Published online 8 October 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D1149–D1155***
doi:10.1093/nar/gkq866

The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl

Aureliano Bombarely, Naama Menda, Isaak Y. Tecle, Robert M. Buels, Susan Strickler, Thomas Fischer-York, Anuradha Pujar, Jonathan Leto, Joseph Gosselin and Lukas A. Mueller*

Boyce Thompson Institute for Plant Research, Tower Road, Ithaca, NY 14853, USA

Received August 19, 2010; Accepted September 13, 2010

ABSTRACT

The Sol Genomics Network (SGN; http://solgenomics.net/) is a clade-oriented database (COD) containing biological data for species in the Solanaceae and their close relatives, with data types ranging from chromosomes and genes to phenotypes and accessions. SGN hosts several genome maps and sequences, including a pre-release of the tomato (*Solanum lycopersicum* cv Heinz 1706) reference genome. A new transcriptome component has been added to store RNA-seq and microarray data. SGN is also an open source software project, continuously developing and improving a complex system for storing, integrating and analyzing data. All code and development work is publicly visible on GitHub (http://github.com). The database architecture combines SGN-specific schemas and the community-developed Chado schema (http://gmod.org/wiki/Chado) for compatibility with other genome databases. The SGN curation model is community-driven, allowing researchers to add and edit information using simple web tools. Currently, over a hundred community annotators help curate the database. SGN can be accessed at http://solgenomics.net/.

INTRODUCTION

The Solanaceae, also known as the Nightshades, are a flowering plant family with important crop species such as potato, tomato and eggplant. The Solanaceae have a unique biology, with highly conserved genomes, yet extraordinarily diverse phenotypes and specialized adaptations. Thus, highly comparative approaches of genome study within the Solanaceae seem likely to yield important discoveries. As a step in this direction, several Solanaceae genomes are currently being sequenced, including a high-quality tomato reference sequence (1), several wild tomato species and a wild potato species (*Solanum phureja*). The Sol Genomics Network (SGN; http://solgenomics.net/) integrates this information in a clade-oriented database (COD), containing genomic, genetic, transcriptomic, phenotypic and taxonomic information with the data of major Euasterid families such as the Solanaceae (tomato, potato, eggplant, pepper and petunia), Plantaginaceae (snapdragon) and Rubiaceae (coffee).

Due to rapid progress in the development of new scientific methods, the database design needs to be constantly adapted and revised to accommodate the ever larger information. Over the last few years, genomics has undergone a significant transformation based on new sequencing technologies that can generate millions of sequences in a single run (2–6), enabling fast and low cost sequencing even of complex genomes and transcriptomes. The speed of sequencing and the resulting amount of sequence data poses novel challenges on how these data can be stored efficiently and presented to the research community.

Whereas model organism databases (MODs) such as the yeast database [saccharomyces genome database (SGD); http://yeastgenome.org/] (7) or Arabidopsis [the arabidopsis information resource (TAIR); http://www.arabidopsis.org/](8) can rely on a large staff of in-house curators who extract relevant information from the literature, providing their databases with deeper, richer information on genes and other data types, this approach is not scalable to CODs, which hold multiple species. Therefore, SGN has developed a powerful and easy to use community-based annotation system that uses a mixed approach of 'trusted users' in which SGN curators

*To whom correspondence should be addressed. Tel: +1 607 255 6557; Fax: +1 607 254 1242; Email: lam87@cornell.edu

assign editor privileges to members of the research community who are gene experts, as evidenced by a publication or a meeting presentation (9). SGN is one of the largest community curated database and an open source project, such that both the database content and the code driving it can be advanced by the respective communities.

In this paper, we examine SGN from three perspectives: tools and data, technology and community participation.

TOOLS AND DATA

The SGN database hosts a wide range of biological data for various species and accessions in the Solanaceae (Figure 1), from genomic sequences to phenotype images. This data originates from a variety of sources including user submissions and data from other curated public databases (Figure 2).

Transcriptomes

For many years, SGN has collected expressed sequence tag (EST) sequences from many sources such as user submissions or public databases and has processed and assembled them into unigene builds, which are annotated using sequence homology and predicted protein domains and then grouped into gene families. Currently there are 14 unigene builds for 18 species with more than 270 000 member sequences used in the assemblies.

New sequencing technologies have made it necessary to change how individual reads are processed, stored and presented online. In the last year, seven Solanaceae

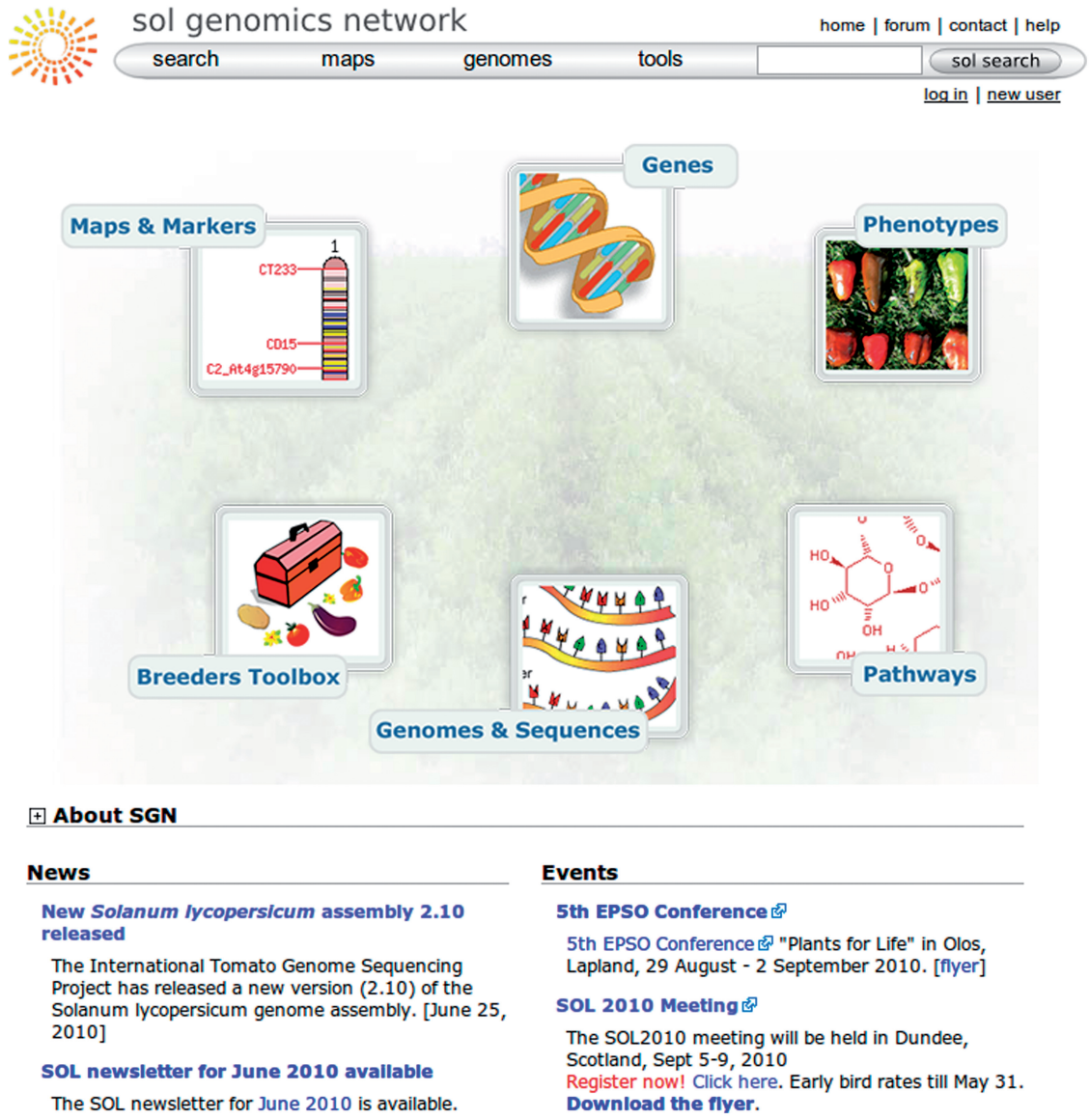

Figure 1. The home page of the SGN. The home page is the main entry page, providing quick access to resources through graphical menus. Every SGN page consistently contains the same toolbar at the top with pull-down menus and links to login and help pages. On the lower part of the home page, the news and events sections keep the community informed and certain elements of the database are highlighted in different feature topics, such as a 'locus of the week'. Links to other important resources are also provided.

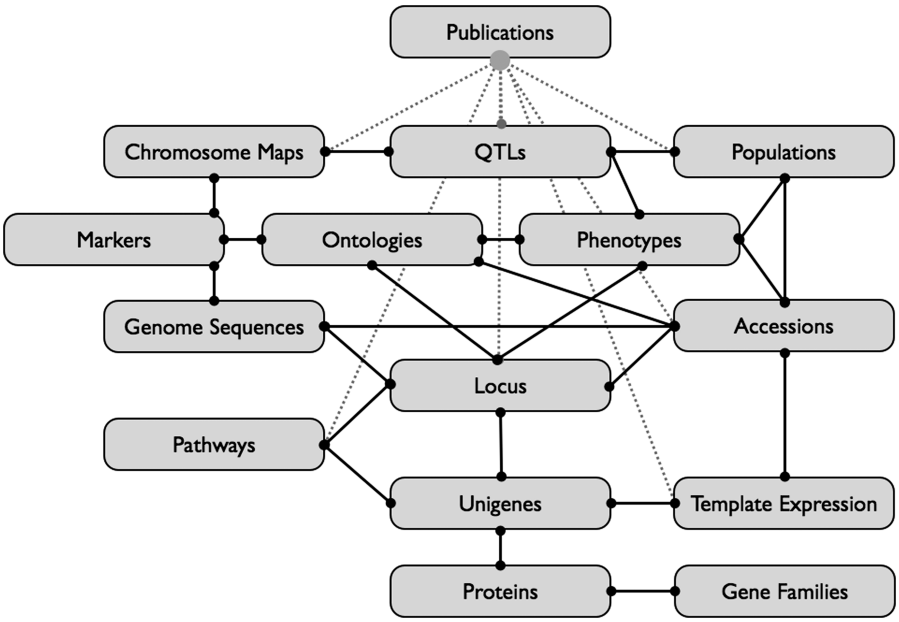

Figure 2. SGN data type relationship diagram, in which the locus data type is a central node, from which most data on SGN data can be accessed with a few clicks. Other important data types include sequences and phenotypes.

RNA-seq datasets have been submitted to the sequence read archive (SRA) database (10) and many more will follow. To better process these larger data sets, a new SGN transcriptome system has been designed and implemented, which uses a hybrid approach for data storage: (i) unigenes, assembly protocol and filenames on the one hand are stored in a relational database, using the Chado schema (11) and sample tables from the SGN biosource component (unpublished); (ii) original reads and assembly data are stored in the filesystem using indexed standard formats such as FASTA and GFF3. This provides enhanced scalability while preserving seamless integration in the user interface.

Next generation sequencing (NGS) data can also be mined for expression and single-nucleotide polymorphisms (SNP) information. Although there are general expression databases like gene expression omnibus (GEO) (12) or ArrayExpress (13) that contain expression data from microarray and RNA-seq analysis, expression information is an important resource for researchers that should be tightly integrated with other data. To this end, SGN has developed a new expression component called the general expression module (GEM), which stores and displays expression data from different technologies such as microarrays and RNA-seq that are directly incorporated into the SGN relational database. Expression data is generated from the NGS data by first associating the expression values with the sequence assembly and loaded the results into the GEM database where they are visible and searchable on SGN.

We have focused on integrating high quality Affymetrix-based expression data (14) for species such as tobacco (15) and tomato (16), for which more than 43 conditions and 167 hybridizations have been loaded. Expression data associated with RNA-seq from 7 tomato trichomes experiments (17) have been loaded. Currently, the system provides a simple interface for searching, querying and visualizing expression data and more advanced functionality and graphical views will be provided in the near future. In addition, plant ontology (PO) annotations (18) can be associated with samples in the expression data to improve downstream sequence annotation.

Protein datasets

Most MODs contain little or no proteomics data, although a number of specialized protein databases such as Pride (19) or the ExPaSy proteomics server (20) exist. Species-specific proteomic databases have also been developed, e.g. the Nottingham arabidopsis stock centre (NASC) proteomics database (21). Usually, such databases collect proteins from different sources like Swiss-Prot (22), protein information resource (PIR) (23) or protein data bank (PDB) (24) and/or predicted proteins from gene models or messenger ribonucleic acid (mRNA) datasets.

In SGN, protein sequences are derived from protein predictions on both unigene transcript sequences and predicted genomic gene models, such as those from the international tomato annotation group (ITAG). For unigenes, several different methods for protein prediction are used, including the analysis of the longest open reading frame (ORF), detection of the coding region using hidden Markov models [using EstScan (25)], or detection of the most probable translation initiation site using NetStart (26). Sequences are stored in a Chado relational database schema and in bulk FASTA files available for download via file transfer protocol (FTP). In addition to standard homology searches using basic local alignment search tool (BLAST) and similar tools, Mascot- (27) and Protein Pilot- (Applied Biosystems) compatible FASTA datasets are generated and published on SGNs FTP site.

Currently SGN hosts 28 unigene-based protein datasets for various species and one gene-model-based protein dataset for tomato from the ITAG. These datasets also contain domain annotations based on InterProScan (28) and signal peptide analysis using SignalP(29).

Gene family analyses are performed on these protein datasets using an SGN-developed pipeline (http://solgenomics.net/about/family_analysis.pl). This pipeline pre-clusters sequences via coarse homology searches with BLAST and then uses TRIBE-MCL (30) for final gene family clustering. Multiple alignments of the sequences in each cluster are performed with Muscle (31) and PAUP (32) is used for the calculation of the phylogenetic trees. The results are loaded into the SGN database and can be searched and visualized on the web. Family sequence alignments can be loaded into the Tree Browser tool to explore the relations between different family members. There are 11 851 gene families stored in the database for the last analysis using eight different species, with families ranging in sizes between 2 and 647 members.

Genomics and genetics

The tomato genome project has been actively sequencing tomato using a BAC-by-BAC approach since 2004, with sequenced BACs continually released on SGN (1). In 2009, a whole-genome-shotgun component was added to complement the BAC-by-BAC sequencing, with both being merged to form a finished assembly. This assembly is being continually refined and corrected with new assembly versions released frequently. The first pre-release of the *S. lycopersicum* whole-genome shotgun was published on 1 December 2009 as a common effort of the international tomato genome sequencing consortium. The previously-sequenced BAC sequences have been incorporated in the assembly covering the 12 tomato chromosomes with 91 scaffolds (release version 2.30) (http://solgenomics.net/genomes/Solanum_lycopersicum/index.pl).

SGN is also involved in the annotation of the *S. lycopersicum* genome as part of the ITAG. The ITAG group has created a distributed annotation pipeline, where each group runs a part of the analysis (http://www.ab.wur.nl/TomatoWiki) (1). Other genome sequences are hosted at the SGN database, such as *S. phureja* (wild potato) or *S. pimpinellifolium* (wild tomato species; http://solgenomics.net/genomes/Solanum_pimpinellifolium/), soon to be followed by new genome sequences (http://solgenomics.net/static_content/solanaceae-project/docs/SOL_newsletter_Jun_10.pdf), e.g. *S. pennellii*, a wild tomato species.

SGN stores genome sequences and annotations using two different approaches. It stores genomic elements both in a Chado relational schema (11) and as GFF3 (http://sequenceontology.org/resources/gff3.html) and FASTA bulk files are available for download. This combination allows tight integration of genomic features with other elements of the SGN database, such as markers and unigenes, but also allows straightforward integration of tools like BLAST (33) for sequence searches based on homology or GBrowse to visualize genome regions and their annotations (34).

Complementing the genome sequence, SGN hosts more than 20 genetic and physical maps for tomato (35), potato (36), pepper and tobacco (37) with thousands of markers. Genetic marker types in the database include AFLP, CAPS, PCR, RFLP, SNP, SSR and dCAPS. Genetic and physical maps are stored in a custom schema and can be accessed from the SGN toolbar or using different tools, including database searches, BLAST (33) or the SGN comparative viewer (38). SGN has also developed tools to facilitate the design of genetic markers, such as the CAPS designer (http://solgenomics.net/tools/caps_designer/caps_input.pl).

Metabolic pathways

Another important component of SGN is the annotation and cataloguing of genes involved in metabolic pathways. SolCyc is a Pathway Genome Database (PGDB) for Solanaceae species, such as tomato, potato, tobacco, pepper, eggplant, petunia, and close relatives, such as coffee (http://solcyc.solgenomics.net/). Currently, SolCyc comprises 7 PGDBs with approximately 1250 pathways, 6200 enzymatic reactions, 8600 enzymes and 4900 compounds for seven different species. SolCyc is based on the pathway tools software suite (39).

Phenomics

One of the most important problems of the post-genomic era is linking sequences to phenotypes. To solve this problem, generation of vast amounts of sequence data must be accompanied by a corresponding amount of phenotypic data for hundreds or thousands of accessions and mutants. SGN has developed an infrastructure for storing, displaying and curating phenotypic data called Phenome, which heavily relies on elements of the Chado schema, and makes significant use of Javascript/JSON(9) to provide a dynamic and responsive user interface.

Phenotypic data can be linked to loci, alleles, accessions, ontology annotations, publications and populations, with loci acting as the central data type linking phenomic and genomic data. A locus can have different alleles responsible for different phenotypes in a given accession or group of accessions. Accessions are also grouped into plant populations, for example, quantitative trait loci (QTL) or mapping populations. A trait is a phenotypic character analyzed in some population to study the distribution among the accessions. Currently SGN contains information on 7100 alleles, 5800 loci, 8200 accessions and 20 populations.

An ontology has been developed over several years to describe the phenotypic traits of the Solanaceae (Solanaceae phenotype ontology, SPO http://solgenomics.net/chado/cvterm.pl?action=view&cvterm_id=23057) with an emphasis on usability by both the scientific and breeder communities. The ontology currently contains about 200 terms and more terms are added as needed. Terms are mapped whenever applicable to the standard PO (18), as well as the phenotype and trait

ontology (PATO, http://obofoundry.org/wiki/index.php/PATO:Main_Page).

A web-based quantitative trait locus (QTL) analysis tool (http://solgenomics.net/qtl) based on R/QTL (40) has been developed for mapping QTLs, experimental crosses and cross-linking putative QTLs to relevant genomic, genetic and expression datasets in SGN (manuscript in preparation). There are currently three QTL populations with more than 40 different quantitative traits stored in the SGN database (41). Users can upload and analyze their own data on the fly and decide whether their data should be publicly visible or kept private.

TECHNOLOGY

From the user's perspective, SGN is a COD containing biological data for Solanaceae and related species, but from a technical perspective, it is a highly complex system for integrating diverse data, standard tools and custom code (Figure 3), written primarily in the Perl programming language. All software source code and daily development logs are publicly viewable at http://github.com/solgenomics/.

Like many websites, SGN is implemented as a three-tiered architecture consisting of user-facing view code, control and data modules and a relational database backend. The site runs on a complete modern Perl software stack: Mason, Catalyst and DBIx::Class, with adaptors providing support for legacy CGI and custom-SQL code. The relational database is PostgreSQL (42). Flat files are used for some storage purposes to complement the relational database, e.g. for storing large assemblies, images and sequence sets.

Many of the site's core functions are provided by generic model organism database (GMOD) tools (http://gmod.org), such as Chado, Bio::Chado::Schema and GBrowse, leaving SGN developers free to integrate more data and custom tools. Currently, over 600 SGN-developed Perl classes underlie the site, with a rich data model (Figure 2), but the codebase is made more concise and powerful by contributing to and integrating code from many open source projects, including GBrowse, Chado, BioPerl, DBIx::Class, Catalyst and Moose.

To produce the site software, SGN uses an Agile software development process, incorporating test-driven development (TDD) (43) and continuous integration (44). Under TDD, detailed test programs are written for each aspect of the system's function and tied together to be run easily, usually many times per day. Consistent adherence to TDD is a powerful tool for accelerating development, since the tests immediately pinpoint most problems, thus greatly reducing debugging time. This increases efficiency by allowing it to produce new site features and open-source software quickly and with high quality.

COMMUNITY PARTICIPATION

Traditional data curation demands tremendous personnel resources for a database. For example, in just tomato research, more than 800 gene-related articles were published in 2009 according to PubMed (http://www.ncbi.nlm.nih.gov/pubmed/). Most databases do not have enough curators to keep the site up-to-date in the face of such overwhelming numbers of publications. Therefore, SGN has developed a community curation

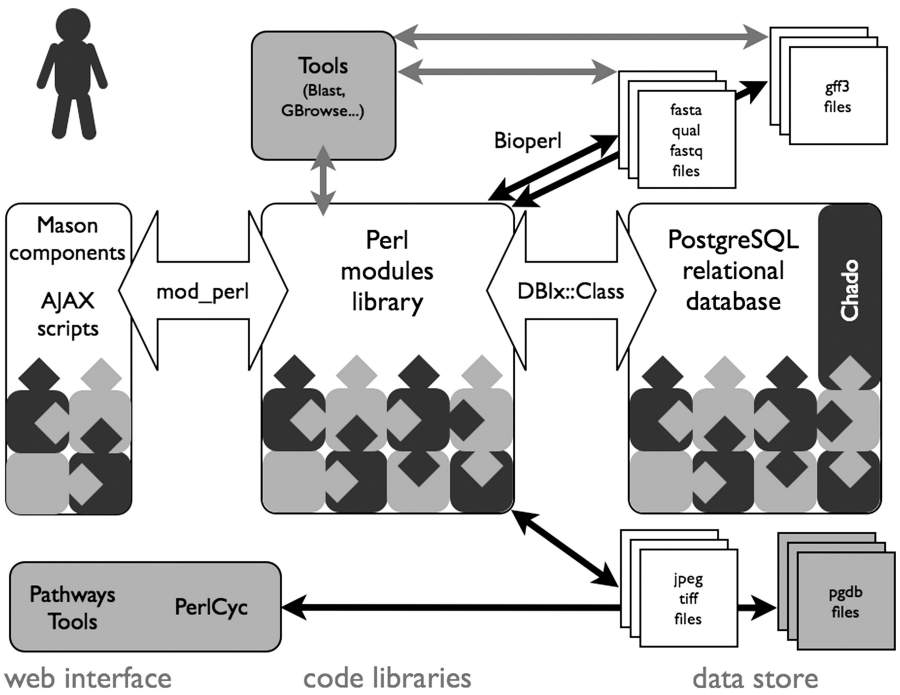

Figure 3. SGN system architecture diagram. SGN is a three-tiered system, consisting of a front-end web interface, back-end code and a data store, which includes both files and a relational database. For example, the GEM component is composed of Javascript and Mason components to create the user-facing web interface, DBIx::Class-based Perl modules to manipulate and model the data and a relational database schema for storage.

system under which certain genes and phenotypes can be curated by domain experts called 'locus editors' who have privileges to add, edit and remove information on genes and phenotypes for a certain gene. Locus editors are chosen by SGN curators, who invite researchers to become locus editors based on journal articles or on meeting presentations (9). As of July 2010, there are over 100 locus editors curating 261 loci.

Community annotations provide important high quality information from experts in their field for both genes and phenotypes, but it also provides a dynamic social network where researchers of different disciplines can meet, share their work and continuously submit updates on their genes of interest.

Other resources provided for SGN users community are database help (http://solgenomics.net/help/index.pl) and SGN tools tutorials such as the community annotation tutorial (http://www.slideshare.net/nm249/sgn-community-annotation-tutorial?type=presentation). SGN also supplies other 'social tools' such as email lists for news announcements (http://rubisco.sgn.cornell.edu/mailman/listinfo/sgn-announce/) and an SGN blog (http://solgenomics.blogspot.com/) where data, community or code topics are discussed.

FUTURE DIRECTIONS

In the near future, over 100 Solanaceae genomes will be sequenced under the SOL-100 project, which will include many less-studied Solanaceae species, varieties and cultivars reflecting the natural biodiversity of this family. In addition, RNA-Seq will also be performed on a much larger scale than ever before in the Solanaceae. The challenge for SGN will be to integrate this information while retaining an easy to use and responsive user interface. In the short term, many of the existing databases and tools will be improved; e.g. a clustering feature will be added to the expression database and improved metabolic and plant ontology searches will be added. In the mid-term, system biology tools for the browsing, curation and visualization of gene network data will be implemented.

Behind the scenes, SGNs codebase is being developed in the direction of creating a generic, reusable, modular and flexible platform suitable for use by other organism databases. In the long run, it is our hope that this will provide opportunities for organizations to cooperate more closely on software development, thereby reducing the endless re-implementation of the same site features at many different databases that is still so common today. SGN actively participates in and contributes to the GMOD project, which has made great strides to combat 'reinventing the wheel', but a lot of work remains to be done.

While SGN has been created primarily with the molecular biologist and the geneticist in mind, a prime focus of current development is on improving the site's usefulness to breeders, who are the crucial link between the advances in the laboratory and improvements in the field, ultimately translating scientific progress into better varieties and contributing to healthier diets and more sustainable agriculture. The recently created breeders' toolbox (http://solgenomics.net/breeders/) will be expanded further, in collaboration with the breeders themselves, to create a comprehensive solution to give breeders easy, intuitive access to the wealth of data in SGN.

CONCLUSIONS

The SGN database is an important resource for Solanaceae scientific research. It currently has over 1000 registered users and more than 6000 unique visitors per month, generating more than 150 000 page views. With many new resources for the Solanaceae coming on-line, usage of SGN can be expected to grow considerably in the future.

Besides the new datasets that have been added, the way SGN interacts with the rest of the world has evolved. SGN actively contributes to open source projects. SGN's 'radically open' software development model offers possibilities for increasing software cooperation with other databases. Most importantly, new community curation tools establish a direct line of communication between the online database and the data producers, with many positive implications for the whole research community.

ACKNOWLEDGEMENTS

The authors would like to gratefully acknowledge Joyce van Eck for her ongoing contribution to the breeders toolbox. And also would like to gratefully acknowledge the National Science Foundation (NSF) and the United States Department of Agriculture (USDA) and ATC Inc. for funding of SGN.

FUNDING

This project was funded by the National Science Foundation (NSF), the United States Department of Agriculture (USDA) and ATC Inc (Advanced Technologies, Cambridge).

Conflict of interest statement. None declared.

REFERENCES

1. Mueller,L.A., Lankhorst,R.K., Tanksley,S.D., Giovannoni,J.J., White,R., Vrebalov,J., Fei,Z.J., Eck,J.v., Buels,R., Mills,A.A. *et al*. (2009) A snapshot of the emerging tomato genome sequence. *Plant Genome*, **2**, 78–92.
2. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
3. Turcatti,G., Romieu,A., Fedurco,M. and Tairi,A.P. (2008) A new class of cleavable fluorescent nucleotides: Synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.*, **36**, e25.
4. Shendure,J., Porreca,G.J., Reppas,N.B., Lin,X., McCutcheon,J.P., Rosenbaum,A.M., Wang,M.D., Zhang,K., Mitra,R.D. and Church,G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.
5. Harris,T.D., Buzby,P.R., Babcock,H., Beer,E., Bowers,J., Braslavsky,I., Causey,M., Colonell,J., Dimeo,J., Efcavitch,J.W. *et al*. (2008) Single-molecule DNA sequencing of a viral genome. *Science*, **320**, 106–109.

6. Travers,K.J., Chin,C.S., Rank,D.R., Eid,J.S. and Turner,S.W. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.*, **38**, e159.
7. Engel,S.R., Balakrishnan,R., Binkley,G., Christie,K.R., Costanzo,M.C., Dwight,S.S., Fisk,D.G., Hirschman,J.E., Hitz,B.C., Hong,E.L. *et al.* (2010) Saccharomyces genome database provides mutant phenotype data. *Nucleic Acids Res.*, **38**, D433–D436.
8. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The arabidopsis information resource (TAIR): Gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
9. Menda,N., Buels,R.M., Tecle,I. and Mueller,L.A. (2008) A community-based annotation framework for linking solanaceae genomes with phenomes. *Plant Physiol.*, **147**, 1788–1799.
10. Shumway,M., Cochrane,G. and Sugawara,H. (2010) Archiving next generation sequencing data. *Nucleic Acids Res.*, **38**, D870–D871.
11. Mungall,C.J. and Emmert,D.B. (2007). FlyBase Consortium. (2007) A chado case study: An ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
12. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A. *et al.* (2009) NCBI GEO: Archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.
13. Parkinson,H., Kapushesky,M., Kolesnikov,N., Rustici,G., Shojatalab,M., Abeygunawardena,N., Berube,H., Dylag,M., Emam,I., Farne,A. *et al.* (2009) ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.
14. Barone,A.D., Beecher,J.E., Bury,P.A., Chen,C., Doede,T., Fidanza,J.A. and McGall,G.H. (2001) Photolithographic synthesis of high-density oligonucleotide probe arrays. *Nucleosides Nucleotides Nucleic Acids*, **20**, 525–531.
15. Edwards,K.D., Bombarely,A., Story,G.W., Allen,F., Mueller,L.A., Coates,S.A. and Jones,L. (2010) TobEA: An atlas of tobacco gene expression from seed to senescence. *BMC Genomics*, **11**, 142.
16. Ozaki,S., Ogata,Y., Suda,K., Kurabayashi,A., Suzuki,T., Yamamoto,N., Iijima,Y., Tsugane,T., Fujii,T., Konishi,C. *et al.* (2010) Coexpression analysis of tomato genes and experimental verification of coordinated expression of genes found in a functionally enriched coexpression module. *DNA Res.*, **17**, 105–116.
17. Schilmiller,A.L., Miner,D.P., Larson,M., McDowell,E., Gang,D.R., Wilkerson,C. and Last,R.L. (2010) Studies of a biochemical factory: Tomato trichome deep expressed sequence tag sequencing and proteomics. *Plant Physiol.*, **153**, 1212–1223.
18. Avraham,S., Tung,C.W., Ilic,K., Jaiswal,P., Kellogg,E.A., McCouch,S., Pujar,A., Reiser,L., Rhee,S.Y., Sachs,M.M. *et al.* (2008) The plant ontology database: A community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Res.*, **36**, D449–D454.
19. Vizcaino,J.A., Cote,R., Reisinger,F., Barsnes,H., Foster,J.M., Rameseder,J., Hermjakob,H. and Martens,L. (2010) The proteomics identifications database: 2010 update. *Nucleic Acids Res.*, **38**, D736–D742.
20. Gasteiger,E., Gattiker,A., Hoogland,C., Ivanyi,I., Appel,R.D. and Bairoch,A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.
21. James,N., Graham,N., Clements,D., Schildknecht,B. and May,S. (2007) AtEnsEMBL. *Methods Mol. Biol.*, **406**, 213–227.
22. Hinz,U. and UniProt Consortium (2010) From protein sequences to 3D-structures and beyond: The example of the UniProt knowledgebase. *Cell Mol. Life Sci.*, **67**, 1049–1064.
23. Wu,C.H., Yeh,L.S., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al.* (2003) The protein information resource. *Nucleic Acids Res.*, **31**, 345–347.
24. Dutta,S., Burkhardt,K., Young,J., Swaminathan,G.J., Matsuura,T., Henrick,K., Nakamura,H. and Berman,H.M. (2009) Data deposition and annotation at the worldwide protein data bank. *Mol. Biotechnol.*, **42**, 1–13.
25. Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 138–148.
26. Pedersen,A.G. and Nielsen,H. (1997) Neural network prediction of translation initiation sites in eukaryotes: Perspectives for EST and genome analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 226–233.
27. Hirosawa,M., Hoshida,M., Ishikawa,M. and Toya,T. (1993) MASCOT: Multiple alignment system for protein sequences based on three-way dynamic programming. *Comput. Appl. Biosci.*, **9**, 161–167.
28. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
29. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.
30. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
31. Edgar,R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
32. Wilgenbusch,J.C. and Swofford,D. (2003) Inferring evolutionary trees with PAUP*. *Curr. Protoc. Bioinformatics*, Chapter 6, Unit 6.4.
33. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
34. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: A building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
35. Gonzalo,M.J. and van der Knaap,E. (2008) A comparative analysis into the genetic bases of morphology in tomato varieties exhibiting elongated fruit shape. *Theor. Appl. Genet.*, **116**, 647–656.
36. Tanksley,S.D., Ganal,M.W., Prince,J.P., de Vicente,M.C., Bonierbale,M.W., Broun,P., Fulton,T.M., Giovannoni,J.J., Grandillo,S. and Martin,G.B. (1992) High density molecular linkage maps of the tomato and potato genomes. *Genetics*, **132**, 1141–1160.
37. Bindler,G., van der Hoeven,R., Gunduz,I., Plieske,J., Ganal,M., Rossi,L., Gadani,F. and Donini,P. (2007) A microsatellite marker based linkage map of tobacco. *Theor. Appl. Genet.*, **114**, 341–349.
38. Mueller,L.A., Solow,T.H., Taylor,N., Skwarecki,B., Buels,R., Binns,J., Lin,C., Wright,M.H., Ahrens,R., Wang,Y. *et al.* (2005) The SOL genomics network: A comparative resource for solanaceae biology and beyond. *Plant Physiol.*, **138**, 1310–1317.
39. Karp,P.D., Paley,S. and Romero,P. (2002) The pathway tools software. *Bioinformatics*, **18(Suppl. 1)**, S225–S232.
40. Broman,K.W., Wu,H., Sen,S. and Churchill,G.A. (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, **19**, 889–890.
41. Brewer,M.T., Moyseenko,J.B., Monforte,A.J. and van der Knaap,E. (2007) Morphological variation in tomato: A comprehensive study of quantitative trait loci controlling fruit shape and development. *J. Exp. Bot.*, **58**, 1339–1349.
42. PostgreSQL Global Development Group. (2010) PostgreSQL documentation. (http://www.postgresql.org/docs/9.0/static/intro-whatis.html).
43. Beck,K. (2003) *Test-driven development by example*, Introduction, xvii. Addison Wesley Longman Publishing Co., Reading, Massachusetts.
44. Larman,C. (2004) Agile, *Agile and Iterative Development: A Manager's Guide*. Addison-Wesley Longman Publishing Co., p. 27.