RNAsum: A tool for personalised genome and transcriptome interpretation for improved cancer diagnostics

Sehrish Kanwal[*,#,†,1,2], Jacek Marzec[*,#,1,4], Joseph H.A. Vissers[1,2], Peter Diakumis[1,2], Leila Varghese[1,2], Luke Tork[1,3], Oliver Hofmann[1,2], Stephen J. Luen[5,6], Sean Grimmond[#,1,2]

* Equal contributions

Corresponding authors

† Lead author

1. **Collaborative Centre for Genomic Cancer Medicine, University of Melbourne, Melbourne, Victoria, Australia**
2. **Department of Clinical Pathology, University of Melbourne, Melbourne, Victoria, Australia**
3. **The Wilson Centre for Blood Cancer Genomics, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia**
4. **AccelBio, Collaborative Laboratory to Foster Translation and Drug Discovery, Cantanhede, Portugal**
5. **Department of Medical Oncology, Peter MacCallum Cancer Centre, Parkville, VIC 3000, Australia**
6. **Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, VIC 3010, Australia**

Abstract

The integration of whole-genome sequencing (WGS) and whole-transcriptome sequencing (WTS) has revolutionized cancer diagnostics, enabling comprehensive molecular profiling of tumours. While WGS uncovers genomic alterations such as single nucleotide variants (SNVs), structural variants (SVs), and copy number changes, WTS provides complementary insights into transcriptomic changes, including gene expression levels, alternative splicing, and fusion events. Together, these technologies offer unparalleled potential to guide precision oncology by identifying actionable biomarkers and stratifying patients for targeted therapies or clinical trials.

However, the integration of WGS and WTS data remains a significant bioinformatics challenge due to biological variability, tumour-specific complexities, and the lack of harmonized reference datasets. To address this, we developed RNAsum, an open-source tool designed to combine and interpret WGS and WTS data from individual cancer patient samples. RNAsum compares patient

data against curated cohorts from The Cancer Genome Atlas (TCGA) and integrates quantitative expression data with genomic insights to enhance diagnostic accuracy, validate genomic findings, and prioritize clinically relevant alterations.

We demonstrate the utility of RNAsum through case studies, showcasing its ability to confirm clinically reportable variants, refine diagnoses, and identify novel therapeutic targets. By bridging the gap between genome and transcriptome analyses, RNAsum represents a step forward in the routine clinical application of multi-omics for personalized cancer care. The tool is freely available as an R package on GitHub, supporting accessibility and scalability for diverse research and clinical settings.

Introduction

High throughput sequencing (HTS) technologies have revolutionized the field of precision oncology by enabling fast and accurate analysis of whole-genome and transcriptome data from tumours (Cuppen et al. 2022). Facilitated with decreasing sequencing cost, whole-genome sequencing (WGS) provides valuable insights into molecular mechanisms underlying oncogenesis and genomic aberrations such as single nucleotide variants (SNVs), small insertions and deletions, structural variants (SVs) and copy number (CN) rearrangements (Alioto et al. 2015) (Rosenquist et al. 2022). Furthermore, it has shown great utility in studying clinically actionable cancer genome alterations including tumour mutational burden (TMB) (Allgäuer et al. 2018), homologous recombination deficiency (HRD) (de Luca et al. 2020) and microsatellite instability (MSI) (Fujimoto et al. 2020). As the field of HTS continues to expand and support precision oncology initiatives globally, whole-transcriptome sequencing (WTS) has gained immense traction in cancer research and personalized medicine (Roychowdhury and Chinnaiyan 2016) due to its ability to provide comprehensive insights into the transcriptome, the complete set of the gene readouts called RNA molecules in a cell or tissue. It allows quantification of expression levels to understand gene activation or silencing in a cancerous cell (Nagy, Munkácsy, and Győrffy 2020) and identify novel transcripts, including alternative splicing variants (Zhang et al. 2021).

WGS combined with WTS, often known as WGTS, has demonstrated the ability to provide a nearly complete fingerprint of patients' tumour profiles and has become a standard of care for molecular testing (Jobanputra et al. 2022). While WGS offers insights into coding and non-coding changes in patients' genome, WTS enhances our understanding of disease aetiology by providing

quantified measurement of gene expression and identifying functional changes such as fusions. This comprehensive diagnostic approach can improve precision oncology practices by providing complete characterisation and molecular profiling for identification of targetable alterations that stratify patients for eligible clinical trials and off-label drug treatment (Gong et al. 2021). By understanding the unique genetic mutations and expression profile of a patient, clinicians can use these data for disease detection, prognosis and tailor therapies to be more effective and minimize side effects.

Although most current clinical genomics tests focus on genotyping mutations in WGS data, comprehensive genome sequencing combined with transcriptome analysis has demonstrated significant value in precision oncology practice (Jobanputra et al. 2022). In several studies ((Zheng et al. 2016), (Bailey et al. 2016), (Kinnersley et al. 2024)), findings from WGS data are cross-validated or confirmed using transcriptome-based results. However, there remains a lack of a comprehensive bioinformatics approach capable of effectively post-processing, combining, summarizing, and visualizing results from both WGS and WTS data to enable personalized interpretation of individual cancer samples. The tumour-only quantified nature of RNA-sequencing (RNA-seq), intrinsic biological variability among samples (even within the same condition), as well as lack of harmonized cancer specific reference data, accurate transcript quantification, and variable methods for data preprocessing makes it difficult to compare against whole-genome based findings.

We propose to address this challenge through RNAsum, a tool that integrates WGS and WTS data from a single patient sample and compares them against a reference cohort from The Cancer Genome Atlas program (TCGA) [9]. It generates an interactive report summarizing key findings to aid data interpretation, as well as supplementary resources that can be used programmatically for further data exploration and cohort comparison. Here, we focus on the practical implementation as well as the utility of RNAsum for WGTS interpretation in routine clinical practice. We describe the utility of RNAsum through clinical case studies, discuss features and limitations of our method and provide detailed usage details, thus establishing the clinical utility of proposed tool.

Methods

Reference Data collection

To assist in the interpretation of gene expression prospectively generated from patient cancer samples, we developed a solution which allows comparison of single-subject tumour-only expression profiles to large cancer-matched expression landscapes taken from the TCGA. Read count data from 10,079 cancer patients, covering 33 cancer types (Table S1) was obtained from the National Cancer Institute (NCI) Genomic Data Commons (GDC) Data Portal using the GDC Application Programming Interface (API) and data release 40.0 (March 29, 2024). These reference profiles were then harmonized with the individual sample to determine the relative gene expression level of any gene of interest in the individual, expressed as a percentile of the relative to the reference patient cohort expression spectrum.

In order to correct for potential batch effects that may arise when integrating the patient data with expression profiles from an external cohort, an internal reference cohort, assembled using in-house WTS samples processed through the same pipeline as the patient's WTS data, can be added to the workflow. The internal cohort can be prepared using scripts available on the GitHub page (https://github.com/umccr/RNAsum). The complete process, as well as the benefits of using an internal reference cohort, are described in supplementary materials.

Additionally, users can provide a customized list of genes-of-interest (GOIs) to the workflow, enabling exploration of their expression profiles in investigated patient relative to corresponding reference cohorts. If no custom gene set is supplied, the workflow defaults to analysing three pre-assembled gene sets: (1) "immune-related genes" (listed in PanelApp Immune markers gene panel and included to monitor potential cancer immune regulation), (2) "homologous recombination deficiency (HRD) genes", (listed in PanelApp Homologous recombination deficiency (HDR) gene panel), and (3) "known cancer genes", a curated set of 1,315 genes associated with tumorigenesis (defined in https://github.com/umccr/gene_panels/releases). These exemplar gene sets are routinely reviewed in individual patients to evaluate their immune response, genomic instability due to HRD, and the expression of key cancer-related genes.

Finally, we collected information from public knowledge bases, including FusionGDB, Cancer Genome Interpreter (CGI), Clinical Interpretation of Variants in Cancer (CIViC), Variant

Interpretation for Cancer Consortium (VICC) and OncoKB, to annotate WGTS findings and to provide clinical significance for identified alterations.

Data integration

With the input data presented in Figure 1 and summarized in Table 1, including both required and optional data, the integration is carried out in two steps. First, mRNA profiles from the patient, internal and external (TCGA) reference cohorts are combined at the WTS level (see Figure S1 and “WTS data integration” section for detailed description). Next, the expression data are overlaid with genomic alterations at the WGTS level, and clinically relevant annotations from public knowledge bases are added (“WGTS data integration” section).

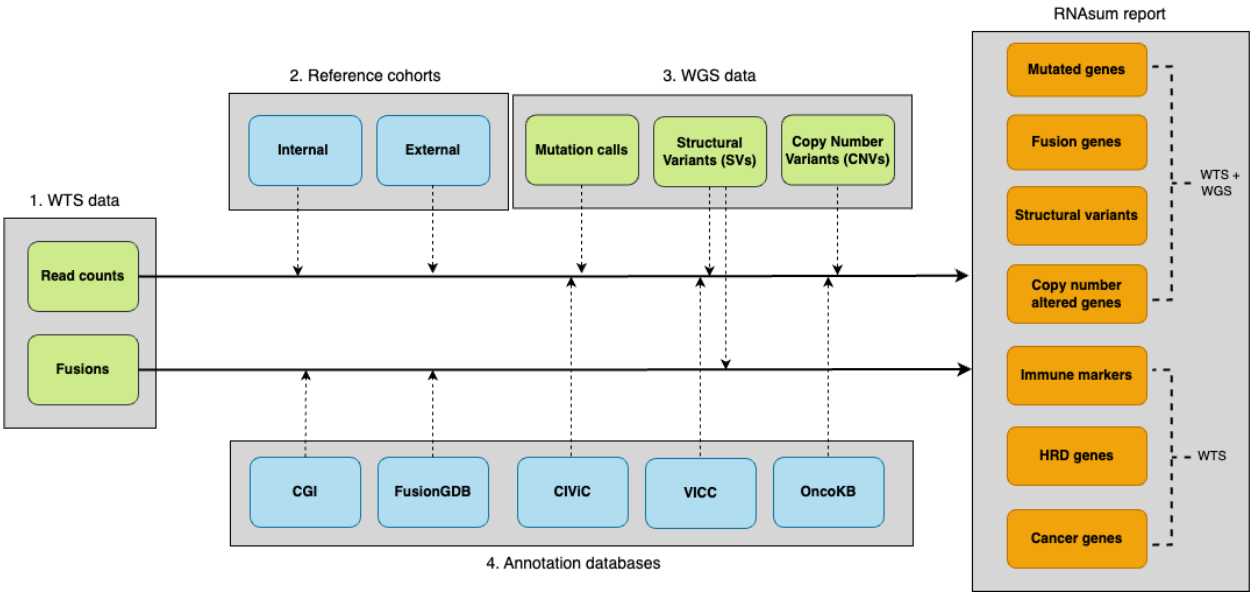


Figure 1. Diagram presenting the main workflow components (green - inputs, blue - reference data, orange - outputs) involved in (1) read counts and gene fusions data collection, (2) reference data processing, (3) integration with WGS-based data and (4) results annotation.

*Table 1. Summary of the input data and their respective sources, utilized in subsequent components of the workflow. * refers to Figure 1.*

Workflow component* (platform)	Input data	Data source	Required	Example tool / database
1 (WTS)	Transcript / gene abundances	Patient	Yes	Salmon (Patro et al. 2017), Kallisto (Bray et al. 2015)
1 (WTS)	Fusion genes	Patient	No	Arriba (Uhrig et al. 2021), DRAGEN
2 (WTS)	Transcript / gene abundances	Internal reference cohort	No	Salmon (Patro et al. 2017), Kallisto (Bray et al. 2015)
2	GOIs	User-defined list	No	OncoKB Cancer Genes (Suehnholz et al. 2024)
3 (WGS)	SNVs/Indels	Patient	No	PCGR (Nakken et al. 2018)
3 (WGS)	CNVs	Patient	No	PURPLE
3 (WGS)	SVs	Patient	No	Manta (Chen et al. 2016)

WTS data integration

The read count data from the following sources can be used as an input for the integration at the WTS level: the mandatory (1) patient sample and (2) external reference cohort, which can consist either of samples corresponding to the patient's cancer type or a collection of all patients available within the TCGA data, and an optional (3) internal reference cohort used for batch effects correction. The input read count data are combined and transformed to counts per million (CPM) values, genes with low counts are filtered out, and then the data are normalized to account for sample-specific effects (Figure S1). Additionally, the potential variation due to technical factors can be removed by performing batch-effects correction upon providing the internal reference cohort. Finally, the WTS data are standardized by converting them into Z-scores, which are comparable by measuring the observations in multiples of the standard deviation of a given sample, to facilitate interpretation of expression values (Figure S2).

WGTS data integration

The WTS data are integrated with WGS by overlaying gene expression levels onto genes bearing detected genomic alterations, including somatic single nucleotide variants (SNVs) and small insertions and deletions (indels) derived from Personal Cancer Genome Reporter (PCGR) tool, structural variants (SVs) detected by MANTA SV caller, and copy-number variants (CNVs) called by PURPLE purity ploidy estimator (Figure 1). Genome-based findings, if provided, serve as the primary source for prioritizing expression profiles, ensuring that the differences in expression levels between patient and reference cohorts for key genes are evaluated first. The integrated WTS and WGS data are then ranked by the difference observed between expression levels in the patient sample and the corresponding reference cohort, highlighting genes with the greatest potential clinical impact. Moreover, detected gene fusions, derived from Arriba or DRAGEN gene fusion callers, supporting WGS-based SVs are prioritized given additional evidence for their functional relevance. All reported alterations are annotated with clinically relevant information from public knowledge bases, including CGI, VICC, OncoKB and CIViC. Furthermore, detected gene fusions are functionally annotated with information from FusionGDB.

Interpretation and reporting

The outputs of the pipeline are an interactive report that prioritizes clinically relevant information to facilitate results interpretation, as well as supplementary resources that can be used programmatically for further data exploration and comparison against other patient cohorts. This report includes searchable tables and plots, organized into several sections as detailed below and illustrated in Figure 2. Each section features an interactive “summary” table and “expression profiles” plots.

The “summary” table contains comprehensive information about individual genes' mRNA expression levels in a patient's sample, the average mRNA expression across samples from TCGA cancer patients, and the difference between these two measurements. Notably, mRNA measurements are presented in both percentiles to facilitate intuitive interpretation, and Z-scores. Genes in these tables are sorted based on section-specific criteria and then by the decreasing absolute values representing the difference between the patient's and TCGA expression levels. The “summary” table also includes hyperlinks to relevant knowledge bases, such as VICC, OncoKB and CIViC, providing additional context, as well as information specific to each report section. The interactive plots comprise two main components: a distribution panel showing

percentile values as a function of expression levels, and a box plot panel displaying expression levels for each gene in the patient's sample compared to the reference cancer cohort. Additionally, we include a bar plot illustrating the read counts of individual genes across all samples.

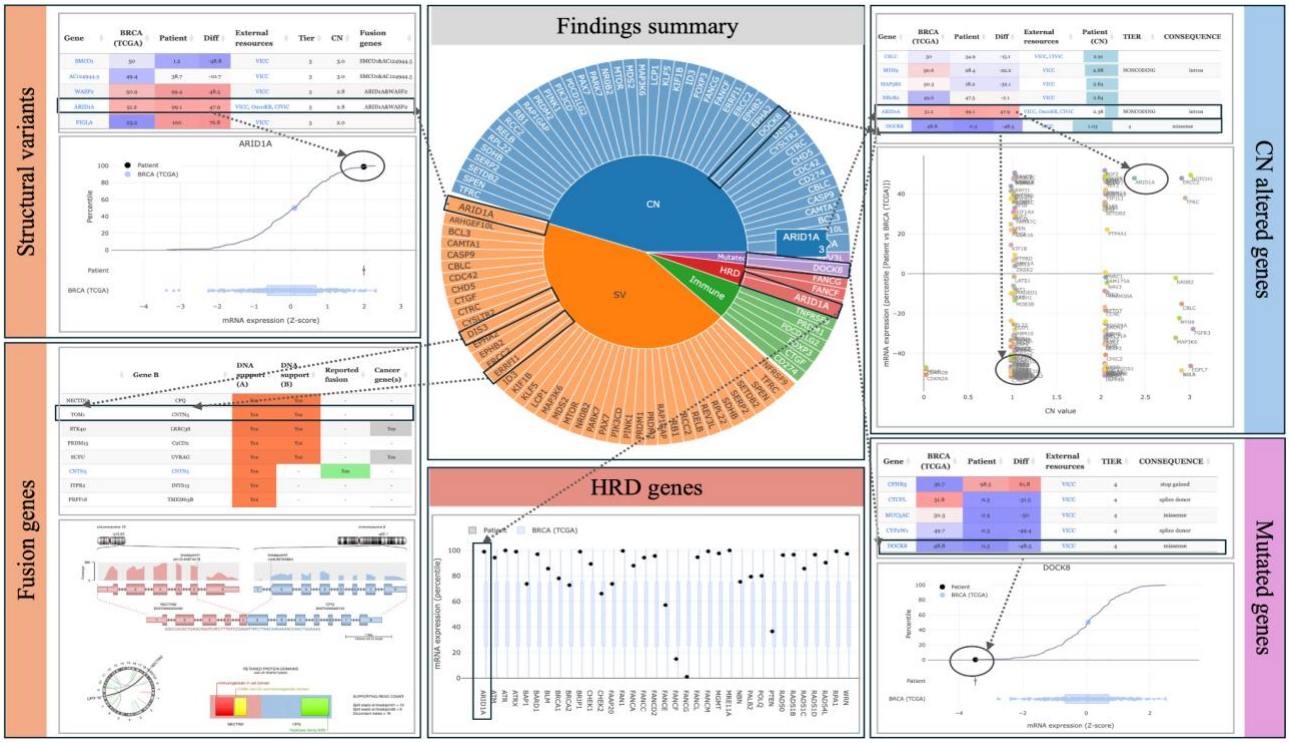


Figure 2. Diagram illustrating an overview of RNAsum interactive report structure and visualization features. It highlights the organization of the report into multiple sections, including the input data summary, findings summary, mutated genes, fusion genes, structural variants, copy number-altered genes, and a custom panel of HRD genes. This report overview showcases examples of summary plots and tables in individual sections, demonstrating how clinically relevant information is prioritized for interpretation.

Input data summary

The WTS input data are summarized at the top of the report, providing a comprehensive overview of the analysis. This section includes information about reference cohorts and patient samples used in the analysis, details about included and excluded genes along with the reasoning behind these decisions, visual representations of the initial library size generated across all samples, read count data filtering, transformation and normalization. Moreover, it contains interactive exploratory data analysis plots, including principal components analysis (PCA) and relative log expression (RLE) plots, generated before and after batch-effect correction, which facilitate identification of key factors affecting variability in the expression data.

Findings summary

The findings from the WGTS data integration are summarized in a form of an interactive doughnut plot and table presenting genes with detected alterations, which are further detailed in corresponding report sections. The genes are ordered by the number of report sections, in which they appear, highlighting those of particular interest due to evidence from multiple sources. Moreover, all affected genes are hyperlinked to databases, including VICC, OncoKB and CIViC, providing additional evidence for their clinical significance.

Mutated genes

The integrated WGTS data for genes with detected SNVs or indels are presented in a form of interactive “summary” table and “expression profiles” plots, as described previously. The “summary” table provides detailed information on mRNA levels for individual genes, as well as additional data about detected variant tiers and corresponding genomic alterations. These alterations encompass variant type and class, allele frequency, and both genomic and proteomic changes. Genes in the table are ordered primarily by variant tier, which reflects their classification and clinical impact, and secondarily by the difference between the investigated patient’s expression levels and those from the TCGA dataset.

Fusion genes

Gene fusions detected in WTS are summarized in an interactive table that provides comprehensive information, including the number of supporting split reads, genomic coordinates, and the locations of the breakpoints in the fused genes. This table also includes data on overlapping SVs identified in WGS data. Gene fusions that support WGS-based SVs or are reported in FusionGDB or CGI databases are prioritized, given the additional evidence for their functional relevance and potential clinical significance in the investigated patient. These fusions are presented in a genomic context using a circos plot, followed by an interactive “summary” table and “expression profiles” plots that demonstrate the mRNA levels of individual fusion genes observed in both the patient and reference cohorts. Additionally, this section provides visualizations (generated by Arriba) that offer detailed information about the transcripts involved in predicted fusions. These visualizations include the orientation of the transcripts, the retained exons in the fusion transcript, and statistics about the number of supporting reads.

Structural variants

The integrated WGTS data for genes located within detected SVs are presented in the “summary” table ordered by the SV score and the difference between investigated patient’s and TCGA expression levels. In addition to the genomic information associated with detected SVs, such as genomic location and CN changes, the table provides details about affected transcripts and potential overlapping gene fusions. The accompanying “expression profiles” plots display the expression levels of the affected genes as measured in the patient’s samples and reference cohort samples.

Copy number altered genes

The "CN altered genes" section overlays mRNA expression data with per-gene somatic CN data and information about SNVs and indels. It begins with an interactive Manhattan-like plot that presents CN values for genes located within gained or lost regions. Additionally, the CN and expression levels for affected genes are displayed in a scatterplot, showing the per-gene difference in mRNA expression between the patient's sample and the reference cohort as a function of corresponding CN values. Following these visualizations, the section includes a "summary" table, ordered by the CN value and the difference between the investigated patient’s and TCGA expression levels, and complemented by information derived from the “Mutated genes” section. The "expression profiles" plots then depict the expression levels of the affected genes, as measured in the patient’s samples and reference cohort samples. These combined visualizations and data tables provide a comprehensive view of the impact of CN alterations on gene expression and their potential clinical significance.

Custom panel of genes-of-interest

The final section of the report presents results derived exclusively from WTS data and focuses on the expression profiles of a customizable list of GOIs provided by the user or three specific sets of genes, (1) “immune-related genes”, (2) “HRD genes” and (3) “known cancer genes”, if the prior is not defined. The mRNA levels of genes involved in these gene sets are presented in the corresponding “summary” tables and “expression overview” box plots.

Results

Clinical applicability evaluation

To evaluate the clinical utility of RNAsum in curating aberrant events observed in WGS, we reviewed 63 cases for which both WGS and WTS data were available (Table S2). Across assessed cases, 152 clinically reportable variants were identified by WGS, of which 141 (92.8%) were confirmed by WTS. Based on corresponding RNAsum reports, WTS confirmed at least half of the variants in 61 cases (96.8%), and all the variants in 54 cases (85.7%). These events covered the spectrum of point mutations, amplifications, deep deletions, inactivating structural variations and gene fusions (Figure 3). Confirmation was based on evidence of increased expression (n=41), loss of expression (n=63), or detection of aberrant splicing or gene fusions (n=37), consistent with WGS findings. This systematic review of clinical cases highlights RNAsum's capability to identify and validate clinically significant findings from WGS, reinforcing its value in clinical data curation.

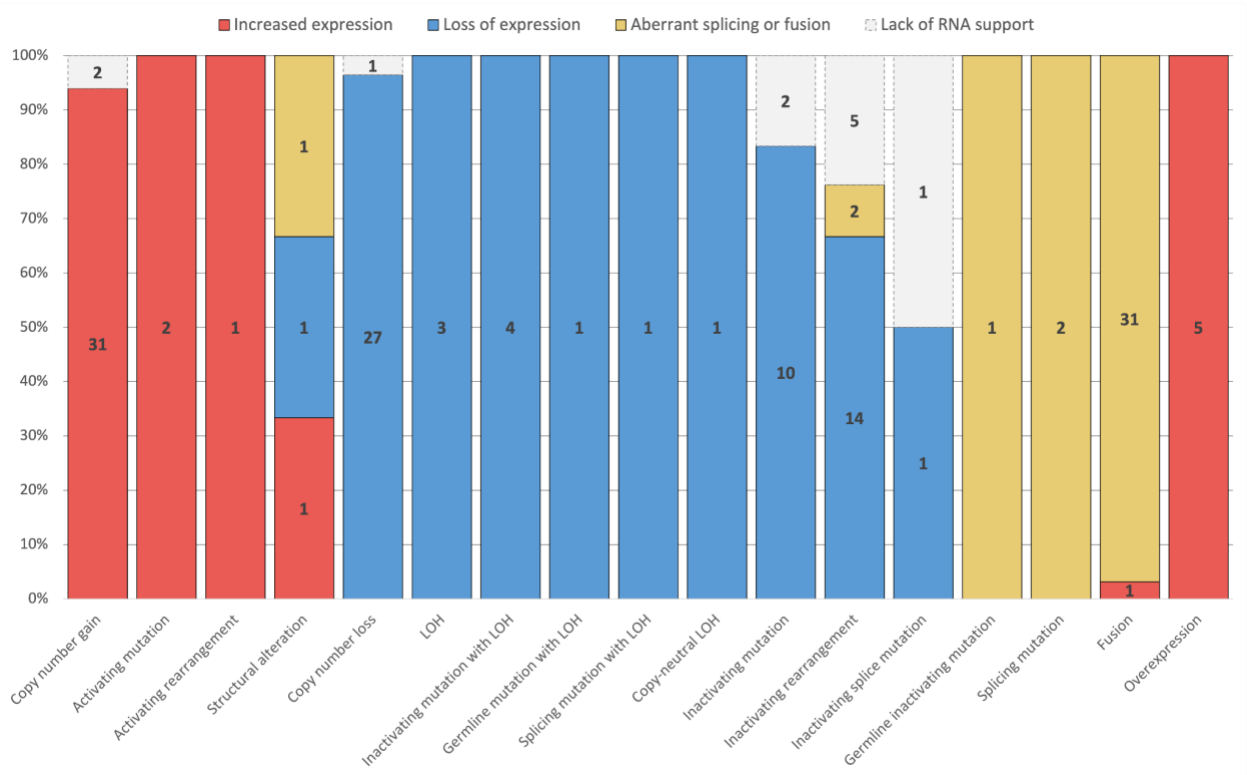


Figure 3. Bar plot illustrating the utility of WGS and WTS integration in validating clinically reportable genomics variants detected by WGS, emphasizing the role of RNAsum in presenting expression changes, splicing events and fusions in the context of these genomic alterations. The bars represent reported genomic variants detected by WGS across the 63 cases, and the proportions within individual bars indicate WTS evidence types, including increased expression (red), loss of expression (blue), aberrant splicing or fusion events (yellow). Genomic alterations which lacked evidence of RNA support are indicated by a dashed pattern. Numbers within the bars indicate the count of alterations supported by each WTS evidence type. LOH - loss of heterozygosity.

Case Studies

To further demonstrate the utility of RNAsum in the clinical interpretation of somatic genomic alterations, we present two cases where the data curated and summarized by RNAsum played a crucial role in guiding clinically significant decisions.

Case study #1

Whole-genome and whole-transcriptome sequencing was performed on a fresh metastatic triple negative breast cancer (TNBC) biopsy. The tumour has a low tumour mutational burden (1.9 mutations/Mb) and contained somatic features typical of this disease: a *TP53* truncating mutation (TP53 p.R213*) accompanied with CN neutral loss of heterozygosity (LOH), chromosomal rearrangements of *PTEN* and *RB1,* and amplification of several genomic regions including chr8q.

RNAsum was run for this case using the TCGA breast cancer reference cohort. The expression of the *TP53* c.637C>T nonsense mutation was confirmed, with 47% of 217 reads carrying the mutation, and the expression level estimated in the 41st percentile, relative to the TCGA cohort (Figure 4 - A). These observations predict expression of a truncated p53 and the mutation was interpreted as a LOF mutation (Malcikova et al. 2010).

The WGS-detected rearrangement in *PTEN* involved a reciprocal translocation with *PDE1C*. RNAsum only detected expression from one of these events (PTEN exon 2 - PDE1C exon 17 fusion). Furthermore, RNAsum detected PTEN::PDE1C fusion transcripts that encode an frameshift at Val54, early in the PTEN phosphatase domain. Coupled with low expression 2nd percentile, these findings strongly supported loss of function of both PTEN loci. *(*Figure 4 - B*)*

The WGS-detected rearrangement in *RB1* involved a tandem duplication of exon 15. RNAsum confirmed expression of transcripts carrying the duplication, resulting in frameshift and predicting truncation after Phe473, N-terminal to the essential RB1 pocket domain. RNAsum also reported

low *RB1* expression (Figure 4 - C) in the 2nd percentile when compared to the TCGA Breast Cancer cohort, adding confidence to the DNA based interpretation. WGS identified amplification of the entire long arm of chromosome 8 (CN range between 10.8 and 14.5). This large amplicon contained 32 cancer-related genes (Figure 4 - D), 3 were known oncogenes based on the OncoKB database: *LYN* (Figure 4 - E), *RUNX1T1* (Figure 4 - F) and *MYC* (Figure 4 - G). RNAsum only identified MYC as highly expressed (92nd percentile relative to the TCGA Breast Cancer cohort). *MYC* amplification and expression was reported as a biomarker for potential inclusion in a clinical trial (Australian and New Zealand Clinical Trial Registry ACTRN12620001146987).

In addition, amplification of the chr3q22 region was detected, encompassing *PIK3CB* (CN of 34.6). *PIK3CB* encodes p110b, an activating subunit of PI3 kinase (PI3K). In contrast to its paralog *PIK3CA*, mutation of *PIK3CB* is rarely observed in TNBC (ref: cohort papers). RNAsum reported high *PIK3CB* expression (100th percentile when compared to the TCGA Breast cancer cohort Figure 4 - H. The amplification was reported as a variant of unknown clinical significance due to the ability of p110b to drive PI3K signalling (Costa et al. 2015).

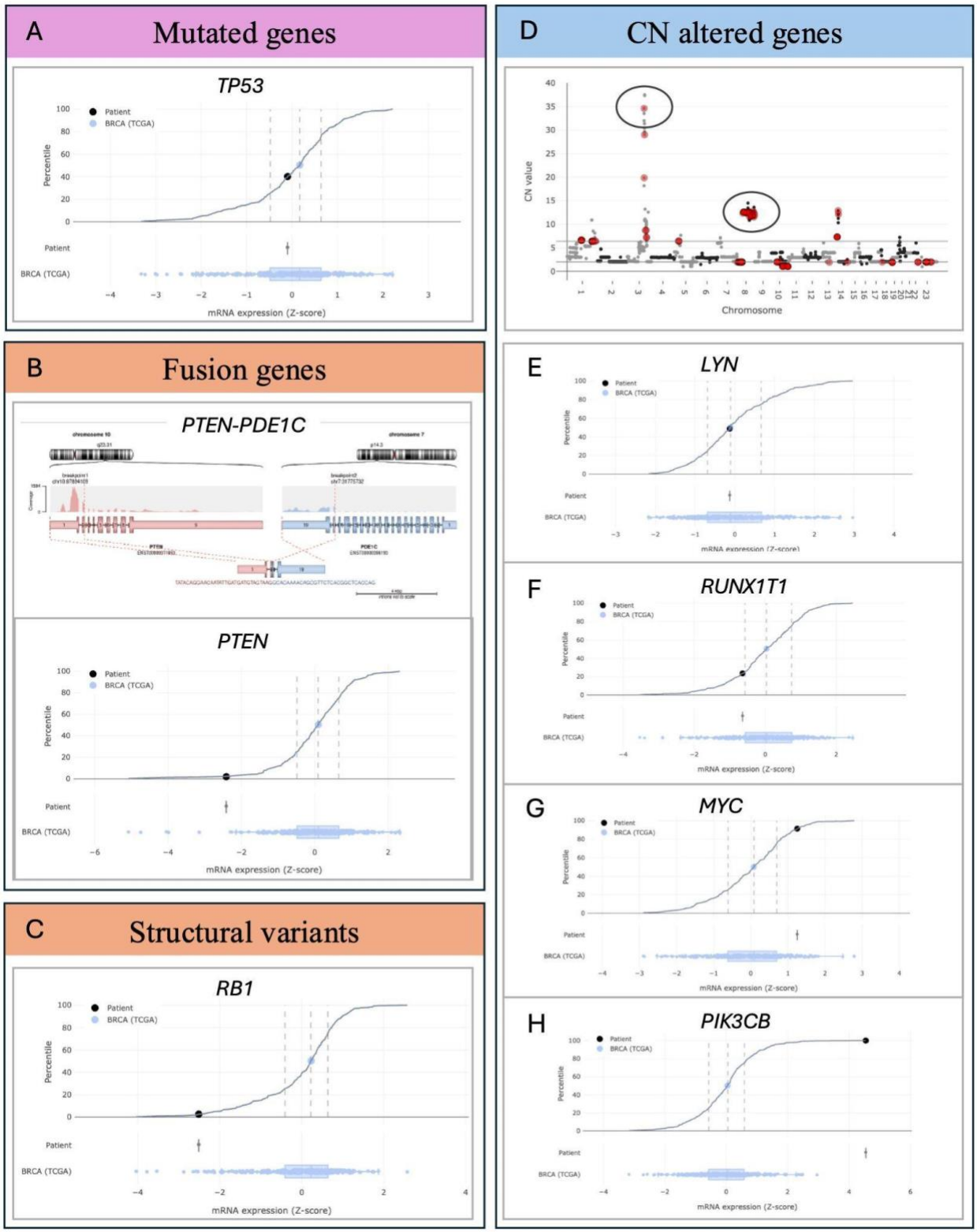
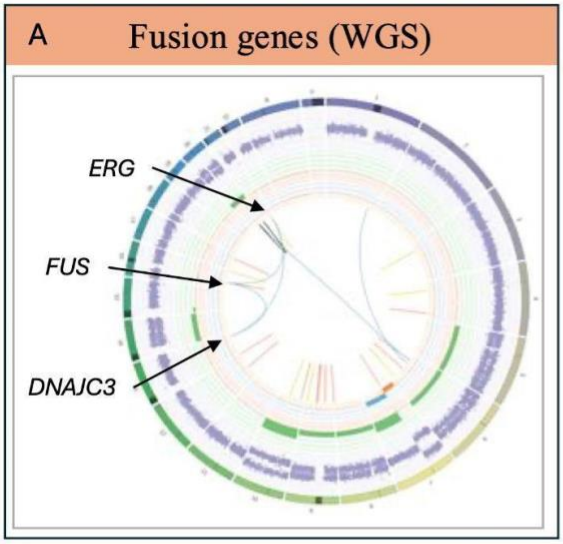

A
Mutated genes
TP53
Patient
BRCA (TCGA)
Percentile
mRNA expression (Z-score)
B
Fusion genes
PTEN-PDE1C
PTEN
Patient
BRCA (TCGA)
Percentile
mRNA expression (Z-score)
C
Structural variants
RB1
Patient
BRCA (TCGA)
Percentile
mRNA expression (Z-score)
D
CN altered genes
CN value
Chromosome
E
LYN
Patient
BRCA (TCGA)
Percentile
mRNA expression (Z-score)
F
RUNX1T1
Patient
BRCA (TCGA)
Percentile
mRNA expression (Z-score)
G
MYC
Patient
BRCA (TCGA)
Percentile
mRNA expression (Z-score)
H
PIK3CB
Patient
BRCA (TCGA)
Percentile
mRNA expression (Z-score)

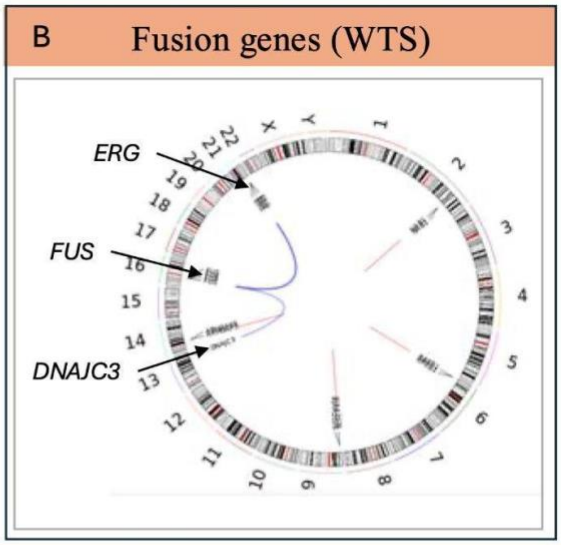
Figure 4. RNAsum visualizations supporting the key genomic and transcriptomic findings from the analysis of a metastatic triple-negative breast cancer (TNBC) case. Panel A shows the expression of the TP53 c.637C>T nonsense mutation in the tumour, with 47% of RNA reads carrying the mutation. The expression level is in the 41st percentile relative to the TCGA breast cancer cohort, supporting the interpretation of a loss-of-function mutation. Panel B highlights the PTEN::PDE1C fusion transcript detected by WTS, resulting in a frameshift at Val54 within the PTEN phosphatase domain. The fusion is coupled with low PTEN expression (2nd percentile), consistent with a loss of PTEN function. Panel C depicts low expression of RB1 (2nd percentile compared to TCGA) supporting a loss of its function following tandem duplication involving exon 15, as confirmed by RNAsum. Panel D displays the amplification of chromosome 8q, including 32 cancer-related genes, with MYC identified as highly expressed (92nd percentile). Panels E-G show expression levels of other genes within the 8q amplification, including LYN and RUNX1T1, which were not highly expressed, and MYC, which was highly expressed. Panel H illustrates the amplification of PIK3CB in the chr3q22 region, with expression at the 100th percentile relative to the TCGA cohort. PIK3CB overexpression is noted as a potential driver of PI3K signalling but remains classified as a variant of uncertain significance.

Case study #2

The second case study involved a woman in her 30's who presented with a tumour of the left lateral skull. Histopathology showed spindle small round blue cell morphology; tumour cells were positive for CD99, S100 and NKX2-2 in keeping with Ewing's sarcoma. However, no *EWSR1* or *FUS* rearrangements were detected by break-apart FISH, calling the diagnosis into question.

WGS demonstrated 2 fusion genes: *FUS::DNAJC3* and *DNAJC3::ERG* (Figure 5 - A). This was intriguing considering the tumour phenotype as typical chromosomal translocations observed in Ewing's sarcoma fuse a member of the FET family of proteins (encoded by *FUS, EWSR1* and *TAF15*), with different members of the ETS (E26-specific) family of transcription factors, such as *FLI1* and *ERG* (PMID: 21872822). *DNAJC3* is a gene on chromosome 13 with no known role in cancer. Manual inspection raised the possibility of a *cis*-arranged three-way fusion where a *DNAJC3* intronic fragment was inserted between *FUS* and *ERG*. However, this could not be definitively determined due to the limitations of short read WGS and presented a diagnostic dilemma. RNAsum confirmed the active expression of a FUS::ERG fusion transcript, which was predicted to encode the FUS N-terminal transcriptional activation domain and the ERG C-terminal DNA-binding ETS domain (Figure 5 - B and C, Figure S4). This was in line with observations in atypical Ewing's sarcoma (PMID: 21872822). The combination of WGS & WTS in this case highlights the utility of RNAsum for interpretation of complex genomic events.


A
Fusion genes (WGS)
ERG
FUS
DNAJC3


B
Fusion genes (WTS)
ERG
FUS
DNAJC3

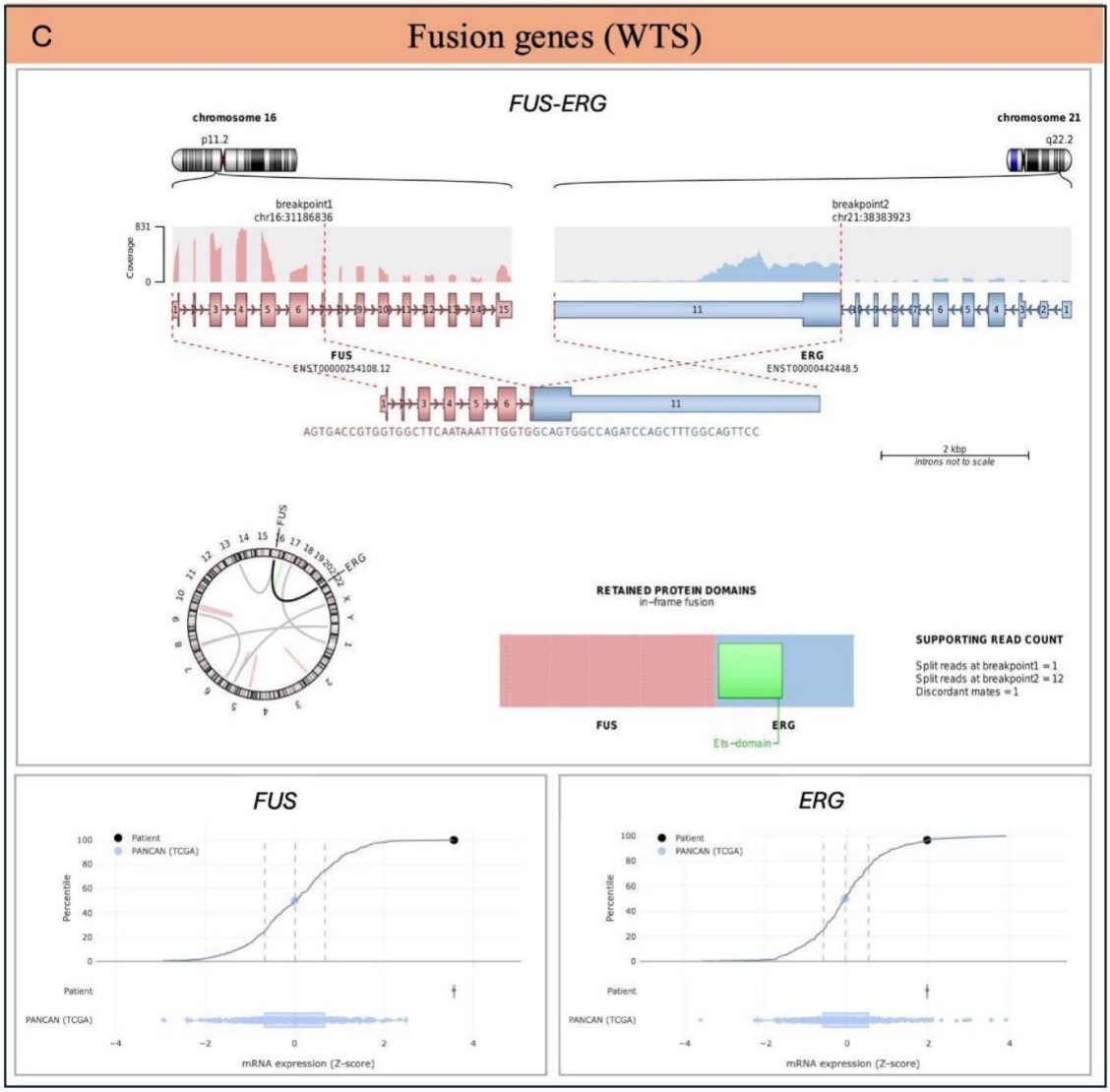

C
Fusion genes (WTS)
FUS-ERG
chromosome 16
p11.2
chromosome 21
q22.2
breakpoint1
chr16:31186836
breakpoint2
chr21:38383923
Coverage
FUS
ENST00000254108.12
ERG
ENST00000442448.5
AGTGACCGTGGTGGCTTCAATAAATTTGGTGGCAGTGGCCAGATCCAGCTTTGGCAGTTCC
2 kbp
introns not to scale
RETAINED PROTEIN DOMAINS
in-frame fusion
FUS
ERG
Ets-domain
SUPPORTING READ COUNT
Split reads at breakpoint1 = 1
Split reads at breakpoint2 = 12
Discordant mates = 1
FUS
Patient
PANCAN (TCGA)
Percentile
mRNA expression (Z-score)
ERG
Patient
PANCAN (TCGA)
Percentile
mRNA expression (Z-score)

Figure 5: RNAsum results providing the genomic and transcriptomic evidence for fusion genes identified in a tumour initially suspected as Ewing's sarcoma but presenting diagnostic challenges. Panels A and B depict two fusion events identified by WGS and WTS: FUS::DNAJC3 and DNAJC3::ERG. These are visualized on circular plots, with FUS and ERG highlighted as known members of the FET and ETS families, respectively, often involved in Ewing's sarcoma translocations. DNAJC3, with no known cancer associations, is shown as the intermediate partner in a potential three-way fusion. Panel C highlights the expressed FUS::ERG fusion transcript detected by WTS, with detailed genomic breakpoints and predicted transcript structure. This fusion retains the FUS N-terminal transcriptional activation domain and the ERG C-terminal DNA-binding ETS domain, consistent with oncogenic activity observed in atypical Ewing's sarcoma cases. The plots presented at the bottom of panel C show expression profiles for the FUS and ERG genes in the patient tumour relative to the TCGA cohort, confirming active transcription and supporting the functional relevance of the FUS::ERG fusion in the tumour.

Discussion

WGTS technologies have revolutionized precision oncology, providing cost-effective and robust data to support personalized cancer treatment (Kumar-Sinha and Chinnaiyan 2018). With the continuous development of robust bioinformatics tools, WGTS has become a standard protocol for investigating cancer genomes and delivering timely, accurate diagnoses (Deyell et al. 2024), (Tessier-Cloutier et al. 2022). The clinical relevance of comprehensive WGTS analysis has also been demonstrated in rare cancers (Horák et al. 2021). Further, use of RNA expression data with genomic profiling has proven particularly informative for identifying precision cancer treatment options, resulting in nearly half of treated patients experiencing positive clinical benefit when guided by combined WGTS analysis (Pleasance et al. 2022). However, integrating and summarizing data from genomics and transcriptomics remains a complex and challenging task, hindered by the tumour-only nature of transcriptome sequencing, the need to harmonize outputs from different analyses, and the lack of cancer-specific reference data and effective integration tools.

To address these challenges, we have developed RNAsum, a robust open-source bioinformatics reporting tool designed to leverage RNA-seq data from cancer patients to either verify and complement whole-genome profiling results or provide summary statistics from WTS data. Our approach incorporates transcriptome profiling into a precision oncology framework by utilizing in-house (or user provided) and external (TCGA) RNA-seq data, thereby increasing confidence in the clinical utility of findings. RNAsum integrates data at multiple levels, combining mRNA profiles from the patient, internal, and external reference cohorts, overlaying expression data with genomic changes, and annotating findings with clinically relevant information from public

knowledge databases. This comprehensive integration process enables the identification and interpretation of cancer-relevant molecular aberrations, facilitating precision oncology by streamlining genomics research, accelerating discoveries, and translating them into actionable clinical insights.

The RNAsum tool generates an interactive HTML-based report and data outputs that summarize clinically relevant information, including findings summary, mutated genes, fusion genes, structural variants, CN altered genes, immune markers, homologous recombination deficiency (HRD), and cancer genes. The report provides a comprehensive summary table and plots showcasing mRNA expression levels, prioritized findings based on whole-genome sequencing results and public databases, and visual data summaries for quality control and interpretation. Instead of solely relying on external knowledge based components to extract biological evidence and therapeutic implication for variants (Reisle et al. 2022), our method complements this approach by focusing on the development of a comprehensive reporting system that enhances confidence in the results from whole-genome analysis tools and provides translational information using whole-transcriptome expression analysis and integration, ultimately improving patient management and health outcomes in precision oncology settings.

In molecular pathology, various techniques are employed to identify or confirm genomic changes such as rearrangements. Among these, fluorescence in situ hybridization (FISH) and immunohistochemistry (IHC) are widely utilized due to their rapid availability and cost-effectiveness. However, the combined analysis of DNA and RNA-seq data has proven to be a more sensitive and clinically informative screening approach, particularly for detecting intricate rearrangements involving multiple partner genes, such as GATM::RAF1, GOLGA4::RAF1, PDZRN3::RAF1, HERPUD1::RAF1, and TRIM33::RAF1 (Prall et al. 2020). The results from WTS are crucial for detecting in-frame chimeric transcripts, as demonstrated in this study, where HERPUD1 exons 1-8 (full ORF) were fused to RAF1 exons 8-17. This fusion event involves the skipping of HERPUD1 exon 8 and including its stop codon, predicting a potentially functional chimeric protein. Consequently, this underscores the necessity for a comprehensive tool like RNAsum, which can aggregate the results from molecular testing and support the detection of such complex fusions and reduce false-positive findings arising from rearrangements that reflect complex chromosomal instability.

In another study by Eden C. Andrew et al. (Andrew et al. 2024), refined diagnosis was achieved for 32% of Adolescents and Young Adults (AYA) with sarcoma based on the identification of gene fusions or fusion partners following WGTS screening. Notably, more than half of the AYA patients with sarcoma who underwent comprehensive molecular profiling exhibited a new potentially actionable variant. These variants encompassed amplifications, homozygous deletions, in-frame deletions, and fusions, which could be orthogonally confirmed through gene expression changes. Such findings as also supported through case study 1 underscore the importance of integrating WGTS data with genomic analysis for refining diagnoses and identifying potential therapeutic targets in cancer patients.

Furthermore, in clinical cases where evaluating the breakpoints and expression levels of fused transcripts is critical, RNAsum can play a pivotal role in providing evidence for functional fusion proteins and assessing the expression of diagnostic fusion genes (Abdelmogod et al. 2023). In a case reported by Holly E. Barker et al., a novel gene fusion (FBN1::WWTR1) was confirmed based on supported reads and evidence of retained exons. This study highlighted the importance of evaluating gene expression data to validate and properly characterize gene fusions, as genomic data alone may not provide a complete picture of the functional consequences. This underscores the potential clinical significance of such novel fusions identified through comprehensive analysis of transcriptomic and genomic data. These fusion events can lead to hyperactivation of transcriptional programs and increased cellular proliferation, which may be potentially inhibited by targeted therapeutic interventions. This is also demonstrated through the results we observed for case study 2 where the expression of a functional fusion protein provided critical genomic information for the diagnosis.

RNAsum offers great flexibility in data processing and parameter selection, but it requires users to provide specific WGTS analysis inputs in a compatible format for integration into the reporting. This presents opportunities for developing modules that can support outputs from various bioinformatics workflows and tools. Although TCGA provides a rich resource of cancer-specific datasets for gene expression comparison, it has been shown to exhibit unwanted variation arising from library size, tumour purity, and batch effects (Molania et al. 2023). While RNAsum implements an internal approach to mitigate unwanted variation (as discussed in the Methods section), the results should be interpreted within the specific biological context and corroborated through orthogonal evidence or supported by the literature.

To ensure efficient data and dependency management, the RNAsum codebase has been developed as an R-package, which can be easily accessed and downloaded from GitHub. It offers flexibility in execution, as it can be run on either an HPC or cloud infrastructure, and is also available as a docker container. The TCGA reference data utilized by RNAsum is conveniently managed through an R data package (Colaprico et al. 2016), also accessible via GitHub. Moreover, RNAsum is a cost-effective solution that can be seamlessly adapted to run on outputs from WTS analysis exclusively or from both WGS and WTS analysis.

Conclusion

We have developed RNAsum, a novel bioinformatics method that deploys a comprehensive approach to data preprocessing, quantification, and integration of multi-omics data sets. RNAsum represents a significant advancement in RNA-seq data analysis, seamlessly integrating WGS and WTS data to provide critical insights for cancer patient genomes. Additionally, we demonstrate that by successfully achieving its aims of delivering clinically relevant information and prioritising results for therapeutic intervention for clinical cases, RNAsum enhances prognosis, patient management, and health outcomes in precision oncology settings. The tool's potential to improve data interpretation and accelerate discoveries in genomics research is substantial, paving the way for streamlining research efforts and driving personalised medicine initiatives. To enable its wider applicability and usage, RNAsum is available as an R-package that can be downloaded from the GitHub https://github.com/umccr/RNAsum/tags.

Author Contributions

S.G and O.H. conceived and supervised the study. S.K. and J.M. extracted and analysed data and wrote the manuscript. S.K., J.M and S.G. analysed primary data. L.T. reviewed genomic and transcriptomic reports available for case studies. L.V. provided clinical interpretation of genomic and transcriptomic alterations in selected case studies. J.H.A.V., O.H., S.J.L. and S.G. provided manuscript review. All authors have read and agreed to the published version of the manuscript.

Funding

We acknowledge the funding support from the National Health and Medical Research Council of Australia (Project Grant APP1160855). Whole genome and transcriptome sequencing were

available through enrolment in the VCCC PRECISION program funded by the Victorian Comprehensive Cancer Centre (VCCC). The VCCC alliance (https://vcccalliance.org.au/about-us/victorian-comprehensive-cancer-centre/) is a joint venture of 10 academic, clinical, and research institutions and is funded by the Victorian Government in Australia.

Institutional Review Board

The VCCC PRECISION program received ethics approval by the Peter MacCallum Cancer Centre Human Research Ethics Committee (HREC/48455/PMCC-2018).

Informed Consent

Informed consent was obtained from all subjects involved in the study.

Data Availability

The data presented in this study are available upon request from the corresponding author and after approval from the VCCC PRECISION program investigators. The data are not publicly available due to ethics limitations.

Code Availability

Codes used in this study were recorded on GitHub at https://github.com/umccr/RNAsum.

Acknowledgements

We would like to acknowledge Kym Pham Stewart and the University of Melbourne Centre for Cancer Research Clinical Genomics Platform for performing whole-genome and whole-transcriptome sequencing analyses.

Conflicts of Interest

The authors declare no conflicts of interest.

References

Abdelmogod, Arwa, Lia Papadopoulos, Stephen Riordan, Melvin Wong, Martin Weltman, Ratana Lim, Christopher McEvoy, et al. 2023. “A Matched Molecular and Clinical Analysis of the Epithelioid Haemangioendothelioma Cohort in the Stafford Fox Rare Cancer Program and Contextual Literature Review.” *Cancers* 15 (17): 4378.

Alioto, T. S., I. Buchhalter, S. Derdak, B. Hutter, M. D. Eldridge, E. Hovig, and Gut I. 2015. “A Comprehensive Assessment of Somatic Mutation Detection in Cancer Using Whole-Genome Sequencing.” *Nature Communications* 6 (1): 1–13.

Allgäuer, Michael, Jan Budczies, Petros Christopoulos, Volker Endris, Amelie Lier, Eugen Rempel, Anna-Lena Volckmar, et al. 2018. “Implementing Tumor Mutational Burden (TMB) Analysis in Routine Diagnostics-a Primer for Molecular Pathologists and Clinicians.” *Translational Lung Cancer Research* 7 (6): 703–15.

Andrew, Eden C., Jeremy Lewin, Jayesh Desai, Lisa Orme, Anne Hamilton, Susie Bae, Wenying Zhu, et al. 2024. “Clinical Impact of Comprehensive Molecular Profiling in Adolescents and Young Adults with Sarcoma.” *Journal of Personalized Medicine* 14 (2): 128.

Bailey, Peter, D. Chang, K. Nones, A. Johns, A. Patch, M. Gingras, David K. Miller, et al. 2016. “Genomic Analyses Identify Molecular Subtypes of Pancreatic Cancer.” *Nature* 531 (February):47–52.

Bray, Nicolas, Harold Pimentel, Páll Melsted, and Lior Pachter. 2015. “Near-Optimal RNA-Seq Quantification.” *arXiv [q-bio.QM]*. arXiv. https://www.nature.com/articles/nbt.3519.

Chen, Xiaoyu, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. 2016. “Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications.” *Bioinformatics (Oxford, England)* 32 (8): 1220–22.

Colaprico, A., T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, and Noushmehr. 2016. “TCGAbiolinks: An R/Bioconductor Package for Integrative Analysis of TCGA Data.” *Nucleic Acids Research* 44 (8): e71–e71.

Costa, Carlotta, Hiromichi Ebi, Miriam Martini, Sean A. Beausoleil, Anthony C. Faber, Charles T. Jakubik, Alan Huang, et al. 2015. “Measurement of PIP3 Levels Reveals an Unexpected Role for p110β in Early Adaptive Responses to p110α-Specific Inhibitors in Luminal Breast Cancer.” *Cancer Cell* 27 (1): 97–108.

Cuppen, Edwin, Olivier Elemento, Richard Rosenquist, Svetlana Nikic, Maarten IJzerman, Isabelle Durand Zaleski, Geert Frederix, et al. 2022. “Implementation of Whole-Genome and Transcriptome Sequencing into Clinical Cancer Care.” *JCO Precision Oncology* 6 (6). https://doi.org/10.1200/po.22.00245.

Deyell, Rebecca J., Yaoqing Shen, Emma Titmuss, Katherine Dixon, Laura M. Williamson, Erin Pleasance, Jessica M. T. Nelson, et al. 2024. “Whole Genome and Transcriptome Integrated Analyses Guide Clinical Care of Pediatric Poor Prognosis Cancers.” *Nature Communications* 15 (1): 4165.

Fujimoto, Akihiro, Masashi Fujita, Takanori Hasegawa, Jing Hao Wong, Kazuhiro Maejima, Aya Oku-Sasaki, Kaoru Nakano, et al. 2020. “Comprehensive Analysis of Indels in Whole-Genome Microsatellite Regions and Microsatellite Instability across 21 Cancer Types.” *Genome Research* 30 (3): 334–46.

Gong, Yiming, Liancheng Fan, Xiaochen Fei, Yinjie Zhu, Xinxing Du, Yuman He, Jiahua Pan, Baijun Dong, and Wei Xue. 2021. “Targeted Next-Generation Sequencing Reveals Heterogenous Genomic Features in Viscerally Metastatic Prostate Cancer.” *Journal of Urology* 206 (2): 279–88.

Horák, P., C. Heining, S. Kreutzfeldt, B. Hutter, A. Mock, Jennifer Hullein, Martina Frohlich, et al.

2021. “Comprehensive Genomic and Transcriptomic Analysis for Guiding Therapeutic Decisions in Patients with Rare Cancers.” *Cancer Discovery*, June. https://doi.org/10.1158/2159-8290.CD-21-0126.

Jobanputra, Vaidehi, Kazimierz O. Wrzeszczynski, Reinhard Buttner, Carlos Caldas, Edwin Cuppen, Sean Grimmond, Torsten Haferlach, Charles Mullighan, Anna Schuh, and Olivier Elemento. 2022. “Clinical Interpretation of Whole-Genome and Whole-Transcriptome Sequencing for Precision Oncology.” *Seminars in Cancer Biology* 84 (September):23–31.

Kinnersley, Ben, Amit Sud, Andrew Everall, Alex J. Cornish, Daniel Chubb, Richard Culliford, Andreas J. Gruber, et al. 2024. “Analysis of 10,478 Cancer Genomes Identifies Candidate Driver Genes and Opportunities for Precision Oncology.” *Nature Genetics*, June, 1–10.

Kumar-Sinha, Chandan, and Arul M. Chinnaiyan. 2018. “Precision Oncology in the Age of Integrative Genomics.” *Nature Biotechnology* 36 (1): 46–60.

Luca, Xavier M. de, Felicity Newell, Stephen H. Kazakoff, Gunter Hartel, Amy E. McCart Reed, Oliver Holmes, Qinying Xu, et al. 2020. “Using Whole-Genome Sequencing Data to Derive the Homologous Recombination Deficiency Scores.” *NPJ Breast Cancer* 6 (1): 33.

Malcikova, Jitka, Boris Tichy, Jiri Damborsky, Jitka Kabathova, Martin Trbusek, Jiri Mayer, and Sarka Pospisilova. 2010. “Analysis of the DNA-Binding Activity of p53 Mutants Using Functional Protein Microarrays and Its Relationship to Transcriptional Activation.” *Biological Chemistry* 391 (2-3): 197–205.

Molania, Ramyar, Momeneh Foroutan, Johann A. Gagnon-Bartsch, Luke C. Gandolfo, Aryan Jain, Abhishek Sinha, Gavriel Olshansky, Alexander Dobrovic, Anthony T. Papenfuss, and Terence P. Speed. 2023. “Removing Unwanted Variation from Large-Scale RNA Sequencing Data with PRPS.” *Nature Biotechnology* 41 (1): 82–95.

Nagy, Á., G. Munkácsy, and Balázs Győrffy. 2020. “Pancancer Survival Analysis of Cancer Hallmark Genes.” *Scientific Reports* 11 (November). https://doi.org/10.1038/s41598-021-84787-5.

Nakken, Sigve, Ghislain Fournous, Daniel Vodák, Lars Birger Aasheim, Ola Myklebost, and Eivind Hovig. 2018. “Personal Cancer Genome Reporter: Variant Interpretation Report for Precision Oncology.” *Bioinformatics (Oxford, England)* 34 (10): 1778–80.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14 (4): 417–19.

Pleasance, E., A. Bohm, L. M. Williamson, J. M. T. Nelson, Y. Shen, M. Bonakdar, E. Titmuss, et al. 2022. “Whole-Genome and Transcriptome Analysis Enhances Precision Cancer Treatment Options.” *Annals of Oncology* 33 (9): 939–49.

Prall, Owen W. J., Violeta Nastevski, Huiling Xu, Christopher R. E. McEvoy, Joep H. A. Vissers, David J. Byrne, Elena Takano, et al. 2020. “RAF1 Rearrangements Are Common in Pancreatic Acinar Cell Carcinomas.” *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc* 33 (9): 1811–21.

Reisle, Caralyn, Laura M. Williamson, Erin Pleasance, Anna Davies, Brayden Pellegrini, Dustin W. Bleile, Karen L. Mungall, et al. 2022. “A Platform for Oncogenomic Reporting and Interpretation.” *Nature Communications* 13 (1): 756.

Rosenquist, Richard, Edwin Cuppen, Reinhard Buettner, Carlos Caldas, Helene Dreau, Olivier Elemento, Geert Frederix, et al. 2022. “Clinical Utility of Whole-Genome Sequencing in Precision Oncology.” *Seminars in Cancer Biology* 84 (September):32–39.

Roychowdhury, Sameek, and Arul M. Chinnaiyan. 2016. “Translating Cancer Genomes and Transcriptomes for Precision Oncology.” *CA: A Cancer Journal for Clinicians* 66 (1): 75–88.

Suehnholz, Sarah P., Moriah H. Nissan, Hongxin Zhang, Ritika Kundra, Subhiksha Nandakumar, Calvin Lu, Stephanie Carrero, et al. 2024. “Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer.” *Cancer Discovery* 14 (1): 49–65.

Tessier-Cloutier, Basile, Jasleen K. Grewal, Martin R. Jones, Erin Pleasance, Yaoqing Shen, Ellen Cai, Chris Dunham, et al. 2022. “The Impact of Whole Genome and Transcriptome Analysis (WGTA) on Predictive Biomarker Discovery and Diagnostic Accuracy of Advanced Malignancies.” *The Journal of Pathology. Clinical Research* 8 (4): 395–407.

Uhrig, Sebastian, Julia Ellermann, Tatjana Walther, Pauline Burkhardt, Martina Fröhlich, Barbara Hutter, Umut H. Toprak, et al. 2021. “Accurate and Efficient Detection of Gene Fusions from RNA Sequencing Data.” *Genome Research* 31 (3): 448–60.

Zhang, Yuanjiao, Jinjun Qian, Chunyan Gu, and Ye Yang. 2021. “Alternative Splicing and Cancer: A Systematic Review.” *Signal Transduction and Targeted Therapy* 6 (1): 78.

Zheng, Siyuan, Andrew D. Cherniack, Ninad Dewal, Richard A. Moffitt, Ludmila Danilova, Bradley A. Murray, Antonio M. Lerario, et al. 2016. “Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma.” *Cancer Cell* 29 (5): 723–36.