Published online 21 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D561–D565***
doi:10.1093/nar/gks1114

BioGPS and MyGene.info: organizing online, gene-centric information

Chunlei Wu*, Ian MacLeod and Andrew I. Su*

Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA 92037, USA

Received September 15, 2012; Revised October 16, 2012; Accepted October 21, 2012

ABSTRACT

Fast-evolving technologies have enabled researchers to easily generate data at genome scale, and using these technologies to compare biological states typically results in a list of candidate genes. Researchers are then faced with the daunting task of prioritizing these candidate genes for follow-up studies. There are hundreds, possibly even thousands, of web-based gene annotation resources available, but it quickly becomes impractical to manually access and review all of these sites for each gene in a candidate gene list. BioGPS (http://biogps.org) was created as a centralized gene portal for aggregating distributed gene annotation resources, emphasizing community extensibility and user customizability. BioGPS serves as a convenient tool for users to access known gene-centric resources, as well as a mechanism to discover new resources that were previously unknown to the user. This article describes updates to BioGPS made after its initial release in 2008. We summarize recent additions of features and data, as well as the robust user activity that underlies this community intelligence application. Finally, we describe MyGene.info (http://mygene.info) and related web services that provide programmatic access to BioGPS.

INTRODUCTION

Genome-scale science is becoming increasingly common for performing unbiased surveys of gene function. Technologies exist for high-throughput interrogation of genetic variation, gene expression, protein expression, protein modifications, epigenetic variation and other molecular features. Using these approaches, scientists can rapidly identify a set of candidate genes that are relevant to their biological system of interest.

Researchers are then faced with the daunting challenge of prioritizing these genes for further study. Most often, the researcher only has direct knowledge of a few, if any, of those candidate genes. The process of learning the current state of knowledge is aided by a variety of online tools. For example, a researcher might visit NCBI's Gene database to get an overview of a gene's genomic context, sequences and annotations (1). But only consulting NCBI Gene would not be sufficient to get a complete picture. One might also visit the Mouse Genome Database for information on the mouse knockout phenotype (2); or the UCSC genome browser for details on local regulatory elements (3); or Reactome for the relevant pathway context (4); or STRING to browse the local protein network (5). There are literally hundreds if not thousands of online sites with gene annotation information, all having some partially overlapping subset of information relative to the other sites.

While the breadth of available gene annotation resources is impressive and valuable, the fragmentation of this landscape of tools significantly impedes scientific research. Users need to spend significant effort learning the search interface for each tool. Even more importantly, staying abreast of the latest gene annotation resources can be quite challenging. For context, the 2012 NAR Database Issue alone included 192 papers. To know which resource, even within a particular domain, is best from among all available resources typically requires either guessing or simply trying them all.

We created BioGPS to simplify navigating the landscape of gene annotation resources. Originally published in 2009 (6), BioGPS promotes two key principles: community extensibility and user customizability. In this article, we review these two design principles as they relate to BioGPS. We summarize recent updates to both the functionality and the data available within BioGPS. As an application that leverages crowdsourcing, we also provide an overview of usage statistics that underlie our critical mass of users. Finally, for bioinformaticians who

*To whom correspondence should be addressed. Tel: +1 858 784 2079; Fax: +1 858 784 2080; Email: asu@scripps.edu
Correspondence may also be addressed to Chunlei Wu. Tel: +1 858 784 2111; Fax: +1 858 784 2080; Email: cwu@scripps.edu

are interested in accessing BioGPS content, we provide an overview of the application programming interface (API).

COMMUNITY EXTENSIBILITY

As we illustrated in our previous paper (6), the landscape of online gene annotation resources can be accurately described as a 'Long Tail'. While there are a few sites that generate a huge amount of content and value (e.g. NCBI, Ensembl, UniProt, etc.), there are a large number of smaller resources that each produce a smaller amount of specialized content. In aggregate, this Long Tail of gene-centric resources comprises an essential source of content for a complete understanding of gene function.

Organizing data from Long Tail resources represents a significant challenge. BioGPS addresses this challenge using crowdsourcing by enabling our user community to directly contribute to our library of Long Tail resources. In BioGPS, we use the concept of a plugin to represent each individual resource. Each BioGPS plugin is primarily defined by the URL template that includes a variable placeholder for a gene-specific identifier (e.g. NCBI Gene, Ensembl Gene, Refseq, Symbol, etc.). BioGPS currently supports more than 30 commonly used identifiers. When a user views a gene report within BioGPS, the identifiers corresponding to that gene are combined with a plugin's URL template to create a resolved URL.

BioGPS maintains an extensive library of such gene-centric plugins. Adding a new plugin into BioGPS is very straightforward, and any registered user is allowed to add to the BioGPS plugin library. The plugin library currently contains more than 280 public plugins (in addition to more than 200 plugins with restricted access) covering a wide diversity of gene-centric resources (Supplementary Table S1). The number of plugins has almost tripled over the past 35 months since we first described the BioGPS plugin library in our previous paper (6). These plugins offer a broad range of gene-centric information, including gene expression data, pathway information, genetics tools, reagents, literature searching and protein information. Any online, gene-centric resource is within scope for inclusion in the plugin library, and virtually all such sites are technically simple to register. In total, more than 113 BioGPS users have contributed to BioGPS by registering at least one plugin, and these plugins span more than 150 hosting domains. The plugin library also has been extended with a tagging, rating and commenting system. In addition, BioGPS users can flag broken or non-functional plugins, which will notify both the plugin owner and BioGPS developers.

The most frequently used plugin is the 'Gene expression/activity chart' viewer. This plugin (maintained by our group) was originally constructed to display the 'Gene Atlas' data, a reference gene expression data set focused on defining the 'normal transcriptome' (7,8) using a simple and intuitive bar-chart visualization. Recently, we added more than 2000 human microarray data sets from NCBI's GEO repository (9). These data records also include extensive metadata for improved searching and visualization. Despite the massive increase in the size of our database, an improved data model and extensive indexing actually resulted in improved performance.

USER CUSTOMIZABILITY

Like other gene-centric websites, BioGPS provides a default gene report page that displays content that we (the BioGPS developers) think will be used by the majority of our users. Most sites limit users to such predefined views. In contrast, BioGPS allows users to customize the content in their gene report, creating a unique page that is tailored to the user's specific use cases. A gene report page is comprised of multiple window widgets in the user's browser, where each window renders content for a single BioGPS plugin (Figure 1). Plugin windows can be arranged and resized according to the user's preferences, and these gene report layouts can be saved for easy access. For example, a structural biologist may find a layout with the PDB, PFAM and UniProt plugins to be most useful, whereas a systems biologist may refer more commonly to a layout that includes the KEGG, Reactome and WikiPathways plugins.

By default, BioGPS provides access to 12 predefined layouts corresponding to common use cases. For example, there are layouts for literature searching, for browsing model organism databases, and for viewing relevant pathway data. In addition, BioGPS users have saved more than 2200 custom layouts.

For mobile devices without a capable browser, we automatically redirect users to our mobile website (http://biogps.org/m/). For iPhone users specifically, a 'BioGPS' iPhone App is also available for installation. Both the mobile website and the iPhone app provide the access to a variety of layouts, and the plugins in each layout are displayed as hyperlinks for easy access.

USAGE STATISTICS

We described earlier how all crowdsourcing applications depend on maintaining a positive feedback loop between utility, usage and contributions (10–12). To summarize, crowdsourcing applications need to provide users some basic utility, which then will attract some number of users to the site. Some percentage of those users will make a contribution of content, thereby making the site more useful and drawing even more users and contributors. Applications that fail to achieve this positive feedback loop eventually stagnate, whereas sites that achieve critical mass continue to grow and expand in a way that naturally scales with the user community.

The utility of BioGPS is defined by the two design principles described earlier—community extensibility and user customizability. The extensive gene-centric plugin library and the system for creating custom gene report layouts provide functionality that is not available elsewhere. In addition, utility is defined by the simple access and visualization of more than 2000 expression data sets from

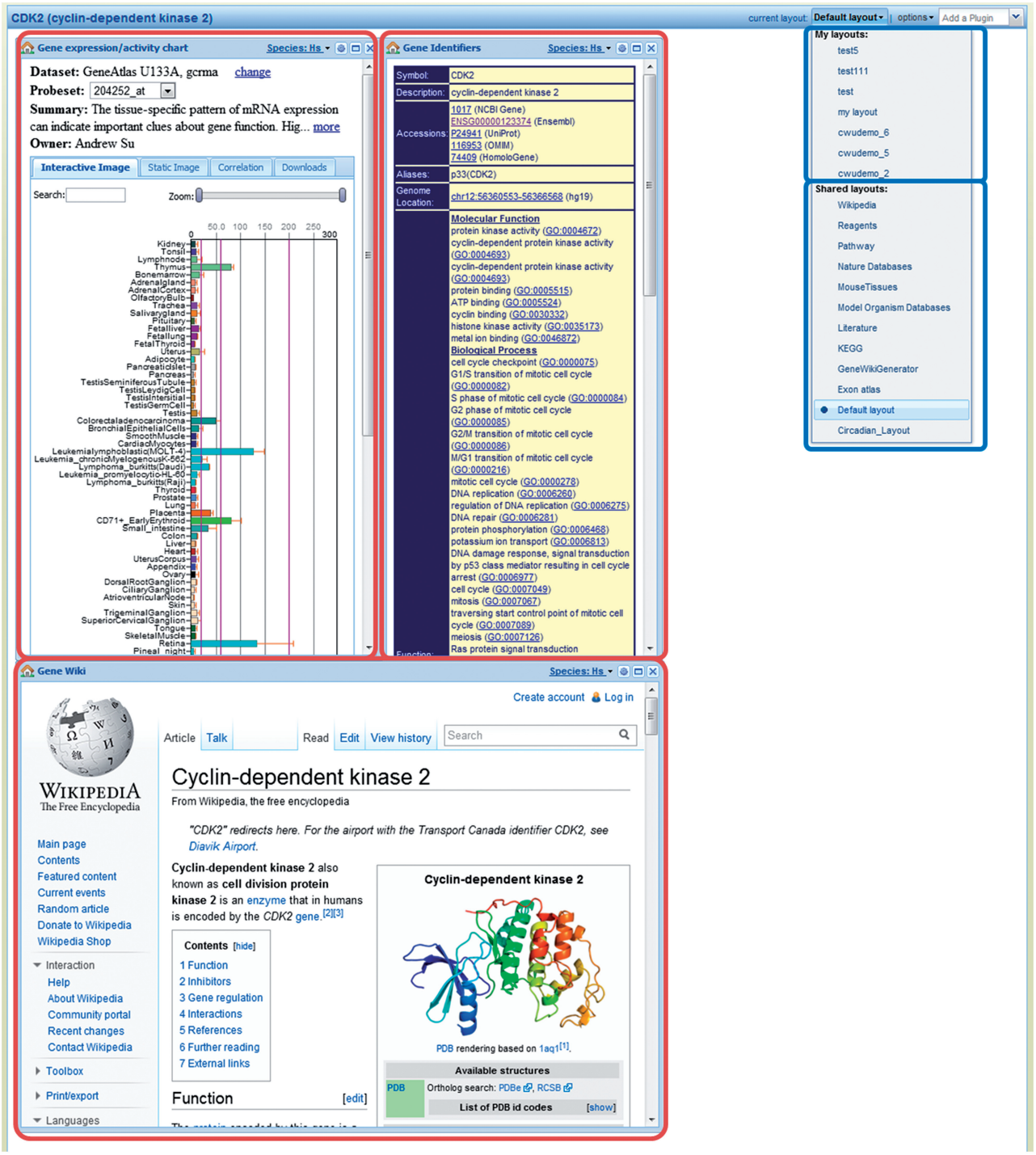

Figure 1. The default BioGPS gene report layout. BioGPS gene reports are composed of collections and arrangements of plugins (red boxes). Each user can customize the selection and configuration of plugins into layouts (selection in the blue boxes). BioGPS provides 12 default layouts (the lower blue box) that correspond to common use cases.

GEO. BioGPS now supports nine species that are commonly used in biomedical research.

BioGPS usage is measured using Google Analytics and our internal logging systems. Each month, we currently average more than 155 000 page views from ~13 500 unique users. For comparison, those numbers grew steadily from 100 000 monthly pageviews and 7000 unique users in 2009. More than 5000 users have registered for a BioGPS user account (up from 900 in 2009).

Contributions from users are instrumental for the growth of BioGPS. As BioGPS employs a simple plugin registration system, more than 100 users have registered at least one plugin that is shared with other BioGPS users, and in total we have more than 280 publicly available

plugins. By tabulating plugin usage across all ~2200 user-saved layouts, BioGPS can tabulate which plugins are most commonly used (Supplementary Table S1). Sorting plugins by popularity is a useful metric when users search the plugin library to discover new resources corresponding to biological keywords (e.g. 'SNP', 'splicing', 'pathway'). These popularity metrics make BioGPS a useful resource discovery tool, in addition to providing convenient access to known resources.

APPLICATION PROGRAMMING INTERFACE

For developers, BioGPS provides an extensive set of REpresentational State Transfer (REST) based APIs for programmatic access of BioGPS resources. BioGPS offers three categories of web service APIs, described later, and documented at http://biogps.org/api.

Gene query API

The BioGPS plugin system depends on a fast and reliable system to search genes by identifiers and keywords. BioGPS supports more than 30 such identifiers across nine common species. As gene searching and resolution are key features for many online biology-oriented sites, we abstracted the underlying features into a public gene annotation web service provider, called MyGene.info (http://mygene.info).

MyGene.info offers two simple REST web services. The Gene Query service allows searching by any commonly used identifier or by genome intervals, and it returns a list of canonical gene identifiers (NCBI Gene or Ensembl Gene IDs). The gene annotation service accepts as input a canonical identifier and returns a comprehensive list of synonymous identifiers and gene annotations. Both services return JavaScript Object Notation (JSON) formatted data, making them easy to use in web applications. Detailed documentation and usage examples can be found at the http://mygene.info website. While most developers will find it straightforward to call the MyGene.info web services directly, we also provide a Python client library ('mygene') for even tighter integration with Python applications (http://pypi.python.org/pypi/mygene/). Web developers can also utilize a customizable JavaScript-based autocompletion widget that easily enables online gene searches (http://mygene.info/doc/widget/autocomplete).

MyGene.info currently contains ~460 k genes from nine common species, and gene annotation data are regularly updated once per month. MyGene.info is built on CouchDB, a document-based database. Unlike the case in more commonly used relational database systems (e.g. Oracle, MySQL), data are stored as key-document pairs. The 'document' is a JSON-formatted gene annotation object, whereas the 'key' is a gene ID (NCBI or Ensembl). The hierarchical structure of gene annotation data can be represented naturally in this key-document model. Using preindexed views and indexing using Lucene (http://lucene.apache.org), MyGene.info can achieve high query performance which sufficiently powers at least 500 concurrent users. MyGene.info source code is available via an open source license (https://bitbucket.org/newgene/genedoc/overview).

Plugin API

The BioGPS plugin library provides a relatively comprehensive collection of online gene-centric resources. Using the Plugin API, external developers can access the metadata of all publicly registered plugins, including the deep-linking syntax, in JSON format. For a given plugin and gene ID, we also provide a web service to get the rendered URL from the URL template registered by the plugin owner.

Data Set API

As described earlier, the BioGPS 'Gene expression/activity chart' plugin provides intuitive visualization of more than 2000 expression data sets from GEO repository. Developers can access all the underlying data via our Data Set API. JSON-formatted data set metadata can be retrieved via a simple GET request with a given GSE ID from GEO (or an internal data set ID for non-GEO data sets). Moreover, developers can easily retrieve the individual data row or a PNG-formatted bar-chart image for a given reporter from any loaded data set.

IMPLEMENTATION

BioGPS was built on top of the Django web framework with PostgreSQL as the database backend. EXTJS and jQuery Javascript libraries, together with the latest HTML5 technology, were extensively used to render the front-end interface.

Up-to-date gene annotation data were retrieved regularly from common data sources (predominantly NCBI and Ensembl), and then stored in the MyGene.info CouchDB instance and indexed via Lucene. BioGPS interfaces with MyGene.info via the REST-based web services for gene annotation queries and retrieval. Both BioGPS and MyGene.info are hosted in Amazon's Elastic Compute Cloud (EC2).

BioGPS is also available as an iPhone app (http://biogps.org/iphone/) that allows users to access all of the key features of the web application through their mobile device.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1.

ACKNOWLEDGEMENTS

The authors acknowledge contributions from Marc Leglise and Camilo Orozco, and many helpful comments and suggestions from the BioGPS user community.

FUNDING

National Institutes of Health [GM089820, GM083924 to A.I.S.]. Funding for open access charge: United States NIH [GM083924].

Conflict of interest statement. None declared.

REFERENCES

1. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
2. Eppig,J.T., Blake,J.A., Bult,C.J., Kadin,J.A., Richardson,J.E. and Mouse Genome Database Group. (2012) The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.*, **40**, D881–D886.
3. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H., Coelho,A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
4. 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.
5. Szklarczyk,D., Franceschini,A., Kuhn,M., Simonovic,M., Roth,A., Minguez,P., Doerks,T., Stark,M., Muller,J., Bork,P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
6. Wu,C., Orozco,C., Boyer,J., Leglise,M., Goodale,J., Batalov,S., Hodge,C.L., Haase,J., Janes,J., Huss,J.W. 3rd *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.
7. Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M., Moqrich,A. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
8. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
9. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
10. Good,B.M., Clarke,E.L., de Alfaro,L. and Su,A.I. (2012) The Gene Wiki in 2011: community intelligence applied to human gene annotation. *Nucleic Acids Res.*, **40**, D1255–D1261.
11. Huss,J.W. 3rd, Lindenbaum,P., Martone,M., Roberts,D., Pizarro,A., Valafar,F., Hogenesch,J.B. and Su,A.I. (2010) The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acids Res.*, **38**, D633–D639.
12. Huss,J.W. 3rd, Orozco,C., Goodale,J., Wu,C., Batalov,S., Vickers,T.J., Valafar,F. and Su,A.I. (2008) A gene wiki for community annotation of gene function. *PLoS Biol.*, **6**, e175.