAY
Soltub.DH_PAY.02_1G023950
DH_PAY
Soltub.DH_PIK.02_1G025620
DH_PIK
Soltub.DH_PIK.02_2G025920
DH_PIK
Soltub.DH_Q29.02_2G027520
DH_Q29
Soltub.DH_Q29.02_1G027050
DH_Q29
Soltub.DH_Q9.02_1_DH_Q9G024920
DH_Q9
Soltub.DH_Q9.02_2_DH_Q9G025500
DH_Q9

a

Soltu.DM.02G025650.1 — DM v6.1 Annotation

Show Soltu.DM.02G025650.1 in the Spud DB Genome Browser

Gene Identification

Putative Function:	TCP-1/cpn60 chaperonin family protein
Locus Name:	Soltu.DM.02G025650
Alternative Splice Form:	Soltu.DM.02G025650.2

Gene Attributes

Scaffold:	chr02
mRNA Genomic Coords (5'-3'):	38790394 - 38798818
CDS length:	1608 nt
Protein length:	535 aa

b

Gene Ontology Classification

GO accession	Type	Name	Code	With
GO:0005515	molecular_function	protein binding	IEA	TAIR:AT3G02530
GO:0042221	biological_process	response to chemical	IEA	TAIR:AT3G02530
GO:0005829	cellular_component	cytosol	IEA	TAIR:AT3G02530

c

BlastP Searches (UniRef 100)

Accession	% Sim	% Cov	Description	P-value
UniRef100_M1AYG3	100	99.8	TCP domain class transcription factor n=1 Tax=Solanum tubero	1e-303
UniRef100_A0A6N2CBJ5	100	99.8	Uncharacterized protein n=1 Tax=Solanum chilense TaxID=4083	3.9e-303
UniRef100_A0A3Q7F9W4	100	99.8	Uncharacterized protein n=1 Tax=Solanum lycopersicum TaxID=4	1.9e-302

d

PFAM hits

Accession	Name	Match Start	Match End	E-value
PF00118	Cpn60_TCP1	29	528	6.1e-165

Interpro hits

Interpro Acc	Method	Method Desc	Match Start	Match End
IPR027413	Gene3D	GROEL	1	185
IPR012722	TIGRFAM	chap_CCT_zeta: T-complex protein 1, zeta subunit	2	533
IPR012722	CDD	TCP1_zeta	6	529

e

Potato Syntelogs

Syntelog ID	Species
Soltub.DH_A07.02_2G026920	DH_A07
Soltub.DH_A07.02_1G026280	DH_A07

f

Coexpression Module Assignment

Module ID	Module Peak Expression
6	tuber_3pm_SD

g

RNA-Seq Gene Expression Values

SRA Run Accession	SRA Exp Accession	Sample Desc	Sample Category	TPM
SRR15401435	SRX11703628	closedflower_control_C_R1	Flower	35.2437
SRR15401434	SRX11703629	closedflower_control_C_R2	Flower	32.994
SRR15401433	SRX11703630	closedflower_control_C_R3	Flower	33.8838