Chunlei Wu, Ph.D.

✉️ cwu@scripps.edu

🐦 @chunleiwu

Associate Professor of Molecular Medicine
Dept. of Molecular Experimental Medicine
The Scripps Research Institute
La Jolla, CA, USA

BD2K-AHM 11/2016
BioThings APIs

Objective:
Building unified APIs for “Bio-Things” (biological entities)
Biological knowledge is a complex network

No one-fit-all database can capture the entire knowledge space
Simplify the knowledge network as entities

Extracting those central hub nodes as flat lists:

- Gene
- Variant
- Pathway
- Disease
- Metabolite

...
Gene and Variant annotations represented in JSON documents

```json
{
  "_id": "chr1:g.196659237C>T",
  "cosmic": {
    "chrom": "1",
    "hg19": {
      "start": 196659237,
      "end": 196659237,
      "ref": "C",
      "alt": "T",
      "tumor_site": "breast",
      "mut_freq": 0.49,
      "mut_nt": "C>T",
      "cosmic_id": "COSM424915"
    }
  },
  "_id": "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"
    ]
  }
}
```
Keep data always up-to-date

Schematic view of MyVariant.info architecture

Python/Tornado

ElasticSearch

Indexing Engine

MongoDB

Loading Scripts

Variant retrieval API
/v1/variant/<id>

Variant query API
/v1/query

Each data source is updated individually. Colors indicate their different updating schedules.
High-performance web service APIs

Schematic view of MyVariant.info architecture
MyGene.info + MyVariant.info

Gene

Variant

/v3/gene/<geneid>
/v3/query?q=<query>

/v1/variant/<hgvsid>
/v1/query?q=<query>

single query on GET, batch query on POST
We focus on building APIs. Try to …
Make it really easy to use

Just two endpoints

No registration/sign-in

No API key
Developer-friendly

JSONP
CORS
https
msgpack
http compression
http caching
JSON-LD

Python/R clients
(also js client for myvariant)

search “mygene” and “myvariant” in PyPI and Bioconductor

Supported!
Aggregate everything about genes and variants

**MyGene.info**
- Support >17M genes for ~18K species
- ~200 annotation fields

**MyVariant.info**
- Support >340 M variants
- ~500 annotation fields
- from 14 sources:
  - ClinVar
  - dbNSFP
  - dbSNP
  ...
Keep up-to-date

MyGene.info

Support >17M genes for ~18K species

~ 200 annotation fields

MyVariant.info

Support >340 M variants

~ 500 annotation fields

from 14 sources:
  ClinVar
dbNSFP
dbSNP...

Weekly

Monthly
High-performance and scalable

>95% queries response < 30ms
High-performance and scalable

Apparently @mygeneinfo was working very hard during last XMas, 5M hits/day, so did our users :-)
High-performance and scalable

Over 100M requests in Nov 2016
High availability

MyGene.info

MyVariant.info

99.999%
over last year

99.935%
over last year

Availability tracked by

UptimeRobot
Who is using Live applications:

- **MyGene²**
  - Uniting families for sharing and discovery

- **Gene Wiki**
- **The Monarch Initiative**

- **BioGPS**

- **JBrowse**

- **MinePath.org**

- **CIViC**
  - Clinical Interpretations of Variants in Cancer

- **ENCODE**
Who is using

Many users use them in their
daily analysis pipelines

or

simply caching annotations locally
Generalized BioThings SDK

BioThings SDK

- JSON data aggregation mechanism
- Well-designed REST API pattern
- Data-updating scheduler
- Python/R clients
- High-performance query engine
- JSON-LD enabled Linked Data
BioThings SDK

A tutorial here (more docs are coming):
BioThings SDK

- g.biothings.io: gene (alias to MyGene.info)
- v.biothings.io: variant (alias to MyVariant.info)
- s.biothings.io: species/taxonomy
- c.biothings.io: drugs/compounds
- d.biothings.io: disease

...
JSON-LD brings the linkage between BioThings APIs
Apply JSON-LD context

**JSON document**

```
{"_id" : "chr6:g.26093141G>A",
"clinvar" :{
  "gene" :{
    "id" : "3077",
    "symbol" : "HFE"
  }
},
"dbsnp" :{
  "rsid" : "rs1800562"
},
"cadd" :{
  "genename" : "HFE"
}
}
```

**JSON-LD Context**

```
{"root":{
  "@context":{
    "clinvar": "http://schema.myvariant.info/datasource/clinvar",
    "dbsnp": "http://schema.myvariant.info/datasource/dbsnp",
    "genename": "http://identifiers.org/hgnc.symbol",
    "cadd": "http://schema.myvariant.info/datasource/cadd",
    "rsid": "http://identifiers.org/dbsnp/",
    "gene": "http://schema.myvariant.info/datanode/gene"},
  "clinvar/gene":{
    "@context":{
      "symbol": "http://identifiers.org/hgnc.symbol"
  }
}
```

**N-Quads Transformation**

```
_:b0 <http://schema.myvariant.info/datasource/clinvar> _:b1 .
_:b0 <http://schema.myvariant.info/datasource/dbsnp> _:b3 .
_:b0 <http://schema.myvariant.info/datasource/cadd> _:b4 .
_:b2 <http://identifiers.org/hgnc.symbol> "HFE".
_:b3 <http://identifiers.org/dbsnp/> "rs1800562".
_:b4 <http://identifiers.org/hgnc.symbol> "HFE".
```
BioThings TEAM

TSRI:

Chunlei Wu
Andrew Su
Jiwen Xin
Cyrus Afrasiabi
Sebastien Lelong
Ginger Tsueng
Julee Adesara
Mike Mayers

U. Washington:

Sean Mooney
Moritz Juchler
Nikhil Gopal
Sicheng Song

Funding and Support

U01HG008473
U54GM114833