Tempus Bioinformatics Pipeline: RNAseq Assays

This document describes the operation of Tempus' bioinformatics pipeline for the RNAseq component of our solid tumor/heme assays, including RNAseq data associated with the xT assays (versions 1-4), the exome-level xE assays (versions 1-2), and our legacy xO assay. It is intended to provide a deeper understanding of our RNAseq analytical pipeline, to describe how we develop the downstream products from our assays, and to help our partners define and build their analytical framework for processing Tempus deliverables. Some of the information provided here has been compiled from prior Tempus publications, a partial list of which is available at the end of the document. In addition, aspects of the RNAseq pipeline that are identical to the DNA process (specifically, sequence demultiplexing, FASTQ file generation, and the associated quality control steps) are presented here as well, for completeness. Please do not hesitate to provide input regarding additional areas that would be helpful for your team or reach out for any additional clarifications as you read through the document.

A general note on Tempus RNAseq data: RNA sequencing data is currently being generated using an updated (RNA.v2) assay developed for the Illumina NovaSeq 6000 Sequencing System, which replaced our original (RNA.v1) RNAseq assay, which was designed for the Illumina HiSeq 4000. Extensive validation studies have been conducted in order to assess and confirm the consistency of the data derived from our current and legacy RNAseq assays. While elements of these two RNA assays are shared, there are some differences, the most significant of which are highlighted below. These and other differences are also noted throughout the document. Each assay is also associated with an assay-specific bioinformatics pipeline, however, our updated second-generation pipeline has been designed to accommodate RNA.v1 sequencing data, and will typically be utilized when RNA.v1 data is delivered.

	Tempus RNA.v1 Assay	Tempus RNA.v2 Assay
Sequencing Platform	Illumina HiSeq 4000	Illumina NovaSeq 6000
Hybridization Probes	IDT xGen Exome Panel v1.0	IDT xGen Exome Panel v2.0, with additional spike-in probes
Reference Genome	GRCh37*	GRCh37

*Legacy RNA.v1 data was aligned to the GRCh38 reference genome, however, all current RNA-seq data (RNA.v1 and RNA.v2) is aligned to GRCh37.

RNAseq Pipeline

Demultiplexing and FASTQ generation

The RNA analysis pipeline uses Illumina BCL2FASTQ demultiplexing software. A sample sheet containing index information is first checked to confirm that there are no adapter pair mismatches and that they map to the expected isolate in the laboratory information management system (LIMS) used by Tempus to perform the assay. Demultiplexing occurs, isolates are tagged by the demultiplexing process unique identifier, the order unique identifier, and the lab accessioning ID and indexed within a FASTQ staging object store file system. Two FASTQ files containing paired RNAseq reads are generated per sample, corresponding to all full length forward and reverse reads.

Demultiplexing quality control includes quality metrics for per-base sequence quality, sequence content, GC content, and relative percentages of unmatched indices. If the sample does not pass the automated quality control step, it is manually reviewed. Cases that do not pass quality control review are referred to the pathology and laboratory teams for re-processing or re-analysis.

Indexing QC check

The potential for index contamination is managed by demultiplexing all sequencing reads for all possible barcodes. If any sample has fewer than 50 megabytes worth of reads assigned, then the undetermined read bin is analyzed to assess for potential barcode assignment errors. If a sample has had an index incorrectly assigned, the demultiplexing process will be restarted after all indices have been re-checked on the sample sheet and index assignment has been confirmed.

RNA sequencing process

The current version of the Tempus RNA exome panel uses the Integrated DNA Technologies (IDT) xGen Exome Research Panel v2, which consists of 415,115 individually synthesized hybridization probes. The IDT panel spans a 34 Mb target region (19,433 genes) of the human genome and covers 39 Mb of end-to-end tiled probe space. (The legacy Tempus RNA pipeline utilized the IDT v1 xGen Panel, for which the respective numbers are 429,826 probes, a 39 Mb target region, and also 19,433 genes.) In addition to the IDT exome probes, Tempus-specific custom probes have been added to the RNA.v2 panel, with probes for additional T-cell and B-cell receptor sequences (TCR/BCR), fusion transcripts, viral/bacterial sequences, and multiple oncofetal genes. Tempus utilizes RNA-sequencing on this transcriptome panel to identify expression evidence of chromosomal rearrangements that result in the expression of fusion RNA species. This assay detects interchromosomal and intrachromosomal rearrangements containing one or more coding RNA sequences.

Tempus performs library preparation using the Roche KAPA RNA HyperPrep Kit. Following a quality and quantity review of the samples, passing libraries are pooled together for hybridization with the IDT xGen probes. After hyb-capture and PCR, Tempus performs post-PCR bead clean-up and assesses quality. The amplified target-captured libraries are sequenced on an Illumina NovaSeq 6000 system using patterned flow cell technology, to a targeted minimum depth of 30 million reads per sample (as noted above, the original RNA panel was sequenced on the HiSeq 4000 system) .

Read alignment and BAM file generation

Tumor RNA BAM files generated from the Tempus bioinformatics pipeline have been aligned to the Ensembl GRCh37 Release 75 (July 2019) reference genome using STAR (version 2.5.4a). Reads containing adapter sequences are trimmed prior to alignment using skewer (version 0.2.2). BAM files are then processed by Opossum (v0.2) where unmapped reads and split reads spanning splice junctions are removed, overlapping reads are merged, and quality scores of the merged reads are adjusted. Duplicate reads are sorted and marked in each BAM file, and each BAM is indexed by a BAI file. In addition to this processed STAR-aligned BAM file, the Bioinformatics pipeline also generates an unprocessed STAR-aligned BAM file, including all mapped as well as unmapped reads.

Quality control

1. The number of unique deduplicated reads should meet or exceed 6 million. For some exploratory and research purposes, a lower threshold of 5 million reads is applied.
2. The mapping rate should be greater than 80%.
3. The average GC content should be between 45% and 59%.
4. Total number of expressed genes detected should exceed 12,000.
5. Fingerprint variant analysis between RNA and matching DNA samples is performed, and should return a matching value.
6. The percent of reads in the proper orientation (RNA library construction is strand-sensitive) should exceed 90%, and the percent of reads failing strand detection should be 6% or below.

RNA Expression

Pseudo-alignment, transcript and gene quantification

Transcript level pseudo-alignment to the Ensembl GRCh37 Release 75 (July 2019) reference genome, as well as subsequent quantification, is performed using Kallisto (version 0.44). Raw counts and transcripts per million (TPM) are calculated for 180,253 transcripts in the Ensembl

reference (134,160 transcripts covered by the exome panel). Protein coding transcripts covered by the exome panel are then summed to obtain raw and TPM gene-level counts (19,146 genes).

RNA fusions

RNA.v2 Pipeline: Two algorithms are used to detect fusions. STAR-Fusion v1.9.0 is run to identify candidate fusion transcripts supported by Illumina reads. This tool further processes the output generated by the STAR aligner to map junction reads and spanning reads to a reference annotation set (https://github.com/STAR-Fusion/STAR-Fusion/wiki). Additionally, Mojo v0.0.5 is run to identify gene fusions at canonical exon-exon junctions from paired-end transcriptome sequencing data. This tool identifies clusters of discordant reads by mapping reads to the transcriptome in iterative steps to maximize sensitivity. Candidate fusion junctions are constructed from the exons predicted to be involved in fusions between the pairs of genes. Reads that cannot be aligned to the canonical transcriptome are mapped to these junctions. High confidence fusions are nominated following rigorous filtering steps designed to capture both technical and biological noise (https://github.com/cband/MOJO/blob/master/README.md). Following an integration step that combines data from both STAR-Fusion and Mojo (selecting fusion calls with maximum support), AGFusion v2.0.2 is run to annotate the integrated fusion calls.

RNA.v1 Pipeline (applicable to legacy RNA.v1 data only): p.arc v2.5.1 was run to identify structural variants from GRCH38-aligned RNA-seq data. This tool quantifies gene-level expression and chimeric transcripts through non-canonical exon-exon junctions which are mapped using split or discordant read pairs. Subsequent to expression quantification, reads are mapped across exon–exon boundaries to unannotated splice junctions, and evidence is computed for potential chimeric gene products. If sufficient evidence is present for the chimeric transcript, a rearrangement is called as detected. Internal tandem duplications and inversions are excluded, for a 'fusion-centric' view. Strand-specificity in the original sequence data is retained.

Tempus References:

1. Beaubier, N., Bontrager, M., Huether, R. et al. Integrated genomic profiling expands clinical options for patients with cancer. Nat Biotechnol 37, 1351–1360 (2019). https://doi.org/10.1038/s41587-019-0259-z
2. Beaubier N., Tell R., Lau D. et al. Clinical validation of the Tempus xT next-generation targeted oncology sequencing assay. Oncotarget. 10: 2384-2396 (2019). https://doi.org/10.18632/oncotarget.26797
3. Reiman, D., Sha, L., Ho, I. et al. Integrating RNA expression and visual features for immune infiltrate prediction. Pac Symp Biocomput. 24:284-295 (2019). https://doi.org/10.1142/9789813279827_0026