**Epigen 2025 – Project Overview**

1. **Introduction** (maybe base it on the project proposal)
2. **Preprocessing**
   1. **Trimming :** Trimmomatic v.0.36 (Bolger et al., 2014)
   2. **Mapping:** aligned to the mm10 genome assembly using the Burrows- Wheeler Aligner v.0.7.12 (Li and Durbin, 2009)
   3. **Filtering:** Reads were filtered to keep only high quality alignments (MAPQ score >20), duplicates were removed using SAMtools v.1.3.1 (Li et al., 2009). 🡪 Did you remove the duplicates?
   4. **Peak calling**: MACS2 v.2.1.1 (Zhang et al., 2008) using piling up of paired- end fragment mode (--format BAMPE) & **annotate the peaks**
   5. **Filtering**: The peak files (bed) were filtered by removing the ENCODE black listed regions (https://www.encodeproject.org/files/ ENCFF547MET) using BEDTools v2.29.1 (Quinlan, 2014). Mitochondrial reads were also removed before the analysis. 🡪 Do we want to do this? I don’t think we did it in the lecture, but at least the mitochondrial reads I think could be a big problem and & I would remove them.
   6. **Get into raw count matrix**
   7. **Normalize:** normalized by scaling factor (--scale Factor) with the deepTools v.2.5.0.1 (Ramírez et al., 2014) bamCoverage function.
3. **A comparison of different types of fiber

   AI-generated content may be incorrect.Check our Results**

Make sure we got meaningful results e.g. through Enrichment of genomic signal around TSS or select peaks in regions that should be active in myofibers like Ckm, Acta 1, etc.

1. **Correlation analysis**

Between the replicates (could maybe add the results with their data in their as well 🡪 How well does it correlate to their pre-processed results)

A graph with numbers and letters

AI-generated content may be incorrect.A comparison of a graph

AI-generated content may be incorrect.A diagram of a graph

AI-generated content may be incorrect.

2.1. Pearson Correlation

2.2. PCA

2.3. Hierarchical Clustering

2.4. maybe comparison Venn Diagram between our & their peaks

1. **Differential accessibility analysis**

For each comparison between the conditions, overlapping and unique accessible regions were identified with DiffBind v.2.16.2 (Stark, 2011) based on the measure of confidence in the peak call by MACS2 v.2.1.1 (Zhang et al., 2008). edgeR v.3.30.1 (Robinson et al., 2010). Log fold changes were calculated, and their associated p- values were corrected for multiple hypothesis testing via the Benja mini–Hochberg procedure to obtain adjusted p- values. The DARs were annotated by their nearest gene using the annotatePeaks.pl function of Homer v.4.11 (Heinz et al., 2010)

A screenshot of a graph

AI-generated content may be incorrect.A graph of fiber fibers

AI-generated content may be incorrect.

Volcano plot of differentially accessible regions/peaks identified by FDR < 0.05 and LFC ≥ 1 between uninjured myofibers and injured myofibers. Each coloured dot represents a differentially accessible region/peak and the distance to the nearest gene is annotated.

Batch correction? WT vs. MDX & uninjured vs. injured are clearly in different batches

1. **Gene set enrichment analysis**

Genes nearby the DARs were ranked based on the log- fold change calculated with edgeR v.3.30.1 (Robinson et al., 2010). This ranked list of genes was used as input to perform gene set enrichment analysis with the fgseaMultilevel function of the R package fgsea v.1.14.0 (Korotkevich et al., 2021). The FGSEA-multilevel method is based on an adaptive multi- level split Monte Carlo scheme, which allows the estimation of very low p- values. The Hallmark gene sets collection from the Molecular Signatures Database (MSigDB) (Korotkevich et al., 2021) was used as a reference to identify the biological processes that were significantly enriched.

A graph of a cell structure

AI-generated content may be incorrect.A graph with text and numbers

AI-generated content may be incorrect.

A graph of fiber and fiber

AI-generated content may be incorrect.A white paper with text and colorful bars

AI-generated content may be incorrect.

(**A–C**) Gene Ontology (GO Biological Process) analysis of genes associated with ATAC-Seq peaks based on association by proximity using Genomic Regions Enrichment of Annotations Tool (GREAT) ([McLean et al., 2010](https://elifesciences.org/articles/72792#bib50)) for all peaks present in the uninjured myofibers, MuSCs and injured myofibers, respectively.

Gene Set Enrichment Analysis performed on genes nearest to the differentially accessible regions/peaks for uninjured myofibers compared to injured myofibers. Top 10 enriched pathways are shown although do not reach significance.

1. **Evt. Motif enrichment analysis**

The identification of known TF motifs found in peaks overlapping the promoter region (±5 kb of TSS) was done using the findMotifsGenome.pl function from HOMER v.4.9.1 (Heinz et al., 2010). The - size parameter was set to given to use the exact peak region as target sequence. Following the screening of HOMER’s reliable motifs library against the target sequences, the motifs enriched with a p- value less than 0.05 are returned.

A screenshot of a computer

AI-generated content may be incorrect.

1. **Diskussion**
2. **Additional Plots we could do**

Peak annotation pie charts

A diagram of fiber fibers

AI-generated content may be incorrect.A diagram of a graph

AI-generated content may be incorrect.

**Mythogenic Pathways (GSE)**

A diagram of a graph

AI-generated content may be incorrect.