POGs2: A Web Portal to Facilitate Cross-Species Inferences About Protein Architecture and Function in Plants

Michael Tomcal, Nicholas Stiffler, Alice Barkan*

Institute of Molecular Biology, University of Oregon, Eugene, Oregon, United States of America

Abstract

The Putative orthologous Groups 2 Database (POGs2) (http://pogs.uoregon.edu/) integrates information about the inferred proteomes of four plant species (*Arabidopsis thaliana*, *Zea mays*, *Orza sativa*, and *Populus trichocarpa*) in a display that facilitates comparisons among orthologs and extrapolation of annotations among species. A single-page view collates key functional data for members of each Putative Orthologous Group (POG): graphical representations of InterPro domains, predicted and established intracellular locations, and imported gene descriptions. The display incorporates POGs predicted by two different algorithms as well as gene trees, allowing users to evaluate the validity of POG memberships. The web interface provides ready access to sequences and alignments of POG members, as well as sequences, alignments, and domain architectures of closely-related paralogs. A simple and flexible search interface permits queries by BLAST and by any combination of gene identifier, keywords, domain names, InterPro identifiers, and intracellular location. The concurrent display of domain architectures for orthologous proteins highlights errors in gene models and false-negatives in domain predictions. The POGs2 layout is also useful for exploring candidate genes identified by transposon tagging, QTL mapping, map-based cloning, and proteomics, and for navigating between orthologous groups that belong to the same gene family.

Citation: Tomcal M, Stiffler N, Barkan A (2013) POGs2: A Web Portal to Facilitate Cross-Species Inferences About Protein Architecture and Function in Plants. PLoS ONE 8(12): e82569. doi:10.1371/journal.pone.0082569

Editor: Joshua L Heazlewood, Lawrence Berkeley National Laboratory, United States of America

Received October 18, 2013; **Accepted** November 4, 2013; **Published** December 10, 2013

Funding: This work was supported by National Science Foundation grant IOS-0922560 to AB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: abarkan@molbio.uoregon.edu

Introduction

The ability to acquire genome sequence data far outpaces the ability to infer gene functions. The vast majority of functional data comes from work with a handful of intensively studied model organisms, each of which is represented by a species-specific database that compiles functional data in the context of predicted and validated gene models. For genes that have not been studied directly, extrapolation of functional data from orthologous genes is a powerful method for inferring function, and comparisons of orthologous gene models can make errors in gene models apparent. The information for such comparisons can be extracted from existing species-specific and phylogenetic resources [1–6], but this requires time-consuming navigation between databases or among numerous pages within the same database. Thus, there is a need for web interfaces that bring together functional and structural annotations for orthologous genes in a manner that simplifies cross-species comparison and functional inference.

In 2007, we developed the Putative Orthologous Groups (POGs) database to meet this need for plant genomes [7]. POGsDB was structured around the inferred proteomes of *Arabidopsis thaliana* (Arabidopsis) and *Oryza sativa* (rice), the two plant species whose complete genome sequences were available at that time. POGsDB collated data from multiple resources into a single-page format that allowed users to view salient functional information for orthologous proteins in rice and Arabidopsis at a glance. Here, we describe a major update, POGs2. POGs2 retains the user-friendly single-page view and flexible search strategies of the original POGsDB, but incorporates two additional plant species [*Zea mays* (maize) and *Populus trichocarpa* (poplar)], two independent sets of ortholog predictions, and numerous interface improvements as described below.

Methods, Results, and Discussion

Overview

The POGs2 database is a MySQL database that organizes information about the inferred proteomes of Arabidopsis, rice, maize, and poplar in the context of Putative Orthologous Groups (POGs). Because large-scale orthology prediction is a developing art [8], we imported orthologous groups from two sources that employed different methodologies: orthologs imported from Gramene (http://www.gramene.org) [2] were calculated with the Ensembl pipeline [9,10], and orthologs imported from PLAZA (http://bioinformatics.psb.ugent.be/plaza/) [11] were calculated with OrthoMCL [12]. Gramene-calculated POGs are displayed by default, accompanied by a statement indicating whether the OrthoMCL prediction from PLAZA is in agreement; if not, the interface allows easy navigation between the alternative POG predictions.

Other data in the POGs2 database was drawn from the following sources. Inferred proteomes with associated gene descriptions and InterPro domains [13] were imported from: *Arabidopsis thaliana* genome v10, ftp://ftp.arabidopsis.org/; *Oryza sativa* genome v7, ftp://ftp.plantbiology.msu.edu; *Zea mays* genome B73 Release 5b (http://ftp.maizesequence.org); *Populus trichocarpa* genome v8 (ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v8.0/Ptrichocarpa/). Experimentally-validated intracellular locations of maize and Arabidopsis proteins were imported from the Plant Proteome Database (http://ppdb.tc.cornell.edu/) [14]. InterPro domains are displayed as graphics drawn from a custom graphics library written in PHP. The library takes the start and end locations of each annotated domain within a gene model, and renders an image scaled to the length of the protein. Each domain is illustrated as a colored ellipse, with the color alternating for each InterPro algorithm. Domain graphics are linked via an image map to the InterPro site for that particular domain.

User Interface: POG-view page

By collaborating closely with experimentalists whose research requires frequent access to this type of data, we developed a user-friendly interface for searches, data visualization, and navigation to complementary web resources. A single-page format allows users to quickly assess key functional and phylogenetic features of POG members without the need to navigate to multiple pages (Figures 1 and 2). Each “POG-view” page includes the following information about each POG member (Figure 1): i) Gene identifiers and imported gene descriptions, with links to each gene in the corresponding species-specific database; ii) Graphical displays of domain architectures, with links to the corresponding InterPro domain descriptions; iii) Experimentally verified intracellular locations; iv) Subcellular targeting predictions for mitochondria, chloroplasts, and the secretory system calculated with TargetP v1.1 [15] and Predotar v1.03 [16], and for the nucleus calculated with NucPred [17]. This information is spatially organized according to data type rather than to gene, in order to facilitate comparisons among orthologs (Figure 1).

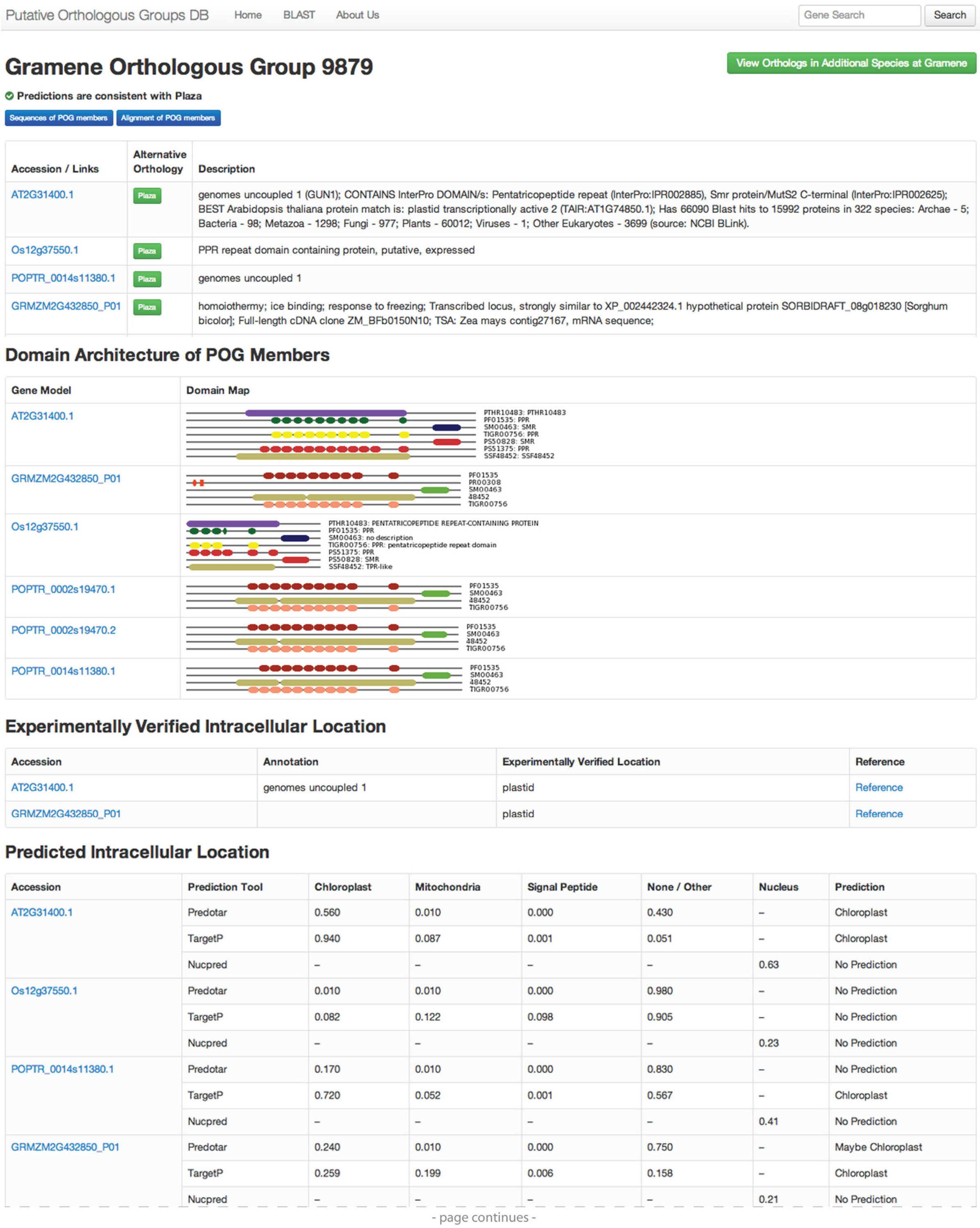
Because the validity of the ortholog predictions is essential to support valid cross-species inference, two tools are provided to help users assess the quality of the predictions. First, a statement at the top of each POG-view page indicates whether the Gramene and PLAZA predictions are in agreement (Figure 1); if not, buttons next to each gene bring up a side-bar showing the corresponding orthologous group from PLAZA (Figure 3). Second, a tree is displayed at the bottom of each POG page with the POG members marked with asterisks (Figure 2); this allows users to quickly assess whether the POG members cluster as expected if they are orthologous. To generate these trees, proteins that are related to POG members were retrieved from an all-against-all BLAST search employing the first gene model for each gene (e.g. Os06g51110.1) as query against the inferred proteomes of the four species in the database. Matches with an E-value < 1e-10 were retained and the top four hits for each POG member (denoted “Closely Related Proteins”) were used to generate multiple sequence alignments with MUSCLE [18]. MUSCLE uses UPGMA to generate phenetic trees during its calculation of multiple sequence alignments. These trees were retrieved, simplified by restricting them to the first two gene models for each gene in the “closely-related” set, and displayed using jsPhyloSVG [19]. Each gene in the tree is represented by a link through which users may navigate to related POGs.

Buttons on each POG-view page allow users to view or download sequences and multiple sequence alignments of POG members (Figure 1, top), as well as sequences, multiple sequence alignments, and domain architectures of closely-related paralogs (Figure 2, bottom). Multiple sequence alignments were calculated with MUSCLE [18] and are displayed with MView [20] using a coloring scheme that reports the chemical properties of each amino acid (i.e. acidic, basic, polar, uncharged, etc) to highlight functional similarity (Figure 4). Gene identifiers for putative orthologs in species beyond the four represented in the POGs2 database can be accessed via links to the corresponding orthologous group in Gramene and PLAZA.

POGs2 employs a client-side JavaScript library AngularJS (http://angularjs.org) to retrieve JSON data files from our PHP API server, and render them into text and graphics By using this method, POGs2 relies largely on the web browser for rendering data rather than relying on the server alone. Thus, the load time is very quick and the server may handle numerous requests simultaneously without a significant decrease in speed.

User Interface: Search Page

POGs2 offers flexible and intuitive options for queries based on gene identifier, keywords, domain names, InterPro domain identifiers, and predicted or established intracellular locations (Figure 5). The keyword search features Boolean operations, with “and” being the default operator when multiple search terms are used. Wildcard searches may be added to a full text query by adding an asterisk (e.g. ‘ribosom*’). A BLAST interface [21] provides access to POGs harboring proteins that are related to a query sequence. To achieve a rapid, responsive search interface we employed Sphinx (http://sphinxsearch.com/) for full text indexing of annotation data, and



Gramene Orthologous Group 9879

✓ Predictions are consistent with Plaza

Accession / Links	Alternative Orthology	Description
AT2G31400.1	Plaza	genomes uncoupled 1 (GUN1); CONTAINS InterPro DOMAIN/s: Pentatricopeptide repeat (InterPro:IPR002885), Smr protein/MutS2 C-terminal (InterPro:IPR002625); BEST Arabidopsis thaliana protein match is: plastid transcriptionally active 2 (TAIR:AT1G74850.1); Has 66090 Blast hits to 15992 proteins in 322 species: Archae - 5; Bacteria - 98; Metazoa - 1298; Fungi - 977; Plants - 60012; Viruses - 1; Other Eukaryotes - 3699 (source: NCBI BLink).
Os12g37550.1	Plaza	PPR repeat domain containing protein, putative, expressed
POPTR_0014s11380.1	Plaza	genomes uncoupled 1
GRMZM2G432850_P01	Plaza	homoiothermy; ice binding; response to freezing; Transcribed locus, strongly similar to XP_002442324.1 hypothetical protein SORBIDRAFT_08g018230 [Sorghum bicolor]; Full-length cDNA clone ZM_BFb0150N10; TSA: Zea mays contig27167, mRNA sequence;

Domain Architecture of POG Members

Gene Model	Domain Map
AT2G31400.1	PTHR10483: PTHR10483 PF01535: PPR SM00463: SMR TIGR00756: PPR PS50828: SMR PS51375: PPR SSF48452: SSF48452
GRMZM2G432850_P01	PF01535 PR00308 SM00463 48452 TIGR00756
Os12g37550.1	PTHR10483: PENTATRICOPEPTIDE REPEAT-CONTAINING PROTEIN PF01535: PPR SM00463: no description TIGR00756: PPR: pentatricopeptide repeat domain PS51375: PPR PS50828: SMR SSF48452: TPR-like
POPTR_0002s19470.1	PF01535 SM00463 48452 TIGR00756
POPTR_0002s19470.2	PF01535 SM00463 48452 TIGR00756
POPTR_0014s11380.1	PF01535 SM00463 48452 TIGR00756

Experimentally Verified Intracellular Location

Accession	Annotation	Experimentally Verified Location	Reference
AT2G31400.1	genomes uncoupled 1	plastid	Reference
GRMZM2G432850_P01		plastid	Reference

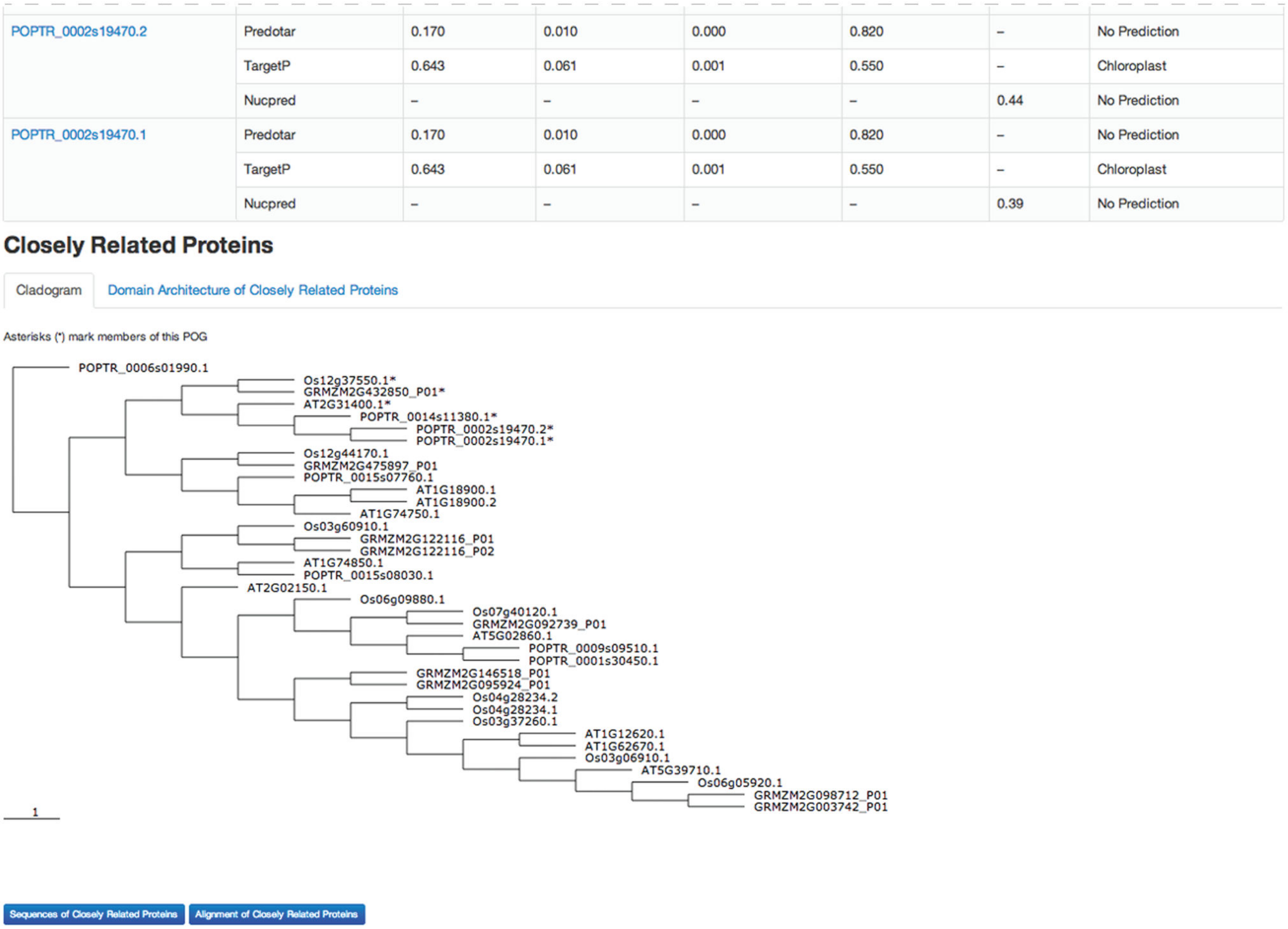
Predicted Intracellular Location

Accession	Prediction Tool	Chloroplast	Mitochondria	Signal Peptide	None / Other	Nucleus	Prediction
AT2G31400.1	Predotar	0.560	0.010	0.000	0.430	–	Chloroplast
	TargetP	0.940	0.087	0.001	0.051	–	Chloroplast
	Nucpred	–	–	–	–	0.63	No Prediction
Os12g37550.1	Predotar	0.010	0.010	0.000	0.980	–	No Prediction
	TargetP	0.082	0.122	0.098	0.905	–	No Prediction
	Nucpred	–	–	–	–	0.23	No Prediction
POPTR_0014s11380.1	Predotar	0.170	0.010	0.000	0.830	–	No Prediction
	TargetP	0.720	0.052	0.001	0.567	–	Chloroplast
	Nucpred	–	–	–	–	0.41	No Prediction
GRMZM2G432850_P01	Predotar	0.240	0.010	0.000	0.750	–	Maybe Chloroplast
	TargetP	0.259	0.199	0.006	0.158	–	Chloroplast
	Nucpred	–	–	–	–	0.21	No Prediction

- page continues -

Figure 1. Screen capture showing upper portion of a representative POG page. Users can quickly assess that this POG prediction is likely to be valid because Gramene and PLAZA predict the same orthologous group (see top), and the members of the group form a clade in the tree (see asterisks in tree at bottom of Figure 2). The layout highlights the orthology of the uncharacterized maize and rice proteins to the characterized Arabidopsis protein GUN1, even though the gene descriptions fail to capture this relationship. The graphical illustration of domain architecture reveals what is likely to be an incorrect gene model in rice. The bottom portion of the same page is shown in Figure 2.

doi: 10.1371/journal.pone.0082569.g001



POPTR_0002s19470.2	Predotar	0.170	0.010	0.000	0.820	–	No Prediction
	TargetP	0.643	0.061	0.001	0.550	–	Chloroplast
	Nucpred	–	–	–	–	0.44	No Prediction
POPTR_0002s19470.1	Predotar	0.170	0.010	0.000	0.820	–	No Prediction
	TargetP	0.643	0.061	0.001	0.550	–	Chloroplast
	Nucpred	–	–	–	–	0.39	No Prediction

Closely Related Proteins

Asterisks (*) mark members of this POG

Figure 2. Screen capture showing bottom portion of a representative POG page. This is the bottom portion of the page shown in Figure 1.

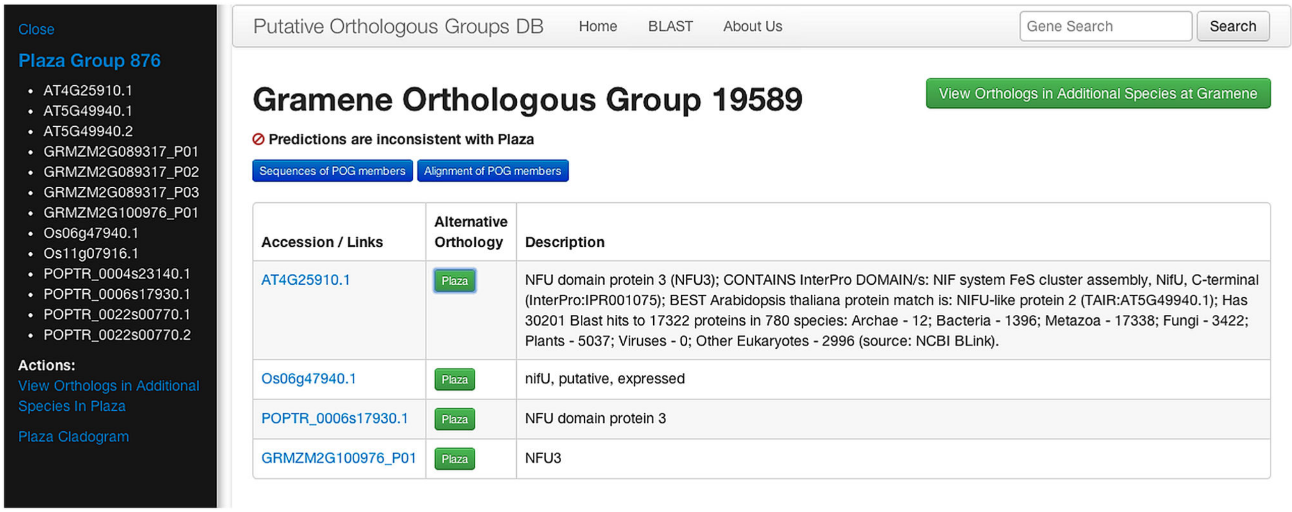
doi: 10.1371/journal.pone.0082569.g002

recent search results are cached in the key-value storage engine Redis (http://redis.io/).

POGs2's Niche Among Genome Database Resources for Plants

POGs2 complements existing database resources for plant genomics by collating a subset of data types available elsewhere in a manner that simplifies protein-centric cross-species functional inferences. POGs2 is distinct in the types of data that are displayed concurrently for orthologous proteins (gene descriptions, domain architectures, intracellular locations, protein trees), in the ease of accessing protein sequences and multiple sequence alignments for orthologous proteins, and in the types of searches that are readily accomplished (e.g. searches for orthologous groups based on both conserved domains and intracellular location). Whereas other resources provide greater depth of information about particular species or greater breadth of phylogenetic data, POGs2 offers a rapid first stop for researchers wishing to glean the essence of what is known about the functions of a particular protein and its orthologs. For this reason, the display is limited to four model plant species, and priority was placed on devising a highly responsive interface that minimizes the need to navigate to multiple pages. Links are provided to species-specific and phylogenetic resources for those wishing to explore more deeply or broadly.

POGs2 is particularly useful as the starting point for evaluating and prioritizing candidate genes that arise through genetic, coexpression, or proteomic analyses. QTL mapping, transposon-tagging, and map-based cloning approaches for identifying genetic variations that underlie phenotypes typically yield multiple candidates. Likewise, coexpression and coimmunoprecipitation data yield sets of genes that are candidates for harboring functions of interest. It is typically both time-consuming and costly to follow up all such candidates, so informed prioritization is a worthwhile first step. By bringing together gene descriptions, domain architectures, and intracellular location data for orthologous proteins in a single page view, and by providing ready link access to coexpression and publication data in Arabidopsis, users working in both "model" and "non-model" plant species can quickly distinguish



Close

Plaza Group 876

- AT4G25910.1
- AT5G49940.1
- AT5G49940.2
- GRMZM2G089317_P01
- GRMZM2G089317_P02
- GRMZM2G089317_P03
- GRMZM2G100976_P01
- Os06g47940.1
- Os11g07916.1
- POPTR_0004s23140.1
- POPTR_0006s17930.1
- POPTR_0022s00770.1
- POPTR_0022s00770.2

Actions:

View Orthologs in Additional Species In Plaza

Plaza Cladogram

Putative Orthologous Groups DB
Home
BLAST
About Us
Gene Search
Search

Gramene Orthologous Group 19589

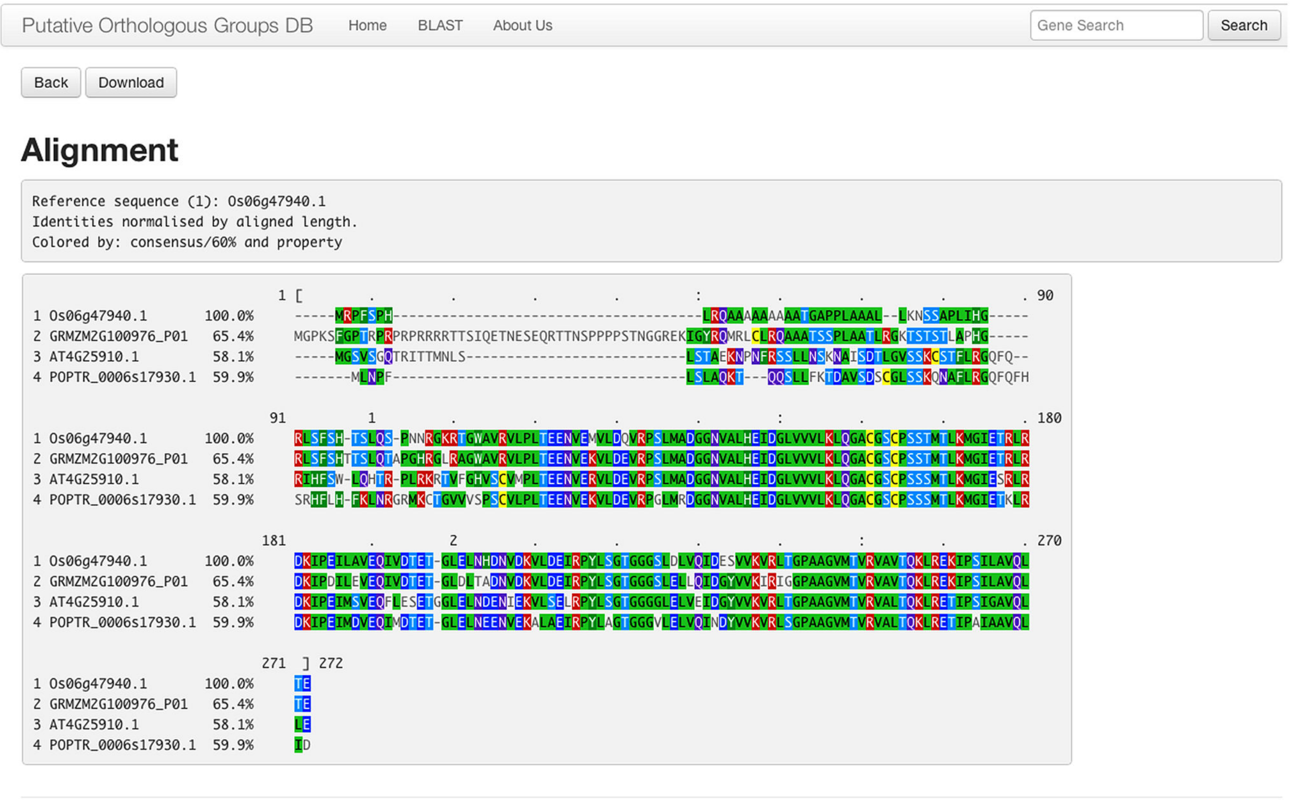
View Orthologs in Additional Species at Gramene

⊘ **Predictions are inconsistent with Plaza**

Accession / Links	Alternative Orthology	Description
AT4G25910.1	Plaza	NFU domain protein 3 (NFU3); CONTAINS InterPro DOMAIN/s: NIF system FeS cluster assembly, NifU, C-terminal (InterPro:IPR001075); BEST Arabidopsis thaliana protein match is: NIFU-like protein 2 (TAIR:AT5G49940.1); Has 30201 Blast hits to 17322 proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).
Os06g47940.1	Plaza	nifU, putative, expressed
POPTR_0006s17930.1	Plaza	NFU domain protein 3
GRMZM2G100976_P01	Plaza	NFU3

Figure 3. Example of sidebar showing alternative orthology predictions when OrthoMCL and Gramene predictions are inconsistent. The PLAZA Group identifier in the side bar is a link to the full POG-view page for that orthologous group.

doi: 10.1371/journal.pone.0082569.g003



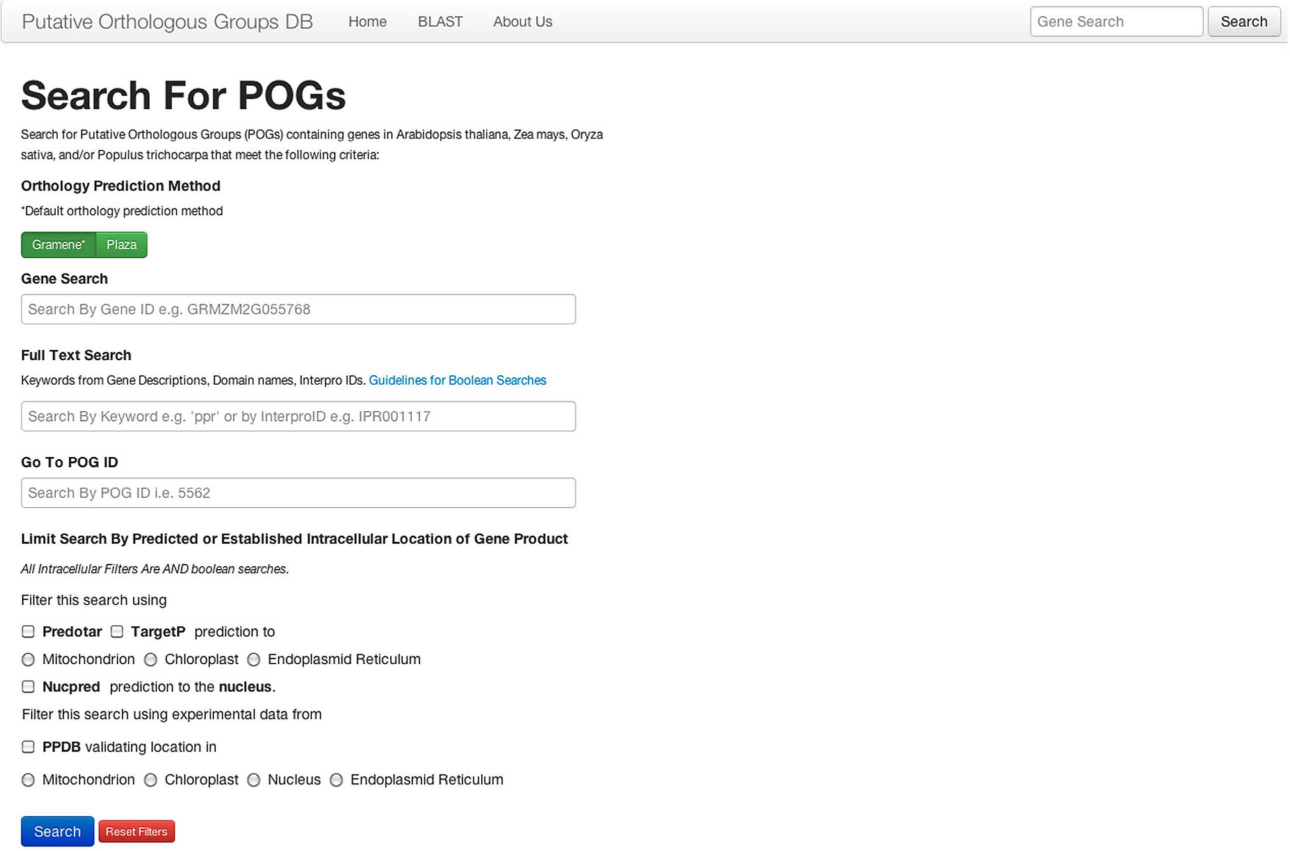
Alignment

```
Reference sequence (1): Os06g47940.1
Identities normalised by aligned length.
Colored by: consensus/60% and property
```

```
                             1 [        .         .         .         .         :         .         .         .         . 90
1 Os06g47940.1       100.0%    -----MRPFSPH--------------------------------------LRQAAAAAAAAATGAPPLAAAL--LKNSSAPLIHG-----
2 GRMZM2G100976_P01   65.4%    MGPKSFGPTRPRPRPRRRRTTSIQETNESEQRTTNSPPPPSTNGGREKIGYRQMRLCLRQAAATSSPLAATLRGKTSTSTLAPHG-----
3 AT4G25910.1         58.1%    -----MGSVSGQTRITTMNLS---------------------------LSTAEKNPNFRSSLLNSKNAISDTLGVSSKCSTFLRGQFQ--
4 POPTR_0006s17930.1  59.9%    -------MLNPF------------------------------------LSLAQKT---QQSLLFKTDAVSDSCGLSSKQNAFLRGQFQFH

                             91         1         .         .         .         .         :         .         .         . 180
1 Os06g47940.1       100.0%    RLSFSH-TSLQS-PNNRGKRTGWAVRVLPLTEENVEMVLDQVRPSLMADGGNVALHEIDGLVVVLKLQGACGSCPSSTMTLKMGIETRLR
2 GRMZM2G100976_P01   65.4%    RLSFSHTTSLQTAPGHRGLRAGWAVRVLPLTEENVEKVLDEVRPSLMADGGNVALHEIDGLVVVLKLQGACGSCPSSTMTLKMGIETRLR
3 AT4G25910.1         58.1%    RIHFSW-LQHTR-PLRKRTVFGHVSCVMPLTEENVERVLDEVRPSLMADGGNVALHEIDGLVVVLKLQGACGSCPSSSMTLKMGIESRLR
4 POPTR_0006s17930.1  59.9%    SRHFLH-FKLNRGRMKCTGVVVSPSCVLPLTEENVEKVLDEVRPGLMRDGGNVALHEIDGLVVVLKLQGACGSCPSSSMTLKMGIETKLR

                            181         .         2         .         .         .         .         :         .         . 270
1 Os06g47940.1       100.0%    DKIPEILAVEQIVDTET-GLELNHDNVDKVLDEIRPYLSGTGGGSLDLVQIDESVVKVRLTGPAAGVMTVRVAVTQKLREKIPSILAVQL
2 GRMZM2G100976_P01   65.4%    DKIPDILEVEQIVDTET-GLDLTADNVDKVLDEIRPYLSGTGGGSLELLQIDGYVVKIRIGGPAAGVMTVRVAVTQKLREKIPSILAVQL
3 AT4G25910.1         58.1%    DKIPEIMSVEQFLESETGGLELNDENIEKVLSELRPYLSGTGGGGLELVEIDGYVVKVRLTGPAAGVMTVRVALTQKLRETIPSIGAVQL
4 POPTR_0006s17930.1  59.9%    DKIPEIMDVEQIMDTET-GLELNEENVEKALAEIRPYLAGTGGGVLELVQINDYVVKVRLSGPAAGVMTVRVALTQKLRETIPAIAAVQL

                            271  ] 272
1 Os06g47940.1       100.0%    TE
2 GRMZM2G100976_P01   65.4%    TE
3 AT4G25910.1         58.1%    LE
4 POPTR_0006s17930.1  59.9%    ID
```

Figure 4. Screen capture of a multiple sequence alignment of POG members. Alignments are viewed by clicking a button at the top of the POG-view page (see Figure 1), and can be downloaded in several different formats. Amino acids are shaded according to their chemical properties to draw attention to potential functional features.

doi: 10.1371/journal.pone.0082569.g004



Search For POGs

Search for Putative Orthologous Groups (POGs) containing genes in Arabidopsis thaliana, Zea mays, Oryza sativa, and/or Populus trichocarpa that meet the following criteria:

Orthology Prediction Method

*Default orthology prediction method

Gramene* Plaza

Gene Search

Search By Gene ID e.g. GRMZM2G055768

Full Text Search

Keywords from Gene Descriptions, Domain names, Interpro IDs. Guidelines for Boolean Searches

Search By Keyword e.g. 'ppr' or by InterproID e.g. IPR001117

Go To POG ID

Search By POG ID i.e. 5562

Limit Search By Predicted or Established Intracellular Location of Gene Product

All Intracellular Filters Are AND boolean searches.

Filter this search using

□ **Predotar** □ **TargetP** prediction to

○ Mitochondrion ○ Chloroplast ○ Endoplasmid Reticulum

□ **Nucpred** prediction to the **nucleus**.

Filter this search using experimental data from

□ **PPDB** validating location in

○ Mitochondrion ○ Chloroplast ○ Nucleus ○ Endoplasmid Reticulum

Search Reset Filters

Figure 5. Screen capture of the POGs2 search page. Gene identifiers from any of the four species represented in the database may be used as a query under Gene Search. Searches for specified domain content, key word, and predicted or established intracellular location may be combined.

doi: 10.1371/journal.pone.0082569.g005

the more promising from the less promising candidates. This aspect has been enormously useful in our own efforts to characterize the gene complement underlying chloroplast biogenesis via a large-scale forward genetic approach in maize [22]. In fact, this was our primary motivation for developing POGs2, as we found that existing online resources (i) required time-consuming multi-page navigation to access the core functional data that is collated in POGs2; and/or (ii) included a wealth of detail that hindered the extraction of the core functional data.

Other benefits of the POGs2 layout include the ease of access to protein sequences and multiple sequence alignments, and graphics that display domain architectures in a manner that highlights likely errors in gene models and false-negatives in domain predictions (see Figure 1). In the future, we plan to incorporate additional proteome data, direct links to selected coexpression databases, button access to graphical representations of expression profile data, and refined POG predictions as methods for large-scale ortholog prediction continue to improve.

Acknowledgements

We are grateful to the many members of the Barkan lab whose suggestions have been critical to developing the POGs2 resource, and to Roz Williams-Carrier for help preparing figures.

Author Contributions

Analyzed the data: MT NS AB. Contributed reagents/materials/analysis tools: MT NS. Wrote the manuscript: MT AB.

References

1. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD et al. (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40: D1178-D1186. doi:10.1093/nar/gkr944. PubMed: 22110026.
2. Jaiswal P (2011) Gramene database: a hub for comparative plant genomics. Methods Mol Biol 678: 247-275. PubMed: 20931385.
3. Rouard M, Guignon V, Aluome C, Laporte MA, Droc G et al. (2011) GreenPhylDB v2.0: comparative and functional genomics in plants.

Nucleic Acids Res 39: D1095-D1102. doi:10.1093/nar/gkq847. PubMed: 20864446.
4. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M et al. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res 35: D883-D887. doi:10.1093/nar/gkl976. PubMed: 17145706.
5. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C et al. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40: D1202-D1210. doi:10.1093/nar/gkr1090. PubMed: 22140109.
6. Sen TZ, Andorf CM, Schaeffer ML, Harper LC, Sparks ME et al. (2009) MaizeGDB becomes 'sequence-centric'. Database (Oxford). p. bap020. PubMed: 21847242
7. Walker NS, Stiffler N, Barkan A (2007) POGs/PlantRBP: a resource for comparative genomics in plants. Nucleic Acids Res 35: D852-D856. doi:10.1093/nar/gkl795. PubMed: 17142226.
8. Trachana K, Larsson TA, Powell S, Chen WH, Doerks T et al. (2011) Orthology prediction methods: a quality assessment using curated protein families. Bioessays 33: 769-780. doi:10.1002/bies.201100062. PubMed: 21853451.
9. Flicek P, Amode MR, Barrell D, Beal K, Brent S et al. (2011) Ensembl 2011. Nucleic Acids Res 39: D800-D806. doi:10.1093/nar/gkq1064. PubMed: 21045057.
10. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R et al. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19: 327-335. PubMed: 19029536.
11. Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C et al. (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol 158: 590-600. doi:10.1104/pp.111.189514. PubMed: 22198273.
12. Li L, Stoeckert CJ Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189. doi:10.1101/gr.1224503. PubMed: 12952885.
13. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK et al. (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40: D306-D312. doi:10.1093/nar/gkr948. PubMed: 22096229.
14. Sun Q, Zybailov B, Majeran W, Friso G, Olinares PD et al. (2009) PPDB, the Plant Proteomics Database at Cornell. Nucleic Acids Res 37: D969-D974. doi:10.1093/nar/gkn654. PubMed: 18832363.
15. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2: 953-971. doi:10.1038/nprot.2007.131. PubMed: 17446895.
16. Small I, Peeters N, Legeai F, Lurin C (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4: 1581-1590. doi:10.1002/pmic.200300776. PubMed: 15174128.
17. Brameier M, Krings A, MacCallum RM (2007) NucPred--predicting nuclear localization of proteins. Bioinformatics 23: 1159-1160. doi:10.1093/bioinformatics/btm066. PubMed: 17332022.
18. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797. doi:10.1093/nar/gkh340. PubMed: 15034147.
19. Smits SA, Ouverney CC (2010) jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. PLOS ONE 5: e12267. doi:10.1371/journal.pone.0012267. PubMed: 20805892.
20. Brown NP, Leroy C, Sander C (1998) MView: a web-compatible database search or multiple alignment viewer. Bioinformatics 14: 380-381. doi:10.1093/bioinformatics/14.4.380. PubMed: 9632837.
21. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402. doi:10.1093/nar/25.17.3389. PubMed: 9254694.
22. Williams-Carrier R, Stiffler N, Belcher S, Kroeger T, Stern DB et al. (2010) Use of Illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy Mutator lines of maize. Plant J 63: 167-177. PubMed: 20409008.