Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

paRDal: Bioinformatics Framework for a Parallel RNA and DNA NGS Analysis in Hereditary Cancer Diagnostics

Kateřina Matějková

katerina.matejkova@lf1.cuni.cz

Charles University

Petr Nehasil

Charles University

Petra Kleiblová

Charles University

Taťána Ptáčková

Charles University

Marta Černá

Charles University

Pavel Just

Charles University

Ondřej Vít

Charles University

Petra Zemánková

Charles University

Mária Giertlová

Pavol Jozef Šafárik University

Markéta Janatová

Charles University

Jana Soukupová

Charles University

Ina Skaara Brorson

Charles University

Zdeněk Kleibl

Charles University

Article

Keywords: targeted DNA/RNA NGS, parallel analysis, multiomics, ACMG criteria, variant classification, splicing analysis

Posted Date: July 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10043502/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Germline next-generation sequencing (NGS) is the gold standard for hereditary cancer diagnostics. However, clinical utility is limited by variants of uncertain significance (VUS), which may disrupt mRNA processing; consequences invisible to DNA NGS alone but identifiable by parallel RNA NGS. Here, we present *paRDaI* (parallel RNA/DNA analytical workflow), a containerized, modular framework that integrates matched germline DNA and RNA NGS data into a single diagnostic pipeline. *paRDaI* implements allele-aware two-pass alignment, automated splicing and expression outlier detection, allele-specific expression analysis, and comprehensive variant annotation. A bidirectionally reactive filtering engine enables both DNA-centric (variant-to-function) and RNA-centric (function-to-variant) workflows. An R/Shiny application consolidates multilayered evidence per patient and across cohorts, linking *in silico* predictions with observed transcriptomic consequences without requiring specialized bioinformatic expertise. Validated on 204 individuals undergoing hereditary cancer genetic testing, *paRDaI* facilitates the application of ACMG/AMP criteria [PVS1(RNA) and PP3], improving diagnostic yield and VUS resolution in clinical germline genomics.

Background

Next-generation sequencing (NGS) of germline DNA has become the gold standard for the analysis of hereditary cancer predisposition, the currently dominant application of Mendelian disease diagnostics in human clinical genetics, increasing diagnostic yield while reducing turnaround time and costs[1–5]. Beyond the identification of causal germline pathogenic variants (gPV) in cancer predisposition genes, which enables the application of tailored cancer prevention strategies for high-risk carriers, NGS reveals many variants of uncertain significance (VUS). These genetic changes, with unknown biological effects and clinical implications, hamper the diagnostic utility of NGS and may lead to inadequate disease surveillance[6]. VUS account for approximately 40% of NGS-identified variants[7,8] and are found in up to 30% of patients[7,9]. Their resolution has accelerated in recent years, largely driven by machine-learning approaches leveraging large public variant datasets to provide more quantitative assessment[10].

RNA NGS offers a complementary layer of evidence for resolving VUS, particularly those affecting mRNA processing. By capturing transcript-level consequences, RNA NGS can reveal the impact of variants on pre-mRNA splicing, nonsense-mediated decay (NMD), and gene expression, none of which are fully assessable by DNA NGS[11–16]. About 13% of all gPVs disrupt canonical splice sites (± 1/2 positions of donor or acceptor sites), leading to aberrant splicing. However, non-canonical splicing variants, such as those at exonic and especially deep-intronic positions, are likely substantially underreported[11,12,17,18], as they often evade detection and prioritization in panel-based or whole-exome sequencing (WES) approaches. Incorporating RNA NGS into clinical workflows has shown considerable potential for reclassifying canonical and non-canonical splice variants alike, yielding mechanistic evidence for pathogenicity that supports more accurate diagnosis and clinical decision-making[8,19]. Furthermore, RNA NGS can identify aberrant allele-specific expression and gene expression outliers.

Several studies have applied RNA NGS to improve variant interpretation in clinical settings, particularly for evaluating aberrant splicing, expression outliers, and allele-specific expression, as well as for variant calling and sample authentication[18–28]. Importantly, the American College of Medical Genetics and Genomics (ACMG) and Association for Molecular Pathology (AMP) have acknowledged the role of RNA NGS and splicing prediction tools as evidence for variant reclassification. Under the ACMG/AMP framework, RNA NGS can support PVS1(RNA) criterion, yet its implementation remains challenging, partly due to the lack of standardized frameworks and practical guidance[29,30].

RNA-NGS-based workflows typically rely on a combination of tools for splice junction or expression outlier detection, and allele-specific expression analysis[19,20,31]. However, currently there is no standardized or widely adopted framework that systematically integrates DNA NGS with RNA NGS into a variant interpretation workflow, particularly in the context of hereditary cancer diagnostics. While some frameworks address isolated aspects, none encompass the full optimized workflow: from raw data processing and read mapping through variant calling to comprehensive variant annotation and functional RNA analysis. This fragmentation drives variability in analytical approaches and inconsistent application of RNA-based evidence accross clinical genomics laboratories. Additionally, the implementation of functional evidence in clinical settings is frequently constrained by a critical technical bottleneck: fragmented workflows generate complex, multilayered data that remain inaccessible to geneticists and clinicians lacking specialized bioinformatic expertise.

Here we present *paRDal* (parallel RNA/DNA analytical workflow), a containerized (Docker) modular framework designed for the systematic integration of matched DNA and RNA NGS data, accompanied by an R/Shiny application, that empowers clinicians and researchers to efficiently examine patient genomic data.

Methods

Dataset and NGS

Paired DNA and RNA samples were collected as pairs from 204 individuals of Czech and Slovak descent who underwent germline genetic testing for hereditary cancer predisposition at the General University Hospital in Prague (GUHP). The study was approved by the Ethics Committee of the GUHP (ethics approval number 22/22). All participants provided informed consent, and the study was conducted in accordance with the Declaration of Helsinki. DNA and RNA were isolated from peripheral blood leukocytes as described previously[28]. DNA and RNA panel NGS libraries were prepared using the custom CZECANCA panel (CZEch CAncer paNel for Clinical Application) according to[28,32], and sequenced using Illumina paired-end sequencing on NextSeq 2000. Fifty matched genomes from 1000 Genomes[33] and transcriptomes from Geuvadis project[34] served as external controls.

General workflow

Raw NGS data were processed using a fully containerized modular workflow (Fig. 1) implemented inside Docker, ensuring complete reproducibility of software versions, dependencies, and reference files. All steps, tools, folder structures, and scripts used in this workflow are defined in our GitHub repository.

All mapping steps, variant calling and annotation steps used the GRCh38 primary assembly (Homo_sapiens.GRCh38.dna.primary.assembly.fa, release 114) and Ensembl GRCh38 GTF annotation file (release 114)[35]. All databases used for annotation, as well as tool and their versions, are listed in Additional File Table S1 and Table S2, respectively.

DNA NGS Data Preprocessing and Variant Calling

Raw DNA NGS paired-end reads were processed according to GATK Best Practices workflow[36]. Raw FASTQ files were aligned to the reference genome using BWA-MEM2 (v2.0)[37]. Sorting, duplicate marking, and read groups assignment were performed with Picard tools (v1.129)[38]. Base Quality Score Recalibration (BQSR) was applied using GATK (v4.3.0.0)[39] BaseRecalibrator, employing sets of known, high-confidence variants (Mills and 1000G gold standard indels; 1000G phase 1 SNPs)[36].

Variant calling was conducted per-sample using GATK (v4.3.0.0) HaplotypeCaller in GVCF mode. Cohort-wide joint genotyping was subsequently performed using GATK (v4.3.0.0) GenotypeGVCFs. Multiallelic variants were decomposed, and indels were left-aligned and normalized using BCFtools (v1.9)[40] and vt (v0.57721)[41]. Standard cohort-level hard filtering was applied to retain high-quality variants (DP ≥ 5, FS < 45, QD > 1, SOR < 4, MQRankSum ≥ -13.5).

Single-nucleotide variants (SNVs) were extracted to support RNA NGS allele-aware alignment using the WASP algorithm and allele-specific expression (ASE) analyses[40]. Annotation was carried out using Ensembl Variant Effect Predictor (VEP)[42] with multiple plugins, as described in Additional File Table S1.

RNA NGS Data Preprocessing and Variant Calling

RNA NGS paired-end reads were aligned using STAR (v2.7.11b) utilizing a rigorous two-pass mapping mode that supports the discovery of novel splice junctions[43]. First pass mapping required an 8-bp overhang for *de novo* junctions, whereas only a 3-bp overhang was needed for annotated junctions. Export to sample-level databases was strictly gated by a 12-bp bilateral overhang; minimum coverage of 3 reads was required for non-canonical junctions, whereas coverage of 1 read sufficed for canonical junctions. High-confidence novel junctions from the whole cohort were subsequently utilized to generate updated genome indices for the second pass. The second pass required a minimum depth of 3 reads across all canonical and non-canonical motifs for the export to the final splice-junctions file. To suppress reference-mapping bias, a critical requirement for accurate ASE analysis, the second alignment pass was executed with the WASP algorithm enabled. During this step, patient-matched variants derived from the DNA variant calling pipeline were supplied as known variant loci.

BAMs were processed using a workflow equivalent to DNA BAMs processing. To make RNA reads compatible with downstream variant calling, spliced reads were processed using SplitNCigarReads (GATK, v4.3.0.0)[36]. Variant calling and VCF processing were conducted in accordance with the DNA framework.

Variant Annotation

Normalized DNA/RNA variants were comprehensively annotated using Ensembl VEP, (offline cache mode, GRCh38, release 114)[42]. A predefined list of clinically significant transcription variants in cancer predisposition genes was used for internal samples. MANE Select transcripts[44] were used for the exomes from the Geuvadis project.

To empower the diagnostic evaluation of multiomics outliers and aberrant events, the annotation framework was augmented with a specialized set of plugins categorized by biological mechanism (Additional file Table S1).

By integrating a multilayered splicing prediction framework, we provided a mechanistic background for aberrant splicing outliers detected by FRASER[45].The SpliceAI and SpliceVault plugins were utilized to predict cryptic donor/acceptor activation, intron retention, and exon skipping events[46,47]. Standard motifs were annotated using the SpliceRegion plugin.

To facilitate interpretation of protein dysfunction in the absence of transcriptomic outliers, deleteriousness-prediction tools were incorporated, including AlphaMissense[48], BayesDel[49], CADD[50], REVEL[51], and SIFT[52] plugins. To further elucidate the variants' impact on protein structure and function, the DOMAINS and ProteinLengthChange plugins were included to assess domain disruption and changes in protein length. MaveDB[53] provides functional experimental variant effects.

Variants were placed in context using global population frequencies and clinical databases, using gnomAD (v4.0) and the ClinVar database (obtained 11/2025), to ensure clinical relevance. Furthermore, the annotation logic prioritizes clinical transcripts, integrating RefSeq and MANE Select identifiers to ensure that HGVSc and HGVSp[54] notations are consistent with established ACMG/AMP variant classification standards.

Gene Expression Quantification and Expression Outlier Detection with OUTRIDER

Gene-level read counts were quantified from the final RNA BAM files using featureCounts from the Subread package (v2.0.2), counting paired reads spanning exons. Reads were not normalized, leaving normalization to the user.

Expression outliers were detected using the OUTRIDER package in R[55]. Genes with low expression were strictly filtered out (FPKM < 10). OUTRIDER's denoising autoencoder was applied to correct for unobserved confounders. Significant expression outliers were defined as events exhibiting an adjusted p-value < 0.05.

Aberrant Splicing Detection and Splicing Outlier Identification using FRASER

Splicing outliers were extracted from RNA NGS BAMs using the FRASER R-package[45], followed by the calculation of the Percent Spliced In (Ψ). Significant aberrant splicing events were strictly defined using an adjusted p-value < 0.05, an absolute $\Delta\Psi$ > 0.1, and a minimum junction coverage of 5 reads. FRASER was used to generate Sashimi plots for the aberrant splice junctions.

Allele-specific Expression (ASE)

ASE profiling was restricted to heterozygous, biallelic SNPs. Allelic counts were quantified directly from the RNA BAM files using GATK ASEReadCounter with a minimum required depth of 10 reads. Statistical assessment of allelic imbalance was performed using the MBASED R-package (1-sample analysis mode, unphased). ASE events were considered significant at an adjusted p-value < 0.05.

Sample Authentication

The sample authentication was performed via SNPRelate R-package[56] using Identity-by-Descent measures in order to detect sample swaps or errors in library preparation. Biallelic SNPs were LD-pruned (MAF > 0.005; missing rate < 0.05), and the KING method was used to confirm autologous pairings. To define the threshold for identifying identical samples, we examined the distribution of kinship coefficients within each pair of technical replicates, and the final cut-off was set to the higher of the two kinship values surrounding the maximum delta if it exceeded 0.4.

Multiomics Data Integration

The final diagnostic table was constructed using the R/*tidyverse*-based framework. Variant annotation was joined with unannotated DNA and RNA VCFs. DNA and RNA sequencing metrics, including depth and variant allele frequency, were aligned side-by-side to allow immediate distinction between technical artifacts and biologically relevant phenomena such as nonsense-mediated decay (NMD), monoallelic expression, or tissue-specific expression (column details in Additional File 3, Table S3).

The Shiny/R application was developed with an AI-assisted coding[57] approach. However, the multilayer integration logic and data architecture were fully defined by the authors. The application underwent rigorous testing and debugging. Results were thoroughly manually checked for biological sanity, using

internal clinical data as positive controls, mimicking real-life manual data curation and evaluation to verify the correctness of filtering and join engines.

Software Implementation and Containerization

The *paRDal* framework is implemented as a fully containerized solution using Docker (20.10.21) to ensure both fast setup and reproducibility across diverse institutional environments. All required bioinformatics tools, reference genomes (GRCh38), and external databases are included within the Docker image and need about 400 GB of free space. The setup process and directory's structure are described on https://doi.org/10.5281/zenodo.19055361 (https://github.com/LOG-CZECANCA/pardal). The application setup process is described at https://doi.org/10.5281/zenodo.19055389 (https://github.com/LOG-CZECANCA/pardal-app).

Code Availability

The underlying code for the *paRDal* workflow and R/Shiny application can be found at https://doi.org/10.5281/zenodo.19055361 and https://doi.org/10.5281/zenodo.19055389.

ResultS

Performance and Software Testing on a Cohort of 204 Individuals

To demonstrate the scalability of *paRDal*, we benchmarked its performance across two distinct hardware architectures using both targeted clinical panels (CZECANCA) and whole genome/transcriptome pairs.

On a high-end server (AMD EPYC 9654P 96-core, 768 GB RAM), the clinical cohort of 204 individuals sequenced on the CZECANCA panel was processed in 9 hours, representing a peak aggregate throughput of 22.7 samples per hour. Processing 50 whole genome/transcriptome pairs from the 1000 Genomes[33] and Geuvadis[34] projects required 106.5 hours, yielding a throughput of 0.47 samples per hour.

Pipeline performance was also validated on a standard workstation (2x Intel Xeon E5-2690 v2, 128 GB RAM), where 30 individuals analyzed by the CZECANCA panel were processed in 11 hours (2.7 samples per hour), demonstrating that *paRDal* performs robustly even on older infrastructure. Mean compressed file sizes were 250 MB for DNA and 215 MB for RNA (targeted panels).

In unfiltered data, we have identified 98,325 unique variants across 204 individuals (839,163 variant-individual pairs). To identify candidate causal variants, we filtered the original dataset to retain those with a minor allele frequency < 1%, a valid feature (e.g., transcript or regulatory feature), and non-benign ClinVar annotations. High-confidence variant calls were defined as those with a minimum read depth > 10 and a maximum genotype quality score (GQ = 99) in either the matched DNA or RNA sequencing data. The final dataset comprised 5,497 unique candidate causal variants across 204 individuals.

Integration of DNA and RNA evidence

To facilitate rapid identification and prioritization of disease-causing germline variants, with an emphasis on splicing and non-coding variants, we developed an interactive, R/Shiny-based application that bridges genomic data (DNA and RNA variants) with functional transcriptomic evidence (RNA NGS). The application is structured around two primary analytical modules: a cohort-wide global ‘All Variants’ view and a patient-centric ‘Patient View’ (Fig. 2).

DNA/RNA variants are annotated with their functional and structural consequences, protein deleteriousness scores, splicing predictions, population frequencies, and clinical relevance. Annotated variants are displayed in the List of DNA/RNA Variants, with selectable columns or column groups tailored to the specific research or diagnostic question (Additional file Table S1).

Integrated Quality Control via Side-by-Side DNA and RNA Variant Metrics

To facilitate rigorous technical validation and biological interpretation of identified variants, the application features a parallel analysis module (Parallel DNA/RNA Info) that resolves variant-level data at the individual patient level. While cohort-level summaries (Fig. 3) are useful for exploratory data analysis, confidently interpreting a causal variant requires assessing its sequencing quality and expression status at the individual level.

This integrated view provides direct comparisons of critical quality parameters, including total depth (DP) for coverage assessment and genotype quality (GQ) for exclusion of low-quality calls. Allelic depth (AD) and variant allele frequency (VAF) enable the direct DNA-to-RNA VAF comparison, facilitating detection of allelic imbalances such as ASE and NMD.

The application embeds the Integrative Genome Viewer (IGV) through igvShiny[58] for visual validation of matched DNA and RNA BAM slices, with genomic coordinates set automatically based on the inspected variant locus.

Functional Data Integration

The application dynamically integrates several distinct layers of functional evidence (Fig. 2):

- called and thoroughly annotated DNA/RNA variants;
- allele-specific expression (ASE);
- read counts (gene expression) with GTEx reference data;
- expression outliers (computed via OUTRIDER); and
- aberrant splicing events (computed via FRASER) with corresponding Sashimi plots (available in Patient View only) of the affected sample and three randomly selected reference samples.

To aid variant interpretation, the platform embeds external APIs, including multi-tissue expression profiling from the GTEx Portal (V8)[59].

Bidirectional Multiomics Evidence Joining

A major challenge in multiomics diagnostics is accurately linking DNA variants to RNA outliers. To resolve this, we implemented a bidirectionally reactive filtering engine. Selecting a DNA or RNA variant (Fig. 2) instantly filters all RNA outlier tables to show only events mechanistically linked to that variant. Conversely, selecting an RNA outlier instantly queries the DNA variant pool for a potential causal variant. This filtering can be applied at the variant level or globally across all evidence layers, ensuring consistent results throughout.

Biologically Bounded Integration of Splicing Outliers

Splicing aberrations present a unique integration challenge, as a causal spliceogenic variant may reside at canonical splice sites (±1,2), within an adjacent exon, or deep intronic sequences.

When a FRASER splicing outlier is selected, the application queries the reference GTF to identify the exact physical coordinates of the nearest upstream (donor) and downstream (acceptor) exons adjacent to the aberrant junction. The platform dynamically constructs a custom, continuous genomic window that precisely spans from the start of the upstream exon to the end of the downstream exon.

By enforcing spatial intersections against this biologically expanded window, the application reliably captures causative splice-site variants, deep intronic cryptic variants, and exonic regulatory disruptions, without relying on rigid distance buffers that risk capturing adjacent and irrelevant gene variants or missing some of the causative ones.

In Patient View, clicking on a splice junction of interest triggers a Sashimi plot alongside an IGV window displaying the corresponding genomic range.

The paRDal framework in real-life examples

Below, we present several cases illustrating the utility of the *paRDal* framework across specific functionalities, selected from analyses of individuals undergoing genetic testing for hereditary cancer predisposition.

Splicing Landscape in a Set of 204 Individuals Tested for Hereditary Cancer Predisposition

We have analyzed the splicing landscape in 204 individuals who underwent genetic testing for hereditary cancer predisposition. Extraction of SNPs from the original dataset yielded 18,804 variants; regulatory variants were excluded. Subsequent filtering for quality metrics and population frequencies (annotated variants with GQ of 99 in either DNA or RNA, with a depth of 10 reads, and average population frequencies < 0.01 or not available) further limited the SNP dataset to 5,034 variants. Of these, 292

(5.8%) (Fig. 4a) mapped to splice regions (in the vicinity of the splice site or in the polypyrimidine tract); however, most were absent from ClinVar. The largest proportion of variants with high SpliceAI scores and positive FRASER hits were classified as pathogenic, with the majority residing at donor or acceptor sites. In contrast, variants classified as likely benign or VUS exhibited moderate to high SpliceAI with a positive FRASER hit predominantly located in the polypyrimidine tract. As shown in Fig. 4b, many FRASER hits lacked an associated causal variant, suggesting causative variants may reside in deep intronic sequences or regulatory exonic elements.

Detection of Complex Splicing Disruptions Caused by Variants with Inconclusive Evidence (ATM:c.6197A > G)

A carrier of a rare variant affecting the penultimate nucleotide in exon 42 (ENST00000675843:c.6197A > G; Fig. 5a, lower panel) exhibited skipping of exon 42 (E42Δ) in *ATM*. While predicted to cause a missense amino acid change (p.Gln2066Arg), this extremely rare variant (gnomAD NFE MAF = 0.0000009) is currently classified as a VUS in ClinVar. Integrated analysis using *paRDal* revealed a severe failure of the splicing machinery that *in silico* models failed to predict. The variant weakened the donor splice site, leading to the use of an upstream splice donor with $\Delta\Psi = 0.11$. Reference samples showed only a negligible amount of exon skipping (Fig. 5b).

This was accompanied by the retention of both flanking introns (i41 and i42; Fig. 5b), present in a substantially more reads ($\Delta\Psi = -0.2$ for i41 and $\Delta\Psi = -0.27$ for i42).

Although a subset of variant-allele reads exhibited wild-type exon-exon junctions, indicating a leaky splicing, the overall impact on transcript stability was substantial. Statistically significant ASE (p-value adjusted $< 2\times10^{-16}$) with the major allele frequency of 0.65 suggested aberrant transcript degradation, further supported by significant transcript heterogeneity (p-value adjusted $< 2\times10^{-16}$), indicating the presence of diverse isoforms. Notably, none of the protein-based prediction tools classified the missense variant observed in the leaky transcripts as deleterious.

Elucidation of the Impact of Deep Intronic Variants (ATM:c.1899-123A > G)

A rare deep intronic variant in *ATM* (ENST00000675843:c.1899-123A > G), located 123 bp upstream of exon 13, was prioritized due to a high SpliceAI score (> 0.5), suggesting cryptic splice site activation. FRASER detected a mixture of pseudo-exonizations (one dominant with $\Delta\Psi = -0.28$) with 3 alternative acceptors, and varying degrees of exon elongation (one dominant $\Delta\Psi = 0.17$), confirmed by IGV visual inspection and Sashimi plots (Fig. 6).

Transcripts stability was assessed using ASE, which revealed a statistically significant MAF of 0.6. This allelic imbalance suggests that a portion of the aberrant transcripts undergo NMD, likely due to the introduction of a premature termination codon (PTC) within the retained intronic sequences, implying

that the true level of aberrant splicing is even higher, as the damaged isoforms are depleted from the transcript pool. Collectively, these findings support PVS1(RNA) evidence, demonstrating that the variant directs the majority of the transcripts into a variety of non-functional isoforms, consistent with the *in silico* predictions (PP3).

Two-pass STAR+WASP Mapping for Novel Splice-Junctions Rescue (APC:c.1627-8A > G)

FRASER detected a complex splicing pattern, described as a (partial) intron retention with $\Delta\Psi = -0.26$ of intron 13 in *APC*, involving two competing alternative acceptor sites. One site, a 7-bp exon elongation supported by only 5 reads, was linked to a variant in the polypyrimidine tract, while a seemingly distinct, better-supported non-canonical junction (46 reads) was detected 12 bp downstream (Fig. 7a), adjacent to a common synonymous polymorphism (ENST00000257430:c.1635G > A, MAX_AF = 0.84).

Despite a SpliceAI score of 0.99 for the polypyrimidine tract variant, visual inspection identified the 46-read junction as an alignment artifact; the c.1635G > A mismatch introduced a reference bias in which the associated mismatch penalty exceeded aligner tolerances during first-pass STAR mapping, causing these reads to be flagged as an 18-bp insertion (Fig. 7b). Recognizing that these 46 reads and the original 5 junction reads represented the same biological event, we unified the signal into a single 7-bp exon elongation with 51 total supporting reads – a ten-fold improvement achieved through two-pass STAR+WASP realignment. Although ASE was detected, the low SNP density (2 SNPs, heterogeneity = 1) renders these results unreliable. Regardless, the 7-bp elongation is predicted to introduce a PTC and trigger NMD, supporting the application of PVS1(RNA) (or PP3) criteria.

To further evaluate the impact of two-pass alignment on splice junction detection sensitivity, we analyzed uniquely mapped read coverage across four clinically actionable genes (*APC*, *ATM*, *BRCA1*, and *MSH3*) in a single run of 204 individuals. On average, the two-pass strategy rescued spanning reads for 23% of splice junctions missed during the initial pass, while only 1.7% of reads were lost (in fact, they were not depleted but transitioned from uniquely mapping reads to multi-mappers). These results demonstrate the utility of cohort-wide junction database construction in two-pass alignment (Fig. 8).

A Cautionary Tale: Identification and Curation of an APC Promoter Variant in Siblings with GAPPS Syndrome

In two siblings, both diagnosed with gastric adenocarcinoma and proximal polyposis of the stomach (GAPPS), *paRDal* identified a regulatory variant (ENSR5_6H8TT4, g.112707527T > C, c.-191T > C) within the *APC* promoter region. Although isolated in a non-coding regulatory element lacking standard gene-level annotation, the framework's integrated table-joining strategy successfully linked these variants to *APC*, demonstrating *paRDal*'s ability to capture the impact of regulatory variants.

Automated ASE analysis flagged *APC* as statistically significant (p-value < 0.05); however, visual inspection of the Parallel DNA/RNA Info table, per-patient variant metric, and matched DNA/RNA BAMs in IGV revealed that this signal was driven by an extremely low density of heterozygous SNPs within the

coding region (specifically, the presence of a synonymous variant c.1212C > T in homozygous state) consistent with the absence of calculated ASE heterogeneity. No expression outlier was detected.

Despite the established pathogenicity of *APC* promoter variants in GAPPS syndrome[60], the lack of informative heterozygous sites precluded detection of a biologically meaningful ASE. This case highlights the importance of the *paRDaI* interactive interface, which enables clinicians to recognize the limitations of automated outputs and conduct the manual investigation necessary to confirm a variant's functional impact.

Discussion

paRDaI is an integrated clinical bioinformatics framework that operationalizes the transition between computational predictions (ACMG/AMP PP3 criterion) and functional clinical evidence (PVS1(RNA)). Validated across extensive internal clinical cohorts and public datasets, *paRDaI* unifies DNA/RNA variant calling and annotation with advanced transcriptomic analyses, including allele-aware two-pass alignment, expression outlier detection, aberrant splicing identification, and allele-specific expression, facilitating the application of ACMG/AMP criteria and accelerating VUS resolution in routine germline diagnostics.

Despite the potential of RNA NGS to resolve VUS[11–16,30], its clinical adoption remains hampered by a lack of standardized bioinformatics workflows (Fig. 9). Existing frameworks suffer from fragmentation or narrow analytical focus, delegating raw RNA NGS preprocessing to the user and thereby compromising technical reproducibility and diagnostic yield. Furthermore, established tools typically rely on a unidirectional trajectory from outliers to the potentially causal variants. To address these systemic bottlenecks, *paRDaI* offers optimized raw data pre-processing and employs a bidirectionally reactive engine, allowing geneticists to initiate queries from a candidate DNA/RNA VUS or, conversely, to anchor analysis on any observed transcriptomic signal to investigate reciprocal molecular interactions.

While frameworks like DROP and Maassen et al.[61] provide robust downstream analysis, they either delegate the critical pre-processing of raw RNA NGS FASTQs to the user or do not optimize those steps for downstream analysis. *nf-core* frameworks[62] address only isolated analytical components, requiring manual synthesis of outputs. This fragmentation promotes non-standardized aligner settings and directly compromises reproducibility and diagnostic yield. *paRDaI* standardizes the entire trajectory from raw FASTQ processing to the final interactive application, implementing two critical mapping advancements: two-pass STAR alignment and WASP reference-allele bias correction, high-fidelity steps that, while individually addressed by specific modules (e.g., *nf-core* rnavar), are routinely omitted from standard pipelines (Fig. 9).

In contrast to most available RNA NGS frameworks, which either do not integrate DNA variants or use them solely as a basis for ASE analysis, *paRDaI* employs DNA/RNA variant calling primarily for variant exploration while also leveraging VCFs for sample authentication to exclude sample swaps.

While diagnostic prioritization increasingly relies on splicing, expression, and ASE as primary drivers, outlier-detection tools like FRASER and OUTRIDER frequently fail to resolve VUS in isolation[19,20,23,61]. Statistical significance alone often remains inconclusive even for known splice-altering variants; the diagnostic utility of expression outliers is primarily limited to structural variants, and ASE rarely serves as a standalone diagnostic driver. Our findings demonstrate that successful variant classification requires interpreting outlier detection and ASE alongside variant quality metrics, *in silico* predictions, and visualization, rather than treating them as independent metrics. Manual curation remains essential when a variant is suspected to alter the transcriptomic profile[23] (e.g., a a splice-site variant should be evaluated in IGV, even in the absence of significant FRASER hits). *paRDaI* specifically enables both DNA-centric (from a variant to a functional effect) and RNA-centric (from an observed effect to a causal variant) analytical perspectives. By consolidating conflicting scores, annotations, visual evidence, and per-sample metrics into a single interface, *paRDaI* enables the time-efficient application of the PVS1(RNA) ACMG/AMP criterion, eliminating the need for manual cross-referencing and multi-step BAM inspection.

paRDaI is subject to several biological and computational limitations. The workflow currently applies predefined parameters and thresholds optimized specifically for Mendelian diagnostics, which may require adjustment for broader research contexts. Computationally, the autoencoders used for the outlier detection perform optimally with a minimum cohort size of 60 samples[55], a requirement that can be mitigated by utilizing tissue-matched public datasets as a background reference. Biologically, diagnostic utility is limited by tissue-specific expression; although integration of GTEx data helps identify signals absent from the examined tissue. Highly efficient NMD may deplete aberrant isoforms, leaving some aberrant splicing events undetected. Finally, will focus on integrating Pindel[63] and a dedicated CNV analysis module to address current limitations in large indel detection and provide a truly comprehensive multi-omics diagnostic solution.

Declarations

Author Contribution

K.M. – conceptualization, pipeline and application design, data analysis, figure preparation, manuscript preparationP.N. – workflow preparation and pipeline performance optimization, Docker preparationP.K. – conceptualization, data analysisT.P. – sample preparation, data analysisM.Č. – sample preparationP.J. – sample preparationO.V. – pipeline reviewP.Z. – conceptualizationM.G. – data collectionM.J. – variant classification, conceptualizationJ.S. – variant classification, conceptualizationI.S.B. – manuscript drafting and reviewZ.K. – conceptualization, figure preparation, manuscript preparation

Acknowledgements:

This work was supported by Ministry of Health of the Czech Republic grant projects (NU23-03-00150 and RVO-VFN 64165); Charles University research projects (SVV 260631, 24/MED/022, and Cooperatio); and the Ministry of Education, Youth and Sports of the Czech Republic project MULTIOMICS_CZ (Programme Johannes Amos Comenius, ID Project CZ.02.01.01/00/23_020/0008540) and the European Union Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V03-00007 Funded by the EU NextGenerationEU.

Data Availability

The restriction on sharing sensitive genomic data from 204 individuals is a standard practice designed to comply with the European Union's General Data Protection Regulation (GDPR) and national legal privacy frameworks. The data for the minimal framework can be downloaded from https://www.ebi.ac.uk/ena/data/view/PRJEB31736 and the corresponding RNA can be accessed at https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-GEUV-3#o3.

References

1. Kuzbari, Z. *et al.* Germline-focused analysis of tumour-detected variants in 49,264 cancer patients: ESMO Precision Medicine Working Group recommendations. *Ann Oncol* 34, 215–227 (2023). https://doi.org:10.1016/j.annonc.2022.12.003
2. Ewans, L. J. *et al.* Whole exome and genome sequencing in mendelian disorders: a diagnostic and health economic analysis. *Eur J Hum Genet* 30, 1121–1131 (2022). https://doi.org:10.1038/s41431-022-01162-2
3. Smedley, D. *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N Engl J Med* 385, 1868–1880 (2021). https://doi.org:10.1056/NEJMoa2035790
4. Wojcik, M. H. *et al.* Genome Sequencing for Diagnosing Rare Diseases. *N Engl J Med* 390, 1985–1997 (2024). https://doi.org:10.1056/NEJMoa2314761
5. Huang, K. L. *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* 173, 355–370.e314 (2018). https://doi.org:10.1016/j.cell.2018.03.039
6. Burke, W., Parens, E., Chung, W. K., Berger, S. M. & Appelbaum, P. S. The Challenge of Genetic Variants of Uncertain Clinical Significance: A Narrative Review. *Ann Intern Med* 175, 994–1000 (2022). https://doi.org:10.7326/m21-4109
7. Federici, G. & Soddu, S. Variants of uncertain significance in the era of high-throughput genome sequencing: a lesson from breast and ovary cancers. *J Exp Clin Cancer Res* 39, 46 (2020). https://doi.org:10.1186/s13046-020-01554-6
8. Chen, E. *et al.* Rates and Classification of Variants of Uncertain Significance in Hereditary Disease Genetic Testing. *JAMA Netw Open* 6, e2339571 (2023). https://doi.org:10.1001/jamanetworkopen.2023.39571

9. Ha, H. I., Ryu, J. S., Shim, H., Kong, S. Y. & Lim, M. C. Reclassification of BRCA1 and BRCA2 variants found in ovarian epithelial, fallopian tube, and primary peritoneal cancers. *J Gynecol Oncol* 31, e83 (2020). https://doi.org:10.3802/jgo.2020.31.e83
10. Kobayashi, Y. *et al.* Clinical Variant Reclassification in Hereditary Disease Genetic Testing. *JAMA Netw Open* 7, e2444526 (2024). https://doi.org:10.1001/jamanetworkopen.2024.44526
11. Horton, C. *et al.* Diagnostic Outcomes of Concurrent DNA and RNA Sequencing in Individuals Undergoing Hereditary Cancer Testing. *JAMA Oncol* 10, 212–219 (2024). https://doi.org:10.1001/jamaoncol.2023.5586
12. Rhine, C. L. *et al.* Hereditary cancer genes are highly susceptible to splicing mutations. *PLoS Genet* 14, e1007231 (2018). https://doi.org:10.1371/journal.pgen.1007231
13. Li, J. *et al.* Point Mutations in Exon 1B of APC Reveal Gastric Adenocarcinoma and Proximal Polyposis of the Stomach as a Familial Adenomatous Polyposis Variant. *Am J Hum Genet* 98, 830–842 (2016). https://doi.org:10.1016/j.ajhg.2016.03.001
14. Buzby, J. S., Williams, S. A., Schaffer, L., Head, S. R. & Nugent, D. J. Allele-specific wild-type TP53 expression in the unaffected carrier parent of children with Li-Fraumeni syndrome. *Cancer Genet* 211, 9–17 (2017). https://doi.org:10.1016/j.cancergen.2017.01.001
15. Evans, D. G. R. *et al.* A Dominantly Inherited 5' UTR Variant Causing Methylation-Associated Silencing of BRCA1 as a Cause of Breast and Ovarian Cancer. *Am J Hum Genet* 103, 213–220 (2018). https://doi.org:10.1016/j.ajhg.2018.07.002
16. Perrin-Vidoz, L., Sinilnikova, O. M., Stoppa-Lyonnet, D., Lenoir, G. M. & Mazoyer, S. The nonsense-mediated mRNA decay pathway triggers degradation of most BRCA1 mRNAs bearing premature termination codons. *Hum Mol Genet* 11, 2805–2814 (2002). https://doi.org:10.1093/hmg/11.23.2805
17. Dragoš, V. *et al.* Identification of Spliceogenic Variants beyond Canonical GT-AG Splice Sites in Hereditary Cancer Genes. *Int J Mol Sci* 23 (2022). https://doi.org:10.3390/ijms23137446
18. Truty, R. *et al.* Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation. *American journal of human genetics* 108, 696–708 (2021). https://doi.org:10.1016/j.ajhg.2021.03.006
19. Yépez, V. A. *et al.* Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *Genome Med* 14, 38 (2022). https://doi.org:10.1186/s13073-022-01019-9
20. Jaramillo Oquendo, C. *et al.* Identification of diagnostic candidates in Mendelian disorders using an RNA sequencing-centric approach. *Genome Med* 16, 110 (2024). https://doi.org:10.1186/s13073-024-01381-w
21. Agiannitopoulos, K. *et al.* Clinical Utility of Functional RNA Analysis for the Reclassification of Splicing Gene Variants in Hereditary Cancer. *Cancer genomics & proteomics* 18, 285–294 (2021). https://doi.org:10.21873/cgp.20259
22. Frésard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med* 25, 911–919 (2019). https://doi.org:10.1038/s41591-019-0457-8

23. De Cock, L. *et al.* Cracking rare disorders: a new minimally invasive RNA-seq protocol. *NPJ Genom Med* 10, 45 (2025). https://doi.org:10.1038/s41525-025-00502-7
24. Sha, Y. *et al.* Refining the interpretation of variants of uncertain significance in hereditary cancer screening through integrated RNA sequencing. *Genet Med Open* 3, 101914 (2025). https://doi.org:10.1016/j.gimo.2024.101914
25. Luo, X. *et al.* Blood RNA-seq in rare disease diagnostics: a comparative study of cases with and without candidate variants. *J Transl Med* 23, 586 (2025). https://doi.org:10.1186/s12967-025-06609-w
26. Yépez, V. A. *et al.* Detection of aberrant gene expression events in RNA sequencing data. *Nat Protoc* 16, 1276–1296 (2021). https://doi.org:10.1038/s41596-020-00462-5
27. Kamps-Hughes, N. *et al.* A Systematic Method for Detecting Abnormal mRNA Splicing and Assessing Its Clinical Impact in Individuals Undergoing Genetic Testing for Hereditary Cancer Syndromes. *J Mol Diagn* 25, 156–167 (2023). https://doi.org:10.1016/j.jmoldx.2022.12.002
28. Kleiblová, P. *et al.* Parallel DNA/RNA NGS Using an Identical Target Enrichment Panel in the Analysis of Hereditary Cancer Predisposition. *Folia Biol (Praha)* 70, 62–73 (2024). https://doi.org:10.14712/fb2024070010062
29. Walker, L. C. *et al.* Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup. *American journal of human genetics* 110, 1046–1067 (2023). https://doi.org:10.1016/j.ajhg.2023.06.002
30. Brnich, S. E. *et al.* Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med* 12, 3 (2019). https://doi.org:10.1186/s13073-019-0690-2
31. Yepez, V. A. *et al.* Detection of aberrant gene expression events in RNA sequencing data. *Nat Protoc* 16, 1276–1296 (2021). https://doi.org:10.1038/s41596-020-00462-5
32. Soukupova, J. *et al.* Validation of CZECANCA (CZEch CAncer paNel for Clinical Application) for targeted NGS-based analysis of hereditary cancer syndromes. *PLoS One* 13, e0195761 (2018). https://doi.org:10.1371/journal.pone.0195761
33. Auton, A. *et al.* A global reference for human genetic variation. *Nature* 526, 68–74 (2015). https://doi.org:10.1038/nature15393
34. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511 (2013). https://doi.org:10.1038/nature12531
35. Dyer, S. C. *et al.* Ensembl 2025. *Nucleic Acids Res* 53, D948-d957 (2025). https://doi.org:10.1093/nar/gkae1071
36. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43, 11.10.11-11.10.33 (2013). https://doi.org:10.1002/0471250953.bi1110s43

37. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 314–324.
38. Picard toolkit (2019).
39. Auwera, G. v. d. & O'Connor, B. D. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. First edition. edn, (O'Reilly Media, 2020).
40. Danecek, P. & McCarthy, S. A. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* 33, 2037–2039 (2017). https://doi.org:10.1093/bioinformatics/btx100
41. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* 31, 2202–2204 (2015). https://doi.org:10.1093/bioinformatics/btv112
42. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 17, 122 (2016). https://doi.org:10.1186/s13059-016-0974-4
43. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013). https://doi.org:10.1093/bioinformatics/bts635
44. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* 604, 310–315 (2022). https://doi.org:10.1038/s41586-022-04558-8
45. Mertes, C. *et al.* Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat Commun* 12, 529 (2021). https://doi.org:10.1038/s41467-020-20573-7
46. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 176, 535–548.e524 (2019). https://doi.org:10.1016/j.cell.2018.12.015
47. Dawes, R. *et al.* SpliceVault predicts the precise nature of variant-associated mis-splicing. *Nat Genet* 55, 324–332 (2023). https://doi.org:10.1038/s41588-022-01293-8
48. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 381, eadg7492 (2023). https://doi.org:10.1126/science.adg7492
49. Feng, B. J. PERCH: A Unified Framework for Disease Gene Prioritization. *Hum Mutat* 38, 243–251 (2017). https://doi.org:10.1002/humu.23158
50. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46, 310–315 (2014). https://doi.org:10.1038/ng.2892
51. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* 99, 877–885 (2016). https://doi.org:10.1016/j.ajhg.2016.08.016
52. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res* 11, 863–874 (2001). https://doi.org:10.1101/gr.176601
53. Esposito, D. *et al.* MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol* 20, 223 (2019). https://doi.org:10.1186/s13059-019-1845-6
54. den Dunnen, J. T. *et al.* HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat* 37, 564–569 (2016). https://doi.org:10.1002/humu.22981

55. Brechtmann, F. *et al.* OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *Am J Hum Genet* 103, 907–917 (2018). https://doi.org:10.1016/j.ajhg.2018.10.025
56. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28, 3326–3328 (2012). https://doi.org:10.1093/bioinformatics/bts606
57. Google. Gemini 3.1 Pro. (2026).
58. igvShiny: igvShiny: a wrapper of Integrative Genomics Viewer (IGV - an interactive tool for visualization and exploration integrated genomic data) v. 1.6.0 (2025).
59. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45, 580–585 (2013). https://doi.org:10.1038/ng.2653
60. Foretová, L. *et al.* GAPPS - Gastric Adenocarcinoma and Proximal Polyposis of the Stomach Syndrome in 8 Families Tested at Masaryk Memorial Cancer Institute - Prevention and Prophylactic Gastrectomies. *Klin Onkol* 32, 109–117 (2019). https://doi.org:10.14735/amko2019S109
61. Maassen, W. T. K. *et al.* Accelerating rare disease diagnostics by linking DNA and RNA through an explainable and interactive RNA-guided workflow. *NAR Genom Bioinform* 8, lqag016 (2026). https://doi.org:10.1093/nargab/lqag016
62. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* 38, 276–278 (2020). https://doi.org:10.1038/s41587-020-0439-x
63. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2865–2871 (2009). https://doi.org:10.1093/bioinformatics/btp394

Figures

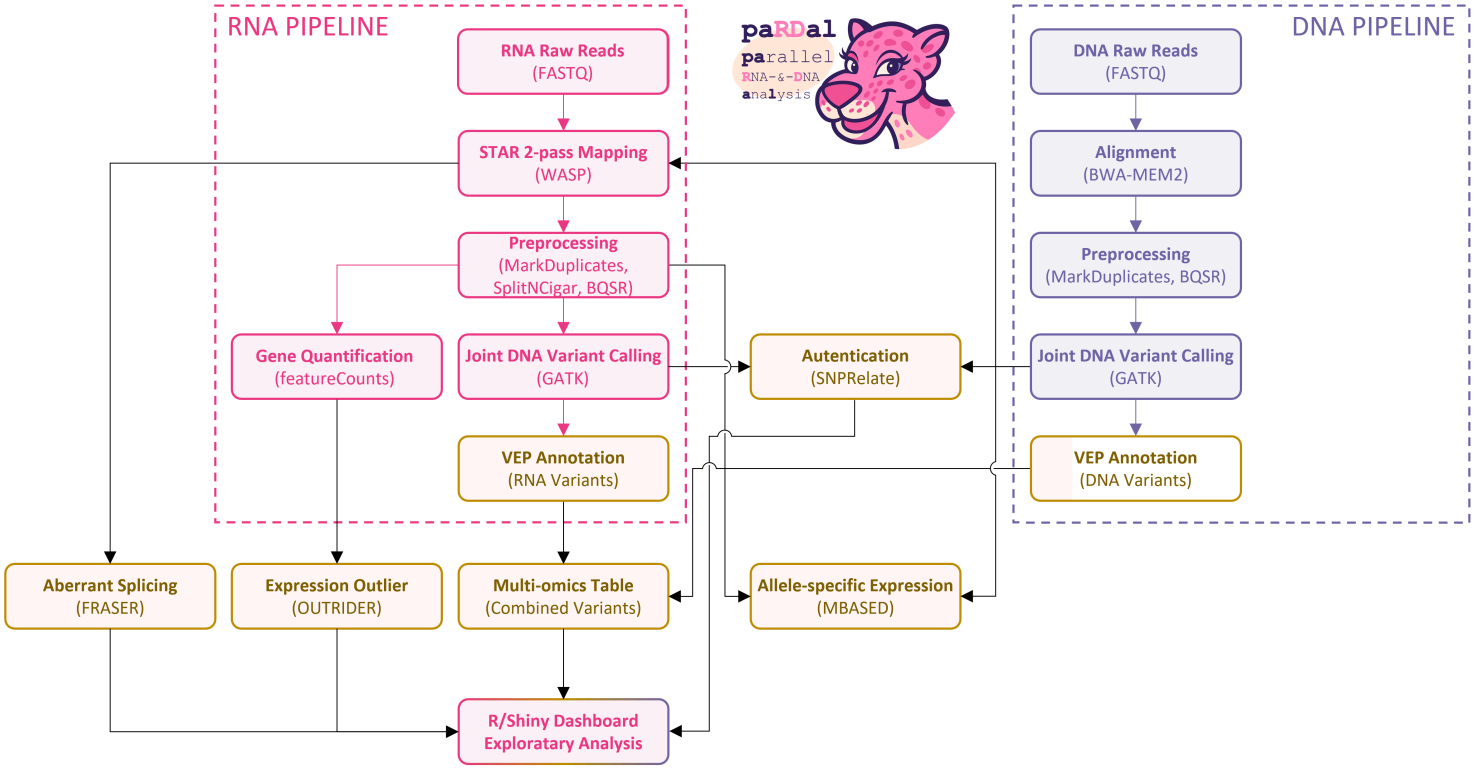


Figure 1

*paRDal*workflow. The flowchart describes basic functionalities of the framework and software tools used.

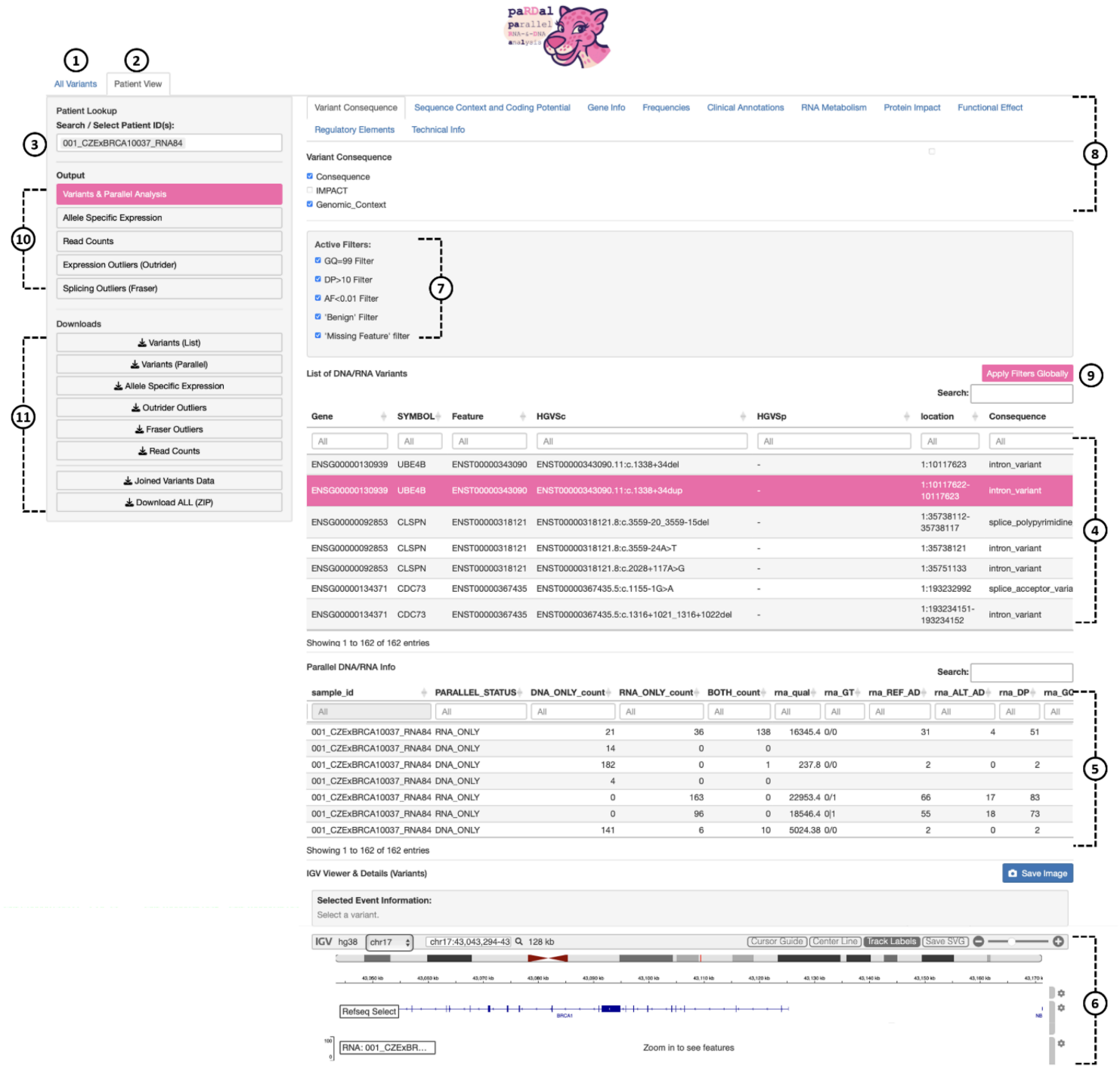


Figure 2

The *paRDal* application layout. Users can toggle between All Variants view (1) displaying all variants across the analyzed cohort (e.g. sequencing run) or Patient View (2), showing variants identified in a given individual (3). List of DNA/RNA Variants (4) includes all variants detected by parallel DNA and RNA NGS, Parallel DNA/RNA Info (5) displays their quality metrics. Clicking a selected variant (not shown) loads it into parallel DNA/RNA tracks in IGV (6). Variants can be prefiltered using switchable filters (7) and annotated using VEP plugins organized into thematically relevant groups (8). The Apply Filters Globally (9) propagates a selected feature (e.g. variant or gene) across all functional output layers ((10), which can be downloaded in csv format (11).

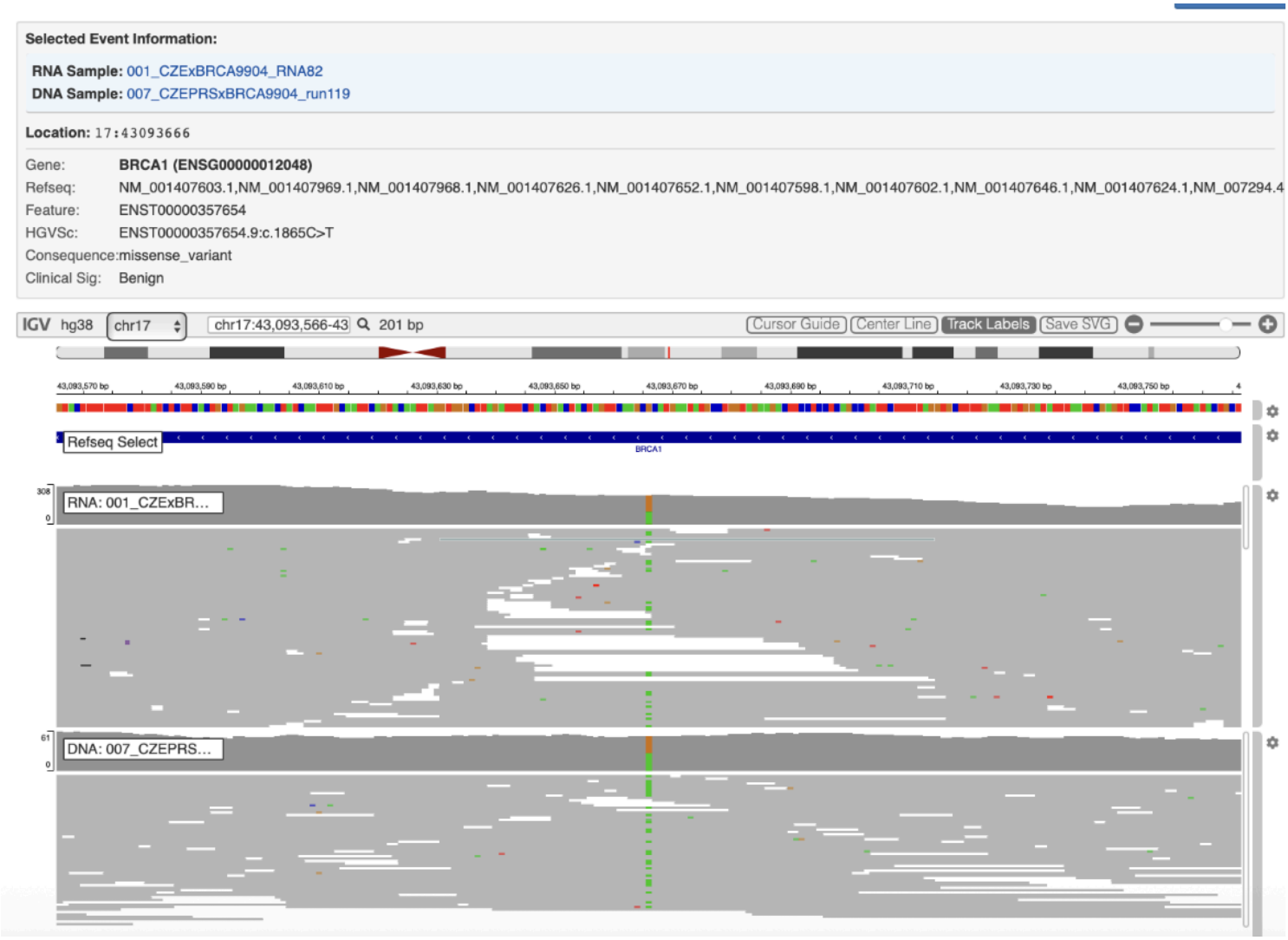


Figure 3

Parallel IGV visualization of a selected variant in Patient View. Upon selection of a variant, matched RNA (upper panel) and DNA (lower panel) BAM file windows are loaded from the NGS analyses (here illustrated for an individual BRCA9904 harboring the BRCA1 missense variant c.1865C>T).

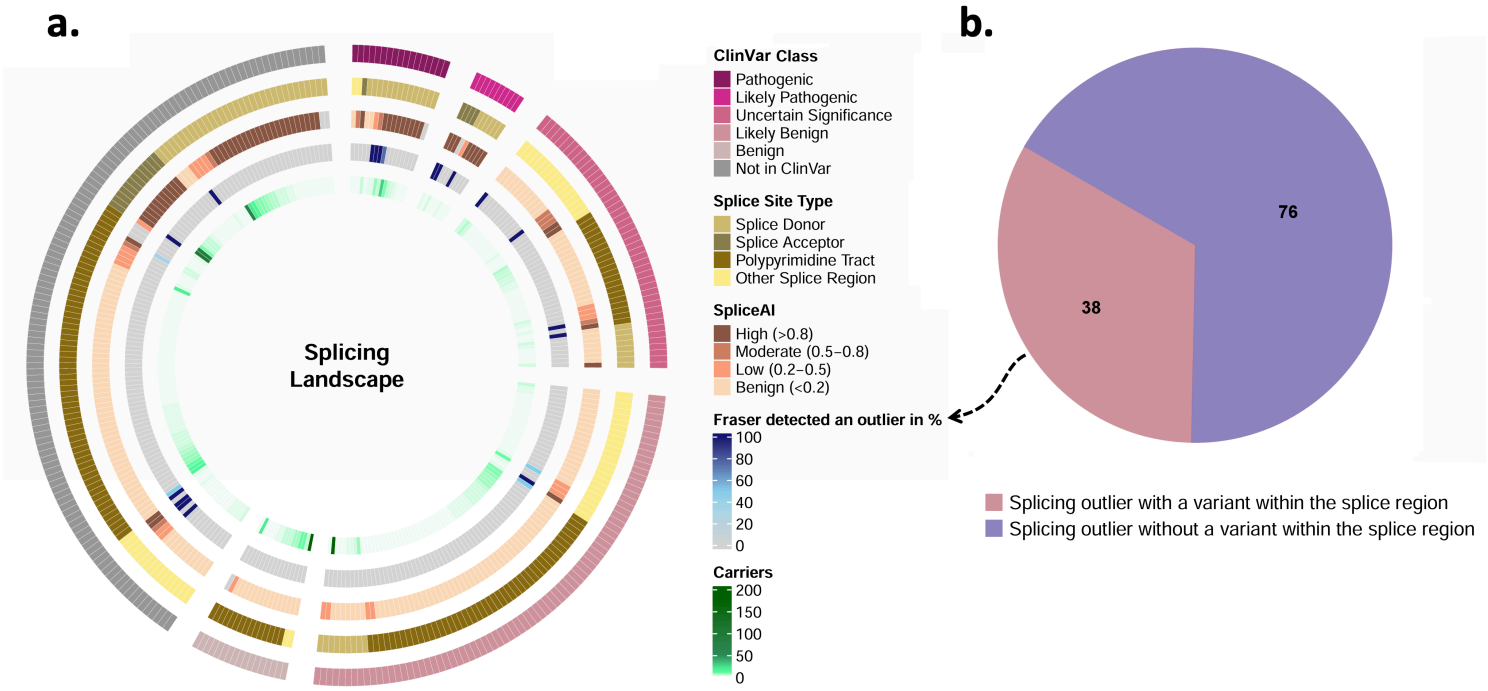


Figure 4

Splicing landscape of 292 VEP-annotated unique variants (a) located in splice regions in 204 individuals tested for hereditary cancer predisposition. FRASER analysis (b) of the same sample set revealed 114 positive hits (splicing outliers), of which 38 (33%) were linked to a causal variant located within the splice region. Splicing outliers with high SpliceAI scores were enriched among pathogenic/likely pathogenic ClinVar variants; however, they were also identified among VUS and few likely benign variants. Eight (7%) FRASER-positive hits in 13 individuals scored to variants not described in ClinVar.

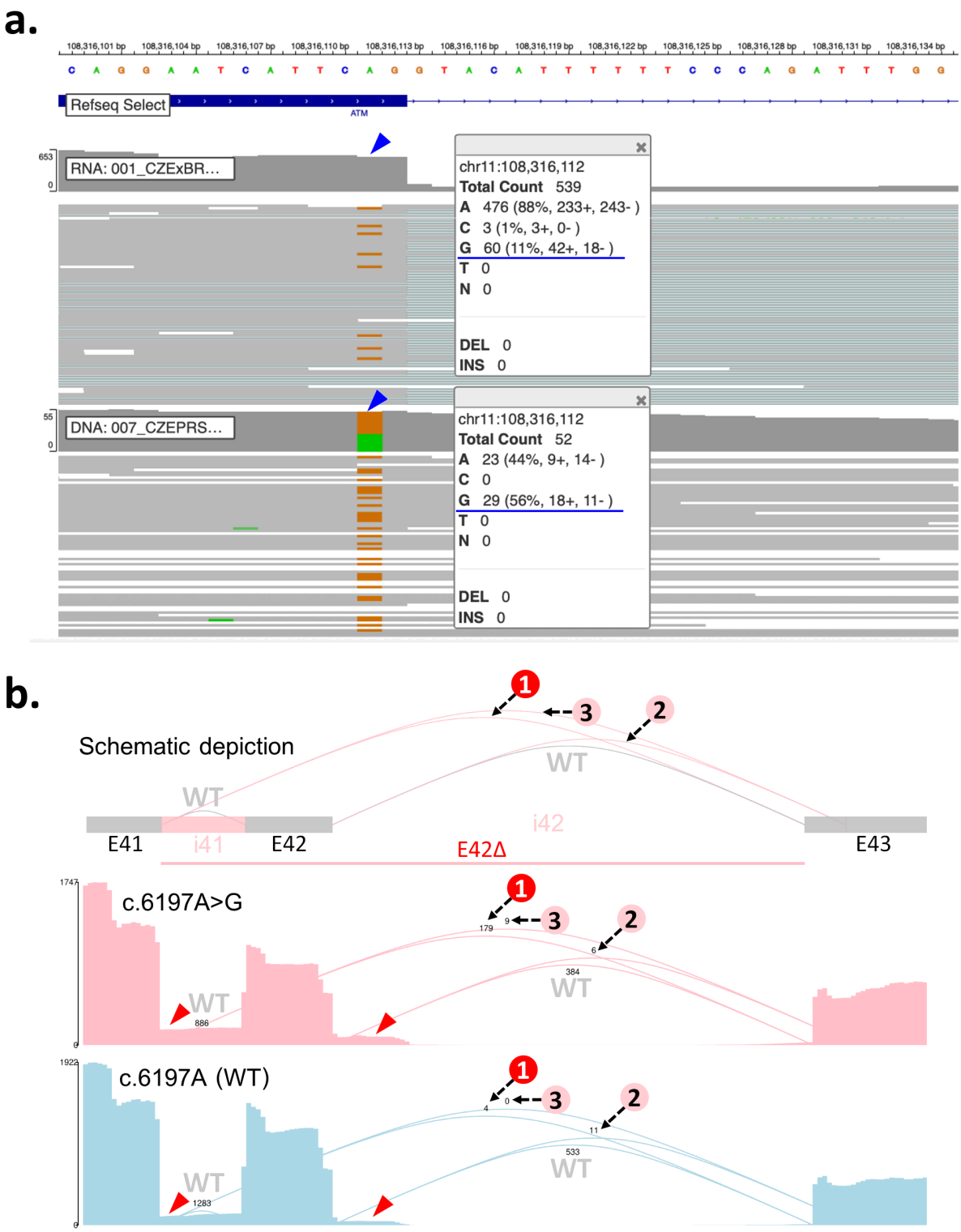


Figure 5

Visualization of the *ATM* c.6197A>G variant (a; blue arrowhead) in IGV showing parallel RNA (upper panel) and DNA (lower panel) NGS data from a carrier in *paRDaI*. RNA VAF is substantially reduced relative to DNA (VAF% underlined in blue). Sashimi plots (b) for the variant carrier and a wild-type (WT) reference sample revealed skipping of E42 (1; 179 vs. 4 events in the carrier and WT sample, respectively) accompanied by enhanced retention of flanking introns (red arrowheads). Minor alternative

splicing events (2 and 3) did not differ significantly between carrier and WT sample (6 vs. 11 and 9 vs. 0 events, respectively).



Figure 6

Sashimi plots (a) and IGV view (b) for an aberrant splicing event caused by *ATM*:c.1899-123A>G in *paRDaI*, shown alongside a wild-type reference. Both panels show pseudo-exonization (*) of in an intrinsic segment of the intron 12 and intron 12 retention (red arrowheads). The c.1899-123A>G variant (b; black arrow) is shown at higher magnification (right inset) to illustrate increased proportion of the variant allele within the retained intron sequence.



Figure 7

Sashimi plots (a) for *APC*:c.1627-8A>G. The carrier RNA sample depicts two novel aberrant acceptor sites absent from the wild-type (WT) reference, producing partial (7 bp) exon elongation (red arrowhead). The IGV plot (b) of the carrier RNA sample shows small number of NGS reads spanning the c.1627-8A>G variant in unprocessed RNA, common homozygous polymorphism c.1635G>A, and, 3 bases downstream, an erroneously mapped 18-nucleotide insertion representing the retained acceptor site sequence of intron 13 (correctly mapped only in the read highlighted by red horizontal arrowhead).

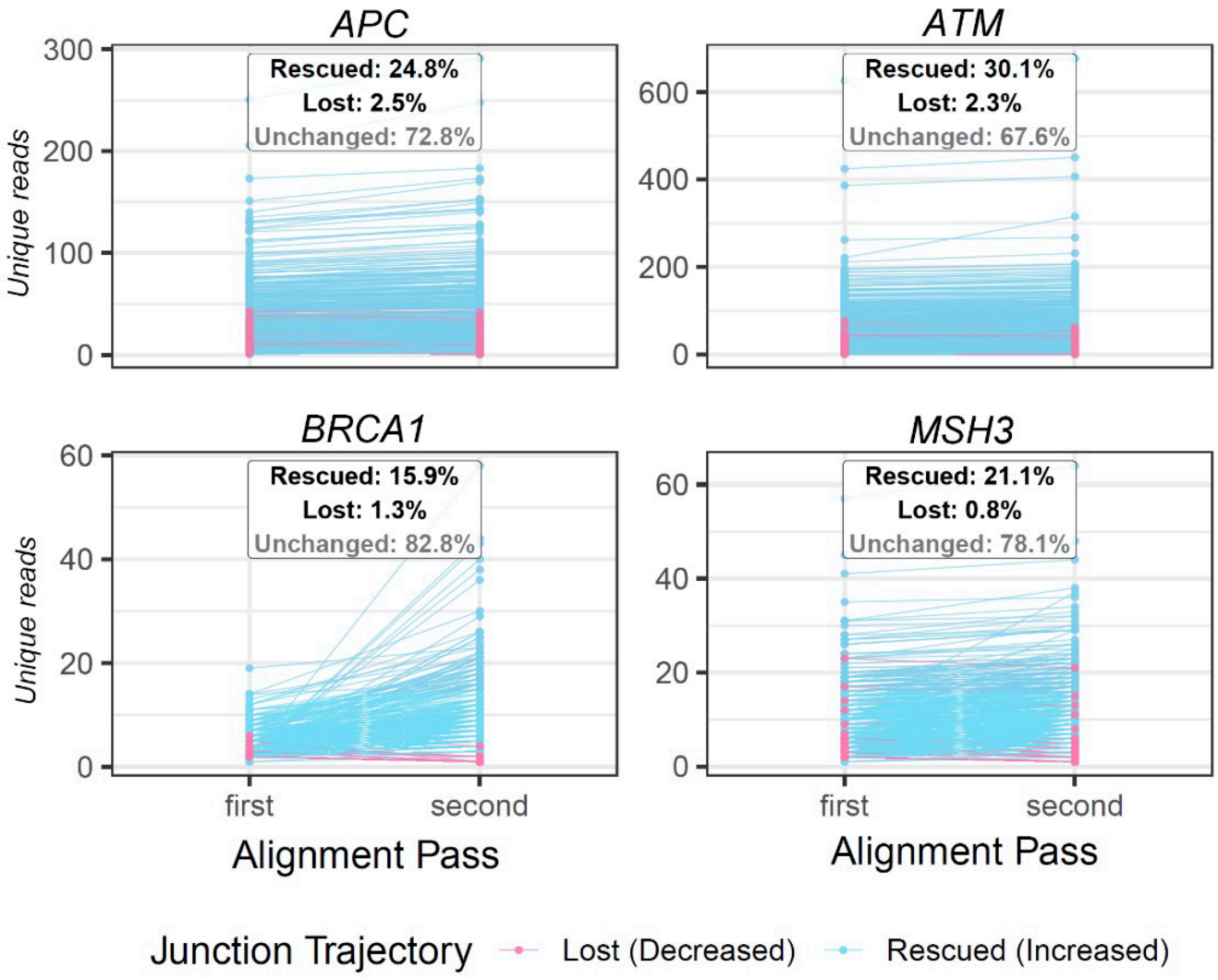


Figure 8

Rescue of novel splice junctions by two-pass STAR alignment across four clinically actionable multi-exon genes. The graphs show the number of novel splice junction after the first (left) and second (right) STAR pass across 204 samples. Junctions newly detected in the second pass and unchanged junctions are not depicted.

	paRDal	DROP [31]	*nf-core* frameworks [62]	*Maassen et al., 2026* [61]
Two-pass STAR	**Yes**, with strict parameters	**No**	**Fragmented**, rnavar (basic)	**No**
WASP+STAR	**Yes**	**No**	**Fragmented**, alleleexpression	**No**
Splicing outlier	**Yes**, FRASER	**Yes**, FRASER	**Fragmented**, FRASER in drop	**Yes**, FRASER
Expression outlier	**Yes**, OUTRIDER	**Yes**, OUTRIDER	**Fragmented**, OUTRIDER in drop	**Yes**, OUTRIDER
Allele specific expression	**Yes**, MBASED	**Yes**, negative binomial test	**Yes**, alleleexpression with phASER	**No**
***In silico* predictions**	**Yes**, splicing, protein deleteriousness	**No**	**Fragmented**, only popular plugins or manual setting	**Yes**, SpliceAI, CAPICE
Clinical annotations and population frequencies	**Yes**, e.g. ClinVar, OMIM, gnomAD	**No**	**Fragmented**, only popular plugins or manual setting	**Yes**, ClinVar, OMIM, HPO
Visual validation and manual curation	**Yes**, in interactive R/Shiny App	**Limited**, in HTML report	**No**	**Limited**, in HTML report
Multi-omics data integration	**Yes**, in Shiny app with multidirectional filtering	**No**	**No**	**Limited**, in HTML report
Containerization	**Yes**, Docker	**Yes**, different kinds	**Yes**, Docker	**Yes**, Apptainer

Figure 9

Comparison of state-of-art tools for integrated DNA/RNA analysis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS2listoftools.docx
- TableS1listofvepplugins.xlsx

- TableS3listofseqmetrics.xlsx