Tempus Bioinformatics Pipeline: Solid Tumor Assays (xT, xT.v2, xT.v3, xT.v4, xE, xE.v2, xO)

This document describes the operation of Tempus' bioinformatics pipeline for our solid tumor assays, including the xT assays (versions 1-4, see Appendix 1), the exome-level xE assays (versions 1-2), and our legacy xO assay. It is intended to provide a deeper understanding of our current 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. Much of the information provided here has been compiled from prior Tempus publications, which are listed at the end of the document (and cited throughout). For information regarding the RNA-seq data associated with these assays, please refer to the Tempus document Tempus Bioinformatics Pipeline: RNAseq Assays. 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.

DNA Pipeline

Demultiplexing and FASTQ generation

The 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 are generated per sample, corresponding to all full-length forward and reverse reads, respectively.

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's resulting FASTQ file has an insufficient number of reads to exceed 50 megabytes in size, 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.

Read alignment and BAM file generation[1]

FASTQ files are aligned to the 19th edition of the human reference genome build (hg19) via the Novoalign alignment algorithm. During this process, remaining adapter sequences are trimmed. A SAM output file is generated which is then converted to a BAM file. This file is then sorted by chromosome, indexed, and PCR duplicate FASTQ entries are marked via Novosort.

Variant calling - SNVs and Indels

All data provided for research use to partners uses an analysis pipeline that identifies two primary classes of short variants: single nucleotide variants (SNVs) and insertion-deletion variants (indels), detected by two variant callers, Freebayes and Pindel. Paired sample variant calling is performed on tumor samples and their respective matched normal controls, when available. Of note, a ploidy

value greater than 2 (ploidy = 3) is used by the Freebayes variant caller, in order to enhance the sensitivity for calling somatic variants. Variant filtering is performed to remove low-quality sequence data and sources of sequencing artifacts.

In the absence of a matched normal sample, called variants are classified as either germline or somatic based on a Bayesian model of the likelihood that the variant is somatic, following an approach similar to methods described by Halperin et al. (BMC Med Genomics. 2017 Oct 19;10(1):61. https://doi.org/10.1186/s12920-017-0296-8) and Sun et al. (PLoS Comput Biol. 2018 Feb; 14(2): e1005965. https://dx.doi.org/10.1371%2Fjournal.pcbi.1005965). Selected genes have amplified prior expectations for somatic variants, to avoid mis-classifying potentially pathogenic somatic variants as germline. The Bayesian prior probability is informed by internal and external databases of variants (including dbSNP, gnomAD, and COSMIC) as they are observed in germline and tumor samples, as well as the copy state, variant allele frequency and tumor purity of the cancer sample. Each variant is classified as germline, somatic, or uncertain by this model. Uncertain variants are subsequently treated as somatic variants for filtering purposes.

Table 1 lists the primary filters typically applied to all variants supplied for research use to our partners, according to the Tempus panel type. Additional custom filtering criteria are also used to evaluate a subset of variants; these are applied in order to account for known or potential artifactual variants, as well as to accommodate variants that should be “rescued” for subsequent analysis despite failing one or more of our standard (or custom) filters. The custom filters are derived from a continuous analysis of pipeline performance, across panels, and are applied to variants associated with strand bias, homopolymers, polynucleotide repeats, pseudogene regions, etc. The somatic variant allele fraction (VAF) filter threshold in the bioinformatics pipeline is set to 1%, although for SNVs and indels we typically observe a minimum of 5% or 10% VAF, respectively, unless previously whitelisted, within regions of enhanced coverage, or otherwise retained, for subsequent analysis. Where not specified, thresholds shown in Table 1 apply to variants called by both the Freebayes and Pindel variant callers.

Table 1: Filter Passing Criteria for Variants Called by Solid Tumor Pipelines

Variant Attribute	Passing Condition	xT/ xT.v2	xT.v3/ xT.v4	xE/ xE.v2	xO
Supporting reads for somatic variants called by Freebayes	>	3	3	3	3
Germline alternate allele supporting reads called by Freebayes[1]	>	3	3	3	3
Supporting reads for somatic variants called by Pindel	>	10	10	3	10
Germline alternate allele supporting reads called by Pindel[1]	>	10	10	3	10
Coverage for somatic or germline variants called by Freebayes	>=	35x	35x	35x	35x
Normal sample coverage for germline variants called by Freebayes[1]	>=	35x	35x	35x	35x
Coverage for somatic or germline variants called by Pindel	>=	75x	75x	35x	75x
Normal sample coverage for germline variants called by Pindel[1]	>=	75x	75x	35x	75x
Allele fraction for somatic variants	>	1%	1%	1%	1%
Allele fraction for germline SNVs	>=	30%	30%	30%	30%
Allele fraction for germline indels	>=	20%	20%	20%	20%
Tumor to normal allele fraction ratio for somatic variants	>=	5	5	6	6
Mapping quality score for supporting reads	>=	30	30	30	30
Base quality score for bases contributing to variants called by freebayes, either as flanking or centered.	>=	20	20	20	20

1) Includes equivalent tumor sample alternate reads for germline variants, in tumor-normal paired analyses.

Quality control

The primary quality control thresholds applied post-analysis are shown in Table 2. In addition to meeting these per-sample QC metrics, when multiple samples have been sequenced from a given patient, we also apply a germline-variant 'fingerprinting' analysis, in order to confirm that such samples have been derived from the same source material. Further, the positive sensitivity control processed and sequenced with the sample is expected to meet internally established QC acceptance criteria for positive controls.

Table 2: Passing Quality Control Criteria for Solid Tumor Analyses

Analysis Attribute	Passing Condition	xT/ xT.v2	xT.v3/ xT.v4	xE	xE.v2	xO
Coverage of target regions at the 90th percentile of regions in the tumor sample	>	N/A	N/A	100x	100x	N/A
Coverage of target regions at the 95th percentile of regions in the tumor sample	>	100x	150x	N/A	N/A	300x
Coverage of target regions at the 95th percentile of regions in the normal sample	>	50x	100x	N/A	N/A	250x
Total read count in the normal sample	>=	10M	10M	100M	50M	30M
Total read count in the tumor sample	>=	10M	10M	100M	100M	30M

Variant annotation[2]

Predicted functional effect for each mutation is curated by automated software using information from multiple databases. Alterations are investigated using criteria that include known evolutionary models, functional data, clinical data, known gene-disease relationships, hotspot regions within genes, internal and external somatic databases, primary literature, and the unique combination of data derived from the availability of paired tumor/normal DNA samples. A weighted-heuristic model predicts the functional effect for each alteration, which groups alterations into one of several categories, including pathogenic, likely pathogenic, variants of unknown significance (VUS), benign, likely benign, etc. The logic of the model is performed by automated software that was developed using the AMP/ASCO/CAP/ClinGen Somatic working group and ACMG guidelines.

To determine that high sensitivity variant calling can be performed with high levels of confidence, there are several quality control analyses performed as part of the assay:

1. Alignment and Coverage Quality Control: This first process computes a series of statistics on the output of the reference genome alignment process such as read mapping rates, on-target rates, and PCR duplication rates. The second process computes the depth and uniformity of coverage for all bases analyzed by the assay. The goal is to confirm that all bases within the target regions are sequenced to sufficient depth so that if a variant were present, it would be called successfully down to the lower limit of detection of the assay.
2. Sequence Data Quality Control: This includes per-base quality scores for both forward and reverse reads output from the sequencer.
3. Variant Quality Control: This process involves computing the depth of coverage for all recurrent hotspot variants as well as therapeutically critical tumor suppressor genes and confirming that there is sufficient sequencing depth at the genomic locus to confidently call a negative or a positive result. Variant calling filter reasons are aggregated and compared against historical data from previous QC passing runs to confirm that variants generated as candidates within the given sample conform to expected parameters regarding base quality, variant allele fraction, and population frequency.

CNV segments[1]

CNV analyses aim to utilize variant calls and sequencing coverage to make copy number estimations of genomic regions. Tempus utilizes two distinct algorithms to generate first-pass and final copy estimation. Both follow the same general pattern outlined below, although specific details are predominantly limited to the second of these, a Tempus-developed copy number algorithm.

Preliminary copy number analysis for datasets delivered to research partners is performed by cnatools and produces a segmentation of the genome into a set of chromosomal regions of specified copy number. Segmentation is based on panel coverage data, but uses a greedy algorithm, which assumes that un-probed regions between probes or at the beginning of a chromosome share the copy number of its neighbors. The segment data are also transformed to provide probe-level copy number values. This algorithm was optimized for a specific gene panel and as such does not perform optimally on all Tempus panels. It is utilized as a first-pass approach to inform variant filtering, as it does not require filtered variants, while the Tempus algorithm does. For details on differences in CNV analysis between versions of xT, see Appendix 2.

The Tempus copy number algorithm attempts to derive approximate integer copy number estimates for total copies and minor allele count by estimating tumor purity, ploidy, and B-allele frequencies within the tumor sample. The following steps are taken by the algorithm:

Inputs:

1. Tumor BAM + index file
2. Normal BAM (optional) + index file
3. Variants file

4. Human reference genome (hg19)
5. Gene targets file
6. Pool of process matched normal samples
7. Blacklist of recurrent problematic areas of the genome (optional)

Processing Steps:

File ingestion: The set of BAM files provided to the algorithm are read in for each genomic target specified in the targets file. Per region coverages are computed and stored along with per variant coverage read in from the variant file. Variants are mapped to genomic positions, and germline variants are classified as heterozygous or homozygous based on variant allele frequency. For heterozygous germline variants present in both the tumor and the germline samples, the deviation of variant allele frequency between the tumor sample and the germline sample is computed as the log of the odds ratio of the count of alternate alleles in tumor versus normal, hereafter referred to as logBAF (Shen, Nucleic Acids Res. 2016 19;44(16):e131 https://doi.org/10.1093/nar/gkw520).

Initialize tumor purity estimates: In order to generate an initial estimate of tumor purity for downstream analytical processes, the highest variant allele fraction somatic variant is selected and its variant allele fraction is stored. Concomitantly, the 90th percentile of b-allele frequency deviation between tumor and germline is taken as a basic first pass estimate of tumor purity. Deviation of variant allele fraction in regions showing loss of heterozygosity will correspond to approximately half tumor purity. The interval between these two measurements is used to bound subsequent best fit tumor purity estimations.

Read data normalization:

1. Depth normalization against normal pool
2. GC-correction across GC percentiles for all target regions
3. Principal components noise correction against the normal pool
4. Log ratio computation against both the normal pool and the matched normal sample

Following normalization, data proceed to segmentation.

Circular binary segmentation (CBS): Tempus uses a modified version of CBS wherein both log ratio and logBAF are integrated into a single estimate. Recursive splitting of each chromosome is computed using the T-squared statistic, the multivariate generalization of the Student's t-statistic (Shen, 2016). At first pass, a tree is generated for each recursive split and each region is assessed for the possibility of focal amplification (large changes in log ratio). Small regions of high deviation are protected, and then the tree is pruned based on a parameterized threshold set by the user on the maximum Hotelling statistic at which splits are acceptable. Following pruning, the tree is stabilized and segment summary statistics are computed. These include median log ratio, median logBAF, number of heterozygous variants, and feature length.

Zygosity shift estimation: Prior to attempting to assess tumor purity, the zygosity shift associated with genome ploidy is computed. In order to correct for this, a correction to the true diploid log ratio must be applied to all segments. First, allelically balanced segments of sufficient length are selected and clustered by k-means. The lowest three clusters, representing the 0-2-4 or 2-4-6 copy clusters are selected for analysis. If the lowest cluster exceeds n% of the genome in size, the 0-2-4 copy state hypothesis will be rejected as high percentages of genomic deep deletions are inimical to cell survival. If both models are plausible, the median log ratio adjustment for the hypothetical two-copy state will be applied for both models and proceed to tumor purity/copy number fitting.

Tumor purity estimation / Copy number calling: Following segmentation, there is an integrated process by which copy number and tumor purity are assessed using a grid search methodology. Starting from the initial estimate of the tumor purity lower bound, a copy number state matrix is generated containing total/minor copy state combinations and their expected log ratios and logBAF given the initialized tumor purity. Each segment is projected into the matrix and the log probability that it belongs to each analyzed copy state is computed based on drawing from a normal probability density function defined by the expected log-ratio and the pre-computed log ratio standard deviation. The same process is applied for logBAF with a sliding-weight scale based on the number of heterozygous germline variants observed within the segment. This is done to account for noise in logBAF in the context of sparse observations.

This analysis occurs for each segment at each tumor purity grid value sliding by user-defined interval. Following the grid computation, each segment will vote on the most likely tumor purity by the estimate on which it has the highest probability fit. Each segment's vote is then projected and peaks within the fit histogram are selected as high confidence tumor purity values. Subsequent to tumor purity candidate selection, a loss function is applied to each purity value wherein the complete target set is fit to that tumor purity estimate and sum squared error is computed against expected b-allele frequencies for unbalanced copy states between zero and four copies. The tumor purity estimate with the lowest model error will be selected as the best fit and integer total/minor copy number will be fit to all segments.

Copy number variants are typically evaluated in samples having greater than 30% tumor purity (or greater than 40% for copy losses called by the xT.v4 and xE.v2 assays), with observed copy number gains of eight or more copies, or when a homozygous loss (zero copies) is detected. Gains of seven copies, or CNVs detected in samples with tumor purities below threshold, may be considered for some genes, in circumstances where corroborative or prior data are available.

Quality Control: A blacklist for recurrent problematic areas of the genome removes cnv calls from areas that have been demonstrated to be inaccurate based on panel-specific issues.

DNA fusions[2]

Overview:
The Tempus DNA fusion detection pipeline runs the SpeedSeq analysis pipeline (Chiang et al. Nat Methods 12, 966–968, 2015. https://doi.org/10.1038/nmeth.3505) to identify the set of structural variants within a sample and do some initial filtering of structural variants to those with adequate read support. Additional filtering is done to identify structural variants with the ability to produce biologically-relevant chimeric (fusion) proteins. Domains associated with fusion proteins are also compiled.

Software used:
Initial alignment of tumor fastq files uses BWA and aligns to hg19. Read pairs aligned directly, with split reads aligning to multiple positions and with read pairs mapped to discordant positions, are identified and separated from one another by samblaster. Structural variants are called by Lumpy, filtered based on supporting read count, and annotated by an enhanced version of AGfusion.

Quality Control:
All fusions detected remain in the dna fusion data file. Those without adequate read support (defined as a total of at least 50 supporting reads, with a minimum of 2 of those reads being discordant pairs and a minimum of 30 being split reads) have a 'filtered' notation included.

Immuno Pipeline[1]

Tumor mutational burden (TMB)

TMB is calculated for each panel by dividing the number of non-synonymous mutations by the Mb size of the appropriate panel. All non-silent somatic coding mutations, including missense, indel, and stop-loss variants, typically with coverage greater than 100x and an allelic fraction greater than 5%, are included in the count of non-synonymous mutations. Hypermutated tumors are considered TMB-high if they have a TMB >9 mut/Mb. This threshold was established by testing for the enrichment of tumors with orthogonally defined hypermutation (MSI-H) in the larger Tempus database.

Human leukocyte antigen (HLA) typing

HLA class I typing for each sample is performed using either Optitype or Kourami on DNA-seq data. Normal samples are used as the default reference for matched tumor-normal sample analyses. Tumor-only determined HLA type is used when the normal sample did not meet internal HLA coverage thresholds or when a sample is run as tumor-only. We are currently extending our HLA typing to include class II variant alleles, which is expected to be available in the near future.

Neoantigen prediction

Neoantigen prediction is performed on all non-silent mutations identified by the bioinformatics pipeline for the solid tumor panels, including indels, SNVs, and frameshifts. For each mutation, the binding affinities for all possible 8-11 amino acid peptides containing that mutation is predicted using MHCflurry. For alleles with insufficient training data to generate an allele-specific MHCflurry model, binding affinities are predicted for the nearest neighbor HLA allele as assessed by amino acid homology. A mutation is determined to be antigenic if any resulting peptide is predicted to bind to any of the patient's HLA alleles using a 500 nM affinity threshold. RNA support is calculated for each variant using varlens (https://github.com/openvax/varlens). Predicted neoantigens are determined to have RNA support if at least one read supporting the variant allele can be detected in the RNA-seq data.

Microsatellite instability (MSI) status

The Tempus xT panels include probes for microsatellites that are frequently unstable in tumors with mismatch repair deficiencies (44 loci on the xT/xT.v2, 239 loci on xT.v3/xT.v4), in order to assess microsatellite instability (MSI). The MSI classification algorithm uses reads mapping to these frequently unstable regions in order to classify tumors into four categories: microsatellite instability-high (MSI-H), microsatellite stable (MSS), microsatellite equivocal (MSE), or undetermined. This test can be performed with paired tumor-normal samples or tumor-only samples. For each microsatellite locus, a minimum of 30 mapped sequencing reads, in tumor and normal (for paired analyses) samples, is required in order to be included in the MSI analysis. For each panel, a predetermined number of microsatellites above this minimum read depth is required for MSI status to be assessed (20 for xT/xT.v2, 150 for xT.v3/xT.v4). Each locus is individually tested for instability, as measured by the difference in the number of repeats in tumor data compared to normal data, using the Kolmogorov-Smirnov test. If $p \leq 0.05$, the locus is considered unstable. The proportion of unstable microsatellite loci is then fed into a logistic regression classifier previously trained on samples obtained from the TCGA colorectal and endometrial cohorts, which have clinically determined MSI statuses. For MSI testing in tumor-only analyses, the mean and variance for the number of repeats are calculated for each microsatellite locus. A vector containing the mean and variance data is entered into a support vector machine classification algorithm. Both algorithms return the probability of the patient being MSI-H. If the probability of MSI-H status exceeds the high probability threshold set for the assay (70% for xT/xT.v2, 84% for xT.v3/xT.v4), the sample is classified as MSI-H. If the probability of MSI-H is below the low probability threshold (50% for xT/xT.v2, 55% for xT.v3/xT.v4), the sample is considered MSS. If the probability of MSI-H is between these two thresholds (0-70% for xT/xT.v2, 55%-84% for xT.v3/xT.v4), the test results are considered too ambiguous to interpret and these samples are classified as MSE. If no MSI call can be made, the status is classified as 'undetermined.' The probability thresholds specifically for tumor-only analyses in xT.v3/xT.v4 are different from the above, see Appendix 3. Assessment of

MSI status has not been validated for use with the xE or xO panels; when available for these panels, MSI data should be interpreted with caution.

Immune infiltration estimation[2]

The relative proportion of immune subtypes is estimated using a support vector regression (SVR) model, which includes an L2 regularizer and an epsilon insensitive loss function, similar to that of Newman et al. (Nat Methods. 2015;12(5):453–457. https://doi.org/10.1038/nmeth.3337). The SVR is implemented in Python using the nuSVR function in the SVM library of scikitlearn, using the leukocyte gene signature reference matrix (LM22) described by Newman et al.

Tempus References cited

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

Appendices

1. The full list of genes included in the current version of xT can be downloaded and viewed at https://www.tempus.com/genomic-profiling/#proprietary-sequencing. Different genes are included in different versions of xT as shown in Table 3. Genes listed under xT.v3 and xT.v4 are *not* in either xT.v3 or xT.v4, but are present in xT.v1 and/or xT.v2. Genes listed under xT.v1 are *not* xT.v1, but are present in xT.v2, xT.v3, and xT.v4. Genes listed under xT.v2 are *not* in xT.v2, but are present in xT.v1, xT.v3, and xT.v4.

Table 3: Gene Panel Differences Across xT Versions

xT.v1	xT.v2	xT.v3	xT.v4
FUS, HSD11B2, PHLPP1, PPARD, OLIG2, CYSLTR2, MN1, C11orf30	FUS, HSD11B2, PHLPP1, PPARD, OLIG2, CYSLTR2, MN1, C11orf30	ATM;C11orf65, C11orf30;EMSY, PDPK1, UGT1A1;UGT1A9,	ATM;C11orf65, C11orf30;EMSY, PDPK1, UGT1A1;UGT1A9,

(EMSY), RHEB, CHD7, HOXA11, PPARG, VEGFB, TRAF7, WNK1, CUL4A, CUL4B, SYNE1, RRM1, CARM1, PHLPP2, ASPSCR1, HSD3B1, PPARA, HSD3B2, EIF1AX, MAGI2, WNK2, TFE3, TFEB, NOTCH4, TGFBR1, LCK, TCEB1 (ELOC), PHGDH, UGT1A1;UGT1A9	(EMSY), RHEB, CHD7, HOXA11, UGT1A1, VEGFB, TRAF7, WNK1, CUL4A, CUL4B, SYNE1, RRM1, CARM1, PHLPP2, ASPSCR1, HSD3B1, PPARA, HSD3B2, EIF1AX, MAGI2, WNK2, TFE3, UGT1A9, TFEB, NOTCH4, TGFBR1, LCK, PHGDH, PTPRT	EMSY;C11orf30	EMSY;C11orf30

xT.v3 contains the same set of genes as xT.v4 with the exception of 7 changes in gene names as outlined here:

- FAM175A has been updated to ABRAXAS1 in xT.v4.
- TCEB1 has been updated to ELOC in xT.v4.
- C11orf30 has been updated to EMSY in xT.v4.
- MRE11A has been updated to MRE11 in xT.v4.
- WHSC1 has been updated to NSD2 in xT.v4.
- PARK2 has been updated to PRKN in xT.v4.
- C10orf54 has been updated to VSIR in xT.v4.

2. Different versions of the xT panel do not target the same set of SNPs used for CNV analysis. In xT.v2, a list of 1723 SNPs are specifically targeted; an additional 61 SNPs were added in xT.v3 and xT.v4 for a total of 1784 SNPs. Specifically targeted SNPs were not incorporated in CNV analysis for xT.v1.
3. Per MSI-testing in tumor-only analyses in xT.v3 and xT.v4, if the probability of MSI-H status exceeds the high probability threshold set for the assay (70%), the sample is classified as MSI-H. If the probability of MSI-H is below the low probability threshold (30%), the sample is considered MSS. And If the probability of MSI-H is between 30%-70%, the test results are considered too ambiguous to interpret and these samples are classified as MSE.