

EBI Search

Biological data search engine

Nicola Buso – Web Production Team

European Bioinformatics Institute - EMBL-EBI

www.ebi.ac.uk

Overview

- EBI Search what it provides
- When and how it started and evolved
- The data we index and search
- Indexing
- Searching
- Log analysis, statistics
- Future plans

EBI Search what it provides

- Apache Lucene based full text search engine
- Open source & EMBL-EBI => free data
- Developed as a bespoke solution for EMBL-EBI data
- Different interfaces: Web, SOAP, REST
- Access to Biological data ~1.1B documents (mainly from EMBL-EBI)
- Easy search solution for many projects

Web Interface

Search results for **brca2**

Showing 21 results out of 15,405 in All results

Filter your results

Source

All results (15,405)

[Genomes](#) (511)

[Nucleotide sequences](#) (3,080)

[Protein sequences](#) (2,451)

[Macromolecular structures](#) (34)

[Small molecules](#) (21)

[Gene expression](#) (125)

[Molecular interactions](#) (250)

[Reactions, pathways & diseases](#) (1,611)

[Protein families](#) (15)

[Literature](#) (5,279)

[Samples & ontologies](#) (2,006)

[EBI web](#) (22)

Gene & protein summaries (includes expression, structures, literature...) (3 results found)



[Breast cancer 2, early onset](#)

BRCA2 (603468, FACD, FAD1, PNCA2, FAD, FANCD1, BRCC2, GLM3, BROVCA2, FANCD, ENSG00000139618)

Human (Homo sapiens)



[Breast cancer 2](#)

Brca2 (RAB163, Fancd1, ENSMUSG00000041147)

House Mouse (Mus musculus)

More...

Reactions, pathways & diseases (1,611 results found)

[BRCA2](#)

Stable Id: REACT_116541

Type: Protein

Species: Gallus gallus

Related data ▾

Views ▾

Source: Reactome
ID: 265880

[View all 1611 results for Reactions, pathways & diseases](#)

Genomes (511 results found)

[breast cancer 2, early onset](#)

Approved Symbol: **BRCA2**

Approved Name: breast cancer 2, early onset

Status: (Approved)

Aliases: FAD FAD1 BRCC2 XRCC11

Locus Type: gene with protein product

Chromosome: 13q12-q13

Related data ▾

Views ▾

Source: HGNC
ID: HGNC:1101

[Summary information is available for this gene](#)

Web Interface 2

Search results for **brca2**

Showing 21 results out of 15,405 in All results

Filter your results

Source

- All results (15,405)
- Genomes (511)
- Nucleotide sequences (3,080)
- Protein sequences (2,451)
- Macromolecular structures (34)
- Small molecules (21)
- Gene
- Molecu
- Reacti
- Protein
- Literat
- Sampl
- EBI we

Gene & protein summaries (includes expression, structures, literature...) (3 results found)



[Breast cancer 2, early onset](#)

BRCA2 (603468, FACD, FAD1, PNCA2, FAD, FANCD1, BRCC2, GLM3, BROVCA2, FANCD, ENSG00000139618)

Human (Homo sapiens)

[Breast cancer 2](#)

Search results for **brca2**

Showing 15 results out of 511 in All results → Genomes

Filter your results

Source

- All results (15,405)
- Genomes (511)
- HGNC (8)
- LRG (2)
- Ensembl Gene (299)
- Ensembl Genomes Gene (148)
- WormBase ParaSite (54)

Organisms

- Homo sapiens (23)
- Triticum aestivum (13)
- Mus musculus (11)
- Danio rerio (8)
- Rattus norvegicus (6)
- Lepisosteus oculatus (6)
- Dipodomys ordii (6)
- Ailuropoda melanoleuca (5)
- Astyanax mexicanus (5)

Genomes (511 results found)

[breast cancer 2, early onset](#)

[Related data](#) [Views](#)

Approved Symbol: **BRCA2**

Approved Name: breast cancer 2, early onset

Status: (Approved)

Aliases: FAD FAD1 BRCC2 XRCC11

Locus Type: gene with protein product

Chromosome: 13q12-q13

Source: HGNC
ID: HGNC:1101

Summary information is available for this gene

[FBgn0050169](#)

[Related data](#) [Views](#)

Breast cancer 2, early onset homolog [Source:FlyBase gene name;Acc:FBgn0050169]

Species: Drosophila melanogaster

Source: Ensembl Genomes Gene
ID: FBgn0050169

[FBgn0075710](#)

[Related data](#) [Views](#)

Species: Drosophila pseudoobscura

Source: Ensembl Genomes Gene
ID: FBgn0075710

When and how it evolved

- Started in 2006, launched in 2007
- Copes with heterogeneous and great volumes of data
- Closely integrated with EMBL-EBI's IT infrastructure:
 - Shared resources
 - Distributed FS: shared across nodes, proprietary sync tools; good performance
 - 2 data centres => data synchronization
 - Deals with changes in data and infrastructure

The data we index and search

- ~3.3TB of data => ~450GB indexes => ~10/12 hours
- “metadata” rather than “data”
- Heterogeneous types of Biological data
 - Genes, proteins, nucleotide sequences: structured/annotated data
 - Literature, patent databases, diseases: text rich
- Specific analysers: Chemical, reaction formulas (e.g. InChi)

EBI Data

Genes, genomes & variation

European Nucleotide Archive
1000 Genomes

Ensembl
Ensembl Genomes

European Genome-phenome Archive
Metagenomics portal

Gene, protein & metabolite expression

ArrayExpress
Expression Atlas

Metabolights
PRIDE

Protein sequences, families & motifs

InterPro

Pfam

UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

ChEMBL

ChEBI

Systems

BioModels

Enzyme Portal

BioSamples

Cross domain resources . Cross domain resources

Literature & ontologies

Europe PubMed Central
Gene Ontology
Experimental Factor Ontology

Reactions, interactions & pathways

IntAct

Reactome

MetaboLights

Indexing

- Data to index is in various formats:
 - data resource specific (flat and XML files)
 - generic XML with a common set of fields/annotations (id, name, description)
- VMs & LSF indexing environments
 - Parallelized
 - Daily update and indexing of data
 - Indexes verification
 - Driven by biological data release cycles

Searching 1

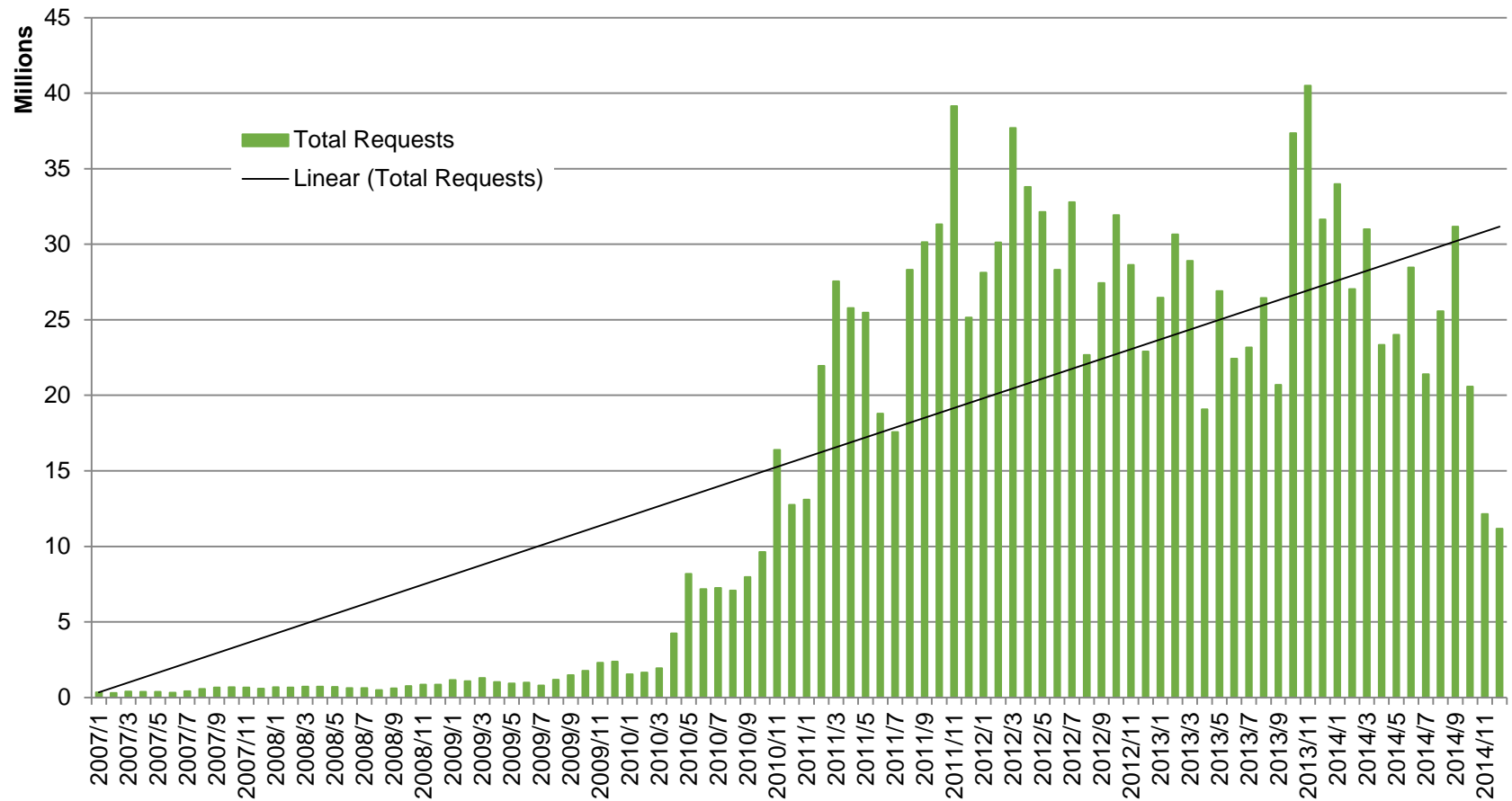
- 97 indexes categorized in a 2 level hierarchy
- Search on categories or leaf indexes; keeping consistent scoring across categories
- Faceting
- Cross reference search
- Autocomplete functionality
- Distributed cache (HazelCast)
- Presentation of heterogeneous results

Searching 2

- Integration with other biological Tools/Services
- Work with UXD experts
- Exploration of different designs
 - Gene & protein summaries

The screenshot displays the EBI Search website interface. At the top, there is a navigation bar with links for 'Services', 'Research', 'Training', and 'About us'. The main header features the 'EBI Search' logo and a search input field containing 'brca1'. Below the search bar, there are links for 'Help & Documentation' and 'About EBI Search', along with 'Print', 'Share', and 'Feedback' options. The search results section shows 'Search results for brca1' and a breadcrumb trail: 'Gene & protein summaries for brca1 > Gene summary for brca1'. The main content area is titled 'Gene & protein summary for brca1'. An 'ORGANISMS' dropdown menu is set to 'human' (Homo sapiens). A sidebar on the left contains navigation buttons for 'Gene', 'Expression', 'Protein', 'Protein Structure', and 'Literature'. The main content area displays the gene name 'Breast cancer 1, early onset' and a 'View in Ensembl' link. Under the heading 'Gene Information and Sequence', it provides details: 'BRCA1 spans 81188 bps of chromosome 17 from 43044295 to 43125483. BRCA1 has 29 transcripts containing a total of 100 exons on the reverse strand. Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article.' It also includes links to 'View the gene sequence in Ensembl' and 'View the chromosome region for this gene in Ensembl'. A 'Variations' section is partially visible at the bottom.

Log Analysis Statistics



Future plans

- Compare against other search engines
- Focus on scalability and reducing search time:
 - n. docs 2010 ~400M => n. docs 2014 ~1.1B
 - Index incremental updates
- Experimenting with novel search features
 - Expanding facets
 - Dealing with new data types
 - Investigate different visualizations
- Collaborations with the open source community

Reference & Acknowledgements

- Valentin F, Squizzato S, Goujon M, McWilliam H, Paern J, Lopez R. Fast and efficient searching of biological data resources--using EB-eye. Brief Bioinform. 2010 Jul;11(4) 375-384. doi:10.1093/bib/bbp065. PubMed PMID: 20150321; PubMed Central PMCID: PMC2905521

Web Production Team:

- Rodrigo Lopez, Silvano Squizzato, Young Mi Park, Tamer Gur, Nicola Buso
- www.ebi.ac.uk/ebisearch/

Thanks

Questions?

Dump formats – XML enriched

```
<database>
  <name>ailuropoda_melanoleuca_core_78_1</name>
  <description>Ensembl Ailuropoda melanoleuca core database</description>
  <release>78</release>
  <entries>
    <entry id="ENSAMEG000000000001">
      <name>ENSAMEG000000000001 (HGNC: HNF4G)</name>
      <description>hepatocyte nuclear factor 4, gamma [Source:HGNC Symbol;Acc:HGNC:5026]</description>
      <cross_references>
        <ref dbname="ncbi_taxonomy_id" dbkey="9646" />
        .....
      </cross_references>
      <additional_fields>
        <field name="species">Ailuropoda melanoleuca</field>
        .....
      </additional_fields>
    </entry>
    .....
  </entries>
</database>
```

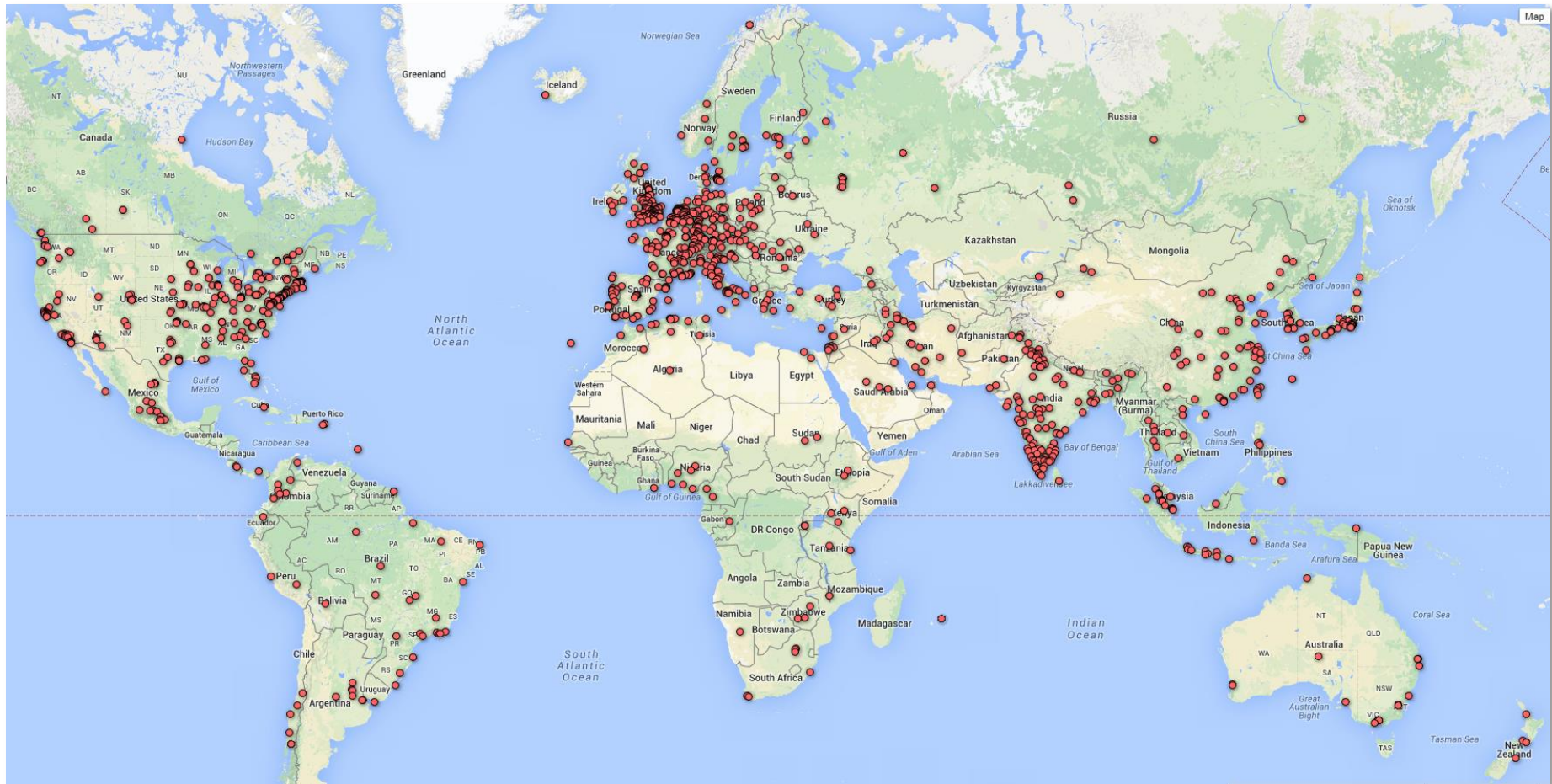

Dump formats – RAW provider (flat file)

OS *Saccharomyces* sp. 'boulardii'
OC Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes;
OC Saccharomycetales; Saccharomycetaceae; *Saccharomyces*.
XX
RN [1]
RP 1-650684
RA Batista T.M., Castro I.M., Araujo F.M., Salim A.C., Brandao R.L.,
RA Drummond M.G., Cardoso D.C., Oliveira G.C., Rosa C.A., Nicoli J.R.,
RA Franco G.R.;
RT "Whole genome sequence of the probiotic yeast *Saccharomyces cerevisiae* var
RT *boulardii* 17 (marketed in Brazil)";
RL Unpublished.
XX
RN [2]
RP 1-650684
RA Batista T.M., Castro I.M., Araujo F.M., Salim A.C., Brandao R.L.,
RA Drummond M.G., Cardoso D.C., Oliveira G.C., Rosa C.A., Nicoli J.R.,
RA Franco G.R.;
RT ;

Dump formats – XML provider 1

```
<?xml version="1.0" encoding="UTF-8"?>
<uniprot xmlns="http://uniprot.org/uniprot"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://uniprot.org/uniprot http://www.uniprot.org/support/docs/uniprot.xsd">
<entry version="23" modified="2015-01-07" created="2011-06-28" dataset="Swiss-Prot">
  <accession>Q6GZV8</accession>
  <name>017L_FRG3G</name>
  <protein>
    <recommendedName>
      <fullName>Uncharacterized protein 017L</fullName>
    </recommendedName>
  </protein>
  <gene>
    <name type="ORF">FV3-017L</name>
  </gene>
  <organism>
    <name type="scientific">Frog virus 3</name>
    <name type="common">isolate Goorha</name>
    <name type="synonym">FV-3</name>
    <dbReference id="654924" type="NCBI Taxonomy"/>
  <lineage>
```

Geolocation distinct IPs (no robots) in 2014



Monitoring and log analysis

- Application log analysis
 - Query log analytics based on Solr
- HTTP access log analysis
 - ELK stack (Elasticsearch, Logstash and Kibana)