Zebrafish Dataset

Before using the dataset, copy "Amp.counts.tsv" and "Oxy.counts.tsv" from "penelopeprime" to your home directory. Alternatively, you can download the files from:
https://funcgen2022.buschlab.org/downloads/Amp.counts.tsv
https://funcgen2022.buschlab.org/downloads/Oxy.counts.tsv

You should also copy the files "Amp.samples.tsv" and "Oxy.samples.tsv" from "penelopeprime" to your home directory. These are used by DESeq2 (see below) and list all the samples along with their corresponding DESeq2 conditions. You can also download the files from:
https://funcgen2022.buschlab.org/downloads/Amp.samples.tsv
https://funcgen2022.buschlab.org/downloads/Oxy.samples.tsv

This dataset comes from a paper we published earlier this year:

- Mech *et al.* (2022) "Behavioral and Gene Regulatory Responses to Developmental Drug Exposures in Zebrafish." Front. Psychiatry https://www.frontiersin.org/articles/10.3389/fpsyt.2021.795175

Zebrafish were exposed to amphetamine, nicotine or oxycodone from 24 hpf to 5 dpf and behavioural assays were performed on the larvae. At 5 dpf, 6 samples, each consisting of pools of 6-7 embryos, were collected for each condition (plus unexposed controls). In total, 24 samples were collected, although two later failed QC and were excluded from the analysis.

RNA was extracted and sequencing libraries were made using Illumina's TruSeq Stranded mRNA kit. They were sequenced on one lane of NovaSeq SP PE50, resulting in 16-24 million reads per sample. The reads were aligned to the GRCz11 reference genome with STAR and differentially expressed genes were determined with DESeq2.

Each of the 22 sampleshas a name like "Cnt_3", where "Cnt" indicates a control sample (the others being "Amp", "Nic" and "Oxy") and 3 is a number indicating the replicate.

The column headings are:

1	**Gene**	Ensembl ID
2	**pval**	p-value
3	**adjp**	Adjusted p-value
4	**log2fc**	Log_2 fold change
5	**Chr**	Chromosome (or scaffold) name
6	**Start**	Gene start (in bp)
7	**End**	Gene end (in bp)
8	**Strand**	Gene strand (1 or -1)
9	**Biotype**	Gene biotype (e.g. protein coding or lincRNA)
10	**Name**	Gene name
11	**Description**	Gene description
12	**Cnt_1 count**	Counts for 1st control replicate
13	**Cnt_2 count**	Counts for 2nd control replicate
…	**…**	…
23	**Cnt_1 normalised count**	Normalised counts for 1st control replicate
24	**Cnt_2 normalised count**	Normalised counts for 2nd control replicate
…	**…**	…

For reference (and only for reference – none of this is necessary for this course), this dataset was generated using STAR and DESeq2 as follows. The files needed to repeat this analysis yourself are available in https://funcgen2022.buschlab.org/downloads/dataset.zip

1. The zebrafish GRCz11 genome and Ensembl 107 transcriptome were downloaded and unzipped using:

```
wget ftp://ftp.ensembl.org/pub/release-
107/fasta/danio_rerio/dna/Danio_rerio.GRCz11.dna_sm.primary_assembly.fa.gz
wget ftp://ftp.ensembl.org/pub/release-107/gtf/danio_rerio/Danio_rerio.GRCz11.107.gtf.gz
gunzip Danio_rerio.GRCz11.dna_sm.primary_assembly.fa.gz
gunzip Danio_rerio.GRCz11.107.gtf.gz
```

2. The genome was indexed using STAR:

```
mkdir grcz11
STAR \
--outFileNamePrefix grcz11. \
--runThreadN 4 \
--runMode genomeGenerate \
--genomeDir grcz11 \
--genomeFastaFiles Danio_rerio.GRCz11.dna_sm.primary_assembly.fa \
--sjdbGTFfile Danio_rerio.GRCz11.107.gtf \
--sjdbOverhang 49
```

3. For each sample ($sample below) a pair of FASTQ files ($fastq1 and $fastq2 below) were aligned to the genome using STAR:

```
mkdir -p star1/$sampleSTAR \
--runThreadN 4 \
--genomeDir ref/grcz11 \
--readFilesIn $fastq1 $fastq2 \
--readFilesCommand zcat \
--outFileNamePrefix star1/$sample/ \
--quantMode GeneCounts \
--outSAMtype BAM SortedByCoordinate
```

4. Each pair of FASTQ files were aligned to the genome for a second round, with improved splice junction annotation, using STAR:

```
mkdir -p star2/$sample
STAR \
--runThreadN 4 \
--genomeDir ref/grcz11 \
--readFilesIn $fastq1 $fastq2 \
--readFilesCommand zcat \
--outFileNamePrefix star2/$sample/ \
--quantMode GeneCounts \
--outSAMtype BAM Unsorted SortedByCoordinate \
--sjdbFileChrStartEnd `find star1 | grep SJ.out.tab$ | sort | tr '\n' ' '`
```

5. DESeq2 was run using the `deseq2.R` file available in https://funcgen2022.buschlab.org/downloads/dataset.zip