

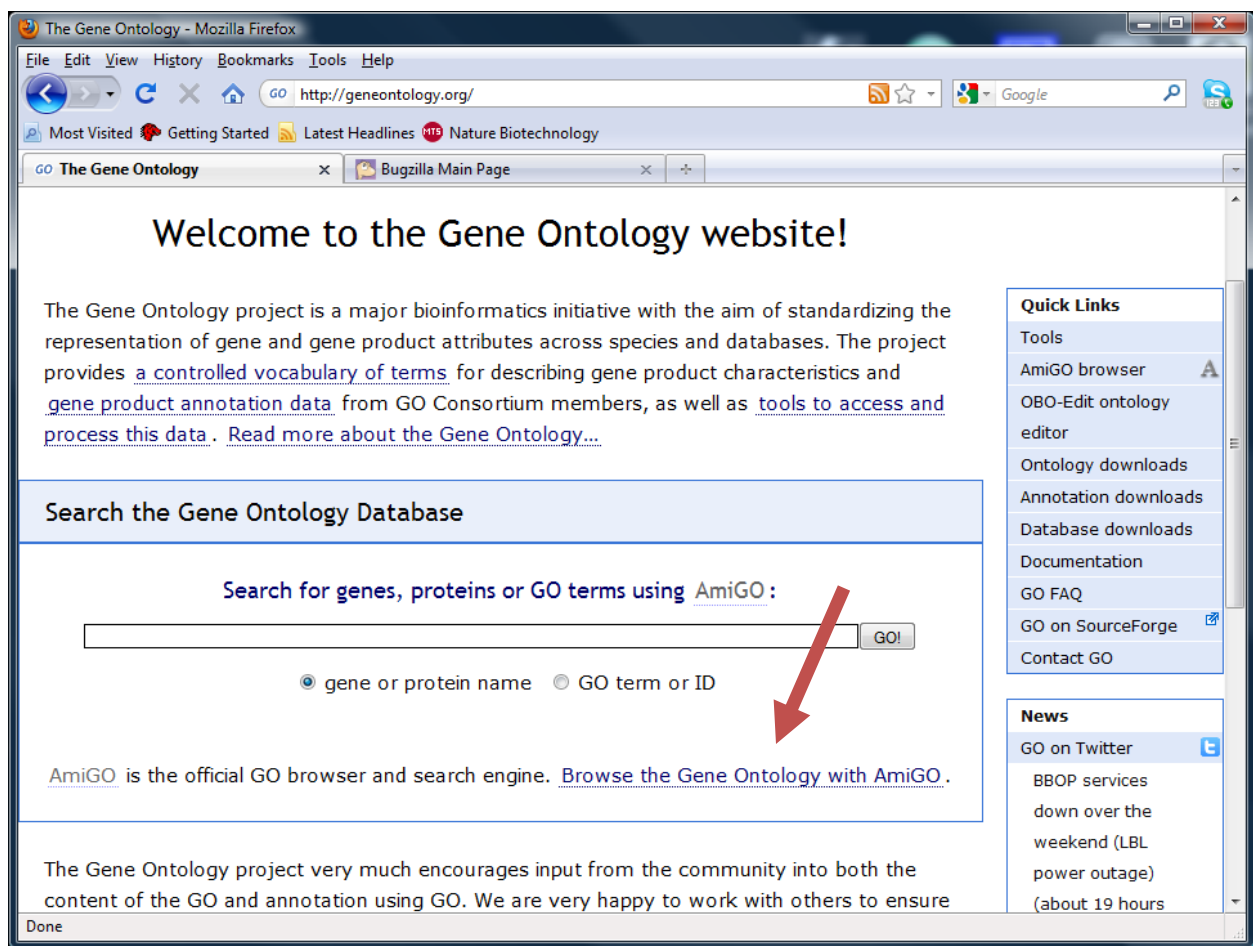
GO, DAVID and iRefScape revisited.

Ian Donaldson

MBV-INF 4410/9410

This exercise will revisit a number of applications introduced in the class. You will learn how to browse the Gene Ontology, select a set of genes that are annotated with some specific GO term and then you will analyze this gene list using DAVID and iRefScape.

Go to <http://geneontology.org/>



The screenshot shows the Gene Ontology website homepage. The main heading is "Welcome to the Gene Ontology website!". Below this, there is a paragraph describing the project's aim. A search bar is prominently displayed with the text "Search for genes, proteins or GO terms using AmiGO:". A red arrow points to the "GO!" button next to the search input field. Below the search bar, there are radio buttons for "gene or protein name" (selected) and "GO term or ID". At the bottom of the search section, there is a link: "AmiGO is the official GO browser and search engine. [Browse the Gene Ontology with AmiGO.](#)". The right sidebar contains "Quick Links" and "News" sections.

Click on "Browse the Gene Ontology with AmiGO".

Spend some time browsing through each of the three GO Ontologies. Look for terms that you are familiar with and see how they relate to terms above (parent terms) and below (child terms) in the GO.

AmiGO: Tree Browser - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://amigo.geneontology.org/cgi-bin/amigo/browse.cgi

Most Visited Getting Started Latest Headlines Nature Biotechnology

AmiGO: Tree Browser Bugzilla Main Page

the Gene Ontology AmiGO

Search Browse BLAST More Tools Help

Search GO terms genes or proteins exact match

Tree Browser

▼ Filter tree view

Filter by ontology

Ontology

All

biological process

cellular component

molecular function

Filter Gene Product Counts

Data source

All

ASAP

AspGD

CGD

Species

All

Arabidopsis thaliana

Bacillus anthraci...

Bacillus subtilis

View Options

Tree view Full Compact

all : all [447301 gene products]

GO:0008150 : biological_process [343853 gene products]

GO:0005575 : cellular_component [305208 gene products]

GO:0003674 : molecular_function [366536 gene products]

Actions...

Last action: Reset the tree

Graphical View

Permalink

Download...

OBO

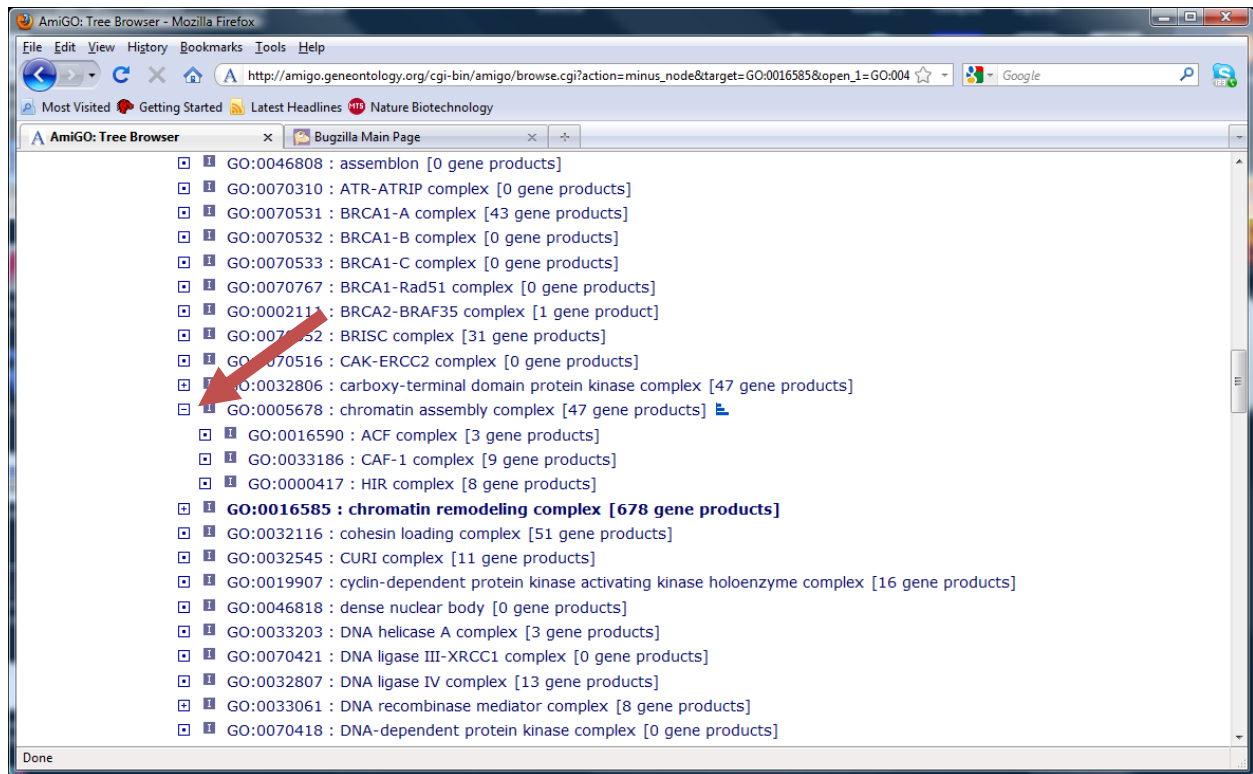
RDF-XML

Graphviz dot

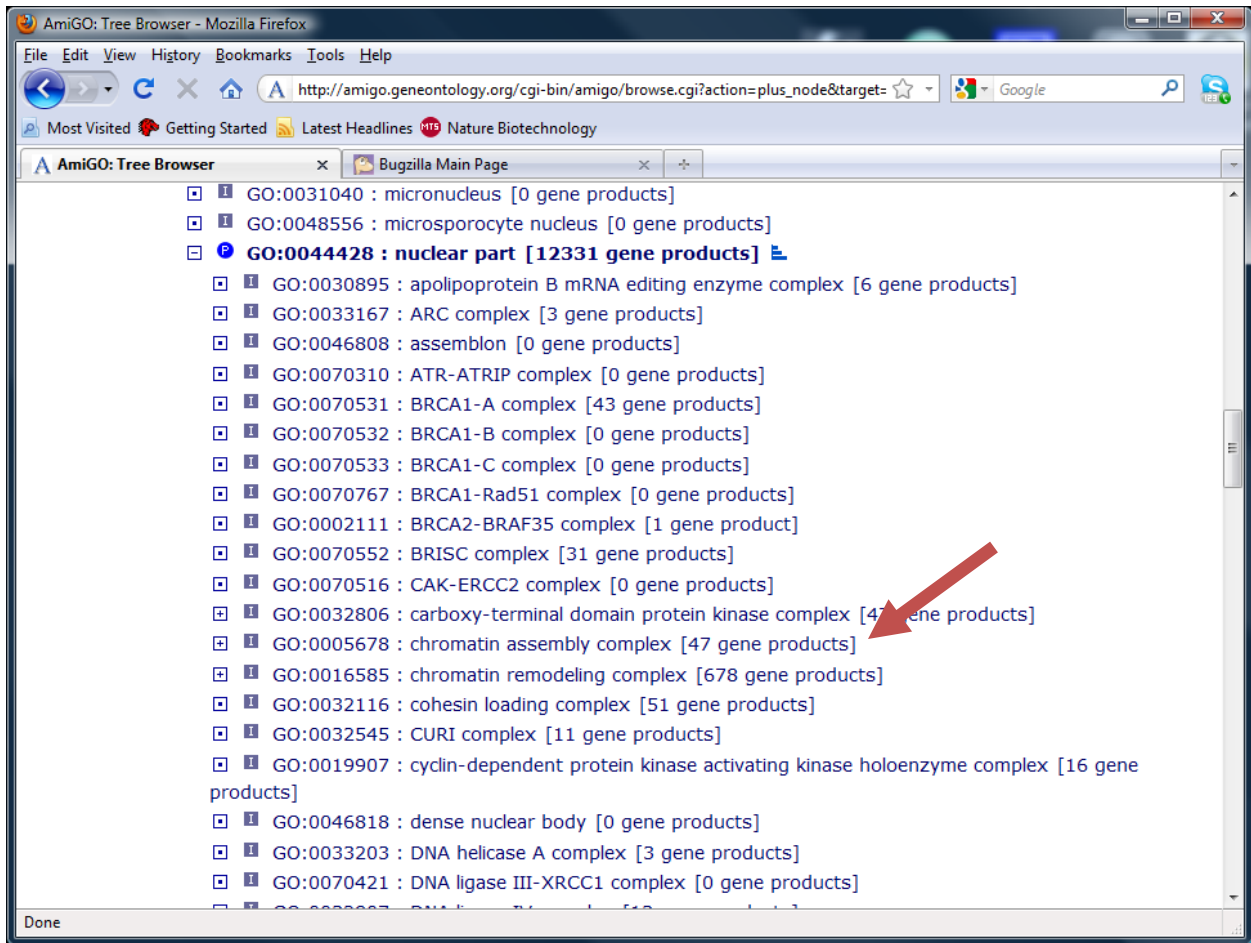
Done

Try to make your way to the term that describes “nucleus” without directly searching for it. Instead, start by expanding the cellular_component “root node” (GO:0005575) by clicking on the + beside it. Look for the next closest thing to “nucleus” in the expanded list and then click on that. If you want to look at the definition of any given term, just click on it.

If you can't find the term entry for “nucleus”, use the next page as a hint or search for nucleus in the “Search GO” box at the top of the interface.



Now go back and click on the 47 gene products that correspond to the “chromatin assembly complex”.



Click on the "47 gene products" to view them.

AmiGO: Term Association Details - Mozilla Firefox

http://amigo.geneontology.org/cgi-bin/amigo/term-assoc.cgi?term=GO:0005678&session_id=9682amigo1284654448

Gene Product Associations to chromatin assembly complex ; GO:0005678 and children

Download all association information in: gene association format RDF/XML

Filter associations displayed

Filter by Gene Product: Gene Product Type: Data source: Species: Evidence Code: View associations: All Direct associations

1 2

chromatin assembly complex ; GO:0005678 [\[show def\]](#) [\[view in tree\]](#)

	Symbol, full name	Information	Qualifier	Evidence	Reference	Assigned by
<input type="checkbox"/>	Acf1 ATP-dependent chromatin assembly factor large subunit	21 associations BLAST protein from <i>Drosophila melanogaster</i>		NAS	FB:FBrf0105495	FlyBase
<input type="checkbox"/>	AN4891	13 associations gene_product from <i>Emericella nidulans</i>		IEA With SGD:5000003651	AspGD REF:ASPL0000000005	ASPGD (via AspGD)
<input type="checkbox"/>	AN8039	6 associations gene_product from <i>Emericella nidulans</i>		IEA With SGD:5000005372	AspGD REF:ASPL0000000005	ASPGD (via AspGD)
<input type="checkbox"/>	ASF1	11 associations gene_product from		IEA	CGD REF:CAL0121033	CGD

Done

Spend some time browsing the links from this page (or just hovering over them). Scroll down. Genes annotated with each of the three leaf terms are grouped separately on the page.

What kinds of evidence are provided for these assignments. What kinds might you be cautious of (less likely to believe) or more likely to believe? Who assigns these GO terms to the genes?

Try filtering the list using the filters at the top. Make selections from the menus (Gene Product Type, Data source, Species and Evidence code and then click "Set filters".

Next, try to export your results for all genes from all data sources from human (Homo sapiens) that have any evidence code. Set filters and then click on "Download all association information in gene association format".

AmiGO: Term Association Details - Mozilla Firefox

http://amigo.geneontology.org/cgi-bin/amigo/term-assoc.cgi?gptype=all&speciesdb=all&taxid=9606&evcode=all&t

Search GO terms genes or proteins exact match Submit Query

chromatin assembly complex

Term associations Term information Term lineage External references

Gene Product Associations to chromatin assembly complex ; GO:0005678 and children

Download all association information in: gene association format RDF/XML

Current filters

Species:

Filter associations displayed

Filter by Gene Product

Gene Product Type	Data source	Species
All	All	Geobacillus stear...
gene	ASAP	Geobacter sulfur...
gene product	AspGD	Haloarcula marism...
mirna	CGD	Homo sapiens

Filter by Association

Evidence Code

View associations: All Direct associations

Set filters Remove all filters

chromatin assembly complex ; GO:0005678 [show def] [view in tree]

Symbol, full name	Information	Qualifier	Evidence	Reference	Assigned by
<input type="checkbox"/> CHAF1A	view associations protein from <i>Homo sapiens</i>		TAS	PMID:7600578	Proteome Inc.

Done

Mozilla Firefox

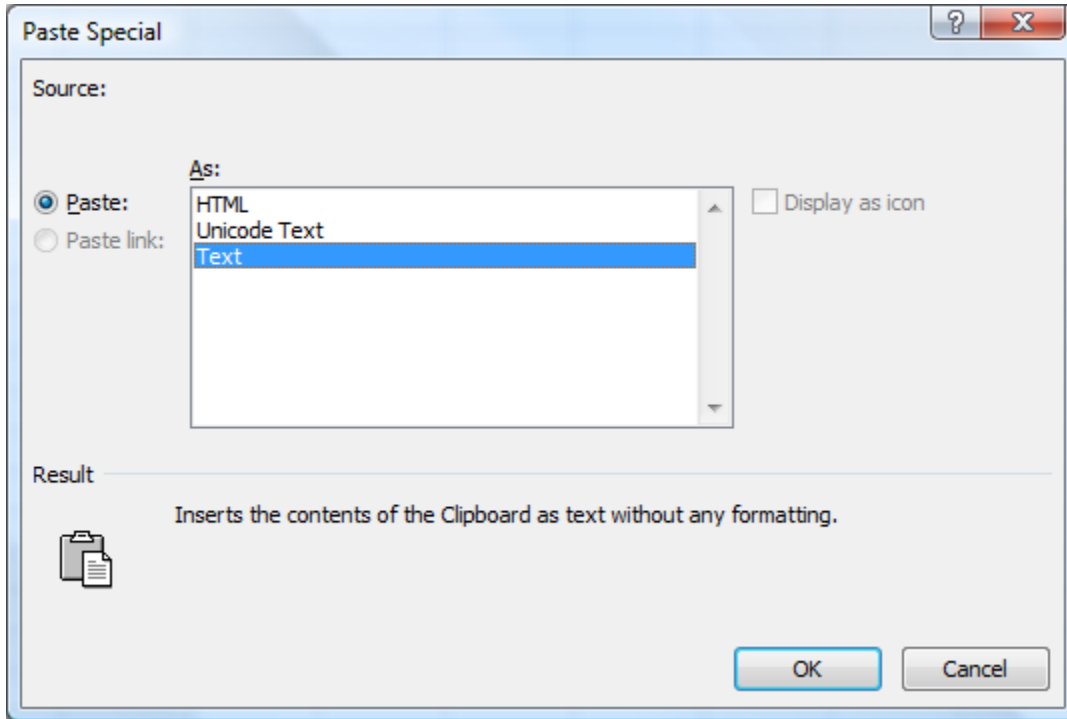
http://amigo.geneontology.org/cgi-bin/amigo/term-assoc.cgi?term=GO:0005678&format=go_assoc&session_id=968:

http://amigo.ge...migo1284654448 x Bugzilla Main Page

UniProtKB	Q13112	CHAF1B	GO:0005678	PMID:7600578	TAS	C	Chromatin assembly factor 1 subunit B	CAF1A CAF1B
UniProtKB	P55209	NAP1L1	GO:0005678	PMID:8297347	TAS	C	Nucleosome assembly protein 1-like 1	IFI000238
UniProtKB	Q99733	NAP1L4	GO:0005678	PMID:9325046	TAS	C	Nucleosome assembly protein 1-like 4	IFI009414
UniProtKB	Q13111	CHAF1A	GO:0005678	PMID:7600578	TAS	C	Chromatin assembly factor 1 subunit A	CAF CAF1A
UniProtKB	Q9ULW6	NAP1L2	GO:0005678	PMID:8789438	TAS	C	Nucleosome assembly protein 1-like 2	B2RE61 B2RE62
UniProtKB	Q99457	NAP1L3	GO:0005678	PMID:8976385	TAS	C	Nucleosome assembly protein 1-like 3	B2RCMO B2RCMP
UniProtKB	Q9NRL2	BAZ1A	GO:0016590	PMID:10662543	TAS	C	Bromodomain adjacent to zinc finger domain prote	

Done

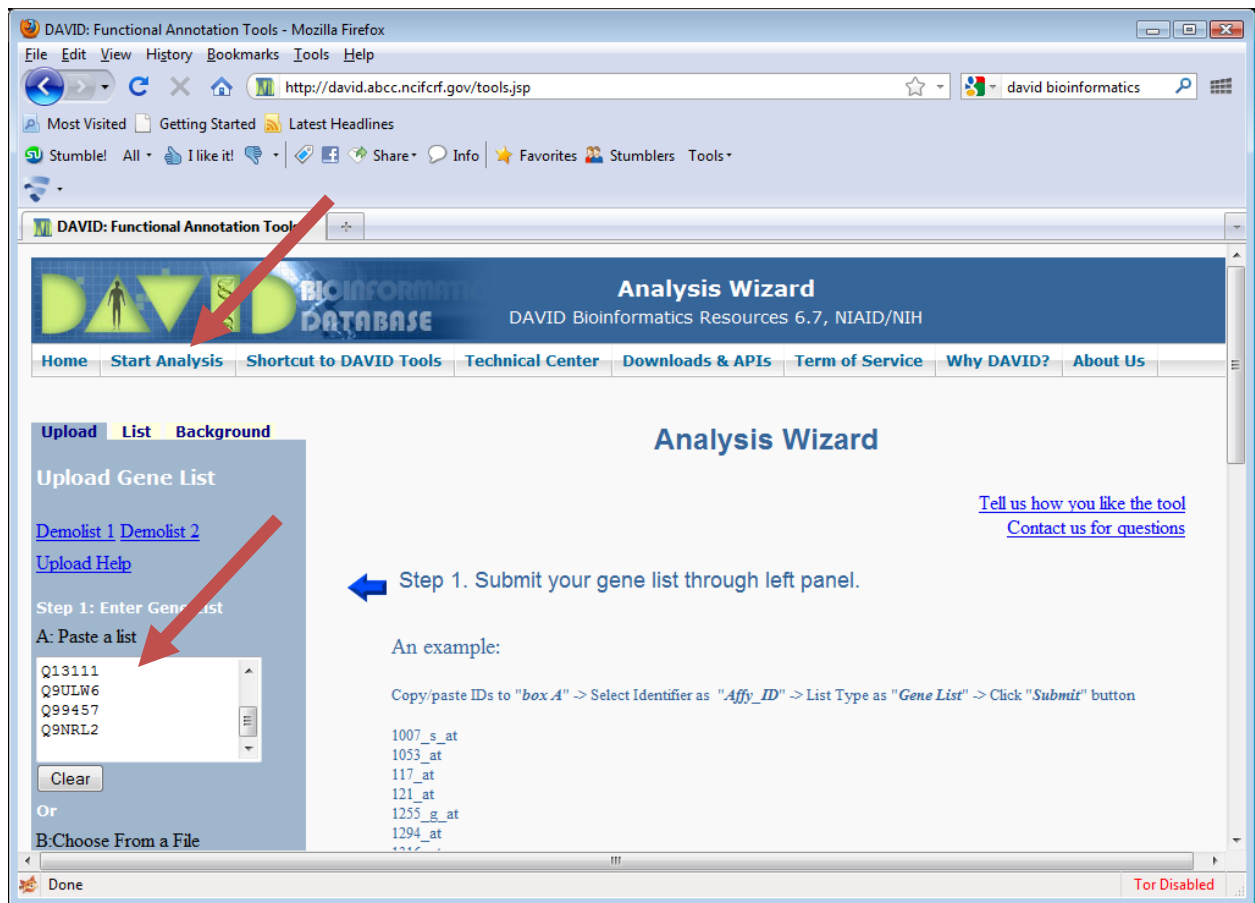
Copy and paste this to an Excel spreadsheet (or similar). Right-click, paste special, as text.



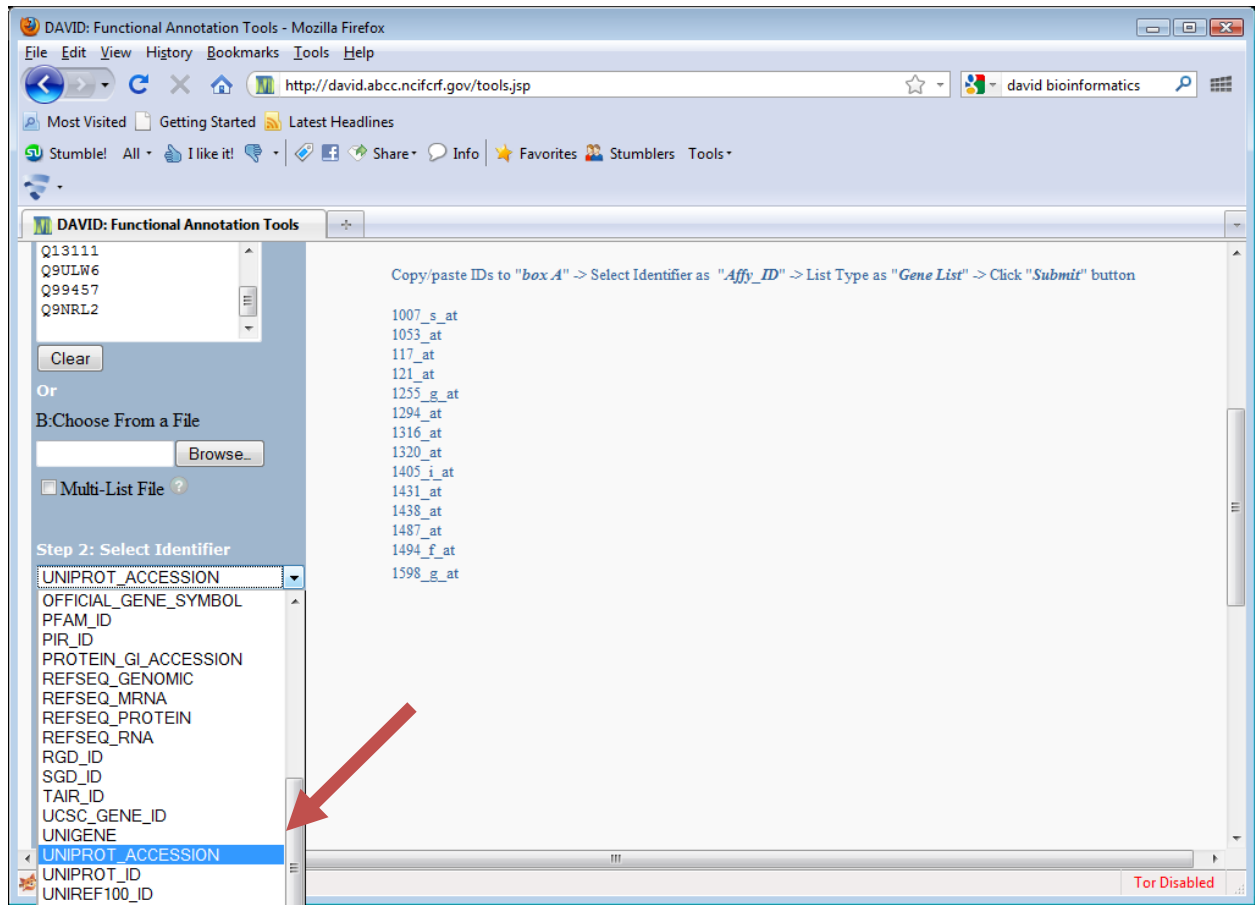
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	UniProtKE Q13112	CHAF1B		GO:00056; PMID:760; TAS					C	Chromatin CAF1A CA protein	taxon:960 20030904			PINC		
2	UniProtKE P55209	NAP1L1		GO:00056; PMID:829; TAS					C	Nucleosor IPI00238; protein	taxon:960 20030904			PINC		
3	UniProtKE Q99733	NAP1L4		GO:00056; PMID:932; TAS					C	Nucleosor IPI009414; protein	taxon:960 20030904			PINC		
4	UniProtKE Q13111	CHAF1A		GO:00056; PMID:760; TAS					C	Chromatin CAF CAF1 protein	taxon:960 20030904			PINC		
5	UniProtKE Q9ULW6	NAP1L2		GO:00056; PMID:878; TAS					C	Nucleosor B2RE61 Bi protein	taxon:960 20030904			PINC		
6	UniProtKE Q99457	NAP1L3		GO:00056; PMID:897; TAS					C	Nucleosor B2RCM0 E protein	taxon:960 20030904			PINC		
7	UniProtKE Q9NRL2	BAZ1A		GO:00165; PMID:106; TAS					C	Bromodor ACF1 BAZ protein	taxon:960 20100527			BHF-UCL		
8																
9																
10																
11																
12																

Select and copy the UniProt accessions from column B.

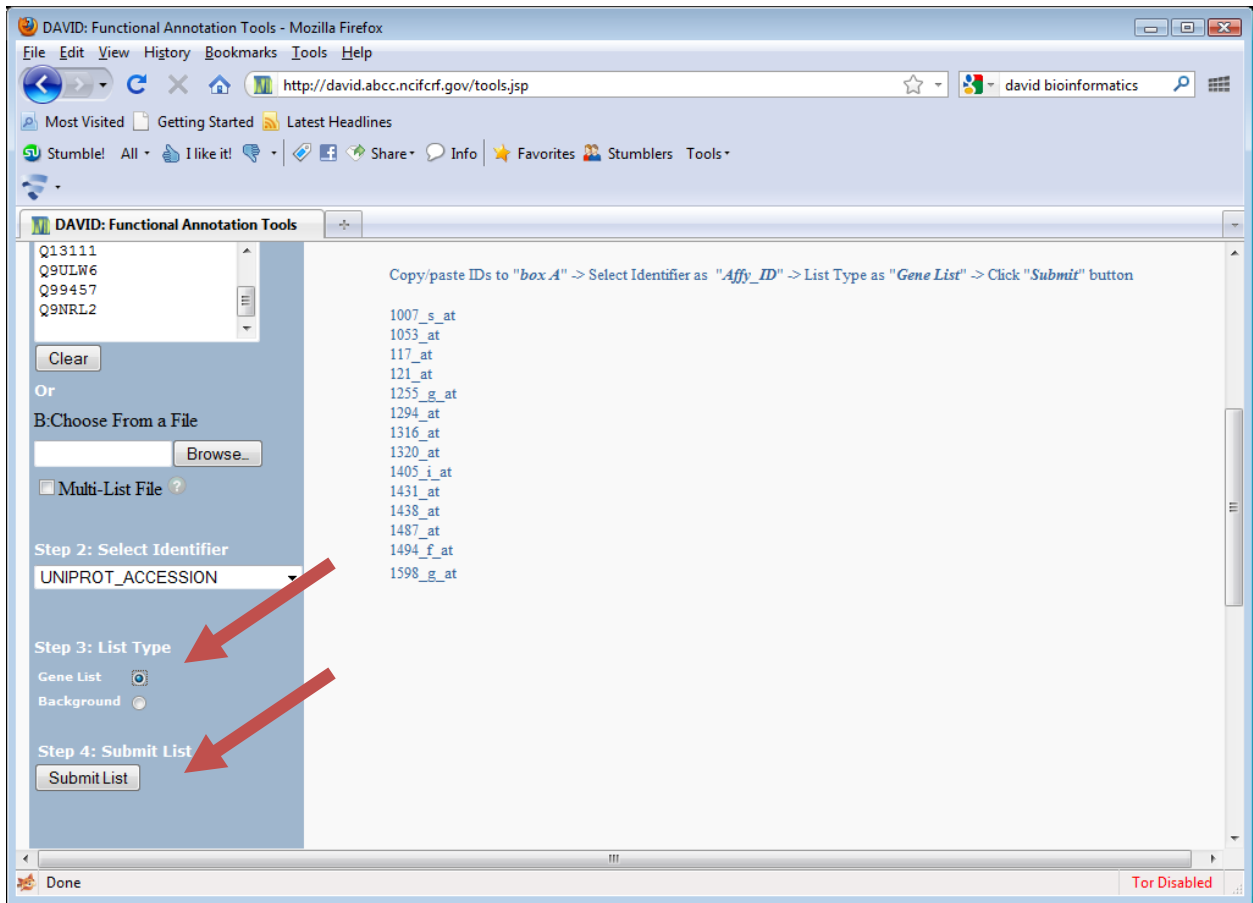
Go to <http://david.abcc.ncifcrf.gov/home.jsp> , click on Start Analysis and then paste the list of accessions into the query box like this:



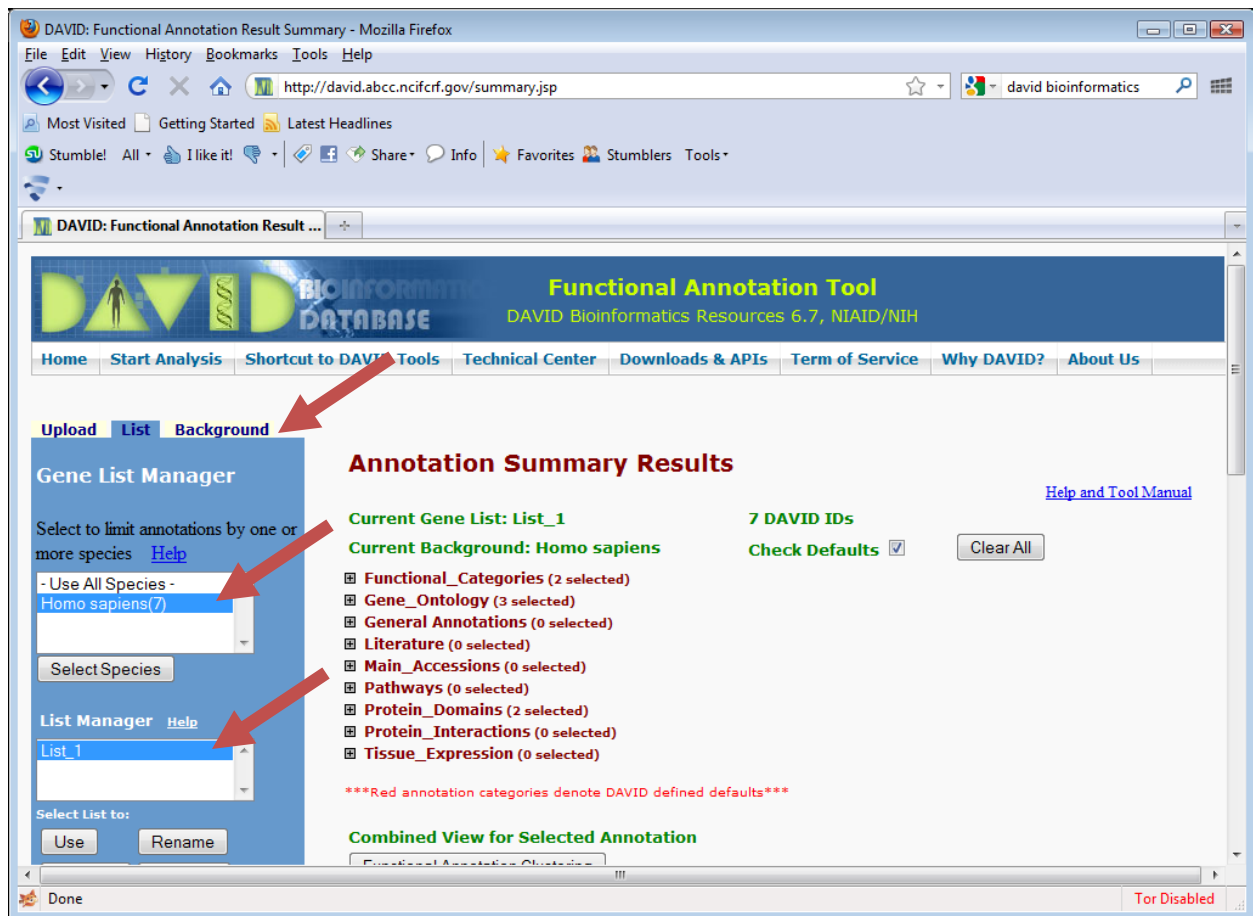
The tell DAVID that you have entered UniProt Accessions. Like this



Then select “Gene List” under List type and click on Submit list:



Your list has been saved by DAVID as List_1 and it has automatically recognized “Homo sapiens” as the species from which the list is derived:



If you click on the “Background” tab, you will see that DAVID has also set “Homo sapiens” as the “background”. Its important that you check these settings and change them if necessary. Discuss why.

You can now look at categories that are over-represented in your list. For example, click on the + beside Gene_Ontology and scroll down to GOTERM_CC_FAT like this:

DAVID: Functional Annotation Result Summary - Mozilla Firefox

http://david.abcc.ncifcrf.gov/summary.jsp

Current Background: Homo sapiens

Functional_Categories (2 selected)

Gene_Ontology (3 selected)

Category	Percentage	Count	Action
<input type="checkbox"/> GOTERM_BP_1	100.0%	7	Chart
<input type="checkbox"/> GOTERM_BP_2	100.0%	7	Chart
<input type="checkbox"/> GOTERM_BP_3	100.0%	7	Chart
<input type="checkbox"/> GOTERM_BP_4	100.0%	7	Chart
<input type="checkbox"/> GOTERM_BP_5	100.0%	7	Chart
<input type="checkbox"/> GOTERM_BP_ALL	100.0%	7	Chart
<input checked="" type="checkbox"/> GOTERM_BP_FAT	100.0%	7	Chart
<input type="checkbox"/> GOTERM_CC_1	100.0%	7	Chart
<input type="checkbox"/> GOTERM_CC_2	100.0%	7	Chart
<input type="checkbox"/> GOTERM_CC_3	100.0%	7	Chart
<input type="checkbox"/> GOTERM_CC_4	100.0%	7	Chart
<input type="checkbox"/> GOTERM_CC_5	100.0%	7	Chart
<input type="checkbox"/> GOTERM_CC_ALL	100.0%	7	Chart
<input checked="" type="checkbox"/> GOTERM_CC_FAT	100.0%	7	Chart
<input type="checkbox"/> GOTERM_MF_1	71.4%	5	Chart
<input type="checkbox"/> GOTERM_MF_2	71.4%	5	Chart

If you then click on the “Chart” button, you see a new window open like this:

DAVID: Database for Annotation, Visualization, and Integrated Discovery (Laboratory of Immunopathogenesis and Bioinformatics (LIB); National Institute of Allergies and Infections (NIAID); National Institute of Allergy and Infectious Diseases (NIAID), NIH)

http://david.abcc.ncifcrf.gov/chartReport.jsp?annot=33

DAVID Bioinformatics Resources 6.7

National Institute of Allergy and Infectious Diseases (NIAID), NIH


Functional Annotation Chart

[Help and Manual](#)

Current Gene List: List_1
 Current Background: Homo sapiens
 7 DAVID IDs
 Options








Rerun Using Options Create Sublist

1 chart records [Download File](#)

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_CC_FAT	chromatin assembly complex	RT		6	85.7	1.3E-17	2.3E-16

1 gene(s) from your list are not in the output.

Please cite [Nature Protocols 2009; 4\(1\):44](#) & [Genome Biology 2003; 4\(5\):P3](#) within any publication that makes use of any methods inspired by DAVID.

[Term of Service](#) | [Contact Us](#) | [Site Map](#)

Done Tor Disabled

6 out of the 7 genes in your list (85.7%) are annotated with the GO term for chromatin assembly complex. The probability of randomly choosing 7 genes from the human genome where 6 of them all have this associated GO term is 1.3E.17. When you correct for multiple hypothesis testing (Benjamini), the probability is 2.3E-16 (still quite surprising – i.e. we would suspect that whatever “process” was used to pick out these 7 genes was not unrelated to this annotation). And of course, we know this to be the case.

Click on the “chromatin assembly complex” to see details about the term that is overrepresented.

Click on the blue bar underneath “Genes” to see the list of genes that had this annotation.

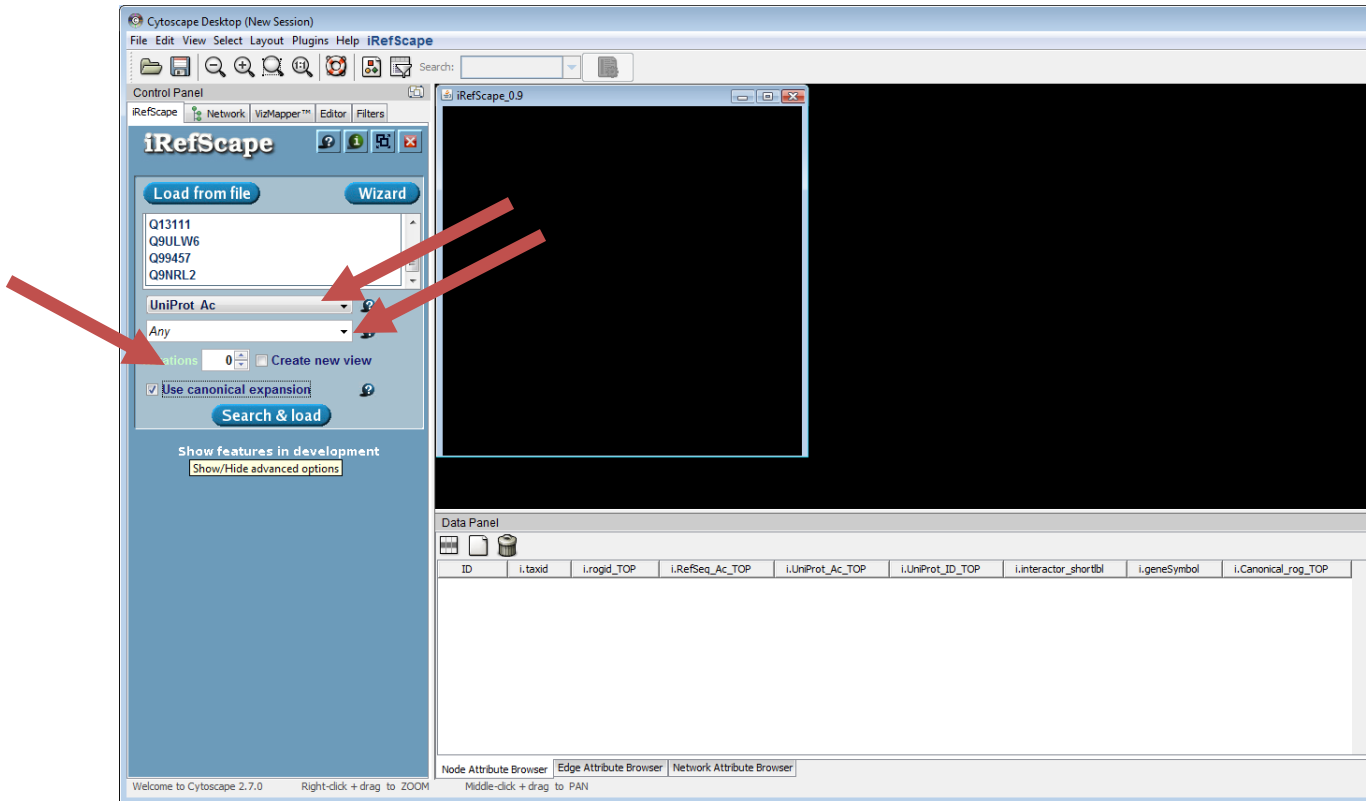
Go back to the “Annotation summary results” and explore whether other categories of annotation were over represented in this list.

There is a lot of other material you can explore on this site. Make a note of the Nature Protocols tutorial on use of DAVID for later.

<http://www.nature.com/nprot/journal/v4/n1/pdf/nprot.2008.211.pdf>

Next, we will explore known interactions between the genes in our list.

Start Cytoscape and the iRefScape plugin from Plugins menu?iRefScape 0.9. Then copy and paste the identifiers from column B of the Excel spreadsheet to the query box. Like this...



Make sure that the search type is set to UniProt Ac, the taxon is set to Any (or Homo Sapiens) and that “Use canonical expansion” is selected and that iterations are set to 0 (see red arrows above). Iterations of zero will only return interactions that occur between proteins in our query list. Then click on search and load.

The initial results will look like this

The screenshot shows the Cytoscape Desktop interface with the iRefScope plugin. The main window displays a network graph with nodes and edges. The left sidebar contains the iRefScope control panel with search and load options. The bottom right shows a Data Panel with a table of protein interactions.

ID	I.taxid	I.rigid_TOP	I.RefSeq_Ac_TOP	I.UniProt_Ac_TOP	I.UniProt_ID_TOP	I.interactor_shortbl	I.geneSymbol
3600472	9606	PtjzXGuxpQqr...	NP_004528	P55209	NP1L1_HUMAN		[NAP1L1]
907988	9606	8ZiQPvmtN...	NP_005474				[CHAF1A]
297207	9606	1N8UuHZKU...	NP_038476	Q9NRL2	BAZ1A_HUMAN		[BAZ1A]
Ijw5K2Gz1ddjKVx1q7ozXnsYJJA							[]
10012887	9606	jje8vhT2pRf0...		Q13111	CAF1A_HUMAN		[]
LQHFADWz4jzpzF7SZGRp5Xxwn8							[]
stD6cjxBWVezrBphXont+9dhsAM							[]
ssje7XqHIV37qc+Ymh7SffY6Q							[]
4601076	9606	VvbW9mpU...	NP_005432	Q13112	CAF1B_HUMAN		[CHAF1B]
hMwbwvpv8YDjgDfLpq5vroalwWo							[]
3997538	9606	sAdePPKwz...	NP_068798	Q9ULW6	NP1L2_HUMAN		[NAP1L2]
DiaoPiddtPeexzGBfslvakWzAA							[]

And you should be able to clean them up (rearrange them) to look like this.

Hint use Layout/align and distribute to align and stack the purple hexagons (nodes that represent complexes) and then iRefScope/ViewTools/Toggle selected multi-edges to hide multi-edges (representing multiple experiments that support the same protein-protein interaction).

Go back to the Gene Ontology pages. Should more of these proteins have been involved in a single complex?

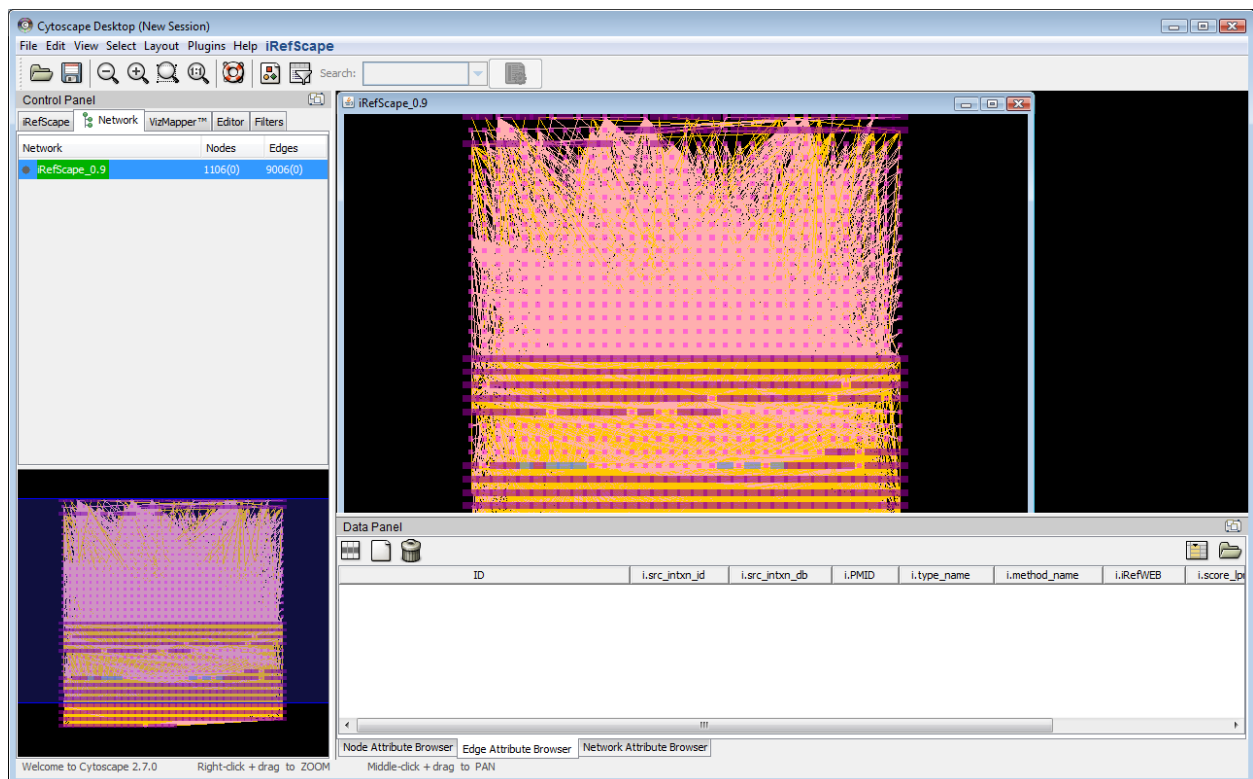
It is left as an exercise to see if you would get different results by querying iRefScape with proteins from different organisms (say *Drosophila*) that are annotated as belonging to the “chromatin assembly complex”.

Next, we will move on to a technique that will identify proteins related to our initial query. We want to find proteins that interact with 2 or more of the proteins in our starting list (i.e. things that are associated with the “chromatin assembly complex”).

Select all nodes in the current view.

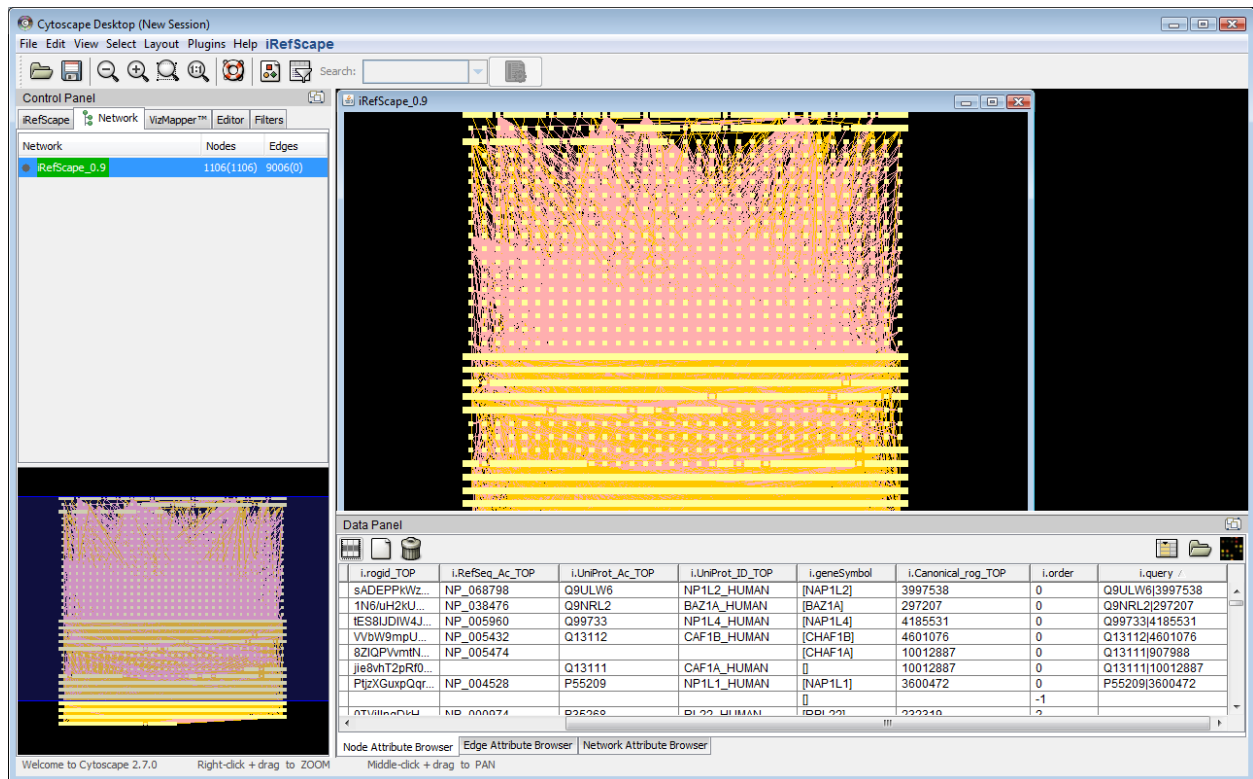
Retrieve their neighbours using iRefIndex menu/Search tools/Retrieve interactions for selected nodes.

You will see (a ridiculously large network – 1106 nodes and 9006 edges) like this:



Select all nodes using control-A.

In the node attribute browser, select to view the i.query node feature and then sort on this feature by clicking on the i.query column heading. You should see all the nodes from your initial search at the top of the node attribute browser. Like this:



The screenshot shows the Cytoscape Desktop interface. The main window displays a network visualization with nodes and edges. The Control Panel on the left shows the network details for 'iRefScale_0.9'. The Data Panel at the bottom displays a table of node attributes, sorted by the 'i.query' column. The table has the following columns: i.rogid_TOP, i.RefSeq_Ac_TOP, i.UniProt_Ac_TOP, i.UniProt_ID_TOP, i.geneSymbol, i.Canonical_rrog_TOP, i.order, and i.query. The table contains several rows of data, with the top entries highlighted in yellow.

i.rogid_TOP	i.RefSeq_Ac_TOP	i.UniProt_Ac_TOP	i.UniProt_ID_TOP	i.geneSymbol	i.Canonical_rrog_TOP	i.order	i.query
sADEPPKWz...	NP_068798	Q9ULW6	NP1L2_HUMAN	[NAP1L2]	3997538	0	Q9ULW6 3997538
1N6/uH2KU...	NP_038476	Q9NRL2	BAZ1A_HUMAN	[BAZ1A]	297207	0	Q9NRL2 297207
tES8LDIW4J...	NP_005960	Q99733	NP1L4_HUMAN	[NAP1L4]	4185531	0	Q99733 4185531
VvbW9mpU...	NP_005432	Q13112	CAF1B_HUMAN	[CHAF1B]	4601076	0	Q13112 4601076
8ZIQVvmtN...	NP_005474			[CHAF1A]	10012887	0	Q13111 907988
jie8vhT2pRf0...		Q13111	CAF1A_HUMAN	[]	10012887	0	Q13111 10012887
PltzXGuxpQqr...	NP_004528	P55209	NP1L1_HUMAN	[NAP1L1]	3600472	0	P55209 3600472
ATWllk+DkH...	NP_000674	P35289	NP1L1_HUMAN	[]	3600472	-1	

Left-click and drag over these table entries to select them with the mouse.

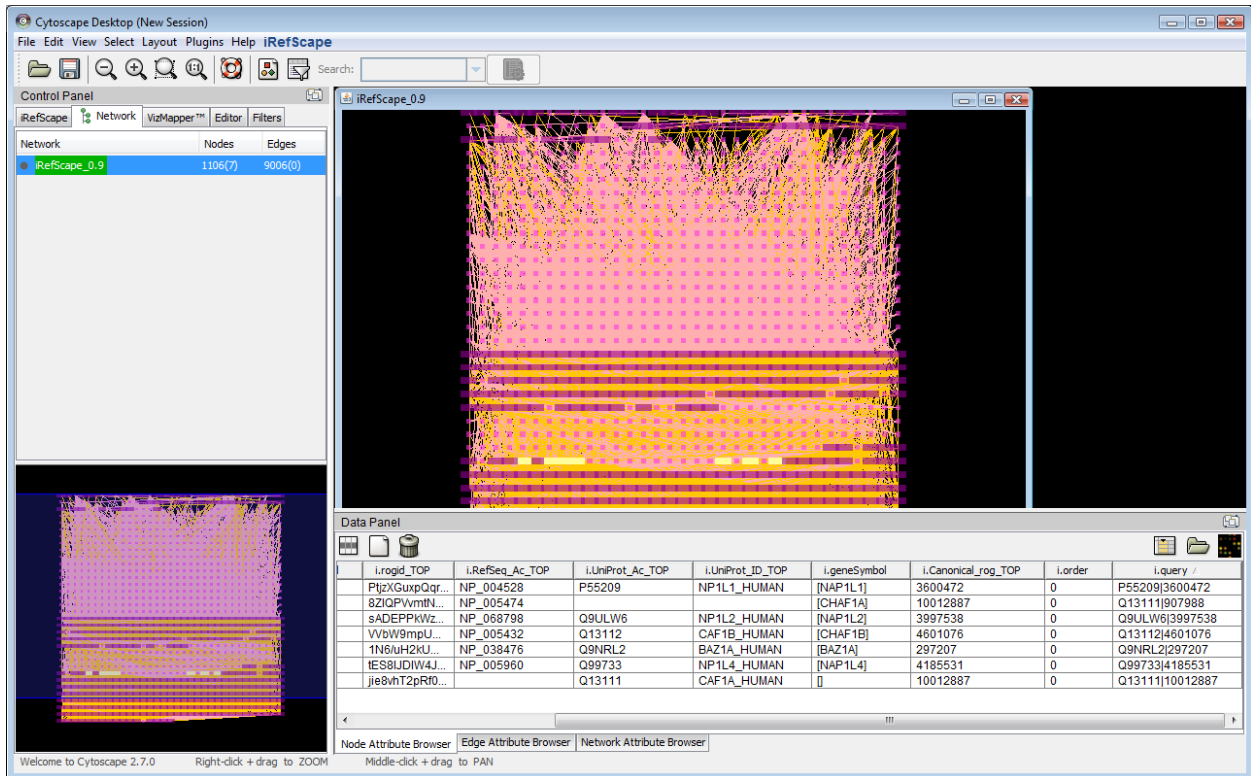
The right-click on one of the hi-lited entries and choose "Select from table".

The screenshot shows the Cytoscape Desktop interface with a network visualization of iRefScapE_0.9. The Control Panel on the left shows the network details: 1106 nodes and 9006 edges. The main visualization window displays a dense network of nodes and edges, with a zoomed-in view of a specific cluster. The Data Panel at the bottom shows a table of node attributes.

i.rogid_TOP	i.RefSeq_Ac_TOP	i.UniProt_Ac_TOP	i.UniProt_ID_TOP	i.geneSymbol	i.Canonical_rog_TOP	i.order	i.query /
sADEPPKwz...	NP_068798	Q9ULW6	NP1L2_HUMAN	[NAP1L2]	3997538	0	Q9ULW6 3997538
1N6/uH2KJ...	NP_038476	Q9NRL2	BAZ1A_HUMAN	[BAZ1A]	297207	0	Q9NRL2 297207
tES8uDIW4J...	NP_005960	Q99733	NP1L4_HUMAN	[NAP1L4]	4185531	0	Q99733 4185531
VvbW9mpU...	NP_005432	Q13112	CAF1B_HUMAN	[CAF1B]	4601076	0	Q13112 4601076
8ZIQPVmtN...	NP_005474			[CAF1A]	10012887	0	Q13111 907988
jie8vhT2pRI0...		Q13111	CAF1A_HUMAN	[]	10012887	0	Q13111 10012887
PltzXGuxpQqr...	NP_004528	P55209	NP1L1_HUMAN	[NAP1L1]	3600472	0	P55209 3600472
				[]		-1	
ATUllwDdW	NP_000074	P28268	BI-22_HUMAN	[BI-22]	222340	2	

At the bottom of the interface, there are tabs for "Node Attribute Browser", "Edge Attribute Browser", and "Network Attribute Browser". A search bar labeled "Lquery" is visible on the right side of the Data Panel, containing the text "P55209|3600472".

After this operation, only the original nodes directly returned by the query will appear in the node attribute browser.



Then go to the iRefScape menu/View tools/Select between nodes.

Use the node attribute browser to select all nodes that are either

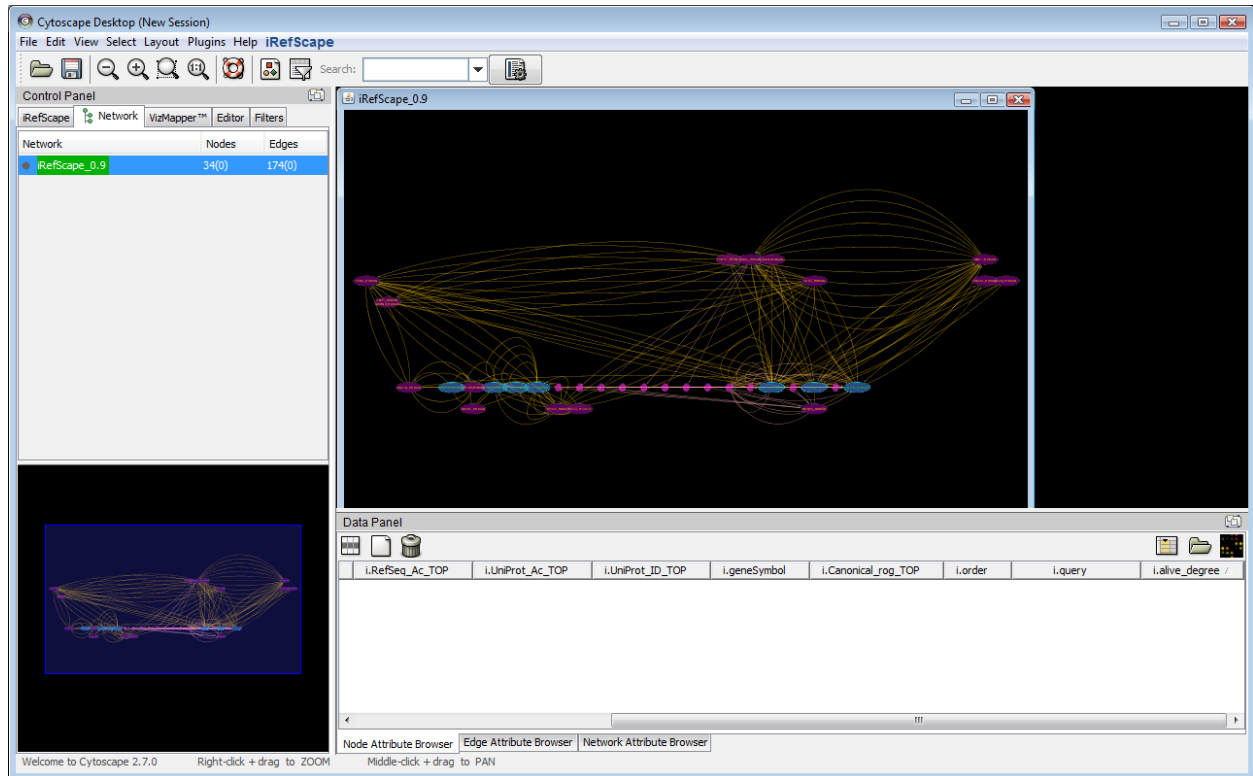
- 1) Returned by the original query (see i.query node feature) or
- 2) Have an i.alive_degree of 2 or more (see the i.alive_degree node feature)

This may require a live demo. Ask if you have problems.

Then select all other nodes and delete them. Use Select/Nodes/Invert node selection

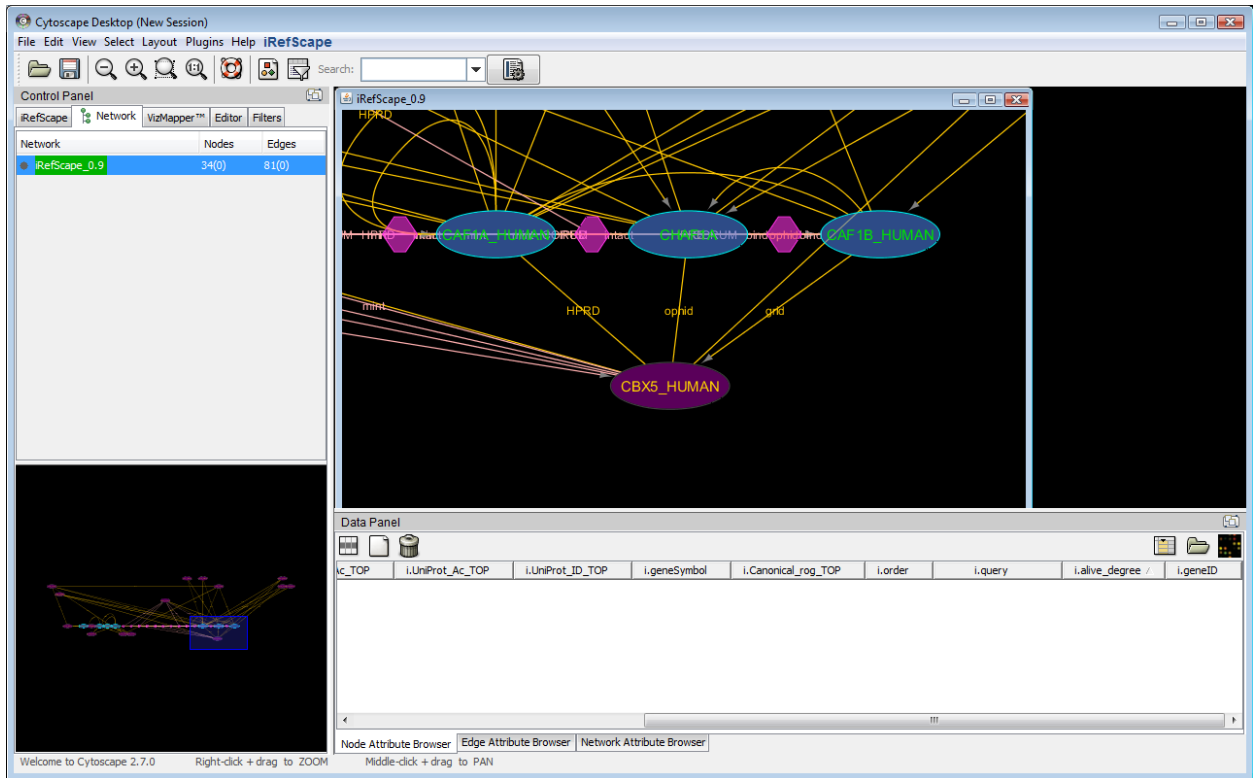
And then hit delete.

You should see something like this.



These new “between nodes” are connected to two or more nodes from your original search list (seed list). Guilt-by-association makes these nodes candidates to look at for things that may be related to the “chromatin assembly complex”. The higher the `i.alive_degree`, the better.

For example, `CBX5_HUMAN` interacts with `CAF1A`, `CAF1B` and `CHAF1A` from our original search.



A brief review of the Entrez Gene record for CBX5 (Entrez Gene ID 23468) shows that there is evidence to support this connection. It is left as an exercise to review the papers that support evidence of an interaction between CBX5 and each of the three original query proteins.

If you have time left, you can use the techniques described in this tutorial to investigate another GO term and its associated genes that are of specific interest to you.

