

Gene Ontology annotation workbook

**March 2020**

www.ucl.ac.uk/functional-gene-annotation

A close up of a logo

Description automatically generated

**Tutors: Ruth Lovering r.lovering@ucl.ac.uk**

**Moodle page:** <https://moodle.ucl.ac.uk/enrol/index.php?id=14481>

[What do you know about bioinformatic databases? 3](#_Toc36209838)

[What data do you want to curate? 3](#_Toc36209839)

[National Center for Biotechnology information (NCBI) Gene 3](#_Toc36209840)

[UniProt KnowledgeBase (UniProtKB) 3](#_Toc36209841)

[Gene Ontology 3](#_Toc36209842)

[IntAct 4](#_Toc36209843)

[Reactome 4](#_Toc36209844)

[Key information before you submit an annotation 4](#_Toc36209845)

[UniProt KnowledgeBase (UniProtKB) 4](#_Toc36209846)

[Exercise 1: Searching UniProt using a text search 5](#_Toc36209847)

[Exercise 2: Exploring a UniProt/SwissProt entry: General Information 6](#_Toc36209848)

[Exercise 3: Exploring a UniProt/SwissProt entry: General Annotation 6](#_Toc36209849)

[Exercise 4: Exploring a UniProt/SwissProt entry: Sequence 6](#_Toc36209850)

[Summary 7](#_Toc36209851)

[Finding the right gene or protein identifier (ID) 7](#_Toc36209852)

[Exercise 5: Finding the right ID 7](#_Toc36209853)

[Gene or protein identifier mapping 7](#_Toc36209854)

[Exercise 6: Mapping identifiers in UniProt 7](#_Toc36209855)

[Exercise 7: Mapping identifiers in BioMart 8](#_Toc36209856)

[Gene Ontology 9](#_Toc36209857)

[The QuickGO browser www.ebi.ac.uk/QuickGO 9](#_Toc36209858)

[Exercise 8: Browsing GO terms in QuickGO 9](#_Toc36209859)

[Exercise 9: Finding GO term information from QuickGO 10](#_Toc36209860)

[Exercise 10: Using the basket options in QuickGO 10](#_Toc36209861)

[Exercise 11: Using the filtering options in QuickGO 10](#_Toc36209862)

[Exercise 12: Viewing GO statistics 11](#_Toc36209863)

[Exercise 13: Viewing GO annotations associated with a publication 11](#_Toc36209864)

[Gene Ontology evidence codes 11](#_Toc36209865)

[Finding papers to annotate 11](#_Toc36209866)

[Annotation Examples 12](#_Toc36209867)

[Exercise 14: Annotation 12](#_Toc36209868)

[Annotation example 2, PMID: 18573874 14](#_Toc36209869)

[Annotation example 3, PMID: 9733515 14](#_Toc36209870)

[Ancestor chart view of part of the kinase and phosphorylation ontology 16](#_Toc36209871)

[Definitions of some of the kinase ontology terms listed above 16](#_Toc36209872)

[Steps to GO annotation 16](#_Toc36209873)

[Annotation extensions and Ontology LookUp Service 17](#_Toc36209874)

[Annotate papers 18](#_Toc36209875)

[Exercise Answers 19](#_Toc36209876)

[Appendix 21](#_Toc36209877)

[Further annotation examples 21](#_Toc36209878)

[Annotation example 4, PMID: 7521911 21](#_Toc36209879)

[Annotation example 5, PMID: 26323318 22](#_Toc36209880)

[Annotation example 6, PMID: 22523093 23](#_Toc36209881)

[Further annotation example answers 25](#_Toc36209882)

[Further Resources 26](#_Toc36209883)

[Table 1. A selection of GO evidence codes used for manual annotation 27](#_Toc36209884)

[Table 2. GO evidence code Decision tree for manual annotation 28](#_Toc36209885)

Note answers to exercises are available at the end of the workbook

## What do you know about bioinformatic databases?

Many researchers believe that the data in bioinformatic resources is imported by text mining, ie by computational extraction of data. This is true for many resources however ‘manual’ curation is an important and essential aspect of the majority of the gold standard bioinformatic resources such as those provided by NCBI (eg NCBI Gene), UniProt, Gene Ontology (GO) Consortium, and molecular interaction resources such as IntAct and the IMEx Consortium.

Manual annotation is considered to be more reliable than text mining (which has potentially only 70% accuracy). However, it relies on the biocurator (the person who is curating a paper) having a good understanding of the data presented in the paper.

At UCL we are focused on the curation of human proteins and microRNAs. We have provided almost 10% of all the manual GO annotations for human genes. We curate a wide range of biology, from biochemistry to development.

## What data do you want to curate?

Before you start consider [registering for an ORCID ID](https://support.orcid.org/hc/en-us/categories/360000663074-Register-your-ORCID-iD). This will enable all your annotations to be attributed to you. In addition, many journals request authors provide an ORCID ID.

There are a wide range of bioinformatic resources available, and your expertise will suggest resource you might want to contribute to. In addition, you may want to focus on curating a bioinformatic resource that you use because you are aware that these are missing key information. Below is a list of possible resources to consider contributing to.

### National Center for Biotechnology information (NCBI) Gene

[www.ncbi.nlm.nih.gov/gene](http://www.ncbi.nlm.nih.gov/gene)

Many scientists have used NCBI Gene ‘GeneRIFs’ to link their own paper to a specific gene record. There is a limit to the number of characters you can include in ‘GeneRIFs’ and often people just paste in the title of the paper. This workbook does not provide more information about this.

* You might find ‘My NCBI’ a useful resource to save papers you are interested in curating <http://www.ncbi.nlm.nih.gov/sites/myncbi/>

### UniProt KnowledgeBase (UniProtKB)

[www.uniprot.org](http://www.uniprot.org)

Similar to NCBI Gene you can add a publication to a protein record using the ‘add a publication’ link at the top of each protein page, on the righthand side.

This enables you to highlight a paper that requires annotation and provide some information about what the paper describes so that the paper is forwarded to the appropriate expert biocurator at UniProt. Look at the [submission page](https://community.uniprot.org/bbsub/sampleform.html) to consider if this meets your requirements.

### Gene Ontology

Gene Ontology captures the molecular role of a gene product, as well as the gene product’s location in a cell. How to contribute to the Gene Ontology (GO) will be described in more detail below. The PomBase curation tool, [Canto](https://curation.pombase.org/uniprot/), has recently been modified to enable researchers to curate genes from all species, rather than just Schizosaccharomyces pombe. In addition, some members of the GO Consortium have their own tools to support community curation:

* *Saccharomyces cerevisiae* <https://www.yeastgenome.org/submitData>
* GO Consortium does not have a form to complete but directs people to contribute using a tracking system or email their help desk <http://geneontology.org/docs/contributing-to-go/>
* WormBase <https://wormbase.org/about/userguide/submit_data#201--10>

### IntAct

<https://www.ebi.ac.uk/intact/>

Provides a [submission form and information](https://www.ebi.ac.uk/intact/submission;jsessionid=E0D68F436BAB6ADA4EA2B03273737E91?conversationContext=1) on how to capture molecular interaction data. In addition, they have recently confirmed that they will be running a community curation training session for IntAct sometime next week (1st week of April). You can contact them on

[intact-help@ebi.ac.uk](mailto:intact-help@ebi.ac.uk).

### Reactome

<https://reactome.org/>

Reactome is a pathway database and encourages users to review their draft pathways on [https://reactome.org/community/collaboration](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Freactome.org%2Fcommunity%2Fcollaboration&data=02%7C01%7C%7C65843e5669c2420d369f08d7d13bb9ab%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C637207926771687070&sdata=n5kcAXLN8F%2B%2B2V%2BsdCllntE3FM9gBBpEBfPPBqvUGNw%3D&reserved=0) and to provide your expertise and to [support curation of additional pathways](https://docs.google.com/presentation/d/1RtKgIq5FSHJlABpDZ9TU5COpbG_QrlBCv990s0kPWUM/edit#slide=id.p1). (and data analysis).

## Key information before you submit an annotation

It is important to make sure that you are curating the genes or gene products you want to curate before you start, as many genes have multiple aliases and the name you are using might be the alias of more than one gene in that species. So check the gene/protein record carefully before you start. From now the word ‘gene’ will be use for protein or gene product etc.

The following information is focused on Gene Ontology (GO) annotation using the excel form. To avoid the association of data with the wrong gene our annotations are associated with the UniProt identifier (ID), which is a unique (ID). So, to start you need to learn how to find the identifier for the genes you would like to curate.

Please also consider including the approved gene symbol if the species has one. For example, human gene symbols are approved by the HUGO Gene Nomenclature Committee (HGNC, [www.genenames.org](http://www.genenames.org)).

## UniProt KnowledgeBase (UniProtKB)

[www.uniprot.org](http://www.uniprot.org/)

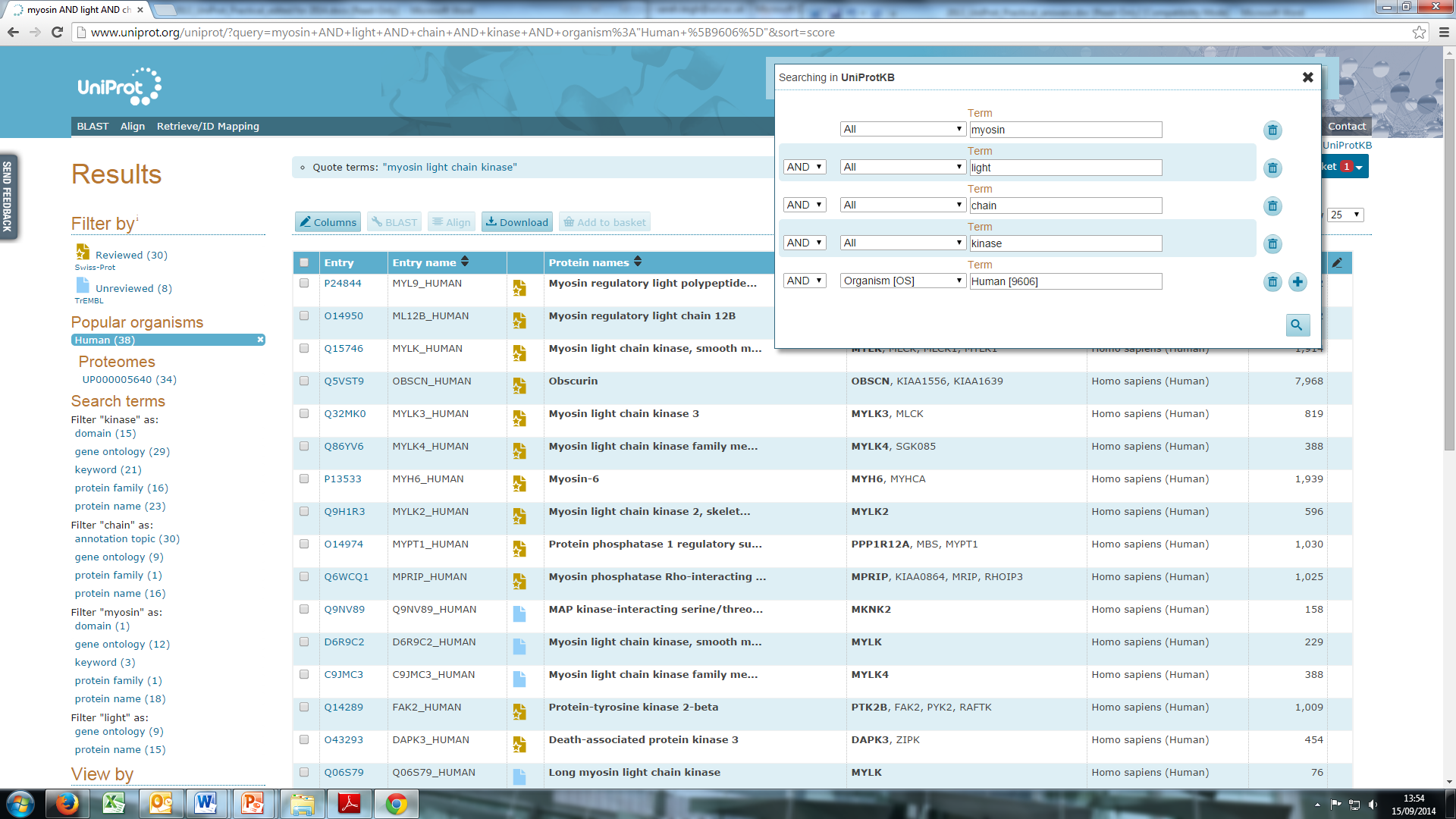
UniProt is the world’s most comprehensive catalogue of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in UniProt/Swiss-Prot, UniProt/TrEMBL and PIR.

This tutorial provides an introduction to the wealth of annotated protein data available within the UniProt database. **The majority of the data investigated below is manually curated from the literature by UniProt.** Follow these instructions to use UniProt to answer the following questions:

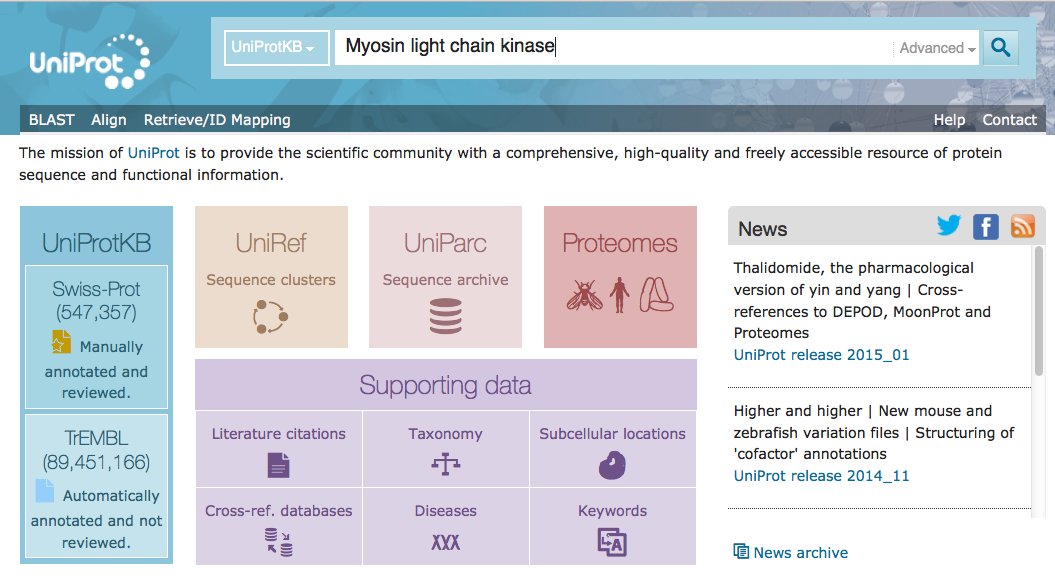
1. *What UniProtKB identifier should be used for the human protein myosin light chain kinase?*
2. *What are the alternative names used to describe myosin light chain kinase?*
3. *What is the role of this protein in the cell?*

### Exercise 1: Searching UniProt using a text search

UniProt can be searched in a number of different ways. The Text Search allows the database to be searched using keywords, similar to how one searches *Google* (logical operators such as “and” and “but not” can be used to restrict search).

1. In the ‘query’ box on the UniProt homepage ([www.uniprot.org](http://www.uniprot.org)), type in “Myosin light chain kinase” (see figure below). Make sure the “Search in” box states “UniProtKB”. Click on the “Search” button .

*Q. How many results did you retrieve?*



* In September 2011, this search retrieved 927 results, and in August 2012 the same search retrieved 1506 results.

*Q. Can you explain why there is a difference in number of results returned?*

1. On the lefthand side are options to filter the results, eg ‘human’ or you can use the other organism field to cut down on the volume of results.

NOTE: the “Query” box at the top of the page now includes the extra search term ‘AND organism:”Homo sapiens (Human) [9606]"’.

1. The results table now has only myosin light chain kinases that occur in humans.

*Q. Now how many human results did you retrieve?*

*Q. Which is the UniProt identifier that you should use to get the most complete information about the human myosin light chain kinase?*

NOTE: UniProt/SwissProt entries have a gold star  and UniProt/TrEMBL entries are a grey-blue . The UniProt/SwissProt gold star  entries have been manually annotated by a curator. Ideally all annotations you submit will be associated with the gold star  entries if there is one for your gene.

NOTE: the search looks for any mention of the symbol you submitted within the protein record. Usually a match with the name or alias fields will be near the top. Many of the returned records are for proteins that interact with the protein you have searched for.

1. Click on the UniProt ID hyperlink for the protein record you have chosen to open. Some proteins in the UniProt database are predicted from analysing the DNA or RNA sequence, while other proteins have direct evidence that they really do exist.
2. Look at the top of the protein record page for the “Status” line.

A screenshot of a computer

Description automatically generated*Q. How does UniProt indicate whether there is evidence that this protein exists at the protein level?*

### Exercise 2: Exploring a UniProt/SwissProt entry: General Information

1. Open the UniProt entry for Q15746.
2. It is useful to add this page to your **basket** (using the blue button ‘Add to basket’ just below the orange protein ID at the top of the page), as you can then open the basket to re-open this entry page during the practical.
3. Click on the “NAMES & TAXONOMY” link (in the left hand blue panel).

*Q. What other names is this protein known by?*

NOTE: the names given in the UniProt/SwissProt entry are specific to a particular species (here, Homo sapiens).

1. Use the back button in your browser to navigate back to the Q15746 entry page.

### Exercise 3: Exploring a UniProt/SwissProt entry: General Annotation

GO is the abbreviation for Gene Ontology. GO terms can be used to get a rough idea of what this protein does. In addition, the GO terms use a hierarchical controlled vocabulary and are very useful for accurately cross-comparing data between databases. The GO annotation data is created by many different annotation teams in Europe and the USA, not just UniProt, but is stored as a single dataset.

1. Click on the ‘FUNCTION’ link (in the blue panel on the left side of the page) and scroll down to the “GO - Molecular function” section.

*Q. What might this protein bind according to the GO - Molecular Function?*

*Q. How many GO - Molecular Functions are listed in this protein record?*

1. More information about GO will be provided later*.*

### Exercise 4: Exploring a UniProt/SwissProt entry: Sequence

1. Navigate to the Sequences section of the Q15746 entry page, by clicking on ‘Sequences’ (in the blue panel on the left side of the page).

NOTE: Q15746 MYLK is a protein for which multiple isoforms have been identified. All of these are mapped within UniProtKB and given stable identifiers. Isoform 1 has been chosen as the 'canonical' sequence, which is why the sequence is shown in full.

1. Scroll down to view the different isoforms (below the canonical sequence).

*Q. How many different isoforms are known for this protein?*

NOTE: **if you wanted to capture information about a specific isoform then you would use a specific isoform ID such as** Q15746-2 but if you do not know which isoform is being described in the paper then you would not include the isoform information, i.e. in this case you would annotate to Q15746.

1. Let’s try aligning the different isoforms. Tick the “Align” button at the top of the “Sequences” section (this will take a few seconds to run). This enables you to see what sequence differences exist between the different isoforms.

NOTE:The sequence, isoform and annotations associated with the protein sequence have all been associated with this protein record by a UniProt curator.

UniProt can also be searched using BLAST (Basic Local Alignment Search Tool), which takes a protein or nucleotide sequence (which it translates) and compares it with those contained in the UniProt database. This facility is available using the BLAST tab in the menu bar at the top of the page.

### Summary

Using UniProt, we can uncover a lot of information about a protein in addition to its sequence, even for UniProt/TrEMBL entries; this includes functional annotation, sequence annotation, structure and isoform information, and much more.

You can return to the UniProt home page by clicking on the UniProt Logo at the top left-hand side of the page. There you will find “Getting started” information and a YouTube link, with a number of videos that you may find useful.

## Finding the right gene or protein identifier (ID)

It is important to know if the identifier you are using corresponds to the gene or protein you are investigating. Some genes may share names or have the same synonyms, this is why unique and stable identifiers are so important. Some tools require an input of a specific identifier type, therefore it is useful to know how to map between identifier types.

### Exercise 5: Finding the right ID

1. For this exercise you can use any gene that you are interested in. If you don’t have a particular gene of interest, these are some examples you can use: *APOB, LIPG* or *TRIB1*.

*Q. How many aliases can you find for your gene in the gene records of:*

* + - HGNC ([www.genenames.org](http://www.genenames.org))?
    - NCBI Gene ([www.ncbi.nlm.nih.gov/gene](http://www.ncbi.nlm.nih.gov/gene))?

*Q. Is there a difference between the two sites?*

*Q. Why might this be?*

1. Search with an alias of your gene at UniProt ([www.uniprot.org](http://www.uniprot.org)).
2. Repeat this search at UniProt with the approved gene symbol.

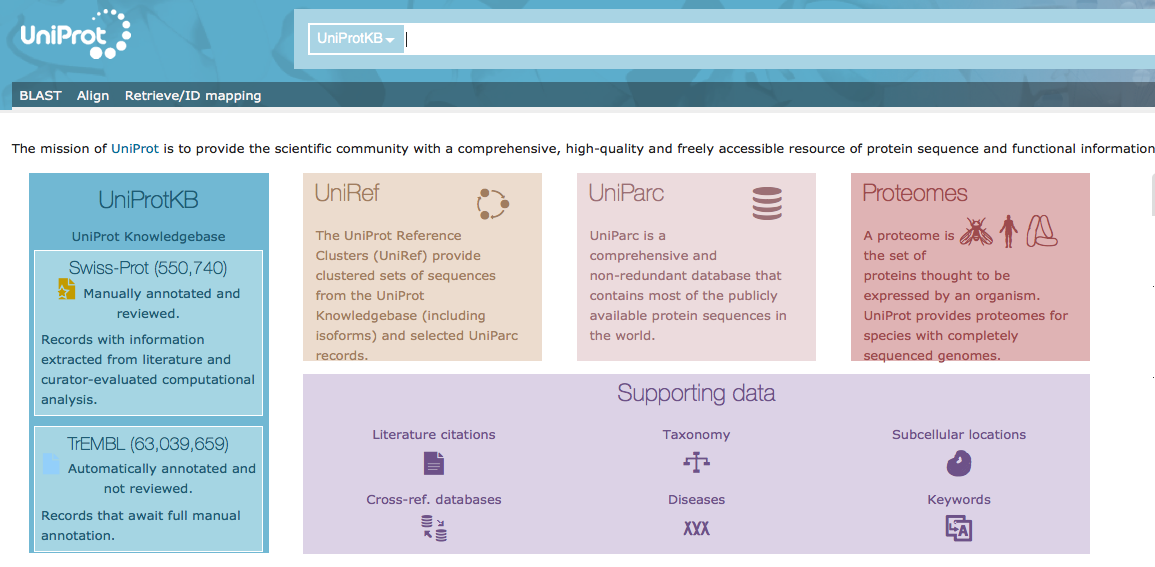
*Q. Which search is easiest for finding the approved symbol or an alias?*

## Gene or protein identifier mapping

### Exercise 6: Mapping identifiers in UniProt

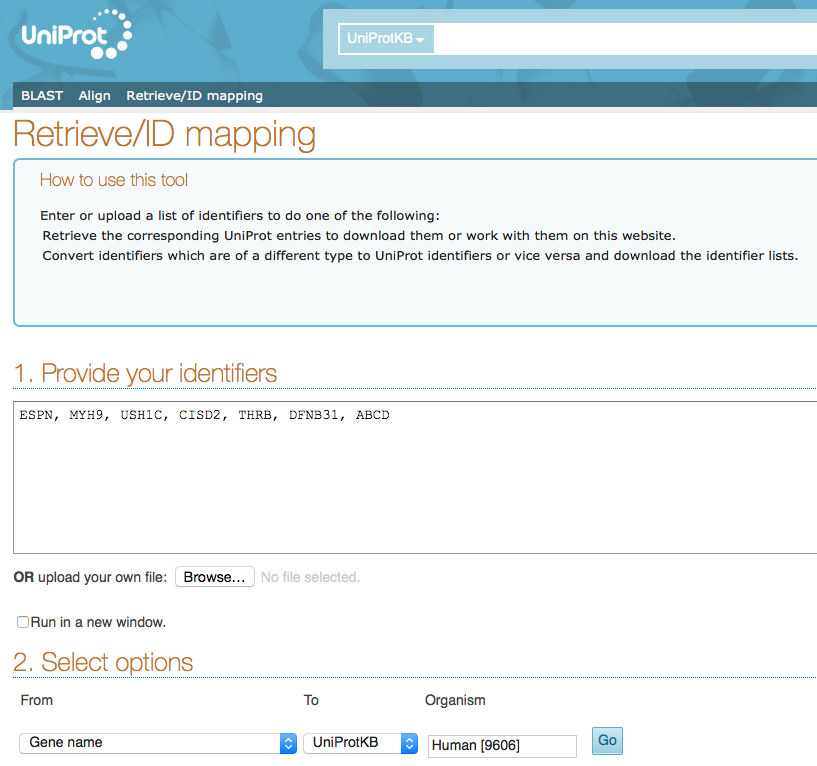
If you have a large list of genes you want to curate then use the UniProt mapping tool to extract these quickly, rather than investigating one by one. The UniProt mapping tool can be used to map various database identifiers to UniProtKB identifiers or to map UniProtKB identifiers to various database identifiers.

1. On the UniProt homepage ([www.uniprot.org](http://www.uniprot.org)), click on the toolbar link “Retrieve/ID mapping”.



1. Paste this list of gene names in the text box:

*ESPN, MYH9, USH1C, CISD2, THRB, DFNB31, ABCD*

1. Select “Gene name” in the ‘**From’** field and “UniProtKB” in the ‘**To’** field and select the ‘Organism’ as “Human [9606]”, click on ‘Submit.

*Q. How many results to you get back?*

*Q. Why do you get more IDs back than you put in?*

1. Click on the “Reviewed” filter to obtain only those entries that have been manually curated.
2. Locate the UniProtKB identifier, protein name and gene name for each entry.

*Q. Why do you think the protein names and gene names differ in some cases?*

*Q. Have any of the gene names not been mapped?*

NOTE: IDs can also be mapped in the other direction, i.e. from UniProtKB to specified database identifiers that are available in UniProt.

1. See if you can modify the selected options to enable you to convert the UniProt identifiers you obtained (or these identifiers Q63369, Q16665, Q07869, Q8NHJ6, P08047, P48729, O94907, P22725, Q9NYF0) to a few different ID types.

### Exercise 7: Mapping identifiers in BioMart

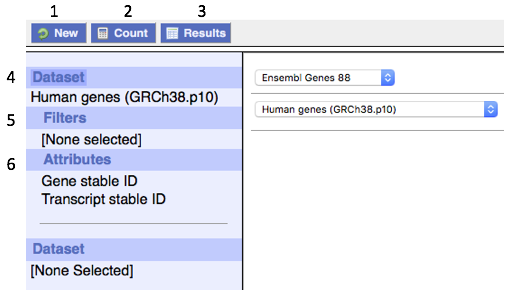
Ensembl BioMart can be used to map between several types of identifier.

Follow this exercise to map the following list of HGNC symbols to NCBI Gene IDs:

*ESPN, MYH9, USH1C, CISD2, THRB, WHRN.*

1. Click on *BioMart* in the top header of a [www.ensembl.org](http://www.ensembl.org) page to go to: [www.ensembl.org/biomart/martview](http://www.ensembl.org/biomart/martview)
2. Select ‘Ensembl Genes 92’ as the primary database (CHOOSE DATABASE)
3. Select ‘Human genes’ as the dataset (CHOOSE DATASET)
4. On the left you will see the control panel, which contains the following:

**Key:** 1. New: allows you to reset the form and start another query

2. Count: gives you a count of the gene identifiers you have added

3. Results: Once you have set up your query criteria and output options, this shows you the results of the query

4. Dataset: Shows you the dataset you have chosen

5. Filters: this is the section where you apply your filters and input your gene/protein list

6. Attributes: this is the section where you choose the output options for your results

NOTE: by default it always outputs Ensembl Gene and Transcript IDs, but these can be switched off)

1. Click ‘Filters’ on the left.
2. Expand the ‘GENE’ panel, by clicking + next to GENE
3. In ‘Input external references ID list’, paste in the gene symbols from above. Change the dropdown menu for this field to read ‘HGNC symbol(s) [e.g. A1BG]’.
4. Click Count to see BioMart is reading 6 genes out of 64561 possible Human genes. Since we entered 6 gene symbols, this confirms that our filters have worked
5. Click on Attributes to select output options
6. Expand the ‘EXTERNAL’ panel (using the + button) and scroll down to select ‘NCBI Gene ID’ and ‘HGNC symbol’. Note that if you don’t select HGNC symbol you will not be able to align (map) the resulting IDs with the symbols you submitted.
7. Click ‘Results’ on the control panel
8. By default only the first 10 rows are displayed initially, you can increase this using the drop-down box under “View” or download all results by clicking on “Go”.

*Q. What is the NCBI Gene ID for each of the HGNC IDs?*

*Q. Why are there multiple rows for one gene ID?*

## Gene Ontology

Gene Ontology (GO) is a dynamic controlled vocabulary that describes gene products’ *Molecular Functions*, *Biological Processes* which they are involved in, and *Cellular Components* to which they localise.

In order to curate a paper you need to ideally provide:

1. the UniProt ID for the protein (finding this is covered above)
2. the GO term
3. an evidence code describing the evidence supporting the annotation
4. the PubMed ID for the paper (PMID:number)

There are 45,000 GO terms to choose from so understanding how to find suitable GO terms is the next step

## The QuickGO browser www.ebi.ac.uk/QuickGO

The QuickGO browser has been developed by the European Bioinformatics Institute (EBI) to browse the GO hierarchy and view annotations for individual gene products. The QuickGO home page provides a text box to start searching for GO information. You may search for any aspect of a GO annotation including; GO term names and synonyms, GO IDs, UniProtKB accessions, or UniProtKB keywords. This practical session uses the GO browser QuickGO, will describe how to find GO terms and how to view protein annotations.

### Exercise 8: Browsing GO terms in QuickGO

1. Open QuickGO in a new window at [**www.ebi.ac.uk/QuickGO**](http://www.ebi.ac.uk/QuickGO)
2. Search QuickGO by entering into the text box a cellular component name, such as ‘nucleus’.

QuickGO will return any relevant GO terms and proteins (“gene products”) associated with the word ‘nucleus’. The first 15 GO terms and proteins are shown by default in the top and bottom search results tables, respectively.

1. To see further terms and/or proteins, click on ‘Show all…’ at the bottom of the list.
2. Use the options in the left-hand side menu to view terms from a particular aspect of the GO, i.e. *Molecular Function*, *Biological Process* or *Cellular Component.*

NOTE: Some terms are retrieved due to information in their synonym or definition fields.

Click on the GO ID for the term will take you to a page providing full details of the selected term. Items in the right-hand side menu provide further information about the term such as Synonyms, Ancestor Chart, or Child Terms. Protein Annotations associated with the term can be viewed by clicking on the number or annotations on the GO term entry page, underneath the GO term definition.

### Exercise 9: Finding GO term information from QuickGO

*Q. What is the GO ID for the term “mitochondrial chromosome”?*

*Q. What are the four direct parents of the term “nucleosome”?*

*Q. What types of relationships are there between the term “nucleosome” and its direct parents?*

*Q. What are the child terms for the term “ensheathment of neurons”?*

### Exercise 10: Using the basket options in QuickGO

1. Search QuickGO by typing the biological process term “endocytosis*”* into the text box.
2. Click on the endocytosis term ID to open this term record.
3. Click on the child term menu item and click ‘show all’ to display all of the child terms.
4. Select any 3-5 of the child terms, using the basket icon buttons (one of the 3 icons next to the child term GO term IDs). Try to select terms which have different relationships to the parent term. (A green tick will appear over the basket icon, when a term has been selected; a red cross will be visible, while a term is being unselected, and then it will immediately disappear, leaving only the basket icon being displayed).
5. To visualise the relationships between the selected child terms and their parent terms, click on the ‘Basket’ in the menu bar at the top of the page. The terms in your basket will be displayed in a table; click on the basket icon underneath the table to display the graph.
6. Your selected terms are highlighted in yellow within the graph. This is a useful tool for checking relationships between terms.

### Exercise 11: Using the filtering options in QuickGO

Filtering allows you to manipulate the dataset according to the attributes you are interested in. Use the filtering tabs in the bar above the table with annotations.

1. Click on annotations (under the GO term definition) in the “endocytosis” QuickGO entry.

*Q. How many ‘annotations’ are listed?*

1. Click on ‘taxon’ and select ‘*Homo sapiens*’; click ‘apply’.

*Q. How many annotations are listed now?*

1. Click on the ‘evidence’ and select ‘evidence used in manual assertion’; click ‘apply’.

*Q. How many annotations are listed now?*

1. Change the filter to ‘experimental evidence used in manual assertion’; click ‘apply’.

*Q. How many annotations are listed now?*

1. Click on ‘GO terms’, and on ‘Options’ at the bottom of the drop-down list. Select the bottom option bullet: ‘is\_a, part\_of, occurs\_in, regulates’ to include annotations to terms associated with endocytosis via regulation relationships; click ‘apply’.

*Q. How many annotations are listed now?*

### Exercise 12: Viewing GO statistics

1. While viewing the filtered “endocytosis” GO annotation record, click on the ‘Statistics’ tab.
2. Select each of the different tabs in the pop-up window:

* Summary: displays the number of annotations to distinct gene products (usually proteins) associated with this term and its child terms.
* GO ID: displays the number of annotations per GO ID, and the number of gene products (usually proteins) associated with each individual GO term in this area of GO.

1. Explore the results in each of the other tabs.
2. Switch back to the ‘Annotations’ tab.

### Exercise 13: Viewing GO annotations associated with a publication

1. Click on ‘View GO Annotations’ on QuickGO homepage.
2. Click on the ‘References’ tab above the Annotations table.
3. Type in e.g. ‘PMID:15919722’ *without any spaces.* Click ‘Add’; click ‘apply’.

## Gene Ontology evidence codes

[www.geneontology.org/GO.evidence.shtml](http://www.geneontology.org/GO.evidence.shtml)

All GO annotations have an associated evidence code which indicates the category of the evidence that was used to make the annotation. There are currently six evidence codes that are used to categorise experimental data **(see Table 1 in the Appendix)**. There are ten evidence codes for computational evidenced annotations and five evidence codes for non-experimentally evidenced annotations. For more information about guidelines for applying the evidence codes see the **GO evidence code decision tree in the Appendix**.

## Finding papers to annotate

1. The paper should have evidence to support at least one of the following
   * molecular function of the gene
   * the role of the gene in a biological process
   * the location of the gene in a cell
2. A GWAS paper or a paper listing co-expression of multiple genes are unlikely to be suitable for annotation.
3. If you are not an author on the paper it is very important to check that it is clear what species the gene you want to curate is associated with. For example, if the paper describes a cDNA construct then the species the cDNA was derived from needs to be clear. It usually does not matter what the species is that the construct is transfected into. Although if you are capturing a protein interaction you will need to consider the species of the transfected cDNA and the bound endogenous protein. These may not be the same species. If you are an author of the paper and this information is not in the paper then you need to clarify the species you have used in your submitted annotation as some additional free text.
4. If you are not sure about the suitability of the paper send it to Ruth r.lovering@ucl.ac.uk

## Annotation Examples

### Exercise 14: Annotation

1. Read through the worked annotation example 1.
2. Work through annotation examples 2-3: The answers can be found at the back of the workbook.
3. If you have time work through examples 4-6 in the Appendix.

Worked example 1, PMID: 19177149

Proteins: O94761 (RECQ4\_HUMAN)

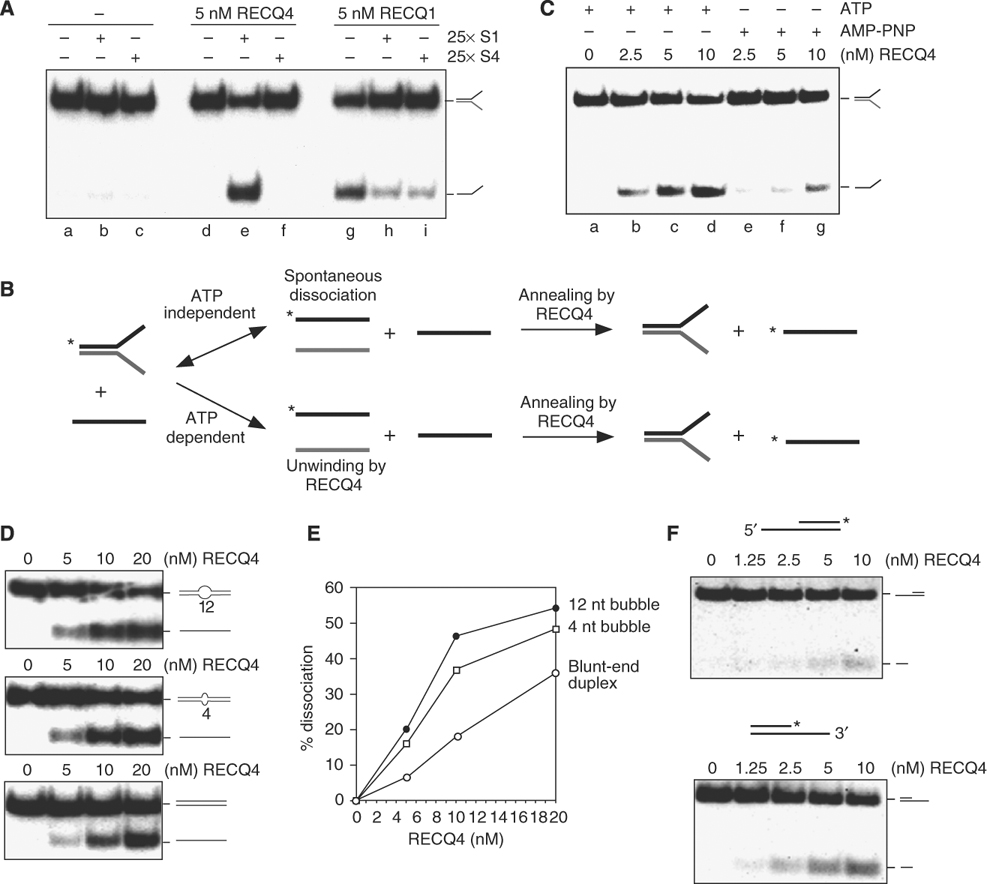
P46063 (RECQ1\_HUMAN)

...we determined whether DNA unwinding by RECQ4 was an ATP-dependent reaction. Helicase assays were carried out in the presence of ATP, or the non-hydrolysable ATP analogue AMP-PNP, leading us to observe a significantly higher level of dissociation in the presence of ATP compared with the reactions containing AMP-PNP ([Figure 2C](http://www.nature.com/emboj/journal/v28/n5/full/emboj200913a.html#f2)). This observation shows that the majority of the dissociated ssDNA products result from ATP-dependent DNA unwinding driven by the RECQ4 helicase ([Figure 2B](http://www.nature.com/emboj/journal/v28/n5/full/emboj200913a.html#f2), lower panel). Therefore, RECQ4 is an active ATP-dependent DNA helicase...

...To determine the polarity of the helicase activity of RECQ4, we analysed its activity on 3'- and 5'-overhang substrates. As RECQ4 is capable of unwinding blunt-ended duplex DNA ([Figure 2D](http://www.nature.com/emboj/journal/v28/n5/full/emboj200913a.html#f2)), reactions were carried out for relatively short times (15 min), as prolonged incubation results in the unwinding of both substrates, as observed previously with *E. coli* RecQ protein ([Umezu](http://www.nature.com/emboj/journal/v28/n5/full/emboj200913a.html" \l "B21) *[et al](http://www.nature.com/emboj/journal/v28/n5/full/emboj200913a.html" \l "B21)*[, 1990](http://www.nature.com/emboj/journal/v28/n5/full/emboj200913a.html" \l "B21)). As expected, although DNA unwinding was observed with both substrates, RECQ4 showed a preference towards the 3'-overhang structure ([Figure 2F](http://www.nature.com/emboj/journal/v28/n5/full/emboj200913a.html#f2), lower panel) compared with the 5'-overhang substrate ([Figure 2F](http://www.nature.com/emboj/journal/v28/n5/full/emboj200913a.html#f2), upper panel). These results show that RECQ4 exhibits a 3' right arrow5' polarity. **See figure below**

|  |  |  |  |
| --- | --- | --- | --- |
| Protein ID | GO ID | GO Term Name | Evidence code |
| RECQ4/O94761 | GO:0003678\* | DNA helicase activity | IDA |
| RECQ4/O94761 | GO:0004003\* | ATP-dependent DNA helicase activity | IDA |
| RECQ4/O94761 | GO:0043140 | ATP-dependent 3'-5' DNA helicase activity | IDA |
| RECQ4/O94761 | GO:0032508 | DNA duplex unwinding | IDA |

\* Parent of a more specific child term that can be also annotated to from this evidence; so will not be added as an annotation.

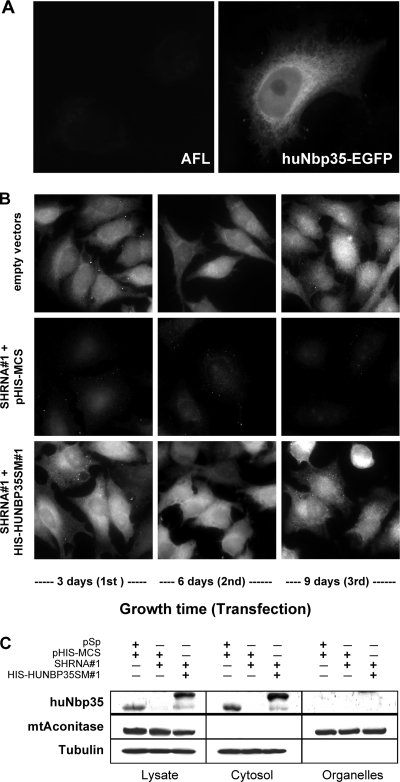
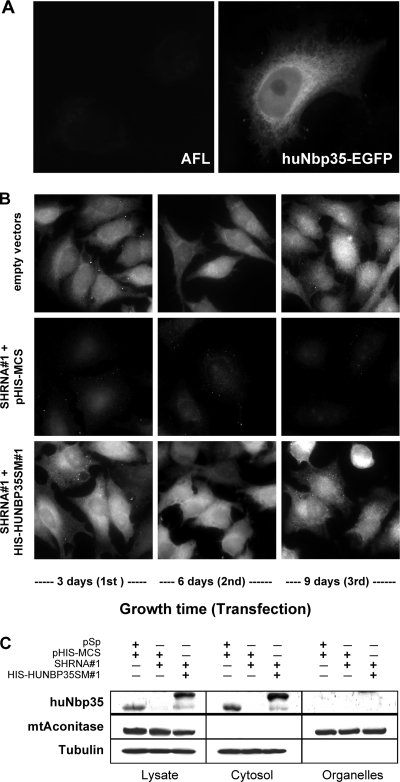


**Figure 2 Legend:** **ATP-dependent DNA unwinding activity of human RECQ4.** (**A**) Comparison of the helicase activities of the recombinant RECQ4 and RECQ1 in the presence of unlabelled S1 or a control oligo, S4, with different sequence composition. (**B**) Schematic diagram of the production of 32P-labelled single-stranded S1 in the DNA dissociation reaction. 32P-end labels are indicated with asterisks. In the ATP-independent event (upper), low level of spontaneous DNA dissociation of splayed arm allows the exchange of 32P-labelled single-stranded S1 with the unlabelled S1, a reaction facilitated by the DNA annealing activity of RECQ4. In the ATP-dependent reaction (lower), DNA dissociation is catalysed by RECQ4 unwinding of splayed arm upon ATP hydrolysis to generate 32P-labelled single-stranded S1, which is then stabilized by the presence of unlabelled S1 to compete for re-annealing back to S2. (**C**) Helicase activities of the recombinant RECQ4 using splay-arm substrates in the presence of ATP or AMP-PNP. (**D**) Helicase assays of the recombinant RECQ4 proteins were carried out as described using duplex DNA with 12-nt bubble (upper panel), duplex DNA with 4-nt bubble (middle panel) and blunt-ended duplex DNA (lower panel). (**E**) Product formation in (D) was quantified by phosphorimaging. Close circle, dissociation product using duplex DNA with 12-nt bubble. Open rectangle, dissociation product using duplex DNA with 4-nt bubble. Open circle, dissociation product using blunt-ended duplex DNA. (**F**) Helicase activities of the recombinant RECQ4 using either a 32P-labelled 5' overhang (upper panel) or 3' single stranded overhang (lower panel).

### Annotation example 2, PMID: 18573874

Protein: P53384 (NUBP1\_HUMAN, alias HuNbp35)

Endogenous huNbp35 was visualized in HeLa cells by indirect immunofluorescence microscopy using an affinity-purified polyclonal rabbit anti-huNbp35 antibody. The fluorescent stain was distributed throughout the entire cell, a finding consistent with a cytosolic localization of the protein (Fig. 2A and C).

**See figure below**

**Figure 2 Legend**: **Human Nbp35 is a cytosolic protein.**

(A) EGFP fluorescence of a HeLa cell transiently transfected with a vector encoding a huNbp35-EGFP fusion protein (right) in comparison to the endogenous autofluorescence (AFL) of control cells (left).

(C) Subcellular localization of huNbp35 by cell fractionation. HeLa cells were transfected with the vectors indicated (see panel B). Six days after the first transfection cells were permeabilized with digitonin, and the cell lysate was centrifuged for 10 min at 15,000 × *g*. Supernatant (cytosol) and pellet (organelles) fractions were analyzed by immunoblotting. HuNbp35 exclusively colocalizes with tubulin in the cytosolic fraction, but not with mitochondrial aconitase (mtAconitase) present in the membrane fraction. pSp, empty vector; pHIS-MCS, empty vector; SHRNA#1, knockdown Nbp35; HIS-HUNBP35SM#1, transfection His-tagged Nbp35.

|  |  |  |  |
| --- | --- | --- | --- |
| Protein ID | GO ID | GO Term Name | Evidence code |
| NUBP1/P53384 |  |  |  |

### Annotation example 3, PMID: 9733515

ATM: approved symbol and name: ATM, ataxia telangiectasia mutated

p53: approved symbol and name: TP53, tumor protein p53 gene

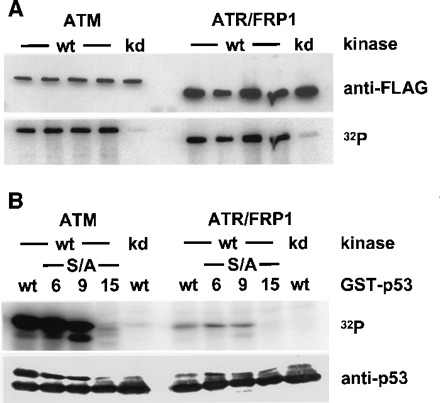
ATR/FRP-1: approved symbol and name: ATR, ataxia telangiectasia and Rad3 related

293T/17 cells, human embryonic kidney cells

IR: ionizing radiation

|  |
| --- |
| The gene mutated in Ataxia telangiectasia (A-T), *ATM* (ataxia telangiectasia–mutated), encodes a 370-kD protein that is a member of a family of proteins related to phosphatidylinositol 3-kinase (PI-3-K) that have either lipid kinase or protein kinase activity. Cell lines derived from A-T patients exhibit hypersensitivity to IR. In response to DNA damage, cells with wild-type ATM accumulate p53 protein and show a subsequent increase in p53 activity, whereas cells with defective ATM show a smaller increase in the amount of p53 protein in response to IR ([4](http://www.sciencemag.org.libproxy.ucl.ac.uk/content/281/5383/1677.full#ref-4),[6](http://www.sciencemag.org.libproxy.ucl.ac.uk/content/281/5383/1677.full#ref-6)). Therefore, ATM appears to act upstream of p53 in a signal transduction pathway initiated by IR. |

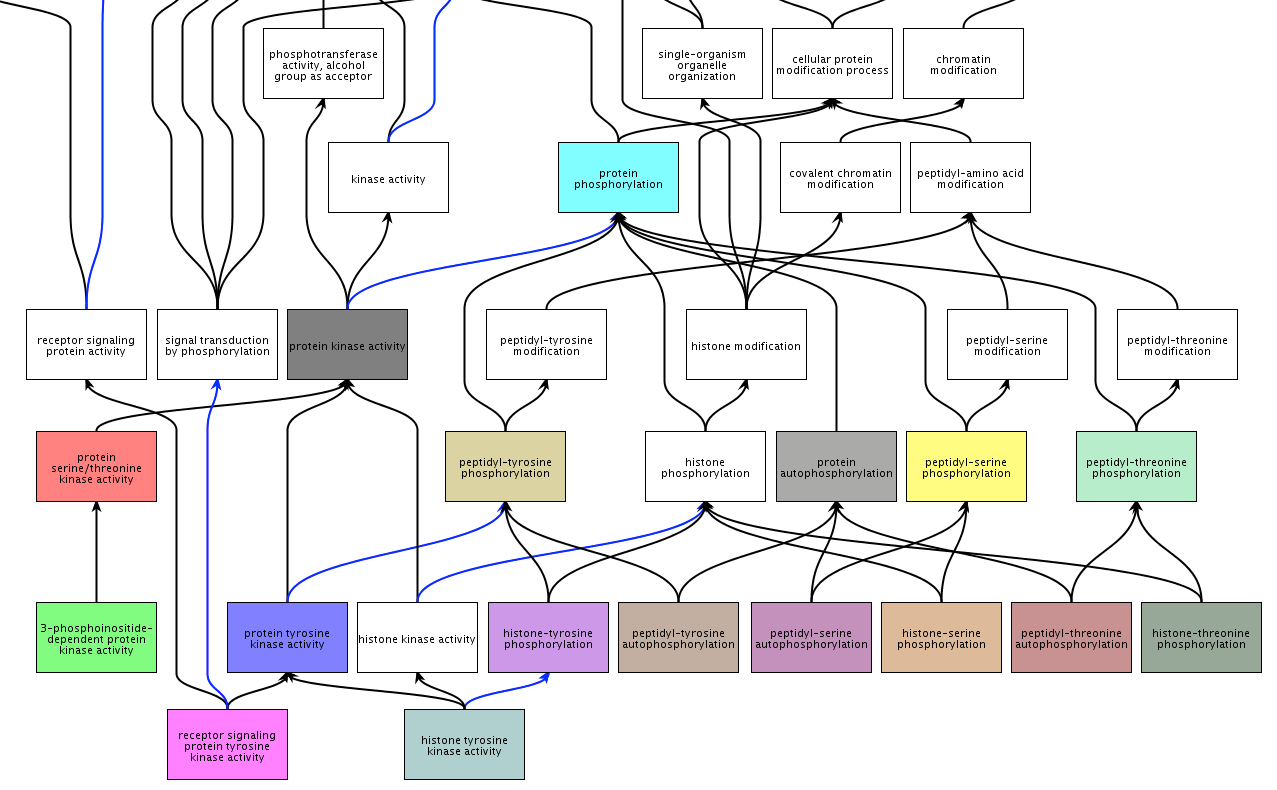
….a recombinant, FLAG peptide–tagged, wild-type human ATM was used as a source of ATM protein, and a FLAG peptide–tagged, mutant ATM expression construct was generated in which two of the three critical amino acid residues required for catalysis were mutated (Asp2870 → Ala and Asn2875 → Lys) (13). Wild-type and mutant recombinant ATM proteins were individually expressed in 293T cells, and in vitro kinase activity was assessed (14). Equivalent amounts of wild-type (wt) and mutant (kd) ATM recombinant proteins were immunoprecipitated and incubated with [γ-32P]adenosine triphosphate (ATP) and recombinant glutathione S-transferase (GST)–conjugated human p53 protein containing the first 101 amino acids of p53 (GSTp531–101) (Fig 1A). Only the wild-type enzyme phosphorylated GSTp531–101 (Fig 1B). Furthermore, wild-type recombinant ATM phosphorylated wt p53, Ser6 →Ala (S6A), and S9A mutant p53, but not S15A mutant p53 protein (Fig 1B). Wild-type ATM kinase also showed autophosphorylation in this assay (Fig 1A).

FLAG-tagged recombinant wt human ATR/FRP-1 also showed autophosphorylation in vitro that was dependent upon the integrity of the catalytic domain. Like ATM, ATR/FRP-1 also phosphorylated p53 on Ser15 in a Mn2+-dependent manner (Fig 1B), though ATR/FRP-1 had at least 20-fold less activity than ATM toward GSTp531–101 when assayed under identical experimental conditions (Fig 1B). Thus, p53 appears to be a better substrate for ATM than ATR/FRP-1.

**Figure 1 Legend: Phosphorylation of Ser15 of p53 by ATM and ATR/FRP1 in vitro**. We transfected 293T/17 cells with expression vectors encoding FLAG-tagged wild-type (wt) or catalytically inactive (kd) ATM or ATR/FRP1. After 48 hours, ATM or ATR was immunoprecipitated with antibody to FLAG and used in an in vitro kinase assay with [γ-32P]ATP and either wt, S6A, S9A, or S15A GSTp531–101 as substrates (14). Proteins from each reaction were separated by SDS-PAGE (7% gel), transferred to nitrocellulose, and analyzed either on a PhosphorImager or by immunoblotting. (**A**) Amounts of FLAG-tagged ATM or ATR in each kinase reaction as measured by immunoblotting with anti-FLAG M2 (top panel) and amount of [γ-32P]phosphate incorporated into ATM or ATR during the reaction (lower panel). (**B**) In vitro kinase assay with wt GSTp531–101 or various mutant GSTp531–101 proteins (S6A, S9A, or S15A) as substrates (top panel). Levels of substrate protein present in each reaction were determined by immunoblotting for p53 (lower panel). The upper immunoreactive band represents phosphorylated GSTp53 fusion protein. ATM did not phosphorylate GST alone. The same exposures are shown for ATM, ATR/FRP1, and corresponding substrate proteins in all panels shown in (A) and (B).

|  |  |  |  |
| --- | --- | --- | --- |
| Protein ID | GO ID | GO Term Name | Evidence code |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

### Ancestor chart view of part of the kinase and phosphorylation ontology



**Function Ontology**

**8**

**7**

**5**

**4**

**3**

**2**

**Process Ontology**

**6**

**1**

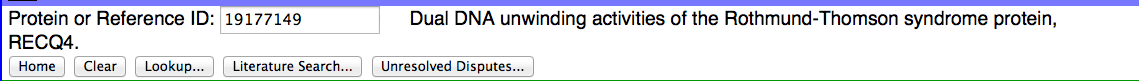
### Definitions of some of the kinase ontology terms listed above

|  |  |  |
| --- | --- | --- |
| 1 | kinase activity | Catalysis of the transfer of a phosphate group, usually from ATP, to a substrate molecule. |
| 2 | protein kinase activity | Catalysis of the phosphorylation of an amino acid residue in a protein, usually according to the reaction: a protein + ATP = a phosphoprotein + ADP. |
| 3 | protein serine/threonine kinase activity | Catalysis of the reactions: ATP + protein serine = ADP + protein serine phosphate, and ATP + protein threonine = ADP + protein threonine phosphate. |
| 4 | 3-phosphoinositide-dependent protein kinase activity | Catalysis of the reaction: ATP + a protein = ADP + a phosphoprotein. This reaction requires the presence of a phosphatidylinositol-3-phosphate. |
| 5 | protein tyrosine kinase activity | Catalysis of the reaction: ATP + a protein tyrosine = ADP + protein tyrosine phosphate. |
| 6 | protein phosphorylation | The process of introducing a phosphate group on to a protein. |
| 7 | peptidyl-serine phosphorylation | The phosphorylation of peptidyl-serine to form peptidyl-O-phospho-L-serine. |
| 8 | histone-serine phosphorylation | The modification of histones by addition of a phosphate group to a serine residue. |

## Steps to GO annotation

The information below provides a brief summary of the information described in the “Steps to GO annotation” presentation.

1. Select experimental paper for annotation, often based on abstract
   1. Check that the PMID has not already been annotated using QuickGO ([www.ebi.ac.uk/QuickGO](http://www.ebi.ac.uk/QuickGO))
   2. If it has already been annotated go to another paper
   3. If it hasn’t been annotated, see if you can annotate it
   4. Add the PMID number, eg: 24704852, to the spreadsheet



1. Read method section of the paper to confirm you can identify species of gene/protein used for the experiment
   1. If not here, check rest of paper for this information (even supplemental data).
   2. If you can find this information then copy and paste the text into the top section ‘Key species information’.
   3. Add the species, (approved gene symbol, eg HGNC symbol) and alias used in paper (if different from the HGNC symbol) to the spreadsheet in the ‘gene information section’.
   4. Find the protein accession ID(s) for the protein(s) you want to annotate, using UniProtKB, these will be 6-10 alphanumeric IDs e.g. P12345. Remember to look for the reviewed IDs, these have a gold star
      1. DO NOT use QuickGO to find the protein accession ID
      2. If there is not a gold star record then choose the ID that represents the longest protein, or represents the isoform you are curating
   5. Add the protein accession ID to the ‘gene information section’.
2. Read results section of the paper and identify GO terms supported by experimental evidence
   1. Use QuickGO to find the most appropriate GO term
      1. check the ancestor chart
      2. check the child terms
      3. read the definition
   2. Copy and paste the GO ID to relevant annotation section, along with the GO term name.

## Annotation extensions and Ontology LookUp Service

* 1. If the results provide additional information, which would add value to the annotation, e.g. cell type, tissue type, or a regulation target, this can be included in the annotation extension (AE) field.

NOTE: the AE relates the primary GO term in the annotation.

* + 1. In order to find the correct identifier for the cell, or tissue type, use the Ontology Lookup Service (<http://www.ebi.ac.uk/ols/index>). This service allows you to search for terms from any biological ontology including GO.
    2. Or just paste the information into this field and the checker will find the relevant IDs to include.

1. Add appropriate evidence code
   1. Use evidence code decision tree, mostly you will just use IDA, IMP or IPI
   2. Use IDA or IMP for experimental evidence **OR**
   3. Use IPI for protein-protein interactions, remember to add the binding partner protein accession ID to the ‘with’ column
2. So that your annotations can be reviewed efficiently, add a supporting statement to the comments field of the annotation row
   1. Add the figure number (there maybe more than one figure that supports your annotation)
   2. Write your supporting statement in the free text field,
3. Optional extra – read paper introduction
   1. Are there any of statements supported by the results section, but not captured by your annotations? E.g. toxin metabolism supported by experiment confirming thiourea metabolism
   2. If there are add these annotations.
4. Optional extra – read paper introduction and discussion
   1. Are there any statements which you consider to be general knowledge which are not already associated with the protein record in QuickGO? E.g. toxin metabolism. If these statements are not supported by the results section, but should be associated with the protein record then create the annotation and use either the TAS or NAS evidence codes:
      1. If you use the evidence TAS (Traceable Author Statement) then include the reference number supporting the statement (eg REF. 1) and paste the sentence into the ‘summary of data’ section
      2. If you use the evidence NAS (Non-traceable Author Statement) not captured by your annotations? paste the sentence into the ‘summary of data’ section.
   2. If there are potential GO terms that are already in the database, then don't worry about them.

## Annotate papers

Using the information gained in this workshop, look at the papers you would like to curate and suggest annotations that can be created based on the experimental data presented.

1. All your annotations will be held in the private database
2. We will review/edit all of your annotations before we release them into the public database

NOTE: Information about evidence codes at the back of the workbook

## Exercise Answers

(Answers current as of April 2019)

**Exercise 1: Searching UniProt using a text search**

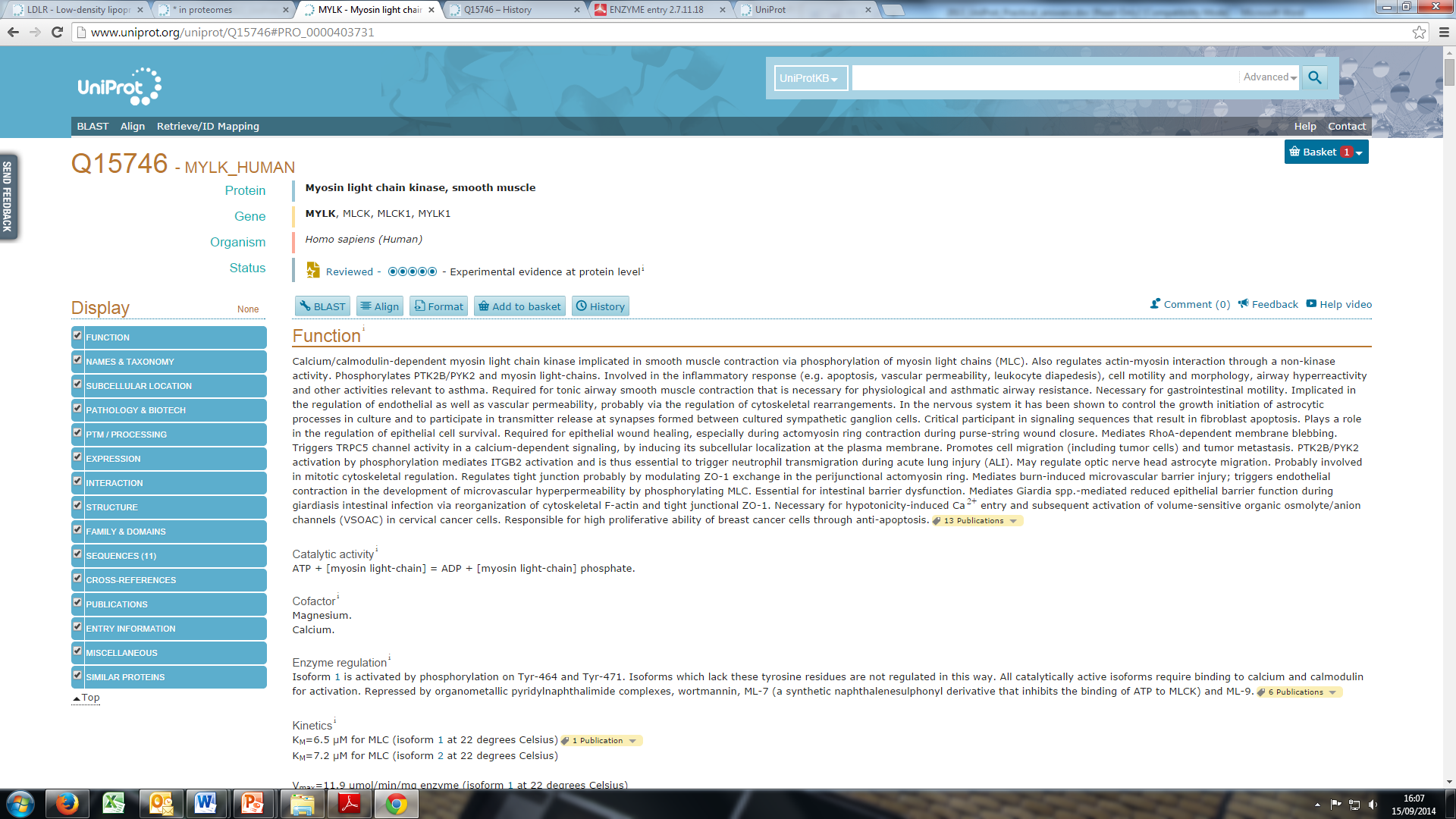
*Q. How many results did you retrieve?* A: 2306

*Q. Can you explain why there is a difference in number of results returned?* A: The data is constantly being updated, and 579 new protein sequences fitting these search parameters were added to UniProtKB between September 2011 and August 2012. In addition, duplicate entries have been removed/merged.

*Q. Now how many results did you retrieve?* A: 45

*Q. Which is the identifier that you should use to get the most complete information about the human myosin light chain kinase?* A: Q15746

*Q. How does UniProt indicate whether there is evidence that this protein exists at the protein level?* A: Denoted by the annotation score:

******

**Exercise 2: Exploring a UniProt/SwissProt entry: General Information**

*Q. What other names is this protein known by?* A: Kinase-related protein (KRP) & Telokin

**Exercise 3: Exploring a UniProt/SwissProt entry: General Annotation**

*Q. What might this protein bind according to the GO - Molecular function list?* A: Actin, calmodulin, ATP and metal ion

*Q. How many GO - Molecular functions are listed in this protein record?* A: 5

**Exercise 4: Exploring a UniProt/SwissProt entry: Sequence**

*Q. How many different isoforms are known for this protein?* A: 11

**Exercise 7: Finding the right ID**

*Q. Is there a difference between the two sites?* A: Different databases get updated at different frequencies and so may not be completely synchronised.

*Q. Why might this be?* Both of these databases are manually curated, leading to slightly different content.

**Exercise 8: Mapping identifiers in UniProt**

*Q. How many results to you get back?* A: 44.

*Q. Why do you get more IDs back than you put in?* Many of the results are unreviewed TrEMBL entries.

*Q. Why do you think the protein names and gene names differ in some cases?* A: The gene name is the officially approved symbol from HGNC. The protein name is assigned by UniProt and, while it often matches the approved symbol, it could be different.

*Q. Have any of the gene names not been mapped?* A: Yes, 1. You can see which one by clicking on the link in the top left of the page.

**Exercise 9: Mapping identifiers in BioMart**

*Q. What is the NCBI Gene ID for each of the HGNC IDs?*

A: HGNC symbol NCBI Gene ID

*ESPN* 83715

*MYH9* 4627

*USH1C* 10083

*CISD2* 493856

*THRB* 7068

*WHRN* 25861

*Q. Why are there multiple rows for one gene ID?* A: The Ensembl gene and transcript IDs are automatically included in the results. If you don’t want these you need to deselect them in the “Attributes” -> “Gene” section.

**Exercise 11: Finding GO term information from QuickGO**

*Q. What is the GO ID for the term “mitochondrial chromosome”?* A: GO:0000262

*Q. What are the four direct parents of the term “nucleosome”?* A: chromatin, chromosomal part, DNA packaging complex, protein-DNA complex

*Q. What types of relationships are there between the term “nucleosome” and its direct parents?* A: Part\_of for chromatin and is\_a for the other three

*Q. What are the child terms for the term “ensheathment of neurons”?* A: axon ensheathment, ensheathment of neuronal cell bodies

**Exercise 13: Using the filtering options in QuickGO**

*Q. How many ‘annotations’ are listed? (Endocytosis, All)* A: 652,464

*Q. How many annotations are listed now? (Endocytosis, Human)* A: 2,014

*Q. How many annotations are listed now? (Endocytosis, Human, Manual All)* A: 856

*Q. How many annotations are listed now? (Endocytosis, Human, Manual Experimental)* A: 206

*Q. How many annotations are listed now? (Endocytosis and regulation terms, Human, Manual Experimental)* A: 348

**Exercise 16: Annotation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Example | Protein | GO ID | GO Term | Evidence |
| 2 | NUBP1 | GO:0005829 | cytosol | IDA |
| 3 | ATM | GO:0018105 | peptidyl-serine phosphorylation | IDA |
|  | ATM | GO:0046777 | autophosphorylation | IDA |
|  | ATM | GO:0004674 | protein kinase activity | IDA |
|  | ATR | GO:0018105 | peptidyl-serine phosphorylation | IDA |
|  | ATR | GO:0046777 | autophosphorylation | IDA |
|  | ATR | GO:0004674 | protein kinase activity | IDA |

## Appendix

### Further annotation examples

### Annotation example 4, PMID: 7521911

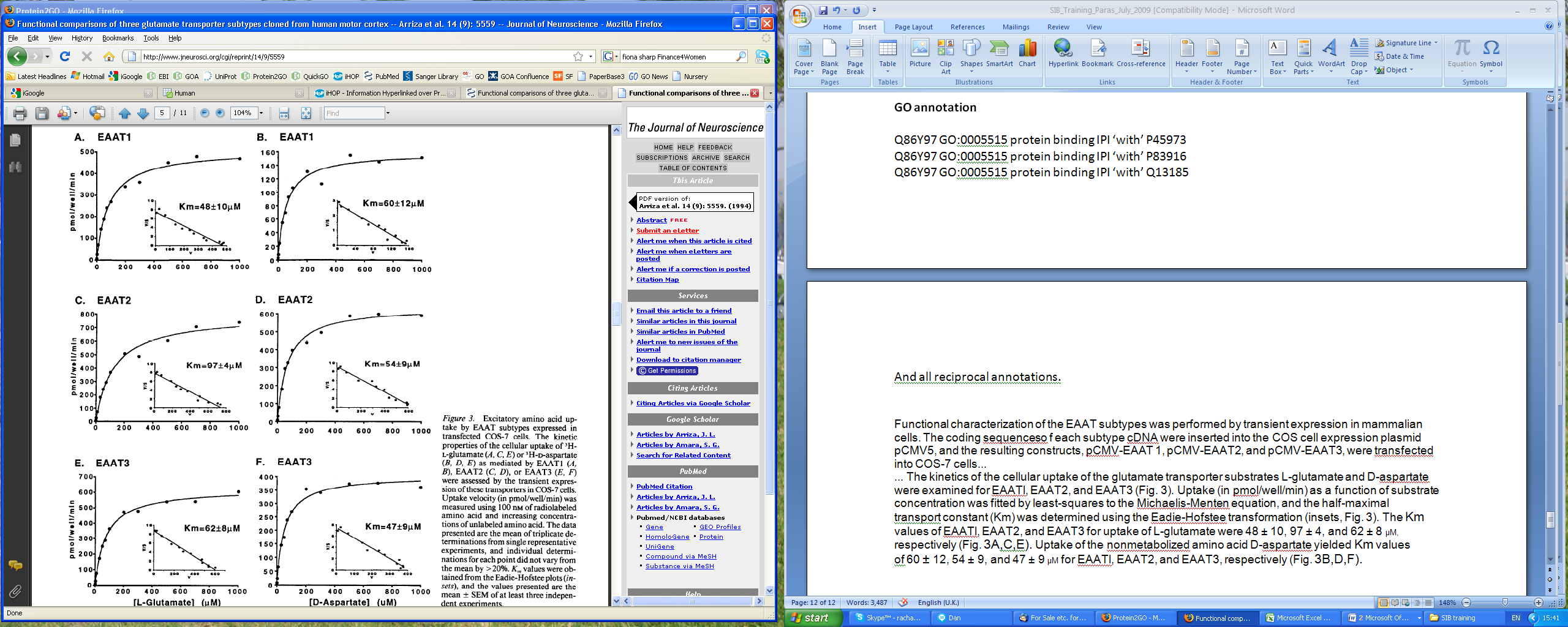
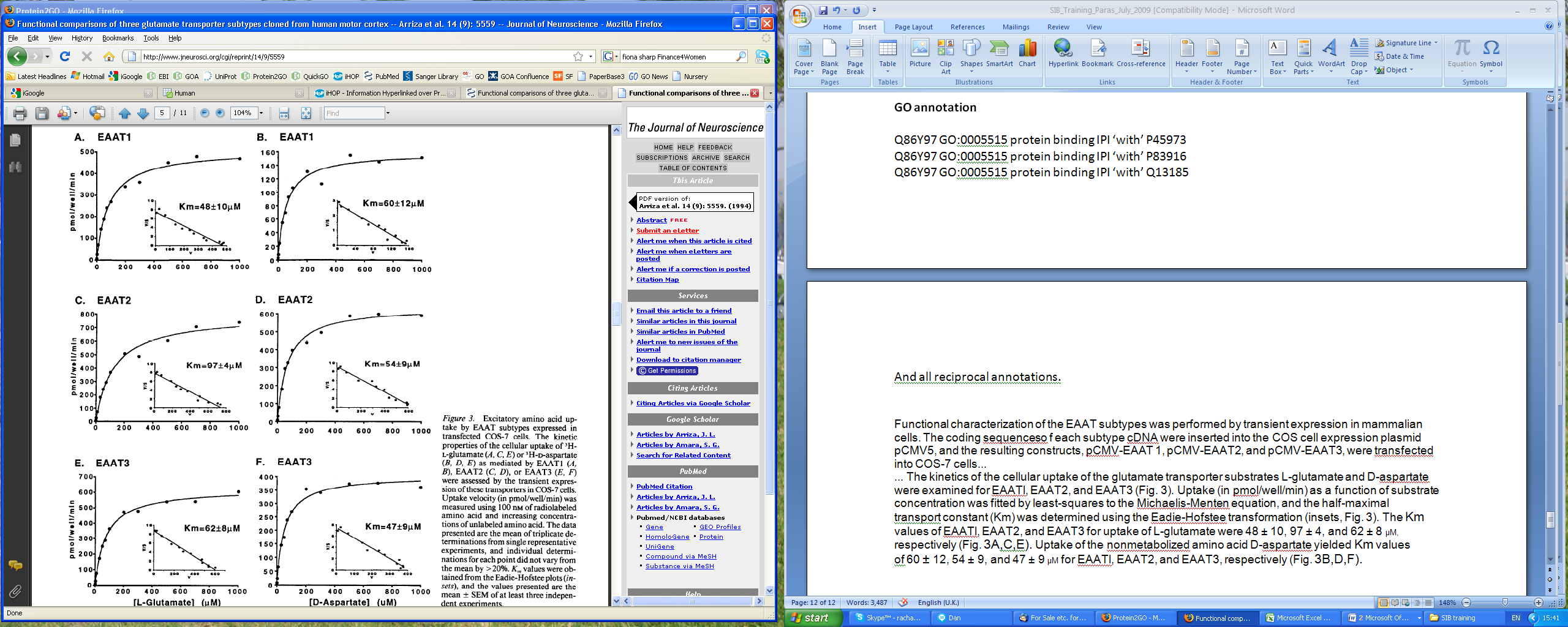
Proteins: P43003 (EAA1\_HUMAN)

P43004 (EAA2\_HUMAN)

P43005 (EAA3\_HUMAN)

Functional characterization of the EAAT subtypes was performed by transient expression in mammalian cells. The coding sequences of each subtype cDNA were inserted into the COS cell expression plasmid pCMV5, and the resulting constructs, pCMV-EAAT 1, pCMV-EAAT2, and pCMV-EAAT3, were transfected into COS-7 cells...

... The kinetics of the cellular uptake of the glutamate transporter substrates L-glutamate and D-aspartate were examined for EAAT1, EAAT2, and EAAT3 (Fig. 3). Uptake (in pmol/well/min) as a function of substrate concentration was determined (insets, Fig. 3). The Km values of EAAT1, EAAT2, and EAAT3 for uptake of L-glutamate were 48 ± 10, 97 ± 4, and 62 ± 8 µM, respectively (Fig. 3A,C,E). Uptake of the nonmetabolized amino acid D-aspartate yielded Km values of 60 ± 12, 54 ± 9, and 47 ± 9 µM for EAATl, EAAT2, and EAAT3, respectively (Fig. 3B,D,F).

|  |  |  |  |
| --- | --- | --- | --- |
| Protein ID | GO ID | GO Term Name | Evidence code |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

### Annotation example 5, PMID: 26323318

Proteins: Q9H1E3 (NUCKS1\_HUMAN)

Q96B01 (RAD51AP1\_HUMAN)

Q86YC2 (PALB2\_HUMAN)

Abbreviations:

HR = homologous recombination

DSB = double-strand break

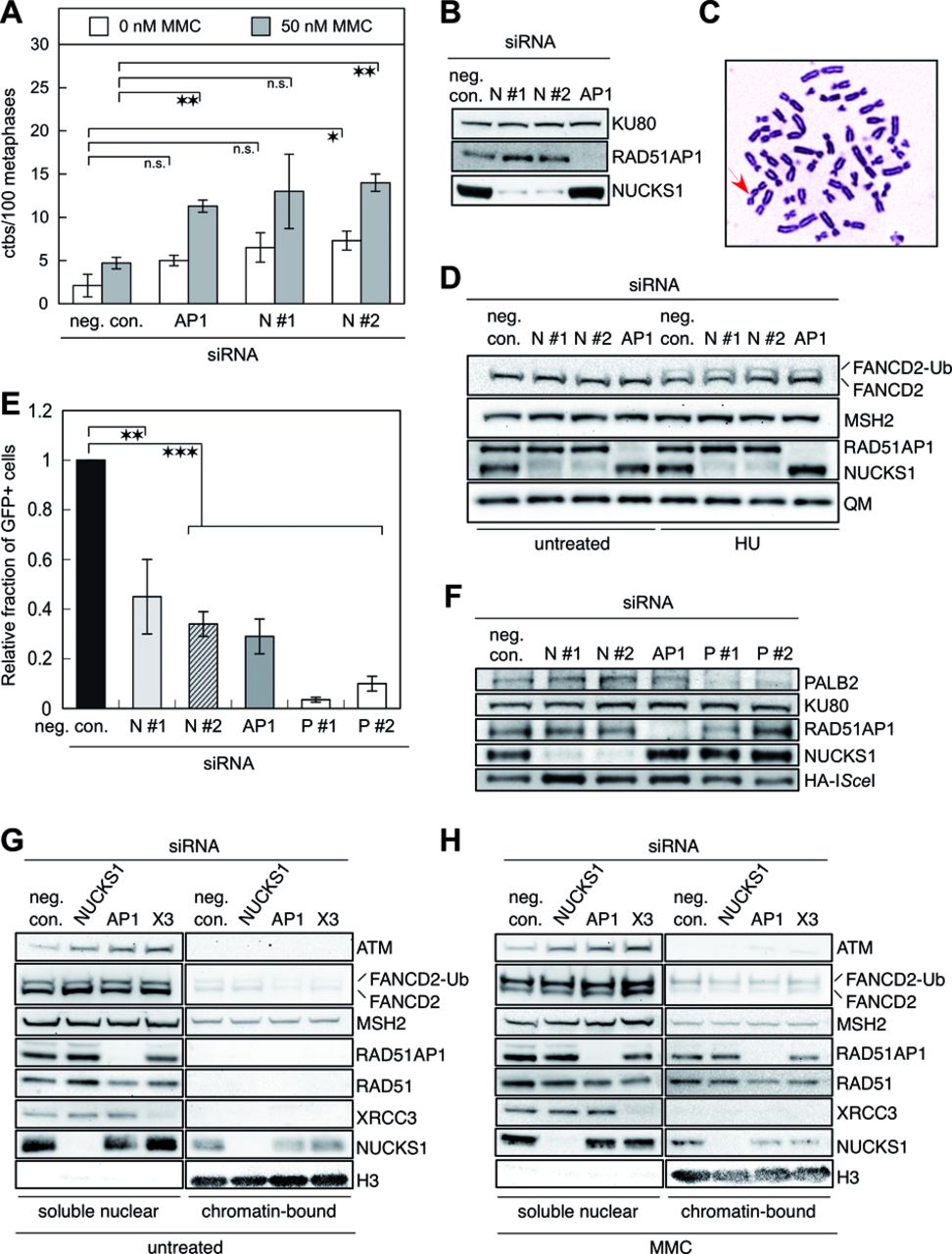
U2OS = a human-derived cell line

siRNA = small interfering RNA

**NUCKS1 and RAD51AP1 are equally important for DNA homology-directed DSB repair**

We used the recombinational reporter cell line U2OS-DRGFP (i.e. DR-U2OS) to directly examine the repair of DSBs induced by I-SceI endonuclease in the context of suppressed expression of NUCKS1. In this system, gene conversion HR triggered by DSB formation generates a functional GFP gene, expression of which can be conveniently monitored by flow cytometry. Importantly, depletion of NUCKS1 by one of two different siRNAs (N #1 or N #2) led to a significantly reduced level of gene conversion (Figure 3E and Supplementary Figure S2G). The level of impaired homology-mediated DSBR in NUCKS1-depleted cells is comparable to that engendered by RAD51AP1 knockdown (Figure 3E). However, we also found that knockdown of PALB2 by one of two different siRNAs elicits a more dramatic effect on gene conversion frequency than depletion of either NUCKS1 or RAD51AP1 (Figure 3E).

|  |  |  |  |
| --- | --- | --- | --- |
| Protein ID | GO ID | GO Term Name | Evidence code |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |



**Figure 3E.** (**E**) Homology-directed repair at DR-GFP is reduced in DR-U2OS cells depleted for NUCKS1 by one of two different siRNAs (N #1 and N #2). \*\*P ≤ 0.01; \*\*\*P ≤ 0.001. The effects of RAD51AP1 depletion (AP1) or of PALB2 depletion by one of two different siRNAs (P #1 and P #2) are shown for comparison purposes. Data are from eight independent experiments ± 1 SD.

### Annotation example 6, PMID: 22523093

Proteins: P90994 (DJ11\_CAEEL, cDJR1.1)

O16228 (DJ12\_CAEEL, cDJR1.2)

Q99497 (PARK7\_HUMAN, hDJ-1)

Abbreviations:

GO = Glyoxal

MGO = Methylglyoxal

Hint: Consider annotating to both Molecular function and Biological Process GO terms

The reactions tested were:

glyoxal + H2O = glycolic acid; methylglyoxal + H2O = D-lactate.

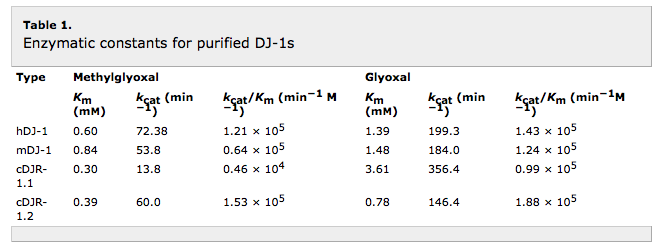
Background:

α-Oxoaldehydes including glyoxal (GO) and methylglyoxal (MGO) are produced by glucose oxidation, lipid peroxidation and DNA oxidation. They react non-enzymatically with amino groups of proteins, forming advanced glycation end-products (AGEs), which were implicated in aging, diabetes and neurodegenerative diseases, such as Parkinson and Alzheimers. Here, we characterized human DJ-1 (PARK7) and its homologs in the mouse and C. elegans (DJR1.1 and DJR1.2) as glyoxalases and investigate their roles in protecting cells, neurons and worms from glyoxals.

Results:

Based on the recent characterization of the bacterial enzyme, we carried out an experiment for animal enzymes to determine whether they exhibit glyoxalase activities. The His-tagged hDJ-1, mDJ-1, cDJR-1.1 and cDJR-1.2 were purified using the Ni-affinity column and were used to determine the enzymatic activity and conditions for the catalytic reaction. The results show typical Michaelis–Menten kinetics for all the enzymes tested (Table S1). The reaction products were identified by 1H-NMR and high-performance liquid chromatography (HPLC), such that lactic and glycolic acids were produced from methylglyoxal and glyoxal, respectively (Fig. 2A and C). Various types of aldehydes, including acetaldehyde, acrolein, glyceraldehydes, 2,3-butanedione and 2-carboxybenzaldehyde, were also tested as substrates, in which none of them was shown to be positive (unpublished data). The DJ-1 homologs have in general slightly higher affinities to methylglyoxal than to glyoxal, while specific activities are higher for glyoxal than for methylglyoxal (Table 1).

|  |  |  |  |
| --- | --- | --- | --- |
| Protein ID | GO ID | GO Term Name | Evidence code |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |





**Figure 2.** **Characterization of the DJ-1 homologs as glyoxalases.** (**A**) Purified human DJ-1 (50 μg) was mixed with GO (5 mM) or MGO (3 mM) for 30 min, and their reaction products, glycolic acid and lactic acid (two peaks), respectively, were analyzed with NMR. The peak appeared at ∼2 ppm is acetate contained in the reagent from Sigma, and (\*) indicates ethylene glycol, a contaminant. (**B**) The glyoxalase reaction of DJ-1 was also analyzed with DNPH assay, from which initial velocities were plotted with Michaelis–Menten and Lineweaver–Burk (inset) equations. Error bars represent standard deviations with triplicate experiments. (**C**) Purified DJ-1s from C. elegans were mixed with GO (10 mM) or MGO (10 mM), from which glycolic and lactic acids were detected by HPLC. The products were formed proportional to reaction time. (**D**) Enzyme reactions with cDJR-1.1 (100 μg) and cDJR-1.2 (100 μg) were monitored and plotted as described in (**B**).

### Further annotation example answers

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Example | Protein | GO ID | GO Term | Evidence |
| 4 | EAA1 | GO:0051938 | L-glutamate import | IDA |
|  | EAA1 | GO:0070779 | D-aspartate import | IDA |
|  | EAA2 | GO:0051938 | L-glutamate import | IDA |
|  | EAA2 | GO:0070779 | D-aspartate import | IDA |
|  | EAA3 | GO:0051938 | L-glutamate import | IDA |
|  | EAA3 | GO:0070779 | D-aspartate import | IDA |
| 5 | NUCKS1 | GO:0000724 | double-strand break repair via homologous recombination | IMP |
|  | RAD51AP1 | GO:0000724 | double-strand break repair via homologous recombination | IMP |
|  | PALB2 | GO:0000724 | double-strand break repair via homologous recombination | IMP |
| 6 | PARK7 | GO:0019172 | glyoxalase III activity | IDA |
|  | PARK7 | GO:1990422 | glyoxalase (glycolic acid-forming) activity | IDA |
|  | PARK7 | GO:0061727 | methylglyoxal catabolic process to lactate | IDA |
|  | PARK7 | GO:1903190 | glyoxal catabolic process | IDA |
|  | PARK7 | GO:0046295 | glycolate biosynthetic process | IDA |
|  | djr-1.1 | GO:0019172 | glyoxalase III activity | IDA |
|  | djr-1.1 | GO:1990422 | glyoxalase (glycolic acid-forming) activity | IDA |
|  | djr-1.1 | GO:0061727 | methylglyoxal catabolic process to lactate | IDA |
|  | djr-1.1 | GO:1903190 | glyoxal catabolic process | IDA |
|  | djr-1.1 | GO:0046295 | glycolate biosynthetic process | IDA |
|  | djr-1.2 | GO:0019172 | glyoxalase III activity | IDA |
|  | djr-1.2 | GO:1990422 | glyoxalase (glycolic acid-forming) activity | IDA |
|  | djr-1.2 | GO:0061727 | methylglyoxal catabolic process to lactate | IDA |
|  | djr-1.2 | GO:1903190 | glyoxal catabolic process | IDA |
|  | djr-1.2 | GO:0046295 | glycolate biosynthetic process | IDA |

### Further Resources

The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2018 47(D1):D330-D338. PMID:[30395331.](https://www.ncbi.nlm.nih.gov/pubmed/?term=30395331)

Kramarz B, Lovering RC. **Gene Ontology: A Resource for Analysis and**

**Interpretation of Alzheimer’s Disease Data.** In: Wisniewski T, editor. Alzheimer’s

Disease [Internet]. Brisbane (AU): Codon Publications; 2019 Dec 20. Chapter 2.

[PMID:31895510](https://www.ncbi.nlm.nih.gov/pubmed/?term=31895510)

Kramarz B, Roncaglia P, Meldal BHM, Huntley RP, Martin MJ, Orchard S, Parkinson H, Brough D, Bandopadhyay R, Hooper NM, Lovering RC. **Improving the Gene Ontology Resource to Facilitate More Informative Analysis and Interpretation of Alzheimer’s Disease Data**. Genes(Basel). 2018 9(12). pii: E593. PMID:[30501127](https://www.ncbi.nlm.nih.gov/pubmed/30501127).

Denny P, Feuermann M, Hill DP, Lovering RC, Plun-Favreau H, Roncaglia P. **Exploring autophagy with Gene Ontology**. Autophagy. 2018 17:1-18. [PMID:29455577.](https://www.ncbi.nlm.nih.gov/pubmed/?term=29455577)

Lovering RC. (2017) **How Does the Scientific Community Contribute to Gene Ontology?**Methods Mol Biol. 1446:85-93. [PMID:27812937](https://www.ncbi.nlm.nih.gov/pubmed/?term=27812937).

Foulger RE, Denny P, Hardy J, Martin MJ, Sawford T, Lovering RC (2016). **Using the Gene Ontology to Annotate Key Players in Parkinson's Disease.** Neuroinformatics. 14(3):297-304. [PMID:26825309](http://www.ncbi.nlm.nih.gov/pubmed/26825309)

Huntley, R.P. *et al.* **A method for increasing expressivity of Gene Ontology annotations using a compositional approach**. BMC Bioinformatics. 2014 May 21;15:155. [PMID:24885854](https://www.ncbi.nlm.nih.gov/pubmed/?term=24885854)

***Introduction to Biomedical Ontologies***, video created by the Rat Genome Database (RGD) <https://rgd.mcw.edu/wg/home/rgd_rat_community_videos/ontology-term-enrichment-using-ratmine>/

***Anatomy of an Annotation***, video tutorials from RGD <https://rgd.mcw.edu/wg/home/rgd_rat_community_videos/anatomy_of_an_ontology_annotation-part_1/>

<https://rgd.mcw.edu/wg/home/rgd_rat_community_videos/anatomy_of_an_ontology_annotation-part_2/>

***Using QuickGO***, tutorials from EBI <https://www.ebi.ac.uk/training/online/course/quickgo-gene-ontology-annotation>

Kinsella, R.J. *et al* **Ensembl BioMarts: a hub for data retrieval across taxonomic space.** Database (Oxford) 2011:bar030 [PMID:21785142](https://www.ncbi.nlm.nih.gov/pubmed/?term=21785142)

### Table 1. A selection of GO evidence codes used for manual annotation

|  |  |
| --- | --- |
| **Evidence code** | **Description** |
| EXP | Experimental: parent evidence code to IDA, IEP, IGI, IMP, IPI |
| IDA | Inferred from Direct Assay   * Enzyme assays * In vitro reconstitution (e.g. transcription) * Immunofluorescence (for cellular component) * Cell fractionation (for cellular component) * Physical interaction/binding assay (sometimes appropriate for cellular component or molecular function) |
| IEP | Inferred from Expression Pattern (do not use, discuss with curator) |
| IGI | Inferred from Genetic Interaction (Ideally has a protein accession ID in the ‘WITH’ field, remember to create the reciprocal annotation)   * "Traditional" genetic interactions such as suppressors, synthetic lethals, etc. * Functional complementation * Rescue experiments * Inference about one gene drawn from the phenotype of a mutation in a different gene |
| IMP | Inferred from Mutant Phenotype   * mutations, natural or introduced, that result in partial or complete impairment or alteration of the function of that gene * polymorphism or allelic variation (including where no allele is designated wild-type or mutant) * any procedure that disturbs the expression or function of the gene, including RNAi or the use of any molecule or experimental condition that may disturb or affect the normal functioning of the gene, such as inhibitors * overexpression or ectopic expression of wild-type or mutant gene that results in aberrant behavior of the system or aberrant expression where the resulting mutant phenotype is used to make a judgment about the normal activity of that gene product |
| IPI | Inferred from Physical Interaction (must have a protein accession ID in the ‘WITH’ field, remember to create the reciprocal annotation)   * 2-hybrid interactions * Co-purification * Co-immunoprecipitation * Ion/protein binding experiments |
| TAS | Traceable Author Statement   1. Any statement in an article where the original evidence (experimental results, sequence comparison, etc.) is not directly shown, but is referenced in the article and therefore can be traced to another source. |
| NAS | Non-traceable Author Statement   1. Statements in papers (abstract, introduction, or discussion) that a curator cannot trace to another publication |
| ISS | Inferred from sequence similarity |

Table 2. GO evidence code Decision tree for manual annotation ****

**Contributions**

This workbook was put together by the UCL Functional Gene Annotation Team

A close up of a logo

Description automatically generated

Some sections of the workbook have been adapted from the **UniProt** workbook provided by EBI

|  |  |
| --- | --- |
| EMBL | Screen Shot 2016-06-29 at 15 |
|  |  |
|  |  |