CCR-Sequencing Facility Illumina Sequencing Report

Project Information

Principal Investigator: Stefan Ambs

PI Laboratory Contact: Gatikrushna Panigrahi

Bioinformatics Contact: Tang, Wei

Project Title: StefanAmbs_CS029103_24RNA_081221

NAS Order ID: CS029103

Samples Total in project: 24

Samples in This Report: 24

Completion of NAS: Yes

Report Date: 10/07/21

Sequencing Details

Flowcell ID:	AAAMV57M5	Sequence Control:	PhiX
Instrument Type:	NextSeq 2000	Control Result:	Pass
Flowcell Type:	P2	Library Protocol:	TruSeq Stranded mRNA Prep
Sequencing Type:	mRNA-Seq	Reference Genome:	hg38
Read Length:	R1:101, i7:8, i5:8, R2:101	Annotation:	GENCODE_30 GTF
Strand Specificity:	Stranded		

Run Comments

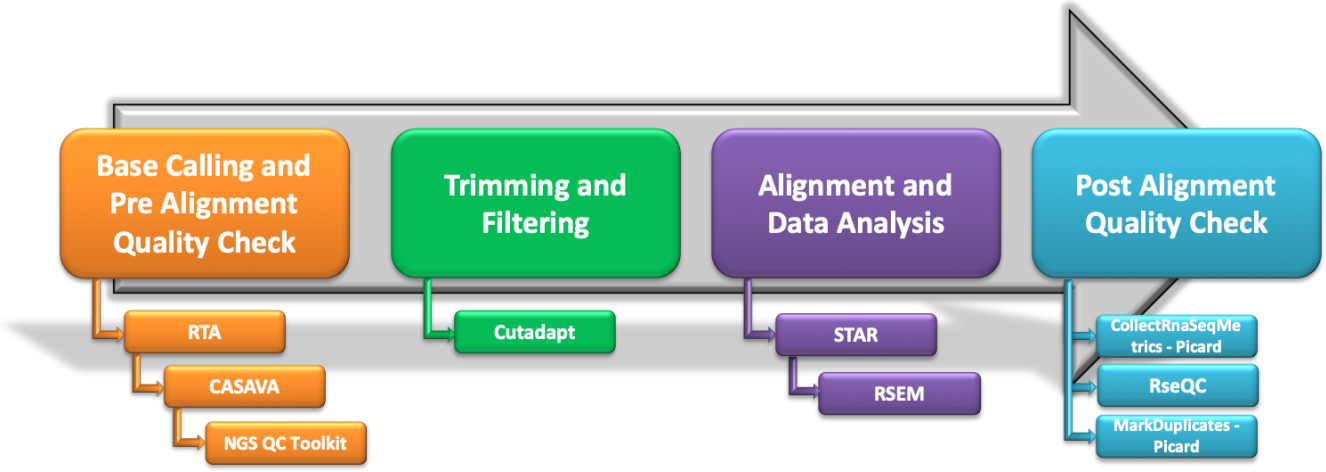
24 mRNA-Seq samples were pooled and sequenced on NextSeq 2000 P2 using TruSeq Stranded mRNA Prep and paired-end sequencing. The samples have 35 to 43 million pass filter reads with more than 92% of bases above the quality score of Q30. Reads of the samples were trimmed for adapters and low-quality bases using Cutadapt before alignment with the reference genome (hg38) and the annotated transcripts using STAR. The average mapping rate of all samples is 97%. Unique alignment is above 85%. There are 2.16 to 8.91% unmapped reads. The mapping statistics are calculated using Picard software. The samples have 0.00% ribosomal bases. Percent coding bases are between 60-64%. Percent UTR bases are 30-34%, and mRNA bases are between 93-95% for all the samples. Library complexity is measured in terms of unique fragments in the mapped reads using Picard's MarkDuplicate utility. The samples have 69-75% non-duplicate reads. In addition, the gene expression quantification analysis was performed for all samples using STAR/RSEM tools. Both the normalized count and the raw count are provided as part of the data delivery.

Note: *Residual samples will be retained up to* ***90 days*** *of the delivery of this report. To avoid shipping charges, please contact SFILLUMINALAB@mail.nih.gov to arrange pickup samples prior to this time.*

Note: *Sequencing data will be available to download for* ***two weeks*** *following delivery of this report. Please download the data files as soon as possible.*

For questions on any aspect of this report please contact CCRSF_IFX@nih.gov.

Analysis Workflow

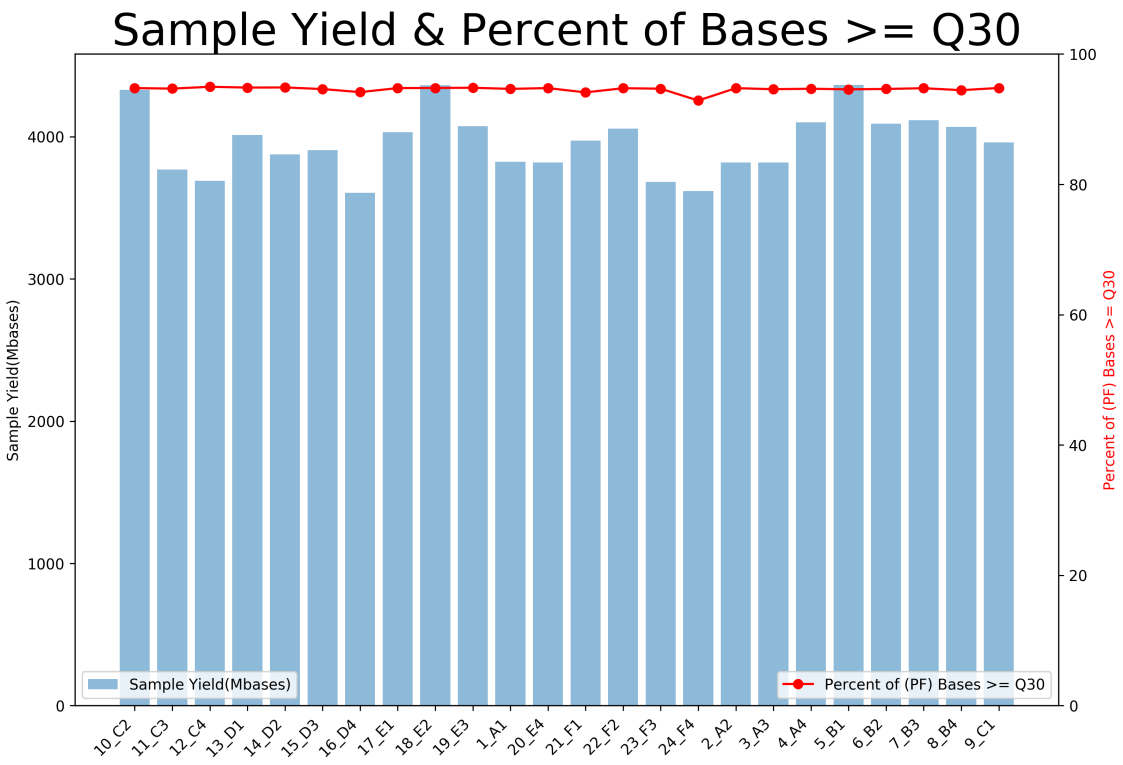


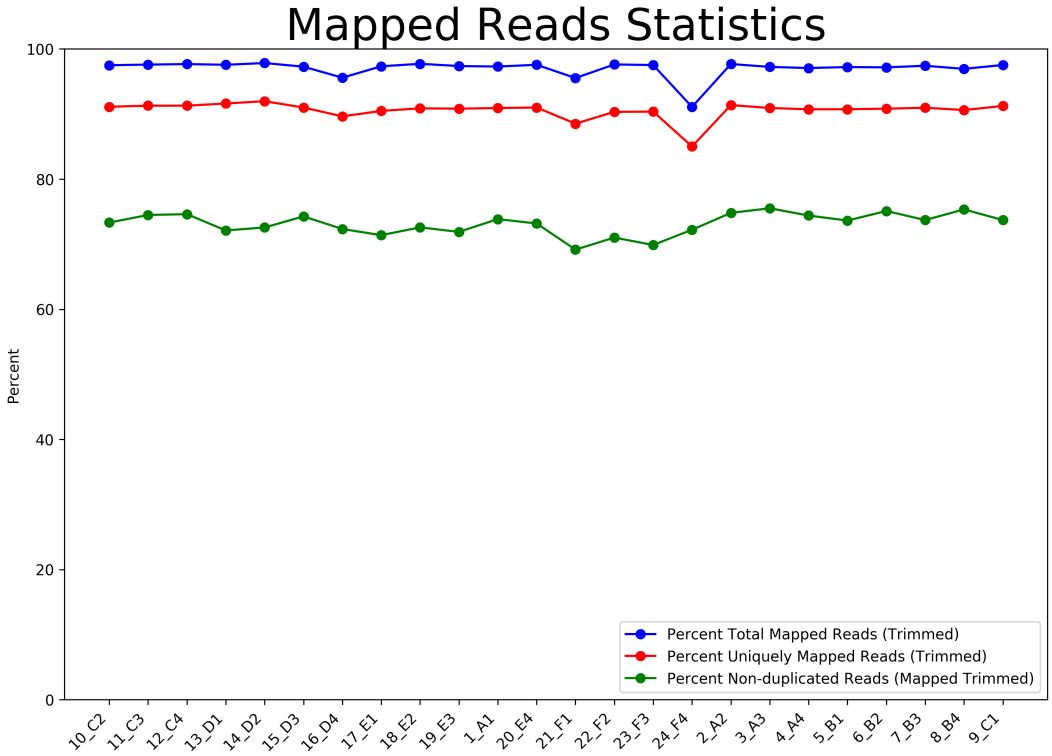
Software and Parameters

Analysis Step	Software	Software Parameters / Notes
Basecalling	RTA 3.9.2	Illumina instrument run time analysis software
Demultiplexing	Bcl2fastq v2.20	--no-lane-splitting -i RunFolder/Data/Intensities/BaseCalls -R RunFolder --barcode-mismatches 1 --ignore-missing-bcls --ignore-missing-filter --ignore-missing-positions --ignore-missing-controls --sample-sheet SampleSheet.csv -o Unaligned
Filtering (Adaptor and quality)	Cutadapt 1.18	-j 8 -b file:adapters.fa -B file:adapters.fa --nextseq-trim=2 --trim-n -n 5 -O 5 -q 10,10 -m 35:35 -o trimmed_R1.fq -p trimmed_R2.fq input_R1.fq input_R2.fq
Alignment	STAR 2.7.0f	1-pass: --genomeDir $star_genome --outSAMunmapped Within --outFilterType BySJout --outFilterMultimapNmax 20 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --sjdbScore 1 --readFilesCommand zcat --readFilesIn $trimmed_R1.fastq.gz $trimmed_R2.fastq.gz --runThreadN numThreads --outFilterMatchNminOverLread 0.66 --outSAMtype BAM Unsorted --quantMode TranscriptomeSAM --peOverlapNbasesMin 10 --alignEndsProtrude 10 ConcordantPair 2-pass: --genomeDir $star_genome --outSAMunmapped Within --outFilterType BySJout --outFilterMultimapNmax 20 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --alignSJoverhangMin 8 --limitSjdbInsertNsj 2500000 --sjdbFileChrStartEnd $input_1-path_sj --alignSJDBoverhangMin 1 --sjdbScore 1 --readFilesCommand zcat --readFilesIn $trimmed_R1.fastq.gz $trimmed_R2.fastq.gz --runThreadN $numhreads --outFilterMatchNminOverLread 0.66 --outSAMtype BAM Unsorted --quantMode TranscriptomeSAM --peOverlapNbasesMin 10 --alignEndsProtrude 10 ConcordantPair

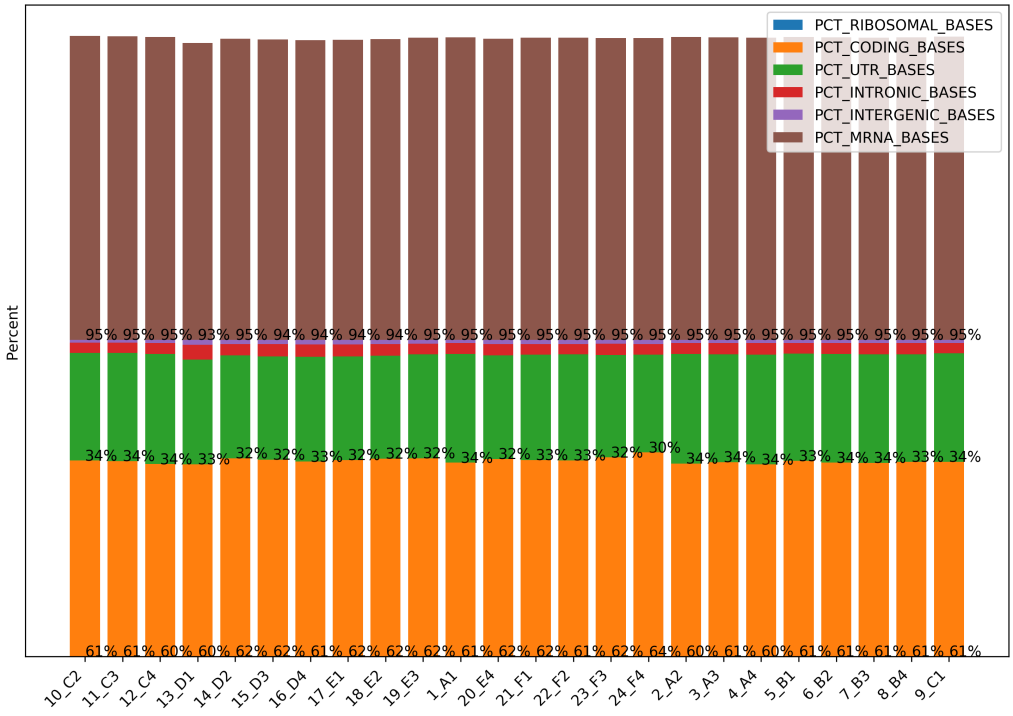
RNAStatistics	Picard 2.18.26	CollectRnaSeqMetrics.jar REF_FLAT=annotation_refFlat.txt INPUT=sample.bam OUTPUT= RnaSeqMetrics.txt RIBOSOMAL_INTERVALS= ribosome_interval_list.txt STRAND_SPECIFICITY=SECOND_READ_TRANSCRIPTION_STRAND VALIDATION_STRINGENCY=LENIENT
Duplication Statistics	Picard 2.18.26	MarkDuplicates.jar INPUT=sample.bam OUTPUT=sample.MKDUP.bam METRICS_FILE=sample.bam.metric ASSUME_SORTED=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000 VALIDATION_STRINGENCY=LENIENT
Quantification	RSEM 1.3.1	rsem-calculate-expression -bam --paired-end --estimate-rspd Transcriptome.out.bam $RSEM_Genome $Sample_Name

Data Statistics




Mapped Reads Statistics
Percent
100
80
60
40
20
0
Percent Total Mapped Reads (Trimmed)
Percent Uniquely Mapped Reads (Trimmed)
Percent Non-duplicated Reads (Mapped Trimmed)
10_C2
11_C3
12_C4
13_D1
14_D2
15_D3
16_D4
17_E1
18_E2
19_E3
1_A1
20_E4
21_F1
22_F2
23_F3
24_F4
2_A2
3_A3
4_A4
5_B1
6_B2
7_B3
8_B4
9_C1

RNA Statistics



Notes

- **Sample Yield –** The sum of all bases in reads that passed filtering per sample. Indicates the output in million bases (Mb) per lane.
- **% >=Q30 –** The percentage of bases called with an inferred accuracy of 99.9% or above, a measure of basecalling quality.
- **% Total (Primary) Alignment –** The percentage of filtered reads that align to the reference; for mRNA-seq, to the reference genome and the splice junctions. Reads aligning to multiple locations are included in the calculation
- **% Unique Alignment –** The percentage of filtered reads that align uniquely to the reference; for mRNA-Seq, the reference genome and known splice junctions. Reads aligning to multiple locations and abundant sequences are not included in the score.
- **% Non-duplicated Reads –** The percentage of aligned reads with non-redundant start coordinate.
- **% RNA Statistics –** Collect metrics about the alignment of RNA to various functional classes of loci in the genome: coding, intronic, UTR, intergenic, ribosomal. Also determines strand-specificity for strand-specific libraries.

  PCT_RIBOSOMAL_BASES: RIBOSOMAL_BASES / PF_ALIGNED_BASES

  PCT_CODING_BASES: CODING_BASES / PF_ALIGNED_BASES

  PCT_UTR_BASES: UTR_BASES / PF_ALIGNED_BASES

  PCT_INTRONIC_BASES: INTRONIC_BASES / PF_ALIGNED_BASES

  PCT_INTERGENIC_BASES: INTERGENIC_BASES / PF_ALIGNED_BASES

  PCT_MRNA_BASES: PCT_UTR_BASES + PCT_CODING_BASES