Title: MyGene.info updates: scalable gene-centric web services with user contributions

Authors: Chunlei Wu and Andrew I. Su (presenting author underlined)
Email: cwu@scripps.edu asu@scripps.edu
Affiliation: The Scripps Research Institute, 10550 N Torrey Pines Rd, La Jolla, CA 92037

Project web site: http://mygene.info
Source code: https://bitbucket.org/sulab/mygene.hub/src
https://bitbucket.org/sulab/mygene.info/src
Open Source License being used: **Apache License**
Considered for a talk, a poster, or both: **both**

Biological applications typically start with a query interface for users to search for their favorite genes. Building such an interface often requires developers to maintain a dedicated gene annotation database to translate user queries into the desired gene annotation objects. Setting up a database server and keeping it updated can be a time-consuming and cumbersome tasks. Since the majority of raw gene annotation data are coming from several large data centers like NCBI and Ensembl, developers are also duplicating their efforts to setup gene annotation databases from essentially the same data providers.

MyGene.info (http://mygene.info) is a cloud-based solution to abstract the task of building a gene annotation database into a set of scalable and extensible web services. End users have access to two simple-to-use REST web services for gene annotation query and retrieval, without worrying about designing, building and maintaining a dedicated database. The gene query service [1] takes the user query string and returns the matching gene objects with desired annotations; and the gene annotation service [2] returns annotation data for given gene IDs. Both services return JSON (Javascript Object Notation) formatted data, making them easy to integrate into applications.

Right before last year's BOSC, we released new v2 MyGene.info API [3]. Thanks to the scalable backend built upon MongoDB and Elasticsearch, we expanded our services to 16M genes from >13K species, while keeping the high query performance [4]. Since the release, MyGene.info has served over 60M requests (July 2013-March 2014). As of now, MyGene.info services steadily serve ~1.5M requests per month, and our Python client *mygene.py* module [5] also achieves ~600 downloads per month.

As a centralized resource hub, one important aspect of MyGene.info is the user contributions. Even though we, as core developers, are always adding more gene annotation data into our system, it's equally important to build a framework to allow our users to contribute data into MyGene.info. Users can now write a simple data importer script, based on the template we provide, to load their own data into MyGene.info, so that they will be accessible via our high-performance query engine. Given the flexibility of JSON data structure, new data under a new field name will not affect existing data, avoiding breaking existing applications.

Another way of user contributions is to write custom query filters. Our users often need to restrict their search scopes, e.g. for a given species, for ncRNA genes only, or simply a specific list of genes. We allow users to contribute their custom query filters (each with a unique name) at server-side, that way they can achieve both higher performance (server-side execution) and cleaner query syntax (with the filter name instead of the actual query). Moreover, other users can benefit from re-using those custom filters relevant to their use cases.

[1] http://mygene.info/doc/query_service.html
[2] http://mygene.info/doc/annotation_service.html
[3] http://sulab.org/2013/07/mygene-info-v2-api-goes-live/
[4] http://mygene.info/#what-s-new-in-v2-api
[5] https://pypi.python.org/pypi/mygene