- Making Elastic Gene API

Chunlei Wu, Ph.D.

The Scripps Research Institute
La Jolla, CA, USA

BOSC 2013
July 20, 2013
Being “Elastic”

- Fast
- Always ON
- Up-to-date
- Scalable
- Extensible
Common procedure for gene data retrieval

Data sources
- Entrez
- Ensembl
- UniProt
- ...

Data parsing

Internal gene object

update regularly
Common procedure for gene data retrieval

Data sources:
- Entrez
- Ensembl
- UniProt
- ...

Data parsing

As A Service?

Internal gene object

update regularly
Working model - 1

Data sources
- Entrez
- Ensembl
- UniProt

Merging

Query Engine

User queries
A dummy merging example:

```
{
  "entrezgene": 1017
}
{
  "ensemblgene": "ENSG00000123374"
}
{
  "uniprot": "P24941"
}
```

Merging

```
{
  "entrezgene": 1017,
  "ensemblgene": "ENSG00000123374",
  "uniprot": "P24941"
}
```
Gene object in noSQL database

```
token: '1017:
{
    "Symbol": "CDK2",
    "Ensembl": "ENSG00000123374",
    "RefSeq": [
        "NM_001798",
        "NM_052827"
    ],
    "Reporter": {
        "U95A": [
            "1792_g_at",
            "1833_at"
        ],
        "U133A": [
            "211804_s_at",
            "2045252_at",
            "211803_at"
        ]
    }
}
```
Syncing from data-hub to query instance

Data sources
- Entrez
- Ensembl
- UniProt

Merging

Public data hub

Public query instance

Query Engine

User queries
Syncing from data-hub to query instances

Data sources
- Entrez
- Ensembl
- UniProt

Merging

Public data hub

Public query instance

Query Engine

User queries

powered by MongoDB

elasticsearch
Public query instance

http://MyGene.info
(currently v2 API, two endpoints)

any query term(s) ➔ http://MyGene.info/v2/query?q=<query> ➔ matching gene hits

gene id(s) ➔ http://MyGene.info/v2/gene/<geneid> ➔ matching gene objects
Public query instance

http://MyGene.info

- Support ALL species, from NCBI (>12K species, >13M genes)
- >40 annotation fields and expanding
- Weekly-updated
- Flexible query interface
  - Simple queries
    - ?q=cdk2
  - Fielded queries
    - ?q=symbol:cdk2
  - Wildcard queries
    - ?q=cdk*
  - Genomic interval queries
    - ?q=chr1:1-100,000&species=human
  - Species filter
    - ?q=cdk2&species=mouse,rat
  - Returning fields filter
    - ?q=cdk2&fields=symbol,homologene
- Support batch queries, JSONP, CORS
- Committed for long-term availability
Public query instance

http://MyGene.info

High-performance host (serving ~500K requests/day)
Public query instance

Third-party packages

MyGene.py - Python wrapper

https://pypi.python.org/pypi/mygene

```
In [1]: import mygene
In [2]: mg = mygene.MyGeneInfo()
In [3]: mg.getgene(1017)
Out[3]:
{'id': '1017', 'entrezgene': 1017, 'name': 'cyclin-dependent kinase 2', ...
In [4]: mg.query('cdk2')
Out[4]:
{'hits': [{'id': '1017',
  'score': 373.24667,
  'entrezgene': 1017,
  'name': 'cyclin-dependent kinase 2',
  'symbol': 'CDK2',
  'taxid': 9606},
  ...
  'max_score': 373.24667,
  'took': 10,
  'total': 28}
```
Public query instance

Third-party packages

MyGene.autocomplete

- Gene query autocomplete widget

https://bitbucket.org/sulab/mygene.autocomplete

```javascript
$("input.mygene_query_target").genequery_autocomplete({
  "species": "human,mouse,rat",
  "limit": 10,
  "gene_label": "{symbol}: {name} ({species})",
  "select": function(event, ui) {
    alert( ui.item ?
      "Selected: " + ui.item.label + '('+ui.item._id+')':
    "Nothing selected, input was " + this.value);
  }
});
```
Public query instance

Third-party packages

MyGene.autocomplete
- Gene query autocomplete widget

https://bitbucket.org/sulab/mygene.autocomplete
Syncing from data-hub to query instances

Data sources:
- Entrez
- Ensembl
- UniProt
...

Merging

Public data hub

Public query instance
Query Engine

User queries

Private query instance
Query Engine

Private data hub

Private 1
Private 2
...

Merging

User queries
Private query instance

- Dedicated host
- Same powerful query interface
  - Third-party packages still work
- Public data still get sync-ed
- Allow to merge private data
To reach us?

Questions on public query instance
or
interested in setting up your own private query instance?

Please let us know:
help@mygene.info
Code repositories

- Web front-end
  https://bitbucket.org/sulab/mygene.info
  Apache 2 licensed

- Data hub
  https://bitbucket.org/sulab/mygene.hub
  GPL v3 licensed
Acknowledgement

**Sulab**

Andrew Su

Benjamin Good
Max Nannis
Salvatore Loguercio
Katie Fisch
Tobias Meissner

**Funding and Support**

R01GM083924