RNA Sequencing & Amplicon Sequencing

Nicholas Chludzinski

RNA-Sequencing Project 1

RNA-Seq
Finding a publication

NCBI (PubMed) publication (RNA seq experiment):
https://pubmed.ncbi.nlm.nih.gov/37649081/

Gene Expression Omnibus Page:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE224152

SRA Run Selector Page:
https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA929981&o=acc_s%3Aa

RNA-Seq
Publication Summary

- This study explores the taxonomy and molecular characteristics of human bone marrow mesenchymal stromal cells (BM-MSCs), endothelial cells (ECs), and mural cells (MCs) using single-cell RNA sequencing (scRNA-seq).
- Researchers identified distinct markers and pathways specific to each cell type, emphasizing the adipogenic transcriptomic pattern of MSCs and their roles in hematopoietic support and extracellular matrix production.
- The study also highlights inter-species conserved biomarkers like MARCKS, CXCL12, PDGFRA, and LEPR, which were validated as pivotal regulators of MSC functions across humans and mice.
- These findings deepen the understanding of BM cell heterogeneity, particularly MSCs, revealing their conserved gene expression and functional signatures related to skeletal development, vascular regulation, and hematopoiesis.
- The research provides insights for regenerative medicine and potential therapeutic strategies targeting age-related bone and hematopoietic dysfunctions.

RNA-Seq
FASTQ Preparation

```
#prefetch the SRR file

(opencv_env) C:\Users\nicho>prefetch SRR23290191 -O D:\SRA_Files -p

#convert the SRR file to a .fastq file

(opencv_env) C:\Users\nicho>fasterq-dump D:\SRA_Files\SRR23290188\SRR23290188.sra -O D:\SRA_Files -t D:\Temp -p

#concatenate/ combine all .fastq files into a single .fastq file for further processing

(opencv_env) C:\Users\nicho>cat D:\SRA_Files\SRR23290187.fastq D:\SRA_Files\SRR23290188.fastq
D:\SRA_Files\SRR23290189.fastq D:\SRA_Files\SRR23290190.fastq D:\SRA_Files\SRR23290191.fastq >
D:\SRA_Files\combined1.fastq
```

- **Data Retrieval**: Retrieved all (5) SRR files from NCBI using SRR_Acc_List.txt as a reference for SRR codes.
- **FASTQ Conversion**: Converted .sra files to .fastq format using fasterq-dump. Alternatively, utilized the Galaxy tool "Faster Download and Extract Reads in FASTQ format from NCBI SRA" for direct extraction due to storage and resource concerns.
- **File Consolidation**: Concatenated all 5 FASTQ files into a single dataset for downstream processing. Final combined file size: 31.2 GB.

RNA-Seq
Quality Control - FastQC Report

- The FastQC report highlighted strong quality metrics for the concatenated RNA-Seq dataset:
  - **Per Base Sequence Quality**: Consistently high quality across all base positions, with average scores exceeding 30, indicative of reliable sequencing accuracy.
  - **Sequence Length Distribution**: Uniform sequence lengths across reads, reflecting a well-prepared library with no truncation or anomalies.
  - **GC Content**: The observed GC content of 48% aligns well with typical RNA-Seq data, suggesting no significant contamination or bias.
  - **No Low-Quality Reads**: Zero sequences flagged as poor quality, further reinforcing the overall integrity of the dataset.

```
(opencv_env) C:\Users\nicho>cd D:\fastqc_v0.12.1\FastQC
```

```
(opencv_env) C:\Users\nicho>D:
```

```
(opencv_env) D:\fastqc_v0.12.1\FastQC>run_fastqc.bat D:\SRA_Files\SRR23290187.fastq -o D:\SRA_Files
```

RNA-Seq
Read Alignment

- This script outlines the workflow I would have implemented had sufficient storage and computational resources been available to process a 31.2 GB FASTQ file locally. It describes the critical steps required for RNA-Seq data alignment and gene expression analysis:
    - Aligning RNA-Seq reads to the human reference genome using HISAT2.
    - Converting the alignment output from SAM to BAM format and sorting it using SAMtools for downstream compatibility.
    - Generating a gene expression count matrix with featureCounts.

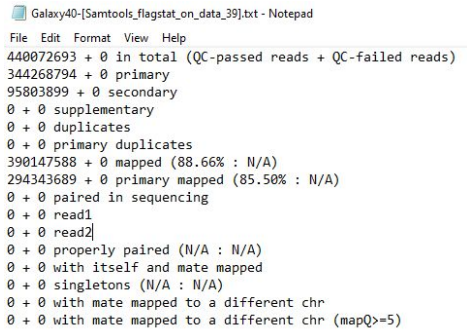
```
# Run HISAT2 for alignment using human reference genome and the concatenated .fastq file

# Output a SAM file and an alignment summary

(opencv_env) C:\Users\nicho> hisat2 -x D:/hg_38_genome -U concatenated_reads.fastq -S aligned_reads.sam
--summary-file alignment_summary.txt


# Convert the SAM file to a BAM file using samtools

(opencv_env) C:\Users\nicho> samtools view -Sb aligned_reads.sam -o aligned_reads.bam


# Sort the BAM file

(opencv_env) C:\Users\nicho> samtools sort -o aligned_reads_sorted.bam aligned_reads.bam


# Check files directory to make sure the BAM file was created

(opencv_env) C:\Users\nicho> dir


# Generate a gene_counts.txt file from the BAM file

(opencv_env) C:\Users\nicho> featureCounts -a annotation.gtf -o gene_counts.txt -t exon -g gene_id -s 0
aligned_reads_sorted.bam
```

RNA-Seq
Read Alignment

- **Resource Constraints**: Local alignment script using a 31.2 GB file was impractical due to high storage and computational requirements.
- **Solution**: HISAT2 on Galaxy: Utilized Galaxy's HISAT2, a fast and sensitive alignment tool optimized for next-generation sequencing (NGS) data. Supports alignment against the general human population or a specific reference genome.
- **Genome Selection**: Downloading a human reference genome was resource-intensive. Opted for Galaxy's built-in genome: *Homo sapiens* (b38), hg38.
- **Output Format**: Alignment results were generated as a BAM file, a standard format for storing aligned sequencing data.

RNA-Seq
Quality Metrics for Alignment - Flagstat

- Using samtools flagstat, a tool for summarizing alignment statistics from BAM files, the RNA-Seq alignment was evaluated. The alignment achieved a total mapping rate of 88.66%, with 85.50% primary reads mapped to the human reference genome. Flagstat confirmed no duplicate or improperly paired reads, highlighting the high quality of the alignment and the reliability of the data for downstream analysis.


Galaxy40-[Samtools_flagstat_on_data_39].txt - Notepad

File Edit Format View Help

```
440072693 + 0 in total (QC-passed reads + QC-failed reads)
344268794 + 0 primary
95803899 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
390147588 + 0 mapped (88.66% : N/A)
294343689 + 0 primary mapped (85.50% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

RNA-Seq
Gene Expression Quantification

- Gene expression quantification was performed using RNA-Seq data via the public tool featureCounts, with the BAM files generated from the HISAT2 alignment.
    - **Input**: The alignment results in BAM format, which contain mapped reads to the reference genome, were used as the foundation for quantification.
    - **Quantification Outputs**:
        - A raw counts table, capturing the number of reads aligned to each gene, was generated.
        - A gene lengths file, detailing the lengths of each gene, was produced to normalize the raw counts.
- The outputs serve as inputs for downstream calculations of RPKM (Reads Per Kilobase of transcript per Million mapped reads) and TPM (Transcripts Per Million). These normalized metrics are essential for comparing gene expression levels within and between datasets.

RNA-Seq
Normalization and Quality Control

- **Tools Used**:
  - Jupyter Notebook and Python were employed to perform comprehensive quality checks and normalization of RNA-Seq data.
- **Normalization Metrics**:
  - RPKM (Reads Per Kilobase of transcript per Million mapped reads): Normalizes read counts by gene length and sequencing depth to compare expression levels.
  - TPM (Transcripts Per Million): Accounts for gene length and sequencing depth, allowing comparison across samples.
- **Process Highlights**:
  - Quality checks were integrated to validate data integrity and ensure accurate quantification.
  - RPKM and TPM values were calculated to normalize gene expression levels.
  - Summary statistics provided insights into the data distribution and highlighted top expressed genes.
- **Output**:
  - .CSV File: A gene expression table containing RPKM and TPM values was generated for further analysis and reporting.

RNA-Seq
Code and Results

Code Summary:

- **Objective**: Process RNA-Seq data for gene expression quantification, ensuring data integrity and normalization.
- **Key Steps**:
  - Input validation: Checked for missing values, duplicate entries, and invalid gene lengths.
  - Calculations: Used information from gene counts and lengths files to compute RPKM (Reads Per Kilobase per Million reads) and TPM (Transcripts Per Million).
  - Data inspection: Identified top expressed genes and potential outliers.
  - Outputs: Generated summary statistics and gene expression metrics.

RNA-Seq
Code and Results contd.

Results Overview:

- **Data Quality**:
  - No Missing Values in counts or gene lengths.
  - No Negative Counts or Invalid Gene Lengths detected.
  - Total TPM is 1,000,000, confirming proper normalization.
- **Top Expressed Genes (TPM)**:
  - Highest TPM: Gene 3043 (90,880.49).
  - Other top genes showed consistently high expression values (e.g., Gene 3514, TPM: 70,101.89).
  - Higher TPM and RPKM values indicate higher gene expression levels.
- **Statistics**:
  - Gene counts ranged from 0 to 11.99 million reads.
  - Average TPM: 35.22 with some outliers indicating highly expressed genes.
- **Output Files**:
  - Normalized gene expression table with RPKM and TPM values.
  - Summary statistics and outlier information for expression levels.

RNA-Seq
Code and Results contd.

Top 10 Genes by Expression (TPM and RPKM)

- Highly Expressed Genes:
  - Gene 3043 exhibits the highest expression levels with TPM: 90,880.49 and RPKM: 91,503.82.
  - Gene 3514 follows closely with TPM: 70,101.89 and RPKM: 70,582.71.
- Consistent TPM and RPKM Values:
  - All top genes demonstrate close alignment between TPM and RPKM, indicating accurate normalization.
- Biological Relevance:
  - These top-expressed genes may represent essential genes or specific markers relevant to the sample.

```
Top 10 Genes by TPM (Higher TPM and RPKM values indicate higher gene expression levels):
        Geneid           TPM          RPKM
4047      3043  90880.492380  91503.824527
15266     3514  70101.892150  70582.707799
9286      6176  22730.718524  22886.624233
23280     6206  20938.380878  21081.993281
27072     4513  18157.806643  18282.347612
27076     4514  15123.080704  15226.807061
6950      6144  14625.015408  14725.325628
9674      3040  14325.377004  14423.632062
17774     3539  14216.525579  14314.034046
17772     3538  12542.763399  12628.791846
```

RNA-Seq
Code and Results contd.

Summary Statistics:

- Key Metrics Overview:
  - Dataset Size: 28,395 genes analyzed.
  - Mean Values:
    - Aligned Reads (BAM): 7,377 reads per gene.
    - RPKM: 35.46.
    - TPM: 35.22.
- Distribution Highlights:
  - High Variability:
    - Standard deviations:
      - BAM: 107,206 reads.
      - RPKM: 810.23.
      - TPM: 804.71.
    - Indicates a wide range of gene expression levels.
- Conclusion:
  - Mix of low-expression genes and highly expressed outliers

Summary Statistics:

	HISAT2 on data 35: aligned reads (BAM)	RPKM	TPM
count	2.839500e+04	28395.000000	28395.000000
mean	7.377300e+03	35.459018	35.217468
std	1.072064e+05	810.226226	804.706893
min	0.000000e+00	0.000000	0.000000
25%	0.000000e+00	0.000000	0.000000
50%	3.000000e+01	0.051625	0.051273
75%	1.006000e+03	1.295920	1.287092
max	1.199922e+07	91503.824527	90880.492380

Amplicon-Sequencing Project 2

Amplicon-Seq
Finding a publication

NCBI (PubMed) publication (Amplicon Sequencing Experiment):
https://pubmed.ncbi.nlm.nih.gov/32024835/

Gene Expression Omnibus Page:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143466

SRA Run Selector Page:
https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA600540&o=acc_s%3Aa

Amplicon-Seq
Publication Summary

- This study introduces a high-throughput method for identifying synthetic riboswitches in human cells through barcode-free cDNA amplicon sequencing (amplicon-seq).
- Researchers screened libraries of riboswitch constructs (~18,000 designs) in HEK-293 cells, identifying tetracycline- and guanine-responsive ON/OFF riboswitches with ligand-dependent gene expression control.
- The study demonstrated the discovery of Twister ribozyme-based switches and U1-snRNP-dependent RNA switches, expanding the functional repertoire of synthetic riboswitches in gene expression regulation.
- Key findings include the identification of high-sensitivity constructs with minimal false positives and insights into sequence motifs affecting riboswitch functionality and ligand responsiveness.
- These advancements highlight the potential of amplicon-seq for designing RNA-based tools for gene therapy and other biotechnological applications, with significant implications for therapeutic gene expression control.

Amplicon-Seq
FASTQ Preparation

- **Data Retrieval**: Retrieved 9/ 125 SRR files from NCBI using SRR_Acc_List.txt as a reference for SRR codes. I opted only for the 30 uM treatment as downloading and processing 125 SRR files and converting them into FASTQ files would be infeasible due to storage and computational concerns.
- **FASTQ Conversion**: Converted .sra files to .fastq format using fasterq-dump. Alternatively, utilized the Galaxy tool "Faster Download and Extract Reads in FASTQ format from NCBI SRA" for direct extraction due to storage and resource concerns.
- **File Consolidation**: Concatenated all 5 FASTQ files into a single dataset for downstream processing. Final combined file size: 6.8 GB.

Amplicon-Seq
Quality Control - FastQC Report

- The FastQC report for the amplicon sequencing dataset indicates overall strong quality metrics:
  - **Per Base Sequence Quality**: High-quality scores across all base positions, with mean values consistently above 30, ensuring accurate sequencing results.
  - **Sequence Length Distribution**: Slight variability in read lengths (152-154 bp), which may require further review but generally indicates good library preparation.
  - **GC Content**: A GC content of 45%, typical for amplicon sequencing, reflecting balanced base composition without significant bias.
  - **No Poor-Quality Reads**: Zero sequences flagged as poor quality, demonstrating the dataset's high reliability and sequencing precision.

Amplicon-Seq
Read Alignment, Variant Calling and VCF Filter

- This script outlines the workflow I would have implemented had sufficient storage and computational resources been available to process amplicon sequencing data locally. It describes the critical steps required for high-depth, targeted variant calling analysis:
    - Aligning Amplicon Reads: The concatenated FASTQ file was aligned to the human reference genome using BWA-MEM2.
    - Converting and Sorting: The alignment output was converted from SAM to BAM format and sorted using SAMtools.
    - Variant Calling with FreeBayes: FreeBayes was used to identify genetic variants within the predefined genomic loci.
    - Filtering Variants: The raw variant call set was refined using VCFfilter, applying stringent quality and depth thresholds to retain only high-confidence variants.

```
# Align the concatenated FASTQ file to the human reference genome using BWA-MEM2

# Output: SAM file

(opencv_env) C:\Users\nicho> bwa-mem2 mem -t 4 -o aligned_reads.sam D:/hg_38_genome.fa
concatenated_reads.fastq


# Convert the SAM file to BAM format using samtools

(opencv_env) C:\Users\nicho> samtools view -Sb -o aligned_reads.bam aligned_reads.sam


# Sort the BAM file

(opencv_env) C:\Users\nicho> samtools sort -o aligned_reads_sorted.bam aligned_reads.bam


# FreeBayes for variant calling

# Output: VCF file

(opencv_env) C:\Users\nicho> freebayes -f D:/hg_38_genome.fa -v variants.vcf aligned_reads_sorted.bam


# Filter the VCF file for high-quality variants

# Use vcffilter to retain variants with a QUAL score >30 and depth >10

(opencv_env) C:\Users\nicho> vcffilter -f "QUAL > 30 & DP > 10" -o high_quality_variants.vcf variants.vcf


# Verify files are in the current directory

(opencv_env) C:\Users\nicho> dir
```

Amplicon-Seq
Read Alignment

- **Resource Constraints**: Local alignment of amplicon-seq data required significant computational power and storage capacity, making standalone analysis less feasible.
- **Solution**: BWA-MEM2 on Galaxy: Implemented BWA-MEM2, a faster and memory-efficient alignment tool designed for high-throughput sequencing data. Galaxy provided a streamlined workflow, eliminating the need for standalone hardware setup.
- **Reference Genome**: Used a pre-indexed human genome reference (GRCh38) available in Galaxy, avoiding the resource-heavy process of downloading and indexing locally.
- **Output Format**: Generated aligned sequencing data in BAM format, a widely accepted standard compatible with downstream analysis tools such as variant calling and visualization platforms.

Amplicon-Seq Variant Calling

FreeBayes

- **Purpose**: FreeBayes identifies genetic variants (SNPs, indels, and structural variations) from aligned sequencing data.
- **Input Data**: BAM file generated from BWA-MEM2, containing aligned reads to a reference genome.
- Sensitive to small indels and variants in highly repetitive regions.
- **Output Format**: Produces VCF files listing variants with details such as position, type, and quality scores.
- **Significance**: Ensures accurate detection of mutations critical for downstream analyses in amplicon sequencing studies.
- **Efficient Workflow**: The BWA-MEM2 and FreeBayes combination offers a streamlined pipeline tailored for identifying variants in predefined genomic loci. In contrast, HISAT2, designed for transcriptomics and large genome-wide alignments, is less suited for the high-depth, small-target focus of amplicon sequencing.

Amplicon-Seq VCF Filter

- **Purpose**: Filtering variant call format (VCF) files refines raw variant calls by removing low-confidence or irrelevant data, ensuring high-quality downstream analysis.
- **Filtering Criteria**:
  - **Quality Score (QUAL)**: Retained variants with a QUAL score > 30, indicating high confidence.
  - **Depth of Coverage (DP)**: Selected variants with a depth > 10 to ensure sufficient read support.
  - **Genotype Quality (GQ)**: Ensured genotype reliability by setting a GQ threshold > 20.
- **Output**: A cleaned and concise VCF file containing only high-quality, relevant variants for downstream interpretation and visualization.

Amplicon-Seq
Convert to .tabular

VCF to Tab-Delimited Conversion

- **Purpose**: Converting VCF files to a tab-delimited format simplifies data handling and enables easier integration with downstream tools for visualization and analysis.
- **Key Features**:
  - Extracts essential variant information, including chromosome, position, reference/alternate alleles, and quality metrics.
  - Includes sample-specific genotype information for detailed analysis.
- **Output Format**:
  - .tabular file with organized columns for straightforward processing.
- **Advantages**:
  - Facilitates rapid filtering, sorting, and statistical analysis.
  - Enhances compatibility with data visualization tools for presenting variant distributions and impacts.

Amplicon-Seq
Code and Results

Code Summary:

- **Objective**: Analyze tab-delimited VCF data to filter high-quality variants and visualize quality metrics.
- **Key Steps**:
  - Data Loading: Imported tab-delimited VCF data into a DataFrame using Pandas.
  - Quality Checks: Inspected total variants, missing values in QUAL, and generated summary statistics.
  - Filtering: Applied thresholds (QUAL > 30 and DP > 10) to retain high-confidence variants.
- **Output**:
  - Saved high-quality variants to a CSV file for downstream analysis.
  - Extracted and displayed the top 10 variants ranked by QUAL.
  - Summary statistics for QUAL and top-ranked variants

Amplicon-Seq
Code and Results contd.

Results Overview:

- **Data Quality**:
  - No Missing Values detected in QUAL or key metrics.
  - All variants passed basic integrity checks (e.g., valid numerical values for QUAL and DP).
  - High-Quality Variants: All 76 variants met the filtering criteria (QUAL > 30 and DP > 10).
- **Top Variants**:
  - Highest Quality Variant: chr15:25338489 (QUAL: 4,700,200), exhibits exceptionally high quality, indicating a strong confidence in true genetic changes.
  - Other top variants include chr11:5225911 (QUAL: 1,701,690) and chr11:5226561 (QUAL: 221,051).

Amplicon-Seq
Code and Results contd.

Top 10 Variants by Quality:

	#CHROM	POS	ID	REF	ALT	QUAL
16	chr15	25338489	.	A	G	4700200.0
6	chr11	5225911	.	G	A	1701690.0
8	chr11	5226561	.	C	G	221051.0
7	chr11	5226503	.	A	C	209966.0
40	chr2	32916486	.	A	G	122493.0
38	chr2	32916441	.	A	G	79224.0
37	chr2	32916401	.	A	G	23559.1
65	chr9	92287865	.	C	T	11298.3
41	chr2	32916557	.	A	T	10466.9
66	chr9	92288225	.	T	G	10417.8

Explanation of Key Columns in Variant Calling Results:

- #CHROM (Chromosome):
  - Represents the chromosome where the variant is located (e.g., chr15, chr11).
  - In this dataset, variants are found across chromosomes 15, 11, 2, and 9, indicating multiple loci with potential genetic changes.
- POS (Position):
  - Indicates the precise base-pair position of the variant on the chromosome.
  - For example, the top variant is located at position 25,338,489 on chromosome 15.
- ID (Identifier):
  - Refers to a unique database identifier for the variant, such as from dbSNP.
  - A value of . means no known identifier is associated with the variant, suggesting a novel or less-studied mutation.
- REF (Reference Allele):
  - The reference allele represents the expected nucleotide base at the given position in the human reference genome.
- ALT (Alternate Allele):
  - The alternate allele shows the observed nucleotide base that differs from the reference genome, describing a change on that single strand.
  - Variants here include changes like A → G or G → A, which may represent a point mutation (a purine is replaced with another purine), and these changes may occur because of replication errors or environmental factors.
- QUAL (Quality Score):
  - Reflects the confidence level of the variant being a true genetic change rather than a sequencing artifact.
  - Scores in this dataset range from 10,417.8 to 4,700,200.0, with higher scores indicating greater reliability.
  - For instance, the top variant on chromosome 15 with a quality score of 4,700,200.0 has the highest confidence of being a true variant.
- Summary:
  - Variants are distributed across multiple chromosomes, with the highest confidence mutation on chromosome 15 (position 25,338,489).
  - Reference (REF) and alternate (ALT) alleles highlight the specific base changes, which may have biological implications.
  - High quality scores (e.g., >1,000,000) indicate strong evidence supporting the validity of these variants.

Amplicon-Seq
Code and Results contd.

High-Quality Variants Summary Statistics:

	QUAL
count	7.600000e+01
mean	9.397043e+04
std	5.707527e+05
min	4.131900e+01
25%	9.268080e+01
50%	2.487570e+02
75%	8.517785e+02
max	4.700200e+06

Summary Statistics:

- Overview:
  - This slide presents the key statistical metrics for the quality scores (QUAL) of high-quality variants identified in the analysis.
- Statistics:
  - Count: 76 variants were classified as high-quality.
  - Mean (Average QUAL): 93,970.43, indicating high confidence across the dataset.
  - Standard Deviation (STD): 570,752.7, reflecting a broad range of quality scores due to the presence of exceptionally high-quality variants.
  - Minimum (Min QUAL): 41.32, representing the threshold of inclusion as a high-quality variant.
  - 25th Percentile (25%): 92.68, indicating that 25% of variants had a quality score below this value.
  - Median (50th Percentile): 248.76, showcasing that half of the variants had a quality score above this value.
  - 75th Percentile (75%): 851.78, demonstrating that the majority of variants had strong quality scores.
  - Maximum (Max QUAL): 4,700,200.0, representing the highest-quality variant (e.g., chr15:25338489).
- Key Takeaway:
  - The wide range of quality scores reflects a mix of moderate to highly confident variants, with an exceptionally high maximum QUAL score validating the detection of very reliable variants.