

EBI Search

Search your biological data at EMBL-EBI

Web Production

Nicola Buso

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 -



ENA

LRG



PomBase



*e!*EnsemblGenomes



Interpro

emboss

RNAcentral

RNAcentral – www.rnacentral.org

- First RESTful user
- Autocompletion
- Facets

The screenshot displays the RNAcentral website interface. At the top, the RNAcentral logo is on the left, and a search bar on the right contains the query '"HOTAIR" AND TAXONOMY:"9606" not rfam not mirbase'. Below the search bar, a navigation bar includes links for 'v4', 'Databases', 'Tools', 'API', 'Downloads', 'Browse', 'About', 'Help', and 'Feedback'. The main content area shows 'Results 9 out of 9 sequences'. On the left, there are two sections: 'Expert databases' with checkboxes for ENA (6), LNCipedia (6), VEGA (5), NONCODE (3), RefSeq (3), and lncRNAdb (1); and 'RNA types' with checkboxes for lncRNA (8) and antisense RNA (6). The right side lists search results for 'Homo sapiens (human) Long non-coding antisense RNA HOTAIR' (URS0000757747, 2,337 nucleotides), 'Homo sapiens HOX transcript antisense RNA (HOTAIR), transcript variant 2, long non-coding RNA' (URS000075C808, 2,364 nucleotides), 'Homo sapiens HOX transcript antisense RNA (HOTAIR), transcript variant 1, long non-coding RNA' (URS0000759B00, 2,370 nucleotides), and 'Homo sapiens lncRNA' (URS000075EF05, 2,337 nucleotides).

RNAcentral

"HOTAIR" AND TAXONOMY:"9606" not rfam not mirbase

Search

Examples: human HOTAIR, Homo sapiens, tRNA, miRBase, 4V4Q

v4 Databases Tools API Downloads Browse About Help Feedback

Q Results 9 out of 9 sequences

Search help Download

Expert databases

- ☐ ENA (6)
- ☐ LNCipedia (6)
- ☐ VEGA (5)
- ☐ NONCODE (3)
- ☐ RefSeq (3)
- ☐ lncRNAdb (1)

RNA types

- ☐ lncRNA (8)
- ☐ antisense RNA (6)

Homo sapiens (human) Long non-coding antisense RNA HOTAIR URS0000757747

2,337 nucleotides

Homo sapiens HOX transcript antisense RNA (HOTAIR), transcript variant 2, long non-coding RNA. URS000075C808

2,364 nucleotides

Homo sapiens HOX transcript antisense RNA (HOTAIR), transcript variant 1, long non-coding RNA. URS0000759B00

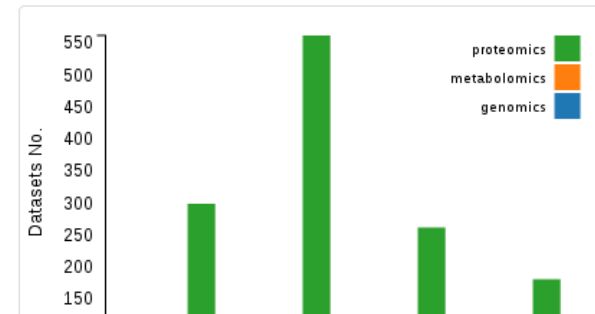
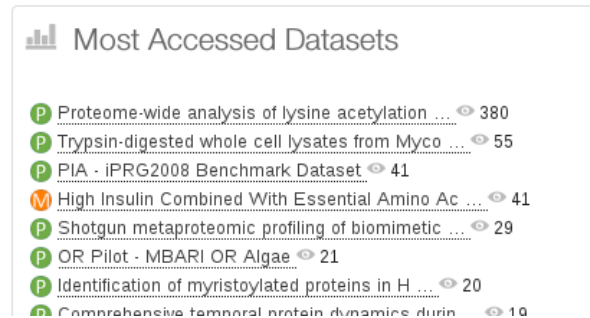
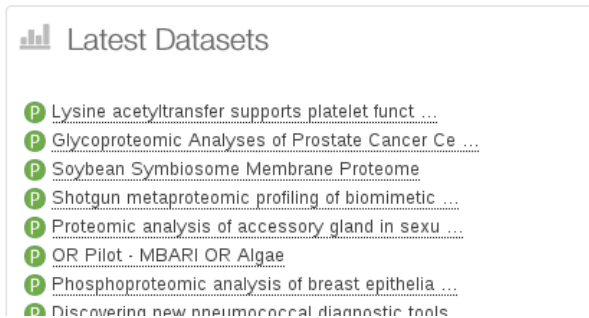
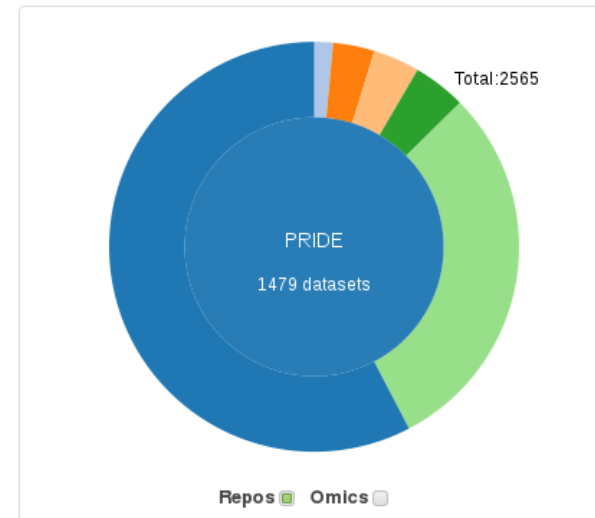
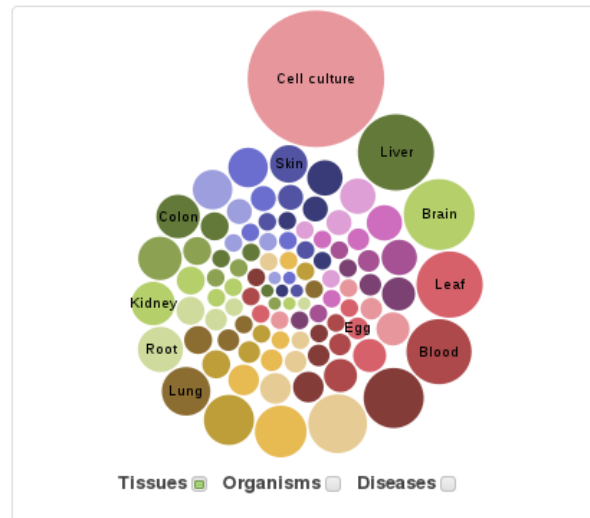
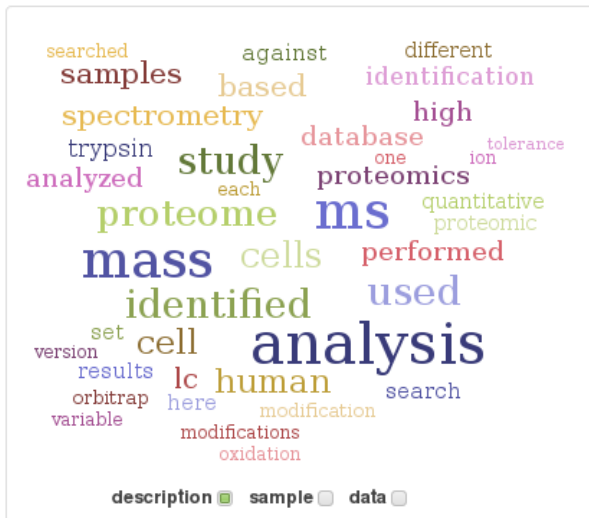
2,370 nucleotides

Homo sapiens lncRNA URS000075EF05

2,337 nucleotides

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

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



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

```
1 <database>
2   <name>MetaboLights</name>
3   <description>MetaboLights is a database for Metabolomics exp
4   <release>4</release>
5   <release_date>2016-02-01</release_date>
6   <entry_count>18659</entry_count>
7   <entries>
8     <entry id="MTBLS67">
9       <name>Metabolomic Analysis of Fission Yeast at the O
10      <description>Microorganisms naturally respond to cha
11      <cross_references>
12        <ref dbkey="24958269" dbname="pubmed"/>
13        <ref dbkey="CHEBI:16551" dbname="ChEBI"/>
14        <ref dbkey="MTBLC16551" dbname="MetaboLights"/>
15        [...]
16        <ref dbkey="MTBLC18413" dbname="MetaboLights"/>
17      </cross_references>
18      <dates>
19        <date type="submission" value="2013-11-19"/>
20        <date type="publication" value="2013-11-26"/>
21      </dates>
22      <additional_fields>
23        <field name="repository">MetaboLights</field>
24        <field name="omics_type">Metabolomics</field>
25        <field name="author">Kenichi Sajiki. G0 Cell Uni
26        <field name="dataset_file">ftp://ftp.ebi.ac.uk/p
27        EMM2-N_15min_2nd_POS.mzML</field>
28        <field name="submitter">Tomas Pluskal</field>
29        <field name="submitter_email">pluskal@oist.jp</f
30        <field name="curator_keywords"/>
31        [...]
32      </additional_fields>
33    </entry>
34    [...]
35  </entries>
36</database>
```

Integration – providing data

```
1 ID      CM003637; SV 1; linear; genomic DNA; CON; VRT; 24927 BP.
2 XX
3 AC      CM003637; AADN04000000;
4 XX
5 PR      Project:PRJNA13342;
6 XX
7 DT      07-JAN-2016 (Rel. 127, Created)
8 DT      07-JAN-2016 (Rel. 127, Last updated, Version 1)
9 XX
10 DE     Gallus gallus isolate RJF #256 breed Red Jungle fowl, inbred line UCD001
11 DE     chromosome 30, whole genome shotgun sequence.
12 XX
13 KW     .
14 XX
15 OS     Gallus gallus (chicken)
16 OC     Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
17 OC     Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda;
18 OC     Coelurosauria; Aves; Neognathae; Galloanserae; Galliformes; Phasianidae;
19 OC     Phasianinae; Gallus.
20 XX
21 RN     [1]
22 RC     Erratum:[Nature. 2005 Feb 17;433(7027):777]
23 RP     1-24927
24 RX     DOI; 10.1038/nature03154.
25 RX     PUBMED; 15592404.
26 RG     International Chicken Genome Sequencing Consortium
27 RA     Hillier L.W., Miller W., Birney E., Warren W., Hardison R.C., Ponting C.P.,
28 RA     Bork P., Burt D.W., Groenen M.A., Delany M.E., Dodgson J.B.,
29 RA     Chinwalla A.T., Cliften P.F., Clifton S.W., Delehaunty K.D., Fronick C.,
30 RA     Fulton R.S., Graves T.A., Kremitzki C., Layman D., Magrini V.,
31 RA     McPherson J.D., Miner T.L., Minx P., Nash W.E., Nhan M.N., Nelson J.O.,
32 RA     Oddy L.G., Pohl C.S., Randall-Maher J., Smith S.M., Wallis J.W., Yang S.P.,
33 RA     Romanov M.N., Rondelli C.M., Paton B., Smith J., Morrice D., Daniels L.,
34 RA     Tempest H.G., Robertson L., Masabanda J.S., Griffin D.K., Vignal A.,
35 RA     Fillon V., Jacobsson L., Kerje S., Andersson L., Crooijmans R.P., Aerts J.,
36 RA     van der Poel J.J., Ellegren H., Caldwell R.B., Hubbard S.J., Grafham D.V.,
37 RA     Kierzek A.M., McLaren S.R., Overton I.M., Arakawa H., Beattie K.J.,
38 RA     Bezzubov Y., Boardman P.E., Bonfield J.K., Croning M.D., Davies R.M.,
39 RA     Francis M.D., Humphray S.J., Scott C.E., Taylor R.G., Tickle C.,
40 RA     Brown W.R., Rogers J., Buerstedde J.M., Wilson S.A., Stubbs L.,
41 RA     Ovcharenko I., Gordon L., Lucas S., Miller M.M., Inoko H., Shiina T.,
42 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>

```
1 {
2   "entries": [
3     {
4       "acc": "Q6ZY51",
5       "id": "PWD_ARATH",
6       "source": "uniprot"
7     },
8     {
9       "acc": "Q2QTC2",
10      "id": "PWD_ORYSJ",
11      "source": "uniprot"
12    },
13    {
14      "acc": "Q13R51",
15      "id": "SELD_BURXL",
16      "source": "uniprot"
17    },
18    {
19      "acc": "B1Z181",
20      "id": "SELD_BURXL"
```

RESTful calls example

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

```
1 {  
2   "entries": [ [ ... ] ],  
3   "facets": [  
4     {  
5       "facetValues": [  
6         {  
7           "count": 30994,  
8           "label": "Daphnia pulex",  
9           "value": "6669"  
10        },  
11        {  
12          "count": 9628,  
13          "label": "Variovorax paradoxus",  
14          "value": "34073"  
15        }, [ ... ]  
16      ],  
17      "id": "TAXONOMY",  
18      "label": "Organisms",  
19      "total": 10650  
20    },  
21  ]
```

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/>

