Confidential v08.2021

Tempus RNA Assays and Pipelines

Tempus has two RNA assays: Tempus|RS and Tempus|RS.v2. Tempus|RS.v2 was launched on Sep 22, 2020 and replaced sequencing of the Tempus|RS assay. This document explains the sequencing process, pipelines, and outputs of each assay version, the extent and types of comparisons that can be performed with these outputs, and the validation strategy used to confirm comparability between assay versions.

Table of Contents:

I. Summary of Major Differences between Tempus|RS and Tempus|RS.v2 Assay Data

Assay and Pipeline Details	Tempus| RS Legacy Deliveries	Tempus| RS Current Deliveries or Deliveries also Including RS.v2	Tempus| RS.v2 Deliveries (launched in September 2020)
IDT exome probe design	Exome Panel v1	Exome Panel v1	Exome Panel v2
Spike in probes	No	No	Common fusions, oncogenic viruses, TCR/BCR, and oncofetal/cancer-testis antigens
Sequencer	HiSeq 4000	HiSeq 4000	NovaSeq 6000
UMI de-duplication	No	No	Yes
Average Read Count:	50 million reads/sample	50 million reads/sample	50 million reads/sample
Gene level normalization approach	**N1** - See Section II.3 Only protein coding transcripts are included	**N2** - See Section III.3 All transcripts regardless of the biotype are included	**N2** - See Section III.3 All transcripts regardless of the biotype are included
# of genes covered in normalization	19,147 genes	20,061 genes	20,061 genes
# of transcripts covered in normalization	134,160 transcripts	180,253 transcripts	180,253 transcripts
Gene level expression metrics	→ `gene_raw` → `gene_tpm`	→ `gene_raw` → `log2_gene_tpm` → `Log2_gene_tpm_corrected` For RS data, values in	→ `gene_raw` → `log2_gene_tpm` → `log2_gene_tpm_corrected` (batch correction for RS.v2 to look like RS data)

		`log2_gene_tpm_corrected` should match `log2_gene_tpm`	
Expression Validation	No	No	Gene expression counts are analytically validated in a CAP/CLIA lab
Fusion Pipeline	p.arc (licensed method from UMichigan)	In-house ensemble method using 2 open source methods (STAR-Fusion and Mojo)	In-house ensemble method using 2 open source methods (STAR-Fusion and Mojo)
Gene Fusion Validation	Targeted gene fusions are clinically and analytically validated	Targeted gene fusions are clinically and analytically validated	Targeted gene fusions are clinically and analytically validated using additional spike-in probes
Expression Workflow IDs*	`tempus_rna_expression`	`tempus_rna_expression_v1`	`tempus_rna_expression_v2`
Fusion Workflow IDs*	`tempus_rs`	`rna_fusion_v1`	`rna_fusion_v2`

*Workflow IDs can be found in the delivered **g_pipeline_version.csv** file where each patient and analysis ID is mapped to a specific workflow and pipeline version.

II. Tempus|RS Assay and Bioinformatics Pipeline

1. Tempus|RS Sequencing Process

The Tempus RNA whole-transcriptome panel uses the IDT xGen Exome Research Panel v1.0, which consists of 429,826 individually synthesized probes. The panel spans a 39 Mb target region (19,396 genes) of the human genome and covers 51 Mb of end-to-end tiled probe space. Tempus utilizes RNA-sequencing from this transcriptome panel to quantify transcript and gene level expression, and identify transcriptional 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 exome capture RNA library preparation using the IDT xGen probes. A minimum of 50ng of total RNA is required for the Tempus|RS test. Following cDNA library preparation, libraries passing quality and quantity review are pooled together for hybridization. After capture and PCR, Tempus performs post-PCR bead clean-up and assesses quality for sequencing. The amplified target-captured libraries are pooled and sequenced on an Illumina HiSeq 4000 System using patterned flow cell technology, to a minimum depth of 30 million reads per sample.

Capture RNA-Seq has a number of benefits when using FFPE samples, as polyA-selection is suboptimal on RNA molecules fragmented during the fixation process. In addition, it obviates the ribosome depletion step, as well as hemoglobin depletion for hematological samples, leading to the capture of a higher percentage of biologically relevant reads.

Specimen Collection and Preparation: Tumor specimens are prepared following standard pathology practices. Archival paraffin embedded specimens subjected to acid decalcification are unsuitable for analysis. Paraffin embedded specimens decalcified in EDTA are acceptable.

Before starting the assay, a Hematoxylin and Eosin (H&E) stained slide is prepared for FFPE tumor specimens and reviewed by a board-certified pathologist to ensure that adequate tissue, tumor content, and sufficient nucleated cells are present to satisfy the minimum tumor content requirements. A minimum tumor content of 20% is required to result in adequate yield at extraction and to proceed with the assay. Macrodissection is carried out when deemed feasible by a pathologist to increase the tumor percentage of a specimen. Macrodissection must be done if the tumor percentage is less than 20%, and may be done to increase tumor content in other instances.

Nucleic Acid Extraction: Total nucleic acid (TNA) is extracted from tissue specimens using a Chemagic 360 or similar instrumentation, and the recovered TNA is quantified and qualified, and may be concentrated if necessary. Tempus|RS uses TNA extracted in a manner identical to extraction of TNA for xT; therefore, TNA extracted for xT will also be suitable for use with Tempus|RS. Extracted TNA that is intended for processing by Tempus|RS is treated with Invitrogen TURBO DNase to degrade DNA, and the remaining RNA is assessed for quantity and quality (sizing). Tempus|RS can also be run using RNA that is extracted directly from an acceptable tissue specimen using a compatible RNA-extraction method, and which meets the RNA input specifications for the assay. The minimum amount of input total RNA required to perform the test is 50 ng. RNA is fragmented using heat and magnesium, with variable parameters, to yield similar sized fragments from RNA inputs with different starting size distributions.

Library Preparation: Strand-specific library preparation is performed using the KAPA RNA HyperPrep Kit for Illumina with KAPA single indexed 6 base pair adapters. This involves first-strand synthesis using a reverse transcriptase (RT) enzyme to create first strand cDNA, followed by treatment with RNAse to degrade RNA, and DNA polymerase to accomplish second-strand synthesis to create double stranded cDNA. Adapters are ligated to cDNA and the adapter-ligated libraries are cleaned using a magnetic bead-based method (0.63X and 0.7X). The libraries are amplified with high fidelity, low-bias PCR using primers complementary to adapter sequences. Amplified libraries are subjected to a 1X magnetic bead based clean-up to eliminate unused primers, and quantity is assessed. Each amplified sample library must contain a minimum of 150 ng of cDNA to proceed to hybridization.

Hybrid Capture: After library preparation and amplification, targets are captured by hybridization, clean-up of captured targets is performed, and unbound fragments are washed away. Library capture is conducted using the xGen Exome Research Panel v1 probe set, along with the xGen Hybridization and Wash Kit and xGen Universal Blockers. The enriched targets are amplified using the KAPA HiFi HotStart ReadyMix and primers supplied in the KAPA Library amplification kit, followed by a magnetic bead-based clean-up. Each post-capture library pool must satisfy a minimum calculated molarity to proceed to sequencing. The molarity is used to load the appropriate concentration of library pools onto sequencing flow cells.

Sequencing: The amplified target-captured libraries are sequenced with a 2x76 read length to an average of 50 million total reads on an Illumina HiSeq4000 System using patterned flowcells. Results from each tumor sample are assessed for quality against a set of metrics including total read count, average GC content, and total number of genes expressed. Per-sample values of these metrics are calculated, and those with questionable and failing values are flagged for quality control review and marked in the bioinformatics database via automated QC systems.

2. Tempus|RS Bioinformatics Pipeline

To date, two bioinformatics pipeline versions have been built for the Tempus|RS assay:

- **Legacy Pipeline:** Pipeline with expression profiling using Kallisto and Fusion calling using a licensed method from UMichigan (p.arc; GRCh38 reference). Only legacy RNA Fusion data deliveries were generated with this pipeline, as described in detail in Section II.4 Tempus|RS Fusions. This pipeline has been deprecated since and is no longer delivered with current RNA data deliveries. If there are any questions regarding this legacy pipeline, please reach out to our team.

- **Current Pipeline**: Updated pipeline with expression profiling using Kallisto and Fusion calling using an in-house method (also applied to the Tempus|RS.v2 assay) that uses GRCh37 reference. Details of this current pipeline is described in the sections below.

a) 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**.**

b) Read Alignment and BAM 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.

c) Quality control

1. The total number of reads aligned to the human genome should meet or exceed 30 million (i.e., 15 million read pairs). For some exploratory and research purposes, a lower threshold is applied (25 million reads; 12.5 million read pairs). This is only evaluated for current RS data. The threshold for unique deduplicated reads has replaced this criterion in RS.v2.
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 10% (RS data only), and the percent of reads failing strand detection should be 6% or below.

d) Key Software Included in Tempus|RS Pipeline

The Tempus RNA bioinformatics pipeline software includes a combination of software developed by Tempus as well as open source and proprietary software applications developed by third parties. All software is version-controlled and managed internally away from public package management systems.

Our RNA pipelines include the following key software:

Software	Current Version	Purpose
JANE	4.4.1	Pipeline Job Definition and Validation
Bedtools	2.27.1	Quality Control
Kallisto	0.44.0	RNA Expression Analysis
Pizzly	0.37.3	RNA Fusion calling
Python3	3.8.102	Execution of Python programming language
Samtools	1.9	Alignment File Manipulation
Skewer	0.2.2	RNA Adapter Trimming

STAR (two-pass)	2.7.75.4a	Reference Sequence Alignment
STAR-Fusion	1.9.0	RNA Fusion Calling
AGFusion-Tempus	2.3.4	Annotate and filter RNA Fusions
Mojo	0.0.5	RNA Fusion Detection
MultiQC	1.11	QC Data Aggregation
BCL2FASTQ	2.17	FASTQ File Conversion
Lore	1.6.3	Data product generation

3. Tempus|RS Transcript and Gene Quantification

a) Pseudo-alignment

Transcript level pseudo-alignment and quantification to the Ensembl GRCh37 Release 97 (July 2019) reference is performed using Kallisto (version 0.44). Raw counts and transcripts per million (TPM) are calculated for 180,253 transcripts in the Ensembl reference.

Tempus|RS-only deliveries included only 134,160 transcripts covered by the Tempus|RS panel rather than the whole raw Kallisto output of 180,253 transcript reference. After Tempus|RS.v2 validation experiments and pipeline improvements, we have changed our deliveries to provide raw transcript level output and can backfill this file for previous deliveries upon request.

b) Normalization

With the launch of RS.v2, Tempus improved its RNA normalization approach. Currently, two normalization strategies exist and are referred to as N1 (legacy) and N2:
i. Legacy RS Normalization Pipeline

Legacy approach (N1) carried out for all Tempus|RS deliveries prior to the development of the Tempus|RS.v2 assay (launched 9/22/2020). This normalization approach was applied to legacy Tempus|RS deliveries even after the Tempus|RS.v2 launch if the delivered cohort did not include any RS.v2 samples.

Transcript level:

Transcript level abundance is normalized using transcripts per million using all 180,253 mapped transcripts obtained from the kallisto (v 0.44.0) pseudoalignment. However, data deliveries only included 134,160 protein coding transcripts covered by the Tempus|RS panel. As a result, transcript level TPMs for a single sample will not add to 1 million for these legacy RS deliveries that only delivered 134,160 protein coding transcripts.

Gene level:

Gene level abundance and TPMs are obtained by summing only 75,404 transcripts labelled protein coding by Ensembl and covered by the RS panel to obtain gene-level counts. These transcripts encompass a total of 19,147 genes.

ii. Current RS Normalization Pipeline

Updated approach (N2) carried out for all Tempus|RS.v2 deliveries and new Tempus RS deliveries. This approach has been applied to Tempus|RS data and and is delivered if the cohort includes both Tempus|RS and Tempus|RS.v2 data. This new normalization strategy can be backfilled for any previous Tempus|RS delivered sample upon request. See Section III.3.b for technical details on normalization.

Attention to the user: Cohort level analyses should not combine data generated with these different normalization methods (N1 vs N2).

4. Tempus|RS Fusions

i. Legacy RS Fusion Pipeline (applicable to legacy RS assay only)

Software 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.

ii. Current RS Fusion Pipeline

Current Tempus|RS samples are processed through the Tempus in-house fusion calling pipeline developed for the Tempus|RS.v2 assay. See details below in Section III.4 - Tempus|RS.v2 fusions.

III. Tempus|RS.v2 Assay and Bioinformatics Pipeline (launched in Q42020)

IDT transitioned to a new version of their exome probe set (v2) in late 2020, requiring Tempus to update our capture and revalidate our assay. Taking advantage of the need to update exome probe design, Tempus has taken the opportunity to improve and expand the capabilities of its RNA-Seq assay and decrease cost by transitioning to the NovaSeq 6000 Illumina platform.

The new Tempus|RS.v2 assay includes the following technical enhancements:

- Upgrade to IDT exome-capture probe v2 design.
- Additional chimeric spike-in probes for improved coverage of common fusions.
- Inclusion of spike-in probes to target oncogenic viruses , T- cell and B-cell receptors, and oncofetal/cancer-testis antigens.
- Robust de-duplication of reads using unique molecular identifier (UMI) adapters.
- Transition to the Illumina NovaSeq platform.

Differences in probe designs between the Tempus|RS and Tempus|RS.v2 assays result in a minor batch effect that benefits from correction when analyzing gene expression from both assays. Tempus will provide a version of corrected gene expression for Tempus|RS.v2 samples to be comparable to legacy Tempus|RS samples.

Pipeline improvements to Tempus|RS.v2 include the following:

- UMI deduplication prior to gene expression quantification.
- Novel in house ensemble method for gene rearrangement detection. The Tempus|RS.v2 pipeline runs two separate open-source structural variant detection algorithms (STAR Fusion and Mojo) and combines calls via a proprietary process.
- Improvements to gene-level normalization strategy. Legacy Tempus|RS samples can be requested for re-normalization and re-delivery.

1. Tempus|RS.v2 Sequencing Process

The Tempus 2nd generation RNA whole-transcriptome panel uses IDT xGen Exome Research Panel v2 backbone, which consists of >415K individually synthesized probes and spans a 34 Mb target region (19,433 genes) of the human genome. Additional Tempus-specific custom spike-ins probes are included to enhance target region detection (e.g., fusion and viral probes).

Specimen Collection and Preparation: Tumor specimens are prepared following standard pathology practices. Archival paraffin embedded specimens subjected to acid decalcification are unsuitable for analysis. Paraffin embedded specimens decalcified in EDTA are acceptable.

Before starting the assay, a Hematoxylin and Eosin (H&E) stained slide is prepared for FFPE tumor specimens and reviewed by a board-certified pathologist to ensure that adequate tissue, tumor content, and sufficient nucleated cells are present to satisfy the minimum tumor content requirement. A minimum tumor content of 20% is required to result in adequate yield at extraction and to proceed with the assay. Macrodissection is carried out when deemed feasible by a pathologist to increase the tumor percentage of a specimen. Macrodissection must be done if the tumor percentage is less than 20%, and may be done to increase tumor content in other instances.

Nucleic Acid Extraction: Total nucleic acid (TNA) is extracted from tissue specimens using a Chemagic 360 or similar instrumentation, and the recovered TNA is quantified and qualified, and may be concentrated if necessary. Extracted TNA is treated with Invitrogen TURBO DNase to degrade DNA, and remaining RNA is assessed for quantity and quality (sizing). Tempus|RS.v2 can also be run using RNA that is extracted directly from an acceptable tissue specimen using a compatible RNA-extraction method, and which meets the RNA input specifications for the assay. The minimum amount of RNA required to perform the test is 50 ng. RNA is fragmented using heat and magnesium, with variable parameters, to yield similar sized fragments from RNA inputs with different starting size distributions.

Library Preparation: Strand-specific library preparation is performed using the KAPA RNA HyperPrep Kit for Illumina with IDT unique dual indexed (UDI) unique molecular identifier (UMI) adapters. This involves first-strand synthesis using a reverse transcriptase (RT) enzyme to create first strand cDNA followed by treatment with RNAse to degrade RNA, and DNA polymerase to accomplish second-strand synthesis to create double stranded cDNA. IDT UDI-UMI adapters are ligated to cDNA and the adapter-ligated libraries are cleaned using a magnetic bead-based method (0.7X and 0.8X). The libraries are amplified with high fidelity, low-bias PCR using primers complementary to adapter sequences. Amplified libraries are subjected to a 1X magnetic bead based clean-up to eliminate unused primers, and quantity is assessed. Each amplified sample library must contain a minimum of 150 ng of cDNA to proceed to hybridization.

Hybrid Capture: After library preparation and amplification, targets are captured by hybridization, clean-up of captured targets is performed, and unbound fragments are washed away. Library capture is conducted using the xGen Exome Research Panel v2 probe set with supplemental custom Tempus-designed probes, along with the xGen Hybridization and Wash Kit and xGen Universal Blockers. The enriched targets are amplified using the KAPA HiFi HotStart ReadyMix and primers supplied in the KAPA Library amplification kit, followed by a magnetic bead-based clean-up. Each post-capture library pool must satisfy a minimum calculated molarity to proceed to sequencing. The molarity is used to load the appropriate concentration of library pools onto sequencing flow cells.

Sequencing: The amplified target-captured libraries are sequenced with a 2x76 read length to an average of 50 million total reads on an Illumina NovaSeq 6000 System using patterned flowcells (SP, S1, S2, or S4). Results from each tumor sample are assessed for quality against a set of metrics including deduplicated total read count, average GC content, and total number of genes expressed. Per-sample values of these metrics are calculated, those with questionable and failing values are flagged for quality control review and marked in the bioinformatics database via automated QC systems.

2. Tempus|RS.v2 Bioinformatics Pipeline

a) Tempus|RS Read Alignment and BAM Generation

Tempus|RS.v2 follows the same read alignment and BAM generation described for Tempus|RS in Section II.2.a.

b) Key Software Included in Tempus|RS.v2

The bioinformatics pipeline of Tempus|RS.v2 includes a combination of software developed by Tempus as well as open source and proprietary software applications developed by third parties. All software is version-controlled and managed internally away from public package management systems.Tempus|RS.v2 includes the following key software:

Software	Current Version	Purpose
JANE	4.4.1	Pipeline Job Definition and Validation

Bedtools	2.27.1	Quality Control
Kallisto	0.44.0	RNA Expression Analysis
Pizzly	0.37.3	RNA Fusion calling
Python3	3.8.102	Execution of Python programming language
Samtools	1.9	Alignment File Manipulation
Skewer	0.2.2	RNA Adapter Trimming
STAR (two-pass)	2.7.75.4a	Reference Sequence Alignment
STAR-Fusion	1.9.0	RNA Fusion Calling
AGFusion-Tempus	2.3.5	Annotate and filter RNA Fusions
Mojo	0.0.5	RNA Fusion Detection
MultiQC	1.11	QC Data Aggregation
BCL2FASTQ	2.17	FASTQ File Conversion
UMItools	1.0.1	UMI deduplication

c) 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 (evaluated for RNA.v2 only).
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% (a threshold of 10% is used for RS assay), and the percent of reads failing strand detection should be 6% or below.

3. Tempus|RS.v2 Transcript and Gene Quantification

a) Pseudo-alignment

Transcript level pseudo-alignment and quantification to the Ensembl GRCh37 Release 75 (July 2019) reference is performed using Kallisto (version 0.44). Raw counts and transcripts per million (TPM) are calculated for 180,253 transcripts in the Ensembl reference. For any data delivery that includes Tempus|RS.v2 all 180,253 transcripts are provided.

b) Normalization

Updated approach (N2) carried out for all Tempus|RS.v2 deliveries. With the launch of RS.v2, Tempus improved its RNA normalization steps. This updated approach (first introduced in Section I.2.d.ii) is carried out for all Tempus|RS.v2 deliveries, it has also been applied to Tempus|RS data and and is delivered if the cohort includes both Tempus|RS and Tempus|RS.v2 data.

Transcript level:

Transcript level abundance is normalized using transcripts per million using all 180,253 mapped transcripts obtained from the kallisto (v 0.44.0) pseudoalignment. Data deliveries provided with this method will include abundance and TPM values for all 180,253 transcripts. As a result, transcript level TPM values for a single sample will add to 1 million.

Gene level:

Gene level abundance and TPMs are obtained by summing 145,199 transcripts independent of the Ensembl biotype label for 20,061 genes with at least one annotated protein coding transcript covered by the RS.v2 assay. Gene level TPMs generated with this method are provided with the $log_2(TPM + 1)$ transformation.

Included with your accompanying documentation you will find the file tempus_rsv1_rsv2_probe_coverage.csv. This file annotates each of the 180,253 transcripts (ESNTs) to genes (ESNGs), annotates the relevant HGNC gene name when available (using BioMart), the number of bases covered by the Tempus|RS (IDT xGEN probes v1) or Tempus|RS.v2 probe panels (IDT xGEN probes v2; excluding spike ins) and whether the transcript was summed to genes in the legacy Tempus|RS (N1) or the Tempus|RS.v2 normalization strategies (N2). Its column descriptions are listed below:

Column name	Description
ensembl_transcript_id	Ensembl transcript ID
ensembl_gene_id	Ensembl gene ID
transcript_biotype	Ensembl transcript biotype
hgnc_symbol	Relevant HGNC gene name when available, obtained from BioMart
transcript_length	Transcript length
n.bases.covered.v1	Number of bases covered by the probes included in the Tempus|RS (IDT xGEN probes v1) assay
perc.cov.v1	Percent coverage by the probes included in the Tempus|RS assay
summed.to.genes.v1	Tag describing if the transcript was included to generate gene level metrics as explained for the normalization method N1 above: 0 if transcript was excluded, 1 if transcript was included.
n.bases.covered.v2	Number of bases covered by the probes included in the Tempus|RS.v2 assay (IDT xGEN probes v2; excluding spike ins)
perc.cov.v2	Percent coverage by the probes included in the Tempus|RS.v2 assay
summed.to.genes.v2	Tag describing if the transcript was included to generate gene level metrics as explained for the normalization method N2 above: 0 if transcript was excluded, 1 if transcript was included.

Attention to the user: Cohort level analyses should not combine data generated with different normalization methods (N1 vs N2).

c) Gene-level Batch Correction to Harmonize Tempus|RS.v2 Expression to be Comparable to Tempus|RS

When comparing cohorts of gene expression with samples sequenced using RS and RSv2, corrected TPM expression will be required. At this time, we are providing TPM corrected data for Tempus|RS.v2 samples to be comparable to Tempus|RS. This correction is performed using the N2 normalization strategy and will require previous Tempus|RS cohorts to have this normalization re-delivered. Details on correction are provided below under the Tempus|RS.v2 assay.

Note that when corrected TPM data is delivered for Tempus|RS samples, this value is equal to the observed TPM value for these RS samples. This is carried out to ensure that using the corrected TPM column will always guarantee comparable values between all Tempus|RS and Tempus|RS.v2 cohorts in a single field.

Due to the sparsity of transcript level abundance, we are not able to perform transcript level correction between Tempus|RS and Tempus|RS.v2 assays. Comparing transcript level abundance across assays should be performed with caution.

i. Description of Batch Correction Applied to RS.v2

To correct for the batch effect between Tempus' RNA assays, we selected ~450 samples from a variety of cancer types for resequencing our new RS.v2 assay. Correction factors were derived from our large reference set of RS.v1 samples and paired RS.v1/RS.v2 samples. The correction strategy was then evaluated on a testing set of ~100 paired samples.

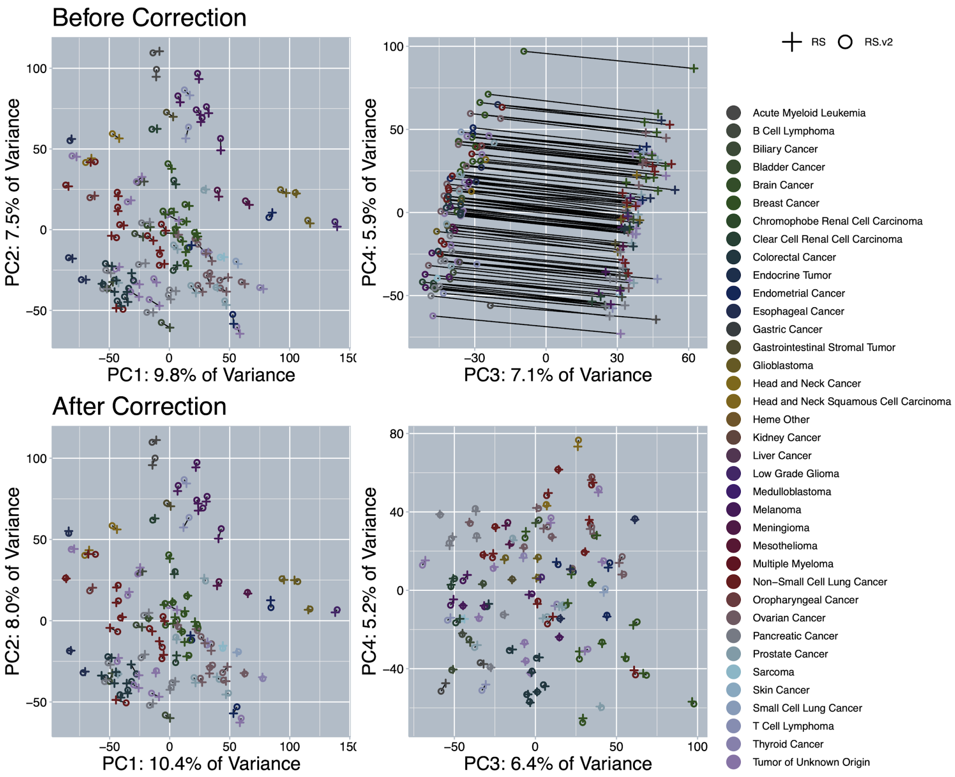
We found per-gene linear corrections were sufficient to remove all systematic differences between the two datasets. For each gene *i*, the *corrected-to-v1* expression value was computed as:

*expression_i * m_i + b_i,*

where *expression_i* is the original uncorrected RS.v2 expression ($\log_2$ TPM) for that sample, *m_i* is the slope term, and *b_i* is the intercept term for *gene_i*. These slope and intercept terms for each gene are learned to match the v1 and *corrected-v2* distributions in our matched datasets. This match was optimized by minimizing a weighted loss function that can account for paired-sample status. By utilizing paired sample information, we ensured that the resulting correction factors will be robust across sequenced samples of any cancer type as the same linear correction can be applied to all RS.v2 samples.

ii. Results of Batch Correction

Principal component analysis of normalized gene expression before and after batch correction. Samples represented on this plot are ~100 paired RS.v1/RS.v2 samples used as a testing set and were not part of training. Samples are colored by cancer cohort and labelled by assay type. Paired samples are connected by a line. Following batch correction, no principal component was associated with RNA assay type, and samples cluster by sample and cancer type.



4. Tempus|RS.v2 Fusions

The Tempus pipeline detects a variety of gene rearrangements using an ensemble methodology. Detection of candidate transcript fusions is performed using two separate open-source structural variant detection algorithms: (1) 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). (2) 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 v2.3.4 is run to annotate the integrated fusion calls including encoded protein domains and reading frames and to facilitate identification of clinically-relevant and/or druggable transcript fusions. Rigorous downstream filtering is applied in the clinical workflow to remove potential technical artifacts and identify high-quality, well-supported reportable fusions.

IV. Appendix

1. List of reference files used in the RNA pipelines

Workflows listed below can be found in the delivered **g_pipeline_version.csv** file where each patient and analysis id is mapped to a specific workflow and pipeline version.

Workflow ID*	Software	Genome build
tempus_rna_expression_v1/v2, rna_fusion_v1/v2	Kallisto	Fasta file: GRCh37.75 GTF file: GRCh37.p13
tempus_rna_expression_v2, tempus_rna_variant	STAR	Fasta file: GRCh37 GTF file: GRCh37, GENCODE version 27 (Ensembl 90) mapped to GRCh37 with gencode-backmap
rna_fusion_v1/v2	STAR, STAR-Fusion	Fasta File: GRCh37 GTF file: GRCh37, version 19 (Ensembl 74), from GENCODE
Tempus_rs (v1)	MCTP, STAR	Fasta file: GRCh38

The FASTA format is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences. (https://en.wikipedia.org/wiki/FASTA_format)

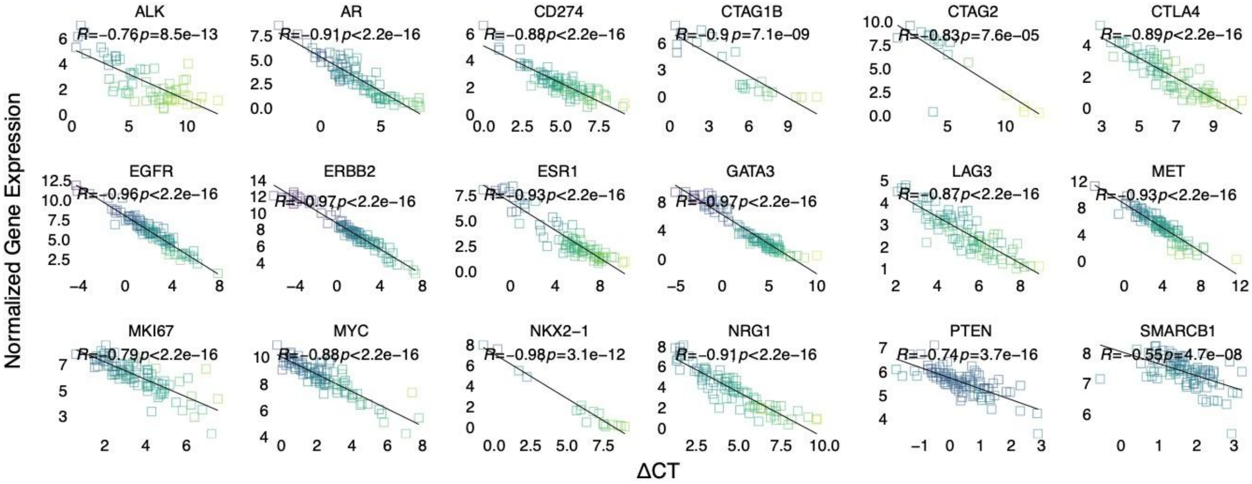
The Gene transfer format (GTF) is a file format used to hold information about gene structure. It is a tab-delimited text format based on the general feature format (GFF), but contains some additional conventions specific to gene information. (https://useast.ensembl.org/info/website/upload/gff.html)

2. Tempus|RS.v2 Validation Methods

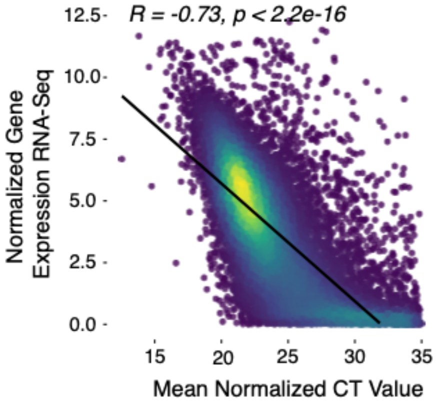
a) Linearity Study

88 samples were utilized in this study. Samples represented multiple cancer types with expression values near the low, midpoint, and high values of the analytical measurement range. qPCR with TaqMan probes was run for 18 genes, 2 housekeeping genes and 2 negative control genes for all samples. Delta CT values were compared to RNA-Seq gene expression normalized values.

Figure 1: Correlation between qPCR CT values and normalized gene expression.
(A). Mean ΔCT values and normalized gene expression in clinical samples for 18 oncogenes. Mean delta CT values for each sample/gene were normalized to the mean of the average CT values of housekeeping genes *AAMP* and *CANX* for that sample.



(B) Universal Human Reference (UHR) correlation between normalized qPCR CT values and gene expression with across 17,321 genes.



b) Concordance with DNA Copy Number Variant (CNV)

Among the top 10 most frequently amplified and deleted genes, we compared expression among patients with deep deletions (0 copies) and patients with amplifications (≥8 copies).

Figure 2: Top 10 Genes with Most Reported Deletions. Patients with only one detected copy were omitted from this plot.

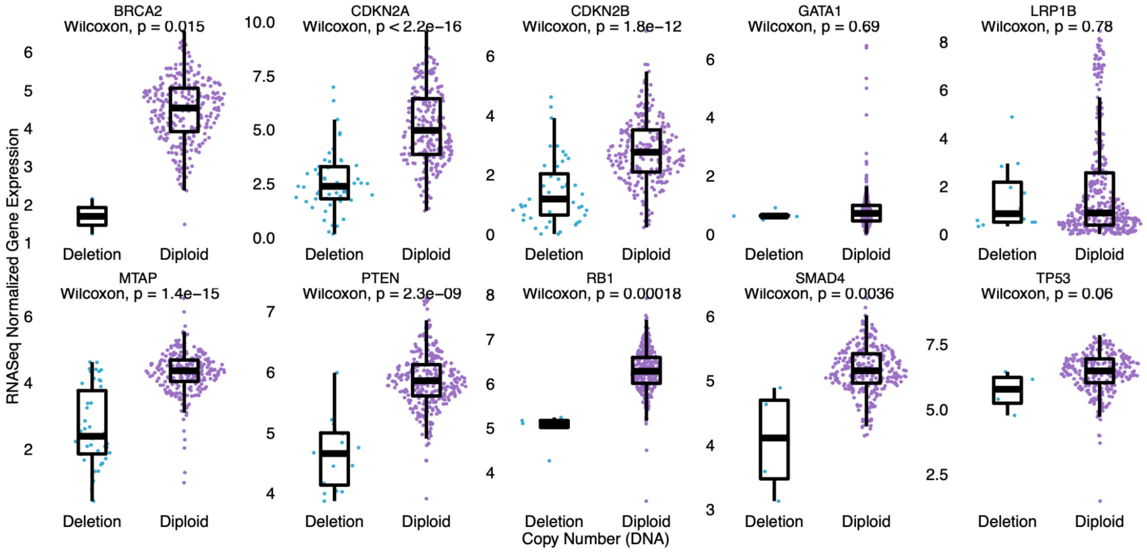
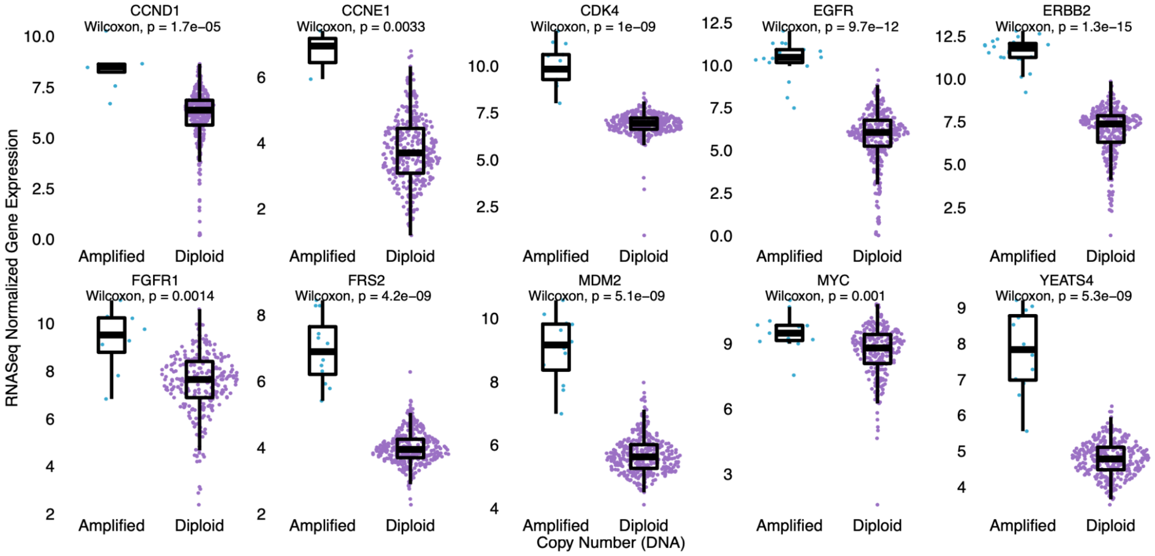


Figure 3: Top 10 Genes with Most Reported Amplifications. Amplified was defined as 8 or more copies.


RNASeq Normalized Gene Expression
CCND1
Wilcoxon, p = 1.7e−05
Amplified
Diploid
CCNE1
Wilcoxon, p = 0.0033
Amplified
Diploid
CDK4
Wilcoxon, p = 1e−09
Amplified
Diploid
EGFR
Wilcoxon, p = 9.7e−12
Amplified
Diploid
ERBB2
Wilcoxon, p = 1.3e−15
Amplified
Diploid
FGFR1
Wilcoxon, p = 0.0014
Amplified
Diploid
FRS2
Wilcoxon, p = 4.2e−09
Amplified
Diploid
MDM2
Wilcoxon, p = 5.1e−09
Amplified
Diploid
MYC
Wilcoxon, p = 0.001
Amplified
Diploid
YEATS4
Wilcoxon, p = 5.3e−09
Amplified
Diploid
Copy Number (DNA)