Author Queries

ARTICLE ID: **DATABASE-bar056**				
Please respond to all queries and send any additional proof corrections. Failure to do so could result in delayed publication				
Query No	**Section**	**Page**	**Line no.**	**Query**
Q1	Author names			Please check that all names have been spelled correctly and appear in the correct order. Please also check that all initials are present. Please check that the author surnames (family name) have been correctly identified by a pink background. If this is incorrect, please identify the full surname of the relevant authors. Occasionally, the distinction between surnames and forenames can be ambiguous, and this is to ensure that the authors' full surnames and forenames are tagged correctly, for accurate indexing online. Please also check all author affiliations.
Q2	Correspondence			Please provide the Telephone and Fax numbers of the corresponding author.
Q3	References			Please note that the Reference 29 is not cited in the text. Please cite it in text or delete from the reference list.
Q4	References			Please update Reference 29.
Q5	Funding			Please provide a Funding statement, detailing any funding received. Remember that any funding used while completing this work should be highlighted in a separate Funding section. Please ensure that you use the full official name of the funding body, and if your paper has received funding from any institution, such as NIH, please inform us of the grant number to go into the funding section. We use the institution names to tag NIH-funded articles so they are deposited at PMC. If we already have this information, we will have tagged it and it will appear as coloured text in the funding paragraph. Please check the information is correct.
Q6	Funding			Please provide the funding agency details of open access publication.
Q7	Figures			Figures have been placed as close as possible to their first citation. Please check that they have no missing sections and that the correct figure legend is present.

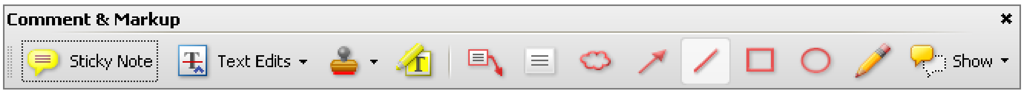
MAKING CORRECTIONS TO YOUR PROOF

These instructions show you how to mark changes or add notes to the document using the Adobe Acrobat Professional version 7.0 (or onwards) or Adobe Reader 8 (or onwards). To check what version you are using go to **Help** then **About**. The latest version of Adobe Reader is available for free from get.adobe.com/reader.

For additional help please use the **Help** function or, if you have Adobe Acrobat Professional 7.0 (or onwards), go to http://www.adobe.com/education/pdf/acrobat_curriculum7/acrobat7_lesson04.pdf

Displaying the toolbars

Adobe Reader 8: Select Tools, Comments & Markup, Show Comments and Markup Toolbar**. If this option is not available, please let me know so that I can enable it for you.**

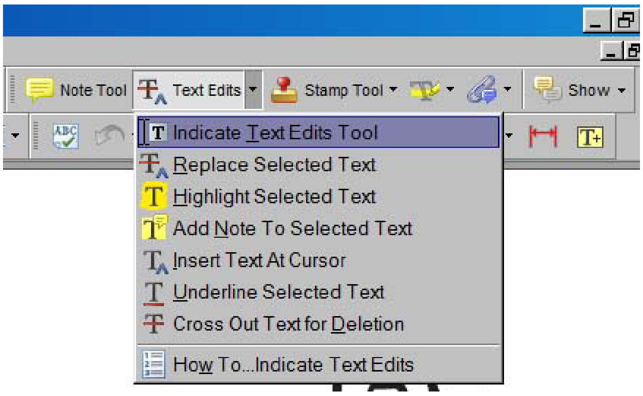


Acrobat Professional 7: Select Tools, Commenting, Show Commenting Toolbar**.**

Adobe Reader 10: To edit the galley proofs, use the comments tab at the top right corner.

Using Text Edits

This is the quickest, simplest and easiest method both to make corrections, and for your corrections to be transferred and checked.

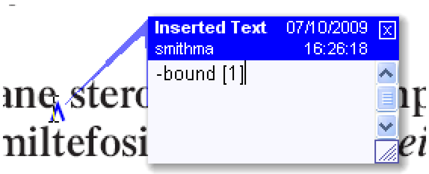


1. Click **Text Edits**
2. Select the text to be annotated or place your cursor at the insertion point.
3. Click the **Text Edits** drop down arrow and select the required action.

You can also right click on selected text for a range of commenting options.

Pop up Notes

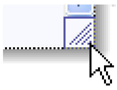
With *Text Edits* and other markup, it is possible to add notes. In some cases (e.g. inserting or replacing text), a pop-up note is displayed automatically.



To **display** the pop-up note for other markup, right click on the annotation on the document and selecting **Open Pop-Up Note**.

To **move** a note, click and drag on the title area.

To **resize** of the note, click and drag on the bottom right corner.



To **close** the note, click on the cross in the top right hand corner.

To **delete** an edit, right click on it and select **Delete**. The edit and associated note will be removed.

SAVING COMMENTS

In order to save your comments and notes, you need to save the file (**File, Save**) when you close the document.
A full list of the comments and edits you have made can be viewed by clicking on the Comments tab in the bottom-left-hand corner of the PDF.

Database, Vol. 2011, Article ID bar056, doi:10.1093/database/bar056

Original article

GrameneMart: the BioMart data portal for the Gramene project

Q1

William Spooner[1,]*, Ken Youens-Clark[1], Daniel Staines[2] and Doreen Ware[1]

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA and [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Q2

***Corresponding author:** wspooner@cshl.edu; gramene@gramene.org

Submitted 30 April 2011; Revised 18 November 2011; Accepted 21 November 2011

Gramene is a well-established resource for plant comparative genome analysis. Data are generated through automated and curated analyses and made available through web interfaces such as GrameneMart. The Gramene project was an early adopter of the BioMart software, which remains an integral and well-used component of the Gramene website. BioMart accessible data sets include plant gene annotations, plant variation catalogues, genetic markers, physical mapping entities, public DNA/mRNA sequences of various types and curated quantitative trait loci for various species.

Database URL: http://www.gramene.org/biomart/martview

Project description

The Gramene project (http://www.gramene.org) was launched in March 2001 as a curated, open-source, Web-accessible data resource for comparative genome analysis (1). Gramene's purpose is to provide added value to data sets available within the public sector, which facilitate researchers' ability to understand plant genomes and take advantage of genomic sequence known in one species for identifying and understanding corresponding genes, pathways and phenotypes in other grass species. This is achieved by building automated and curated relationships between species that can be queried and displayed using web-based interfaces such as GrameneMart.

In the 10 years since its release, the Gramene database has increased in scope and scale. The 34th build released in October 2011 hosts 49M genetic markers and associated DNA sequences from hundreds of plant species, and 22 assembled plant genomes (14 completed and 8 partial). Figure 1 shows the growth of the Gramene database in numbers of complete and partially sequenced species represented in Gramene's genomes module.

Alongside code developed at Gramene, the project has since January 2002 used code developed by the Ensembl project (2) for the genome browser. Gramene's first release including BioMart software (3) was version 18 in July 2005, which used the data transformation tools provided by Ensembl.

With the launch of the Ensembl Genomes project in 2009 (4), Gramene has entered into a close collaboration over the generation of plant Ensembl databases, including the sharing of Plants Genes and Plants Variations BioMarts; these databases, and the interfaces used to query them, are mirrored at both Gramene and Ensembl Genomes websites. The first coordinated release was Gramene v30, Ensembl Genomes release 3 in October 2009.

For the October 2011 release of Gramene, v34, GrameneMart was built with BioMart version 0.7 software, and the Gramene website was using BioMart version 0.7 software.

Query examples

The MartView web interface for GrameneMart is found at http://www.biomart.org/biomart/martview, with access also available from the central BioMart portal http://www.biomart.org/biomart/martview. Mirrors of the Plants Genes Mart and Plants Variations Mart are available through

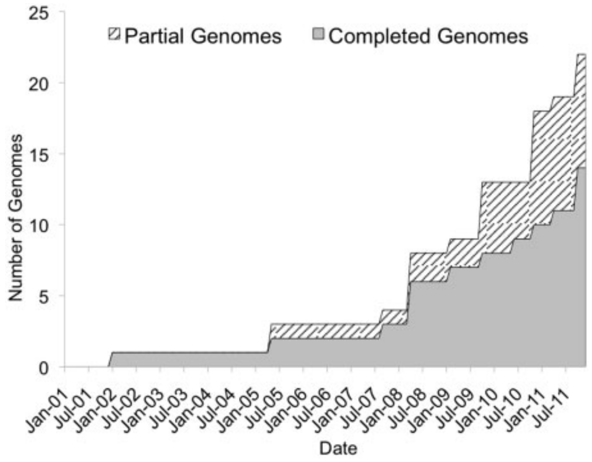


Figure 1. Growth in the number of species represented in Gramene's genomes module.

Ensembl Genomes; http://plants.ensembl.org/biomart/martview.

To demonstrate various features of the GrameneMart database we present various example queries in Table 1.

Data content

There are five BioMart databases in Gramene; Plants Genes Mart, Plants Variations Mart, Gramene Markers, Gramene Mappings and Gramene QTL. Of these, the two Plants Marts are developed with and mirrored by Ensembl Genomes, whereas the three others, Gramene Markers, Gramene Mappings and Gramene QTL, are unique to Gramene. Each database is described below.

Plant Gene Mart database

As of October 2011, Gramene's version of the Plants Genes Mart database holds a data set for each of the 14 fully sequenced plant genomes represented in Gramene. One useful feature of the Plant Genes database is the ability to map gene identifiers of one type to those of another using the extensive set of gene-anchored cross-references. Cross-reference sources from the Plant Gene Mart include EMBL (5), EntrezGene (6), IPI (7), PDB (8), RefSeq (9), UniProt (10) and UniGene (11), PlantGDB transcript assemblies (12) and identifiers from The Gene Index (13). There are also a number of species-specific identifiers, including BGI-RIS (14), the Rice Genome Annotation Project (15) and RAP-DB (16) gene identifiers for rice; IGGP (17) gene identifiers for grape; JGI gene identifiers for *Arabidopsis lyrata* (18), sorghum (19) and poplar (20); and TAIR (21) identifiers for *Arabidopsis thaliana*.

Cross-references to other Gramene modules, such as Genes, Markers/Sequences (see below) and Pathways (1) are also represented and are used to make other gene assignments to EC Number and terms from various ontologies including the Plant Ontology and Gene Ontology (22).

The Plant Gene Mart allows users to filter by, and export attributes related to, genomic region, gene biotype, database cross references, ontology terms, orthology relationships with genes in other species or paralogy with genes in the same species, protein domain annotation and, where available, consequences of any underlying genomic variants. An example query against the Plant Gene Mart database is included as 'Query 1' in Table 1.

Plant Variation Mart database

The Plant Variation Mart database holds a catalogue of DNA variants including single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) for *A. thaliana*, rice (Japonica Group), rice (Indica Group) and grape. The *A. thaliana* data set contains over 8 700 000 variants compiled from a number of studies (23,24), representing SNP discovery and genotyping across over 1000 *Arabidopsis* accessions. In addition to the variants, their genomic locations and corresponding genetic consequences, a number of trait association data (23) are also exposed. The rice data sets, both Indica and Japonica, contain approx. 5 500 000 variants, predominantly from dbSNP (25), but also 150 000 SNPs discovered in 20 accessions by OryzaSNP (26) and data from a 1536 SNP panel genotyped across 395 accessions (27). The grape variation data set contains 460 000 SNPs discovered by next-generation sequencing of 18 grape cultivars (28).

The Plant Variation Mart allows users to filter by, and export attributes related to, genomic region, variation ID, phenotype association, variation set/study, strain/accession, gene association and consequence. An example query against the Plant Variation Mart database is included as 'Query 2' in Table 1.

Gramene Markers database

This database contains all of the genetic marker and associated DNA/mRNA sequence records represented in Gramene, which number 49M entries as of October 2011. Unlike the Gene and Variation Mart databases that aim to be comprehensive catalogues for their respective genomes, the Markers database represents a wide range of entities from many species that have been made available in the public domain, e.g. through databases such as GenBank.

Entities in the Gramene Markers database are classified by type, with a data set created for each type. The different types and the numbers of each are shown in Table 2. Markers are connected to each other via correspondences, e.g. individual Expressed Sequence Tags correspond to EST Clusters of which they are members, and these associations are represented in the database and exposed as filters/attributes. The database also represents groupings of markers within a type into libraries.

Table 1. Example GrameneMart queries

Database	Data set	Filters	Attributes
Query 1: Find the *A. thaliana* genes with orthologs in grape, the EntrezGene IDs of the grape orthologs, and the degree of homology			
Plants Genes	1. *Arabidopsis thaliana* genes	1.a. MULTI SPECIES COMPARISONS: homolog filters: orthologous *Vitis vinifera* Genes: only	1. Homologs: 1.a. GENE: Ensembl, Ensembl gene ID 1.b. ORTHOLOGS: *Vitis vinifera* orthologs, homology type 1.c. ORTHOLOGS: *Vitis vinifera* orthologs, % identity
	2. *Vitis vinifera* genes[a]		2. Features: 2.a. EXTERNAL: external references, EntrezGene ID
Query 2: Which stop codon-introducing SNPs in *Arabidopsis thaliana* are associated with scorable phenotypes, and with what probability?			
Plants Variations	1. *Arabidopsis thaliana* variations	1.a. GENERAL VARIATION FILTERS: variation source, with phenotype data: only 1.b. GENE ASSOCIATED VARIATION FILTERS: consequence type: STOP_GAINED	1. Variation: 1.a. SEQUENCE VARIATION: variation information, variation ID 1.b. SEQUENCE VARIATION: variation annotation, phenotype name 1.c. SEQUENCE VARIATION: variation annotation, *P*-value
Query 3: List all sorghum RFLP markers and their associated GenBank sequences			
Gramene Markers	1. RFLP markers	1.a. Species: Sorghum bicolor 1.b. Marker: analytical correspondence type : entrez 1.c. Marker: analytical correspondence marker type : GSS	1.a. Markers: species 1.b. Markers: marker name 1.c. Markers: analytical correspondence, analytical correspondence type 1.d. Markers: analytical correspondence, marker name
Query 4: List all clones and their locations on the Maize chromosome-anchored Finger Print Contig (FPC) map from 2006.			
Gramene Mappings	1. Physical mappings	1.a. LOCATION: map species : *Zea mays* 1.b. LOCATION: map set : FPC Chr maize may 2006 1.c. MARKER: marker type : clone	1.a. Location: map set name 1.b. Location: map name 1.c. Location: mapping start 1.d. Location: mapping end 1.e. Marker: marker type 1.f. Marker: marker name
Query 5: Which rice QTL are associated with phenotypes related to yield?			
Gramene QTL	1. qtl	1.a. Species : *Oryza sativa* 1.b. Trait category : yield	1.a. qtl: Qtl accession ID 1.b. qtl: trait category 1.c. qtl: trait name

Notes: labels on the GrameneMart interface may change from release to release. The labels presented were correct for Gramene release 34 (October 2011).

[a] To add the second data set to the query, follow the second 'Dataset' link in the left-hand panel of the interface and select from the '-CHOOSE ADDITIONAL DATASET-' options.

Table 2. Marker types and number of records of each represented as data sets in the Gramene Markers biomart database

Marker type/BioMart data set	Number of records
Amplified Fragment Length Polymorphism, AFLP	8150
Breakpoint interval	303
Centromere	57
Clone	2 242 577
Deletion	333
EST cluster	6 154 296
Expressed sequence Tag, EST	20 690 805
Fluorescent In Situ Hybridization, FISH, Probes.	37
Fingerprint Contig, FPC	17 479
Genome Survey Sequence, GSS	10 653 993
Gene prediction	354 564
Gene	10 781
Genomic DNA	5 263 129
Insertion Site-Based Polymorphism, ISBP	691
Insertion	310
Microarray probe	260 656
mRNA	651 207
Overgo hybridization probe	24 464
Oligonucleotide	2 396 466
Point	332
Primer	80 555
Probed site	11 532
Quantitative Trait Locus, QTL	11 625
Random Amplification of Polymorphic DNA, RAPD	175
Restriction fragment length polymorphism	18 761
Simple Sequence Repeat, SSR	24 422
Sequence Tagged Site, STS	3437
Telomere	20

The Gramene Markers database allows users to filter by, and export attributes related to species, germplasm, name/synonym, library/source and related (corresponding) entities. An example query against the Gramene Markers database is included as 'Query 3' in Table 1.

Gramene Mappings database

The Gramene Mappings database models the mappings between markers (described above) and molecular maps. The various map types, bin, cytogenetic, deletion, genetic, physical, quantitative trait loci (QTL), sequence, are each modelled as a separate data set.

The Gramene Mappings database allows users to filter by, and export attributes related to species, map set/map (e.g. chromosome), map position, marker name, marker type and analysis. An example query against the Gramene Mappings database is included as 'Query 4' in Table 1.

Gramene QTL database

The QTL database contains details of all QTL in Gramene; currently 11 624 from 10 species. The focus of this database is querying and reporting QTL by Trait Ontology (22) term. An example query against the Gramene QTL database is included as 'Query 5' in Table 1.

Discussion and future directions

Gramene has been a long-standing user of the BioMart software. We have deployed both the Ensembl data transformation and interface configuration, and also developed Gramene-specific transformations from our MySQL-based data resources on custom schemas for which we made extensive use of the BioMart MartBuilder and MartEditor software. GrameneMart has become an integral and well-used component of the Gramene website. We will adopt software updates from both Ensembl and BioMart projects as appropriate.

Gramene's future BioMart efforts will be focused on data federation, both internally between Gramene databases, and externally with third-party data sets. We anticipate that this will significantly increase the utility of Gramene's extensive ontology and phenotypic linkage data.

Acknowledgements

We would like to thank our users for their feedback and support as well as our collaborators and contributors who have supplied Gramene with data. We would also like to thank the BioMart team at the Ontario Institute for Cancer Research (OICR) for developing and supporting the BioMart software, and also to the Ensembl and Ensembl Genomes teams at the European Bioinformatics Institute and Wellcome Trust Sanger Institute for developing and supporting the Ensembl software.

Funding

National Science Foundation (Grant numbers 0703908, 0851652). Funding for open access charge: XXX.

Q5

Conflict of interest. None declared.

References

1. Youens-Clark,K., Buckler,E., Casstevens,T. *et al.* (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Res.*, **39**, D1085–1094.

2. Flicek,P., Amode,M.R., Barrell,D. *et al*. (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
3. Smedley,D., Haider,S., Ballester,B. *et al*. (2009) BioMart - biological queries made easy. *BMC Genomics*, **10**, 22.
4. Kersey,P.J., Lawson,D., Birney,E. *et al*. (2010) Ensembl Genomes: Extending Ensembl across the taxonomic space. *Nucleic Acids Res.*, **38**, D563–D569.
5. Cochrane,G., Akhtar,R., Aldebert,P. *et al*. (2007) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **36**, D5–D12.
6. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
7. Kersey,P.J., Duarte,J., Williams,A. *et al*. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
8. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
9. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
10. UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
11. Sayers,E.W., Barrett,T., Benson,D.A. *et al*. (2011) Database resources of the National Center for biotechnology information. *Nucleic Acids Res.*, **39**, D38–D51.
12. Duvick,J., Fu,A., Muppirala,U. *et al*. (2008) PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.*, **36**, D959–D965.
13. Quackenbush,J., Cho,J., Lee,D. *et al*. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.*, **29**, 159–164.
14. He,X. and Wang,J. (2007) BGI-RIS V2. *Methods Mol. Biol.*, **406**, 275–299.
15. Ouyang,S., Zhu,W., Hamilton,J. *et al*. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.
16. Tanaka,T., Antonio,B.A., Kikuchi,S. *et al*. (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.*, **36**, D1028–D1033.
17. Jaillon,O., Aury,J., Noel,B. *et al*. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.
18. Hu,T.T., Pattyn,P., Bakker,E.G. *et al*. (2010) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.*, **43**, 476–481.
19. Paterson,A.H., Bowers,J.E., Bruggmann,R. *et al*. (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.
20. Tuskan,G.A., Difazio,S., Jansson,S. *et al*. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.
21. Swarbreck,D., Wilks,C., Lamesch,P. *et al*. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
22. Yamazaki,Y. and Jaiswal,P. (2005) Biological ontologies in rice databases. An introduction to the activities in Gramene and Oryzabase. *Plant Cell Physiol.*, **46**, 63–68.
23. Atwell,S., Huang,Y.S., Vilhjálmsson,B.J. *et al*. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **465**, 627–631.
24. Weigel,D. and Mott,R. (2009) The 1001 Genomes Project for Arabidopsis thaliana. *Genome Biol.*, **10**, 107.
25. Sherry,S.T., Ward,M.H., Kholodov,M. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
26. McNally,K.L., Bruskiewich,R., Mackill,D. *et al*. (2006) Sequencing multiple and diverse rice varieties. Connecting whole-genome variation with phenotypes. *Plant Physiol.*, **141**, 26–31.
27. Zhao,K., Wright,M., Kimball,J. *et al*. (2010) Genomic diversity and introgression in O. sativa reveal the impact of domestication and breeding on the rice genome. *PLoS One*, **5**, e10780.
28. Myles,S., Chia,J-M., Hurwitz,B. *et al*. (2011) Rapid genomic characterization of the genus Vitis. *PLoS One.*, **5**, e8219. Q4
29. Kinsella,R.J., Kähäri,A., Haider,S. *et al*. (2011) Ensembl BioMarts: a hub for data retrieval across the taxonomic space. *Database*, in press. Q3