**1. What is edgeR?**

* **edgeR** is a tool in R (a programming language) used to analyze biological data, like your ATAC-seq data. It helps find differences in how much DNA is "open" (accessible) between different samples, such as your fish at different growth stages (3dpf, 7dpf, 12dpf).
* Think of it as a calculator that compares samples to spot which DNA regions are more or less active.

**2. What is Differential Expression Analysis?**

* **Differential expression analysis** looks for differences in activity (like DNA accessibility in ATAC-seq) between two or more groups (e.g., 3dpf vs. 7dpf fish).
* In ATAC-seq, it identifies DNA regions that are more open (accessible to proteins, often linked to active genes) in one group compared to another.
* It’s like comparing two rooms to see which doors are open more in one room than the other.

**3. Why Do This Analysis in Your Project?**

* Your project studies *Astatotilapia calliptera* (a fish) at three developmental stages: 3 days post-fertilization (dpf, early embryo), 7dpf (developing), and 12dpf (swimming). ATAC-seq measures which DNA regions are open, which can control gene activity.
* You’re looking for DNA regions that change accessibility between these stages to understand how genes drive development (e.g., genes for movement, like GO:0007626 for locomotory behavior).
* This analysis helps find which DNA regions are active at each stage, revealing how the fish grows and develops.

**4. Steps in the Protocol for Differential Analysis**

The protocol outlines these main steps to find differences in DNA accessibility:

1. **Load Data**: Read ATAC-seq data (BAM files) from your six samples (two replicates per stage: 3dpf\_1a, 3dpf\_1b, 7dpf\_2a, 7dpf\_2b, 12dpf\_3a, 12dpf\_3b).
2. **Count Reads**: Divide the genome into small windows (150 bp) and count how many DNA fragments are in each window for each sample.
3. **Filter Windows**: Keep only windows with enough fragments (more than background noise) to focus on meaningful regions.
4. **Normalize**: Adjust counts to account for differences in total reads between samples, making comparisons fair.
5. **Group Samples**: Define your groups (3dpf, 7dpf, 12dpf) for comparison.
6. **Statistical Testing**: Use edgeR to compare groups (e.g., 3dpf vs. 7dpf) and find windows with significant accessibility differences (FDR < 0.05, meaning reliable differences).
7. **Visualize Results**: Create plots (volcano, MA, MDS) to see the differences and check data quality.

**5. What Did We Do with the Code?**

We followed the protocol to analyze your data and create plots, broken into these simple steps:

1. **Loaded Data**: Used your BAM files to read ATAC-seq data for six samples.
2. **Counted and Filtered**: Split the genome into 231,158 windows (150 bp each), counted fragments, and kept windows with enough data (filtered to ~231,158 windows).
3. **Normalized and Grouped**: Adjusted counts and grouped samples by stage (3dpf, 7dpf, 12dpf).
4. **Compared Stages**: Compared pairs (3dpf vs. 7dpf, 3dpf vs. 12dpf, 7dpf vs. 12dpf) to find windows with different accessibility, saving results in TSV files (differential\_peaks\_\*.tsv).
5. **Found Significant Windows**: Identified 78,675 (3vs7), 76,959 (3vs12), and 10,341 (7vs12) significant windows (FDR < 0.05).
6. **Tried Promoter Overlaps**: Attempted to link significant windows to promoter regions (near genes) using a BED file, but hit issues (zero overlaps due to wrong BED file).
7. **Created Plots**: Generated volcano and MA plots from TSV files to visualize differences. Prepared to create an MDS plot to check sample similarity (requires dgel).

**6. What Are Volcano, MA, and MDS Plots?**

* **Volcano Plot**:
  + **What it shows**: A scatter plot where each point is a window. The x-axis shows the difference in accessibility (logFC: positive means more open in the first stage, e.g., 3dpf; negative means more open in the second, e.g., 7dpf). The y-axis shows significance (-log10(FDR), higher means more reliable).
  + **Why use it**: Shows which windows have big, reliable differences. Red points (FDR < 0.05) are significant.
* **MA Plot**:
  + **What it shows**: A scatter plot with x-axis as average accessibility (logCPM, how open the window is overall) and y-axis as logFC (same as volcano). Red points are significant.
  + **Why use it**: Shows if differences occur in highly or lowly accessible regions, helping understand their importance.
* **MDS Plot**:
  + **What it shows**: A 2D plot of your six samples, where similar samples (e.g., 3dpf\_1a and 3dpf\_1b) are close together, and different stages (3dpf vs. 12dpf) are far apart.
  + **Why use it**: Checks if replicates are consistent (close together) and stages are distinct, ensuring your data is reliable.

**7. Interpreting Your Results Based on Images**

You provided a PDF with six images (volcano and MA plots for 3vs7, 3vs12, 7vs12). Here’s how to interpret them based on your significant peak counts (78,675, 76,959, 10,341):

* **Volcano Plots** (image1: 3vs7, image2: 3vs12, image3: 7vs12):
  + **What you see**:
    - **3vs7 and 3vs12**: Many red points (~34% and 33%, ~78,675 and 76,959 windows) spread in a “V” shape, with points at high y-values (-log10(FDR) > 2) and logFC from -5 to 5. This means many DNA regions have strong, reliable differences in openness between early (3dpf) and later stages (7dpf, 12dpf).
    - **7vs12**: Fewer red points (~4.5%, 10,341), closer to logFC = 0 and lower y-values, indicating smaller, less significant differences.
  + **What it means**:
    - Big differences between 3dpf and 7dpf/12dpf suggest major changes in DNA accessibility during early development, possibly linked to genes for growth or movement (e.g., GO:0007626).
    - Smaller differences between 7dpf and 12dpf make sense, as these stages are closer in development.
    - **Concern**: 34% and 33% significant windows are high (typically 1–10% in ATAC-seq). If nearly all points are red or y-values are very high (>10), it might mean your replicates are too similar, making small differences look significant.
* **MA Plots** (image4: 3vs7, image5: 3vs12, image6: 7vs12):
  + **What you see**:
    - **3vs7 and 3vs12**: Many red points across logCPM (0–10), especially at moderate-high values (4–8), with logFC from -5 to 5. This shows significant differences in both highly and lowly accessible regions.
    - **7vs12**: Fewer red points, mostly at moderate logCPM, with smaller logFC (-2 to 2).
  + **What it means**:
    - Significant changes in 3vs7 and 3vs12 occur in active DNA regions, likely controlling important genes for development.
    - 7vs12 has subtler changes, focused in moderately active regions, consistent with closer stages.
    - **Concern**: If red points dominate or cluster at low logCPM, it could indicate noise or technical issues.
* **MDS Plot** (not yet generated, pending MDS\_plot.R):
  + **What to expect**: Six points (3dpf\_1a, 1b in blue; 7dpf\_2a, 2b in green; 12dpf\_3a, 3b in red). Replicates should be close (e.g., 3dpf\_1a near 1b), and stages should be apart (3dpf far from 12dpf).
  + **What it means**:
    - If replicates cluster tightly and stages are separate, your data is reliable, and high significant counts reflect true developmental differences.
    - If replicates are far apart or stages overlap, your data may have issues, and the high counts (34%, 33%) might be inflated.
  + **Action**: Run MDS\_plot.R and check library sizes. If sizes are very similar (e.g., 3dpf\_1a ≈ 3dpf\_1b), low variability may explain the high counts.
* **Overall Interpretation**:
  + Your results show strong DNA accessibility changes between early (3dpf) and later stages (7dpf, 12dpf), likely driving developmental gene activity. The 7vs12 comparison shows fewer changes, as expected.
  + The high number of significant windows (34%, 33%) is unusual and needs checking with the MDS plot and library sizes. If replicates are too similar, you may need to re-run the analysis with the protocol’s normalization (normOffsets).
  + Once you get the correct BED file, you can link significant windows to genes (promoters) for GO analysis (e.g., movement-related genes).

**Summary**

* **What we did**: Used edgeR to compare DNA accessibility across three fish developmental stages, found many significant differences, and created volcano/MA plots to visualize them. We’re waiting to generate the MDS plot and link results to genes.
* **Why it matters**: This shows which DNA regions change during fish growth, helping understand gene control in development.
* **What to do next**: Run the MDS plot script, check library sizes, and describe the plot (replicate clustering, stage separation). If the MDS plot looks good, your results are reliable. Then, use the correct BED file for promoter analysis and GO enrichment.

**Errors**

> # Normalize for biases > normfacs <- normOffsets(filtered.data) > filtered.data <- filtered.data > # Define groups (one stage per BAM, no replicates in merged BAMs) > cell\_types <- c("3dpf", "7dpf", "12dpf") > > # Convert to DGEList > dgel <- asDGEList(filtered.data, group = cell\_types) > dgel An object of class "DGEList" $counts Sample1 Sample2 Sample3 1 12 154 64 2 3 170 45 3 4 148 49 4 6 153 64 5 7 174 54 260 more rows ... $samples group lib.size norm.factors Sample1 3dpf 939488 1 Sample2 7dpf 3097161 1 Sample3 12dpf 687411 1 > > # Design matrix > moma <- model.matrix(~0 + group, dgel$samples) > colnames(moma) <- levels(factor(cell\_types)) > # Fit model > fit <- glmQLFit(dgel, moma) Error in glmFit.default(y, design = design, dispersion = dispersion, offset = offset, : NA dispersions not allowed In addition: Warning message: In estimateGLMCommonDisp.default(y[i, , drop = FALSE], design = design, : No residual df: setting dispersion to NA

**Why the Error Occurred**

The error message:

text

Error in glmFit.default(y, design = design, dispersion = dispersion, offset = offset, :

NA dispersions not allowed

In addition: Warning message:

In estimateGLMCommonDisp.default(y[i, , drop = FALSE], design = design, :

No residual df: setting dispersion to NA

indicates that glmQLFit cannot estimate dispersions (a measure of biological variability) because your data lacks replicates, leading to zero residual degrees of freedom (df). Here’s why this happens:

* **No Replicates**: Your dgel object shows three samples (Sample1, Sample2, Sample3) corresponding to one merged BAM file per stage (3 dpf, 7 dpf, 12 dpf). The dgel$samples output confirms this:

text

$samples

group lib.size norm.factors

Sample1 3dpf 939488 1

Sample2 7dpf 3097161 1

Sample3 12dpf 687411 1

With only one sample per group (cell\_types), edgeR cannot estimate biological variability (dispersion) between replicates, which is required for the quasi-likelihood (QL) method in glmQLFit.

* **Residual Degrees of Freedom**: The warning No residual df means there are no extra samples beyond what’s needed to fit the model, so dispersion estimation fails (set to NA), causing the error.
* **Merged BAMs**: Your supervisor provided merged BAMs (3dpf\_ATAC...merged.chr1subset.bam, etc.), which combine replicates (e.g., 1aAc\_3dpf and 1bAc\_3dpf into one 3 dpf BAM). This simplifies analysis but removes replicate information, making statistical testing challenging.

**Addressing the Error**

To proceed with differential analysis despite having no replicates, we have two options:

1. **Use a Fixed Dispersion**: Since glmQLFit requires dispersion estimates, we can set a fixed dispersion value (e.g., 0.1, a common choice for ATAC-seq data with low variability) and use glmFit instead of glmQLFit, followed by a likelihood ratio test (glmLRT). This is a workaround when replicates are unavailable, though it assumes uniform variability across windows.
2. **Use Replicate BAMs**: If possible, use the individual replicate BAMs (e.g., 1aAc\_3dpf\_ATAC.nochrM.nodup.filt.shifted.bam, 1bAc\_3dpf\_ATAC...) to include replicate variability, which is ideal for glmQLFit. Your directory (/home/hsmidhuk/hc-storage/ATAC\_Acalliptera/2.Annotation/) contains these files:
   * 1aAc\_3dpf\_ATAC.nochrM.nodup.filt.shifted.bam, 1bAc\_3dpf\_ATAC...
   * 2aAc\_7dpf\_ATAC..., 2bAc\_7dpf\_ATAC...
   * 3aAc\_12dpf\_ATAC..., 3bAc\_12dpf\_ATAC...

**Recommendation**: Option 1 (fixed dispersion) is the quickest fix to continue with your merged BAMs, but Option 2 (replicates) is statistically better and aligns with standard ATAC-seq analysis. I’ll provide code for Option 1 now, as it uses your current dgel object, and include Option 2 as an alternative if you can access replicates. You should confirm with your supervisor whether to use replicates for the final analysis.