A. SPECIFIC AIMS

In this proposal, we describe the development of a new gene portal called BioGPS (Biological Gene Portal Services), which specifically targets a particular segment of users and developers called "The Long Tail". In short, we will develop a gene portal platform which leverages small individual contributions from the large scientific community to create a more complete view of gene annotation. BioGPS will provide several interfaces to support community-generated content: for data providers to contribute new data sets, for researchers to contribute gene annotation, and for developers to contribute new functionality. These interfaces effectively eliminate any centralized roadblocks from the critical path for addition to and improvement of BioGPS. Moreover, the focus on The Long Tail allows BioGPS to remain complementary to, and not in competition with, existing gene portal resources. We propose the following four Specific Aims:

Specific Aim #1: Incorporate community-generated data by allowing users to upload custom numeric data sets for analysis and visualization. Although we will continue to provide many useful reference gene expression data sets (e.g., Gene Atlas, NCI60), BioGPS will allow end-users to upload their own gene-centric numeric data for display in gene reports.

A) Create data import tool to upload structured data matrices into BioGPS. We expect that these data will primarily consist of data from gene expression experiments, but will likely also include data from large-scale RNAi and cDNA genomic screens.
B) Enable users to easily import published microarray data sets from public repositories. Since all journals now require submission to microarray data repositories like NCBI's GEO, this function will allow users to easily access and upload existing microarray data for searching and visualization in BioGPS.

Specific Aim #2: Incorporate community-generated gene annotation by seeding a "gene wiki" with structured gene portal content. This effort will form the foundation for creating a continually-updated peer-reviewed review article for every gene in the mammalian genome.

A) Use structured annotation data to seed gene stubs in Wikipedia. These gene stubs will contain structured gene annotation (e.g., genomic location, GO annotation) from public databases, and will serve as a scaffold for subsequent free-text editing by the BioGPS and Wikipedia communities.
B) Wrap Wikipedia gene content in BioGPS gene reports. In anticipation of a growing amount of content (and scientific value) in the Wikipedia gene pages, BioGPS will display this rich, unstructured, and community-generated gene annotation in a dedicated section of each corresponding gene report.

Specific Aim #3: Incorporate community-generated plugins by creating simple programmatic interfaces for external developers to extend BioGPS functionality. This plugin architecture will enable the BioGPS platform to be extended with new functionality by third-party bioinformaticians and data providers.

A) Provide plugin interface to add custom gene report content. BioGPS will allow external developers to develop plugins which add gene-specific content in the main annotation area of the gene report.
B) Provide plugin interface to add custom search interfaces. This interface will allow external developers to develop custom search functions, the results of which can be sent as gene lists to BioGPS for viewing annotation.

Specific Aim #4: Enable users to share and customize the usage and layout of BioGPS plugins through optional user accounts. We envision an open marketplace where anyone can contribute new content for community use, and each user can customize how that community content can best serve their needs.

A) Create data set and plugin libraries from which users can easily browse and select community-contributed content. Developers and data providers will also be able to easily and independently register their content for public display.
B) Create plugin containers and a layout manager that gives users complete control over gene report display. Using a rich user interface, users will be able to reposition and resize each plugin according to their individual needs and use cases.

Successful completion of these aims will result in a gene portal platform that enables The Long Tail of researchers to collaboratively contribute to our understanding of mammalian gene function.

B. BACKGROUND AND SIGNIFICANCE

Web 2.0 and The Long Tail. The moniker "Web 2.0" was first coined in 2004 and quickly became a popular phrase to describe a new wave of web application principles. In the opening talk of the first Web 2.0 conference, Tim O'Reilly cited a defining principle which is particularly relevant to this proposal – leveraging the power of "The Long Tail". This principle is derived from the observation that many patterns in economics and society (and indeed, biology [2]) follow a power law, and that recent trends are shifting emphasis toward the tails of those distributions (**Figure 1**).

This principle behind The Long Tail is often illustrated by comparing internet commerce giant Amazon with traditional physical bookstores [3]. Physical bookstores tend to only carry very popular titles, since high overheads require that a certain number of units be sold to make a profit. This business model focuses on selling a large number of units of a relatively small number of products ("The Short Head"). In contrast, online businesses like Amazon can offer a greater variety of products, including a large selection of relatively obscure products. Analysis of Amazon's sales showed that a large proportion of total sales were attributable to these obscure products, selling a small number each of a relatively large number of products ("The Long Tail") [3]. This phenomenon has also been observed other areas of Web 2.0 commerce, from video rentals (Blockbuster versus Netflix) to music sales (music stores versus Apple's iTunes). (A more detailed overview can be found at http://www.longtail.com/about.html.)

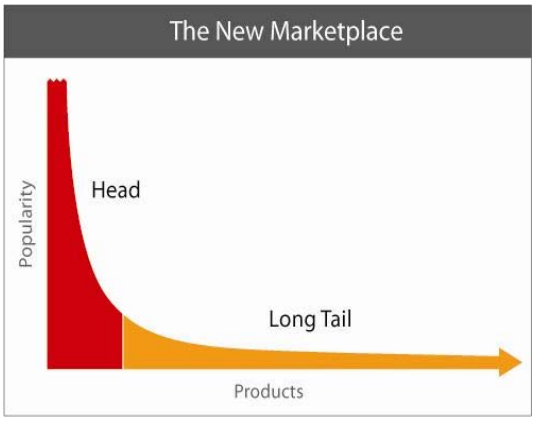

Figure 1. Illustration of "The Long Tail". [1]

Importantly, the principle of The Long Tail also extends beyond internet economics. In the area of mass media, The Long Tail is equally-well typified by another Web 2.0 star – Wikipedia. In the traditional media world, production of information relies on large contributions by a few key players – newspapers, encyclopedias, textbooks. In contrast, Wikipedia inverts the control of information by inviting everyone to contribute to the creation of a comprehensive online encyclopedia. In Wikipedia, the majority of overall content is created through small contributions by a huge number of authors – in short, by harnessing The Long Tail. In the Wikipedia world, the large audience of information *consumers* is also the community of information *producers*. In summary, **The Long Tail** embodies the principle of collaboratively harnessing small individual contributions from a large community.

In the biological community, the two most widely-used gene portals are Entrez Gene (hosted by NCBI) and Ensembl (hosted by the EBI and Sanger Institute). Both of these resources are popular and useful tools for researching gene annotation. These databases provide each gene's protein and transcript sequences, genome location and genomic structure, aliases, and gene function. These sites are considered the definitive resources for these types of gene annotation. However, these existing gene portals are almost entirely aimed at The Short Head of gene annotation. Because these resources are considered authoritative sources, all data is carefully scrutinized prior to incorporation, and only the most broadly appealing data sources are considered. These restrictions effectively limit contributions of annotation to The Short Head – the small number of scientists who each contribute a large amount data.

In contrast, this proposal will leverage The Long Tail and apply it to all facets of the BioGPS gene portal – data generation, gene annotation and application development. Traditionally, gene portals only display large data sets which are presumed to be universally interesting; **Specific Aim 1** describes how BioGPS will solicit and display more biologically specific data sets that come from and appeal to scientific niches. Traditionally, gene annotation is generated by a few large genome centers and annotation authorities; **Specific Aim 2** describes how BioGPS will leverage Wikipedia to allow the entire community to easily and actively participate in the gene annotation process. Traditionally, development of new functionality in gene portals has rested solely in the hands of portal administrators; **Specific Aim 3** describes simple plugin interfaces to the BioGPS platform to allow new features to be developed and deployed completely autonomously by the community. Finally, given the additional complexity and flexibility that is created by this distributed extensibility, **Specific Aim 4** describes how users will be able to individually tailor the BioGPS site to their needs to avoid data overload.

Gene expression as gene annotation. While both Entrez Gene and Ensembl have been successful in organizing and visualizing textual "tag-value" data, they have not put much emphasis in displaying numeric data. For example, neither gene portal displays microarray gene expression data in gene annotation reports. Both Entrez Gene and Ensembl have sister projects (GEO and ArrayExpress) that specifically handle gene expression data sets; however, these resources primarily function as data repositories rather than analysis and visualization tools in the context of a gene portal. We believe that microarray data can often be used as another form of gene annotation. This is particularly true in the context of large reference data sets like the Gene Atlas, which interrogated gene expression across diverse anatomic regions in human and mouse tissues [4, 5]. Smaller gene expression data sets can also provide valuable information. For example, insight into gene function can be revealed from knowing which biological perturbations or experimental conditions have resulted in induction of a gene's expression.

Despite the potential for using gene expression data as additional gene annotation, the currently-available options are limited. Although all major journals require submission of microarray data to a data repositories, as mentioned above, these sites are not designed for viewing these data in the context of other gene annotation. Authors often choose to create web pages or supplementary data files where data can be downloaded (for examples, [6, 7]). However, these raw data files are also not amenable to browsing and visualization.

In a few cases, specific organizations have created rudimentary gene portals for the purposes of incorporating numeric data. These online resources include Stanford SOURCE [8], the Gene Expression Database (GXD) from The Jackson Laboratory [9], and our SymAtlas site [10] (described in more detail below). All of these sites are well used in the community, and many of the Letters of Support from the SymAtlas community reference the power of gene expression patterns as one element of studying gene function. Although these sites are well-tailored to their specific goal of displaying gene expression data, there are several drawbacks. First, there is often significant duplication of effort with the main gene portals. Second, sites often fall into disrepair as graduate students and postdocs who created them move on. Third, these sites to date have not been customizable so that users can upload their own gene expression data. And fourth and most importantly in the contest of this proposal, this option is only open to The Short Head of scientists which have the expertise and resources to devote to creating and maintaining these sites.

We believe there exist an opportunity and a need to create a public resource to view gene expression data as a form of gene annotation in a gene portal. In **Specific Aim 1**, we describe our plans to harness The Long Tail by providing an easy mechanism for data providers to independently upload gene expression data to BioGPS for analysis, publication, and visualization.

Defining gene annotation. Gene reports retrieved from both Entrez Gene and Ensembl are filled with useful gene annotation, including essential information like protein domain structure, genome location, and gene function. These annotations are generally produced by large-scale genome centers and genome-wide analyses. In short, existing annotation is the result of efforts by The Short Head – a relatively few contributors each producing a large amount of content.

As alluded to above, Wikipedia has shown that The Long Tail can also be a fruitful source of information content by soliciting many small contributions from a large population of contributors. This model has produced a quite active and dynamic community. As of January 2008, the English Wikipedia contains over two million articles edited by over six million user accounts. Importantly, a recent study found that the number of monthly contributions from new editors (less than 100 total edits) equals the number of contributions from the most established editors (greater than 10K edits) [11], illustrating the collective importance of The Long Tail. Equally importantly, a study by the journal *Nature* showed that the Wikipedia model of harnessing the Long Tail produces content on scientific topics that rivals the online *Encyclopaedia Britannica* in accuracy [12].

In addition to allowing contributions from The Long Tail, wiki systems have an additional significant difference relative to current gene portals. As mentioned above, gene portals are primarily focused on tag-value annotation types and extensively use ontologies and controlled vocabularies. In short, gene portals are excellent resources for *structured* data. On the other hand, Wikipedia is primarily used for *unstructured* data, which includes free text, images, and figures. In the context of gene annotation, this type of unstructured data

is often presented in review articles as interpretation of data and summaries of community views. Since a wiki system provides the opportunity to effectively manage unstructured contributions from a large community, a gene wiki has the potential for resulting in a collection of gene-focused review articles which have been collaboratively authored, community-reviewed, and continuously updated.

Just as Wikipedia leveraged unstructured contributions from The Long Tail to produce a general-topic encyclopedia, we suggest that The Long Tail of the scientific community will have a similar impact on the quantity and diversity of gene annotation. In **Specific Aim 2**, we describe the development of a gene wiki that will fill this unoccupied but needed niche in gene annotation resources. Although existing gene portals have served and will continue to serve as important resources, this gene wiki will complement and improve our community efforts at annotating the human genome.

Web services interfaces for gene portals. The term "Web services" is most generically defined as "a software system designed to support interoperable machine-to-machine interaction over a network" [13]. The idea of gene portals using Web services (both as providers and as consumers) to extend and customize core functionality is not a new idea. In fact, both popular gene portals mentioned above use Web services interfaces to create "plugins" to their standard gene reports. In this proposal, we focus on the use of Web services by gene portals to incorporate third party content. The programmatic use of BioGPS through Web services is addressed in the Research Design and Methods section.

NCBI's plugin system, called "LinkOut", allows third-party content providers to register links for specific entries in the Entrez database. In Entrez Gene, a link to a page showing available LinkOuts is shown on every gene page. This LinkOut feature allows Entrez Gene users to easily access third-party resources. However, adoption by external data providers for gene resources has been relatively limited. To date, only 23 third-party sites have been registered for LinkOuts from Entrez Gene entries [14], perhaps due to the extensive LinkOut registration process. For example, the registration instructions describe several steps, which include "Applying for inclusion" (a process which takes up to one week) and preparation of an "Identity File" and "Resource File" to submit to NCBI for review [15]. LinkOuts are also constrained by several notable limitations. First, LinkOuts are applied site-wide with no capability for user customization. Second, LinkOuts can only be displayed on the dedicated LinkOut page. Third, display of LinkOuts is limited to text-based hyperlinks to external resources.

Ensembl's plugin interface is based on the Distributed Annotation System (DAS) [16, 17]. Content from DAS sources can be directly included in the standard gene report, and inclusion is user-customizable. The DAS protocol is also a very structured interface in which each element of the transferred data is strongly typed. This highly structured protocol enables the development of DAS clients which rely on specific types of genomic data (for example, genome browsers presenting genomic features [18]). However, the DAS protocol is quite complex for both servers and clients. In the official DAS specification [19], the section covering genomic sequence and annotation alone is over 9000 words. This complexity likely limits DAS users and providers to relatively sophisticated bioinformatics groups, a hypothesis that is supported by an analysis of the services listed at the main DAS registry [20]. Of the 53 functional mouse and human DAS services, 64% were provided by just three institutions: Ensembl (21), Sanger (8), and the EBI (5). Moreover, participation on the BioDAS mailing lists has also not expanded significantly beyond these core groups [21, 22]. Although the entire biological community benefits from DAS-enabled tools, direct use of the DAS protocol to extend Ensembl has not been widely adopted by The Long Tail of bioinformaticians and data providers.

While both LinkOuts and DAS servers provide important functionality for extending gene portals beyond single-site providers of annotation, the limitations highlighted above also represent opportunities for alternate mechanisms for external plugins and Web services. Most notably, both NCBI and Ensembl have chosen to use interfaces which are highly structured. While these highly structured interfaces have several advantages noted above, they also are relatively complex. This complexity effectively limits the use of these interfaces to The Short Head of sophisticated bioinformatics groups. These strongly-typed interfaces also are relatively constrained in how plugin providers can control data rendering, where data providers must delegate all presentation formatting to the portal itself.

We believe there exists an opportunity to create a plugin interface tailored to The Long Tail, the large community with lots of data but comparatively less bioinformatics expertise, to enable these data users to

become data producers. In **Specific Aim 3**, we describe a more flexible and open plugin interface for extending BioGPS which is no more complex than standard HTML and CGI.

The Power Law. Recently, we showed that the pattern of publications on genes in the genome (as well as mention of genes in NIH-funded grant summaries) follows a power law [23]. Simply put, a small number of genes are very well studied (e.g., P53, TNF, VEGF), while the vast majority of genes are mentioned rarely or never in the literature. We suggested that the community's collective goal to annotate every gene in the human genome is hindered by our attention to The Short Head of gene annotation, driven by our tendency to study genes which are easy to study. In the context of BioGPS, we believe that a gene portal which effectively leverages The Long Tail of scientists will be a powerful tool for annotating The Long Tail of genes.

Summary. BioGPS presents a model for gene annotation that is starkly different from, but complementary to, the model employed by existing gene portals. Gene portals and model organism databases serve as authoritative references, and therefore place a high priority on data curation by expert data managers and adherence to data standards. In contrast, BioGPS uniformly lowers the activation barriers to participation in the annotation process, thereby greatly increasing the community of annotators. While each person individually contributes less to the final product, the sum of these small contributions in total will represent a substantial increase in knowledge. While each individual contribution is subject to less oversight, BioGPS will enable the community as a whole (and each user individually) to evaluate the utility of each piece of data. In summary, whereas gene portals offer authoritative resources for well-curated data, BioGPS will offer a rich diversity of data that will be a starting point for further study. With adequate understanding of the relative strengths and weaknesses, we believe that the scientific community will be well served by having both types of tools available.

C. PRELIMINARY STUDIES

Lessons from SymAtlas. Although this proposal is a new application for NIH funding, the proposed work builds on several years of internally funded development of a public gene portal. Our institute built the first instance of our gene portal in 2001 under the name "GNF Gene Expression Atlas" [4, 24]. In a major revision in 2004, we released a second version of the portal under the name "SymAtlas" [5, 10]. Both of these web applications provided the basic functions of a gene portal – integration of data from a variety of primary data sources, synonym and ID resolution, and keyword search capabilities. In addition, these gene portals also served as a simple interface for searching and visualizing gene expression patterns from several published microarray studies [4, 5, 25, 26] (**Figure 2**). In particular, we view our "Gene Atlas" microarray data sets (which show expression patterns across a diverse panel of anatomic tissues) as a useful form of gene annotation.

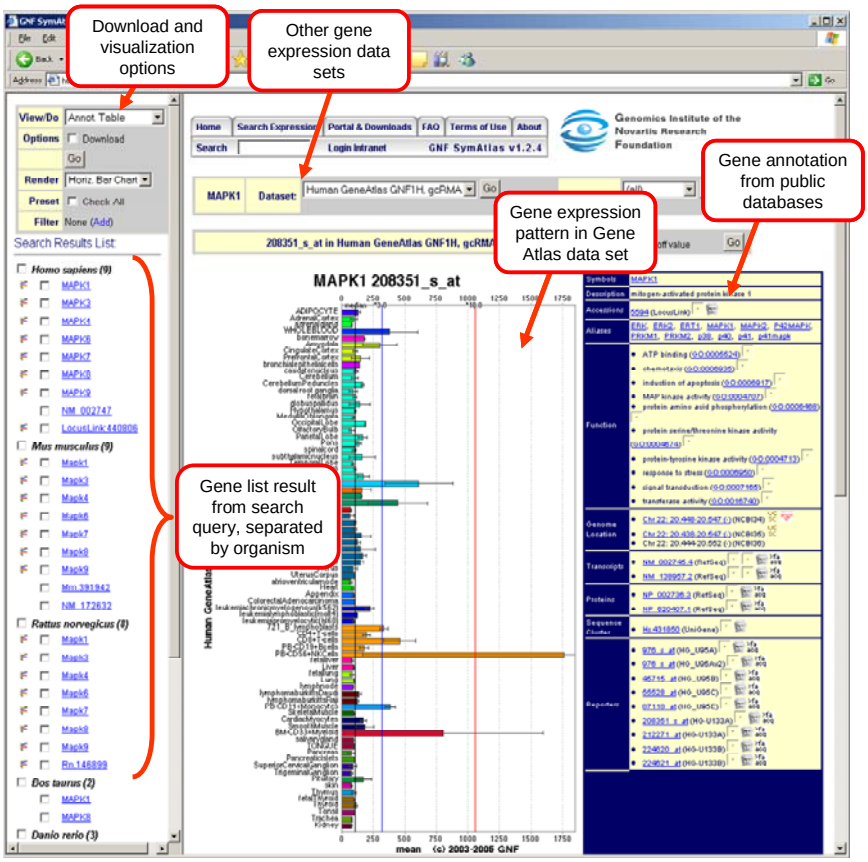

Figure 2. Screen shot from the SymAtlas gene portal (http://symatlas.gnf.org/SymAtlas), the precursor to the BioGPS system.

SymAtlas, like all other gene portals, clearly focuses on content from The Short Head. Annotation data comes from large genome centers, and the display and choice of microarray data is unilaterally controlled by us (the administrators). However, we feel that <u>our experience developing and maintaining a Short Head gene portal provides essential perspective on developing a Web 2.0 gene portal based on The Long Tail.</u> Specifically, we have learned that <u>first</u>, the ability to upload custom data sets is the most common request from users. <u>Second</u>, the most commonly cited feature that distinguishes SymAtlas from other gene portals is its simple user interface that "needs no instruction manual". <u>Third</u>, many technical challenges exist (and have been solved) for maintaining a gene portal like SymAtlas, including, for example, effective methods for synonym resolution and scalable database schema design. <u>Fourth</u>, long-term user growth is exponential, whereas long-term growth in portal developer resources tends to be linear or static.

Figure 3. SymAtlas usage statistics from Google Analytics.

Despite the challenges of developing and maintaining SymAtlas, we have drawn upon this experience in the design of this BioGPS proposal. Moreover, our SymAtlas efforts have resulted in a large, world-wide user community. Analysis of web

site traffic shows an average of 40,000 page views and 3000 user visits per week (over 2 million page views and almost 150,000 visits per year), only counting traffic from outside our institute (**Figure 3**). In addition, the two manuscripts describing our gene portal and the accompanying microarray data have been cited over 900 times in the scientific literature (Science Citation Index). Finally, in a period of two weeks between April 24 and May 10, 2007, we received over 60 letters (36 faculty members, 14 students or postdocs, 11 scientists in industry) in response to a posted solicitation for letters of support for the plan to expand and improve SymAtlas as described herein (see Letters of Support from SymAtlas users). These letters establish the importance of updating and replacing the aging SymAtlas web application. More importantly, because efforts to utilize The Long Tail heavily depend on having critical mass of interested users, we expect that this sizeable SymAtlas user base will greatly enhance the adoption and success of BioGPS.

We believe our experience developing and maintaining the SymAtlas site provides a solid foundation for implementing the proposed BioGPS gene portal, as well as an acute understanding of the need.

Creating gene wiki "stubs". Specific Aim 2 describes how BioGPS will utilize Wikipedia to create a continually-updated peer-reviewed review article for every mammalian gene. The first step in this process, described in **Specific Aim 2A**, is to use BioGPS content to seed "stubs" in Wikipedia for each gene in the BioGPS database. Wikipedia, of course is an online and freely-editable encyclopedia. Wikipedia stubs are short entries that are not yet complete encyclopedic articles, yet they still provide useful information. Importantly, the creation of stubs is often the first step in seeding contributions from the broader Wikipedia. Experience has shown that the activation barrier to editing an existing article is much lower than creating a new article altogether. Within Wikipedia, a nascent "Molecular and Cellular Biology WikiProject" is comprised of a consortium of approximately 100 "Wikipedians" whose focus is on creating and maintaining articles related to molecular and cellular biology. Although significant progress has been made on articles for basic biological concepts (e.g., DNA repair, enzyme, antioxidant), there are relatively few articles on specific genes.

This proposal to integrate content between BioGPS and Wikipedia was presented and discussed in several forums within the Wikipedia community. The idea was first introduced to the MCB WikiProject where it overall was met with enthusiasm and many constructive suggestions for improvement (see Letter of Support from Tim Vickers, MCB WikiProject Coordinator). The format of a gene stub was outlined using IL2-inducible T-cell kinase (ITK) as a sample gene [27]. The ITK gene stub was populated with links to Entrez Gene, Ensembl, Gene Ontology annotations, PDB structures, the Gene Atlas expression pattern, and the UCSC genome browser. Although the data were assembled and the interface was initially designed by our team, many Wikipedians contributed to the final format of the ITK gene stub (**Figure 4**). (Notably, researchers from the Rfam database have proposed a similar project creating stubs for non-coding RNAs with similar enthusiasm.)

After consensus was reached on the value and format of the gene stubs, approval for the automated creation of gene stubs was obtained. Although no approval process is required for manual editing by the community at large, automated editing by computer programs ("bots") involves some oversight to manage server load and prevent large-scale destructive editing. A bot user page was created ("ProteinBoxBot" [28]) which describes the specifications for the bot. Next, the proposal was submitted to the Wikipedia Bot Approval Group for a trial run of 10 genes, which was approved. Finally, to encourage collaboration with local universities, we established a joint project with San Diego State University to develop the bot itself (see Letter of Support from Faramarz Valafar, SDSU).

The first version of the ProteinBoxBot is nearly complete. Given a gene or set of genes, the ProteinBoxBot first downloads all available gene annotation from BioGPS via the gene annotation Web service described above. The XML gene report is then parsed into the "wiki code" to create the gene stub shown in **Figure 4** and subsequently uploaded to Wikipedia. In cases where the ProteinBoxBot detects an existing gene page, the wiki code is written to a log file for manual inspection and integration. As of February 27, 2008, the ProteinBoxBot has been used to create or amend 8634 Wikipedia gene stubs (complete list at [29]). Approximately 650 existing gene pages were amended with the standardized ProteinBoxBot content. ProteinBoxBot had no discernable effect on the edit rate of these pages, with approximately 0.2 edits / page / week in the weeks both immediately before and after the ProteinBoxBot run. ProteinBoxBot also created ~8000 new gene pages in Wikipedia. Although the edit rate on these pages is approximately ten-times lower than for the preexisting gene pages, 1500 edits were subsequently made to these newly created pages.

Because the effort required to create a page far exceeds the effort to edit an existing one, we believe that most of these additions would not have occurred unless ProteinBoxBot had previously created these gene stubs. Moreover, those 1500 edits account for approximately half of all edits to all gene pages after the ProteinBoxBot run, indicating substantial activity on these new pages relative to overall activity. Finally, we found that approximately 66% of all gene wiki pages are in the first page of returned hits at Google when searching by gene symbol (**Figure 5**). In summary, we believe that the gene wiki effort is already reaping substantial benefits, and that activity will increase when these pages are linked from BioGPS.

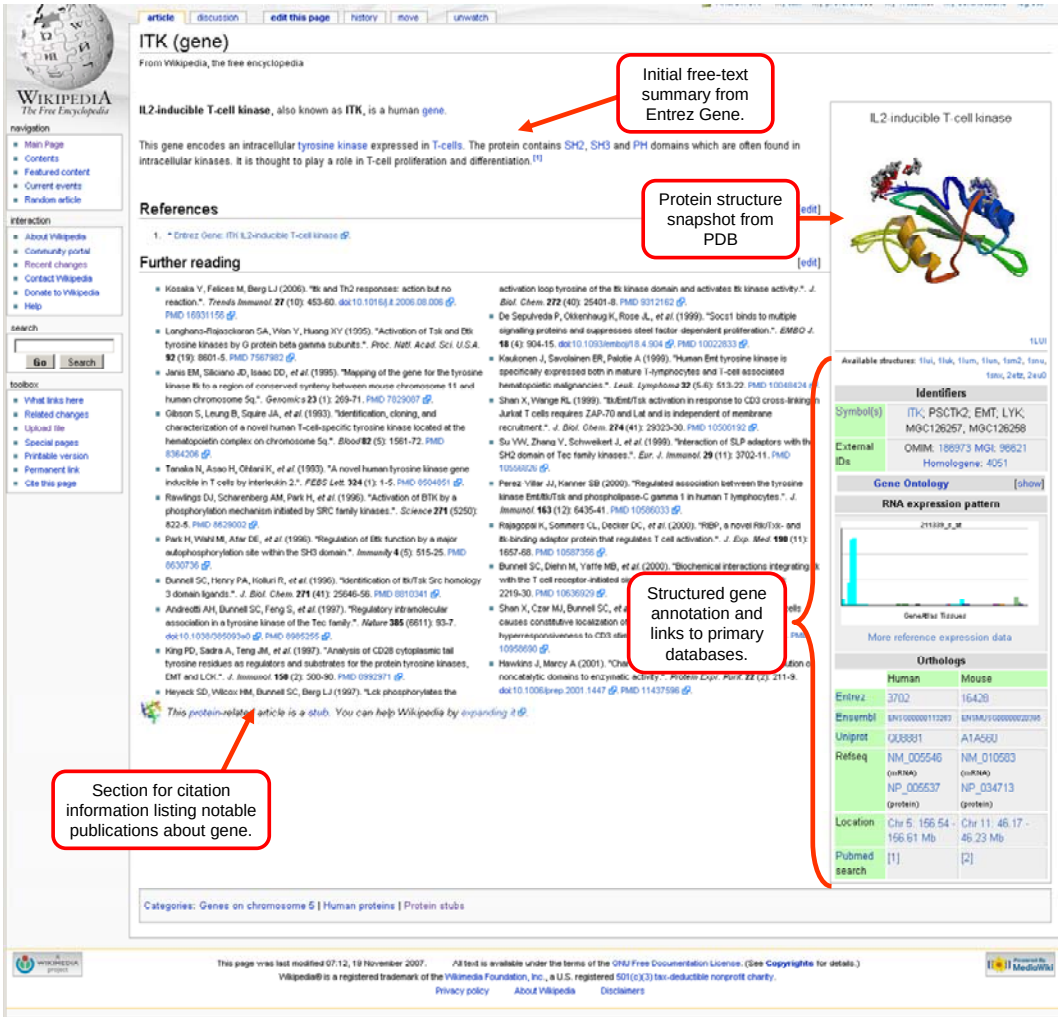

Figure 4. Screen shot of the ITK gene stub at Wikipedia (http://en.wikipedia.org/wiki/ITK_(gene)). To view a gene report with more free text, see the entry for p53 (http://en.wikipedia.org/wiki/p53).

BioGPS working prototype. A working prototype of the BioGPS system is currently being test by a small set of alpha users in our institute. BioGPS, like SymAtlas, performs all basic gene portal functions, including searching by keyword, ID, and genome location, and displaying gene annotation data in a gene report. BioGPS is populated with data from a wide variety of sources, including identifiers and annotation from Entrez Gene, Ensembl, Refseq, Uniprot, Gene Ontology, PDB, MGI, and Pubmed. Importantly, both Entrez Gene and Ensembl release data tables describing identifier relationships with many other databases [30, 31]. These mappings are easily parsed and uploaded to the BioGPS database. To support a primary function of displaying gene expression data from our reference Gene Atlas data set, BioGPS also stores Affymetrix identifiers and gene relationships as computed and updated by the manufacturer. We are confident that the details of storing, searching, and maintaining these relationships will be feasible based on our six years of experience hosting these same data in our SymAtlas site.

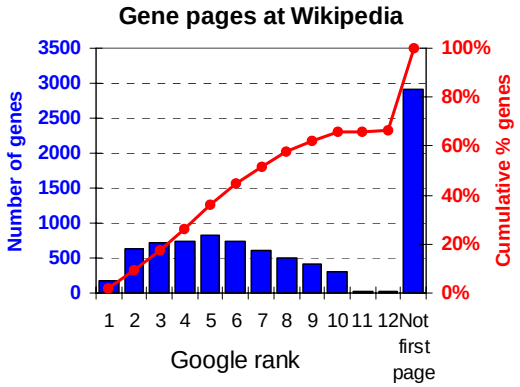

Figure 5. Google rank for gene wiki pages at Wikipedia.

Utilizing the pre-computed relationships released by existing resources allows us to easily update the data in BioGPS with each release of data from the primary gene portals. Moreover, using these data directly from the gene portals allows the BioGPS development effort to focus on the new and complementary features described in the Specific Aims. Importantly, addition of new data sources does not require any changes in the database schema. New relationships can easily be added by simply loading an association

table that relates each ID from the new data source to some existing identifier. Additional details of the design of this prototype are described in the Research Design and Methods section.

Although we currently only have an internal development instance of BioGPS, we expect to have a public prototype version soon (likely before this proposal is reviewed), accessible at http://biogps.gnf.org. Although we recognize that reviewers often choose not to view external URLs associated with grant applications for confidentiality reasons, reviewers who do choose to visit the BioGPS prototype should be assured that access logging is currently disabled on the server. Until we begin actively encouraging our community of SymAtlas users to migrate to BioGPS (as will be clearly noted at http://symatlas.gnf.org/SymAtlas), no efforts will be made to track usage statistics of the BioGPS prototype.

D. RESEARCH DESIGN AND METHODS

Prior to presenting the details of each Specific Aim, we present here four sections that generally apply to all aims of the proposal. Specifically, we discuss The Long Tail and long-term viability, general architecture and design issues, Web services access, and utilization of and contribution to open-source efforts.

The Long Tail and long-term viability. Given the large number of gene portals that already exist, it seems reasonable that any grant proposal to create yet *another* portal must assume a high burden of proof that the new portal would be a significant advance. In addressing this issue, a fundamental question to answer is – *Why are there already so many largely-redundant gene portals in existence?* We believe that this redundancy exists because no one gene portal application satisfies the needs of all users. Hence, we hypothesize that most current developers of gene portals (like ourselves) went through a similar history: 1) encounter a problem well-suited to a gene portal, 2) research existing solutions and find many options that satisfy ~80% of their needs, 3) discover no easy way incorporate the 20% that they need with an existing solution, and 4) resign themselves to duplicating the 80% in order to achieve the unsolved 20%. It would take a monumental effort for a single group to construct a centralized gene portal that anticipates and serves all the diverse needs of the biomedical community, so that a universal solution hasn't been created is not surprising.

Therefore, the goal of this proposal is not to build an *application* that is all things to all people. Rather, we aim to build a *platform* which each user individually, and the biomedical community as a whole, can extend to suit their needs. The open BioGPS platform will enable The Long Tail of users to extend all facets of the gene portal. **Specific Aim 1** will enable the entire community to participate in the accumulation of data, allowing users to contribute or import their own microarray data sets. **Specific Aim 2** will enable the entire community to participate in the gene annotation process, allowing users to contribute free-text knowledge to gene reports. **Specific Aim 3** will enable the entire community to participate in the extension of the BioGPS portal, allowing other bioinformatics programmers to write plugins to enhance BioGPS functionality. Finally, **Specific Aim 4** will provide each user the ability to cherry-pick the specific components of the resulting system that are relevant for their work.

Our experience developing SymAtlas illustrates the challenges to long-term success of a popular web application. As the user community was rapidly growing, so were the number of requests for new data, data sources, and features. Although large genome centers may have the ability to expand developer resources to match demand, we did not. Growing requests for enhancements quickly exceeded our static developer capacity. Therefore, this proposal will specifically target The Long Tail to ensure that BioGPS can grow as quickly and dynamically as its community of users. (It is this same principle that largely led to the widespread success of the social networking site Facebook.) Importantly, we as the BioGPS administrators will not be the rate-limiting step for adding data to BioGPS. Annotators at the large genome centers will not be the rate-limiting step for adding gene annotation. And BioGPS developers will not be the rate-limiting step in improving the functionality of BioGPS. By enabling these mechanisms for users to contribute to the gene portal, the utility of BioGPS will naturally grow as the size of the network of users grows. At the completion of the work described in this proposal, we anticipate that the amount of data, annotation, and code developed and maintained by BioGPS will be far outweighed by the extensions contributed by the biology community. Moreover, we believe that BioGPS can be reasonably maintained for the long term with a relatively small developer core.

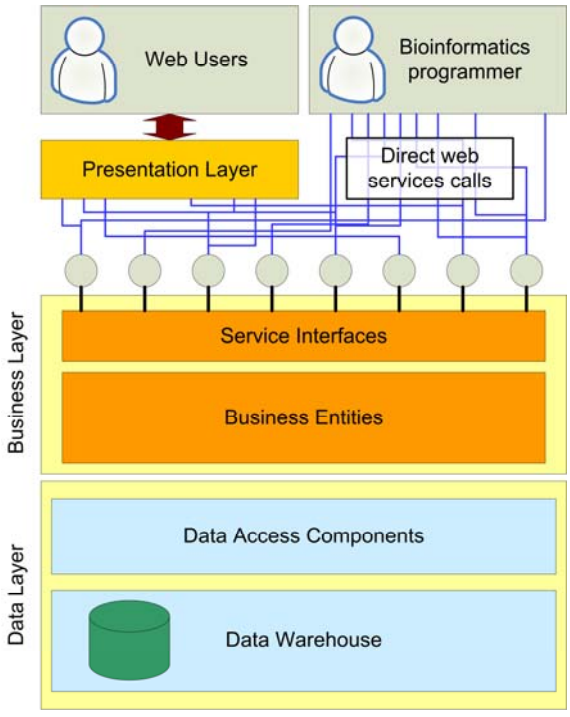

Figure 6. General architecture design of the BioGPS system.

General architecture and design. The BioGPS application is constructed of three modular layers: Presentation, Business and Data Access (**Figure 6**). The **Presentation Layer** provides the application's user interface (UI), composed of visual controls

that render data prepared by the Business Layer. The Presentation Layer is also responsible for binding HTML visual controls to the appropriate objects and methods in the Business Layer. The Presentation Layer is implemented using Django, a high-level web framework based on the Python programming language. While there are many similar application frameworks for rapid web development and prototyping, Django is one of the most mature and actively developed frameworks and also leverages existing expertise.

The **Business Layer** implements the main application logic of BioGPS. It is comprised of two components – the Business Entities and the Services Interface. The Business Entities employ a data tree structure to represent data retrieved from the database through the Data Access Layer. Within a data tree, every node represents a piece of data found for a certain gene, and relationships between objects (e.g., an exon and its corresponding genome location) are represented as a parent-child relationship. This data tree representation allows efficient data searching using mathematical tree traversal algorithms. The Service Interface will expose the search and annotation retrieval functionality as Web services. The Service Interface provides is used for all communication with the Presentation Layer, and exposed Web services can also be directly accessed by external programmers (described in more detail below).

Finally, the **Data Layer** serves as the repository for all data accessible in BioGPS. The Data Warehouse component is implemented in an Oracle schema. The Data Warehouse is an Oracle database schema, where gene objects are stored in hierarchical data trees similar to the business entities described above. The Data Access component provides programmatic access to all search and data retrieval functions to the Business Layer. This component is implemented using ADO.NET, a freely available .NET Enterprise Library. The Data Access layer isolates the Business Layer from the details of data storage, minimizing the impact of changes in data representation or database providers and simplifying testing and maintenance.

The database schema is optimized to enable searching by any keyword or annotation type (**Figure 7**). All annotation data are stored in the XREF table, and each unit of annotation is basically characterized by a source node (XREF_ID), a parent node (XREF_PARENT_ID), and an XREF type (represented by a system ontology, XREF_SYS_ONTOLOGY_ID). For example, an entry in the XREF table could be used to link a Refseq ID to a parent Entrez Gene node according to an ontology classification of “mRNA”. Similarly, a GO annotation could be linked to a parent Refseq ID with an ontology classification of “function”. Heavy indexing the XREF table results in efficient search performance. Each XREF also relates back to the parent of the data tree using the ROOT_NODE attribute, enabling quick retrieval of all annotation for a given gene and fast traversal from any leaf node to the corresponding root. Numeric data are stored separately in the TABULAR_DATA data tables, and are also included in returned gene objects.

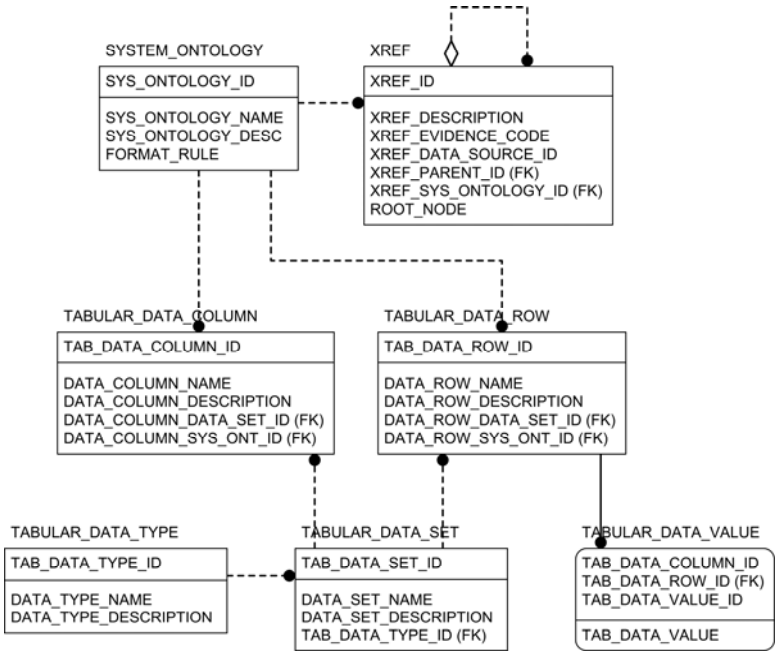

Figure 7. BioGPS database schema describing the data tree representation of gene annotation.

Three application servers will be employed for hosting and developing BioGPS – two production servers and one testing server. One production server will host the BioGPS web application, and the second production server will exclusively serve requests through the Services Interface. Three database instances will also be used – production, staging/loading, and development. Data loading will always be performed on the staging/loading database instance and copied to production to eliminate any impact of data loading on server performance.

Web services access. Although not a Specific Aim, BioGPS will offer Web services access to all search, data retrieval, and data visualization operations. In fact, the use of a technology to build the presentation layer

(Django) that is separate and distinct from the technology used to build the business layer (.NET) necessitates this requirement – all communication between the presentation layer and the business layer will occur through the same Services Interface that is publicly exposed. We anticipate exposing dozens of Web services in the Services Interface. For illustration, we outline three core Web services used for basic gene portal function:

1) Searching for genes according to an ID or keyword query: Inputs to this Web service will include the search term (e.g., "ITK", "MAPK?"), as well as a search type (symbol/ID or keyword). The Web service returns a gene list of all entries that satisfy that criteria.
2) Retrieving a gene annotation report: The external application constructs a Web services query that includes the gene ID of the gene of interest as a parameter. The Web service returns a full gene report with all annotation available in the BioGPS database. (One version of this Web service has already been created for the BioGPS/Wikipedia collaboration, as described in the Preliminary Data section and **Specific Aim 2**.)
3) Retrieving Gene Atlas expression pattern: The external application constructs a Web services query that includes a probe ID or list of probe IDs (parsed from the gene annotation report). The Web service returns the expression data from the Gene Atlas data set (or, if specified, any other data set in BioGPS).

These three Web services are extensively used by our BioGPS presentation layer, and all web services will also be exposed for programmatic access by external bioinformatics users. There are of course many options for data transfer protocols using Web services. There is considerable ongoing debate regarding relative advantages and disadvantages, and many service providers are moving towards accommodating Web services based on both SOAP and REST (the two most common styles) [32-34]. We are utilizing the .NET architecture to create the business layer, and the Windows Communication Framework (WCF) provides simple and native support for creating both SOAP and REST Web services, as well as alternative data formats like RSS and JSON [35]. Therefore, through the use of WCF, BioGPS will easily provide access to all services via a variety of protocols and data formats.

Utilization of and contribution to open-source efforts. We recognize that there are many open source programming libraries (BioPerl, BioJava, etc.) for representing biological objects, as well as open database schemas (BioSQL, Ensembl) for warehousing biological data [36, 37]. However, we have deliberately chosen to develop our own solutions in these areas. We view these open source efforts as solutions for maximizing flexibility. These tools are invaluable for many bioinformaticians, who on any given day may be called upon to conduct analyses as diverse as microarray analysis, gene prediction, phylogenetics, and sequence similarity searches. Open source bioinformatics libraries allow these diverse tasks to be handled within a common framework, maximizing productivity and interoperability. In contrast however, a web-based gene portal platform is tasked with relatively well-defined use-cases, but has much more stringent requirements for efficiency. Because of the interactive nature of a web site's user interface, responses must be delivered in seconds, and efficiency savings of even fractions of seconds can greatly improve usability. With these ideas in mind, we strongly feel that this architecture decision is best for the BioGPS project.

Nevertheless, in addition to flexibility, open source applications and libraries also have the benefits of extensibility and code transparency. Therefore, although we are not significantly utilizing open source projects, the BioGPS project itself will be released via an open source license. This code availability will allow other groups, for example, to mirror a parallel web instance of BioGPS with internal data, to extend the BioGPS application with additional functionality, or to utilize specific components in other projects. Moreover, it is worth noting here that **Specific Aim 3** addresses many of these issues of application extensibility. We anticipate that most users will utilize the plugin interface to extend the main BioGPS instance and benefit from the substantial user community, rather than creating a separate parallel web site. By analogy, many developers build on Google Maps technology using its well-documented API and not direct access to the source code.

Specific Aim #1: Incorporate community-generated data by allowing users to upload custom numeric data sets for analysis and visualization. Although we will continue to provide many useful reference gene expression data sets (e.g., Gene Atlas, NCI60), BioGPS will allow end-users to upload their own gene-centric numeric data for display in gene reports.

Significance and Rationale. The use of gene expression data as a form of gene annotation has been discussed extensively in the Background and Significance section. The significance of gene expression data in biological studies is underscored by the widespread use of our SymAtlas web site and citation of the corresponding manuscripts, by the requirement of all major journals to submit primary data to public data repositories, and by the enthusiastic Letters of Support from existing SymAtlas users. However, despite this importance, microarray data repositories are primarily focused on data warehousing functions, and few gene portals even attempt to use gene expression as a form of gene annotation. Therefore, BioGPS will allow The Long Tail of users to upload their own expression data to a community gene portal for the purposes of searching, visualization, and dissemination. Although BioGPS is initially targeting gene expression data, the capabilities described in this Specific Aim will be equally applicable to large-scale genomic screens on standardized libraries [38, 39].

This functionality offers many advantages relative to existing solutions. First, incorporation of gene expression data into a public gene portal allows for centralization of many portal functions, relieving experimental biologists of many tedious and error-prone operations. These functions include synonym resolution, matching of gene identifiers across different data sources, collection of gene annotation from public databases, and simple searching by expression pattern. Second, display of microarray data alongside other sources of gene annotation reinforces and facilitates the use of gene expression patterns as a form of functional annotation. Although BioGPS will only offer rudimentary analysis capabilities relative to dedicated desktop analysis software, SymAtlas demonstrates how even basic searching and visualization capabilities aid researchers. Third, the contribution of data into an open community-supported gene portal provides all labs the capabilities to publish their data in a common platform, regardless of the level of their own bioinformatics expertise. These capabilities in BioGPS address the current situation where availability and quality of online data access varies widely. In short, successful completion of this Specific Aim allows The Long Tail of the biological community to participate in the generation of data in the BioGPS gene portal.

Issues and Obstacles. Through our experience with the SymAtlas site, we have extensive expertise developing and maintaining a gene portal with the abilities for visualizing and searching gene expression data. Therefore we expect very few significant technical roadblocks to completion of this Specific Aim.

One important conceptual concern for any proposal which aims to harness The Long Tail (and one which we will address for **Specific Aims 1**, **2**, and **3**) is how to draw a critical mass of users. For this Specific Aim, we note that we have a large community of users who already use gene expression patterns as gene annotation. In fact, our most common request by users is exactly the goal of this Specific Aim – to allow users to upload and view custom data sets. We believe that the SymAtlas user community will enthusiastically participate in uploading or importing new data sets into BioGPS to satisfy their own desire to view custom data sets in a gene annotation portal.

One possible issue which would result from tremendous user demand for the ability to upload large scale microarray data would be the exhaustion of our currently allocated disk space capacity. In this case, our first solution will be to expand our storage capacity given the continually falling cost of disk space. Currently, consumer hard disks can be purchased for $0.24 per gigabyte, and storage in our environment (including back-up and high-speed network connections) costs less than 10-fold more. Since BioGPS does not need to store raw data (a function delegated to a true data repository), we estimate that the cost for storing processed data from one modern microarray (~500 KB) to be much less than $0.01. In the face of truly overwhelming demand, the second recourse would be to implement user quotas on uploaded data, though these quotas would likely be generous enough for users to store dozens or hundreds of data sets.

A) Create data import tool to upload structured data matrices into BioGPS. Since the BioGPS database schema already accommodates numeric data, we have begun the customization required to expand the schema to accommodate custom user data. The format for data upload will be standardized according to a simple comma- or tab-separated value text file, a common format for manipulating gene expression data. Parsers for Excel files could also be added if user demand justifies it. Each row will represent a given gene's expression across many samples, and the first column will contain any gene name or ID which can be resolved by the BioGPS database. Each column will contain the expression pattern of a single sample across many genes, and the first row will contain a sample name or ID. Columns which have a common sample

name will be averaged, and error bars will be calculated from the standard deviation. Data validation procedures will flag improperly formatted data and highlight gene names or IDs which do not map to any recognized identifier in BioGPS. Although we expect the most common data type to be gene expression data, users will also be able to upload similarly formatted numeric data sets from, for example, large-scale gain- or loss-of-function cellular screens [40-42].

Moreover, we will implement a "data set library" to which users can optionally contribute their custom data sets for public display (more details in **Specific Aim 4B**). This mechanism for "publishing" microarray datasets in a browsable form can also be used as a supplement to microarray publications. Importantly, the BioGPS gene portal focuses on visualization of numeric data, and therefore it is complementary to the data warehousing functions of public microarray data repositories [43-46].

B) Enable users to easily import published microarray data sets from public repositories. A huge amount of microarray data already exists in public databases like GEO and ArrayExpress [43-46]. Therefore, we will also create a mechanism for users to easily import data from those repositories into BioGPS for searching and visualization. We will enable this utility using two strategies. First, users will be able to upload files which were downloaded from GEO or ArrayExpress without having to reformat the data into the comma- or tab-separated value text file described above. Both repositories provide downloadable files which are similar in format to the simple BioGPS input format (GEO's SOFT format, and ArrayExpress' "processed" format) but differ in several important and sometimes subtle ways. BioGPS will provide a parser to process these files automatically, relieving users of this tedious and error-prone data transformation. Second, users will also be able to simply enter a GEO or ArrayExpress accession number in BioGPS to add a new data set. Because both repositories store processed data in well defined locations on their FTP sites, BioGPS can perform the data download directly if provided with a valid accession number. This second mechanism also provides the added benefit that popular data sets which are requested by multiple users need only be downloaded and stored once in the BioGPS database, alleviating storage requirements.

Specific Aim #2: Incorporate community-generated gene annotation by seeding a "gene wiki" with structured gene portal content. This effort will form the foundation for creating a continually-updated peer-reviewed review article for every gene in the mammalian genome.

Significance and Rationale. In the past decade or more, biology has made huge advances in the ability to generate data in high-throughput. These recently-developed technologies include large scale DNA sequencing, highly-parallel gene expression analysis, functional inference using computational methods, and screening libraries of RNAi reagents. Given their structured nature, these data can easily be stored in a relational database, often in some close variant of a "Gene ID" / "data value" pair. Moreover, by resolving synonymous Gene IDs across databases and technology platforms, gene portals are able to summarize and integrate structured data across many different data sources. Most often, these portals utilize tables and charts for visualization and summarization. In short, gene portals provide excellent support for *structured presentation of structured data*.

Despite this easy access to structured data, the resulting *knowledge* is inherently unstructured, best represented as free-text, illustrative schematics, or pathway models. Most often this sort of unstructured knowledge is only encapsulated in scientific publications. Because the scientific literature is not constructed around a gene-centric model, this knowledge is not amenable to summarization in the context of a gene portal. Moreover, this unstructured knowledge is not easily visualized using the classic tools of structured data, namely tables and charts. Therefore, while existing gene portals primarily focus on presenting *structured data*, there exists an opportunity and need for gene portals to also present *unstructured knowledge* using new media technologies. In addition, the ability to display unstructured knowledge is not useful without a method to assemble the knowledge in the first place. As alluded to above, summarizing knowledge about genes cannot be done automatically, but instead is best done manually by domain experts. However, genome-wide efforts to systematically harness this expert knowledge have been limited.

Wiki systems like Wikipedia solve the problems above by providing a system which utilizes a "bottom-up" community approach. This Specific Aim describes a collaboration between BioGPS and Wikipedia to handle unstructured gene annotation by leveraging The Long Tail. Specifically, we will use structured data in BioGPS

to create gene stubs for all mammalian genes. We anticipate that these gene stubs, like the wealth of other topics at Wikipedia, will seed manual annotation and summarization by the biological community. Historically, as more users view content, more users contribute content and link from other web sites and Wikipedia pages, which in turn draws more users and editors. As evidence of this positive-feedback effect, the example ITK stub we created to outline the format of the gene stub is now the top-ranked hit in Google when searching for "ITK gene". Moreover, sixty-five percent of gene pages created by this effort are highly ranked in search engines, as described in the Preliminary Data section. (See also the Letter of Support from Tim Vickers, Washington University, on behalf of the Molecular Cellular Biology effort at Wikipedia.)

To summarize, successful completion of this Specific Aim offers two significant benefits. First, BioGPS opens the gene annotation process to The Long Tail of the biological community by enabling two-way information transfer between the portal and its users. Second, the use of a wiki system expands the scope of existing gene portals from structured data to unstructured knowledge.

Issues and Obstacles. The most commonly cited concern about anonymous, community-contributed content is accuracy. To assess of this issue, Nature editors recently conducted a study comparing the accuracy of Wikipedia to the online Encyclopedia Britannica and showed that the error rates were comparable between the two sources (an average of four errors per Wikipedia article compared to three errors per Britannica article) [47]. These results suggest that although errors are undoubtedly introduced, the community of fellow contributors and copy editors are reasonably proficient at ensuring that these errors are short-lived. These efforts to root out errors are aided by the maintenance of an exact revision history of every article and a discussion area where interested parties can reach consensus. Nevertheless, concerns over Wikipedia's accuracy have prompted a parallel wiki encyclopedia effort called Citizendium [48], and the idea of starting a parallel gene wiki effort at Citizendium is also a possibility that has been discussed with Citizendium organizers. Finally, we believe that the vast majority of users will understand that content in any openly-editable wiki must be treated differently than content from a standard gene portal. An informal sampling of colleagues reveals that people generally use Wikipedia as a generally good starting point and an overview, but the content of which must ultimately be verified. We believe that this is the appropriate mindset when using Wikipedia, and we also believe that most users will be sophisticated enough to understand this nuance relative to gene portals.

It is also worth noting that other biology wikis currently exist, and it bears consideration how the BioGPS/Wikipedia proposal fits into that landscape. The appeal of creating a biology-themed wiki is not particularly new or novel (biowiki.org puts its creation date in 2000), and there are many other individuals and groups that have started their own since then (partial list in **Table 1**). The idea of a gene wiki has also been discussed in commentaries in biomedical journals [49-51]. Even in light of these other biology wikis, there are two primary reasons why this gene wiki effort is worth pursing. First, many of the other biology wikis have a different focus. Several of these wiki sites are primarily used as web pages for research labs. Some are primarily focused on assembling biological and bioinformatics protocols. Still others focus on non-mammalian gene annotation. Therefore the *relatively* unique focus on mammalian gene annotation is a distinguishing factor from other biology wikis. Second, and perhaps more importantly, the BioGPS/Wikipedia effort has a critical mass of users. While the positive-feedback characteristics of collective intelligence efforts have been emphasized previously (more users leads to more edits, producing more content, and attracting more users), wiki efforts which rely on collective intelligence *but lack critical mass* are heavily anchored by inertia. If no users are drawn to a site, then there is no community of editors, no new content, and hence no new users. The BioGPS/Wikipedia effort combines two substantial user communities. SymAtlas/BioGPS users access our gene portal and view gene expression data, and they bring biology domain expertise to the gene wiki. Wikipedia of course has a huge user base of users, from which it draws a large number of editors with expertise in copyediting, wiki governance, and fighting vandalism. As demonstrated by the ITK example gene stub, the synergy between these two communities will create the necessary momentum to start the positive feedback loop.

Table 1. Partial list of currently available biology-themed wiki sites.

Name	URL
BioDirectory	biodirectory.com
BioWiki	biowiki.org
BioWiki	biowiki.net
EcoliWiki	colimod.org
Genomewiki	genomewiki.org
Open Wet Ware	openwetware.org
Wikiomics	wikiomics.org
WikiProteins	wikiprofessional.info

Finally, it is worth acknowledging that many other valid criticisms exist of wikis in general and of Wikipedia in particular. Nevertheless, despite these criticisms, it is difficult to dispute the success of Wikipedia in achieving its goal. This success is likely attributable to a combination of a detailed revision history, easy editing, light governance, and enthusiasm for a new-found model of collaborative writing.

A) Use structured annotation data to seed gene stubs in Wikipedia. As alluded to above, a successful wiki needs to be seeded with some useful content, which will draw users and begin the positive feedback loop of users and contributors. The BioGPS/Wikipedia effort began with creating gene stubs based on structured annotation from BioGPS. The format of the gene stubs was designed using the gene *ITK* as an example (**Figure 4**). Gene stubs include links to primary sequence databases (NCBI Entrez Gene, Ensembl, Refseq, Uniprot) and primary annotation sources (PDB, Homologene, OMIM, MGI). Thumbnail PDB images showing protein structure are also included when available. In addition, a thumbnail image from our Gene Atlas expression data set is included with a link to SymAtlas (and later BioGPS). All structured BioGPS data appears in a table layout which appears in the right margin. All subsequent free-text edits are designed to go in the central area of the gene page. This area of the gene report is seeded with additional data from NCBI when available, specifically the Gene Summary and links to relevant review articles.

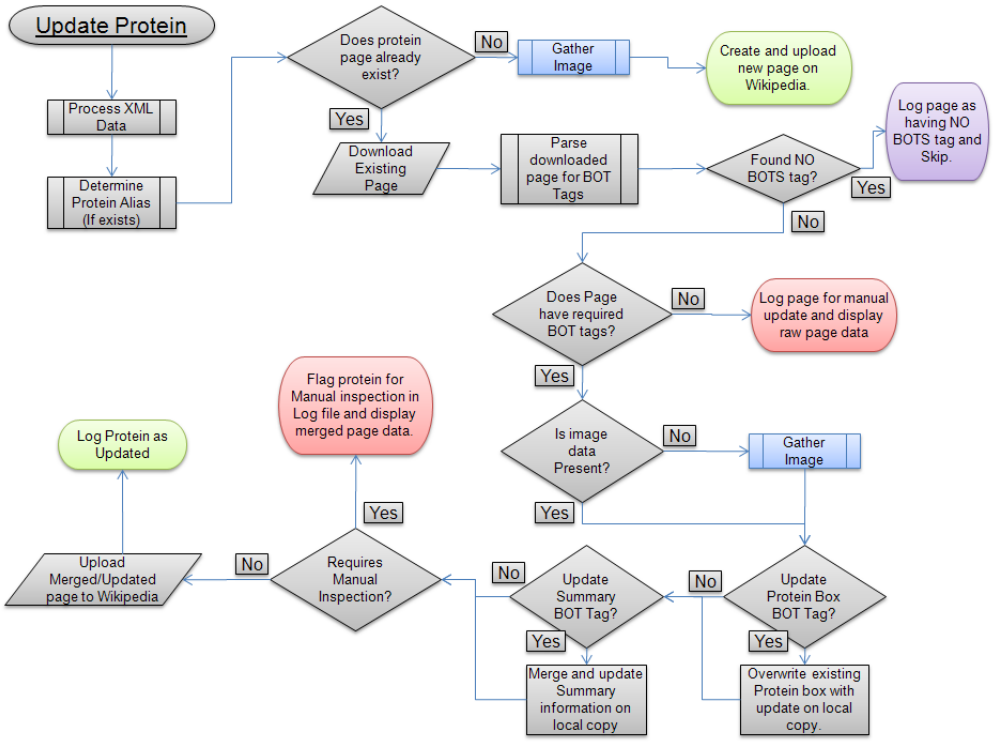

Figure 8. Flowchart representation of ProteinBoxBot program flow.

Having defined the format of the gene stubs, a program called the "ProteinBoxBot" was created to follow the sequence of steps shown in **Figure 8**. Briefly, gene annotation data is downloaded from BioGPS via the gene report Web service described above. The gene report XML file is parsed for the appropriate fields found in the gene stub design. ProteinBoxBot then attempts to determine if a gene page already exists for the gene being processed. If not, then a gene page is created using the official HUGO gene symbol. If a potential conflict is detected, the ProteinBoxBot logs the formatted wiki code to a log file for manual inspection and integration. The Java Wiki Bot Framework (JWBF) will be used to handle all communication with the Wikipedia server.

Comments will be added to the wiki code to clearly delineate sections of the gene stub that will be automatically updated and maintained by the ProteinBoxBot (which is currently planned to run quarterly). Instructions will also be posted in the wiki code comments for how to disable automatic updates by the ProteinBoxBot. The full specs for the ProteinBoxBot are maintained on the bot's Wikipedia user page [28]. The ProteinBoxBot will be created in collaboration with the Bioinformatics program at San Diego State University (see Letter of Support from Faramarz Valafar, SDSU).

B) Wrap Wikipedia gene content in BioGPS gene reports. As a free and open online encyclopedia, Wikipedia content is free to be reproduced according to the GNU Public License (GPL). The unstructured knowledge summarized in the Wikipedia page is relevant to BioGPS users as a source of unstructured and community-summarized knowledge. Therefore, we will directly incorporate Wikipedia content in our BioGPS gene reports. The incorporation of Wikipedia gene pages will be accomplished using the plugin interface described in

Specific Aim 3A. Briefly, external URLs will be rendered in each gene report based on a template URL and variable substitution. However, since gene pages are not always found at systematic locations in Wikipedia (at the gene symbol or gene description, for example), we will utilize the protein directory created by the ProteinBoxBot which relates a gene ID to a specific Wikipedia page title. The Wikipedia plugin will then return an HTML redirect to the appropriate gene page. For example, for the gene ITK, the Wikipedia plugin may be queried by BioGPS via a URL based on Gene ID (http://pluginserver/wp.cgi?gene=3702) and return the correct redirection (http://en.wikipedia.org/wiki/ITK_(gene)).

Alternatively, we may also consider a plugin which actively extracts and parses content from the relevant Wikipedia page. Since BioGPS data was used to seed the Wikipedia gene page, the "infobox" containing BioGPS content could be stripped from the BioGPS display. Extraction of content would be done using the XML export interface provided by Wikipedia (e.g., http://en.wikipedia.org/wiki/Special:Export/ITK_(gene)). After removing the infobox, the remaining text contained in the XML file would be translated to HTML using parsers that are widely available (for example, [52]), and the plugin will then simply display formatted HTML for contributed content. BioGPS would also display a link to edit the free-text content section which redirects the user to the Wikipedia editing page.

Specific Aim #3: Incorporate community-generated plugins by creating simple programmatic interfaces for external developers to extend BioGPS functionality. This plugin architecture will enable the BioGPS platform to be extended with new functionality by third-party bioinformaticians and data providers.

Significance and Rationale. While the two most popular gene portals (Entrez Gene and Ensembl) provide plugin interfaces to customize gene annotation reports, participation in developing portal plugins has largely been limited to The Short Head. The limited utilization by the broader community could possibly be due to the high level of technical knowledge required to develop plugins, the constraints on the types of data that can be displayed using plugins, or some combination of these two factors. In this Specific Aim, we intend to tailor the BioGPS plugin interface toward The Long Tail of plugin developers. By utilizing a simple plugin interface that is based only on HTML and CGI, BioGPS will both increase the flexibility of plugin content display and lower the technical barrier to plugin creation.

In the context of plugin development, we believe that The Long Tail is primarily comprised of labs that collectively generate huge amounts of data but that generally do not have a great deal of bioinformatic sophistication. There is ample evidence that The Long Tail is an untapped resource for gene annotation and data. Many diverse techniques for genome-wide profiling are currently being developed which are often not amenable to the traditional "tag" / "value" annotation format used in most gene portals. Some of these large-scale data sets are accessible online or through custom-built data display sites (e.g., [53, 54]), and many are only accessible through downloadable primary data files (e.g., [6, 7]). The BioGPS plugin interface will grant the developer complete control over plugin display, including heterogeneous and multimedia data types. Although complete enumeration of appropriate data types would be impractical, even very recent scans of the literature identify many diverse possibilities for genome-wide annotation that could be linked to BioGPS. These gene annotation sources include histone acetylation [55], prediction of microRNA targets [56], chromatin immunoprecipitation studies [57, 58], in situ hybridization [59], gene function prediction [60], and cell morphology [61]. In addition to the simple numeric data handled in **Specific Aim 1**, these examples demonstrate that a wide range of gene annotation from The Long Tail cannot easily be incorporated into centralized gene portals, and targeting sources like these is the primary goal of this Specific Aim. (For example, see Letter of Support from Benjamin Cravatt, Scripps Research Institute.)

In addition to the flexibility of display, the BioGPS plugin interface emphasizes ease of use for plugin developers. As noted in the Background and Significance section, the plugin interfaces to Entrez Gene and Ensembl are quite complex, and we suggest that the limited utilization of these interfaces by the broader community is, in part, due to that complexity. The BioGPS plugin interface targets The Long Tail of potential developers who have relatively less bioinformatics experience. Due to the nature of The Long Tail, even slight reductions in the sophistication required has the potential to drastically increase the number of scientists who actively participate in the gene annotation process. Since HTML and CGI are among the most common programming skills learned, we expect that this simplicity will result in the availability of a broad range of BioGPS plugins and an expansion of the community who contribute to gene portals.

The BioGPS plugin interface will also allow for custom search queries. Although BioGPS will natively handle searching by a wide variety of identifiers and keywords, more complex queries are often not amenable to the simple search interfaces that BioGPS will provide. To extend two examples introduced above, data for histone acetylation [55] and prediction of microRNA targets [56] are also amenable to searching, but not through the standard search interface that BioGPS uses for text-based queries. Therefore, the BioGPS plugin interface will also support custom search pages. The common feature of all search functions is that they return a gene list. Therefore, the BioGPS plugin interface will permit external developers to register any HTML-based search plugin which returns a gene list to BioGPS. Although this mechanism is slightly more complex to implement, this custom search plugin mechanism will match the flexibility described above for the gene report plugin display. Applications which utilize this search plugin interface could include DNA motif searches, gene function predictions, and non-textual searches.

In summary, successful completion of this Specific Aim will create a mechanism for The Long Tail of scientists to easily contribute heterogeneous data sources and novel search functions to a common BioGPS platform, thereby advancing our collective view of gene annotation.

Issues and Obstacles. It is first and foremost important to not over-interpret the goals of this Specific Aim. Most notably, there are many advantages of the existing plugin interfaces for Entrez Gene and Ensembl that are not replicated here. Specifically, since all transferred data in those protocols is typed, those interfaces offer the ability to create clients which intelligently handle plugin content. For example, various genome browsers accessing a single DAS server can choose to display content differently. In contrast, BioGPS delegates all presentation decisions to the plugin developer. Similarly, targeting very sophisticated bioinformatics developers enables greater inter-plugin communication, a capability which is not easily addressed using the simple interface proposed here. In summary, this Specific Aim seeks to be complementary to, and not in competition with, other gene portals and gene portal plugins.

Second, there are many liabilities which plague plugin interfaces in general which are not uniquely addressed by BioGPS. Most notably, because plugins are maintained by third parties, it is likely that some external servers underlying BioGPS plugins will eventually grow obsolete and/or fall into disrepair. However, this issue is equally important to other plugin interfaces as well as stand-alone web sites. (For example, at the time of this proposal's submission, 76 human and mouse DAS servers were listed at dasregistry.org, of which 23 were listed as inaccessible for more than two days.) Importantly, BioGPS loads plugins asynchronously using AJAX technology, so outages of a third-party plugin does not affect the loading of the rest of BioGPS gene report (including other plugins).

Third, we again address the question of how we expect this proposal to attract a critical mass of plugin developers. In short, we believe that the design choice to make the plugin interface as simple as possible will be the strongest attraction for plugin developers. We note that many bioinformaticians commonly set up web sites to access gene-centric data, most often using an HTML/CGI interface. Given that these sites already exist, the final step of registering the third-party site as a BioGPS plugin is trivially simple. Moreover, wrapping their site in a BioGPS plugin instantly offers many advantages for searching across synonyms, and any additional users they receive from common BioGPS users can only increase the visibility of their application. Finally, the plugin interface is so simple to set up and so commonly used that plugins can often be added by interested users themselves, independent of even the developers of the third-party site.

Fourth, and perhaps most importantly, the embedding of an external plugin within a parent web site raises potential browser security issues. Improper interactions between code on different websites or in different frames are commonly referred to as cross-site scripting (XSS) or cross-frame scripting (XFS). Attacks based on XSS or XFS typically involve the case where one site seeks to secretly intercept communication between the web user and another site, or when one site seeks to maliciously interact with another site on behalf of the user. Although these types of security threats are serious, we believe that the design of BioGPS will mitigate these concerns on several levels. First, all modern browsers are designed to eliminated XSS and XFS, and the latest releases usually plug all known bugs for serious threats. Second, most known threats are of the type where the "parent frame" attempts to spy on the "child frame". In this case, BioGPS is the parent, and therefore plugin developers can choose to include as little or as much data according to their level of trust in

the BioGPS developers. (A note on these liabilities will be placed in our plugin developers document.) Third, with the aid of Tony Doan (GNF's network security administrator), we will employ web application security testing software (AppScan, Watchfire/IBM) to periodically scan BioGPS to identify and close any exploitable holes by malicious users or plugin developers. In summary, we are confident that the BioGPS server will remain reasonably secure from attack, and that BioGPS users will be no more vulnerable to attack by visiting third-party sites via our plugin interface than directly through their web browser.

A) Provide plugin interface to add custom gene report content. BioGPS will allow external developers to develop plugins which add gene-specific content in the main annotation area of the gene report. Given the emphasis on simplicity and flexibility, this interface will be implemented using REST-base Web services and simple rules for URL variable substitution. For example, plugin developers will register a URL template in BioGPS that will follow a pattern like this: http://myPluginServer/my_data.cgi?id={{EntrezGeneID}}. Users will have the option of including any combination of plugins in their gene reports, as detailed in **Specific Aim 4B**. Each time a user requests a gene report, BioGPS will follow the four-step outline shown in **Figure 9**. First, the client user interface will request from the BioGPS Services Interface the list of all plugins stored in the user's profile. Second, the server will then return the list of URL templates for each plugin. Third, the client UI will perform a simple variable substitution based on the requested gene to create a gene-specific URL. For example, if the user has requested the gene report for ITK (Entrez Gene ID = 3702), then the URL template above would be translated to http://myPluginServer/my_data.cgi?id=3702. Fourth, the client sends individual HTTP requests for all plugins in the plugin list, and content is rendered asynchronously in the gene report as each plugin server responds. The plugin "containers" will be described in more detail in **Specific Aim 4A**.

In the example above, we describe a plugin whose URL template is rendered into a gene-specific URL based on substitution of the Entrez Gene ID. Variable substitution will be implemented for a variety of other fields indexed by BioGPS, including official gene symbol and identifiers from Ensembl, Refseq, PDB, MGI, Affymetrix, Uniprot, etc. In practice, any identifier which has been stored in the BioGPS database will be available to use as a variable placeholder in the URL template. Moreover, given the flexible database schema described in the Preliminary Data section, new identifiers can easily be added to BioGPS. In the case where a many-to-one relationship exists between a substitution variable and a gene (for example, many Refseq mRNA entries can link to one gene), all the identifiers will simply be passed to the plugin using a comma-separated format, leaving it to the plugin developer to properly display the appropriate content.

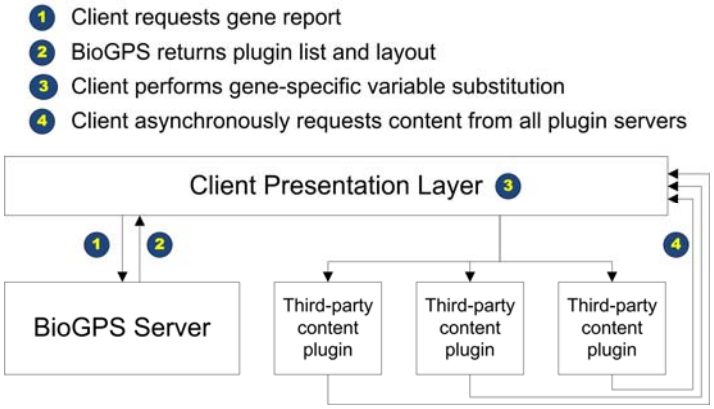

Figure 9. Outline of BioGPS plugin interface for third-party content.

From the perspective of a plugin developer, we expect that the architecture above will be easily utilized by The Long Tail of bioinformatics scientists. Developers only need to relate their plugin data to identifiers from one of the BioGPS-indexed data sources, and then create a CGI server which provides HTML content in a gene-specific manner. Plugin servers can easily be tested using simply formed URLs. Since HTML and CGI are among the most basic of programming skills, we expect to greatly increase the population of scientists who are able to utilize this plugin interface relative to existing gene portals. We expect plugin developers to create plugins which connect to well-established existing resources (e.g., Pubmed or Google search, Entrez Gene and Ensembl, PDB, HUGO), as well as to custom-built websites hosted by the developers themselves.

B) Provide plugin interface to add custom search interfaces. As mentioned above, the BioGPS plugin interface will also allow external developers to integrate custom search functions. Search plugins need to only satisfy two criteria. First, a search query needs to accept input via an HTML-based form. Second, the search results need to be represented as a gene list using identifiers recognized by the BioGPS system. The search plugin mechanism will also utilize a very simple HTML- and CGI-based mechanism, outlined in **Figure 10**. First, when a client activates an external search plugin, the user interface requests the input form from the BioGPS server. Second, the BioGPS requests and retrieves the appropriate HTML form from the external plugin

server. The HTML received from the external server should contain the necessary CGI form elements to transmit the query back to some external server. Importantly, this code will be stripped of all embedded scripts to eliminate the cross-site scripting vulnerability described above. Third, the HTML form is transmitted back to the client for rendering in the browser. Fourth, the client completes the form and submits the search parameters back to the BioGPS server. Fifth, BioGPS sends the query to the external search server on the client's behalf, and sixth, the plugin provider returns a gene list corresponding to the search results. Finally, those search results are displayed to the client as a BioGPS gene list, exactly as if a native BioGPS keyword or identifier query were run.

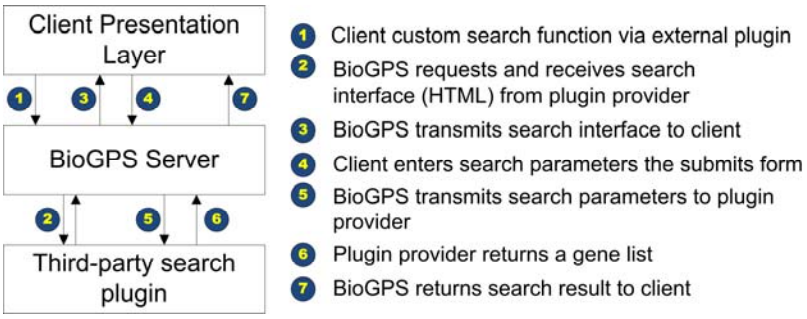

Figure 10. Outline of BioGPS plugin interface for third-party search.

From the perspective of the search plugin developer, utilization of this interface will again be as simple as programming HTML and CGI. In addition to programming the details of the search algorithm, the developer will need only to create an appropriate web form to collect input. The plugin server would then return a gene list in some BioGPS-defined CSV or XML format. Satisfying these criteria, the developer would be able to register the plugin server at BioGPS and test the plugin service, all independently of BioGPS administrators. The developer then has the option to publish this plugin to the search plugin library, as described in **Specific Aim 4B**.

Specific Aim #4: Enable users to share and customize the usage and layout of BioGPS plugins through optional user accounts. We envision an open marketplace where anyone can contribute new content for community use, and each user can customize how that community content can best serve their needs.

Significance and Rationale. The previous three Specific Aims all describe BioGPS features that engage The Long Tail of scientists. **Specific Aim 1** enables users to upload their own numeric data sets. **Specific Aim 2** enables users to contribute their own gene annotation. And **Specific Aim 3** enables developers to incorporate new functionality via plugins. These aims will enable the research community to independently incorporate custom content into BioGPS to extend their own experience beyond the base functionality provided. In this Specific Aim, we will implement custom user accounts in BioGPS which extend the ability to *customize* and provide the ability to *collaborate*.

As users individually extend their own user accounts to better conduct their own research, we expect that users will be motivated to share their customizations with the broader BioGPS community. For example, numeric data which is uploaded to BioGPS could be published to the "data library" simultaneously with the publication of a manuscript, allowing readers to easily search and analyze these data. Similarly, new visualizations of genome-wide annotation are likely to be of interest not just to data generators, but also other scientists as well. The ability for all users to collaboratively share content in public data and plugin libraries has the potential to greatly increase the utility of BioGPS.

Importantly, as the user community which contributes data and plugins grows, these new components cannot be universally applied to the gene reports for all users. Not all custom data sources require the same amount of screen real estate, making a single-sized container impractical. Furthermore, not every user will equally weight the importance of each data source, so allowing the content owner to specify size and layout for all users is not practical. Any one-size-fits-all policy would quickly make the gene report difficult to navigate and also discourage contributions from very specialized providers (the extreme tail of The Long Tail). Therefore, BioGPS will offer users the ability to individually customize the selection and layout of content from the data and plugin libraries. This feature is essential for end-user usability given the free and open BioGPS plugin model.

Therefore, successful completion of this Specific Aim will enable users to precisely tailor their own BioGPS gene report layout to suit their individual use cases.

Issues and Obstacles. Developing the BioGPS components which allow creation of user accounts and integration of access controls will require substantial time and attention to detail. However, technically these objectives are straightforward. We will integrate existing packages which have active development communities and proven track records for robustness, as described in the next section. The remaining potential challenges are non-technical in nature and relate to the user community's willingness to participate in the collaborative community we envision.

For example, it is possible that users may be reluctant to the public BioGPS data and plugin libraries for fear of losing credit for their contribution. Therefore, BioGPS will allow users to use customized URLs by which their dataset is accessed. For example, users can reference in their manuscript a URL like http://MyDataSet.biogps.org, where the prefix "MyDataSet" can be set by the contributor using the "saved layout" mechanism described in Sub-aim B. This option allows users to preserve "branding" of their data set in publications and in hyperlinks. The use of such a URL will indicate to BioGPS to automatically display the contributor's content in each gene report, without the need for users to navigate through preferences or data set libraries. We believe that this solution will effectively encourage users to contribute to the BioGPS community portal.

The challenges involved in the creation of a fully customizable user interface, on the other hand, are considerably more technical in nature. Specifically, BioGPS will utilize many leading-edge technologies in web design, including advanced JavaScript libraries for UI development and AJAX for client-server communication. Other UI tools such as Adobe Flex will also be considered. The use of these newer technologies puts greater emphasis on the browser capabilities of users' web clients. To address this potential problem, we will engage in significant cross-browser and cross-platform testing. Importantly, examination of SymAtlas users confirms that the number of combinations can be reasonably limited. For example, 80% of web users use either Internet Explorer (version 6 or greater) or Firefox (version 2 or greater). An additional 10.4% use one of the two most recent versions of the Safari browser (which come standard on the Tiger and Leopard versions of Mac OS X). Regarding operating systems, 72% of users use Microsoft Windows (Windows 2000 or later), and 24% use a Macintosh OS. These statistics demonstrate that the common and recent user platforms described above account for greater than 90% of users, greatly simplifying our testing protocols.

Finally, it is worth noting that the BioGPS model relies on third-party developers to continually maintain their plugin content. Because BioGPS enables external developers to independently distribute new functionality via the plugin library, BioGPS also necessarily delegates the maintenance of those plugin servers to those same developers. Undoubtedly, this design feature of BioGPS will lead to nonfunctional plugins being found in the plugin library. However, the maintenance responsibilities for the external developers are no different than if they set up their own stand-alone web site. BioGPS provides an efficient marketplace of plugins where the utility and popularity of each plugin can be quickly evaluated by the broad user community. BioGPS usage information may even provide evidence to developers to convince themselves of the importance of further maintenance of their plugin. Moreover, we feel that the advantages of an open and distributed plugin model (most notably, the ability for The Long Tail to independently extend BioGPS) far outweigh this potential liability.

A) Create data set and plugin libraries from which users can easily browse and select community-contributed content. The first step in enabling contribution to and utilization of content libraries is developing the security infrastructure. All functionality for account registration, email confirmation, password retrieval, and authentication will be addressed using the .NET security providers and Active Directory Application Mode (ADAM) libraries which are freely available in the .NET framework. ADAM will securely maintain a database of all users and passwords, and the .NET security providers will communicate with ADAM using the Lightweight Directory Access Protocol (LDAP), a protocol that is widely used by many modern operating systems. ADAM will also maintain security groups and all user group memberships, and will feed the .NET security providers with enough information to authorize users for centrally-defined user roles. As an alternative security model, we may also consider using Open ID authentication (http://openID.net) as that effort to unify online identities matures.

Second, user profiles will be implemented using the .NET Profiles Provider pattern and will be implemented in two parts. One XML configuration file will specify the scope of the application's user profile information, and the other part will be implemented directly in the application's data store in an Oracle server. These profiles will allow users to customize the application by saving preferences and user-specific data. The Profiles Provider will maintain all these settings and will retrieve them once a user is logged on to the application.

Finally, with the authentication and profiles mechanism in place, we will create an open data and plugin library system to which all users are able to contribute. Each user will have their own user content manager which will allow all users to view private content contributed under their own account. Any data set or plugin in the user content manager can be "published" to the global content library by simply marking a flag. The owner of each data set and plugin will be noted in the global content library, and owners retain full control to edit or delete their contributed content.

All public data sets and plugins will be managed in the global content manager. In addition to their own content shown in their user content manager, each user will be able to include any content from the global content library in their personalized gene report. Since the selected content (composed of any combination of public and private content sources) will be stored in the user's profile, users will see their customized gene report regardless of which computer they log in from.

Since we expect the community to collectively contribute a wide selection of data sets and plugins, we will also enable convenient methods to search the global content library. Specifically, the owner of each data source will be able to tag their contribution with any free-text tags, and tags will be visualized in a "tag cloud" format. Additionally, public content will be ranked according to popularity based on how many users have included it in their gene report. We expect that the combination of tags and popularity-based measures will be an effective means for users to navigate the global content library.

B) Create plugin containers and a layout manager that gives users complete control over gene report display.
The user interface will be built using the Django web framework. Based on the python programming language, Django provides high-level libraries for web design. Although full web applications (from database access to the user interface) can be built using the Django framework, BioGPS will exclusively utilize the presentation layer of Django which provides for fast prototyping using concise syntax. Business logic will exclusively be performed in the .NET business layer, with which Django will communicate through Web services. In addition, the presentation layer will use the Ext JS JavaScript library. This library enables fast development of rich user interfaces for web applications.

Using the technologies above, BioGPS will feature a fully-customizable plugin layout. Each plugin that is selected by the user will be rendered in its own plugin container. This plugin container renders content as if in an HTML IFRAME, essentially a browser within a browser. This decoupling between BioGPS and the plugin providers ensures that plugin content does not corrupt any aspect of BioGPS presentation (or the presentation of other plugins). All plugin containers will be rendered asynchronously (as detailed in **Figure 9**) so that BioGPS and plugin content is displayed as it becomes available. Finally, each plugin container will be fully customizable in terms of location and size through a "drag-and-drop" interface (**Figure 11**). This flexibility allows users to tailor their plugin layout based on the importance with which they view each plugin.

Users will also be able to save their layout in their user profile, and all gene reports will be subsequently rendered using the same layout of plugins. Moreover, BioGPS will allow users to save multiple layouts to correspond to multiple use cases. For example, a user may have one layout for general gene overviews, one layout for concise viewing of gene expression thumbnails only, and one layout focused on viewing and/or editing Wikipedia content. Finally, in the same spirit as the data and plugin libraries described above, BioGPS will also provide a layout library for sharing saved layouts between users. This feature will also incorporate the tagging and popularity features described for the data and plugin libraries, and will allow new users to quickly survey available content. Public layouts will also be the mechanism by which we implement the "customized URL" feature described above. For example, developers may release a public layout titled "MyDataSet" (which presumably features their contributed content prominently), and all users who access BioGPS through http://MyDataSet.biogps.org will view using that public layout by default.

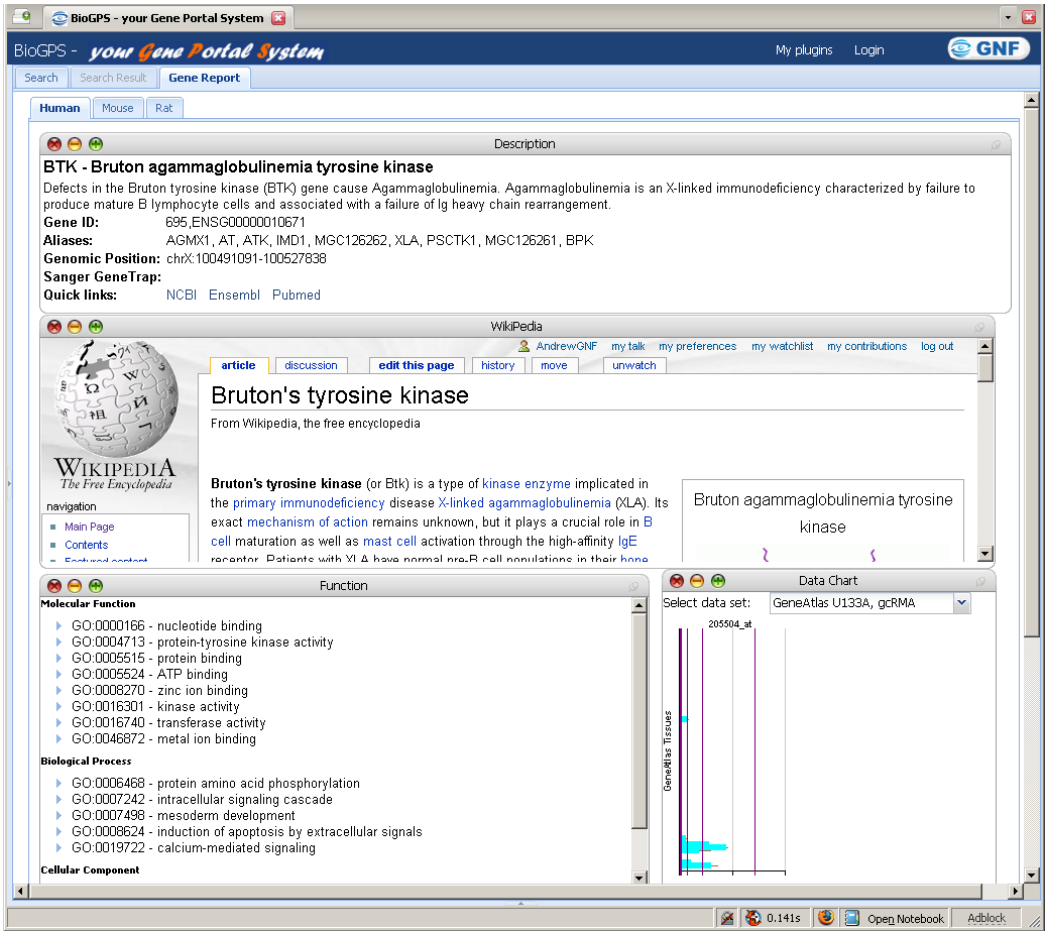

Figure 11. Screenshot of BioGPS prototype showing customizable user interface. Each window can be individually moved and resized, and all layout parameters can be stored in the users' profile. For aesthetic reasons, the frames around each plugin can also be hidden to present a more unified gene report.

Summary. BioGPS will serve the very basic functions of a gene portal. Specifically, BioGPS will allow users to search by any number of keywords and identifiers from public databases and technology platforms and return a list of gene entities. Beyond this capability, BioGPS differs substantially in design and function from existing gene portals. Gene portals heavily emphasize curation, quality, and oversight of all data that is presented. In contrast, BioGPS allows all users to independently and collaboratively aggregate data sets, gene annotation, and plugins. This proposal is precisely aimed at harnessing The Long Tail of scientists, to **synergistically leverage community knowledge and effort,** and to ultimately allow researchers across biological disciplines to **efficiently annotate the human genome**. We believe that these two models will be complementary, and that all biologists will benefit from the addition of BioGPS to the landscape of online gene resources.

Timeline. Although the most novel aspects of the BioGPS gene portal have been highlighted in the Specific Aims, we have requested five years of support since basic portal functionality also needs to be developed and maintained. This initial portal development will incorporate the design pattern described in the introduction to the Research Design and Methods section and is essential to the long-term maintenance and extensibility of BioGPS.

Aim	Year:	1	2	3	4	5
1	1A) Data import tool	■	■			
	1B) GEO browser				■	■
2	2A) Create gene wiki stubs	■	■	■		
	2B) Integrate gene wiki in BioGPS		■	■		
3	3A) Plugin interface for annotation		■	■	■	■
	3B) Plugin interface for searching			■	■	■
4	4A) Create community-content libraries		■	■	■	■
	4B) Create plugin containers	■	■	■		
Basic portal development		■	■			
Maintenance and support		■	■	■	■	■

BIBLIOGRAPHY AND REFERENCES

1. **The Long Tail, in a nutshell** [http://www.longtail.com/about.html]
2. Albert, R: **Scale-free networks in cell biology**. *J Cell Sci* 2005, **118**(Pt 21):4947-4957.
3. Brynjolfsson, E, Hu, YJ, Smith, MD: **Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers**. *Management Science* 2003, **49**(11):1580-1596.
4. Su, AI, Cooke, MP, Ching, KA, Hakak, Y, Walker, JR, Wiltshire, T, Orth, AP, Vega, RG, Sapinoso, LM, Moqrich, A *et al*: **Large-scale analysis of the human and mouse transcriptomes**. *Proc Natl Acad Sci U S A* 2002, **99**(7):4465-4470.
5. Su, AI, Wiltshire, T, Batalov, S, Lapp, H, Ching, KA, Block, D, Zhang, J, Soden, R, Hayakawa, M, Kreiman, G *et al*: **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proc Natl Acad Sci U S A* 2004, **101**(16):6062-6067.
6. Barrera, LO, Li, Z, Smith, AD, Arden, KC, Cavenee, WK, Zhang, MQ, Green, RD, Ren, B: **Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs**. *Genome Res* 2008, **18**(1):46-59.
7. **Cancer Program Data Sets** [http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi]
8. **SOURCE Search** [http://source.stanford.edu.]
9. **The Gene Expression Database (GXD)** [http://www.informatics.jax.org/menus/expression_menu.shtml]
10. **SymAtlas** [http://symatlas.gnf.org/SymAtlas]
11. Ortega, JF, Gonzalez-Barahona, JG: **Quantitative Analysis of the Wikipedia Community of Users.** In: *International Symposium on Wikis: 2007*; 2007.
12. Giles, J: **Internet encyclopaedias go head to head**. *Nature* 2005, **438**(7070):900-901.
13. **Web Services Architecture** [http://www.w3.org/TR/ws-arch/]
14. **LinkOut Home Page** [http://www.ncbi.nlm.nih.gov/projects/linkout/jnlist/loprovlink.html]
15. **Information for Other Resource Providers** [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helplinkout.chapter.nonbib]
16. Dowell, RD, Jokerst, RM, Day, A, Eddy, SR, Stein, L: **The distributed annotation system**. *BMC Bioinformatics* 2001, **2**:7.
17. Prlic, A, Down, TA, Kulesha, E, Finn, RD, Kahari, A, Hubbard, TJ: **Integrating sequence and structural biology with DAS**. *BMC Bioinformatics* 2007, **8**:333.
18. **What can I do with DAS?** [http://www.biodas.org/wiki/Main_Page#What_can_I_do_with_DAS_.3F]
19. **DAS2 Protocol** [http://biodas.org/documents/das2/das2_protocol.html]
20. **Available DAS sources** [http://www.dasregistry.org/listServices.jsp]
21. **BioDAS:Community Portal** [http://www.biodas.org/wiki/BioDAS:Community_Portal]
22. **Das_registry_announce Info Page** [http://lists.sanger.ac.uk/mailman/listinfo/das_registry_announce]
23. Su, AI, Hogenesch, JB: **Power-law-like distributions in biomedical publications and research funding**. *Genome Biol* 2007, **8**(4):404.
24. **GNF Gene Expression Atlas** [http://expression.gnf.org]
25. Panda, S, Antoch, MP, Miller, BH, Su, AI, Schook, AB, Straume, M, Schultz, PG, Kay, SA, Takahashi, JS, Hogenesch, JB: **Coordinated transcription of key pathways in the mouse by the circadian clock**. *Cell* 2002, **109**(3):307-320.
26. Walker, JR, Su, AI, Self, DW, Hogenesch, JB, Lapp, H, Maier, R, Hoyer, D, Bilbe, G: **Applications of a rat multiple tissue gene expression data set**. *Genome Res* 2004, **14**(4):742-749.
27. **ITK (gene) - Wikipedia, the free encycopedia** [http://en.wikipedia.org/wiki/ITK_(gene)]
28. **User:ProteinBoxBot - Wikipedia, the free encyclopedia** [http://en.wikipedia.org/wiki/User:ProteinBoxBot]
29. **List of gene pages at Wikipedia created or amended by ProteinBoxBot** [http://tinyurl.com/29u9l4]
30. **NCBI FTP site** [ftp://ftp.ncbi.nih.gov/gene/DATA]
31. **Ensembl FTP site** [ftp://ftp.ensembl.org/pub/current_mart/data/mysql/ensembl_mart_44/]
32. Labarga, A, Valentin, F, Anderson, M, Lopez, R: **Web services at the European bioinformatics institute**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W6-11.
33. **Web API for Bioinformatics** [http://xml.nig.ac.jp/index.html]
34. **Entrez Programming Utilities** [http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html]
35. **What Is Windows Communication Foundation?** [http://msdn2.microsoft.com/en-us/library/ms731082.aspx]
36. Mangalam, H: **The Bio* toolkits--a brief overview**. *Brief Bioinform* 2002, **3**(3):296-302.

37. Stajich, JE, Block, D, Boulez, K, Brenner, SE, Chervitz, SA, Dagdigian, C, Fuellen, G, Gilbert, JG, Korf, I, Lapp, H *et al*: **The Bioperl toolkit: Perl modules for the life sciences**. *Genome Res* 2002, **12**(10):1611-1618.
38. Carpenter, AE, Sabatini, DM: **Systematic genome-wide screens of gene function**. *Nat Rev Genet* 2004, **5**(1):11-22.
39. Orth, AP, Batalov, S, Perrone, M, Chanda, SK: **The promise of genomics to identify novel therapeutic targets**. *Expert Opin Ther Targets* 2004, **8**(6):587-596.
40. Chanda, SK, White, S, Orth, AP, Reisdorph, R, Miraglia, L, Thomas, RS, DeJesus, P, Mason, DE, Huang, Q, Vega, R *et al*: **Genome-scale functional profiling of the mammalian AP-1 signaling pathway**. *Proc Natl Acad Sci U S A* 2003, **100**(21):12153-12158.
41. Root, DE, Hacohen, N, Hahn, WC, Lander, ES, Sabatini, DM: **Genome-scale loss-of-function screening with a lentiviral RNAi library**. *Nat Methods* 2006, **3**(9):715-719.
42. Silva, JM, Mizuno, H, Brady, A, Lucito, R, Hannon, GJ: **RNA interference microarrays: high-throughput loss-of-function genetics in mammalian cells**. *Proc Natl Acad Sci U S A* 2004, **101**(17):6548-6552.
43. Barrett, T, Suzek, TO, Troup, DB, Wilhite, SE, Ngau, WC, Ledoux, P, Rudnev, D, Lash, AE, Fujibuchi, W, Edgar, R: **NCBI GEO: mining millions of expression profiles--database and tools**. *Nucleic Acids Res* 2005, **33**(Database issue):D562-566.
44. Barrett, T, Troup, DB, Wilhite, SE, Ledoux, P, Rudnev, D, Evangelista, C, Kim, IF, Soboleva, A, Tomashevsky, M, Edgar, R: **NCBI GEO: mining tens of millions of expression profiles--database and tools update**. *Nucleic Acids Res* 2007, **35**(Database issue):D760-765.
45. Parkinson, H, Kapushesky, M, Shojatalab, M, Abeygunawardena, N, Coulson, R, Farne, A, Holloway, E, Kolesnykov, N, Lilja, P, Lukk, M *et al*: **ArrayExpress--a public database of microarray experiments and gene expression profiles**. *Nucleic Acids Res* 2007, **35**(Database issue):D747-750.
46. Parkinson, H, Sarkans, U, Shojatalab, M, Abeygunawardena, N, Contrino, S, Coulson, R, Farne, A, Lara, GG, Holloway, E, Kapushesky, M *et al*: **ArrayExpress--a public repository for microarray gene expression data at the EBI**. *Nucleic Acids Res* 2005, **33**(Database issue):D553-555.
47. Giles, J: **Wikipedia rival calls in the experts**. *Nature* 2006, **443**(7111):493.
48. **Citizendium: The Citizens' Compendium** [http://www.citizendium.org]
49. Salzberg, SL: **Genome re-annotation: a wiki solution?** *Genome Biol* 2007, **8**(1):102.
50. Wang, K: **Gene-function wiki would let biologists pool worldwide resources**. *Nature* 2006, **439**(7076):534.
51. Yager, K: **Wiki ware could harness the Internet for science**. *Nature* 2006, **440**(7082):278.
52. **A Wiki parser -- Flex 2.01** [http://livedocs.adobe.com/flex/201/html/12_Using_Regular_Expressions_168_12.html]
53. **Genomics, evolution and function of protein kinases** [http://kinase.com]
54. **Welcome to the CREB Target Gene Database** [http://natural.salk.edu/CREB/]
55. Sadikovic, B, Andrews, J, Carter, D, Robinson, J, Rodenhiser, DI: **Genome-wide H3K9 Histone Acetylation Profiles Are Altered in Benzopyrene-treated MCF7 Breast Cancer Cells**. *J Biol Chem* 2008, **283**(7):4051-4060.
56. Gaidatzis, D, van Nimwegen, E, Hausser, J, Zavolan, M: **Inference of miRNA targets using evolutionary conservation and pathway analysis**. *BMC Bioinformatics* 2007, **8**:69.
57. Hatzis, P, van der Flier, LG, van Driel, MA, Guryev, V, Nielsen, F, Denissov, S, Nijman, IJ, Koster, J, Santo, EE, Welboren, W *et al*: **Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells**. *Mol Cell Biol* 2008.
58. O'Geen, H, Squazzo, SL, Iyengar, S, Blahnik, K, Rinn, JL, Chang, HY, Green, R, Farnham, PJ: **Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs**. *PLoS Genet* 2007, **3**(6):e89.
59. Lein, ES, Hawrylycz, MJ, Ao, N, Ayres, M, Bensinger, A, Bernard, A, Boe, AF, Boguski, MS, Brockway, KS, Byrnes, EJ *et al*: **Genome-wide atlas of gene expression in the adult mouse brain**. *Nature* 2007, **445**(7124):168-176.
60. Chen, XW, Liu, M, Ward, R: **Protein Function Assignment through Mining Cross-Species Protein-Protein Interactions**. *PLoS ONE* 2008, **3**(2):e1562.
61. Mukherji, M, Bell, R, Supekova, L, Wang, Y, Orth, AP, Batalov, S, Miraglia, L, Huesken, D, Lange, J, Martin, C *et al*: **Genome-wide functional analysis of human cell-cycle regulators**. *Proc Natl Acad Sci U S A* 2006, **103**(40):14819-14824.