1 Group Task 1 Answers

Given an RNA-Seq experiment with a knocked out gene, you were asked to answer the following questions:

- **What is the name of the knock out gene?**
- **What influence does it have?**
- **How did you determine those?**

1.1 Dataset

You were given the following files in order to conduct the analysis:

- Wild type sample reads in FASTQ format
  `WT[replicate]_[1|2].fastq.gz`
- Knockout sample reads in FASTQ format
  `KO[replicate]_[1|2].fastq.gz`
- *P. berghi* genome in FASTA format
  `PbANKA_v3.fasta`
- *P. berghi* transcripts in FASTA format
  `Pb.CDS.fasta`
- *P. berghi* annotations in GFF format
  `PbANKA_v3.gff3.gz`
- *P. berghi* gene descriptions in TSV format
  `Pb.names.txt`
- R script to run sleuth
  `sleuth.R`

1.1.1 Important questions

Can you summarise the experimental design?

The experimental design should explain what each sample represents, i.e. the conditions that were applied and how many replicates there were. In this experiment, there are two conditions, wild type (WT) and knock out (KO), each of which has three biological replicates.

Sample name	Condition	Replicate
WT	wild type	1
WT	wild type	2
WT	wild type	3
KO	knock out	1
KO	knock out	2
KO	knock out	3

1.2 Aligning sample reads to the genome

First, we need to move into the directory containing the data.

```
cd ~/course_data/group_projects/RNASeq_1
```

Then, we need to build our HISAT2 index for the genome.

```
hisat2-build PbANKA_v3.fasta PbANKA_v3_hisat2.idx
```

Next, we can use a loop to align all of our sample files to the genome.

Be patient, this will take a while!

```
for fname in *_1.fastq.gz
       do
           # Get sample name from file name
           sample=`echo "$fname" | cut -d'_' -f1`

           # Align sample to genome
           echo "Aligning sample..."${sample}
           hisat2 -max-intronlen 10000 -x PbANKA_v3_hisat2.idx \
           -1 ${sample}_1.fastq.gz -2 ${sample}_2.fastq.gz \
           -S ${sample}.sam

           # Convert SAM to sorted BAM
           echo "Converting sample SAM to sorted BAM..."${sample}
           samtools view -b ${sample}.sam | \
           samtools sort -o ${sample}.sorted.bam

           # Index sorted BAM
           echo "Indexing sample BAM..."${sample}
           samtools index ${sample}.sorted.bam
       done
```

1.2.1 Important questions

What is the overall alignment rate of each of the samples?

It is important to look at the overall alignment rate (for the genome) as this can give an idea of whether there are any issues with the experiment (e.g. contamination - like we saw in the practical).

Sample name	Alignment rate
WT1	97.63%
WT2	83.60%
WT3	97.83%
KO1	97.14%
KO2	88.88%
KO3	97.12%

This looks good, all of the samples have a relatively similar alignment rate >80%.

1.3 Visualising the genome alignments

Before you can use IGV to visualise the genome, you must first index the genome using `samtools faidx`.

Index the genome with `samtools`.

```
samtools faidx PbANKA_v3.fasta
```

Once that's finished, you need to load your genome (`PbANKA_v3.fasta`), annotation (`PbANKA_v3.gff3.gz`) and sorted alignment files (`[sample].sorted.bam`) into IGV.

First, start IGV.

```
igv.sh &
```

Load the genome file `Genomes -> Load Genome from File`

Load the annotation (gff) file `File -> Load from File`

Load the sorted sample BAM files `File -> Load from File`

Make sure to set the alignment tracks to "squished" and to view reads as "paired".

Type 'PBANKA_KO' in the search box and click 'Go'.

This will give you a view like the one below. Here, we have coloured the WT coverage plots blue and the KO coverage plots red to make it a little easier to see the difference.

images/PBANKA_KO_IGV.png

Do you notice anything unusual about any of the alignments to PBANKA_KO? Should there be any reads mapping to this knocked out gene?

Here, we've hidden the alignment tracks so that we are just looking at the coverage plots. As this is a knock out experiment, you would typically expect to see expression of the gene of interest in the WT samples and no expression of the gene in the KO samples.

In this case, we can see reads mapping to PBANKA_KO in our WT samples as expected. There appears to be a complete knock down in samples KO1 and KO2. However, reads are mapping to PBANKA_KO in the KO3 sample. This suggests that it may have been an incomplete knock down of PBANKA_KO in KO3.

images/PBANKA_KO_IGV_coverage.png

1.3.1 Important questions

Where in the genome is PBANKA_KO located?

You can get the co-ordinates of PBANKA_KO from the annotation file (`PbANKA_v3.gff3.gz`) using `grep` (first uncompressing the file with `gunzip`) or `zgrep`.

```
zgrep gene.*PBANKA_KO PbANKA_v3.gff3.gz
```

PBANKA_KO is located on the **forward strand** of **PbANKA_14_v3** between **1416412** and **1423431**.

How many exons does PBANKA_KO have?

In IGV, you can see that PBANKA_KO has **one exon**.

images/PBANKA_KO_IGV_exon.png

This can be confirmed by looking for the PBANKA_KO CDS annotations the GFF file.

```
zgrep CDS.*PBANKA_KO PbANKA_v3.gff3.gz
```

1.4 Aligning sample reads to the transcriptome

Before we can use `kallisto` to align the sample reads to the transcriptome, we first need to build a kallisto index of the transcriptome using `kallisto index`.

Build a Kallisto index of the tanscriptome (`Pb.CDS.fasta`) using `kallisto`.


```
kallisto index -i Pb.CDS.kallisto Pb.CDS.fasta
```

As with the genome alignments, we can run the transcriptome alignments for all the samples using a loop.

Align your samples to the transcriptome using `kallisto quant`.

```
for fname in *_1.fastq.gz
       do
           # Get sample name from file name
           sample=`echo "$fname" | cut -d'_' -f1`

           # Quantify transcript expression in sample
           echo "kallisto quantification for sample..."${sample}
           kallisto quant -i Pb.CDS.kallisto -o ${sample} -b 100 \
           ${sample}_1.fastq.gz ${sample}_2.fastq.gz
       done
```

1.4.1 Important questions

How many transcripts are there?

There are **5077** transcripts.

```
grep -c '>' Pb.CDS.fasta
```

1.5 Run DE analysis in sleuth

To identify differentially expressed genes you can use the R package, `sleuth`.

Run the sleuth R script (`sleuth.R`).

```
Rscript sleuth.R
```

This should give you an error which contains:

```
cannot open file 'hiseq_info.txt': No such file or directory
```

So, let's take a look at the R script and see what's going on.

```
cat sleuth.R
```

Look at the second line:

```
s2c <- read.table("hiseq_info.txt", header = TRUE, stringsAsFactors=FALSE)
```

The script is looking for a file called `hiseq_info.txt`.

Let's see if the file has been given to you.

```
ls hiseq_info.txt
```

Nope. Well...we did warn you that some files might be missing! But, that still doesn't tell us what the `hiseq_info.txt` file contains...

Let's take a look at the one we used in the practical.

```
cat ~/pathogen-informatics-training/Notebooks/RNA-Seq/data/hiseq_info.txt
```

So, it looks like this file indicates which condition was applied to each of the samples.

Copy the file from the practical into the same directory as your sleuth R script.

```
cp ~/pathogen-informatics-training/Notebooks/RNA-Seq/data/hiseq_info.txt .
```

Let's check that worked.

```
cat hiseq_info.txt
```

Good, now let's update this file so it contains our sample names.

You can edit the file manually by typing the following `nano` command in your terminal. Be careful which order you put the samples in.

```
nano ~/course_data/group_projects/RNASeq_1/hiseq_info.txt
```

Alternatively, you can make the edits using `sed`.

```
sed -i -e 's/MT/WT/g' hiseq_info.txt
       sed -i -e 's/SBP/KO/g' hiseq_info.txt
       sed -ie '/^WT2/a WT3\tWT' hiseq_info.txt
```

And, check that it's worked.

```
cat hiseq_info.txt
```

Perfect, our six samples are now in the file.

So, let's try running that R script again.

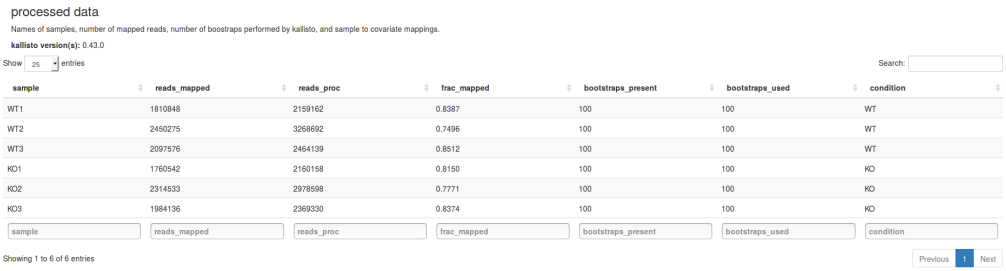
```
Rscript sleuth.R
```

Click http://127.0.0.1:42427 to open the sleuth results in your web browser.

1.5.1 Important questions

Can you summarise the data that's been processed (i.e. number of reads processed and the proportion of reads mapping to the genome and transcriptome)?

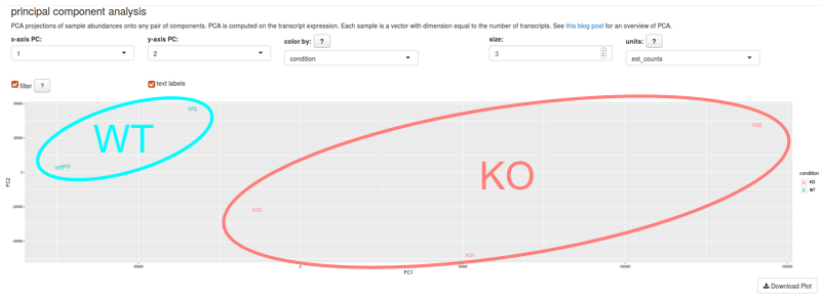
You can get a summary of the processed data by going to `summaries -> processed data`.



images/kallisto_processed_data.png

Looking at the PCA plot (`maps -> PCA`) do you think the samples form tight, distinct clusters based on the condition (WT or KO) that was applied?

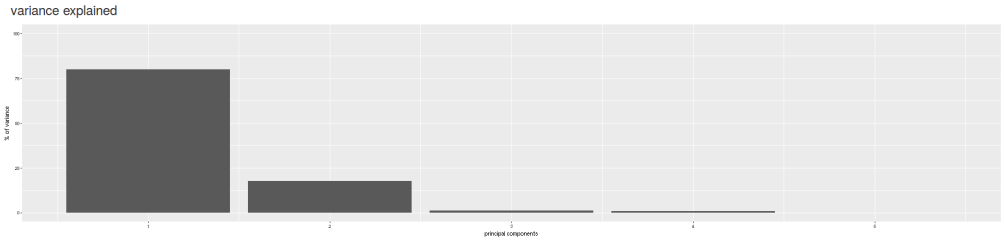
Not really. There is a vertical split between the WT and the KO samples. However, KO3 is reasonably close to the WT samples because of the incomplete knock out, which prevents tighter clustering.



images/sleuth_original_pca.png

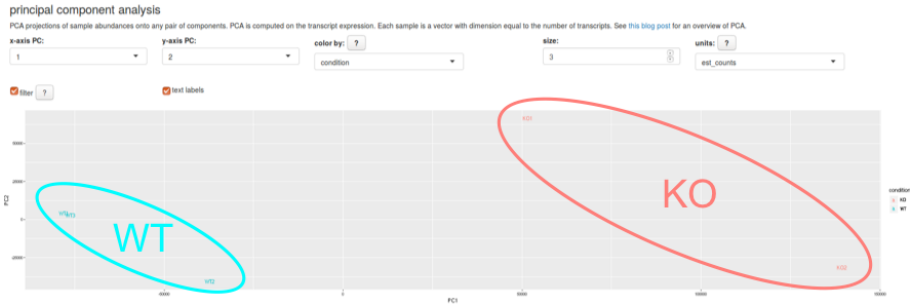
You can also look at the proportion of variance explained by each principal component (PC). As this is a single factor experiment, we would expect that if there were variation, most of this would be explained by the first principal component, PC1. Broadly speaking, this represents the variation resulting from the difference in condition (WT vs KO).

You can see here that >75% of the variance is explained by PC1 (the vertical axis of the PCA plot above). However, there's 15-20% of the variance which is explained by the second principal component. It's possible this is linked to the replicate number and, if you were particularly worried about it, can be accounted for in downstream analyses.



images/sleuth_original_pca_bar.png

If you were to rerun the analysis with KO3 removed, the PCA plot does become a little clearer.

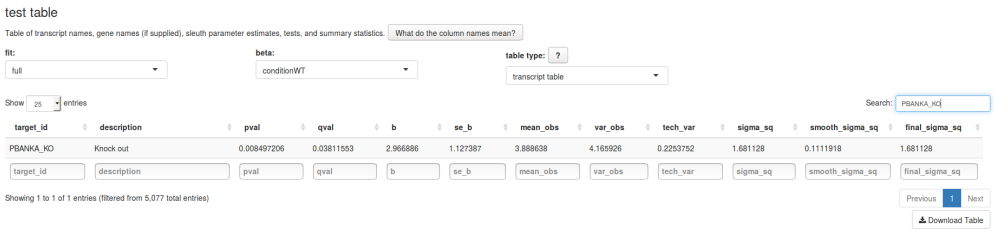


images/sleuth_processed_pca.png

Was PBANKA_KO differentially expressed?

Yes as it's significantly (q-value < 0.05) more highly expressed (b > 0) in the WT samples.

For this, you need to go to `analyses -> test table` and enter PBANKA_KO in the search box.

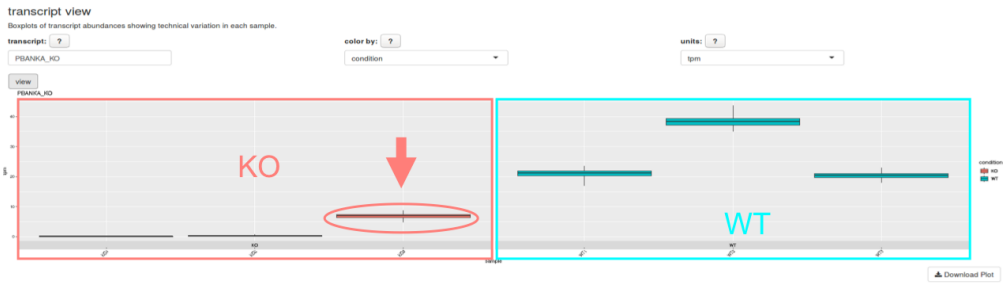


images/sleuth_test_table.png

Is there anything unusual about the PBANKA_KO expression levels in any of the samples?

You'll already have seen an indication of this in the genome alignments. It seems there is partial expression of PBANKA_KO in KO3.

You can look at the expression profiles by going to `analyses -> transcript view` and typing PBANKA_KO in the search box.



images/sleuth_transcript_view.png

How many genes are more highly expressed in the WT samples than in the KO samples?

299

```
awk -F'\t' '$4 < 0.01 && $5 > 0' kallisto.results | wc -l
```

Can you identify which 10 genes are most upregulated in the WT samples?

```
awk -F'\t' '$4 < 0.01 && $5 > 0  {OFS="\t"; print $1,$2,$4,$5}' \
       kallisto.results | sort -t$'\t' -k4 -nr | head
```

How many genes are more highly expressed in the KO samples than in the WT samples?

410

```
awk -F'\t' '$4 < 0.01 && $5 < 0' kallisto.results | wc -l
```

Can you identify which 10 genes are most upregulated in the WT samples?

```
awk -F'\t' '$4 < 0.01 && $5 < 0  {OFS="\t"; print $1,$2,$4,$5}' \
       kallisto.results | sort -t$'\t' -k4 -nr | head
```

Write the gene IDs of the significantly differentially expressed genes to files for the next part of the analysis.

```
awk -F'\t' '$4 < 0.01 && $5 > 0  {print $1}' \
       kallisto.results > kallisto.WT.sig.genes
```

```
awk -F'\t' '$4 < 0.01 && $5 < 0  {print $1}' \
       kallisto.results > kallisto.KO.sig.genes
```

1.6 GO term enrichment analysis

Gene ontology (GO) terms are a dictionary which can be used to assign functions to a gene or transcript. You can use `http://www.plasmodb.org` to perform a GO term enrichment analysis (i.e. which terms are significantly more abundant in your differentially expressed genes than in all of the genes as a whole).

Go to `http://www.plasmodb.org` in your web browser.

Go to `My Strategies -> New`.

Go to `Annotation, curation and identifiers -> Gene IDs`.

Upload your file of gene IDs that were more highly expressed in the WT samples.

Go to `Analyse results` (blue button) and `GO enrichment`.

You want to do a GO analysis using the biological processes (BP).

1.6.1 Important questions

Which GO terms (biological processes) are enriched in genes with higher expression in the WT samples?

You can get this from the table that the analysis generates. You could say that broadly speaking that this gene is involved in the regulation of motility, adhesion and the cell cycle.

Analysis Results:

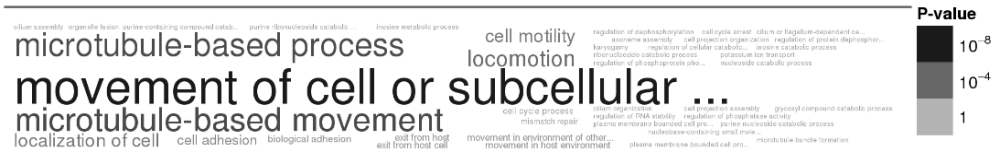
Got a total of 42 results Filter :

Open in **Revigo** Show **Word Cloud** Download

GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd Genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonferroni
GO:0006928	movement of cell or subcellular component	39	14	35.9	7.93	13.14	4.32e-10	1.08e-7	1.08e-7
GO:0007018	microtubule-based movement	18	8	44.4	9.81	17.93	4.25e-7	3.78e-5	1.06e-4
GO:0007017	microtubule-based process	38	11	28.9	6.39	9.31	4.53e-7	3.78e-5	1.13e-4
GO:0040011	locomotion	31	8	25.8	5.7	7.76	4.65e-5	2.91e-3	1.16e-2
GO:0048870	cell motility	21	6	28.6	6.31	8.8	2.38e-4	9.90e-3	5.94e-2
GO:0051674	localization of cell	21	6	28.6	6.31	8.8	2.38e-4	9.90e-3	5.94e-2
GO:0007155	cell adhesion	2	2	100.0	22.08	inf	2.04e-3	7.27e-2	5.09e-1
GO:0022610	biological adhesion	4	2	50.0	11.04	21.4	1.15e-2	2.70e-1	1.00e+0
GO:0035890	exit from host	12	3	25.0	5.52	7.17	1.48e-2	2.70e-1	1.00e+0
GO:0035891	exit from host cell	12	3	25.0	5.52	7.17	1.48e-2	2.70e-1	1.00e+0
GO:0052192	movement in environment of other organism involved in symbiotic interaction	12	3	25.0	5.52	7.17	1.48e-2	2.70e-1	1.00e+0
GO:0052126	movement in host environment	12	3	25.0	5.52	7.17	1.48e-2	2.70e-1	1.00e+0
GO:0022402	cell cycle process	13	3	23.1	5.1	6.45	1.86e-2	2.70e-1	1.00e+0
GO:0006298	mismatch repair	6	2	33.3	7.36	10.69	2.71e-2	2.70e-1	1.00e+0
GO:0060271	cilium assembly	1	1	100.0	22.08	inf	4.53e-2	2.70e-1	1.00e+0
GO:0048284	organelle fusion	1	1	100.0	22.08	inf	4.53e-2	2.70e-1	1.00e+0
GO:0072523	purine-containing compound catabolic process	1	1	100.0	22.08	inf	4.53e-2	2.70e-1	1.00e+0
GO:0046130	purine ribonucleoside catabolic process	1	1	100.0	22.08	inf	4.53e-2	2.70e-1	1.00e+0
GO:0046102	inosine metabolic process	1	1	100.0	22.08	inf	4.53e-2	2.70e-1	1.00e+0
GO:0044782	cilium organization	1	1	100.0	22.08	inf	4.53e-2	2.70e-1	1.00e+0
GO:0043666	regulation of phosphoprotein phosphatase activity	1	1	100.0	22.08	inf	4.53e-2	2.70e-1	1.00e+0
GO:0043487	regulation of RNA stability	1	1	100.0	22.08	inf	4.53e-2	2.70e-1	1.00e+0
GO:0042454	ribonucleoside catabolic process	1	1	100.0	22.08	inf	4.53e-2	2.70e-1	1.00e+0
GO:0120031	plasma membrane bounded cell	1	1	100.0	22.08	inf	4.53e-2	2.70e-1	1.00e+0

images/WT_BP_table.png

You can also use some of the other output options to find interesting ways of displaying this data.



images/WT_BP_words.png

Which GO terms (biological processes) are enriched in genes with higher expression in the KO samples?

You'll need to run the same analysis with your KO file and look at the results table. It looks like there are changes in ribosomal processes.

Analysis Results:

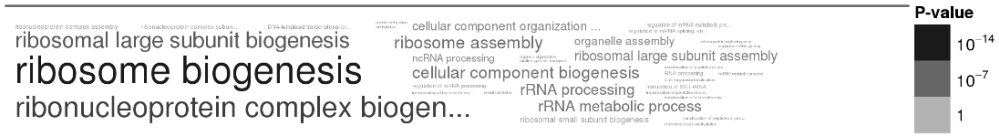
Got a total of 35 results Filter :

Open in **Revigo** Show **Word Cloud** Download

GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd Genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonferroni
GO:0042254	ribosome biogenesis	64	30	46.9	5.04	9.58	2.21e-15	1.08e-12	1.08e-12
GO:0022613	ribonucleoprotein complex biogenesis	82	30	36.6	3.93	6.22	6.57e-12	1.60e-9	3.21e-9
GO:0042273	ribosomal large subunit biogenesis	15	11	73.3	7.88	27.93	3.63e-9	5.90e-7	1.77e-6
GO:0042255	ribosome assembly	10	8	80.0	8.59	40.17	1.94e-7	2.37e-5	9.48e-5
GO:0044085	cellular component biogenesis	136	32	23.5	2.53	3.28	3.37e-7	3.29e-5	1.64e-4
GO:0006364	rRNA processing	36	14	38.9	4.18	6.5	1.51e-6	1.06e-4	7.39e-4
GO:0016072	rRNA metabolic process	36	14	38.9	4.18	6.5	1.51e-6	1.06e-4	7.39e-4
GO:0000027	ribosomal large subunit assembly	7	6	85.7	9.21	59.81	3.98e-6	2.43e-4	1.94e-3
GO:0071840	cellular component organization or biogenesis	199	37	18.6	2.0	2.43	1.75e-5	9.51e-4	8.56e-3
GO:0070925	organelle assembly	16	8	50.0	5.37	10.02	3.37e-5	1.64e-3	1.64e-2
GO:0034470	ncRNA processing	63	16	25.4	2.73	3.47	1.32e-4	5.85e-3	6.43e-2
GO:0042274	ribosomal small subunit biogenesis	9	5	55.6	5.97	12.4	6.20e-4	2.52e-2	3.02e-1
GO:0022618	ribonucleoprotein complex assembly	28	8	28.6	3.07	3.99	3.01e-3	1.13e-1	1.00e+0
GO:0071826	ribonucleoprotein complex subunit organization	30	8	26.7	2.86	3.62	4.81e-3	1.68e-1	1.00e+0
GO:0070897	DNA-templated transcriptional preinitiation complex assembly	2	2	100.0	10.74	inf	8.63e-3	2.16e-1	1.00e+0
GO:0050684	regulation of mRNA processing	2	2	100.0	10.74	inf	8.63e-3	2.16e-1	1.00e+0
GO:0048024	regulation of mRNA splicing, via spliceosome	2	2	100.0	10.74	inf	8.63e-3	2.16e-1	1.00e+0
GO:1903311	regulation of mRNA metabolic process	2	2	100.0	10.74	inf	8.63e-3	2.16e-1	1.00e+0
GO:0030490	maturation of SSU-rRNA	2	2	100.0	10.74	inf	8.63e-3	2.16e-1	1.00e+0
GO:0006396	RNA processing	157	24	15.3	1.64	1.83	8.87e-3	2.16e-1	1.00e+0
GO:0042000	translocation of peptides or proteins into host	3	2	66.7	7.16	19.63	2.43e-2	4.39e-1	1.00e+0
GO:0051836	translocation of molecules into other organism involved in symbiotic interaction	3	2	66.7	7.16	19.63	2.43e-2	4.39e-1	1.00e+0
GO:0051808	translocation of peptides or proteins into	3	2	66.7	7.16	19.63	2.43e-2	4.39e-1	1.00e+0

images/KO_BP_table.png

And again, there are several useful ways to visualise your results.



images/KO_BP_words.png

1.7 What is PBANKA_KO?

So, we've seen the influences of the knock out gene, but what is it?

For this group task, we removed the real name of PBANKA_KO from all of the files we gave you. How mean! To get the real name of PBANKA_KO, we need the real genome annotation file.

Download the real annotation file from the FTP site (`Pberghei.gff3.gz`).

```
wget ftp://ftp.sanger.ac.uk/pub/project/pathogens/gff3/CURRENT/Pberghei.gff3.gz
```

Now, earlier on, you will have jotted down the location of PBANKA_KO in the genome.

Search for a gene with the same co-ordinates as PBANKA_KO in the reall annotation file.

```
zgrep "PbANKA_14_v3.*gene.*1416412.*1423431" Pberghei.gff3.gz
```

Looks like PBANKA_KO is really **PBANKA_1437500**, better known as **AP2-G**, in disguise!

Looking into the literature, you will find that AP2-G encodes a transcription factor and plays a role in gametocyte development (gametogenesis). Thus, it makes sense that knocking out this gene will result in differential expression of genes involved in cell structure, cycle and motility.