

## PomBase advanced search

The advanced search has several filtering options that allow you to find all genes annotated to terms of interest.

Let's start by finding all genes annotated to "DNA repair":

- Click on the GO filter
- Start typing 'DNA repair', select this term from the dropdown list
- Click submit

The results appear at the bottom of the search interface.

### Advanced search

New query  
Commonly used queries  
Gene IDs

GO  
Phenotype  
Product type  
Protein modification  
Protein domain  
Protein feature  
Protein length  
Protein mol. weight  
Disease  
Number of TM domains  
Genome location  
Number of exons  
Taxonomic conservation  
Characterisation status

**Retrieve genes via annotated Gene Ontology terms the location of their products (protein or ncRNA):**

DNA repair

The process of restoring DNA after damage. Genomes are subject to damage in their natural environment (e.g. UV and ionizing radiations, chemical mutagens, fungal alkylating agents endogenously generated in metabolism. DNA is also damaged by a variety of different DNA repair pathways have been reported that include base excision repair, photoreactivation, bypass, double-strand break repair pathways.

Submit

### Combine queries:

Union / or  Intersect / and  Subtract / not  Delete

	Results	Query history
<input type="checkbox"/>	183	DNA repair (GO:0006281)

Clicking on the hyperlinked results number takes you to a page where you can view and download the results. Clicking on the "product" description header sorts the genes on their description, which can be handy when scanning a long list of genes.

Result: 183 genes

Results for: DNA repair (GO:0006281)

Visualise

Slim

Select subset

Download ...

Gene name	Systematic ID	Product
<a href="#">pfh1</a>	<a href="#">SPBC887.14c</a>	5' to 3' DNA helicase Pif1/Pfh1
<a href="#">myh1</a>	<a href="#">SPAC26A3.02</a>	adenine DNA glycosylase Myh1
<a href="#">atl1</a>	<a href="#">SPAC1250.04c</a>	alkyltransferase-like protein Atl1
<a href="#">apn1</a>	<a href="#">SPCC622.17</a>	AP endonuclease, minor transcript isoform Apn1
<a href="#">apn2</a>	<a href="#">SPBC3D6.10</a>	AP-endonuclease Apn2
<a href="#">hnt3</a>	<a href="#">SPCC18.09c</a>	aprataxin Hnt3
<a href="#">ath2</a>	<a href="#">SPAC139.01c</a>	Ath1 complex protein Ath2 nuclease, XP-G family
<a href="#">tel1</a>	<a href="#">SPCC23B6.03c</a>	ATM checkpoint kinase
<a href="#">fml2</a>	<a href="#">SPAC20H4.04</a>	ATP-dependent 3' to 5' DNA helicase (predicted)
<a href="#">fml1</a>	<a href="#">SPAC9.05</a>	ATP-dependent 3' to 5' DNA helicase, FANCM ortholog Fml1
<a href="#">chl1</a>	<a href="#">SPAC3G6.11</a>	ATP-dependent DNA helicase Chl1 (predicted)
<a href="#">rdh54</a>	<a href="#">SPAC22F3.03c</a>	ATP-dependent DNA helicase Rdh54
	<a href="#">SPBC582.10c</a>	ATP-dependent DNA helicase Rhp16b (predicted)
<a href="#">srs2</a>	<a href="#">SPAC4H3.05</a>	ATP-dependent DNA helicase, UvrD subfamily

The '**visualise**' tool lets you explore your gene list based on a number of criteria. In this gene list we can see, for example, that most localize to the nucleus and do not have transmembrane domains. There is a sorting function on the left hand side to sort categories of interest. Clicking on a segment will show you the list of genes in that segment.

## Visualising 183 genes

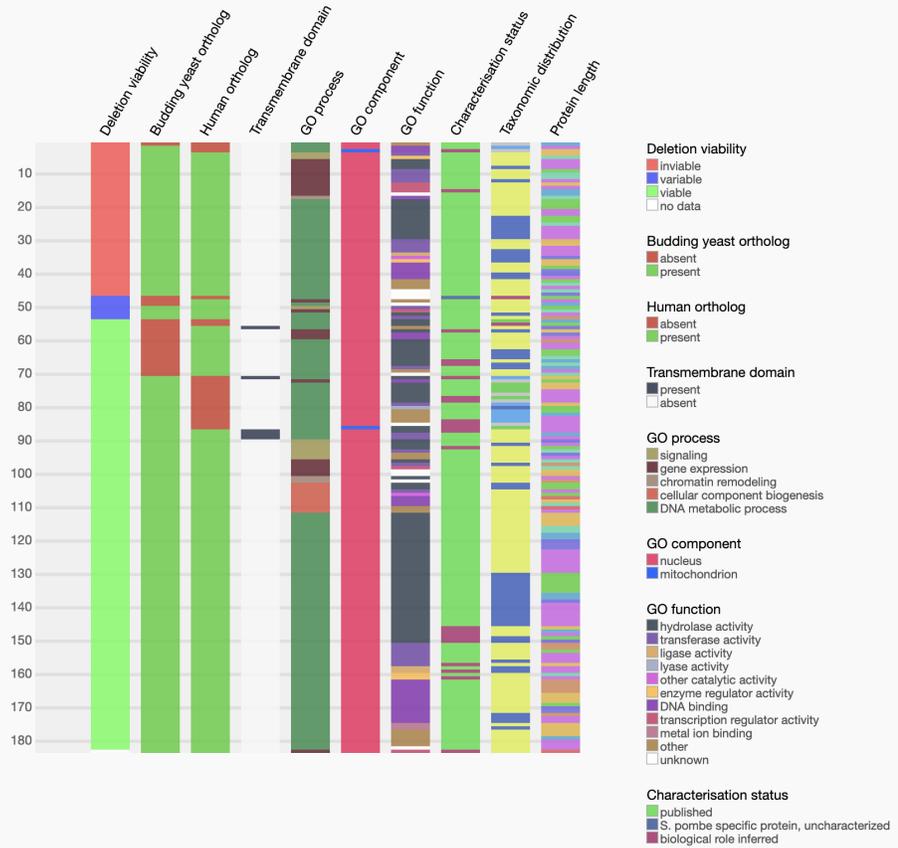
Visualisation for: DNA repair (GO:0006281)

Finish v

- Deletion viability:  [sort]
- Budding yeast ortholog:  [sort]
- Human ortholog:  [sort]
- Transmembrane domain:  [sort]
- GO process:  [sort]
- GO component:  [sort]
- GO function:  [sort]
- Characterisation status:  [sort]
- Taxonomic distribution:  [sort]
- Protein length:  [sort]

[Documentation](#)

[Download image ...](#)



The slim tool will show you how your gene list distribute in the *S. pombe* GO slim.

## S. pombe high level GO biological process terms

### Results for: DNA repair (GO:0006281)

A "GO slim" is a subset of the Gene Ontology terms selected for a specific purpose in interpreting the functional annotations in an entire organism, or a gene product list derived from an experiment. PomBase has created a GO slim to provide a simple summary of S. pombe's biological capabilities by grouping gene products using broad biological process classifiers.

### Terms and genes

Name	Term	Genes
actin cytoskeleton organization	GO:0030036	1
ascospore formation	GO:0030437	1
chromatin organization	GO:0006325	22
conjugation with cellular fusion	GO:0000747	2
DNA recombination	GO:0006310	71
DNA repair	GO:0006281	183
lipid metabolic process	GO:0006629	1
microtubule cytoskeleton organization	GO:0000226	2
mitochondrion organization	GO:0007005	7
mitotic cytokinesis	GO:0000281	1
mitotic sister chromatid segregation	GO:0000070	10
mRNA metabolic process	GO:0016071	2
nucleocytoplasmic transport	GO:0006913	1
protein catabolic process	GO:0030163	14
protein modification by small protein conjugation or removal	GO:0070647	21

Now use the advanced term to find all genes annotated to the GO term "nucleus".

The advanced search has three different operators that allow you to combine queries using union (AND), intersect (OR), and subtract (NOT).



If you were to combine the two searches "DNA repair" and "nucleus", you could combine these two searches in the following way:

- OR - you get a list of all gene products involved in DNA repair plus all gene products that localize to the nucleus
- AND - you get a list of all genes involved in DNA repair that also localize to the nucleus (e.g. would exclude those that exclusively are involved in mitochondrial DNA repair)
- NOT - you get a list involved in DNA repair that do not localize to the nucleus (check the direction that you run the query in e.g. 'DNA repair NOT nucleus' - the query 'nucleus NOT DNA repair' will get you a different result.
  - At the moment (in PomBase) you cannot force the direction so for "query A minus query B" query A needs to be the most recent query (higher up in the list) - to make a query jump in in the list you can click on the hyperlinked results number - this will make the query in question jump to the top of the list.

How many genes localize to the nucleus but are NOT involved in DNA repair?

## Enrichment analysis

There are many different tools for performing enrichment analysis. Because GO annotations are frequently updated (new inferences are added, and problematic ones are removed) it is important to ensure that the tool is up to date - some popular tools use GO data that is years out of date! In this exercise we are going to try three different tools:

1. The princeton enrichment tool. This tool allows you to upload your own annotation file (GAF) so can be used for species outside of the ones they provide direct support for.
2. G:Profiler, this tool supports several fungal species
3. Angeli. An *S. pombe* specific tool that allows you to enrich on phenotype annotations as well as GO annotations.

First we are going to use the advanced search to find genes that when mutated make cells more sensitive to staurosporine:

- Go to the PomBase advance search, <https://www.pombase.org/query>
- Using the phenotype filter, search the phenotype “sensitive to staurosporine” and click submit
- In the results section you can see that 16 genes were found. Click on the result and download the systematic IDs of these 16 genes

### **Princeton enrichment tool**

<https://go.princeton.edu/cgi-bin/GOTermFinder>

Paste the gene list into the box at the top and select the *S. pombe* annotation set from the drop down list. Enrich for GO Biological Process terms and inspect the list of results.

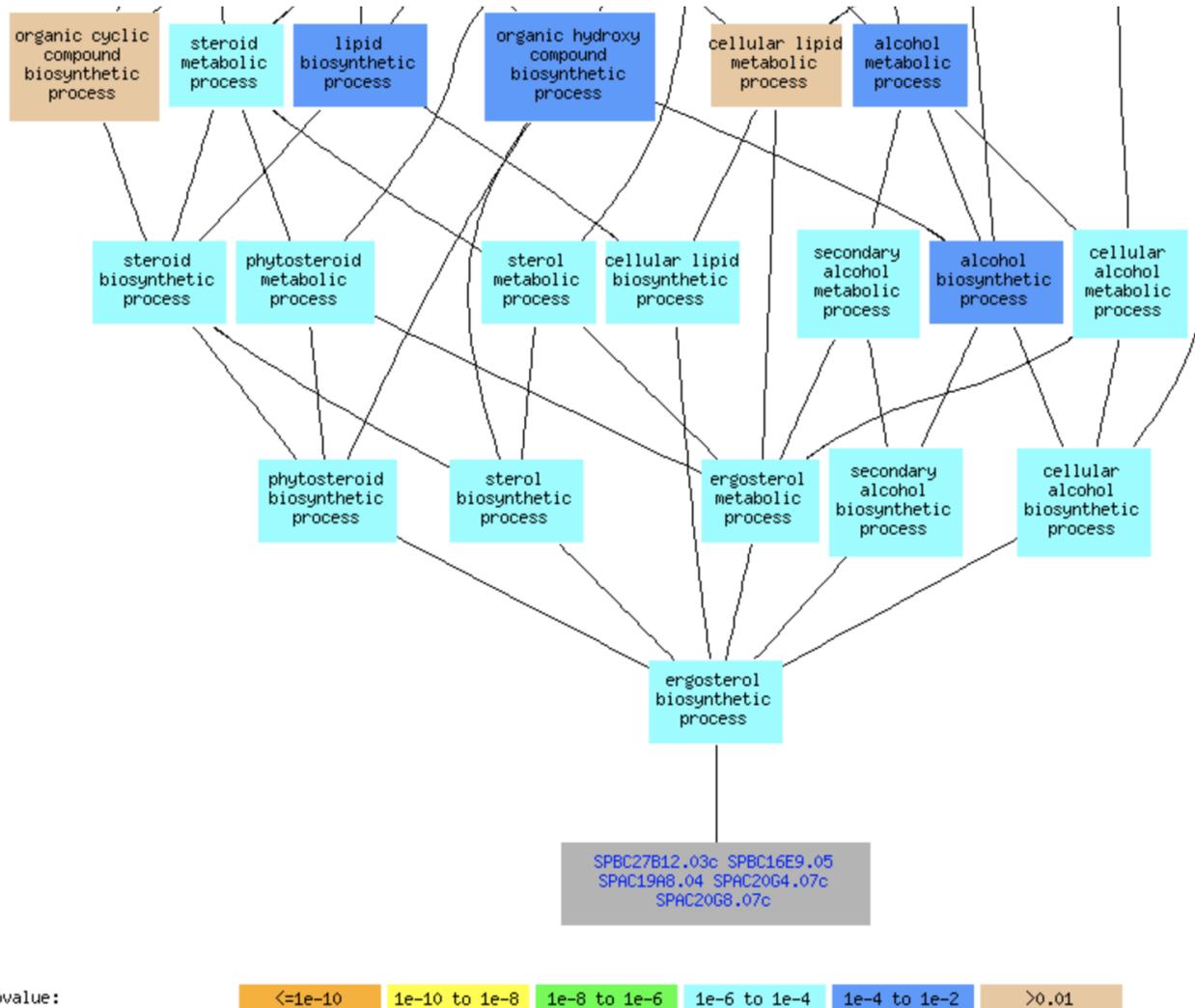
The results table shows you what genes are annotated to the enriched terms, P values and the their relative frequencies

- The cluster frequency specifies the number and % of genes in your list annotated to a term
- The genome frequency specifies the number and % of genes in the genome annotated to the term

What is the cluster frequency and genome frequency of the term ‘ergosterol biosynthetic process’?

Cluster frequency/genome frequency tells you the proportion of genes in your list annotated to the term, compared to the genome overall - 5/38 genes annotated to ‘ergosterol biosynthetic process’ are present in the list.

The result of an enrichment analyses can sometimes look ‘cluttered’. For example, in this list of enriched terms, both ‘steroid metabolic process’ and ‘steroid biosynthetic process’ appear. An ontology “tree view” can be used to visually identify terms of interest (more specific terms are shown at the bottom of the graph). In the Princeton tool this view is displayed at the bottom of the results page.



**G:profiler**

Paste the same gene list into G:profiler and select *S. pombe* from the organism filter. Click the tab for detailed results.

G:profiler only show you the P values, not the cluster or genome frequencies.

You can calculate those using the detailed stats information.

In the GO:BP section, click the >> to the right of ‘stats’.

Q corresponds to number of genes in list

U corresponds to the number of genes in the genome

T corresponds to number of genes annotated to the term in the genome

TnQ corresponds to the number of genes in the list annotated to the term

Using the numbers from this tool, calculate the cluster frequency ( $TnQ/Q$ ) and genome frequency ( $T/U$ ) for 'ergosterol biosynthetic process'.

Thus the cluster frequency ( $TnQ/Q$ ) of 'ergosterol biosynthetic process' is 5/16

The genome frequency ( $T/U$ ) for 'ergosterol biosynthetic process' is 43/5052

- This differs from the numbers obtained using the Princeton tool! The Princeton tool gets its GO data from GO and uses PomBase genes with 'frequent' updates, whereas G:profiler gets the data from Ensembl (and is updated quarterly).

G:profiler doesn't provide a tree view of the enriched terms. If you are interested in manually inspecting the relationship between terms you can use the EBI QuickGO tool, it has a handy "basket" function allowing you to search for terms and see them in an ontology tree view.

Go to <https://www.ebi.ac.uk/QuickGO/>

Click on the "basket" menu item in the top menu, and paste in the GO IDs

GO:0006696

GO:0006694

GO:0008610

GO:0006066

GO:0046165

These correspond to the GO terms

ergosterol biosynthetic process

steroid biosynthetic process

lipid biosynthetic process

alcohol metabolic process

alcohol biosynthetic process

Click this symbol



The ontology tree view is displayed, showing how terms are connected in the ontology. The terms entered are shown in yellow.

Paste the same gene list into Angeli

[http://bahlerweb.cs.ucl.ac.uk/cgi-bin/GLA/GLA\\_input](http://bahlerweb.cs.ucl.ac.uk/cgi-bin/GLA/GLA_input)

- select to enrich for Biological Process terms (you can deselect the other categories, they make the query slower to run)
- click “analyze gene list”

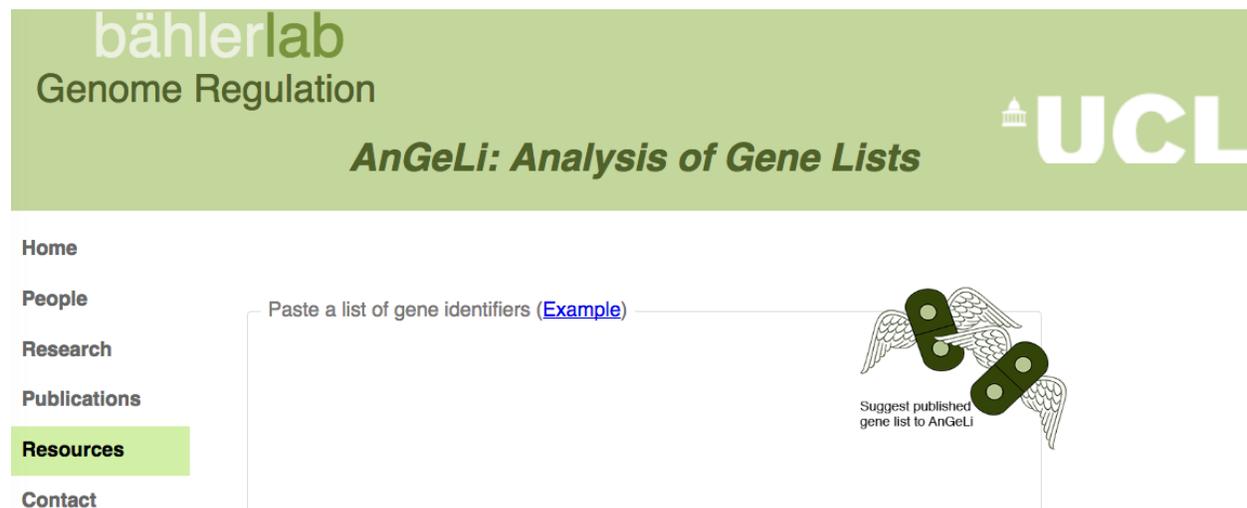
## **Angeli**

Lets use a different gene list.

Using the PomBase advanced search, select the “GO” filter and search for the GO term GO:0007005 mitochondrion organization, click submit.

Click on the hyperlinked number in the results list and download the systematic IDs.

Go to the gene list analyzer tool AnGeLi [http://bahlerweb.cs.ucl.ac.uk/cgi-bin/GLA/GLA\\_input](http://bahlerweb.cs.ucl.ac.uk/cgi-bin/GLA/GLA_input)



- Paste the IDs of your gene list into the box
- Select to enrich for phenotypes (you can deselect the other categories, they make the query slower to run)
- click “analyze gene list”

Which phenotype term is at the top of list of enriched phenotypes?

Go back to the PomBase advanced search. Using the phenotype filter, find all genes annotated to the 'tapered cell' phenotype. How many of these (around 143) also localize to the mitochondrion?

(hint: Use the GO filter and search for mitochondrion)

Q6: how many genes annotated to this phenotype localize to mitochondrion?

(hint: search for GO term mitochondrion, tick the two queries (the one for mitochondrion and the one for the phenotype) in the results list and use the intersect/and button to intersect the queries - you should get around 137 genes).