EBI Search

Search your biological data at EMBL-EBI

Web Production

Nicola Buso

EMBL-EBI

EBI Search

EBI Search is a full text search engine based on Apache Lucene library.

It's scalable and provides easy and uniform access to the biological data resources hosted at the European Bioinformatics Institute (EMBL-EBI).

- Domains (indices) are organized in biological categories
- Search executed across domains or categories
- Navigation throughout cross-references

Development, deployment and maintenance are carried on in the Web Production Team.

Indexing

- Indexing
  - Data dumps
    - Splitting
    - Distributed indexing
  - Index verification/validation
  - Deployment in production
  - Daily indexing of updated resources
- Exposed interfaces
  - Web interface from main EMBL-EBI site
  - RESTful interface
    - Java client
    - Python client
    - Perl client

Data indexed in EBI Search

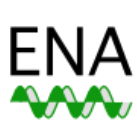
- Categories
  - Genomes
  - Nucleotide Sequences
  - Protein Sequences
  - Macromolecular Structures
  - Gene Expression
  - Molecular Interactions
  - Samples and Ontologies
  - Literature
  - …
- Domains
  - Ensembl
  - ENA
  - Uniprot
  - PDBe
  - Medline
  - Array Express
  - GO
  - BioSamples
  - …

Data indexed in EBI Search

Domains	106
Documents	~1.2bilions (1,191,499,029)
Indices size	534 GB (534,573,875 KB)

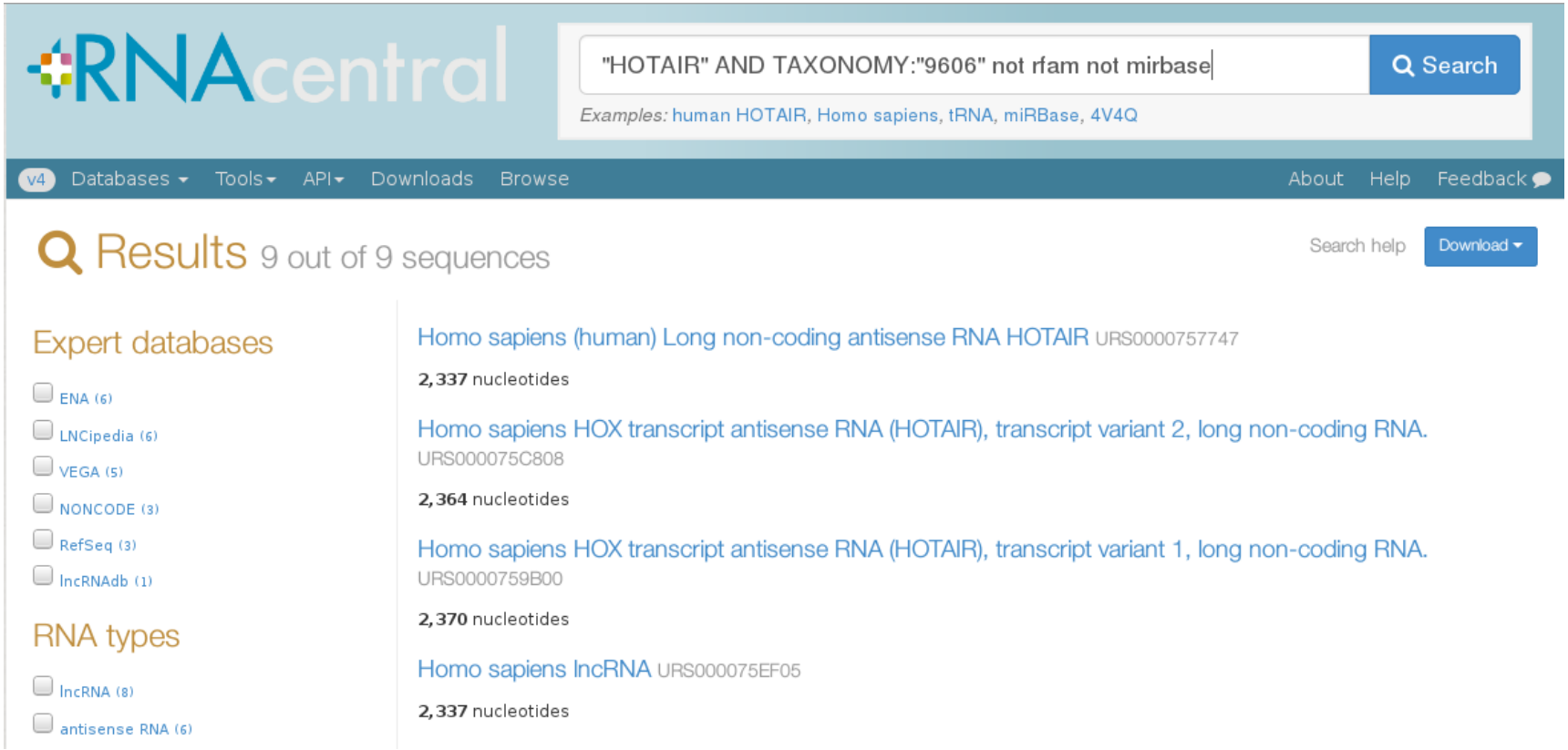
Who is using EBI Search

- European Nucleotide Archive (ENA)
- Ensembl Genomes
- RNAcentral
- InterPro
- MetaboLights
- Enzyme portal
- Expression Atlas
- Omics DI
- - Others -

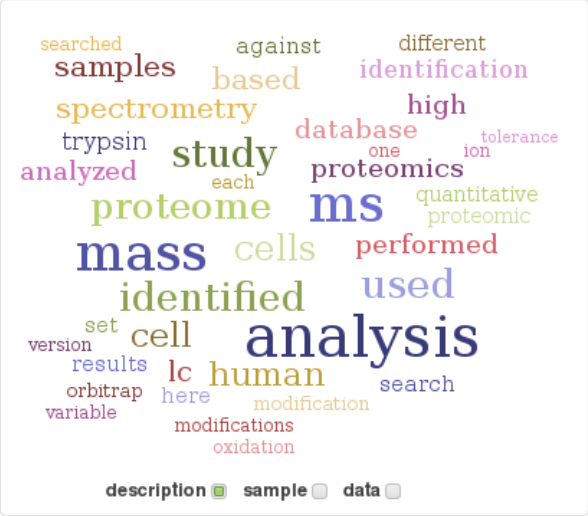

RNAcentral – www.rnacentral.org

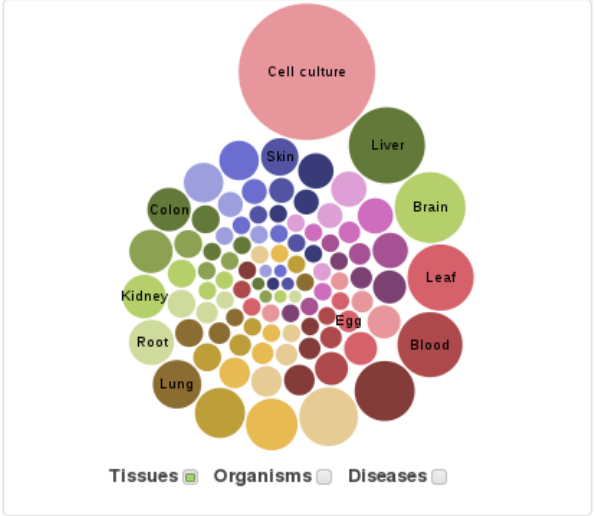
- First RESTful user
- Facets
- Autocompletion

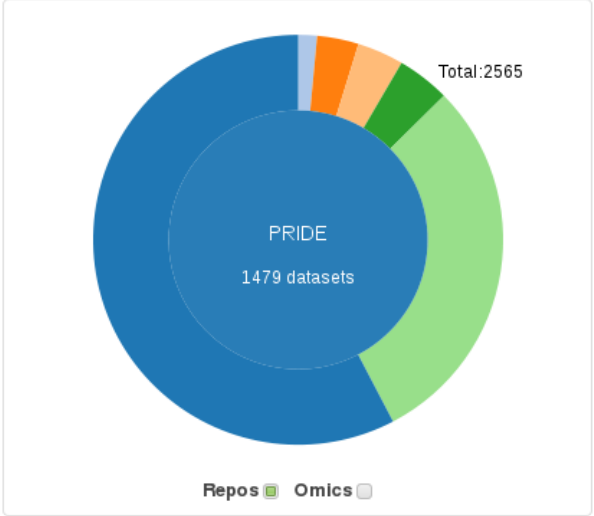


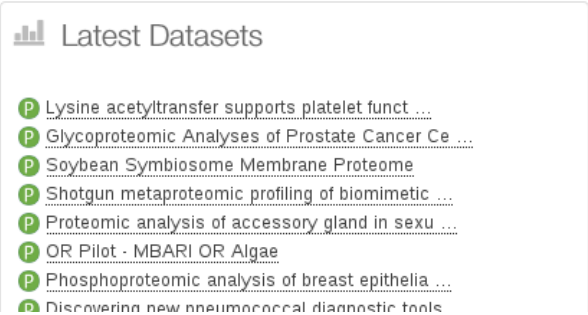
Omics DDI – www.ebi.ac.uk/Tools/omicsdi

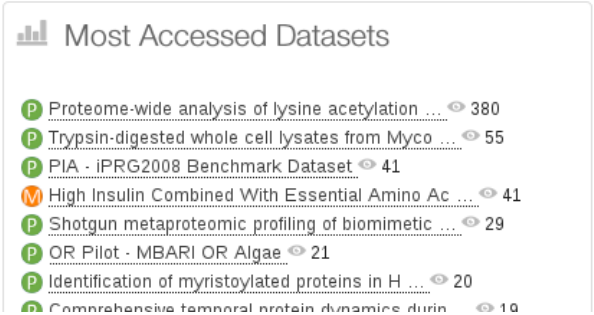
- Experimentation on visualization patterns
- Term frequencies
- More Like this queries

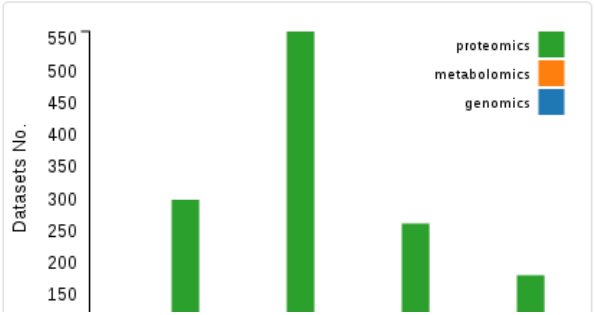












EMBL-EBI

Integration – providing data

- Make a request at www-prod@ebi.ac.uk
- Create dump of your dataset in one of the formats:
  - EBI Search XML dump
  - Flat file (EMBL HGNC)
  - Custom dumps
- EBI Search format preferred

```
<database>
    <name>MetaboLights</name>
    <description>MetaboLights is a database for Metabolomics exp
    <release>4</release>
    <release_date>2016-02-01</release_date>
    <entry_count>18659</entry_count>
    <entries>
        <entry id="MTBLS67">
            <name>Metabolomic Analysis of Fission Yeast at the O
            <description>Microorganisms naturally respond to cha
            <cross_references>
                <ref dbkey="24958269" dbname="pubmed"/>
                <ref dbkey="CHEBI:16551" dbname="ChEBI"/>
                <ref dbkey="MTBLC16551" dbname="MetaboLights"/>
                [...]
                <ref dbkey="MTBLC18413" dbname="MetaboLights"/>
            </cross_references>
            <dates>
                <date type="submission" value="2013-11-19"/>
                <date type="publication" value="2013-11-26"/>
            </dates>
            <additional_fields>
                <field name="repository">MetaboLights</field>
                <field name="omics_type">Metabolomics</field>
                <field name="author">Kenichi Sajiki. G0 Cell Uni
                <field name="dataset_file">ftp://ftp.ebi.ac.uk/p
EMM2-N_15min_2nd_POS.mzML</field>
                <field name="submitter">Tomas Pluskal</field>
                <field name="submitter_email">pluskal@oist.jp</f
                <field name="curator_keywords"/>
                [...]
            </additional_fields>
        </entry>
        [...]
    </entries>
</database>
```

Integration – providing data

```
ID   CM003637; SV 1; linear; genomic DNA; CON; VRT; 24927 BP.
XX
AC   CM003637; AADN04000000;
XX
PR   Project:PRJNA13342;
XX
DT   07-JAN-2016 (Rel. 127, Created)
DT   07-JAN-2016 (Rel. 127, Last updated, Version 1)
XX
DE   Gallus gallus isolate RJF #256 breed Red Jungle fowl, inbred line UCD001
DE   chromosome 30, whole genome shotgun sequence.
XX
KW   .
XX
OS   Gallus gallus (chicken)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda;
OC   Coelurosauria; Aves; Neognathae; Galloanserae; Galliformes; Phasianidae;
OC   Phasianinae; Gallus.
XX
RN   [1]
RC   Erratum:[Nature. 2005 Feb 17;433(7027):777]
RP   1-24927
RX   DOI; 10.1038/nature03154.
RX   PUBMED; 15592404.
RG   International Chicken Genome Sequencing Consortium
RA   Hillier L.W., Miller W., Birney E., Warren W., Hardison R.C., Ponting C.P.,
RA   Bork P., Burt D.W., Groenen M.A., Delany M.E., Dodgson J.B.,
RA   Chinwalla A.T., Cliften P.F., Clifton S.W., Delehaunty K.D., Fronick C.,
RA   Fulton R.S., Graves T.A., Kremitzki C., Layman D., Magrini V.,
RA   McPherson J.D., Miner T.L., Minx P., Nash W.E., Nhan M.N., Nelson J.O.,
RA   Oddy L.G., Pohl C.S., Randall-Maher J., Smith S.M., Wallis J.W., Yang S.P.,
RA   Romanov M.N., Rondelli C.M., Paton B., Smith J., Morrice D., Daniels L.,
RA   Tempest H.G., Robertson L., Masabanda J.S., Griffin D.K., Vignal A.,
RA   Fillon V., Jacobbson L., Kerje S., Andersson L., Crooijmans R.P., Aerts J.,
RA   van der Poel J.J., Ellegren H., Caldwell R.B., Hubbard S.J., Grafham D.V.,
RA   Kierzek A.M., McLaren S.R., Overton I.M., Arakawa H., Beattie K.J.,
RA   Bezzubov Y., Boardman P.E., Bonfield J.K., Croning M.D., Davies R.M.,
RA   Francis M.D., Humphray S.J., Scott C.E., Taylor R.G., Tickle C.,
RA   Brown W.R., Rogers J., Buerstedde J.M., Wilson S.A., Stubbs L.,
RA   Ovcharenko I., Gordon L., Lucas S., Miller M.M., Inoko H., Shiina T.,
RA   Kaufman J., Salomonsen J., Skjoedt K., Wong G.K., Wang J., Liu B., Wang J.,
```

Integration – searching data

- RESTful interface
  - Domain (index) based search
  - Free text (Lucene) queries
  - Facets
  - Pagination
  - Cross reference queries
  - Autocomplete
  - More Like This queries
  - Terms frequency
  - …

RESTful calls example

- http://www.ebi.ac.uk/ebisearch/ws/rest/uniprot?query=water&format=JSON


```
{
  "entries": [
    {
      "acc": "Q6ZY51",
      "id": "PWD_ARATH",
      "source": "uniprot"
    },
    {
      "acc": "Q2QTC2",
      "id": "PWD_ORYSJ",
      "source": "uniprot"
    },
    {
      "acc": "Q13R51",
      "id": "SELD_BURXL",
      "source": "uniprot"
    },
    {
      "acc": "B1Z181",
```

RESTful calls example

- http://www.ebi.ac.uk/ebisearch/ws/rest/uniprot?query=water&facetcount=10&format=JSON


```
{
  "entries": [ [...] ],
  "facets": [
    {
      "facetValues": [
        {
          "count": 30994,
          "label": "Daphnia pulex",
          "value": "6669"
        },
        {
          "count": 9628,
          "label": "Variovorax paradoxus",
          "value": "34073"
        }, [...]
      ],
      "id": "TAXONOMY",
      "label": "Organisms",
      "total": 10650
    },
    {
```

Contacts

- Web Production Team - South building - Studio 15
- www-prod@ebi.ac.uk
- https://www.ebi.ac.uk/seqdb/confluence/display/EXT/EBI+Search+Engine (needs a confluence user)
- https://www.ebi.ac.uk/ebisearch/documentation.ebi

- https://www.ebi.ac.uk/ebisearch/