CCR-Sequencing Facility Illumina Sequencing Report

Project Information

Principal Investigator: Jing Wu

PI Laboratory Contact: Madison Butler

Bioinformatics Contact: CCBR, Parthav Jailwala

Project Title: JingWu_CS026381_44RNA_021320

NAS Order ID: CS026381

Samples Total in project: 44

Samples in This Report: 44

Completion of NAS: Yes

Report Date: 03/14/20

Sequencing Details

Flowcell ID:	HJJKWDRXX	Sequence Control:	PhiX
Instrument:	NovaSeq	Control Result:	Pass
Sequencing Type:	mRNA-Seq	Library Protocol:	**TruSeq Stranded mRNA Library Prep**
Read Length:	151 (2x151 cycles)	Sequencing Chemistry:	NovaSeq
Multiplexing:	44 per lane	Reference Genome:	Human_hg38
Strand Specificity:	**Stranded**	Annotation:	**Ensembl96_30 GTF**

Run Comments

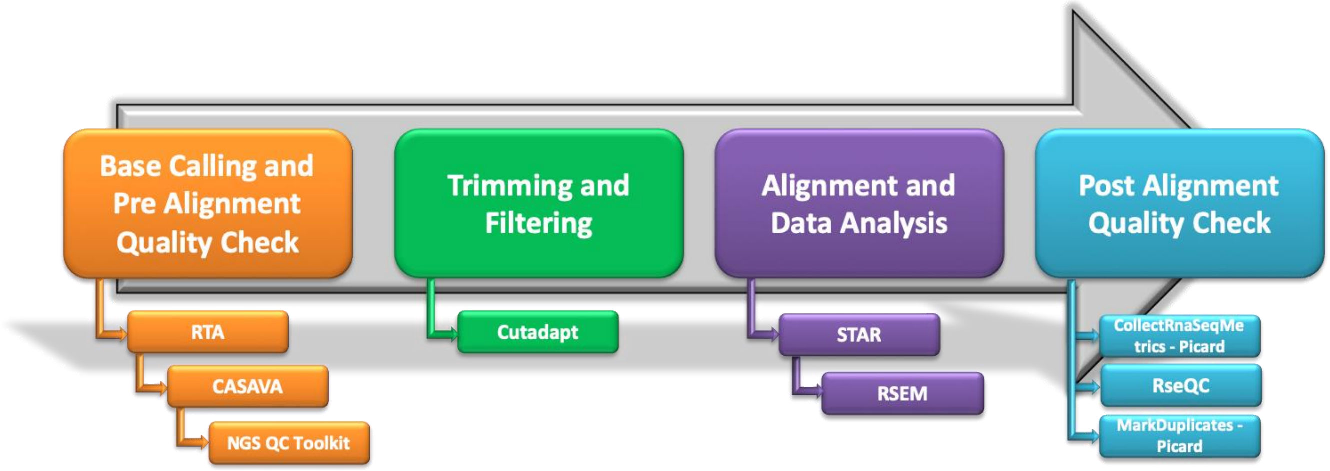
44 mRNA-Seq samples were pooled and sequenced on NovaSeq_SP using Illumina TruSeq Stranded mRNA Library Prep and paired-end sequencing. The samples have 33 to 50 million pass filter reads with more than 94% of bases above the quality score of Q30. Reads of the samples were trimmed for adapters and low-quality bases using Cutadapt before alignment with the reference genome (Human_hg38) and the annotated transcripts using STAR. The average mapping rate of all samples is 95%. Unique alignment is above 90%. There are 3.35 to 5.93% unmapped reads. The mapping statistics are calculated using Picard software. The samples have 0.00% ribosomal bases. Percent coding bases are between 54-60%. Percent UTR bases are 32-35%, and mRNA bases are between 89-92% 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 72-77% 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

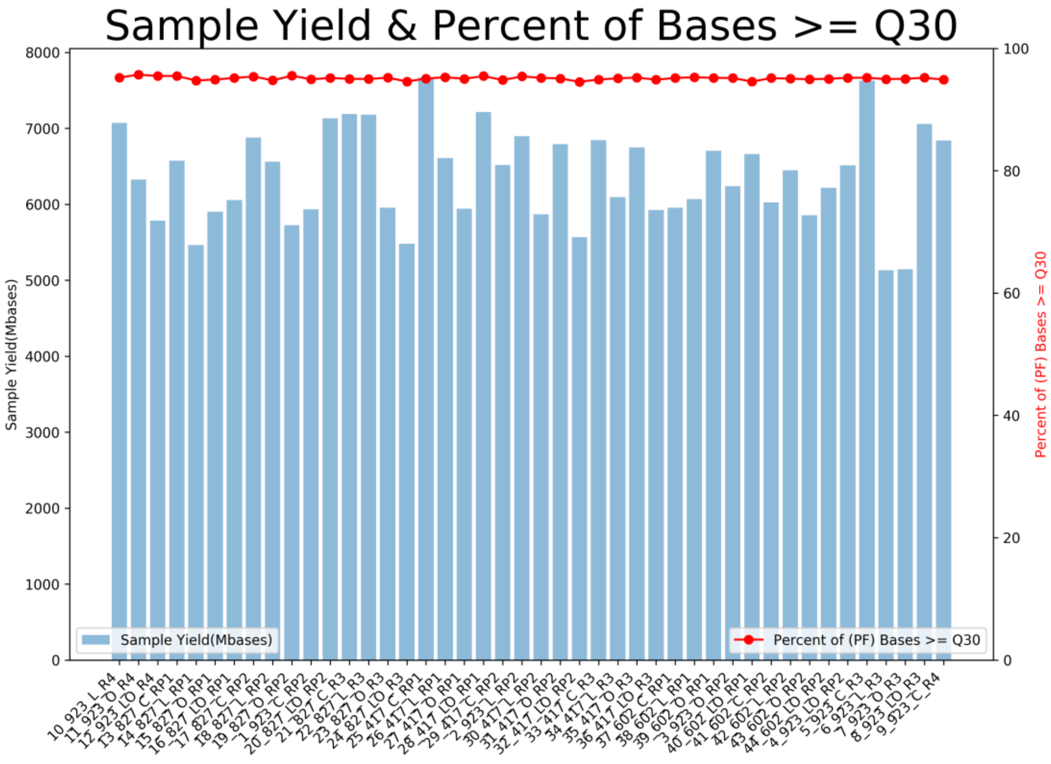

Base Calling and Pre Alignment Quality Check
RTA
CASAVA
NGS QC Toolkit
Trimming and Filtering
Cutadapt
Alignment and Data Analysis
STAR
RSEM
Post Alignment Quality Check
CollectRnaSeqMetrics - Picard
RseQC
MarkDuplicates - Picard

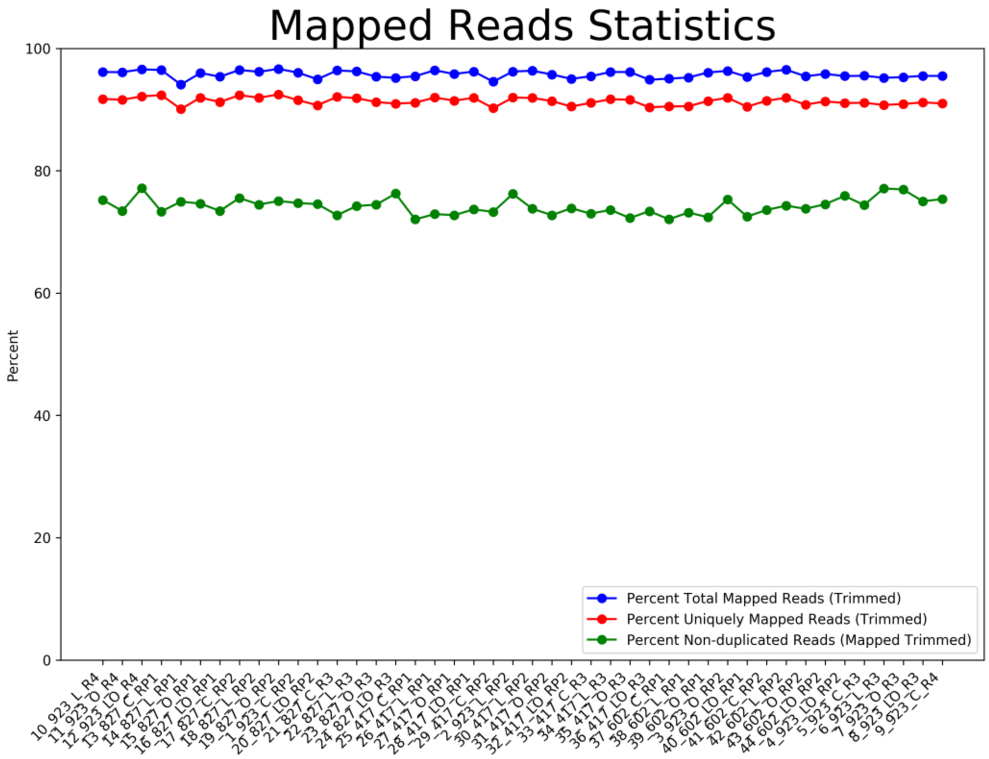
Software and Parameters

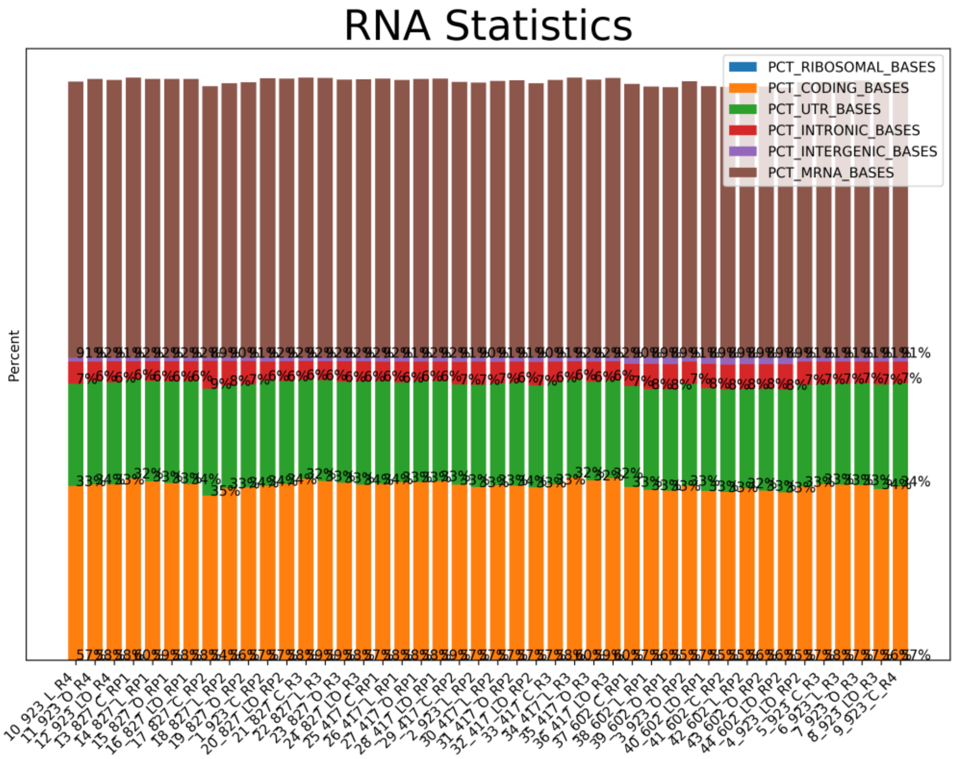
Analysis Step	Software	Software Parameters / Notes
Basecalling	RTA v3.4.4	Illumina instrument run time analysis software
Demultiplexing	Bcl2fastq v2.17	--no-lane-splitting -i RunFolder/Data/Intensities/BaseCalls -R RunFolder --barcode-mismatches 1 --ignore-missing-bcls --ignore-missing-filter --ignore-missing-positions --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_923_L_R4
11_923_O_R4
12_923_LO_R4
13_827_C_RP1
14_827_L_RP1
15_827_O_RP1
16_827_LO_RP1
17_827_C_RP2
18_827_L_RP2
19_827_O_RP2
1_923_C_RP2
20_827_LO_RP2
21_827_C_R3
22_827_L_R3
23_827_O_R3
24_827_LO_R3
25_417_C_RP1
26_417_L_RP1
27_417_O_RP1
28_417_LO_RP1
29_417_C_RP2
2_923_L_RP2
30_417_L_RP2
31_417_O_RP2
32_417_LO_RP2
33_417_C_R3
34_417_L_R3
35_417_O_R3
36_417_LO_R3
37_602_C_RP1
38_602_L_RP1
39_602_O_RP1
3_923_O_RP2
40_602_LO_RP1
41_602_C_RP2
42_602_L_RP2
43_602_O_RP2
44_602_LO_RP2
4_923_LO_RP2
5_923_C_R3
6_923_L_R3
7_923_O_R3
8_923_LO_R3
9_923_C_R4


RNA Statistics
Percent
PCT_RIBOSOMAL_BASES
PCT_CODING_BASES
PCT_UTR_BASES
PCT_INTRONIC_BASES
PCT_INTERGENIC_BASES
PCT_MRNA_BASES
57% 58% 58% 60% 59% 58% 58% 54% 56% 57% 57% 58% 59% 59% 58% 57% 58% 58% 58% 59% 57% 57% 57% 57% 57% 58% 60% 59% 60% 57% 56% 55% 57% 55% 55% 56% 56% 55% 57% 58% 57% 57% 56% 57%
10_923_L_R4
11_923_O_R4
12_923_LO_R4
13_827_C_RP1
14_827_L_RP1
15_827_O_RP1
16_827_LO_RP1
17_827_C_RP2
18_827_L_RP2
19_827_O_RP2
1_923_C_RP2
20_827_LO_RP2
21_827_C_R3
22_827_L_R3
23_827_O_R3
24_827_LO_R3
25_417_C_RP1
26_417_L_RP1
27_417_O_RP1
28_417_LO_RP1
29_417_C_RP2
2_923_L_RP2
30_417_L_RP2
31_417_O_RP2
32_417_LO_RP2
33_417_C_R3
34_417_L_R3
35_417_O_R3
36_417_LO_R3
37_602_C_RP1
38_602_L_RP1
39_602_O_RP1
3_923_O_RP2
40_602_LO_RP1
41_602_C_RP2
42_602_L_RP2
43_602_O_RP2
44_602_LO_RP2
4_923_LO_RP2
5_923_C_R3
6_923_L_R3
7_923_O_R3
8_923_LO_R3
9_923_C_R4

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

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