Proteomics Resources at EBI

Michele Magrane
magrane@ebi.ac.uk
Introduction

- EBI provides a number of databases and tools for the deposition, distribution and analysis of proteomics data
- All resources are freely available
- Contribute to the development of community standards for proteomics data to enhance data standardisation and exchange
- Close collaboration and regular data exchange between the resources ensures interoperability and data synchronisation
Proteomics resources

Protein sequence and functional data
www.uniprot.org

Protein interaction data
www.ebi.ac.uk/intact

Pathways and reactions
www.reactome.org

Protein/peptide identifications
www.ebi.ac.uk/pride
UniProt Consortium
UniProt databases

UniProtKB
Protein Knowledgebase

UniProtKB/Swiss-Prot
Reviewed ★

Expert manual annotation

UniProtKB/TrEMBL
Unreviewed ★

Automatic annotation

UniMES
Metagenomic and environmental sequences

UniMES Clusters
Cluster 100
Cluster 90

UniParc
Sequence archive
New, revised and obsolete sequences

UniRef
Sequence clusters
UniRef100
UniRef90
UniRef50

EMBL/GenBank/DDBJ (Metagenomics), Ensembl, VEGA, RefSeq, PDB, MODs, other sequence resources
UniProtKB consists of two sections

UniProtKB/Swiss-Prot
- Non-redundant
- High-quality
- Manually curated
  *Reviewed*

UniProtKB/TrEMBL
- Redundant
- Computationally generated
- Automatically annotated
  *Unreviewed*
UniProtKB sequence data

Submissions

Literature scanning
Extensively cross-referenced

2D-gel
3D-structure
Enzyme/pathways
Families/domains
Gene expression
Genome annotation
Ontologies
Organism-specific
Phylogenomics
Polymorphisms
Protein interactions
Proteomics
PTMs
Sequence
Manual curation in UniProtKB/Swiss-Prot

Splice variants

Sequence

Sequence features

UniProt

Annotations

Protein names

Protein quaking

Also known as:

Nomenclature

Gene names

Name: Qki
Synonyms: Qk, Qk1, Qka1

References


[Cited twice]

[2] "Qki Kim" Unlab. unreviewed. "quaking (mouse) [UniProtKB/Swiss-Prot]" [GO:0006785] [SO:0001107] [SO:0001042] [SO:0001041] [SO:0001043] [SO:0001044] [SO:0001045] [SO:0001046]

(Cited: 120)

Ontologies

Cell cycle
DNA damage
DNA repair
Fatty acid biosynthesis
Lipid synthesis
Nucleus
Polymerase
Disease
Repeat
Zinc-finger
DNA-binding
Metal-binding
Zinc
Anti-oncogene
Phosphorylation
3D-structure

References


[Cited twice]

[2] "Qki Kim" Unlab. unreviewed. "quaking (mouse) [UniProtKB/Swiss-Prot]" [GO:0006785] [SO:0001107] [SO:0001042] [SO:0001041] [SO:0001043] [SO:0001044] [SO:0001045] [SO:0001046]

(Cited: 120)
UniProtKB data growth

**UniProtKB/Swiss-Prot** contains 532,146 manually curated entries

**UniProtKB/TrEMBL** contains 16,886,838 unreviewed entries
Automatic annotation in UniProtKB/TrEMBL

- Allows annotation of UniProtKB/TrEMBL in an efficient and scalable manner with a high degree of accuracy

- Based on annotation rules which are created, tested and validated against published experimental data in UniProtKB/Swiss-Prot

- Rules are linked to InterPro member database signatures and define annotations to be added and conditions which must be fulfilled

- Signatures identify family members in UniProtKB

- Common annotation in Swiss-Prot is transferred to related family members in TrEMBL if they fulfil rule conditions
Two complementary approaches

**UniRule**
Manually curated rules devised and tested by curation team

**SAAS**
Automatic decision tree-based rule-generating system

---

Rules are reapplied to UniProtKB/TrEMBL as part of each four-weekly release with both automatic and manual QA procedures ensuring they are still valid.
Evidence attribution

• System which allows linking of all information in an entry to its original source

• Allows users:
  • to trace origin of all data
  • to differentiate easily between experimental and computational data
  • to assess data reliability
### General annotation (Comments)

**Function**
Required for maintenance of euploidy during cleavage-stage embryogenesis. Ensures proper spindle assembly by regulating the localization of AURKA via RHOA signaling and of PLK1 via an RHOA-independent process. Required for the localization of MAD2L1 to kinetochores to enable spindle assembly checkpoint function. (Ref1) (Ref2) (Ref3) (Ref4)

**Subunit structure**
Component of the subcortical maternal complex (SCMC) which is essential for progression of zygotes beyond the first embryonic cell division. The complex also contains NLRP5, OOEPE and TLE6. Within the complex, interacts with NLRP5. (Ref1) (Ref2) (Ref3) (Ref4)

**Subcellular location**
Cytoplasm > cell cortex. Note: Located throughout the cell cortex of ovulated eggs in a complex with NLRP5. After fertilization, restricted to the apical cortex and excluded from regions of cell-cell contact. (Ref1) (Ref2) (Ref3) (Ref4)

**Tissue specificity**
Detected in ovary, but not in testis or somatic tissues. In the ovary, expressed in growing oocytes. (Ref1) (Ref2) (Ref3) (Ref4)

**Developmental stage**
Isoform 2 is detected in growing oocytes, ovulated eggs and preimplantation embryos up to the morula stage and decreases markedly at the blastocyst stage (at protein level). Isoform 1 is detected in growing oocytes but diminishes in fully grown oocytes. Detected at very low levels in morula and early blastocysts. (Ref1) (Ref2) (Ref3) (Ref4)

**Disruption phenotype**
Reduced fecundity and impaired preimplantation embryo development with a high incidence of aneuploidy due to abnormal spindle assembly, chromosomal misalignment and spindle assembly checkpoint inactivation. (Ref1) (Ref2) (Ref3) (Ref4)

**Caution**
Considered by a number of resources to be the ortholog of human CGorf221/ECAT1. However, sequence similarity is low and synteny is not conserved. According to PMID 17913455, CGorf221 has been lost in rodents. (Ref1) (Ref2) (Ref3) (Ref4)

### Sequence annotation (Features)

<table>
<thead>
<tr>
<th>Feature key</th>
<th>Position(s)</th>
<th>Length</th>
<th>Description</th>
<th>Graphical view</th>
<th>Feature identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Molecule processing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chain</td>
<td>1 - 440</td>
<td>440</td>
<td>Protein Filla</td>
<td></td>
<td>PRO_0000407378</td>
</tr>
<tr>
<td><strong>Regions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compositional bias</td>
<td>129 - 369</td>
<td>241</td>
<td>Ala-n-rch</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Natural variations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternative sequence</td>
<td>148 - 171</td>
<td>24</td>
<td>Missing in isoform 1</td>
<td></td>
<td>VSP_040960</td>
</tr>
<tr>
<td>Alternative sequence</td>
<td>341 - 346</td>
<td>6</td>
<td>VREAAT → ESQRT in isoform 2</td>
<td></td>
<td>VSP_0404951</td>
</tr>
<tr>
<td>Alternative sequence</td>
<td>347 - 440</td>
<td>94</td>
<td>Missing in isoform 2</td>
<td></td>
<td>VSP_040952</td>
</tr>
</tbody>
</table>
Complete proteomes in UniProtKB

• Complete proteome sets are available for more than 2700 species:

  http://www.uniprot.org/taxonomy/complete-proteomes

• All IPI species are included in response to the imminent closure of IPI: human, mouse, rat, zebrafish, Arabidopsis, chicken, cow

• A subset of 455 complete proteomes have been manually defined as reference proteomes which have been selected to provide broad taxonomic coverage

  http://www.uniprot.org/faq/15
Finding a complete proteome in UniProtKB

COMPLETE PROTEOMES AND REFERENCE PROTEOMES

A complete proteome consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced.

A reference proteome is the complete proteome of a representative, well-studied model organism or an organism of interest for biomedical research.

These organisms can be searched via the taxonomy pages, which provide links to download complete and reference proteome sets when available, as well as links to the HAMAP and/or IntegR8 web sites.

Browse or list organisms with:

Complete proteomes
- Browse by hierarchy
- List all Bacteria
- List all Archaea
- List all Eukaryota
- List all Viruses

Search organisms with complete proteomes:

Reference proteomes
- Browse by hierarchy
- List all Bacteria
- List all Archaea
- List all Eukaryota
- List all Viruses

Search organisms with reference proteomes:
Protein existence

<table>
<thead>
<tr>
<th>Names and origin</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Protein names</strong></td>
</tr>
<tr>
<td><em>Recommended name:</em> E3 ubiquitin-protein ligase Mdm2</td>
</tr>
<tr>
<td><em>EC:</em> 6.3.2.7</td>
</tr>
<tr>
<td><em>Alternative name(s):</em> Double minute 2 protein</td>
</tr>
<tr>
<td><em>Short name:</em> Mdm2</td>
</tr>
<tr>
<td><em>Oncoprotein Mdm2</em></td>
</tr>
<tr>
<td><em>p53-binding protein Mdm2</em></td>
</tr>
<tr>
<td><strong>Gene names</strong></td>
</tr>
<tr>
<td><em>Name:</em> MDM2</td>
</tr>
<tr>
<td><strong>Organism</strong></td>
</tr>
<tr>
<td><em>Homo sapiens (Human)</em></td>
</tr>
<tr>
<td><strong>Taxonomic identifier</strong></td>
</tr>
<tr>
<td><em>9606 [NCBI]</em></td>
</tr>
<tr>
<td><strong>Taxonomic lineage</strong></td>
</tr>
<tr>
<td><em>Eukaryota &gt; Metazoa &gt; Chordata &gt; Craniata &gt; Vertebrata &gt; Euteleostomi &gt; Mammalia &gt; Eutheria &gt; Euarchontoglires &gt; Primates &gt; Haplorhini &gt; Catarrhini &gt; Hominidae &gt; Homo</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence length</strong></td>
</tr>
<tr>
<td><em>491 AA.</em></td>
</tr>
<tr>
<td><strong>Sequence status</strong></td>
</tr>
<tr>
<td><em>Complete.</em></td>
</tr>
<tr>
<td><strong>Protein existence</strong></td>
</tr>
<tr>
<td><em>Evidence at protein level</em></td>
</tr>
</tbody>
</table>

http://www.uniprot.org/docs/pe_criteria
What UniProtKB can be used for

- Access wide range of manually reviewed experimental data
- Explore the published literature for a protein
- Browse and download complete proteome sets
- Link to information in more than 140 other resources
- Use range of tools such as simple and advanced querying, Blast, sequence alignment, batch retrieval and identifier mapping
Protein interactions

- Proteins rarely fulfill their function in isolation.
- Instead, they typically form transient or stable interactions with other molecules.
- Determination and analysis of these interactions are essential to our biological and biomedical understanding of cellular function.
- Large number of experimental approaches are used to determine molecular interactions.
www.ebi.ac.uk/intact

• Freely available, open source database system and analysis tools for protein interaction data

• All interactions are derived from literature curation or direct user submissions

• Deep annotation model which provides a high level of detail

• Part of the International Molecular Exchange (IMEx) Consortium
IMEx Consortium

- International collaboration between major public interaction data providers
- Share curation effort
- Utilise common standards for non-overlapping curation of scientific literature
- Each database covers distinct set of journals
Accessing interaction data in IntAct

- IntAct contains:
  - 275,144 binary interactions
  - 57,857 proteins
  - 14,108 experiments

- Database can be searched with range of search terms:
  - gene names, UniProtKB accessions, PubMed identifiers, drug names
Experiment (2 interactions)
Accession: EBI-1784897
Name: boeddrich-2006-9
Host organism: *in vitro* [*in vitro*]
Interaction Detection Method: competition binding
Participant Identification Method: anti tat western

<table>
<thead>
<tr>
<th>Cross References</th>
<th>Database</th>
<th>Identifier</th>
<th>Secondary identifier</th>
<th>Qualifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pubmed</td>
<td>16525503</td>
<td>-</td>
<td>primary-reference</td>
</tr>
<tr>
<td></td>
<td>newt</td>
<td>5605</td>
<td>human</td>
<td>target-species</td>
</tr>
</tbody>
</table>

Annotations:
- Topic: author-list
- Topic: journal
  - *EMBO J.* (C261-4198)
- Topic: publication year
  - 2006
- Topic: contact-email
  - ewanker@mdc-berlin.de

Publication
Title: An arginine/lysine-rich motif is crucial for VCP/p97-mediated modulation of ataxin-3 fibrillogenesis.
Journal: *EMBO J.* (C261-4198)
Year of Publication: 2006
PubMed id: 16525503

Interaction
Accession: EBI-1784905
Name: atxn3-vcp-1
Description: -
Type: directed interaction
Annotations:
- Topic: figure legend
  - 4B, 4F

Participants (3)
What IntAct can be used for

• Find molecules that interact with your protein of interest

• Examine details of individual experiments to study the specific interactions and their functional consequences

• Graphically display interaction networks

• Visualise minimal connecting networks for protein sets

• Download data in PSI-MI XML and tabular formats

• Download curated data sets relevant to a specific area of biology e.g. cancer, apoptosis
Interaction data in UniProtKB

- All proteins which participate in an interaction curated in IntAct contain a cross-reference to the relevant IntAct entry

- Subset of high quality interaction data is imported directly into UniProtKB from IntAct

- In addition, UniProtKB curators curate interaction data directly into the IntAct database
## UniProtKB interaction data for BRCA1

### Protein-protein interaction databases

<table>
<thead>
<tr>
<th>Database</th>
<th>Identifier</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIP</td>
<td>DIP-5971N.</td>
<td></td>
</tr>
<tr>
<td>IntAct</td>
<td>P38398</td>
<td>44 interactions.</td>
</tr>
<tr>
<td>MINT</td>
<td>MINT-90433.</td>
<td></td>
</tr>
<tr>
<td>STRING</td>
<td>P38398</td>
<td></td>
</tr>
</tbody>
</table>

### Binary interactions

<table>
<thead>
<tr>
<th>With</th>
<th>Entry</th>
<th>#Exp.</th>
<th>IntAct</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACACA</td>
<td>Q13085</td>
<td>2</td>
<td>EBI-349905,EBI-717681</td>
<td></td>
</tr>
<tr>
<td>BAP1</td>
<td>Q92560</td>
<td>3</td>
<td>EBI-349905,EBI-1791447</td>
<td></td>
</tr>
<tr>
<td>BARD1</td>
<td>Q99728</td>
<td>8</td>
<td>EBI-349905,EBI-473181</td>
<td></td>
</tr>
<tr>
<td>BRAP</td>
<td>Q7Z569</td>
<td>3</td>
<td>EBI-349905,EBI-349900</td>
<td></td>
</tr>
<tr>
<td>BRIP1</td>
<td>Q9BX03</td>
<td>3</td>
<td>EBI-349905,EBI-3509650</td>
<td></td>
</tr>
<tr>
<td>CCND1</td>
<td>P24385</td>
<td>3</td>
<td>EBI-349905,EBI-375001</td>
<td></td>
</tr>
<tr>
<td>CCNE1</td>
<td>P24864</td>
<td>2</td>
<td>EBI-349905,EBI-519526</td>
<td></td>
</tr>
<tr>
<td>CHEK1</td>
<td>Q14757</td>
<td>3</td>
<td>EBI-349905,EBI-974488</td>
<td></td>
</tr>
<tr>
<td>COBRA1</td>
<td>Q8WX92</td>
<td>5</td>
<td>EBI-349905,EBI-347721</td>
<td></td>
</tr>
<tr>
<td>ESR1</td>
<td>P03372</td>
<td>12</td>
<td>EBI-349905,EBI-78473</td>
<td></td>
</tr>
<tr>
<td>FAM175A</td>
<td>Q6UWZ7</td>
<td>10</td>
<td>EBI-349905,EBI-1263451</td>
<td></td>
</tr>
<tr>
<td>KPNA2</td>
<td>P52292</td>
<td>3</td>
<td>EBI-349905,EBI-349938</td>
<td></td>
</tr>
<tr>
<td>PPP1CA</td>
<td>P62136</td>
<td>2</td>
<td>EBI-349905,EBI-357253</td>
<td></td>
</tr>
<tr>
<td>PPP1CB</td>
<td>P62140</td>
<td>3</td>
<td>EBI-349905,EBI-352350</td>
<td></td>
</tr>
<tr>
<td>PPP1CC</td>
<td>P36873</td>
<td>2</td>
<td>EBI-349905,EBI-356283</td>
<td></td>
</tr>
<tr>
<td>RBBP8</td>
<td>Q99708</td>
<td>9</td>
<td>EBI-349905,EBI-1263531</td>
<td></td>
</tr>
<tr>
<td>TRRAP</td>
<td>Q9Y4A5</td>
<td>8</td>
<td>EBI-349905,EBI-399128</td>
<td></td>
</tr>
<tr>
<td>UIMC1</td>
<td>Q96RL1</td>
<td>9</td>
<td>EBI-349905,EBI-725300</td>
<td></td>
</tr>
<tr>
<td>ZCCHC8</td>
<td>Q6NZY4</td>
<td>2</td>
<td>EBI-349905,EBI-1263058</td>
<td></td>
</tr>
<tr>
<td>ZNF350</td>
<td>Q9GZX5</td>
<td>3</td>
<td>EBI-349905,EBI-396421</td>
<td></td>
</tr>
</tbody>
</table>
Pathways

- UniProtKB provides a high level overview of the metabolic pathway(s) in which a protein is involved

- Controlled vocabulary with a structured hierarchy

- Each pathway is split into ‘super-pathway’, ‘pathway’ and/or ‘sub-pathway’. Step number mediated by the protein within the pathway is indicated when known

| Pathway | Carbohydrate degradation, pentose phosphate pathway; D-ribulose 5-phosphate from D-glucose 6-phosphate (oxidative stage): step 2/3. |

- Links to specialised pathway collections also provided

<table>
<thead>
<tr>
<th>Enzyme and pathway databases</th>
<th>tgfbrpathway. TGF-beta receptor signaling.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathway_interaction_DB</td>
<td>REACT_21257. Metabolism of RNA.</td>
</tr>
<tr>
<td>Reactome</td>
<td></td>
</tr>
</tbody>
</table>
Collaboration between EBI, Ontario Institute for Cancer Research, Cold Spring Harbor Laboratory and New York University School of Medicine

- Open-source, open access, pathway database
- Manually curated and peer-reviewed
- Pathways are authored by expert biologists in collaboration with Reactome editorial staff
- Biological pathways are provided in a detailed, computationally accessible format
- Pathways include signaling, innate and acquired immune function, transcriptional regulation, translation, apoptosis and classical intermediary metabolism
What Reactome can be used for

- View expression values overlaid on a pathway
- Compare events in different species
- Link to source databases
- Export pathway to your favourite modelling software
- Interaction overlay on a pathway diagram

- Participating molecules:
  - 14-3-3 protein (cytosol)
  - 3',5'-Cyclic AMP (cytosol)
  - 3-phosphoinositide-dependent protein kinase-1 (plasma membrane)
  - 40S small ribosomal protein S2 (cytosol)
  - 4E-BP (cytosol)
  - 4E-BP1 (cytosol)
  - Activated PI3K (endosome membrane, plasma membrane)
  - ADP (cytosol)
  - ATP (cytosol)
  - ATP (cytosol)
  - ...
High-throughput proteomics

• High-throughput proteomics enables the identification of ever-increasing numbers of proteins

• Requirement for a resource to store protein and peptide identifications including mass spectra, derived peptide identifications and associated metadata

• Information should be stored using controlled vocabularies so that it is human and computer-readable
• Centralized, standards compliant, open source, public data repository for protein and peptide identifications and the evidence supporting these identifications.

• Also captures details of post-translational modifications coordinated relative to the peptides in which they have been found.

• Data is not reprocessed or altered after submission

• Allows data to remain private while anonymously sharing it with journal editors and reviewers

• Recommended submission point for several journals such as *Nature Biotechnology*, *Nature Methods* and *Proteomics*
Summary

• EBI provides a range of proteomics resources

• Each resource is designed to store and handle specific data types and fulfils a unique role

• The resources are highly interlinked with regular data exchange

• Allows for seamless navigation between resources
Contact us

- General queries about any EBI resources: support@ebi.ac.uk

- There is also a feedback form linked from the EBI home page at www.ebi.ac.uk
Resource-specific contacts

- UniProt: help@uniprot.org
- IntAct: intact-help@ebi.ac.uk
- Reactome: help@reactome.org
- PRIDE: pride-support@ebi.ac.uk
Acknowledgements

• All programmers and curators involved in production and maintenance of EBI proteomics resources

• Funders:
  • UniProt: EMBL, NIH, EU, Swiss Federal Government, NSF
  • IntAct: EU
  • Reactome: EMBL, NIH, EU
  • PRIDE: EMBL, BBSRC, Wellcome Trust, EU
Tutorial

• UniProt hands-on tutorial tomorrow 2pm

• Will show how to use the resource in more detail
  • Simple and advanced searching
  • Navigating a UniProtKB entry
  • UniProt tools
  • Complete proteomes
  • Finding documentation