CIDC 2.0 Bioinformatic Pipelines
RNA-Seq Validation

Essex Management

National Cancer Institute (NCI)

03/20/2024

Version 1.0

FINAL

Submitted To:
Daoud Meerzaman
Computational Genomics and Bioinformatics Branch (CGBB)
National Cancer Institute
Center for Biomedical Informatics & Information Technology
Rockville, MD 20850

Submitted By:
Nick Renzette, Jennifer Harvey
Essex Management, LLC
11140 Rockville Pike | Suite 332
Rockville, MD 20852-3149
DUNS: 829872345
Cage Code: 5CYC9

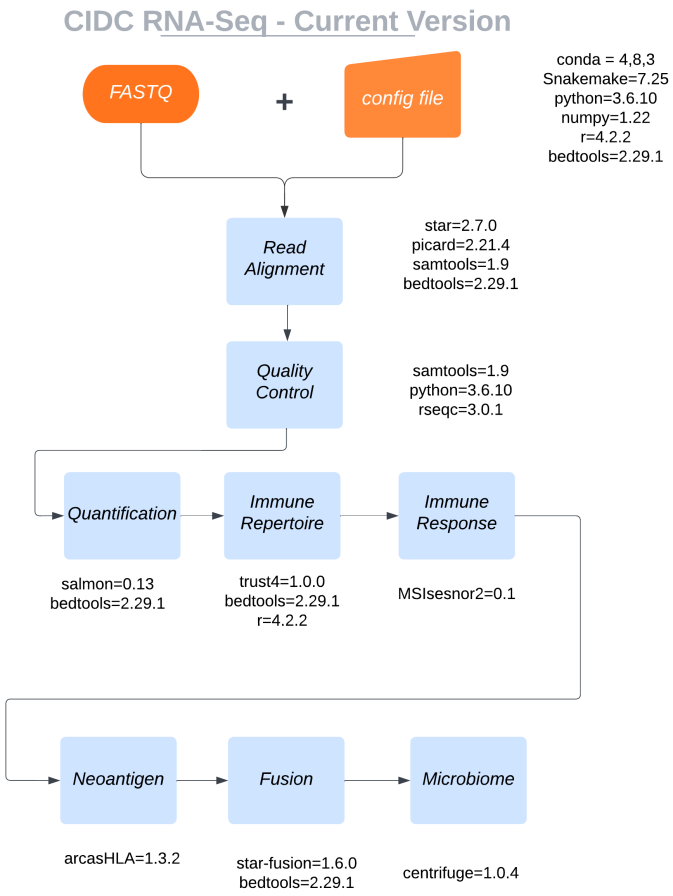
1. Introduction 3
2. RNA-SEQ Pipeline – Validation Dataset 6
3. RNA-SEQ Pipeline – Validation Method 12
4. RNA-Seq Pipeline – Validation Results 14

1. INTRODUCTION

As part of the planned CIDC enhancements after the migration to the National Cancer Institute (NCI), the bioinformatic pipelines were reviewed by members the National Cancer Institute Computational Genomics and Bioinformatics Branch (NCI-CGBB) and Essex Management (EM). These reviews were carried out to determine which, if any, changes could be made to the pipeline to satisfy the following goals and objectives:

Goals	Objectives
Update and clean up code so that it is easy to read and understand for everyone working in the same code base, thus, making it easier to maintain, debug, and update.	Provide code review to determine areas for clean up and refactoring
Maintain industry standard software to optimize the best combination of biochemistry, mathematics, computer science, data science, and modern data analytics tools.	Review the performance of current packages to determine if current functionality is meeting customer needs (CIMAC Input)
Maintain current software versioning to optimize vendor support and application performance.	Provide gap analysis on current software versions to determine what tools to upgrade
Provide enhancements to pipeline to improve current functionality/performance and better support the analysis of DNA, while maintaining backward compatibility with previous versions.	Add new functionality and features as desired by the stakeholder community

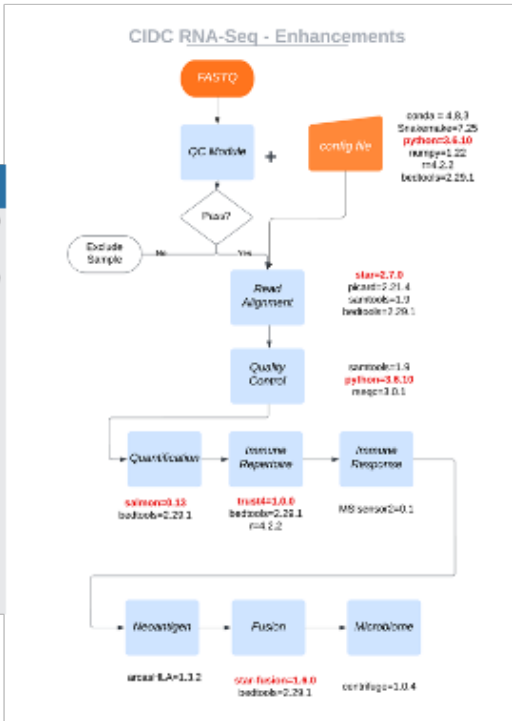
The reviews created schematics of the extant pipelines stood up after the migration from Dana Farber Cancer Institute (DFCI) to NIH's Cloud System. The **RNA-Seq** pipeline was divided into 8 distinct modules:



After thorough review, the following enhancements were planned for the RNA-Seq Pipeline:

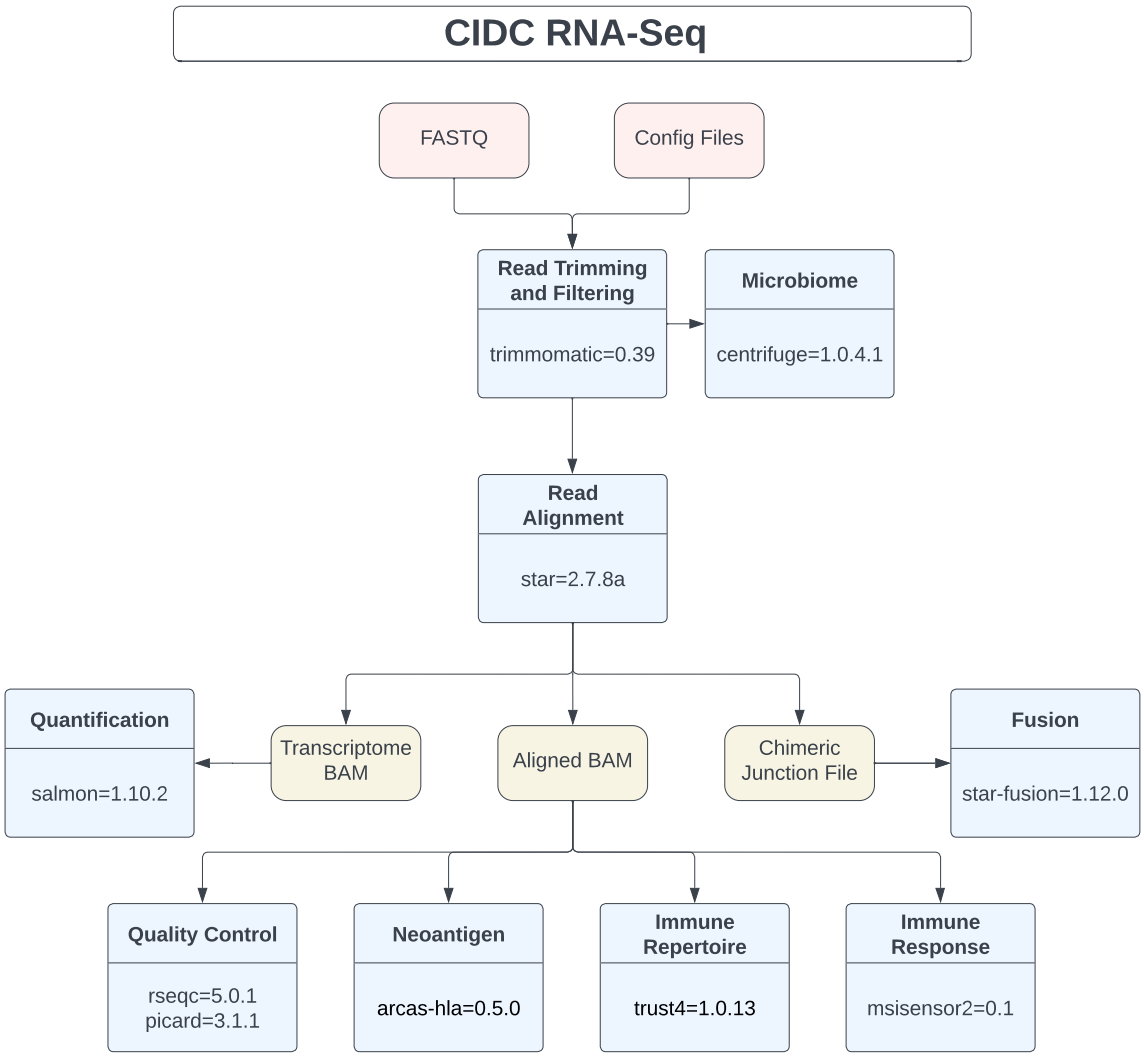
CIDC RNA-Seq – Recommended Enhancements

Pipeline pathway	RNA-Seq Software Recommendations	Reason
Programming language	Upgrade python	Current version unsupported; need to get past 3.7
Read Alignment	Upgrade STAR to 2.7.8a	Fixed a bug causing wrong sequence length in the UB SAM. Increased accuracy and speed
Fusion	Upgrade STAR-Fusion to 1.12.0	Deprecated genome libs as of Mar, 2023. Decreased false positives
Immune repertoire analysis	Upgrade TRUST4 to v1.0.12	Fixes a serious bug that may crash the program, improves the robustness in testing files, and improves the annotation accuracy.
Transcript Quantification	Upgrade Salmon to 1.10	Improve mapping and quantification accuracy
Quality Control	Add file validation (QC Module)	To identify corrupt or incorrectly formatted files



The specific changes associated with the pipeline are described in the figures above. For each module, code refactoring was also performed to increase readability and ease future maintenance and/or upgrade efforts.

Final Design



2. RNA-SEQ PIPELINE – VALIDATION DATASET

Validation Dataset:

For validation, it would be possible to compare the output of the enhanced pipeline to the original pipeline. While this method would be suitable to determine result continuity associated with the transition, it would not be suitable for determining result accuracy. The aligner and fusion caller (STAR and STAR-Fusion, respectively) will be updated in the enhanced pipeline, which could lead to vastly different results (e.g., reduced false positives and removal of 'red herrings', see https://github.com/alexdobin/STAR/releases and https://github.com/STAR-Fusion/STAR-Fusion/releases for STAR and STAR-Fusion release notes). Thus, to verify the accuracy of the enhanced pipeline, we will use truth sets obtained from external sources. The key criteria for evaluating the enhanced pipeline are the quantification steps and fusion calling steps,

because these are core results delivered from the pipeline to the portal and are most likely to be affected by the code changes associated with the enhancements.

To evaluate the quantification accuracy of the pipeline, we will use an RNA-Seq reference standard dataset reported previously[1]. These include 9 replicates samples of hepatocellular cell line MHCC97H with paired-end 2x150 paired end sequencing data. The expression data is presented as reads per kilobase per million reads (RPKM). The GEO entry for the dataset can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE234201
A second set from ENCODE will also be used for evaluation. These samples were selected from ENCODE's deeply profiled cell lines and have been processed through their uniform processing pipelines using defined pipelines and parameters. The specimens were selected based on availability of paired-end total RNA-Seq data

To measure the accuracy of the fusion calls, simulated data will be used. The reads are generated by the Broad for fusion caller benchmarking and are generated as 101 bp paired-end reads: https://data.broadinstitute.org/Trinity/STAR_FUSION_PAPER/SupplementaryData/sim_reads/sim_101_fastq/ . The benchmarking truth set are provided here: https://github.com/STAR-Fusion/STAR-Fusion_benchmarking_data/tree/master/simulated_data/sim_101/samples

Sample	Description	Result Evaluated	Source – Sequencing Reads	Source – Benchmarking Data
ENCSR000CVT_Rep1	ENCODE Dataset – GM12878	Quantification	https://www.encodeproject.org/files/ENCFF000FAG/@@download/ENCFF000FAG.fastq.gz https://www.encodeproject.org/files/ENCFF000FAH/@@download/ENCFF000FAH.fastq.gz	https://www.encodeproject.org/experiments/ENCSR000CVT/
ENCSR000CVT_Rep2	ENCODE Dataset – GM12878	Quantification	https://www.encodeproject.org/files/ENCFF000EZZ/@@download/ENCFF000EZZ.fastq.gz https://www.encodeproject.org/files/ENCFF000FAK/@@download/ENCFF000FAK.fastq.gz	https://www.encodeproject.org/experiments/ENCSR000CVT/

[1] Lu S, Lu H, Zheng T, Yuan H, Du H, Gao Y, Liu Y, Pan X, Zhang W, Fu S, Sun Z, Jin J, He QY, Chen Y, Zhang G. A multi-omics dataset of human transcriptome and proteome stable reference. Sci Data. 2023 Jul 13;10(1):455. doi: 10.1038/s41597-023-02359-w. PMID: 37443183; PMCID: PMC10344951.
https://pubmed.ncbi.nlm.nih.gov/37443183/

ENCSR000AEE_Rep1	ENCODE Dataset – GM12878	Quantification	https://www.encodeproject.org/files/ENCFF001RDI/@@download/ENCFF001RDI.fastq.gz https://www.encodeproject.org/files/ENCFF001RDA/@@download/ENCFF001RDA.fastq.gz	https://www.encodeproject.org/experiments/ENCSR000AEE/
ENCSR000AEE_Rep2	ENCODE Dataset – GM12878	Quantification	https://www.encodeproject.org/files/ENCFF001RDH/@@download/ENCFF001RDH.fastq.gz https://www.encodeproject.org/files/ENCFF001RCZ/@@download/ENCFF001RCZ.fastq.gz	https://www.encodeproject.org/experiments/ENCSR000AEE/
ENCSR000AEC_Rep1	ENCODE Dataset – GM12878	Quantification	https://www.encodeproject.org/files/ENCFF001RFH/@@download/ENCFF001RFH.fastq.gz https://www.encodeproject.org/files/ENCFF001RFG/@@download/ENCFF001RFG.fastq.gz	https://www.encodeproject.org/experiments/ENCSR000AEC/
ENCSR000AEC_Rep2	ENCODE Dataset – GM12878	Quantification	https://www.encodeproject.org/files/ENCFF001RFB/@@download/ENCFF001RFB.fastq.gz https://www.encodeproject.org/files/ENCFF001RFA/@@download/ENCFF001RFA.fastq.gz	https://www.encodeproject.org/experiments/ENCSR000AEC/
ENCSR860RDT_Rep1	ENCODE Dataset – K562	Quantification	https://www.encodeproject.org/files/ENCFF816ZQN/@@download/ENCFF816ZQN.fastq.gz https://www.encodeproject.org/files/ENCFF446GAM/@@download/ENCFF446GAM.fastq.gz	https://www.encodeproject.org/experiments/ENCSR860RDT/

ENCSR860RDT_Rep2	ENCODE Dataset – K562	Quantification	https://www.encodeproject.org/files/ENCFF991YOP/@@download/ENCFF991YOP.fastq.gz https://www.encodeproject.org/files/ENCFF409UVL/@@download/ENCFF409UVL.fastq.gz	https://www.encodeproject.org/experiments/ENCSR860RDT/
ENCSR100JNS_Rep1	ENCODE Dataset – K562	Quantification	https://www.encodeproject.org/files/ENCFF386HOV/@@download/ENCFF386HOV.fastq.gz https://www.encodeproject.org/files/ENCFF258NRG/@@download/ENCFF258NRG.fastq.gz	https://www.encodeproject.org/experiments/ENCSR100JNS/
ENCSR100JNS_Rep2	ENCODE Dataset – K562	Quantification	https://www.encodeproject.org/files/ENCFF109EVR/@@download/ENCFF109EVR.fastq.gz https://www.encodeproject.org/files/ENCFF594UVX/@@download/ENCFF594UVX.fastq.gz	https://www.encodeproject.org/experiments/ENCSR100JNS/
ENCSR601DZY_Rep1	ENCODE Dataset – K562	Quantification	https://www.encodeproject.org/files/ENCFF993RLZ/@@download/ENCFF993RLZ.fastq.gz https://www.encodeproject.org/files/ENCFF113BOL/@@download/ENCFF113BOL.fastq.gz	https://www.encodeproject.org/experiments/ENCSR601DZY/
ENCSR601DZY_Rep2	ENCODE Dataset – K562	Quantification	https://www.encodeproject.org/files/ENCFF255XLW/@@download/ENCFF255XLW.fastq.gz https://www.encodeproject.org/files/ENCFF701RGA/@@download/ENCFF701RGA.fastq.gz	https://www.encodeproject.org/experiments/ENCSR601DZY/

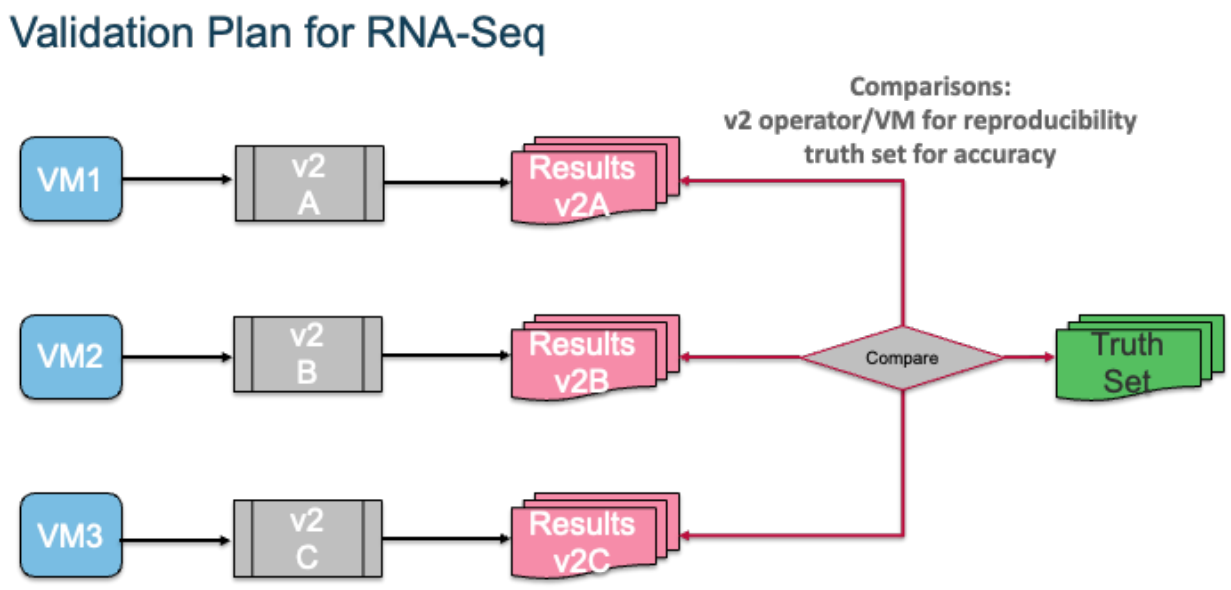
ENCSR062FHL_Rep1	ENCODE Dataset – K562	Quantification	https://www.encodeproject.org/files/ENCFF447PJB/@@download/ENCFF447PJB.fastq.gz https://www.encodeproject.org/files/ENCFF760XHT/@@download/ENCFF760XHT.fastq.gz	https://www.encodeproject.org/experiments/ENCSR062FHL/
ENCSR062FHL_Rep2	ENCODE Dataset – K562	Quantification	https://www.encodeproject.org/files/ENCFF966UCK/@@download/ENCFF966UCK.fastq.gz https://www.encodeproject.org/files/ENCFF160CNN/@@download/ENCFF160CNN.fastq.gz	https://www.encodeproject.org/experiments/ENCSR062FHL/
MHCC97H cells, generation1	Cultured hepatocellular cells	Quantification	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7454068	https://ftp.ncbi.nlm.nih.gov/geo/samples/GSM7454nnn/GSM7454068/suppl/GSM7454068%5FV350003741%5FL01%5F81.txt.gz
MHCC97H cells, generation2	Cultured hepatocellular cells	Quantification	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7454069	https://ftp.ncbi.nlm.nih.gov/geo/samples/GSM7454nnn/GSM7454069/suppl/GSM7454069%5FV350003741%5FL01%5F82.txt.gz
MHCC97H cells, generation3	Cultured hepatocellular cells	Quantification	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7454070	https://ftp.ncbi.nlm.nih.gov/geo/samples/GSM7454nnn/GSM7454070/suppl/GSM7454070%5FV350003741%5FL01%5F83.txt.gz
MHCC97H cells, generation4	Cultured hepatocellular cells	Quantification	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7454071	https://ftp.ncbi.nlm.nih.gov/geo/samples/GSM7454nnn/GSM7454071/suppl/GSM7454071%5FV3500

				03741%5FL01%5F84.txt.gz
MHCC97H cells, generation5	Cultured hepatocellular cells	Quantification	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7454072	https://ftp.ncbi.nlm.nih.gov/geo/samples/GSM7454nnn/GSM7454072/suppl/GSM7454072%5FV350003741%5FL01%5F85.txt.gz
MHCC97H cells, generation6	Cultured hepatocellular cells	Quantification	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7454073	https://ftp.ncbi.nlm.nih.gov/geo/samples/GSM7454nnn/GSM7454073/suppl/GSM7454073%5FV350003741%5FL01%5F86.txt.gz
MHCC97H cells, generation7	Cultured hepatocellular cells	Quantification	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7454074	https://ftp.ncbi.nlm.nih.gov/geo/samples/GSM7454nnn/GSM7454074/suppl/GSM7454074%5FV350003741%5FL01%5F87.txt.gz
MHCC97H cells, generation8	Cultured hepatocellular cells	Quantification	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7454075	https://ftp.ncbi.nlm.nih.gov/geo/samples/GSM7454nnn/GSM7454075/suppl/GSM7454075%5FV350003741%5FL01%5F88.txt.gz
MHCC97H cells, generation9	Cultured hepatocellular cells	Quantification	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7454076	https://ftp.ncbi.nlm.nih.gov/geo/samples/GSM7454nnn/GSM7454076/suppl/GSM7454076%5FV350003741%5FL01%5F73.txt.gz
Fusion_Sim_reads1	Simulated dataset with known fusions	Fusion	https://data.broadinstitute.org/Trinity/STAR_FUSION_PAPER/SupplementaryData/sim_reads/sim_101_fastq/	https://github.com/STAR-Fusion/STAR-Fusion_benchmarking_data/blob/master/simulated_data/si

				m_101/sim_101.fusion_TPM_values.dat
Fusion_Sim_reads2	Simulated dataset with known fusions	Fusion	https://data.broadinstitute.org/Trinity/STAR_FUSION_PAPER/SupplementaryData/sim_reads/sim_101_fastq/	https://github.com/STAR-Fusion/STAR-Fusion_benchmarking_data/blob/master/simulated_data/sim_101/sim_101.fusion_TPM_values.dat
Fusion_Sim_reads3	Simulated dataset with known fusions	Fusion	https://data.broadinstitute.org/Trinity/STAR_FUSION_PAPER/SupplementaryData/sim_reads/sim_101_fastq/	https://github.com/STAR-Fusion/STAR-Fusion_benchmarking_data/blob/master/simulated_data/sim_101/sim_101.fusion_TPM_values.dat
Fusion_Sim_reads4	Simulated dataset with known fusions	Fusion	https://data.broadinstitute.org/Trinity/STAR_FUSION_PAPER/SupplementaryData/sim_reads/sim_101_fastq/	https://github.com/STAR-Fusion/STAR-Fusion_benchmarking_data/blob/master/simulated_data/sim_101/sim_101.fusion_TPM_values.dat
Fusion_Sim_reads5	Simulated dataset with known fusions	Fusion	https://data.broadinstitute.org/Trinity/STAR_FUSION_PAPER/SupplementaryData/sim_reads/sim_101_fastq/	https://github.com/STAR-Fusion/STAR-Fusion_benchmarking_data/blob/master/simulated_data/sim_101/sim_101.fusion_TPM_values.dat

3. RNA-SEQ PIPELINE – VALIDATION METHOD

Validation Design: The validation design is shown below. It was used to compare the results of the enhanced pipeline (v2) to industry-accepted truth sets as well as assess inter-operator and inter-VM variability.



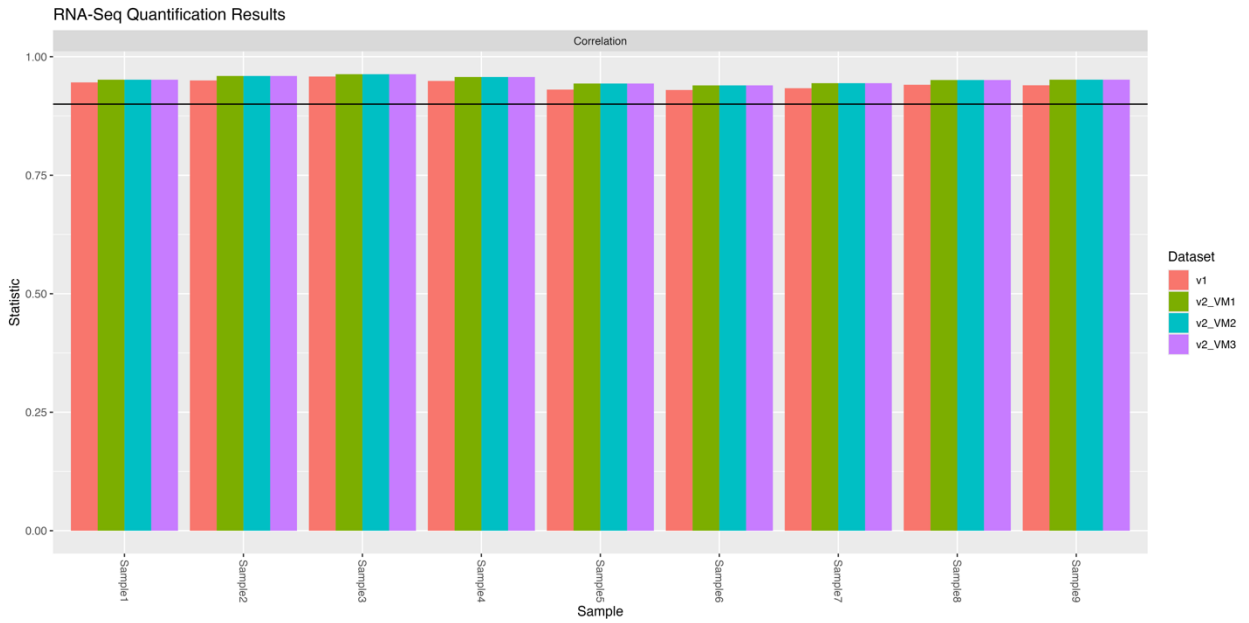
Acceptance Criteria:
Correlation – Transcript quantification data: $R^2 > 0.9$
Sensitivity – Fusion Calls: > 0.90
Specificity – Fusion Calls: > 0.90
Accuracy – Fusion Calls: > 0.90

Investigation of Discrepancies:
If the analysis metrics do not meet the acceptance criteria, an investigation will be carried out by members of NCI-CGBB and EM. The members to perform the investigation will be designated by Daoud Meerzaman, based on expertise and availability. The investigation should last no longer than 10 business days, at which time a report that outlines the problem and suggests solutions will be presented to Daoud Meerzaman.

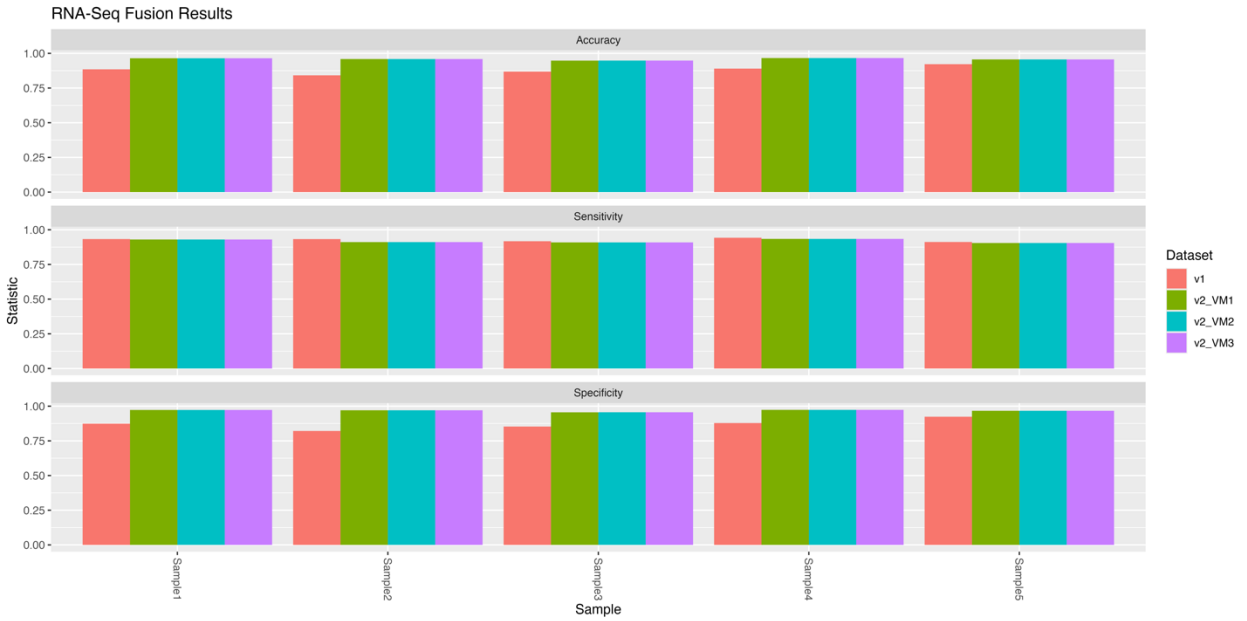
Comparison to v1 Output: The original (v1) and enhanced (v2) pipelines will likely produce different results and thus comparing the outputs would create unacceptable statistics for pipeline validation. However, it is also important to compare the performance to provide a quantive measure of the differences (e.g., improvements) in pipeline specifications for the purposes of documentation. Thus, v1 outputs for both transcript quantification and fusion calls will be analyzed in parallel to the v2 outputs. Performance differences will be noted in the final report and where applicable, these differences will be linked back to software/version updates associated with the v2 pipeline. Further, relevant documentation associated with the core pipeline software will be cited to aid in communicating the source of the differences, noting to highlight

the beneficial aspects of the differences (i.e., fewer false positives or false negatives, increased sensitivity, etc..).

4. RNA-SEQ PIPELINE – VALIDATION RESULTS



Quantification Performance Specifications										
Dataset	**Metric**	**Sample1**	**Sample2**	**Sample3**	**Sample4**	**Sample5**	**Sample6**	**Sample7**	**Sample8**	**Sample9**
v1 / Truth Set	**Correlation**	0.946	0.950	0.958	0.949	0.931	0.930	0.934	0.941	0.940
v2_VM1 / Truth Set	**Correlation**	0.952	0.959	0.963	0.957	0.944	0.940	0.944	0.951	0.952
v2_VM2 / Truth Set	**Correlation**	0.952	0.959	0.963	0.957	0.944	0.940	0.944	0.951	0.952
v2_VM3 / Truth Set	**Correlation**	0.952	0.959	0.963	0.957	0.944	0.940	0.944	0.951	0.952
VM1 / VM3	**Correlation**	0.999	0.991	0.999	0.999	0.999	0.999	0.999	0.997	0.999
VM2 / VM3	**Correlation**	0.999	0.999	0.999	0.995	0.999	0.998	0.999	0.999	0.999
VM1 / VM2	**Correlation**	0.999	0.991	0.999	0.995	0.999	0.998	0.999	0.997	0.999



Fusion Performance Specifications						
Dataset	**Metric**	**Sample1**	**Sample2**	**Sample3**	**Sample4**	**Sample5**
v1	**Sensitivity**	93.3%	93.3%	94.7%	94.2%	91.2%
	Specificity	87.4%	82.1%	85.3%	87.8%	92.4%
	Accuracy	88.5%	84.1%	86.8%	89.0%	92.2%
v2_VM1	**Sensitivity**	93.0%	91.1%	90.9%	93.4%	90.5%
	Specificity	97.3%	97.0%	95.6%	97.3%	96.7%
	Accuracy	96.5%	95.9%	94.8%	96.6%	95.6%
v2_VM2	**Sensitivity**	93.0%	91.1%	90.9%	93.4%	90.5%
	Specificity	97.3%	97.0%	95.6%	97.3%	96.7%
	Accuracy	96.5%	95.9%	94.8%	96.6%	95.6%
v2_VM3	**Sensitivity**	93.0%	91.1%	90.9%	93.4%	90.5%
	Specificity	97.3%	97.0%	95.6%	97.3%	96.7%
	Accuracy	96.5%	95.9%	94.8%	96.6%	95.6%
VM1_VM3	**Overlap**	100.0%	100.0%	100.0%	100.0%	100.0%
VM2_VM3	**Overlap**	100.0%	100.0%	99.8%	100.0%	100.0%
VM1_VM2	**Overlap**	100.0%	100.0%	99.8%	100.0%	100.0%

As shown, all specifications met or surpassed the established acceptance criteria for the evaluation. The analysis also revealed the expected results of the enhanced pipeline. Specifically, the transcript quantification was marginally improved compared to the v1 results, likely owing to improved alignment accuracy associated with the Salmon quantification software (https://github.com/COMBINE-lab/salmon/releases). The fusion calling showed significant improvements in both specificity and accuracy. This result is expected as the fusion caller (STAR-Fusion) was upgraded to a more recent version that addressed known issues with false positive calls seen in previous versions (https://github.com/STAR-Fusion/STAR-Fusion/releases).

Based on these results, it recommended that the production RNA-Seq pipeline be upgraded to the v2 (enhanced) version.