

## AnGeLi trouble shooting / frequently asked questions

**Q: I would like to learn more about AnGeLi**

A: Read the help file that starts on page 3 of this document.

**Q: Web interface does not look pretty or does not work at all.**

A: Install Firefox, Chrome or Safari web browsers

**Q: I did not get any results.**

A: Use only systematic gene identifiers.  
Choose at least one category.  
Paste only one gene per line.  
Don't use gene names in quotes.

**Q: AnGeLi is very slow.**

A: You will have to wait patiently if all categories have been selected. AnGeLi performs thousands of statistical tests (1277 GO Molecular Function, 797 GO Cellular Component and 4 Genetic and Physical Interactions (BioGRID) features are excluded by default). Average searches should take approximately 2 minutes if all categories are selected. If GO and Phenotypes are de-selected, the analysis takes about 1 minute. These are only rough estimates, and times will vary depending on the number of genes in the query list, the number of user-defined gene sets uploaded, and the number of permutations selected.

We suggest selecting fewer categories. GO categories, Phenotypes, and Protein Domains are the most time consuming.

**Q: I do not know what the results mean.**

A: Use the explain button to get some information on the calculation.  
Follow the link to the feature (if given) or read the cited paper.  
Email your input gene list, a screen copy of the input screen and a description of the problem to [d.bitton@ucl.ac.uk](mailto:d.bitton@ucl.ac.uk).

**Q: I would like to contribute a gene list to AnGeLi**

A: That is a great idea. If your work is published, you could submit your list via the online form: [http://bahlerweb.cs.ucl.ac.uk/cgi-bin/GLA/Angeli\\_submission](http://bahlerweb.cs.ucl.ac.uk/cgi-bin/GLA/Angeli_submission) or by clicking on AnGeLi's logo.



Note that you can only submit gene lists via the form, for any numerical data please contact [d.bitton@ucl.ac.uk](mailto:d.bitton@ucl.ac.uk).

**Q: What is the background distribution? Can I change the background distribution?**

A: By default, AnGeLi uses all protein-coding genes from Pombase/Ensembl as its background. However, AnGeLi also includes 5 additional options: non-coding genes only, all annotated genes, all genes with associated GO- or phenotypic terms, and non-essential genes. Alternatively, you can upload your own background file. AnGeLi accepts only text files (txt), using one systematic ID per line.

**Q: `Can I search my query list against my own gene sets?**

A: Yes, AnGeLi offers full user flexibility. You can upload your own gene sets as shown in this example: [http://bahlerweb.cs.ucl.ac.uk/Example\\_Gene\\_Set\\_File.txt](http://bahlerweb.cs.ucl.ac.uk/Example_Gene_Set_File.txt) . Note that AnGeLi only accepts text (txt) files. You can also de-select all of AnGeLi's pre-defined gene sets and utilise the statistical framework to search your own gene sets only.

**Q: I uploaded background or customised gene sets, but AnGeLi does not work or does not recognise the file**

A: Ensure that the file is in the correct format. We suggest the following:

- Ensure that the file is not open in another programme.
- To ensure that the correct format is used, open the file in Excel and save it as Windows formatted text (txt)
- If the problem persists, email your files to [d.bitton@ucl.ac.uk](mailto:d.bitton@ucl.ac.uk)

**Q: Can I export my results?**

A: Yes, you can do this by using the export button.

**Q: When I copy and paste my results into Excel, data are not formatted as a table.**

A: Use paste special and paste it as text. Alternatively, you could simply right click while on the browser and save the exported page (Angeli\_result\_tsv) as a text file.

**Q: I do not understand what some gene sets mean.**

A: Follow the link provided in the first column of the results table, it will take you to the relevant paper or database.

**Q: How do I cite AnGeLi?**

A: Please cite the following paper:

## AnGeLi help

AnGeLi is a web tool that enables a comprehensive analysis of a list of *S. pombe* genes. AnGeLi is freely accessible from the following URL: <http://bahlerlab.info/AnGeLi>


### Input screen

AnGeLi: Analysis of Gene Lists

A

Paste a list of gene identifiers ([Example](#))

```
SPCC1494.07
SPCC1672.05c
SPCC1672.11c
SPCC18B5.01c
SPCP25A2.02c
SPCC1281.06c
SPCC1450.16c
```



Suggest published gene list to AnGeLi

Set parameters below:

B

Choose a pre-defined background

Protein-coding genes

C

OR: upload your own background file (one systematic ID per line)

Browse... No file selected.

D

Optional: upload files containing additional gene sets to compare against ([format](#))

Browse... No file selected.

E

Choose Gene ontology category

Biological Process
  Cellular Component
  Molecular Function
  None

F

Adjust method for multiple testing

FDR
  Holm
  Bonferroni
  None (not recommended)

G

Show results with a p-values smaller than

0.01

H

Perform pairwise interactions enrichment analysis (slow)

No
  Yes

Select Yes to enable the options below:

Set number of permutations (for pairwise interactions)

1000

Adjust p-value (for pairwise interactions)

Yes
  No

#### A: Gene identifiers

The gene identifiers are systematic names from PomBase (<http://www.pombase.org/>). Gene identifiers that are not (yet) known to AnGeLi will be excluded from the analysis and shown.

#### B-C: Pre-defined and user-defined background

The background gene population is the reference for the statistical analyses of your query gene list. Here you have 6 options: use the default (all annotated *S. pombe* genes), only protein-coding or only

ncRNA transcripts, genes associated with GO or phenotypes. Alternatively you could upload your own background file (C); follow the guidelines and upload in the right format.

#### **D: User-defined gene sets**

Here you can upload as many gene sets as you like to also compare against your query list: follow the guidelines and upload the right format, AnGeLi will do the rest.

#### **E: GO category**

Here you can choose which GO category to search against: GO Biological Process (default, 3516 features), GO Molecular Function (1275 features), GO Cellular Component (797 features) or none of them.

#### **F-G: Options for multiple testing adjustment**

**F:** AnGeLi can be used to test a large number of features against a gene list. Given a (standard) alpha of 0.05 and random data, the probability that a single test will give a significant result is 5%. An analysis of more than 20 features given random data will therefore likely return false positive results. The easiest way to deal with this problem is to perform each test at an alpha of 0.05 divided by the number of tests (Bonferroni correction). The drawback of this solution is that subtle differences in a feature are no longer detectable. The Holm procedure has slightly better power and still corrects the familywise error rate (FWER). The False Discovery Rate (FDR) is less conservative and has a greater power, and is therefore the default option. The basic idea is to control the expected proportion of false positive results given all positive results. Given a FDR of 0.01 and a list of 100 significant features, the expected number of wrongly called features would be 1.

**G:** The result screen of AnGeLi will show only features that have a corrected p-value of less than 0.01 (default).

#### **References**

- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289–300.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, **46**, 561–576.

#### **H: Options for pairwise interactions enrichment**

The following paragraphs require a basic understanding of statistical methods. However, you can stick to the default values if you are unsure. The underlying test of pairwise features is a so-called permutation test. Let us start with an example. Imagine you have posted 10 gene IDs. The number of protein-protein-interactions within this list is seven. Next, the number of protein-protein-interactions is determined in 500 lists of randomly chosen genes and length 10. Genes for these random lists are drawn from the set of all genes minus the genes in the gene list. The number of permutations determines how many random lists are generated. Generating thousands of random lists will slow down the analysis.

On the other hand, the number of permutations affects the p-value. The lowest p-value given 1000 permutations and a two-sided test is 0.002. A p-value of 0.002 may not be small enough to be detectable in a multiple testing setting. The default is set to 1000 permutations. If you would like to adjust for multiple tests, use the dropdown menu and select 10,000 or 100,000. Note that this relatively high p-value is unlikely to be significant following correction for multiple testing (7422 tests), therefore we suggest setting a higher number of permutations on the expense of analysis time which could be potentially combined with de-selection of categories that are not of interest to increase the statistical power.

#### **Reference**

- Good, P. (2000): *Permutation Tests. A Practical Guide to Resampling Methods for Testing Hypotheses*.

Select/deselect categories against which to compare your gene list:

**GO Biological Process**

- mitochondrion inheritance
- single strand break repair
- maltose metabolic process
- mitochondrial genome maintenance
- regulation of DNA recombination
- ribosomal large subunit assembly
- reproduction
- regulation of mitotic recombination
- ribosomal small subunit assembly
- vacuole inheritance
- mitotic spindle elongation
- cell wall mannoprotein biosynthetic process

[Show More](#)

**GO Cellular Component**

- phosphopyruvate hydratase complex
- histone deacetylase complex
- transcription factor TFIIIB complex
- nucleotide-excision repair complex
- RNA polymerase I transcription factor complex
- transcription factor TFIIIC complex
- nucleotide-excision repair factor 1 complex
- histone acetyltransferase complex
- polarisome
- nucleotide-excision repair factor 3 complex
- SAGA complex
- alpha-1,6-mannosyltransferase complex

[Show More](#)

**GO Molecular Function**

- alpha-1,6-mannosyltransferase activity
- adenine deaminase activity
- S-adenosyl-L-methionine transmembrane transporter activity
- single-stranded DNA endodeoxyribonuclease activity
- peptidyltransferase activity
- sulfur amino acid transmembrane transporter activity
- alpha-1,2-mannosyltransferase activity
- tRNA binding
- succinate dehydrogenase activity
- mannosyltransferase activity
- fatty-acyl-CoA binding
- microfilament motor activity

[Show More](#)

**Gene Expression**

- Caffeine and Rapamycin induced
- Induced in pab2 mutant splicing efficiency
- Repressed pat1.114 mutant
- Caffeine and Rapamycin repressed
- Induced Pab2 mutant exon intron signal
- Repressed diploid cells
- Induced Dbr1 deletion
- Induced pat1.114 mutant
- Scw1 RNA targets
- Repressed Dbr1 deletion
- Induced diploid cells
- Zfs1 RNA targets

[Show More](#)

**Gene Features**

- Number of introns
- Intron-containing genes
- Chromosome 3
- Average intron length
- Intron-less genes
- Mitochondria
- Length of first intron
- Chromosome 1
- Abs. distance from telomere
- GC contents of first intron
- Chromosome 2
- Abs. distance from centromere

[Show More](#)

## I: AnGeLi curated gene sets

Here you can simply choose against which of our curated gene sets your query list will be searched. Simply follow the link to the corresponding paper or the external database to learn more about each feature. You could deselect any general theme, and AnGeLi even allows deselecting all themes and searching only against your own curated gene sets.

## Processing screen

Please wait...Loading files...

### AnGeLi: Analysis of Gene Lists

Preparing data...

A total of 120 gene identifiers were considered in the query list and 5134 in the background population...

No user user-defined gene sets were uploaded...

Performing Fisher's exact tests ...

Performing Wilcoxon tests ..

Performing Permutation tests ...

As soon as you click the 'Analyse gene list' button, AnGeLi will remind you of the settings you chose and report the stages of the analysis. In this example, 120 genes were in the query list and searched against a customised background list of 5134 genes as well as against a single user-defined gene set.

## Results screen

AnGeLi: Analysis of Gene Lists				
Export Results		Perform new analysis		
GO Biological Process	List frequency	Background frequency	Details	p- value
ncRNA metabolic process	19.17% (23/120)	6.29% (323/5134)	<a href="#">Explain!</a>	0.000783083
ncRNA processing	15.83% (19/120)	5.30% (272/5134)	<a href="#">Explain!</a>	0.006891
Gene Expression	List frequency	Background frequency	Details	p- value
Core Environmental Stress Response repressed	27.50% (33/120)	8.55% (439/5134)	<a href="#">Explain!</a>	8.86348e-07
Caffeine and Rapamycin repressed	19.17% (23/120)	5.92% (304/5134)	<a href="#">Explain!</a>	0.00029773
Repressed Dbr1 deletion	12.50% (15/120)	3.23% (166/5134)	<a href="#">Explain!</a>	0.00330685
Phenotypes (FYPO)	List frequency	Background frequency	Details	p- value
inviable spore	23.33% (28/120)	9.15% (470/5134)	<a href="#">Explain!</a>	0.00139038
Protein Features	List frequency	Background frequency	Details	p- value
Molecular weight (kDa)	78.892	51.7569	<a href="#">Explain!</a>	1.50347e-09
Number of amino acids	702.225	458.819	<a href="#">Explain!</a>	1.50347e-09
<i>S.cerevisiae</i> orthologs	91.67% (110/120)	75.89% (3896/5134)	<a href="#">Explain!</a>	0.00598098
Transcript Features	List frequency	Background frequency	Details	p- value
mRNA copies per proliferating cell	10.5293	8.08127	<a href="#">Explain!</a>	1.50347e-09
Ribosomal density (rb/kb)	2.85601	4.54688	<a href="#">Explain!</a>	5.45616e-07
mRNA level (WT)	4195.57	2767.15	<a href="#">Explain!</a>	2.74408e-06
mRNA copies per quiescent cell	1.7416	1.45623	<a href="#">Explain!</a>	2.74408e-06
Annotated transcript length	2769.75	2133.58	<a href="#">Explain!</a>	2.26978e-05
Relative Pol II occupancy	1.54699	1.33573	<a href="#">Explain!</a>	0.000783083

The Results screen of AnGeLi shows all significant features in a table. Results can be exported as a tab delimited file. A feature is considered as significant if its corrected p-value is lower than the threshold (default 0.01). The underlying options are summarised in the status line (H). All features are ordered according to their p-value in ascending order. If no significant feature has been found in a theme, it will not be shown.

The colour of a binary feature reflects whether a feature is enriched (red) or under-enriched (green) in the gene list compared to the background population. In the second column, the percentage of genes with this feature in the gene list is given along with the actual values. It also shows the proportion of genes within this feature in the background population.

### Example:

Protein Features	List frequency	Background frequency	Details	p- value
<i>Homo sapiens</i> orthologs	100.00% (173/173)	66.15% (3396/5134)	<a href="#">Explain!</a>	2.21615e-29
<i>S.cerevisiae</i> orthologs	61.85% (107/173)	75.89% (3896/5134)	<a href="#">Explain!</a>	0.00381183

In this case, 100% of all genes in the list (173/173) are human orthologs compared to 66.15% (3396/5134) of genes in the background population. The term *Homo sapiens* orthologs is therefore significantly enriched within the gene list. On the other hand, *S. cerevisiae* orthologs are under-enriched.

For numeric features, red shows that a feature is higher in the gene list compared to all other genes and green that a feature is lower in the gene list compared to all other genes. The fourth column shows the average of the feature within the gene list and the sixth column the average of the feature given all other genes.

**Example:**

Transcript Features	List frequency	Background frequency	Details	p- value
mRNA copies per proliferating cell	10.5293	8.08127	<a href="#">Explain!</a>	1.50266e-09
Ribosomal density (rb/kb)	2.85601	4.54688	<a href="#">Explain!</a>	5.45322e-07
mRNA level (WT)	4195.57	2767.15	<a href="#">Explain!</a>	2.7426e-06
mRNA copies per quiescent cell	1.7416	1.45623	<a href="#">Explain!</a>	2.7426e-06
Annotated transcript length	2769.75	2133.58	<a href="#">Explain!</a>	2.26855e-05
Relative Pol II occupancy	1.54699	1.33573	<a href="#">Explain!</a>	0.000782661

The average mRNA copies per proliferating cell of all genes within the gene list is 10.5293 and significantly higher than the average mRNA copies per proliferating cell of all other genes (8.08127).

**Explain button**

If a feature is significant within your list, you can hit the explain button. It provides more information about the underlying feature and the statistical test. Furthermore, it shows all the genes of interest (with links to PomBase).

For example,

- It shows all interaction pairs within your gene list if an interaction is significant
- It shows all transcription factors within your list if the term transcription factor is enriched within your gene list
- It shows all genes with long ORFs if the average ORF length is significantly higher within your gene list.

**P-value and warnings**

Each feature is analysed using an appropriate statistical test:

- Binary feature: two-sided Fisher's exact test
- Metric feature: two-sided Wilcoxon rank sum test
- Pairwise feature: two-sided permutation test.

The number of tests performed is given in the status line. For example, 7545 statistical tests (49 two-sided Wilcoxon rank sum tests, 7505 two-sided Fisher's exact tests) have been performed, corrected according to FDR and evaluated given an alpha of 0.01.

The p-values of the tests are sometimes only approximately correct; more information about the approximation of a p-value may be given next to it. A star implies an approximation problem whereas an exclamation mark refers to a serious problem (like an empty gene list or a run-time error). The Wilcoxon rank test uses a normal approximation with a continuity correction or an exact test (depending on the number of permutations and p-value estimation). A warning will be shown for very short gene lists and p-values based on the normal-approximation.