Project 2: Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

Authors: Preshita Dave (Analyst), Monica Roberts (Programer), Italo Duran (Data Curator)

Introduction

Self-renewal of heart muscle tissue following injury is a phenomena observed in adult vertebrates including zebrafish and newts[2,3], however, this process only occurs shortly after birth in mammals[4]. This is because the cardiac myocytes exit the cell cycle early in life and the heart continues to grow through enlargement of the existing cells, rather than continuing to undergo mitosis. It has been shown that neonatal mice who underwent resection of the left ventricular apex were able to regenerate the myocytes and repair the damage. Further genetic studies revealed the regenerated tissue came from existing heart tissue, not stem cells[5]. Identifying the process by which these cells re-enter the cell cycle could provide a means to activate this process in adult cells after injuries such as myocardial infarction.

The authors of the study measured gene expression patterns of mouse embryonic cells that were undergoing differentiation to become myocytes in vitro and neonatal cardiac cells that were maturing to adult myocytes in vivo[1]. They then compared these profiles to a transcriptional profile measured from a myocyte explant that exhibited loss of differentiation. Any changes observed in gene or gene network expression could explain the changes in the phenotypes of the cells. The results of the study suggested that cardiac myocytes regenerate after injury through regression in the cell cycle and highlighted signaling pathways that were implicated in the process.

Data

The data from the 8 samples used in this study was given (P0_2, P4_1, P4_2, P7_1, P7_2, Ad_1, Ad_2)[1]. Except for P0_1, which was extracted as an SRA file from the sample data GSM1570702 (vP0_1) from GEO Series GSE64403 from NCBI. These samples are cardiac myocytes from neonatal and adult mice collected at different cell stages[1]. The SRA file was uploaded to the university shared computing cluster (SCC). We constructed a file(run_extract.qsub), that utilized sratoolkit and used the fastq-dump function to convert the P0_1.sra file into two pair end reads FASTQ files. The two FASTQ files were assessed for quality control metrics by using the FastQC tool package within a created file (run_qc.qsub), as shown in the supplementary figures section. Within both FastQC reports, they show an equal number of pair-end reads, length and GC%. We also used the given data for the reference mouse genome (mm9) FASTA and Bowtie2 indexes in the SCC.

Methods

The first step of the analysis was to align the reads of sample P0_1 produced by the RNA-Seq to the reference mouse genome, mm9. The unaligned reads were in a fastq format and the reference genome was in fasta format. To accomplish this, indexes of the reference sequence were also available. The indexes were important for the alignment algorithm to run correctly and efficiently, as it narrows down the location a sequence could be and reduces computational complexity. The algorithm used for alignment was called Tophat[9]. Tophat is a splice junction mapper, meaning it is aware of reads that may map to splice junctions and can still align them properly. It was built using the short read aligner bowtie[10], which uses Burrows-Wheeler transform (BWT) to create an alignment while maintaining efficient memory usage. BWT works by rearranging strings of DNA characters into similar strings, which is easier to

compress for memory. Bowtie is best for aligning short reads to a large genome, which was fitting for this study since the reads produced were typically around 75 base pairs long.

To carry out the alignment of the P0 samples, the modules samtools[11], bowtie2, boost[12], and tophat were loaded into the terminal. The two fastq files produced from the RNAseq, containing the paired reads, along with the reference genome fasta file were used as input for the tophat command. The other parameters included a segment length of 20, segment mismatch threshold of 1, expected inner distance between mate pairs of 200, disallowment of novel junctions, an output directory, number of threads of 16, and file location for the gene model annotations, which supplies the known splice junctions. The tool produced an accepted_hits.bam file that was used for subsequent analysis.

Using samtools, the bam file was run through flagstat, which produced a statistics summary on the bit flags for 13 categories. This allowed for an evaluation of the alignment. The summary was as follows:

```
49706999 + 0 in total (QC-passed reads + QC-failed reads)
8317665 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
49706999 + 0 mapped (100.00% : N/A)
41389334 + 0 paired in sequencing
20878784 + 0 read1
20510550 + 0 read2
29422646 + 0 properly paired (71.09% : N/A)
39936472 + 0 with itself and mate mapped
1452862 + 0 singletons (3.51% : N/A)
1387382 + 0 with mate mapped to a different chr
704916 + 0 with mate mapped to a different chr (mapQ>=5)
```

All reads in the alignment passed quality control measures, indicated by the first row. There were 49,706,999 reads in total. 100% were mapped and 8,317,665 were mapped more than once. 71.09% of the reads were properly paired between the two files. These statistics indicate the alignment was sufficient since 100% of the reads were mapped. The percentage of properly paired reads could indicate discrepancies between the two fastq files, such as one being sorted slightly different than the other. Another problem could have been the inner distance length. If the inner distance between two reads exceeded what was set in the parameters, it could be marked as discordant.

Next, the accepted hits bam file was analyzed using several tools from RseQC that output different quality control metrics. These tools required an indexed bam file, which was produced using the index function in samtools. First, the bam file was run through the geneBody_coverage.py command. This file calculated the RNA-seq read coverage over the gene body. Other parameters included were a bed file, which stores genomic regions as coordinates and its associated annotations, and the output file name prefix. The next RseQC tool used was inner_distance.py. This function was used to calculate the inner distance between read pairs. If the two reads spanned a splice junction, the inner distance was calculated by subtracting the size of the intron. If the inner distance was negative, the two reads overlapped. The input of the function was the accepted hits bam file, along with the bed file and output file prefix.

The last RseQC tool used was bam_stat.py, which summarized the mapping statistics of the bam file. It included information on uniquely mapped reads and the probability that a read was mapped incorrectly. The input for the function was the bam file. The output produced the summary below:

```
Total records:                 49706999
QC failed:                     0
Optical/PCR duplicate:         0
Non primary hits               8317665
Unmapped reads:                0
mapq < mapq_cut (non-unique):  2899954
mapq >= mapq_cut (unique):     38489380
Read-1:                        19409941
Read-2:                        19079439
Reads map to '+':              19236824
Reads map to '-':              19252556
Non-splice reads:              33099839
Splice reads:                  5389541
Reads mapped in proper pairs:  27972916
Proper-paired reads map to different chrom:4
```

The summary indicated that 8,317,665 hits were not primary, meaning they were multi-mapped, aligning to more than one loci. Multimapped reads can cause abundance estimation bias. 38,489,380 reads were unique according to a minimum mapping quality threshold and 2,899,954 were non-unique. 5,389,541 reads were spliced and 27,972,916 were mapped in proper pairs, which was slightly lower than the amount indicated in the flagstat output.

After the alignment was created and analyzed, gene expression was quantified using the cufflinks module[8]. The cufflinks tool counted how many reads mapped to annotated genomic regions. The input file for the tool was the accepted hits bam file and the annotated reference genome along with parameters that indicated to only count hits compatible with the reference, to use the 'rescue method' for multi reads, an output directory name, and number of threads. The output of the tool was a GTF file that contained the assembled isoforms, a file that contained the estimated isoform-level expression values, and a file that contained the gene-level expression values. The gene-level expression values file was loaded into R Studio 4.0.3 and plotted, as shown in Figure . FPKM values below 1 and over 10,000 were filtered out prior to plotting. The values below 1 were likely not actually expressed since their values were substantially lower than the others and the values above 10,000 were also extreme outliers that skewed the distribution. In total, 37,469 differentially expressed genes remained.

The last component of the analysis was carried out with the cuffdiff tool of the cufflinks module. The purpose of the tool was to identify differentially expressed transcripts between different samples. The 3 remaining samples, P0_2, Ad_1, and Ad_2) were processed and available in the /project/bf528/project_2/data/samples directory. The script that contained the commands for running cuffdiff was in the /project/bf528/project_2/scripts directory. The output of the cuffdiff command was used for subsequent analysis.

Results

The output of the RseQC tools used and described in methods resulted in several plots of quality control metrics. The first tool, geneBody_Coverage.py, measured the RNA-seq read coverage over the gene body.

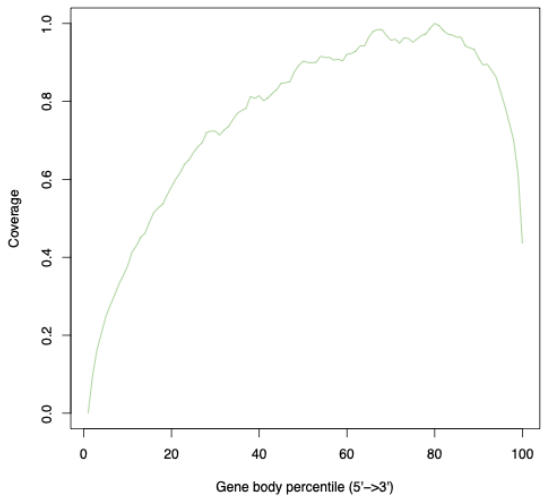

Figure 1: Gene Body Coverage. Output of the GeneBody_Coverage.py tool produced a plot showing the relationship between coverage and gene body percentile.

The plot of coverage shown in Figure 1 indicated there was a drop off in coverage on the ends, which was expected and is normal for high-throughput sequencing. The coverage was higher on the 3' end than the 5', which was a consequence of the samples being prepared with poly-A selection.
The next tool used, inner_distance.py was used to calculate the inner distance between read pairs.

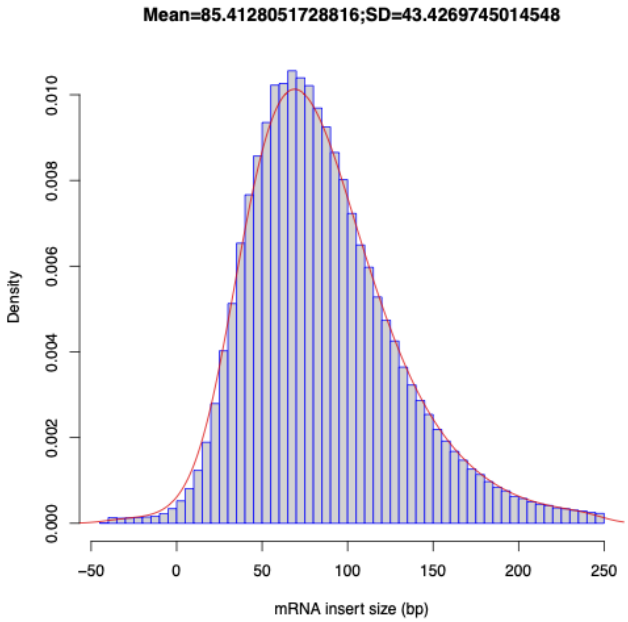

Figure 2: Distribution of inner distance between read pairs. Output of the inner_distance.py tool produced a histogram showing the density mRNA insert sizes.

The density plot of the inner distance between paired reads followed a normal distribution with median 85.413. Most of the distances were above zero, indicating not many reads overlapped. The amount of reads that overlapped was influenced by the fragment size chosen in sample preparation. Too much overlap would have indicated the fragments were longer than needed.

The cufflinks tool produced several outputs, one being the FPKM values for each differentially expressed gene. After threshold values were applied, there were 37,469 differentially expressed genes remaining. The distribution of the values was slightly skewed, with a long right tail, indicating there were less genes with high FPKM values. Most of the values were concentrated near 10.

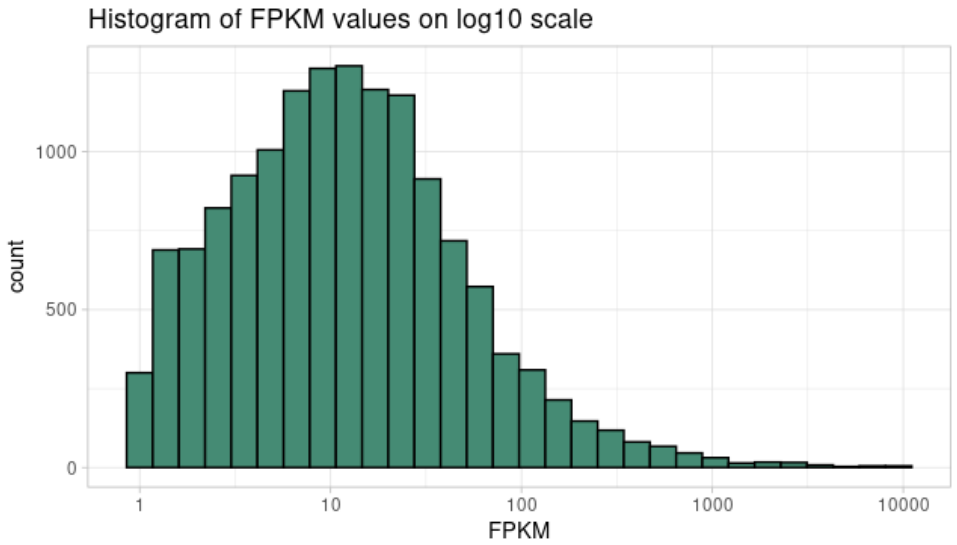

Figure 3: Histogram of FPKM values for differentially expressed genes. The cufflinks tool calculated FPKM values for all genes. Genes with values below 1 and greater than 10,000 were filtered out.

After obtaining the differential gene expression results from the Cuffdiff output, we reproduced the comparison of postnatal day 0 (P0) versus Adult from Figure 1B[1]. The cuffdiff_out/gene_exp.diff file was loaded in and sorted in an ascending order by their q-values. The table with the top 10 differentially expressed genes sorted with the smallest q-value at the top is shown below.

S No.	Gene	P0 FPKM	Adult FPKM	log2.fold_change	p_value	q_value
1	Plekhb2	22.5679	73.5683	1.70481	5E-05	0.00106929
2	Mrpl30	46.4547	133.038	1.51794	5E-05	0.00106929
3	Coq10b	11.0583	53.3	2.26901	5E-05	0.00106929
4	Aox1	1.18858	7.09136	2.57682	5E-05	0.00106929
5	Ndufb3	100.609	265.235	1.39851	5E-05	0.00106929
6	Sp100	2.13489	100.869	5.56218	5E-05	0.00106929

7	Cxcr7	4.95844	32.2753	2.70247	5E-05	0.00106929
8	Lrrfip1	118.997	24.6402	-2.27184	5E-05	0.00106929
9	Ramp1	13.2076	0.691287	-4.25594	5E-05	0.00106929
10	Gpc1	51.2062	185.329	1.8557	5E-05	0.00106929

Table 1: Log2 Fold-Change of Top10 Genes sorted in ascending order of q-value

Next, histograms of the log2 fold-change of all genes and genes with significant expression level were plotted. The significant genes were determined using the significant column which were marked "yes" in the gene_exp.diff file. By "significant genes" Cuffdiff implies whether the p-value is greater than the q-value.

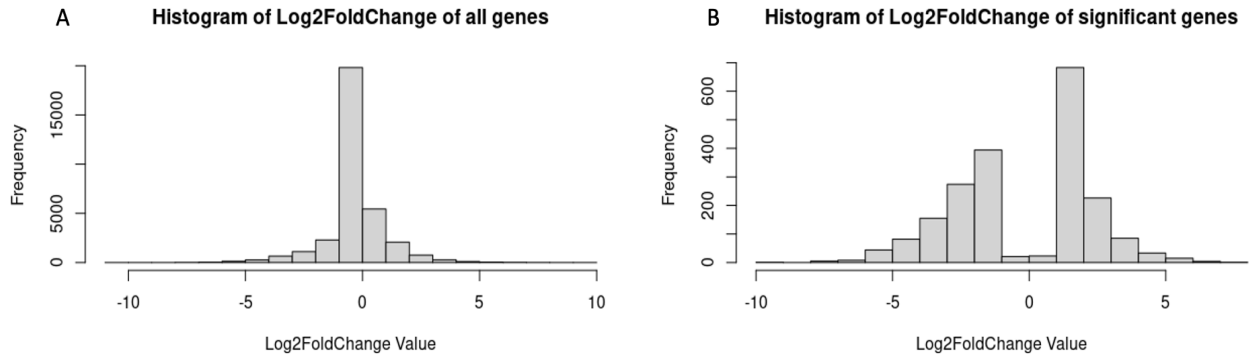

Figure 4: (A) Histogram of Log2 Fold-Change for all of the genes **(B)** Histogram of Log2 fold-change differentially expressed genes (Cufflinks significant).

From the figure above, we can observe that the significant genes have non-zero Log2 Fold change values, which is apparent as these genes need to have a difference in their Log2 Fold change values in order to be classified as significantly differentially expressed genes. Before filtering the genes into up and down-regulated genes, only genes with p-value < 0.01 were considered. A p-value indicates the probability that a fold change as strong as the observed one, or even stronger, would be seen under the situation described by the null hypothesis.

Up-regulated genes are those genes which have a Log2 Fold-Change > 0 and down-regulated genes are those genes which have a Log2 Fold-Change < 0. Based on the Cufflink analysis, we identified 2139 genes of the genes differentially expressed between postnatal P0 and adult. Among them, 1084 (50.7%) genes were up-regulated and 1055 (49.3%) genes were down-regulated. However, the article reported a total of 9779 significant genes with 2409 genes up-regulated and 7570 genes down-regulated[1].

These genes were saved in separate files and Gene Ontology (GO) Analysis using DAVID[7] was performed. The gene ontology groups used for the analysis were GOTERM_BP_FAT, GOTERM_MF_FAT, and GOTERM_CC_FAT and through Functional Annotation Clustering, the top 5 clusters with the highest enrichment scores of up and down-regulated genes were summarized and then compared to the original article.

Cluster	Biological Pathway	Enrichment Score	Similarity to manuscript
1	Organic acid metabolic process	28.12	*
2	Small molecule catabolic process	25.88	*
3	Response to organic substance	20.89	*
4	Mitochondria	19.61	*
5	Metal ion binding	12.86	*

Table 2: Table with top 5 upregulated GO clusters representing common biological pathways in comparison to the reference paper O'Meara, et al[1]

Cluster	Biological Pathway	Enrichment Score	Similarity to manuscript
1	Metal ion binding	20.51	*
2	Cell proliferation	16.92	*
3	Regulation of RNA processing	12.02	*
4	Cell death	11.85	
5	Cytoskeleton organization	11.5	*

Table 3: Table with top 5 downregulated GO clusters representing common biological pathways in comparison to the reference paper O'Meara, et al[1]

Enriched GO analysis between postnatal P0 and adult samples showed primarily up-regulation in mitochondrion and metabolic-related genes, and down regulation in cell cycle genes relating to cell proliferation and death. The result for the ontology analysis is similar to the original paper. However, from Figure 1C of the original article[1], the enrichment terms of down-regulated genes should have higher enrichment scores. As per O'Meara, et al.[1], cardiac myocytes exit the cell cycle, which can be interpreted from the down-regulation of the cell cycle genes from P0 to Adult phase in Figure 7[1]. A complete overlap of GO terms is not observed between the O'Meara, et al.[1] due to the fact that specific terms for Gene Ontology such as GOTERM_BP_FAT, GOTERM_MF_FAT, and GOTERM_CC_FAT were used and the terms that were used in the O'Meara, et al.[1] paper were not defined.

Discussion

According to the evaluation of the FastQC quality report, the data passed with good quality marks. Although, in the report we did obtain one failed metric, per base sequence content (figure E), this failed metric should show parallels between basepairs.This might be due to commonality of the RNA-seq libraries that tend to contain common biases, as we can see indicated there was an imbalance at beginning of the read; in the first 10 positions; which was expected for RNA-seq high-throughput sequencing. With sequence duplication (figure I), we received a warning. A high level of duplication can signify a technical error on part of the PCR or biological duplicates, which can happen by natural collision where different

copies of the same exact sequence are randomly chosen. Despite this warning, it means that this metric test was not significant enough to interfere with the downstream analysis.

For the TopHat Fastq files alignment with the reference genome, it proved to be valid with the resulting BAM files. By analyzing the data using figure 1, gene body coverage, we can see that the same problem we had with figure E, affected the drop in figure 1, as expected from a consequence of the samples being prepared with poly-A selection. The CuffDiff data had several outputs by expressing 37,469 differentially expressed genes. Out of those genes, we filtered for significance. We filter the top 10 genes between postnatal versus adult. Log2 fold change was evaluated for FPKM_ADULT and FPKM_PO, which measures change of expression level between those samples.

Histograms were plotted for log2 fold change of all genes and significant genes. Discovered that for all the genes the values were centered and near 0 (figure 4 A). For significant genes it was the opposite, it distributed towards opposite extremes. This means that we can see the difference in up and down regulated genes and that the up regulated genes were more strongly pronounced than the down regulated genes. Based on the filtered p_values for all the genes, we discovered there were 1084 (50.7%) genes that were up-regulated and 1055 (49.3%) genes were down-regulated. Thus, the research paper reported a total of 9779 significant genes with 2409 genes up-regulated and 7570 genes down-regulated[1]. A reason for this difference might be that the analyst in the paper must have applied a different threshold than what we used in this replication study.

Conclusion

The O'Meara, et al.[1] paper suggests that cardiac regeneration is a transcriptional reversion of the differentiation process. This paper intends to characterize cardiac regeneration at the molecular level. In this project, we aimed to reproduce the results by processing raw mRNA-seq data and comparing the transcriptional states of P0 and adult mice. Results from the Gene Ontology analysis that were generated from the study are similar to the O'Meara, et al.[1] paper, especially clusters with higher enrichment scores for up- and down-regulated genes were common. The differences in results may appear due to the selection of certain GO terms or also updation of the database with more GO terms at the time of the analysis. This study provided a valuable framework for understanding transcriptional changes for cardiac myocyte repair to study adult cardiac regeneration and identify potential stimulators.

References

1. O'Meara, C. C., Wamstad, J. A., Gladstone, R. A., Fomovsky, G. M., Butty, V. L., Shrikumar, A., Gannon, J. B., Boyer, L. A., & Lee, R. T. (2015). Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration. Circulation Research,116(5),804–815.
2. Senyo SE, Steinhauser ML, Pizzimenti CL, Yang VK, Cai L, Wang M, Wu TD, Guerquin-Kern JL, Lechene CP, Lee RT. Mammalian heart renewal by pre-existing cardiac myocytes. Nature. 2013;493:433–436.
3. Bergmann O, Bhardwaj RD, Bernard S, Zdunek S, Barnabe-Heider F, Walsh S, Zupicich J, Alkass K, Buchholz BA, Druid H, Jovinge S, Frisen J. Evidence for cardiac myocyte renewal in humans. Science. 2009;324:98–102.
4. Bicknell KA, Coxon CH, Brooks G. Can the cardiac myocyte cell cycle be reprogrammed? J Mol Cell Cardiol. 2007;42:706–721.
5. Porrello ER, Mahmoud AI, Simpson E, Hill JA, Richardson JA, Olson EN, Sadek HA. Transient regenerative potential of the neonatal mouse heart. Science. 2011;331:1078–1080.
6. Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics (Oxford, England), 28(16), 2184–2185.
7. Huang, D.W., Sherman, B.T. and Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4(1), pp.44-57.
8. Trapnell, C., Williams, B., Pertea, G. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511–515 (2010).
9. Cole Trapnell, Lior Pachter, Steven L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, Bioinformatics, Volume 25, Issue 9, 1 May 2009, Pages 1105–1111, https://doi.org/10.1093/bioinformatics/btp120
10. Langmead, B., Trapnell, C., Pop, M. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.GenomeBiol10,R25(2009). https://doi.org/10.1186/gb-2009-10-3-r25
11. Li, Heng et al. "The Sequence Alignment/Map format and SAMtools." Bioinformatics (Oxford, England) vol. 25,16 (2009): 2078-9. doi:10.1093/bioinformatics/btp352
12. Schäling Boris. (2014). *The Boost C++ Libraries*. XML Press.

Supplementary Figures

This section shows the FastQC quality report figures and its descriptions for the two files:

Measure	Value	Measure	Value
Filename	P0_1_1.fastq	Filename	P0_1_2.fastq
File type	Conventional base calls	File type	Conventional base calls
Encoding	Sanger / Illumina 1.9	Encoding	Sanger / Illumina 1.9
Total Sequences	21577562	Total Sequences	21577562
Sequences flagged as poor quality	0	Sequences flagged as poor quality	0
Sequence length	40	Sequence length	40
%GC	49	%GC	49

Figure A: FastQC basic statistics report summary. Showing total sequence reads, no sequence flagged as poor, the length of the reads and the percentage of GC content in the sequence.

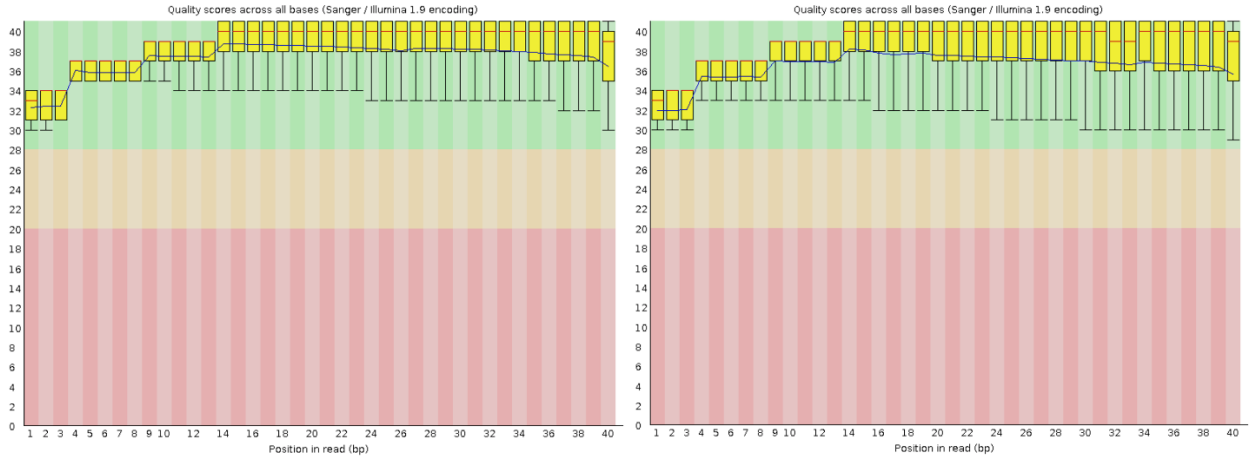

Figure B: Per base sequence quality. P0_1_1 (left), P0_1_2 (right)

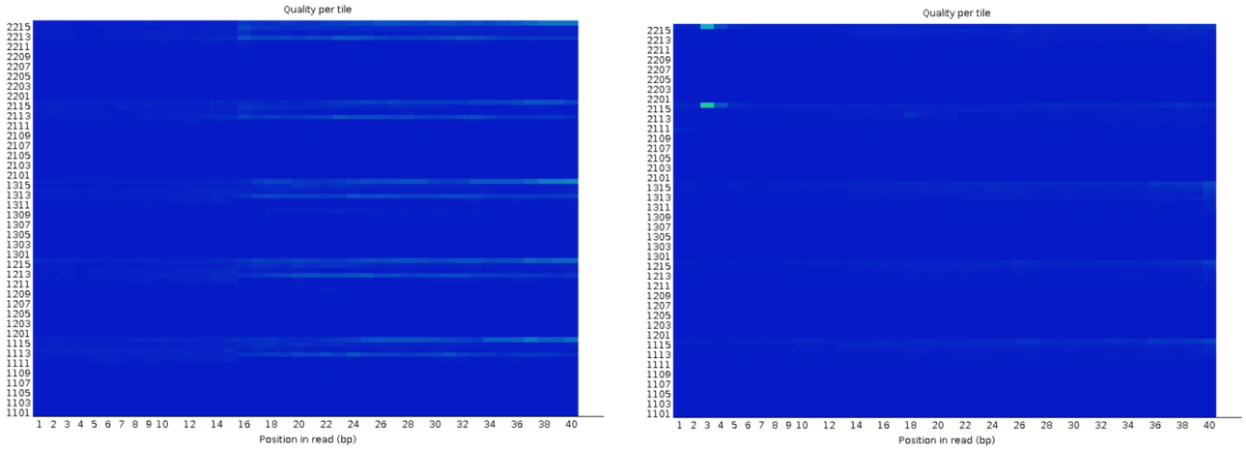

Figure C: Per tile sequence quality. P0_1_1 (right), P0_1_2 (left)

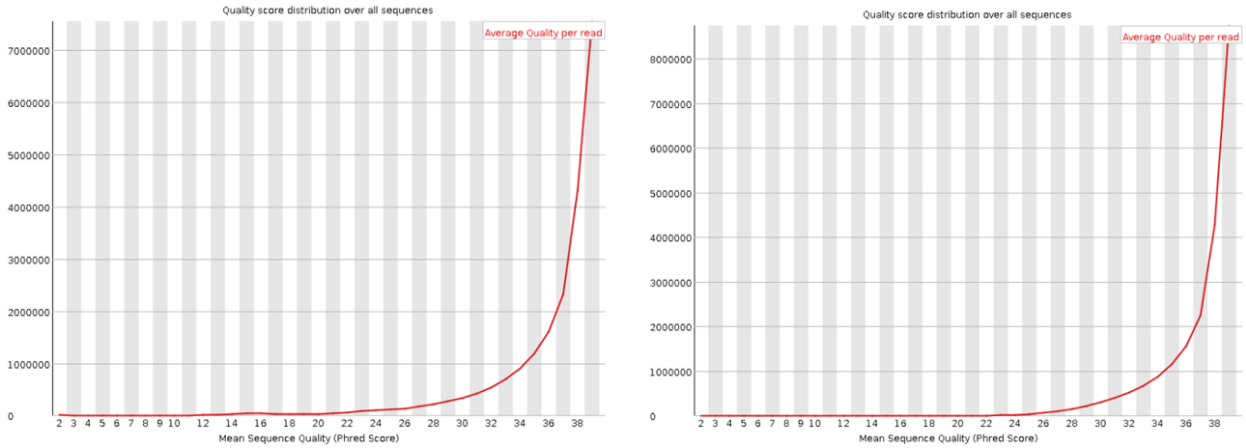

Figure D: Per sequence quality score. P0_1_1 (left), P0_1_2 (right)

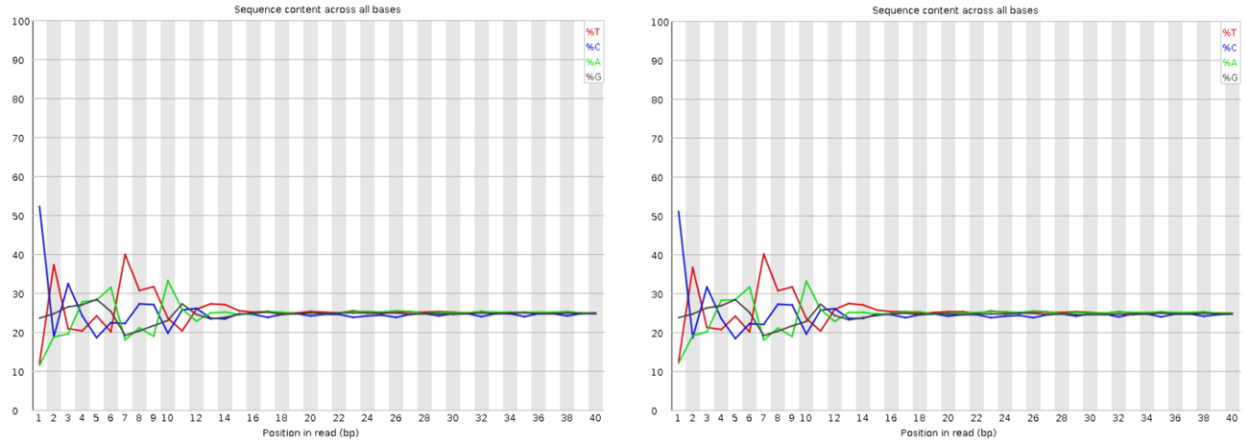

Figure E: Per base sequence content. P0_1_1 (left), P0_1_2 (right)

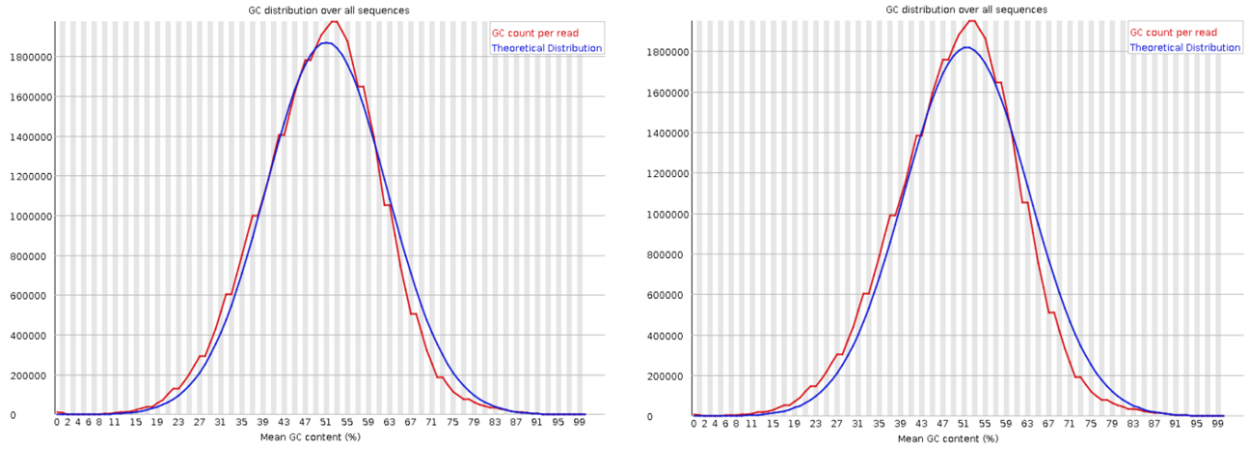

Figure F: Per sequence GC content. P0_1_1 (left), P0_1_2 (right)

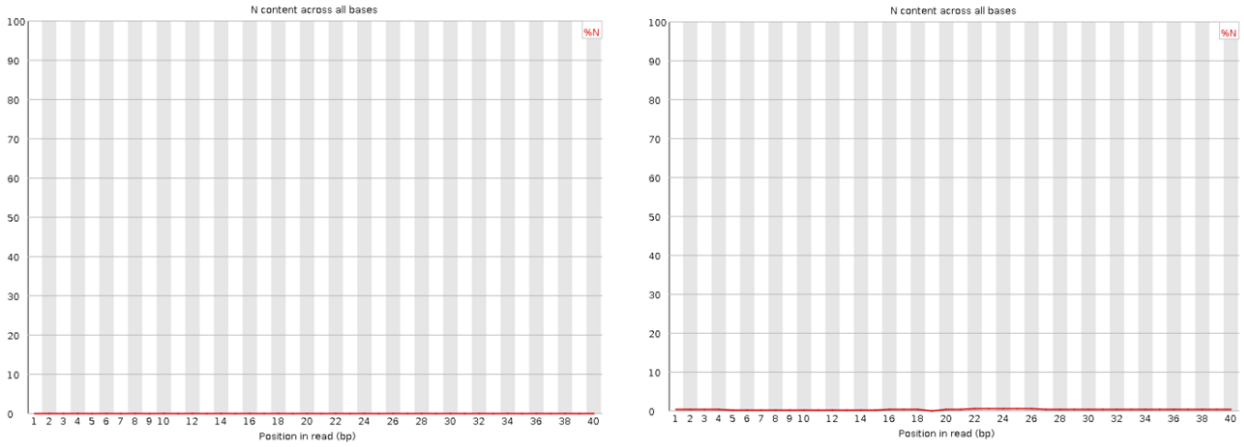

Figure G: Per base N content. P0_1_1 (left), P0_1_2 (right)

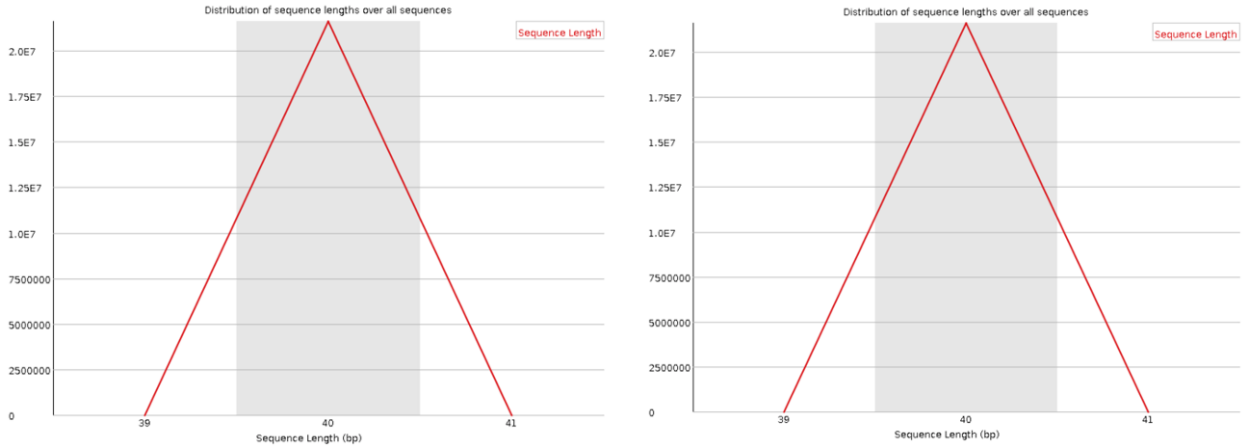

Figure H: Sequence length distribution. P0_1_1 (left), P0_1_2 (right)

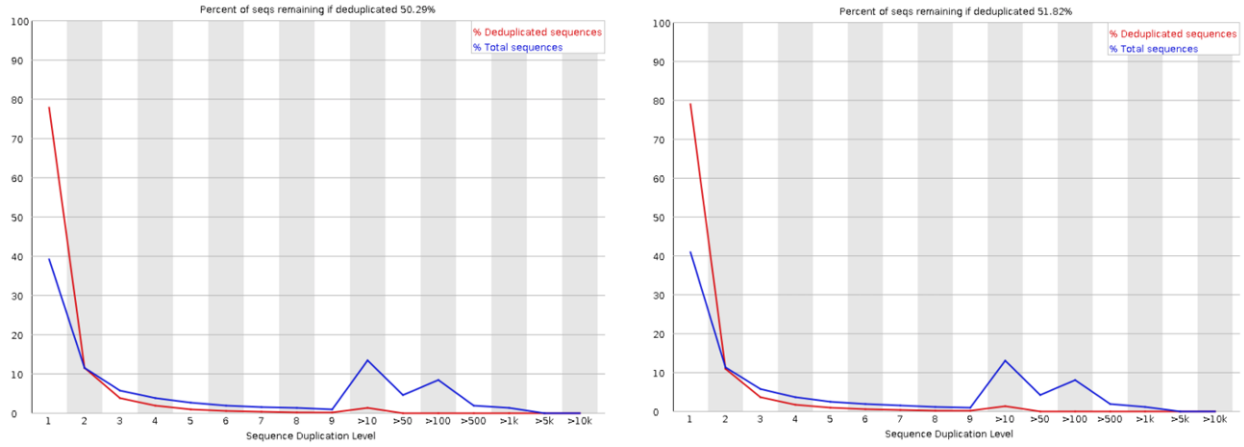

Figure I: Sequence duplication levels. P0_1_1 (left), P0_1_2 (right)

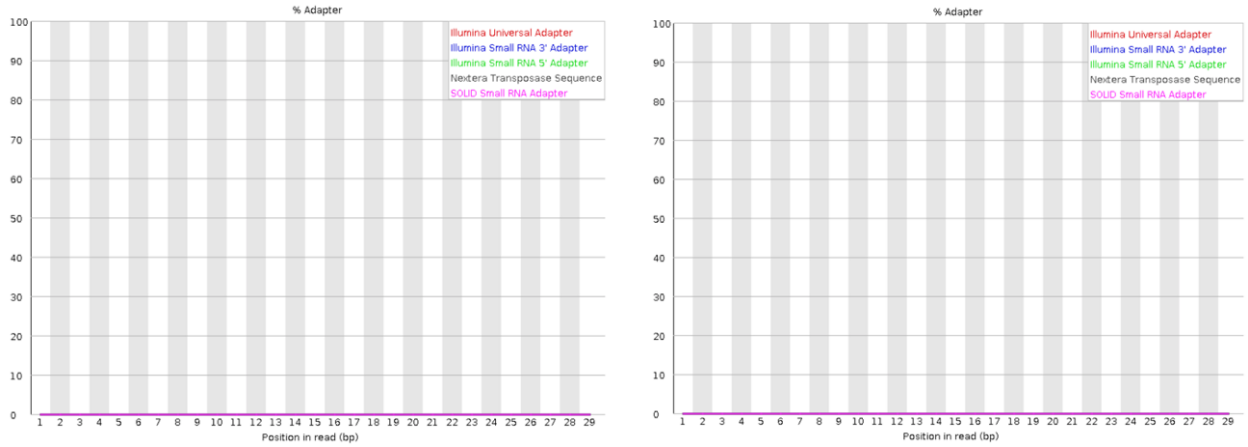

Figure J: Adapter content. P0_1_1 (left), P0_1_2 (right)