

# Zebrafish-Specific Dataset

**Before using the dataset, copy “deseq2-results.tsv” from “penelopeprime” to your home directory. Alternatively, you can download the file from:**

<https://funcgen2019.buschlab.org/downloads/deseq2-results.tsv>

This dataset comes from a current collaboration and it's not yet published so please **DO NOT** share the data outside of this course.

The dataset consists of four comparisons, each of which is between 12 homozygous zebrafish embryos and 12 of their heterozygous and wild-type siblings. The four comparisons vary according to the age of the embryos (either 3, 5 or 7 dpf) and as to whether or not they have been infected with *Mycobacterium marinum*. The mutation was identified in a genetic screen for loci affecting infection susceptibility.

Each of the 96 samples (4 x 12 vs 12) has a name like “inf\_5dpf\_hom\_repX”, where “inf” indicates the sample was infected (as opposed to “uninf”), “5dpf” indicates the embryo is 5 days post fertilisation, “hom” indicates the embryo is homozygous for the mutation (as opposed to “het” or “wt”) and X is a number indicating the replicate.

The column headings are:

1	<b>GeneID</b>	Ensembl ID
2	<b>inf_5dpf_hom_vs_sib_pval</b>	p-value for homozygote vs sibling comparison in infected 5 dpf embryos
3	<b>inf_5dpf_hom_vs_sib_adjp</b>	Adjusted p-value for homozygote vs sibling comparison in infected 5 dpf embryos
4	<b>inf_5dpf_hom_vs_sib_log2fc</b>	Log <sub>2</sub> fold change for homozygote vs sibling comparison in infected 5 dpf embryos
5	<b>uninf_3dpf_hom_vs_sib_pval</b>	p-value for homozygote vs sibling comparison in uninfected 3 dpf embryos
6	<b>uninf_3dpf_hom_vs_sib_adjp</b>	Adjusted p-value for homozygote vs sibling comparison in uninfected 3 dpf embryos
7	<b>uninf_3dpf_hom_vs_sib_log2fc</b>	Log <sub>2</sub> fold change for homozygote vs sibling comparison in uninfected 3 dpf embryos
8	<b>uninf_5dpf_hom_vs_sib_pval</b>	p-value for homozygote vs sibling comparison in uninfected 5 dpf embryos
9	<b>uninf_5dpf_hom_vs_sib_adjp</b>	Adjusted p-value for homozygote vs sibling comparison in uninfected 5 dpf embryos
10	<b>uninf_5dpf_hom_vs_sib_log2fc</b>	Log <sub>2</sub> fold change for homozygote vs sibling comparison in uninfected 5 dpf embryos
11	<b>uninf_7dpf_hom_vs_sib_pval</b>	p-value for homozygote vs sibling comparison in uninfected 7 dpf embryos
12	<b>uninf_7dpf_hom_vs_sib_adjp</b>	Adjusted p-value for homozygote vs sibling comparison in uninfected 7 dpf embryos
13	<b>uninf_7dpf_hom_vs_sib_log2fc</b>	Log <sub>2</sub> fold change for homozygote vs sibling comparison in uninfected 7 dpf embryos
14	<b>Chr</b>	Chromosome (or scaffold) name
15	<b>Start</b>	Gene start (in bp)
16	<b>End</b>	Gene end (in bp)

17	<b>Strand</b>	Gene strand (1 or -1)
18	<b>Biotype</b>	Gene biotype (e.g. protein coding or lincRNA)
19	<b>Name</b>	Gene name
20	<b>Description</b>	Gene description
21	<b>inf_5dpf_wt_rep1_count</b>	Counts for 1 <sup>st</sup> inf_5dpf_wt replicate
22	<b>inf_5dpf_wt_rep2_count</b>	Counts for 2 <sup>nd</sup> inf_5dpf_wt replicate
...	...	...
116	<b>uninf_7dpf_hom_rep12_count</b>	Counts for 12 <sup>th</sup> uninf_7dpf_hom replicate
117	<b>inf_5dpf_wt_rep1_normalised_count</b>	Normalised counts for 1 <sup>st</sup> inf_5dpf_wt replicate
118	<b>inf_5dpf_wt_rep2_normalised_count</b>	Normalised counts for 2 <sup>nd</sup> inf_5dpf_wt replicate
...	...	...
212	<b>uninf_7dpf_hom_rep12_normalised_count</b>	Normalised counts for 12 <sup>th</sup> uninf_7dpf_hom replicate

For reference (and only for reference – none of this is necessary for this course), this dataset was generated as follow:

1. The zebrafish GRCz11 genome and Ensembl 98 transcriptome were downloaded and unzipped using:

```
wget ftp://ftp.ensembl.org/pub/release-98/fasta/danio_rerio/dna/Danio_rerio.GRCz11.dna_sm.primary_assembly.fa.gz
wget ftp://ftp.ensembl.org/pub/release-98/gtf/danio_rerio/Danio_rerio.GRCz11.98.gtf.gz
gunzip Danio_rerio.GRCz11.dna_sm.primary_assembly.fa.gz
gunzip Danio_rerio.GRCz11.98.gtf.gz
```

2. The genome was indexed using STAR:

```
mkdir grcz11 genome-generate
STAR \
--outFileNamePrefix genome-generate/ \
--runThreadN 4 \
--runMode genomeGenerate \
--genomeDir grcz11 \
--genomeFastaFiles Danio_rerio.GRCz11.dna_sm.primary_assembly.fa \
--sjdbGTFfile Danio_rerio.GRCz11.98.gtf \
--sjdbOverhang 74
```

3. For each sample (\$sample below) a pair of FASTQ files were aligned to the genome using STAR:

```
mkdir -p star1/$sample
STAR \
--runThreadN 1 \
--genomeDir grcz11 \
--readFilesIn fastq/$sample.1.fastq.gz fastq/$sample.2.fastq.gz \
--readFilesCommand zcat \
--outFileNamePrefix star1/$sample/ \
--quantMode GeneCounts \
--outSAMtype BAM SortedByCoordinate
done
```

4. Each pair of FASTQ files were aligned to the genome for a second round using STAR:

```
mkdir -p star2/$sample
STAR \
--runThreadN 1 \
--genomeDir grcz11 \
--readFilesIn fastq/$sample.1.fastq.gz fastq/$sample.2.fastq.gz \
--readFilesCommand zcat \
--outFileNamePrefix star2/$sample/ \
--quantMode GeneCounts \
--outSAMtype BAM SortedByCoordinate \
--sjdbFileChrStartEnd `find star1 | grep SJ.out.tab$ | sort | tr '\n' ' '`
```

5. DESeq2 input files were made using:

```
wget https://raw.githubusercontent.com/iansealy/bio-misc/master/make\_deseq\_from\_star.pl
cat /dev/null > counts-files.txt
find star2 | grep ReadsPerGene.out.tab | sort -V | grep 5dpf | grep -v uninf | grep wt >>
counts-files.txt
```



```

echo -ne "Chr\tStart\tEnd\tStrand\tBiotype\tName\tDescription" >> deseq2_results.tsv
for sample in `head -1 deseq2/counts.txt`
do
echo -ne "\t${sample}_count" >> deseq2_results.tsv
done
for sample in `head -1 deseq2/counts.txt`
do
echo -ne "\t${sample}_normalised_count" >> deseq2_results.tsv
done
echo >> deseq2_results.tsv
join -j1 -t $'\t' <(sort deseq2/inf_5dpf_hom_vs_inf_5dpf_sib/output.txt) \
<(sort deseq2/uninf_3dpf_hom_vs_uninf_3dpf_sib/output.txt) \
| join -j1 -t $'\t' - <(sort deseq2/uninf_5dpf_hom_vs_uninf_5dpf_sib/output.txt) \
| join -j1 -t $'\t' - <(sort deseq2/uninf_7dpf_hom_vs_uninf_7dpf_sib/output.txt) \
| join -j1 -t $'\t' - <(sort annotation.txt) \
| join -j1 -t $'\t' - <(sort deseq2/inf_5dpf_hom_vs_inf_5dpf_sib/counts.txt) \
| join -j1 -t $'\t' - <(sort deseq2/uninf_3dpf_hom_vs_uninf_3dpf_sib/counts.txt) \
| join -j1 -t $'\t' - <(sort deseq2/uninf_5dpf_hom_vs_uninf_5dpf_sib/counts.txt) \
| join -j1 -t $'\t' - <(sort deseq2/uninf_7dpf_hom_vs_uninf_7dpf_sib/counts.txt) \
| join -j1 -t $'\t' - <(sort deseq2/inf_5dpf_hom_vs_inf_5dpf_sib/normalised-counts.txt) \
| join -j1 -t $'\t' - <(sort deseq2/uninf_3dpf_hom_vs_uninf_3dpf_sib/normalised-counts.txt) \
| join -j1 -t $'\t' - <(sort deseq2/uninf_5dpf_hom_vs_uninf_5dpf_sib/normalised-counts.txt) \
| join -j1 -t $'\t' - <(sort deseq2/uninf_7dpf_hom_vs_uninf_7dpf_sib/normalised-counts.txt) \
>> deseq2_results.tsv

```