



Integrate Salmon to MultiQC

CSE – 549 COMPUTATIONAL BIOLOGY

MM – DD - YYYY

Alok Thatikunta |

Pratik Zambani |

Saahil Khurana |

Sweta Kumari | 111497926

# PURPOSE

MultiQC is a tool used for analyzing multiple sequencing samples together. The tool itself can parse through different types of outputs and generate reports. So, for example, if a folder contains output of multiple runs of Salmon, a transcript quantification and mapping tool for RNA sequences, MultiQC will be able to parse through them, read the log files, determine which tool was run on the samples, filter out features and generate a report using them. The purpose of this project is to make modules for MultiQC that are Salmon-specific.

# SCOPE

* Plot GC bias related output of Salmon on MultiQC
* Plot 5’ and 3’ sequence biases output of Salmon on MultiQC
* Generate Heatmap

# APPROACH

## GC-BIAS

* Parsed and read the given binaries of observed and expected GC bias (from Salmon output)
* Processed the linear combination of weights for observed and expected values (W1\*a1 + w2\*a2 + w3\*a3)
* Then, calculated the ratio of observed values to expected values and plotted the ratio for each sample using the MultiQC inbuilt line plot module.

## SEQUENCE BIAS

* Wrote a parser for the sequence bias binaries from salmon output.
* Read the binary and made 4 lists representing the observed values and expected values for both 3’ and 5’
* Then, calculated the ratio of observed values to expected values for both 3’ and 5’ and plotted 4-line graphs for 4 nucleotides (A, C, G, T), each representing the marginalized 0th - order probabilities of the context.

## Heatmap

# RESULTS

## GC-BIAS

#Add Screenshots for each