1 Group Task 2 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].fq.gz`
- Knockout sample reads in FASTQ format
  `KO[replicate]_[1|2].fq.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 /home/manager/course_data/group_projects/RNASeq_2
```

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.fq.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.fq.gz -2 ${sample}_2.fq.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	97.04%
WT2	97.62%
WT3	92.69%
KO1	96.80%
KO2	96.33%
KO3	90.44%

This looks good, all of the samples have a relatively similar alignment rate >90%.

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`

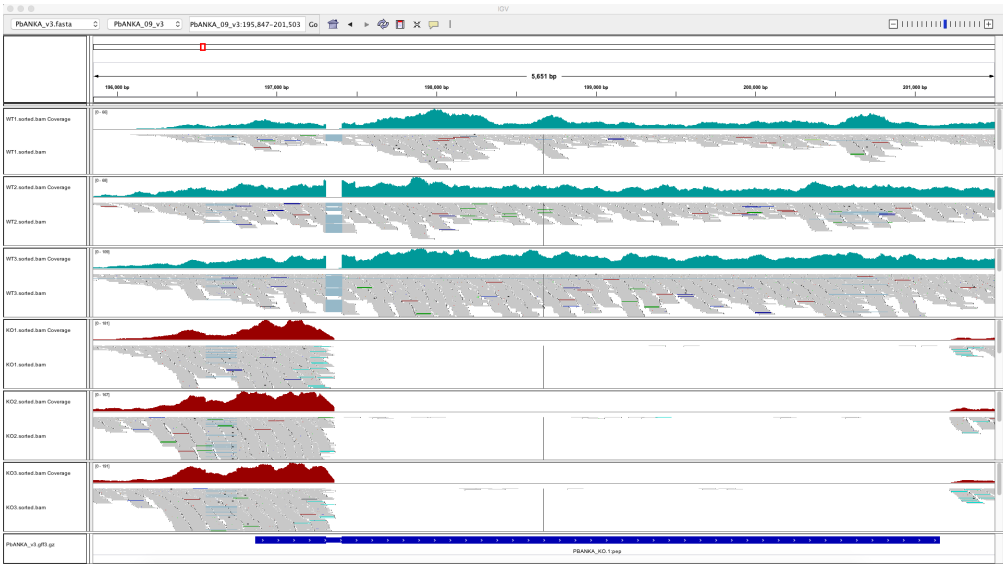
Load the annotation (gff) file `File -> Load from File`

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

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_09_v3** between **196868** and **201157**.

How many exons does PBANKA_KO have?

In IGV, you can see that PBANKA_KO has **two exons**.

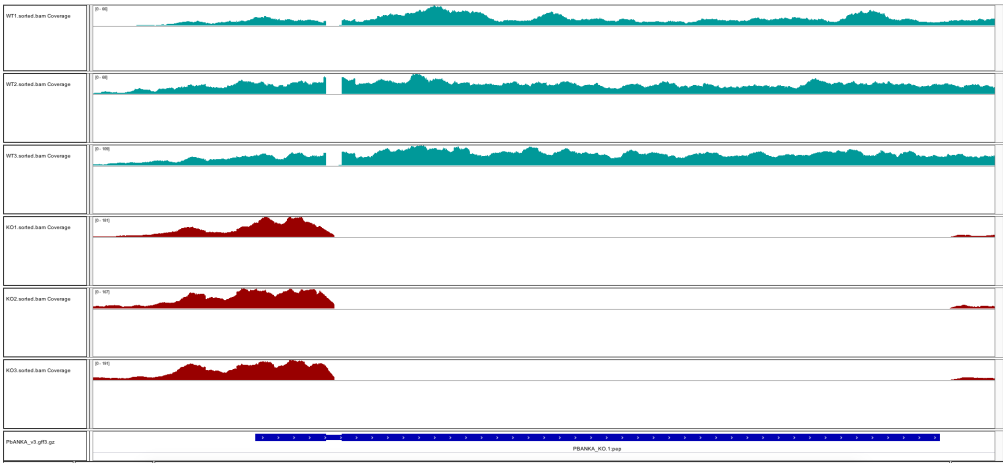
images/PBANKA_KO_IGV_exon.png

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

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

Were all of the exons in PBANKA_KO knocked out?

No. Only the second exon was knocked out.



images/PBANKA_KO_IGV_coverage.png

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.fq.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.fq.gz ${sample}_2.fq.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 -ie 's/MT/WT/g' hiseq_info.txt
       sed -ie '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.

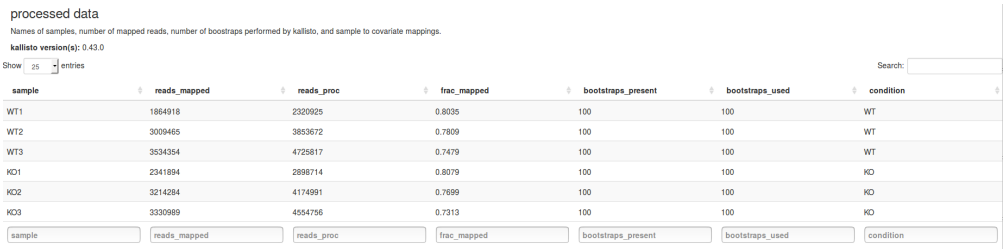
```
Rscript sleuth.R
```

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`.



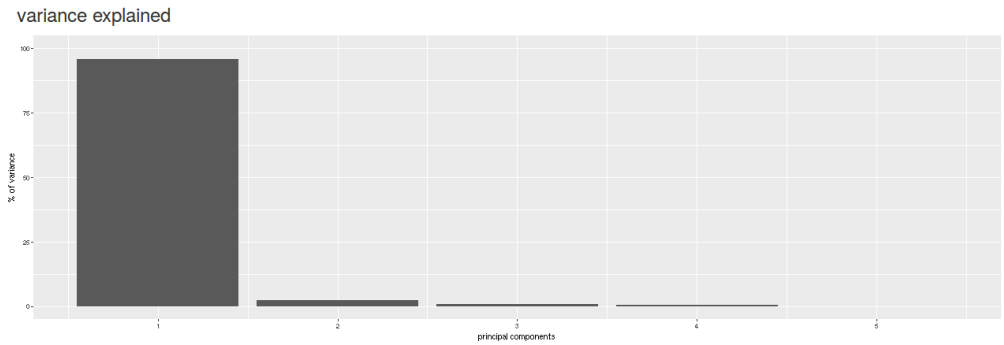
images/kallisto_processed_data.png

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?

Reasonably, yes. There is a clear vertical split between the WT and the KO samples.

images/sleuth_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 >90% of the variance is explained by PC1 (the vertical axis of the PCA plot above).

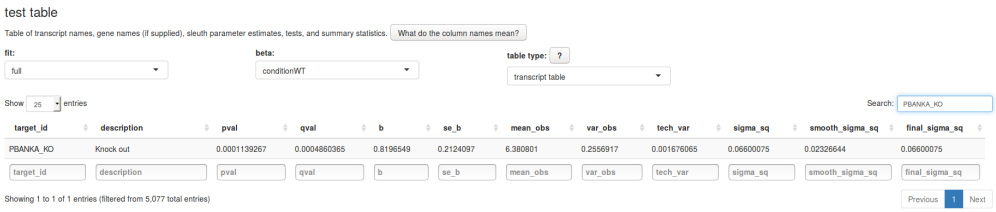


images/sleuth_pca_bar.png

Was PBANKA_KO differentially expressed?

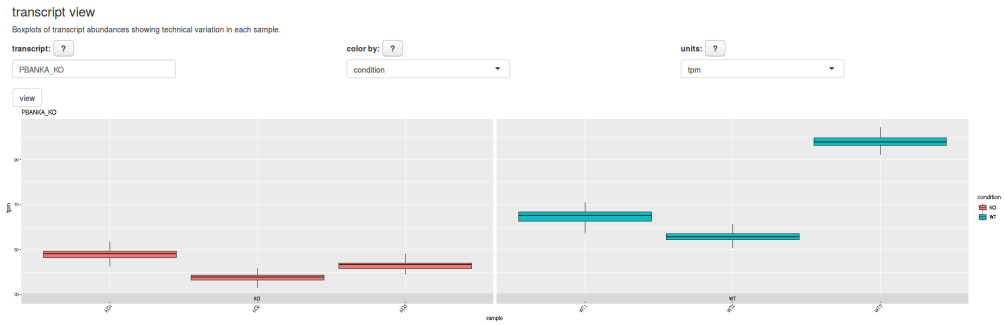
Yes as it's significantly (q-value < 0.05) more highly expressed (b > 0) in the WT samples. The beta value may be lower than you expect as only one of the two exons was knocked out (i.e. the reads mapping to the first exon will be counted towards the PBANKA_KO gene expression in the KO samples).

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



images/sleuth_test_table.png

You can also 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?

828

```
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?

748

```
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 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:

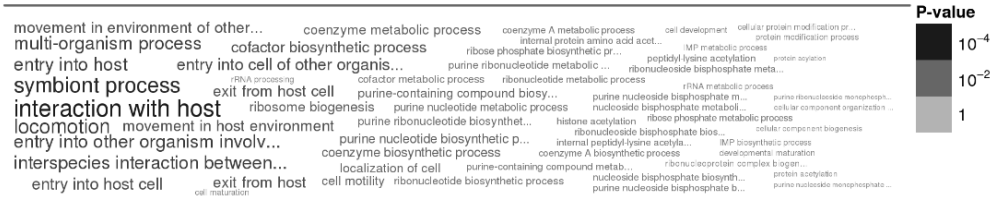
Got a total of 58 results Filter :

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:0051701	interaction with host	56	22	39.3	2.36	3.36	3.62e-5	2.73e-2	2.73e-2
GO:0044403	symbiont process	59	22	37.3	2.24	3.08	9.23e-5	3.48e-2	6.97e-2
GO:0040011	locomotion	31	14	45.2	2.72	4.23	1.72e-4	3.62e-2	1.30e-1
GO:0044409	entry into host	40	16	40.0	2.41	3.43	3.35e-4	3.62e-2	2.53e-1
GO:0051828	entry into other organism involved in symbiotic interaction	40	16	40.0	2.41	3.43	3.35e-4	3.62e-2	2.53e-1
GO:0051704	multi-organism process	72	24	33.3	2.01	2.59	3.36e-4	3.62e-2	2.53e-1
GO:0044419	interspecies interaction between organisms	72	24	33.3	2.01	2.59	3.36e-4	3.62e-2	2.53e-1
GO:0030260	entry into host cell	35	14	40.0	2.41	3.42	7.88e-4	6.61e-2	5.95e-1
GO:0051806	entry into cell of other organism involved in symbiotic interaction	35	14	40.0	2.41	3.42	7.88e-4	6.61e-2	5.95e-1
GO:0052192	movement in environment of other organism involved in symbiotic interaction	12	7	58.3	3.51	7.12	1.24e-3	7.17e-2	9.33e-1
GO:0052126	movement in host environment	12	7	58.3	3.51	7.12	1.24e-3	7.17e-2	9.33e-1
GO:0035890	exit from host	12	7	58.3	3.51	7.12	1.24e-3	7.17e-2	9.33e-1
GO:0035891	exit from host cell	12	7	58.3	3.51	7.12	1.24e-3	7.17e-2	9.33e-1
GO:0051188	cofactor biosynthetic process	37	14	37.8	2.28	3.12	1.50e-3	8.11e-2	1.00e+0
GO:0042254	ribosome biogenesis	64	20	31.3	1.88	2.34	2.54e-3	1.26e-1	1.00e+0
GO:0006732	coenzyme metabolic process	35	13	37.1	2.23	3.02	2.68e-3	1.26e-1	1.00e+0
GO:0006164	purine nucleotide biosynthetic process	28	11	39.3	2.36	3.3	3.41e-3	1.43e-1	1.00e+0
GO:0009108	coenzyme biosynthetic process	28	11	39.3	2.36	3.3	3.41e-3	1.43e-1	1.00e+0
GO:0051674	localization of cell	21	9	42.9	2.58	3.82	4.02e-3	1.52e-1	1.00e+0
GO:0048870	cell motility	21	9	42.9	2.58	3.82	4.02e-3	1.52e-1	1.00e+0
GO:0072522	purine-containing compound biosynthetic process	29	11	37.9	2.28	3.12	4.69e-3	1.68e-1	1.00e+0
GO:0009152	purine ribonucleotide biosynthetic process	26	10	38.5	2.31	3.18	6.20e-3	2.13e-1	1.00e+0
GO:0006163	purine nucleotide metabolic process	35	12	34.3	2.06	2.66	8.13e-3	2.65e-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 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:

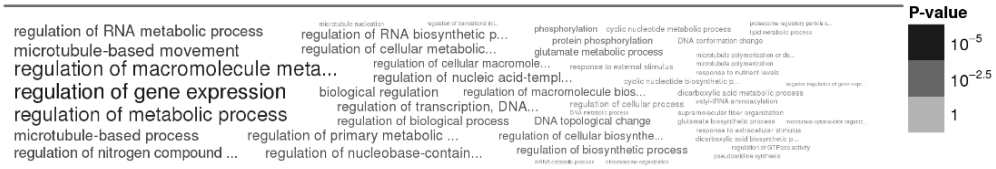
Got a total of 48 results Filter :

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:0010468	regulation of gene expression	82	28	34.1	2.28	3.08	8.98e-6	4.13e-3	6.06e-3
GO:0060255	regulation of macromolecule metabolic process	92	30	32.6	2.18	2.88	1.22e-5	4.13e-3	8.27e-3
GO:0019222	regulation of metabolic process	94	30	31.9	2.13	2.79	1.98e-5	4.45e-3	1.34e-2
GO:0007018	microtubule-based movement	18	10	55.6	3.71	7.24	7.25e-5	1.22e-2	4.89e-2
GO:0007017	microtubule-based process	38	15	39.5	2.64	3.8	1.87e-4	2.24e-2	1.26e-1
GO:0051252	regulation of RNA metabolic process	59	20	33.9	2.26	3.01	1.99e-4	2.24e-2	1.34e-1
GO:0051171	regulation of nitrogen compound metabolic process	83	25	30.1	2.01	2.54	2.78e-4	2.52e-2	1.88e-1
GO:0080090	regulation of primary metabolic process	84	25	29.8	1.99	2.49	3.42e-4	2.52e-2	2.31e-1
GO:0019219	regulation of nucleobase-containing compound metabolic process	62	20	32.3	2.15	2.79	4.22e-4	2.52e-2	2.85e-1
GO:0031323	regulation of cellular metabolic process	81	24	29.6	1.98	2.47	4.83e-4	2.52e-2	3.26e-1
GO:0065007	biological regulation	220	51	23.2	1.55	1.81	5.13e-4	2.52e-2	3.46e-1
GO:2001141	regulation of RNA biosynthetic process	54	18	33.3	2.23	2.92	5.23e-4	2.52e-2	3.53e-1
GO:0006355	regulation of transcription, DNA-templated	54	18	33.3	2.23	2.92	5.23e-4	2.52e-2	3.53e-1
GO:1903506	regulation of nucleic acid-templated transcription	54	18	33.3	2.23	2.92	5.23e-4	2.52e-2	3.53e-1
GO:0050789	regulation of biological process	183	43	23.5	1.57	1.83	1.07e-3	4.83e-2	7.25e-1
GO:2000112	regulation of cellular macromolecule biosynthetic process	72	21	29.2	1.95	2.41	1.36e-3	5.39e-2	9.15e-1
GO:0010556	regulation of macromolecule biosynthetic process	72	21	29.2	1.95	2.41	1.36e-3	5.39e-2	9.15e-1
GO:0031326	regulation of cellular biosynthetic process	73	21	28.8	1.92	2.36	1.65e-3	5.85e-2	1.00e+0
GO:0009889	regulation of biosynthetic process	73	21	28.8	1.92	2.36	1.65e-3	5.85e-2	1.00e+0
GO:0006265	DNA topological change	5	4	80.0	5.34	22.91	2.19e-3	7.39e-2	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_09_v3.*gene.*196868.*201157" Pberghei.gff3.gz
```

Looks like PBANKA_KO is really **PBANKA_0905900**, better known as **AP2-O**, in disguise!

Looking into the literature, you will find that AP2-O encodes a transcription factor and plays a role in ookinete development. Thus, it makes sense that knocking out this gene will result in differential expression of genes involved in host-pathogen interactions and locomotion.