Spud DB: A Resource for Mining Sequences, Genotypes, and Phenotypes to Accelerate Potato Breeding

Cory D. Hirsch, John P. Hamilton, Kevin L. Childs, Jason Cepela, Emily Crisovan, Brieanne Vaillancourt, Candice N. Hirsch, Marc Habermann, Brayden Neal, and C. Robin Buell*

Abstract
Potato is the world’s third most important crop, and is becoming increasingly important in developing countries. Cultivated potato is a highly heterozygous tetraploid ($2n = 4x = 48$) and suffers from significant inbreeding depression when selfed. As potato can be vegetatively propagated, breeding has been based primarily on phenotypic selection in F1 populations. However, recent advances in genome sequencing and genotyping methods have resulted in the development of large genomic, genetic, and phenotypic datasets that will enable more efficient and rapid breeding approaches. We have developed Spud DB (http://potato.plantbiology.msu.edu/) for the community to access the potato genome sequence and associated annotation datasets, along with phenotypic and genotypic data from a diversity panel of 250 potato clones. The Breeder’s Assistant is a web tool to retrieve pertinent phenotypic and genotypic data in a user-guided manner, and query polymorphic markers such as single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs) to identify custom sets of markers for a gene or region of interest. To browse and query the potato genome, a genome browser with 94 tracks of genome annotation, sequence variants, and expression abundance has been deployed. Spud DB also provides a comprehensive search page to data mine the potato genome through tools that query sequence identifiers, functional annotation, gene ontology (GO), InterPro domains, and basic local alignment search tool (BLAST) databases. Collectively, this resource links potato genomic data with phenotypic and genotypic data from a large collection of potato lines for use by the potato community, especially breeders and geneticists.

Published in The Plant Genome 7
© Crop Science Society of America
5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication
All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.


Abbreviations: BAC, bacterial artificial chromosome; BLAST, basic local alignment search tool; DAIT, diversity arrays technology; DM, Solanum tuberosum Group Phureja DMI-3 S16 R44; EST, expressed sequence tag; GO, gene ontology; HTML, HyperText Markup Language; ITAG, International Tomato Annotation Group; NCBI, National Center for Biotechnology Information; PGSC, Potato Genome Sequencing Consortium; PUT, PlantGDB-assembled unique transcript; RH, Solanum tuberosum Group Tuberosum RH89-039-16; RNA-Seq, ribonucleic acid sequencing; SolCAP, Solanaceae Coordinated Agricultural Project; SGN, Sol Genomics Network; SNP, single nucleotide polymorphism; SolCAP, Solanaceae Coordinated Agricultural Project; SSR, simple sequence repeat.
The potato genome serves as the central reference for Spud DB, with the most recent version of the potato pseudomolecules (v. 4.03, Sharma et al., 2013) used in the initial release. We have included both the PGSC (Potato Genome Sequencing Consortium, 2011) and the ITAG (Tomato Genome Sequencing Consortium, 2012) annotations of the potato genome throughout Spud DB (Table 1). One component of the genome browser is comparative analyses with other species as transcript assemblies and annotated gene models for 13 Solanaceae species and several dicotyledonous relatives, *Arabidopsis thaliana*, *Populus trichocarpa*, and *Vitis vinifera*, are included to extend structural and functional annotations of annotated potato genes (Table 1). We have also included genotypic and phenotypic datasets for the SolCAP Diversity Panel (Hirsch et al., 2013).

**Species, Sequences, Genotypic, and Phenotypic Datasets Available through Spud DB**

With respect to resources with a primary focus on potato, the PoMaMo database contains potato genetic maps and sequences (http://www.gabipd.org/projects/Pomamo/) and the Potato Pedigree Database houses pedigree information for potato cultivars (http://www.plantbreeding.wur.nl/potatopedigree/).

Following the release of the potato genome sequence in 2011, multiple large-scale genomic, genetic, and phenotypic datasets have been generated for potato (Felcher et al., 2012; Hamilton et al., 2011; Hirsch et al., 2013; Potato Genome Sequencing Consortium, 2011). To provide a centralized resource to mine these datasets, we constructed Spud DB (http://potato.plantbiology.msu.edu/), a potato-centric database that hosts the Potato Genome Sequencing Consortium (PGSC) genome sequence and associated annotation data and provides access to recently generated large-scale potato datasets. Information within Spud DB includes functional annotations of the potato genome, expression data from a set of 56 developmental stages and stress treatments, a genome browser that includes comparative data with other Solanaceae species in a potato-centric manner, and an array of genetic marker data. To facilitate breeding efforts in potato, we have developed a Breeder’s Assistant that allows users to access the SolCAP potato 8303 Infinium SNP array genotypic data in conjunction with phenotypic data. The Breeder’s Assistant also contains search tools to identify candidate SSRs and SNP markers for genes or regions of interest. Spud DB has been designed to be a user-friendly website to facilitate access to the potato genome as well as large-scale omic datasets to aid in advancing potato breeding and research.

**Resources, Tools, and Data within Spud DB**

The Spud DB web resource has a simple and intuitive layout based on the Biofuel Feedstock Genomics
Table 1. Summary of species and sequences used in comparative analyses of the potato genome.

<table>
<thead>
<tr>
<th>Species</th>
<th>Source¹</th>
<th>Version</th>
<th>No. of representative gene models</th>
<th>No. of transcript assemblies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>TAIR</td>
<td>TAIR10</td>
<td>27,416†</td>
<td></td>
</tr>
<tr>
<td>Capsicum annuum</td>
<td>PlantGDB</td>
<td>171a</td>
<td>–</td>
<td>29,507</td>
</tr>
<tr>
<td>Nicotiana benthamiana</td>
<td>NCBI</td>
<td>TSA</td>
<td>–</td>
<td>35,724</td>
</tr>
<tr>
<td>Nicotiana benthamiana</td>
<td>PlantGDB</td>
<td>173a</td>
<td>–</td>
<td>19,650</td>
</tr>
<tr>
<td>Nicotiana langsdorffii ×</td>
<td>PlantGDB</td>
<td>157a</td>
<td>–</td>
<td>5,120</td>
</tr>
<tr>
<td>Nicotiana sanderae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nicotiana sylvestris</td>
<td>PlantGDB</td>
<td>163a</td>
<td>3,651</td>
<td></td>
</tr>
<tr>
<td>Nicotiana tabacum</td>
<td>PlantGDB</td>
<td>173a</td>
<td>115,649</td>
<td></td>
</tr>
<tr>
<td>Petunia × hybrida</td>
<td>PlantGDB</td>
<td>159a</td>
<td>–</td>
<td>8,499</td>
</tr>
<tr>
<td>Physalis penovina</td>
<td>PlantGDB</td>
<td>187a</td>
<td>–</td>
<td>26,699</td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>Phytozome⁴</td>
<td>Assembly v. 2.2, Annotation v. 2.2</td>
<td>45,033</td>
<td>750,000</td>
</tr>
<tr>
<td>Solanum chacoense</td>
<td>PlantGDB</td>
<td>163a</td>
<td>–</td>
<td>2,634</td>
</tr>
<tr>
<td>Solanum habrochaites</td>
<td>PlantGDB</td>
<td>175a</td>
<td>–</td>
<td>10,779</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>SGN</td>
<td>Assembly SL2.40, Annotation ITAG2.3</td>
<td>34,727</td>
<td></td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>PlantGDB</td>
<td>171a</td>
<td>–</td>
<td>42,933</td>
</tr>
<tr>
<td>Solanum melongena</td>
<td>PlantGDB</td>
<td>175a</td>
<td>–</td>
<td>24,576</td>
</tr>
<tr>
<td>Solanum pennellii</td>
<td>PlantGDB</td>
<td>175a</td>
<td>–</td>
<td>4,251</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>PGSC</td>
<td>Assembly v. 4.03, Annotation v. 3.4</td>
<td>39,031</td>
<td></td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>ITAG</td>
<td>1</td>
<td>35,004</td>
<td></td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>PlantGDB</td>
<td>157a</td>
<td>–</td>
<td>65,456</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>Phytozome</td>
<td>12×</td>
<td>26,346</td>
<td></td>
</tr>
</tbody>
</table>

¹Total gene models excluding pseudogenes and transposable elements.

Breeder’s Assistant
To provide the potato community with easy and quick access to potato genotypic and phenotypic data, we developed the Breeder’s Assistant page. This page houses five separate tools to query and mine the available data (Fig. 1). Three of the tools provide access to genotypic and phenotypic data from the SolCAP Diversity Panel (Hirsch et al., 2013). The other two tools are focused on identifying candidate genetic markers in the form of SSRs and SNPs in the potato genome.

Data-Mining the SolCAP Diversity Panel
The SolCAP Diversity Panel (Hirsch et al., 2013) consists of 250 potato clones that represent cultivars, breeding lines, genetic stocks, and wild species. These lines were genotyped using the SolCAP potato 8303 Infinium SNP array (Felcher et al., 2012; Hamilton et al., 2011) and phenotyped for various agronomic traits including carbohydrate traits and tuber features (Hirsch et al., 2013). We developed tools to mine genotypic and/or phenotypic data based on the users’ needs. These three tools all employ selectable filters to examine all or only a subset of the SolCAP Diversity Panel based on market class, ploidy, breeding program, or accession name. When searching solely for phenotypic data, users select desired accessions and select filters to return all phenotypes, a subset of phenotypes, or a single phenotype. When utilizing the genotype tool, users are able to filter based on SNP identifier, SNP annotation, or SNP physical position. In GenomeStudio (Illumina, San Diego, CA), geno-type calls from Infinium SNP arrays can be outputted in 3-cluster or 5-cluster calls. A diploid model is assumed when using 3-cluster calls (AA, AB, BB) and a tetraploid model is assumed for 5-cluster calls to account for dosage (AAAA, AAAB, ABBB, BBBB). A filter has been made available to report genotypes using 3-cluster calls (cluster) or 5-cluster calls (dosage). When the tool is used to select both genotypic and phenotypic data, all of the filters are available. Provided in Fig. 1 is an example query for phenotypic and genotypic data from the SolCAP Diversity Panel using the Breeders Assistant.

SSR and SNP Marker Search Tools
In release v. 4.03 of the DM genome, a total of 190,589 mono-, di-, tri-, tetra-, penta-, and hexanucleotide SSRs were annotated using the SSRIT tool (Temnykh et al., 2001) (Fig. 2A) using a minimum repeat unit number of 10 for mononucleotide repeats, six for dinucleotides, and five for tri-, tetra-, penta-, and hexanucleotides. Examination of the nucleotide composition of the mono-, di-, and trinucleotide repeats in the genome revealed a large bias for A or T mononucleotides, AT dinucleotides, and AAT/ATT tri-nucleotides (Fig. 2B). We also characterized SSRs within the annotated PGSC and ITAG gene models (i.e., transcripts) and found variation in the number and prevalence of SSRs in the two annotation datasets. A search of the PGSC transcript sequences identified 7268 SSRs, with the most abundant being mononucleotide SSRs, whereas only

Resource (Childs et al., 2012), to provide an easy user experience. The Spud DB website consists of eight main components including the Home page, the Breeder’s Assistant, the Annotation Report pages, the Genome Browser, the Search Tools, the Download page, the Links page, and the Contact page. Each component is accessible from all pages for easy navigation. The supporting databases for Spud DB are based on the Biofuel Feedstock Genomics Resource (Childs et al., 2012) and utilize a modified Chado schema (Mungall and Emmert, 2001; Zhou et al., 2006), with an additional database to support keyword searches.

Home Page
The Home page contains an overview of the Spud DB resource and a quick search form to query by gene identifier or functional annotation keyword. The Home link in the dropdown menu bar also contains links to a News page containing updates for the website as well as a Frequently Asked Questions page for answers to common questions about the resource.
1829 SSRs were identified in the ITAG transcripts, with the most abundant being trinucleotide SSRs (Fig. 2A). The difference in SSRs between these two annotation sets is most likely attributable to the lack of annotated untranslated regions in the ITAG transcripts. Primer pairs to amplify each SSR marker were designed using Primer3 (Rozen and Skaletsky, 2000) and are available in the Breeder’s Assistant. The reported primers were chosen using preference towards smaller product sizes between 100 and 1000 nucleotides, primer length between 18 and 27 nucleotides, and primer melting temperatures between 57 and 63°C. A maximum of three primer pairs were reported for each SSR.

The Breeder’s Assistant SSR marker search tool searches for SSRs within the genome, the PGSC or ITAG potato annotations, and allows users to filter SSR search results based on motif, minimum and maximum length, and output format (HyperText Markup Language [HTML] or tab delimited text). Genome level SSRs can also be searched via genomic region. The returned data consist of transcript identifier, SSR type, length, start and end positions, as well as primer pairs to amplify the SSR, primer start, stop, length, and the estimated product size that would result from the primers. If HTML is the output format, and the SSR is within a transcript, the transcript identities link to their respective annotation report pages (see below). If HTML is the output format, and the SSR was searched at the genome level, the returned SSR identities link to their location on the genome browser (see below).

Single nucleotide polymorphisms present within the Breeder’s Assistant database include the 8303 Infinium SolCAP array SNPs (Felcher et al., 2012), 69,011 SNPs identified from transcriptome sequences of six cultivars (Hamilton et al., 2011), and 2754,111 SNPs identified by aligning whole genome shotgun reads from S. tuberosum Group Tuberosum RH89-039-16 (RH; van Os et al., 2006) to the v. 4.03 DM pseudomolecules (Table 2). For SNPs...
previously identified from six cultivated potato transcriptomes and the 8303 SolCAP Infinium array (Felcher et al., 2012; Hamilton et al., 2011), the SNPs (with their respective context sequence) were mapped to the PGSC v. 4.03 pseudomolecules using Exonerate est2 genome (Slater and Birney, 2005). The SNPs were retained if they fit the following criteria: alignments had to have >95% coverage and identity, contain no insertions or deletions, and have two or fewer alignments to the pseudomolecules. For genomic SNPs, total DNA from the heterozygous diploid RH was sheared to 270 bp, ligated to Illumina TruSeq adaptors, and 100 nucleotide reads generated on the Illumina HiSeq 2000 platform (Illumina, San Diego, CA; SRP019978). Cleaned reads were aligned to the DM v. 4.03 pseudomolecules and superscaffolds using Bowtie v. 0.12.7 (Langmead et al., 2009) and variant positions were filtered using SAMtools v. 0.12.7 (Li et al., 2009), requiring a minimum read depth of five, a maximum read depth of 250, and a maximum of three SNPs in a 100 nucleotide window (to filter dense SNPs).

The SNP finder tool in the Breeder’s Assistant allows SNP searching based on genome location, annotated locus, or transcript identifier with both the PGSC and ITAG transcript identifiers. Returned data consist of a sortable table with SNP names, the chromosome and position of the SNP, the alleles, and the origin of the SNP. All SNP names and any transcript identifiers are linked to their genome browser location to provide the genome context of the SNP. Collectively, the tools within the Breeder’s Assistant allow breeders and other researchers access to genotypic, phenotypic, and marker data easily and in a custom manner, leading to quicker and more informative research for potato cultivar improvement.

Table 2. Summary of genetic markers, expression data, and comparative analyses of the potato genome available within Spud DB.

<table>
<thead>
<tr>
<th>Feature†</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic markers</td>
<td></td>
</tr>
<tr>
<td>SolCAP 8303 SNPs</td>
<td>8,040</td>
</tr>
<tr>
<td>Transcript-derived SNPs</td>
<td>66,498</td>
</tr>
<tr>
<td>WGS SNPs</td>
<td>275,111</td>
</tr>
<tr>
<td>Whole genome SSRs</td>
<td>190,589</td>
</tr>
<tr>
<td>Expression data</td>
<td></td>
</tr>
<tr>
<td>Number of RNA libraries</td>
<td>56</td>
</tr>
<tr>
<td>Genes expressed in RNA libraries</td>
<td>33,133</td>
</tr>
<tr>
<td>Comparative alignments</td>
<td></td>
</tr>
<tr>
<td>Anchored BAC end sequences</td>
<td>202,310</td>
</tr>
<tr>
<td>Species with comparative alignments</td>
<td>15</td>
</tr>
</tbody>
</table>

†BAC, bacterial artificial chromosome; SolCAP, Solanaceae coordinated agricultural project; SNP, single nucleotide polymorphism; SSR, simple sequence repeat; WGS, whole genome shotgun.

Figure 2. Simple Sequence Repeat (SSR) data accessible within Spud DB. (A) The percentage of each SSR motif computationally predicted at the genome level and at the transcript level in the PGSC (Potato Genome Sequence Consortium) and ITAG (International Tomato Annotation Group) annotation datasets. The total number of SSRs predicted for each annotation is listed above the columns. (B) The percentage of each type of SSR within mono-, di-, and trinucleotide motifs predicted at the genome level. The type and total number of each type is listed for the three motifs.
Annotation Report Pages

Researchers are highly focused on genes and the proteins encoded in the annotated gene models/transcripts. Both the PGSC and ITAG projects annotated the potato genome for genes; however, the level of functional annotation provided by both groups is limited. Thus, to facilitate discovery and interpretation of the potato genome sequence, we have further functionally annotated the PGSC and ITAG potato gene models and provide these data in Spud DB through individual Annotation Report pages. Each Annotation Report page not only provides a centralized resource for the sequence and functional annotation of each gene model, transcript, and/or predicted protein, but also provides additional annotation in the form of expression profiles, polymorphisms, candidate genetic markers (SSRs and SNPs), and comparative alignments to other Solanaceae and dicotyledonous species to aid in functional interpretation (Fig. 3).

Functional Annotation and Comparative Analyses of Annotated Potato Genes

Potato predicted protein sequences from the PGSC and ITAG annotations of the potato genome were subjected to protein domain and structural analysis using InterProScan (Mulder et al., 2007). Within InterProScan, analyses were performed using BlastProDom, FPrintScan, Gene3D, HAMAP, HMMPanther, HMMPfam, HMMPIR, HMMSmart, HMMTigr, SuperFamily, PatternScan, ProfileScan, TMHMM, SignalP, seg, and coils programs. Several of the protein domain databases contain relevant GO terms (Camon et al., 2004) and those data were noted and associated with the relevant potato sequences. A total of 69,823 GO associations were made to 13,658 PGSC genes and 16,067 ITAG genes with a similar distribution of associations between the two annotation datasets (Fig. 4). Functional descriptions were retained from the original genome projects, curated by the PGSC or ITAG, and were not recomputed. On the Annotation Report pages, we have included BLAST search results from an all-versus-all search of the PGSC and ITAG annotated proteomes to permit cross-referencing of equivalent loci, including self hits. As insight in determining gene function can be gained through sequence comparisons, all PGSC and ITAG annotated predicted proteins were searched against the A. thali-ana predicted proteins and the ITAG annotated tomato
predicted proteins using BLASTP. These matches are linked to the \textit{A. thaliana} gene pages at the Arabidopsis Information Resource (Swarbreck et al., 2008) and the tomato gene pages at SGN (Bombarely et al., 2011).

**Expression Profiles**

Previously published high throughput RNA-Seq data were used to generate expression profiles for the PGSC annotated gene set. In total, 56 libraries were surveyed, including 40 from DM potato and 16 from RH potato, representing >15 different tissue types (Potato Genome Sequencing Consortium, 2011). Read quality was determined using FastQC v. 0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and adaptor and quality trimming was performed using Cutadapt v. 1.1 (Martin, 2011). Reads with quality scores of 10 or below at the 3' end were trimmed, and quality and adaptor trimmed reads that were <30 nucleotides were removed from the analysis. The reads were mapped to the v. 4.03 DM pseudomolecules using Bowtie v. 0.12.7 (Langmead et al., 2009) and TopHat v. 1.4.1 (Trapnell et al., 2009), with a single mismatch permitted for DM-derived reads, two mismatches for RH-derived reads, and a permitted intron length of 10 to 15,000 bp. Gene expression values (fragments per kilobase exon model per million mapped reads) were calculated using Cufflinks v. 1.3.0 (Trapnell et al., 2010) with the PGSC gene annotation for gene modeling and allowing a maximum intron length of 15,000 bp. By using a variety of tissues for the transcriptome analysis, we were able to report expression for 33,133 genes from the v. 4.03 DM pseudomolecules, respectively (Table 2).

**Gene Model, Transcript, and Predicted Protein Search Tools**

Users can retrieve information within Spud DB by using the Search Tools page (Fig. 5). Investigators are able to perform BLAST alignments of their own sequences to search the potato pseudomolecules, superscaffolds, predicted genes, annotated gene models and/or transcripts, or predicted proteins. Additional datasets of interest including the potato organellar genomes, tomato genome assembly, annotated loci, transcripts, and predicted proteome, Solanaceae assembled transcripts, and \textit{A. thaliana} loci, transcripts, coding sequence, and predicted proteome are provided to facilitate comparative analyses within the Solanaceae and with the model species, \textit{A. thaliana}. The type of BLAST to conduct, as well as several modifiable parameters, are available for user selection to optimize the search.

Another useful feature within the Search Tools is the sequence retriever. This tool allows users to obtain potato v. 4.03 sequences based on different search criteria. Users can retrieve sequences by querying chromosome and start and stop positions (maximum region size of 100,000 bp) or by searching by PGSC potato gene or transcript identifier. In addition, users are able to retrieve sequences upstream or downstream of PGSC genes with a maximum region size of 50,000 bp.

Users are also able to query based on sequence identifier, functional annotation keyword, GO identifier and keyword, and InterPro identifier and keyword. In all cases, searches can be performed with either the PGSC and/or ITAG annotations (Fig. 5).
Figure 5. The options available to query and mine potato data within the Search Tools page in Spud DB. The main category headers are shown in dark green boxes with white text, user input parameters are in light green boxes with black text, and outputs are shown in the blue boxes. (A) Basic local alignment search tools (BLASTs) search options. (B) Options to retrieve sequence information, the maximum region size for an inputted genomic region is <100,000 bp, whereas the maximum region size for queries up/downstream of a gene or transcript is <50,000 bp. (C) Search tools resulting in links to transcript annotation report pages. An example annotation report page is shown, whereas Fig. 3 is a more detailed depiction of an example annotation report page.
Genome Browser

Potato Genome Browser

To support graphical views of the potato genome, as well as to permit querying of the potato genome, a browser containing the most recent release of the DM potato pseudomolecules, v. 4.03, is available in Spud DB along with 94 tracks of annotation (Fig. 6). The browser is based on the Generic Genome Browser v. 1.70 (Stein et al., 2002) and is served by a MySQL database using the Bio::DB::SeqFeature::Store adaptor. The browser contains both the PGSC v. 3.4 gene models (Potato Genome Sequencing Consortium, 2011), as well as the gene models from the ITAG annotation of the potato genome (Tomato Genome Sequencing Consortium, 2012). Each locus, as well as annotated gene models, is linked to their cognate Annotation Report page to provide users with detailed information on each gene and gene model.

To provide comparative information on each annotated gene, the PlantGDB putative unique transcripts [PUTs; (Duvick et al., 2008)] from 12 Solanaceae species and shotgun transcript assemblies from Nicotiana benthamiana downloaded from the National Center for Biotechnology Information (NCBI) were aligned to the potato pseudomolecules using Exonerate v. 2.2.0 (Slater and Birney, 2005), with the highest scoring alignment for each transcript sequence displayed on the browser. The N. benthamiana (Bombarely et al., 2012) and ITAG Tomato Gene Models [v. 2.3, (Tomato Genome Sequencing Consortium, 2012)] were aligned to the PGSC v. 4.03 pseudomolecules with Exonerate est2 genome (Slater and Birney, 2005), and the highest scoring alignment for each gene model was reported.

Bacterial artificial chromosome (BAC) end sequences of DM and RH were downloaded from the NCBI (http://www.ncbi.nlm.nih.gov/nucgss). The BAC end sequences were searched against the PGSC v. 4.03 pseudomolecules using BLASTN. The results for the DM BAC ends were filtered to retain hits with an E-value < 1e−10, a minimum identity of 97%, and a minimum coverage of 90%, whereas RH BAC alignments with an E-value < 1e−10, a minimum of 95% identity, and 90% coverage were retained. After the initial filtering, only the unique top hit for each BAC end sequence was retained. If multiple top hits existed, the sequence was not reported within Spud DB. Using this filtering we were able to anchor 202,310 (120,101 DM and 82,209 RH) BAC end sequences to v. 4.03 (Table 2).

The aforementioned RNA-Seq read mapping results were used to create tracks displaying heat maps for the 56 potato transcriptome libraries to view the concordance of transcript evidence with gene models, and to visually assess transcript abundances. This was done using the wiggles tool, available within TopHat v. 1.4.1 (Trapnell et al., 2009).

The repeat track was generated for the PGSC v. 4.03 pseudomolecules using RepeatMasker v. 4.0.1 (Chen, 2004) with a custom potato repeat library provided by the PGSC (Potato Genome Sequencing Consortium, 2011). The SSRs, Infinium SolCAP array SNPs, transcriptome sequencing-derived SNPs, and RH SNPs described above are provided as tracks on the genome browser. Additional marker datasets available on the potato v. 4.03 pseudomolecule genome browser include a set of unambiguously mapped potato diversity array technology (DArT) markers (Sharma et al., 2013) and SNPs used to develop an oligo-nucleotide pooled assay (Sharma et al., 2013).

Supporting Browsers

Although the v. 4.03 release is the most current release of the potato genome, many researchers in the potato community continue to use the v. 2.1.11 release, as it was available when a number of ongoing genome-based research projects began. To allow users continued access to this legacy version of the pseudomolecules, we have maintained a v. 2.1.11 genome browser, which is selectable on the Potato Genome Browser via a toggle.

The potato and tomato genomes are highly syntenic ( Tanksley et al., 1992), and to support comparative analyses between these two species, we constructed a tomato genome browser containing the annotated v. 2.3 ITAG tomato genome models (Tomato Genome Sequencing Consortium, 2012) and aligned the potato gene models from both the PGSC and ITAG to the tomato genome to provide a comparative framework to interrogate potato genes of interest. As with potato, the tomato genome browser is based on the Generic Genome Browser v. 1.70 (Stein et al., 2002) and served by a MySQL database using the Bio::DB::SeqFeature::Store adaptor. The Solanaceae PUTs, N. benthamiana transcriptome shotgun assemblies, and N. benthamiana, A. thaliana, P. trichocarpa, and V. vinifera gene models were aligned to the tomato genome as described above to provide additional comparative data. The repeat track for the tomato pseudomolecules was also generated with RepeatMasker using the tomato repeat library downloaded from SGN (v. 5, ftp://ftp.solgenomics.net/tomato_genome/repeats/repeats_master.v5.fasta.gz). Additional tracks on the tomato genome browser include alignment of BAC end sequences from the tomato genome project to aid in assessing the quality of the assembly, and alignment of SNPs identified from the SolCAP tomato Infinium array (Hamilton et al., 2012) and from the tomato genome project (Tomato Genome Sequencing Consortium, 2012).

Download Page

Data available to researchers can be downloaded from two sources, an FTP site and through the official PGSC Data Release page. The FTP site provides access to flat files of SNP data as well as sequences and functional annotation of all species covered in Spud DB including the set of filtered mRNA and translated sequences for the species with PUT sequences, and the transcript and peptide sequences for A. thaliana, potato, and tomato. The PGSC public data download page provides links for the v. 2.1.11 and v. 4.03 DM pseudomolecules, annotation, RNA-Seq based gene expression data, potato BAC sequences, and the potato DArT marker data.
Figure 6. A graphical representation of selected tracks on the v. 4.03 DM genome browser for PGSC0003DMT400076546. From top to bottom, the first track shows the chromosome location followed by the Potato Genome Sequence Consortium (PGSC) pseudomolecule tiling path, loci, and the PGSC and International Tomato Annotation Group (ITAG) representative gene models. The next four tracks show related Solanaceae species sequence alignments with the ITAG tomato (Solanum lycopersicum) gene model, and PlantGDB-assembled unique transcript (PUT) alignments from Nicotiana benthamiana, Nicotiana sylvestris, and Physalis peruviana. Following these tracks are four marker tracks, SNP information for Infinium high confidence SNPs (Hamilton et al. 2011), SNPs used on the potato 8303 Infinium SNP array (Felcher et al. 2012), SNPs called by genomic sequence comparison with RH (this study), and diversity arrays technology (DArT) markers (Sharma et al. 2013). The last four tracks are heat maps of RNA-Seq expression data from flower, leaf, mature tuber, and petal tissue. The brackets on the far right indicate where the tracks are linked or where their information can be found within Spud DB; blue, are tracks that link to the annotation report page within Spud DB; red, a track that links to the Sol Genomics Network (SGN) search results for the gene; green, are tracks that link to the PlantGDB PUT sequence page; brown, not direct links, but brackets identify information available on the annotation report page for the transcript.
Table 3. Metrics of Spud DB access from 9/1/2013 through 11/25/2013.

<table>
<thead>
<tr>
<th>Metric</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total visits</td>
<td>9,193</td>
</tr>
<tr>
<td>Unique visitors</td>
<td>3,998</td>
</tr>
<tr>
<td>Total pages viewed</td>
<td>60,196</td>
</tr>
<tr>
<td>v. 4.03 Potato Browsers pages viewed</td>
<td>27,051</td>
</tr>
<tr>
<td>Pages viewed per visit</td>
<td>6.55</td>
</tr>
<tr>
<td>Average duration of visit in minutes</td>
<td>8.30</td>
</tr>
</tbody>
</table>

Links and Contact Pages

The Links page contains numerous links to outside web sources of interest for potato researchers. The outside sources are split into categories of potato research areas, including agriculture and pathogen research, genome projects, cultivar information, organizations, and other Solanaceae-based sites. We also provide a Contact page for users to provide comments, questions, and suggestions for Spud DB.

Use and Future Plans

Access Metrics

Spud DB was released in September 2013 and has been substantially used by the community (Table 3). Nearly 4000 unique visitors have viewed over 60,000 pages during a 3-mo period. The potato v. 4.03 genome browser was the most viewed resource on Spud DB, accounting for nearly half of the total page views.

Future Plans

We will update Spud DB content as new datasets are made available. Of special interest are genotypic and phenotypic datasets for potato. A large number of the 8303 SolCAP Infinium arrays have been purchased by the community (D. Douches, personal communication, 2013) and public availability of these datasets would provide a great resource for the community. We anticipate new Solanaceae genomes and transcriptomes to be made available in the near future, as sequencing advances have enabled a wide scope of access to genome and transcriptome sequencing. Alignment of these transcripts and annotated gene models to the potato genome will provide a richer set of comparative resources to aid in interpretation of the potato genome.

Conclusions

We have developed a potato-centric resource for easy access of potato functional genomics data by the community. The resource also allows quick and user-customizable access to phenotypic and genotypic data from a potato diversity panel within the Breeder’s Assistant page. The datasets within the resource include potato genomic, transcript, and protein sequences, SNP and SSR marker data, expression data, and numerous search tools based on BLAST, functional annotations, GO annotations, and InterPro annotations. The resource is broadly accessible to many researchers in the fields of breeding, functional genomics, and basic potato research.

Acknowledgments

This work was supported by grants from the U.S. Department of Agriculture National Institute of Food & Agriculture (Grant No. 2008-35300-18671) and the National Science Foundation (Grant No. DBI-0604907 and DBI-0834044 to CRB). CDH was supported by a National Science Foundation National Plant Genome Initiative Postdoctoral Fellowship in Biology Fellowship (Grant No. 1202724).

References


