Published online 11 November 2013 Nucleic Acids Research, 2014, Vol. 42, Database issue **D1193–D1199**
doi:10.1093/nar/gkt1110

Gramene 2013: comparative plant genomics resources

Marcela K. Monaco[1], Joshua Stein[1], Sushma Naithani[2], Sharon Wei[1], Palitha Dharmawardhana[2], Sunita Kumari[1], Vindhya Amarasinghe[2], Ken Youens-Clark[1], James Thomason[1], Justin Preece[2], Shiran Pasternak[1], Andrew Olson[1], Yinping Jiao[1], Zhenyuan Lu[1], Dan Bolser[3], Arnaud Kerhornou[3], Dan Staines[3], Brandon Walts[3], Guanming Wu[4], Peter D'Eustachio[5], Robin Haw[4], David Croft[3], Paul J. Kersey[3], Lincoln Stein[4], Pankaj Jaiswal[2] and Doreen Ware[1,6,*]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA, [2]Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, USA, [3]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK, [4]Informatics and Bio-computing Program, Ontario Institute of Cancer Research, Toronto M5G 1L7, Canada, [5]Department of Biochemistry & Molecular Pharmacology, NYU School of Medicine, New York, NY 10016, USA and [6]NAA Plant, Soil & Nutrition Laboratory Research Unit, USDA-ARS, Ithaca, NY 14853, USA

Received September 25, 2012; Accepted October 21, 2013

ABSTRACT

Gramene (http://www.gramene.org) is a curated online resource for comparative functional genomics in crops and model plant species, currently hosting 27 fully and 10 partially sequenced reference genomes in its build number 38. Its strength derives from the application of a phylogenetic framework for genome comparison and the use of ontologies to integrate structural and functional annotation data. Whole-genome alignments complemented by phylogenetic gene family trees help infer syntenic and orthologous relationships. Genetic variation data, sequences and genome mappings available for 10 species, including *Arabidopsis*, rice and maize, help infer putative variant effects on genes and transcripts. The pathways section also hosts 10 species-specific metabolic pathways databases developed in-house or by our collaborators using Pathway Tools software, which facilitates searches for pathway, reaction and metabolite annotations, and allows analyses of user-defined expression datasets. Recently, we released a Plant Reactome portal featuring 133 curated rice pathways. This portal will be expanded for *Arabidopsis*, maize and other plant species. We continue to provide genetic and QTL maps and marker datasets developed by crop researchers. The project provides a unique community platform to support scientific research in plant genomics including studies in evolution, genetics, plant breeding, molecular biology, biochemistry and systems biology.

INTRODUCTION

Gramene is an integrated web resource for accessing, visualizing, and comparing plant genomes and biological pathways. Each hosted genome features community-based gene annotations from primary sources to which we add Supplementary annotations, functional classification and comparative phylogenomics analysis. For an increasing number of species, with particular focus on *Arabidopsis*, rice and maize, Gramene also annotates and displays variation data derived both from data repositories and through collaboration with large-scale re-sequencing and genotyping initiatives. Another mandate of this project is to build plant pathway databases by applying both manual curation and automated methods. By using a core set of consistently applied protocols, Gramene offers a reference resource for basic and translational research in plants.

Gramene is powered by several platform infrastructures that are linked to provide a unified user experience. Our genome browser (http://www.gramene.org/genome_browser) takes advantage of the Ensembl infrastructure (www.ensembl.org) to provide an interface for exploration

*To whom correspondence should be addressed. Tel: +1 516 367 6979; Fax: +1 516 367 6851; Email: ware@cshl.edu

Published by Oxford University Press 2013. This work is written by US Government employees and is in the public domain in the US.

Downloaded from http://nar.oxfordjournals.org/ at Cold Spring Harbor Laboratory on April 11, 2014

of genome features, functional ontologies, variation data and comparative phylogenomics. Since 2009 Gramene has partnered with the Plants division of Ensembl Genomes (http://www.plants.ensemb.orgl) to jointly produce this resource, each benefitting from the other's proximity to research communities in the USA and Europe. This collaboration has also facilitated timely adoption of innovative tools and software updates that accompany frequent version releases by the Ensembl project (1).

Gramene is also a portal for pathway databases developed and curated internally or mirrored from external sources. Since our last NAR update, Gramene developed and released BrachyCyc and MaizeCyc (2), the latter in collaboration with the MaizeGDB organismal database. We also incorporated many updates to RiceCyc (3) and have continued to maintain SorghumCyc. Built upon the Pathway Tools (BioCyc) platform (4,5), these databases emphasize the annotation of metabolic and transport pathways. Recently Gramene has adopted the Reactome data model and visualization platform (6) to develop the Plant Reactome (http://plantreactome.oicr.on.ca), currently available as a beta release. Over the next 2 years this resource will continue to grow with the addition of new species data and broader coverage of molecular interactions.

These platforms provide region-specific (e.g., genome browser) or pathway-specific data downloads (e.g., pathways portal and Plant Reactome). In addition, project data are available for customizable downloads from the GrameneMart (7), BLAST search, bulk downloads by FTP (ftp://ftp.gramene.org/pub/gramene), and programmatic access via Ensembl API and public MySQL (8).

This article summarizes the updates to the Gramene website and database through the 38th release of the Gramene database in August 2013, since last reported in this journal (8). Starting March 2013, the website, database and its contents are being updated five times during the year and changes can be followed from the Gramene news portal (http://news.gramene.org) and by browsing the site's release notes (http://www.gramene.org/db/help?state=current_release_notes).

NEW PLANT GENOMES AND ANNOTATION

Since our previous NAR report (8), Gramene has tripled its number of complete reference genomes to 27. As shown in Supplementary Table S1, the species list broadens taxonomic representation and increases resolution with the inclusion of 14 monocots, 9 core eudicots and 4 primitive non-flowering plants, while serving both crop and model organism research communities. Notable additions to the monocot list include maize (*Zea mays*) and foxtail millet (*Setaria italica*), which along with *Sorghum bicolor* contribute to biofeedstock research owing to their C4 photosynthetic metabolism. Supporting wheat research, we added two diploid progenitor species *Triticum urartu* and *Aegilops tauschii* representing the AA and DD genome types, respectively. Until recently the monocot collection included only grasses (*Poacea*). This changed with the addition of banana (*Musa acuminate*), among the first non-grass monocots to be sequenced.

We have more than doubled core eudicots. Addition of two members of the *Solanaceae*, tomato (*Solanum lycopersicum*) and potato (*S. tuberosum*), represent the first asterids to join this resource, thus broadening eudicots beyond the rosid subclass. Addition of soybean (*Glycine max*) and *Medicago truncatula* represent two ends of the spectrum within legumes and provide complementary resources for crop breeding and research. In order to broaden the base of the species tree, we now include aquatic algae (*Cyanidioschyzon merolae* and *Chlamydomonas reinhardtii*), an early land plant moss (*Physcomitrella patens*) and an early vascular non-seed plant spikemoss (*Selaginella moellendorffii*).

Although inclusion of basal species aids the investigation of early events in plant evolution, the study of rapidly evolving characteristics requires dense species representation within a more shallow clade. In recent years, Gramene has accomplished this goal by building a rice-genus-level resource that now includes 13 of the estimated 24 species within the *Oryza* genus (9) (Supplementary Table S1). In addition to the two subspecies of Asian cultivated rice, this resource includes complete reference assemblies for cultivated African rice *Oryza glaberrima*, its wild progenitor *Oryza barthii*, and the distantly related wild species *Oryza punctata* and *Oryza brachyantha*. An additional eight *Oryza* species, including one polyploid, plus the outgroup species *Leersia perrieri*, are available as chromosome 3 short-arm assemblies and were contributed through collaboration with the NSF-funded Oryza Map Alignment Project (OMAP) and Oryza Genome Evolution (OGE) projects (http://www.genome.arizona.edu/modules/publisher/item.php?itemid=7). In the coming year, many of these will be replaced with complete reference assemblies provided through various international consortia.

Gramene performs base-line annotation of repeat sequences, est/mRNA alignments and *ab initio* gene prediction (8). The community-recognized gene annotations are characterized for InterPro domains and cross-referenced to entries in third-party databases. Functional information is assigned using ontologies (Supplementary Table S2) through a variety of methods (10), which now include projection from one species to another using Compara gene ortholog assignments.

PLANT COMPARATIVE GENOMICS

The value of individual genomes is vastly enhanced by the provision of genomic and phylogenetic comparisons that elucidate ancestral relationships and evolutionary histories. We accomplish this by employing two Ensembl Compara analysis pipelines that provide: (i) pairwise whole-genome alignments at the DNA level (1,8,10); and (ii) Phylogenetic gene trees with classification of ortholog and paralog gene relationships (8,11). Output from either method may be subsequently used to build synteny maps (8). In the past year we increased the number of pairwise whole-genome alignments from 31 to 64, as shown in Supplementary Table S3. By default each species is aligned to rice and Arabidopsis, as well-annotated

Downloaded from http://nar.oxfordjournals.org/ at Cold Spring Harbor Laboratory on April 11, 2014

references. Additional pairwise comparisons were strategically selected to enrich this resource. Among the eudicots, *Vitis vinifera* (grapevine) is the only species not to have undergone whole-genome duplication since divergence from a common ancestor; hence, grapevine serves as the best eudicot reference to identify ancestral regions. As an example, synteny maps in Figure 1A illustrate the better clarity that grapevine provides in identifying duplicated regions of poplar compared to using Arabidopsis as the reference. Other species combinations were chosen to serve specific research needs in the community, such as comparisons between the three C4 grasses and comparison among solanaceous crops.

The standard gene-tree protocol includes annotated protein-coding genes from the complete reference genomes plus several non-plant species to give broader taxonomic context (8,10). Recent Gramene releases have synchronized this resource from Ensembl Plants. Independently, Gramene produces a second set of gene trees that focus on the *Oryza* genus. The '*Oryza*-centered' gene trees incorporate gene predictions from all *Oryza* species (Supplemental Table S1), including those of the chromosome 3 short-arm assemblies, along with a select set of informative outgroup species.

A recent enhancement of the Compara method is automated detection of putative split-gene models that can arise from error in assembly or annotation (1) (Supplementary Table S4), as exemplified in Figure 1B. To serve community annotation efforts, we provide a list of putative split genes available by FTP (ftp://ftp.gramene.org/pub/gramene/CURRENT_RELEASE/data/split_genes/).

PLANT GENETIC DIVERSITY AND SEQUENCE VARIATION

Genomics research is increasingly driven by the collection of polymorphism data from both natural and controlled plant populations. Gramene currently incorporates SNP and/or structural variation datasets for nine genomes (Supplementary Table S5): *Arabidopsis* (14–16), japonica and indica rice (13,17), maize (12,18), barley (12,18,19), grape (20), *Brachypodium* (21), African rice and sorghum (22). The Ensembl variant effect predictor (VEP) pipeline (23) classifies variants according to functional consequences using Sequence Ontology terms (24). These can be visualized in the context of transcript structure and protein domains. For many studies we also capture genotypes of individual plant accessions and phenotype data. A notable addition to this resource was the maize HapMap2 dataset, containing 55 million SNPs and indels across 103 accessions (12,25).

NEW ENSEMBL BROWSING CAPABILITIES

Each Gramene release brings new features through advances in the Ensembl software infrastructure. Since our previous report (8), users are now able to upload their own private datasets (e.g., genome-wide SNP associations, QTLs, linkage data, ESTs, microarray data,

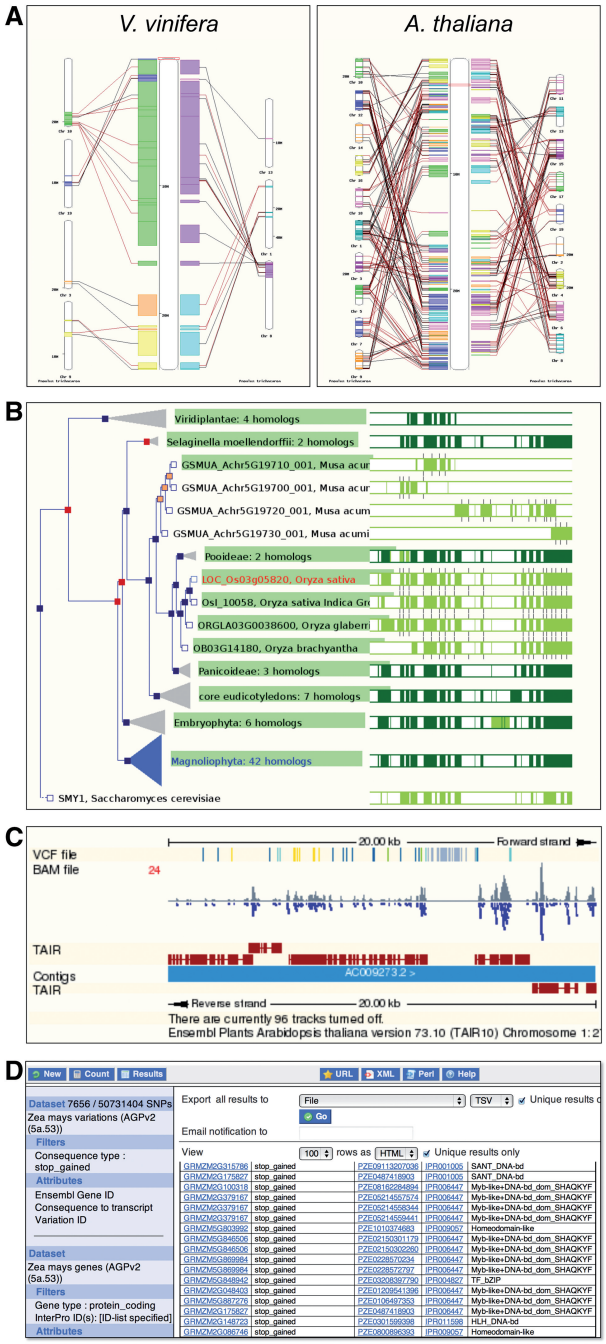

Figure 1. Gramene's Ensembl and Biomart interfaces. (**A**) The synteny map browser facilitates navigation from one genome to another across orthologous regions. In the left panel grapevine chr. 13 is represented in the center surrounded by mapped chromosomes of poplar. Duplicated regions of poplar are obvious in this view but more difficult to trace when using Arabidopsis (chr. 5) as the reference (right panel). (**B**) Compara gene trees help identify putative split-gene models, here showing fragmented alignments of consecutive banana genes, and absence of a conserved InterPro domain (green highlighting) in these fragments. (**C**) The genome browser allows uploading of private user data, here showing the Arabidopsis browser with variation data from a VCF file and transcriptome alignment data from a BAM file. The browser automatically predicted the effect of each SNP on overlapping annotated gene transcripts as indicated by color-coding (legend omitted). (**D**) GrameneMart includes five databases for gene, variation, marker, mapping and QTL-oriented searching. Here maize transcription factor genes were mined for the existence of possible detrimental alleles in HapMap2 data (17,25).

Downloaded from http://nar.oxfordjournals.org/ at Cold Spring Harbor Laboratory on April 11, 2014

RNA-Seq, proteomic sets) to view alongside reference annotations in the genome browser, Figure 1C (1,10). In addition to common file formats such as GFF and BED, users can upload BAM to view short-read alignment data, or VCF files to view variant calls. In the latter case, the web-service automatically performs variant effect prediction and color-codes the displayed SNPs accordingly (Figure 1C). Other supported formats are listed at http://www.gramene.org/info/website/upload/index.html. Gramene has incorporated the ability to dynamically highlight genes sharing the selected GO annotation or InterPro domain into its gene-tree viewer. This allows trees to be evaluated for consistency of annotation across clades (Figure 1B).

DATA MINING USING GRAMENEMART

The Gramene Project was an early adopter of the BioMart data management system and web interface (7,26–28). GrameneMart helps users to rapidly download custom datasets. For example, a user can request a list of maize genes, along with genomic coordinates, protein domains, GO classes, and corresponding orthologs in rice, Arabidopsis and sorghum. More powerful still, users can apply filters to advance specific research questions. In Figure 1D, GrameneMart was used to screen transcription factor genes having putative SNPs of premature stop-codon. As the mart interface is linked to the browser, it is easy for users to quickly navigate to the corresponding gene or variation pages, and drill down to the list of maize strains that can carry the predicted detrimental alleles.

PLANT PATHWAYS

Gramene currently hosts 10 species-specific pathways databases (http://www.gramene.org/pathway; Supplementary Table S6) developed using Pathway Tools software (4,5). Of these, RiceCyc (*Oryza sativa japonica*) (3), SorghumCyc (*Sorghum bicolor*), MaizeCyc (*Z. mays*) (2) and BrachyCyc (*Brachypodium distachyon*) were developed and continue to be maintained by Gramene (Figure 2).

The pathway databases can be browsed online or locally installed and navigated using the Pathway Tools latest software version 17.0 (4). Both desktop and online versions are searchable by gene, enzyme, metabolite or pathway name as shown with the RiceCyc example (Figure 2). Pathway databases also provide information directly or via web links on peptide sequences, gene homologs, chemical structures of metabolites, literature citations and comparative data across multiple species, as described in Dharmawardhana *et al*. (2013) and Monaco *et al*. (2012). The Omics-Viewer tool therein provides a cellular overview of the metabolic networks as a schematic diagram where nodes represent metabolites (with shape indicating class of metabolite) and lines represent reactions (Figure 2). The details of the pathway are accessible by clicking on a node (metabolite icon) or a line (reaction). The Omics-Viewer also allows users to upload and visualize high-throughput experimental datasets (e.g., microarray, RNA-Seq, proteome, metabolomics, reaction flux data, etc.) to compare various samples (e.g., experimental conditions, treatments, tissue types, time series, etc.) in the context of the overall cellular metabolic network (Figure 2F; (29)).

Since our last publication (8) the major updates to the pathway databases include manual curation of metabolic pathways in MaizeCyc and RiceCyc. MaizeCyc (2) currently projects a total of 428 metabolic pathways and transport reactions with ~9000 genes acting as enzymes and transporters, and 1450 compounds. Manually curated pathways in MaizeCyc include carotenoid biosynthesis (from lycopene to carotene and xanthophylls) and flavonoid and flavonol biosynthesis leading to anthocyanin biosynthesis (2), and vitamin B biosynthesis and degradation pathways (30,31). RiceCyc (3) version 3.3 features 311 pathways (Figure 2) and includes the recently curated terpenoid biosynthesis instances of momilactone biosynthesis, Oryzalexin A-F biosynthesis, Oryzalexin S biosynthesis and phytocassane biosynthesis (31–34). SorghumCyc and BrachyCyc are maintained as computational projections from automated builds as described by Monaco *et al*. (2012). The updates also include updated mapping of genes and gene products to the known pathways and reactions and as well as removing the false mappings that were inferred by the automated annotation workflows designed on the gene homology platform. The pathway databases such as RiceCyc provide a platform for building novel hypothesis for experimental validation. As illustrated in Dharmawardhana *et al*. (2013), the link between circadian control and activation of core tryptophan pathway genes under pathogen treatment was a novel finding which may open up opportunities to look for novel sets of genes and networks involved in building new strategies for biotic stress resistance.

To further improve functional annotation and reconstruction of metabolic and regulatory networks in plants, we developed the Plant Reactome (http://plantreactome.oicr.on.ca) in collaboration with the Human Reactome project (35). The rationale behind the Reactome platform is to convey the extensive amount of information available for metabolic and signaling networks in visual representations that are intuitively navigable via a web interface, and are computationally accessible for advanced users via the APIs. Currently in its beta version, the Plant Reactome includes 133 rice pathways. Functionality updates and curation of Arabidopsis and maize pathways in Plant Reactome are in progress. Projections for maize and other species will follow in future Gramene releases.

GRAMENE TUTORIALS

Gramene offers YouTube video tutorials on topics including an overview of current datasets, features and tools (http://www.youtube.com/watch?v=wEaoJTTqWvI), describing Gramene's pathways portal (http://www.youtube.com/watch?v=umlpHVon1OM) and to learn more about the Plant Reactome (http://www.youtube.com/watch?v=wbkuTeIcKjI).

Downloaded from http://nar.oxfordjournals.org/ at Cold Spring Harbor Laboratory on April 11, 2014

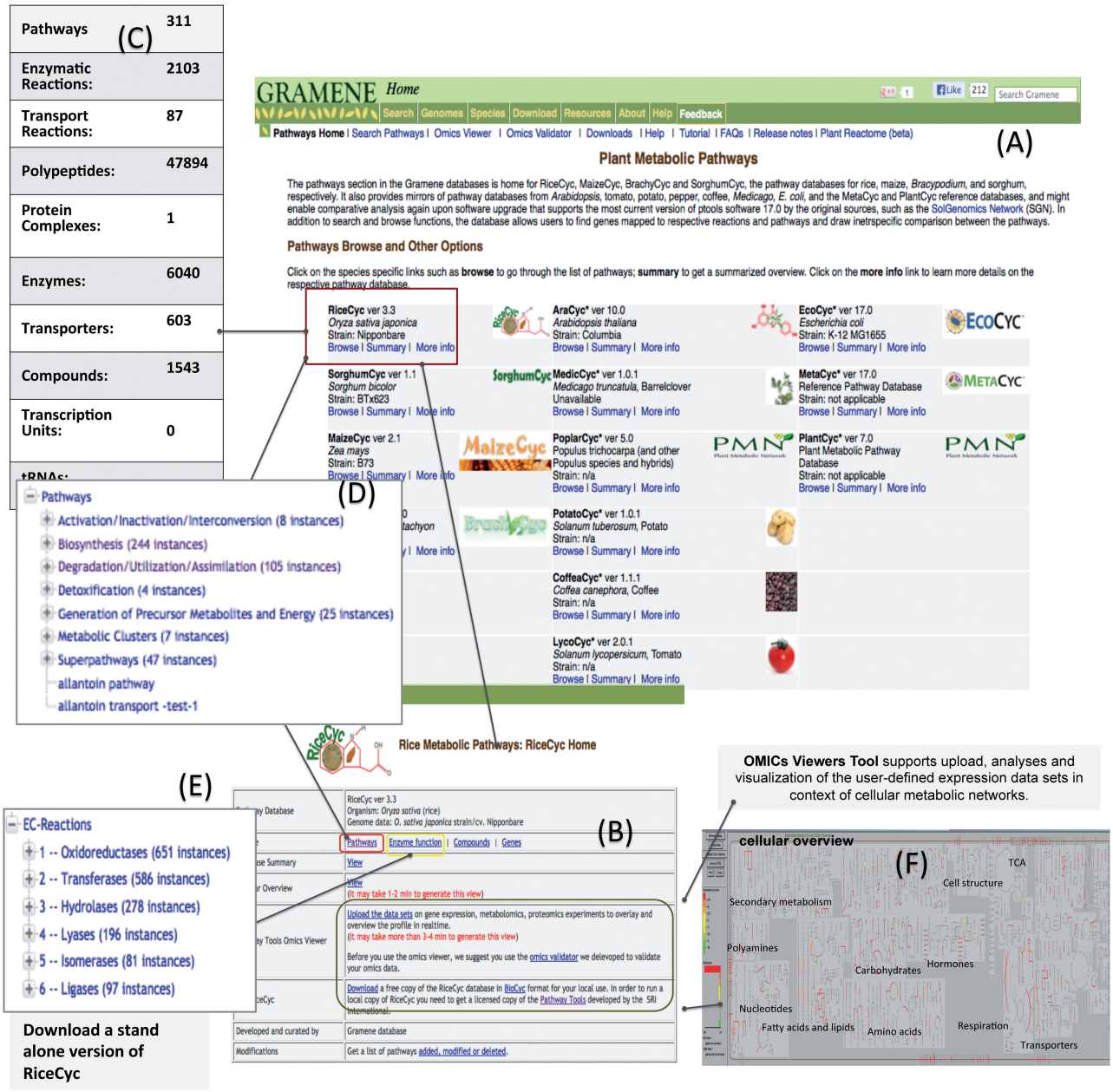

Figure 2. Gramene's Pathway Database module. A view of Gramene's BioCyc-based Plant Metabolic Pathways (http://www.gramene.org/pathway) entry page listing species wise pathways databases (**A**), detail information (**B**) on the RiceCyc example including accessibility of various features such as summary (**C**), browse (**D**, **E**) and Omics-Viewer tool (**F**) allows visualization of user-defined expression data on cellular overview diagram. For example, in the cellular overview diagram rice genes that show high expression at 0 h during diurnal cycle are highlighted in red color (*Data source*: Filichkin *et al.*, 2011).

DISCUSSION AND FUTURE PERSPECTIVE

As this report attests, the plant community has enjoyed enormous success establishing new reference genomes for important crops and model species using Gramene resources. Although this list will continue to grow, a greater opportunity—and challenge—will be presented by new data describing the transcriptome, epigenome and variome within existing reference species. Furthermore, it is common knowledge that factors that are external (i.e., environmental) or internal (i.e., genetic and epigenetic) can cause a perturbation of a biological system. For example a gene mutation may cause an alteration of a protein function leading to a systems-level change. Such a change can be captured and deciphered only through a systems- or network-level approach involving additional components like gene expression, metabolomics and network analysis. Thus, in collaboration with the ATLAS project (36), we are in the process of developing a capacity to map and display expression of genes in response to various environmental conditions, such as drought and salinity, and identify gene functions in order to elucidate biochemical and signaling pathways

Downloaded from http://nar.oxfordjournals.org/ at Cold Spring Harbor Laboratory on April 11, 2014

which underlie the plant's response to abiotic and biotic stress during the course of plant development. With regards to the integration of transcriptomics and epigenomics data, projects such as the Encyclopedia of DNA Elements (ENCODE; (37)) have demonstrated the value of comprehensive analysis of transcription and chromatin structure on understanding gene regulation. As the Ensembl project participated in ENCODE and other large-scale functional genomics projects in human, it is anticipated that Gramene will be also able to adapt infrastructure, such as the Regulatory Build (1), into future developments. Lastly, the large degree of intra-species variation has shown that a single reference's assembly is insufficient to represent the genome of a species (38–40). Following the trend in microbial genomics, the concept of a single reference genome is giving way to that of the 'pan-genome' in both animals and plants (41,42) in order to describe the full-complement of genes and variants in a species by capturing both the conserved 'core' genome as well as the 'dispensable' genome that is specific to populations or single individuals. Hence the systems/network-level approach that we envision will not only answer fundamental biological questions on such mechanisms of adaptation and speciation, but is expected to revolutionize the methodological approaches for crop improvement.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [12–22].

ACKNOWLEDGEMENTS

The authors are evermore grateful to Gramene's users for their valuable suggestions and feedback in improving the overall quality of Gramene as a community resource. We would also like to thank the Cold Spring Harbor Laboratory (CSHL), the Center for Genome Research and Biocomputing (CGRB) at Oregon State University and the Ontario Institute for Cancer Research (OICR) for infrastructure support. We acknowledge our fellow researchers, and their respective organizations for sharing genomic-scale datasets. We also thank Peter van Buren from CSHL for excellent system administration support, undergraduate students Dylan Beorchia, Kindra Amoss and Teague from Oregon State University for their help on Reactome curation.

FUNDING

National Science Foundation [IOS-0703908 and IOS-1127112]; United States Department of Agriculture—Agricultural Research Service [413089, 418046 and 418047 to D.W.]; European Community's 7th Framework Programme (FP7/2007-2013; Infrastructures) [contract # 283496 to P.K.]; United Kingdom Biotechnology and Biosciences Research Council [BB/J000328X/1, I008071/1 and H531519/1 to P.K.]; The infrastructure and intellectual support for the development and running the Plant Reactome is supported by the Reactome database project via a grant from the US National Institutes of Health [P41 HG003751 to L.S.], EU grant [LSHG-CT-2005-518254] 'ENFIN', Ontario Research Fund and the EBI Industry Programme. The funders had no role in the study design, data analysis or preparation of the manuscript. Funding for open access charge: Gramene Project NSF grant [IOS-1127112].

Conflict of interest statement. None declared.

REFERENCES

1. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
2. Monaco,M.K., Sen,T.Z., Dharmawardhana,P.D., Ren,L., Schaeffer,M., Naithani,S., Amarasinghe,V., Thomason,J., Harper,L., Gardiner,J. *et al.* (2012) Maize metabolic network construction and transcriptome analysis. *Plant Genome*, **6**, 1–12.
3. Dharmawardhana,P., Ren,L., Amarasinghe,V., Monaco,M., Thomason,J., Ravenscroft,D., McCouch,S., Ware,D. and Jaiswal,P. (2013) A genome scale metabolic network for rice and accompanying analysis of tryptophan, auxin and serotonin biosynthesis regulation under biotic stress. *Rice*, **6**, 1–15.
4. Caspi,R., Altman,T., Dreher,K., Fulcher,C.A., Subhraveti,P., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al.* (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **40**, D742–D753.
5. Karp,P.D., Paley,S.M., Krummenacker,M., Latendresse,M., Dale,J.M., Lee,T.J., Kaipa,P., Gilham,F., Spaulding,A., Popescu,L. *et al.* (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.*, **11**, 40–79.
6. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
7. Spooner,W., Youens-Clark,K., Staines,D. and Ware,D. (2012) GrameneMart: the BioMart data portal for the Gramene project. *Database*, **2012**, bar056.
8. Youens-Clark,K., Buckler,E., Casstevens,T., Chen,C., Declerck,G., Derwent,P., Dharmawardhana,P., Jaiswal,P., Kersey,P., Karthikeyan,A.S. *et al.* (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Res.*, **39**, D1085–D1094.
9. Wing,R.A., Ammiraju,J.S., Luo,M., Kim,H., Yu,Y., Kudrna,D., Goicoechea,J.L., Wang,W., Nelson,W., Rao,K. *et al.* (2005) The oryza map alignment project: the golden path to unlocking the genetic potential of wild rice species. *Plant Mol. Biol.*, **59**, 53–62.
10. Kersey,P.J., Staines,D.M., Lawson,D., Kulesha,E., Derwent,P., Humphrey,J.C., Hughes,D.S., Keenan,S., Kerhornou,A., Koscielny,G. *et al.* (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.*, **40**, D91–D97.
11. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
12. Atwell,S., Huang,Y.S., Vilhjalmsson,B.J., Willems,G., Horton,M., Li,Y., Meng,D., Platt,A., Tarone,A.M., Hu,T.T. *et al.* (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **465**, 627–631.
13. Clark,R.M., Schweikert,G., Toomajian,C., Ossowski,S., Zeller,G., Shinn,P., Warthmann,N., Hu,T.T., Fu,G., Hinds,D.A. *et al.* (2007) Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*, **317**, 338–342.
14. Gan,X., Stegle,O., Behr,J., Steffen,J.G., Drewe,P., Hildebrand,K.L., Lyngsoe,R., Schultheiss,S.J., Osborne,E.J., Sreedharan,V.T. *et al.* (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, **477**, 419–423.
15. McNally,K.L., Childs,K.L., Bohnert,R., Davidson,R.M., Zhao,K., Ulat,V.J., Zeller,G., Clark,R.M., Hoen,D.R., Bureau,T.E. *et al.*

Downloaded from http://nar.oxfordjournals.org/ at Cold Spring Harbor Laboratory on April 11, 2014

(2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl Acad. Sci. USA*, **106**, 12273–12278.
16. Zhao,W., Wang,J., He,X., Huang,X., Jiao,Y., Dai,M., Wei,S., Fu,J., Chen,Y., Ren,X. *et al.* (2004) BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Res.*, **32**, D377–D382.
17. Chia,J.M., Song,C., Bradbury,P.J., Costich,D., de Leon,N., Doebley,J., Elshire,R.J., Gaut,B., Geller,L., Glaubitz,J.C. *et al.* (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.*, **44**, 803–807.
18. Gore,M.A., Chia,J.M., Elshire,R.J., Sun,Q., Ersoz,E.S., Hurwitz,B.L., Peiffer,J.A., McMullen,M.D., Grills,G.S., Ross-Ibarra,J. *et al.* (2009) A first-generation haplotype map of maize. *Science*, **326**, 1115–1117.
19. Mayer,K.F., Waugh,R., Brown,J.W., Schulman,A., Langridge,P., Platzer,M., Fincher,G.B., Muehlbauer,G.J., Sato,K., Close,T.J. *et al.* (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.
20. Myles,S., Boyko,A.R., Owens,C.L., Brown,P.J., Grassi,F., Aradhya,M.K., Prins,B., Reynolds,A., Chia,J.M., Ware,D. *et al.* (2011) Genetic structure and domestication history of the grape. *Proc. Natl Acad. Sci. USA*, **108**, 3530–3535.
21. Fox,S.E., Preece,J., Kimbrel,J.A., Marchini,G.L., Sage,A., Youens-Clark,K., Cruzan,M.B. and Jaiswal,P. (2012) Sequencing and de novo transcriptome assembly of Brachypodium sylvaticum (Poaceae). *Appl. Plant Sci.*, **1**, 1200011.
22. Zheng,L.Y., Guo,X.S., He,B., Sun,L.J., Peng,Y., Dong,S.S., Liu,T.F., Jiang,S., Ramachandran,S., Liu,C.M. *et al.* (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biol.*, **12**, R114.
23. Chen,Y., Cunningham,F., Rios,D., McLaren,W.M., Smith,J., Pritchard,B., Spudich,G.M., Brent,S., Kulesha,E., Marin-Garcia,P. *et al.* (2010) Ensembl variation resources. *BMC Genomics*, **11**, 293.
24. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
25. Hufford,M.B., Xu,X., van Heerwaarden,J., Pyhajarvi,T., Chia,J.M., Cartwright,R.A., Elshire,R.J., Glaubitz,J.C., Guill,K.E., Kaeppler,S.M. *et al.* (2012) Comparative population genomics of maize domestication and improvement. *Nat. Genet.*, **44**, 808–811.
26. Guberman,J.M., Ai,J., Arnaiz,O., Baran,J., Blake,A., Baldock,R., Chelala,C., Croft,D., Cros,A., Cutts,R.J. *et al.* (2011) BioMart Central Portal: an open database network for the biological community. *Database*, **2011**, bar041.
27. Kasprzyk,A. (2011) BioMart: driving a paradigm change in biological data management. *Database*, **2011**, bar049.
28. Kinsella,R.J., Kahari,A., Haider,S., Zamora,J., Proctor,G., Spudich,G., Almeida-King,J., Staines,D., Derwent,P., Kerhornou,A. *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, **2011**, bar030.
29. Filichkin,S.A., Breton,G., Priest,H.D., Dharmawardhana,P., Jaiswal,P., Fox,S.E., Michael,T.P., Chory,J., Kay,S.A. and Mockler,T.C. (2011) Global profiling of rice and poplar transcriptomes highlights key conserved circadian-controlled pathways and cis-regulatory modules. *PLoS ONE*, **6**, e16907.
30. Gerdes,S., Lerma-Ortiz,C., Frelin,O., Seaver,S.M., Henry,C.S., de Crecy-Lagard,V. and Hanson,A.D. (2012) Plant B vitamin pathways and their compartmentation: a guide for the perplexed. *J. Exp. Bot.*, **63**, 5379–5395.
31. Shimura,K., Okada,A., Okada,K., Jikumaru,Y., Ko,K.W., Toyomasu,T., Sassa,T., Hasegawa,M., Kodama,O., Shibuya,N. *et al.* (2007) Identification of a biosynthetic gene cluster in rice for momilactones. *J. Biol. Chem.*, **282**, 34013–34018.
32. Wilderman,P.R., Xu,M., Jin,Y., Coates,R.M. and Peters,R.J. (2004) Identification of syn-pimara-7,15-diene synthase reveals functional clustering of terpene synthases involved in rice phytoalexin/allelochemical biosynthesis. *Plant Physiol.*, **135**, 2098–2105.
33. Xu,M., Hillwig,M.L., Prisic,S., Coates,R.M. and Peters,R.J. (2004) Functional identification of rice syn-copalyl diphosphate synthase and its role in initiating biosynthesis of diterpenoid phytoalexin/allelopathic natural products. *Plant J.*, **39**, 309–318.
34. Xu,M., Wilderman,P.R., Morrone,D., Xu,J., Roy,A., Margis-Pinheiro,M., Upadhyaya,N.M., Coates,R.M. and Peters,R.J. (2007) Functional characterization of the rice kaurene synthase-like gene family. *Phytochemistry*, **68**, 312–326.
35. Croft,D. (2013) Building models using Reactome pathways as templates. *Methods Mol. Biol.*, **1021**, 273–283.
36. Kapushesky,M., Adamusiak,T., Burdett,T., Culhane,A., Farne,A., Filippov,A., Holloway,E., Klebanov,A., Kryvych,N., Kurbatova,N. *et al.* (2012) Gene Expression Atlas update–a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **40**, D1077–D1081.
37. Lai,J., Li,R., Xu,X., Jin,W., Xu,M., Zhao,H., Xiang,Z., Song,W., Ying,K., Zhang,M. *et al.* (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.*, **42**, 1027–1030.
38. Santuari,L., Pradervand,S., Amiguet-Vercher,A.M., Thomas,J., Dorcey,E., Harshman,K., Xenarios,I., Juenger,T.E. and Hardtke,C.S. (2010) Substantial deletion overlap among divergent Arabidopsis genomes revealed by intersection of short reads and tiling arrays. *Genome Biol.*, **11**, R4.
39. Zhang,P., Dreher,K., Karthikeyan,A., Chi,A., Pujar,A., Caspi,R., Karp,P., Kirkup,V., Latendresse,M., Lee,C. *et al.* (2010) Creation of a genome-wide metabolic pathway database for Populus trichocarpa using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol.*, **153**, 1479–1491.
40. Dooner,H.K. and Weil,C.F. (2007) Give-and-take: interactions between DNA transposons and their host plant genomes. *Curr. Opin. Genet. Dev.*, **17**, 486–492.
41. Morgante,M., De Paoli,E. and Radovic,S. (2007) Transposable elements and the plant pan-genomes. *Curr. Opin. Plant Biol.*, **10**, 149–155.
42. Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Downloaded from http://nar.oxfordjournals.org/ at Cold Spring Harbor Laboratory on April 11, 2014