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.

varianTalker: a Nextflow Framework for Scalable, Portable and Clinically Oriented Variant Annotation, Prioritization, and Review

Gianluca A. Mariani
European Institute of Oncology, IRCCS
Marco Barreca
European Institute of Oncology, IRCCS
Elena Belloni
European Institute of Oncology, IRCCS
Valerio Fassi
European Institute of Oncology, IRCCS
Ruggero Maria
Catholic University of the Sacred Heart
Camilla Nero
Catholic University of the Sacred Heart
Enrico Derenzini
European Institute of Oncology IRCCS
Corrado Tarella
European Institute of Oncology IRCCS
Luca Mazzarella
European Institute of Oncology, IRCCS
Pier Giuseppe Pelicci
European Institute of Oncology, IRCCS
Chiara Ronchini
European Institute of Oncology, IRCCS
Gianmaria Frigè
European Institute of Oncology, IRCCS
Yinxiu Zhan
Yinxiu.Zhan@ieo.it

European Institute of Oncology, IRCCS

Method Article

Keywords: clinical genomics, precision oncology, variant annotation, variant prioritization, molecular tumour board, reproducibility, source available, Nextflow

Posted Date: October 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11052444/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

Accurate variant interpretation is essential for molecular tumour boards and precision oncology, yet available tools are often fragmented, proprietary, or difficult to maintain. We present varianTalker, a framework implemented in Nextflow DSL2 that integrates ten complementary annotation resources (Funcotator, ANNOVAR, CancerVar, InterVar, ReNOVo, CIViC, AlphaMissense, PharmGKB, ESCAT, and ClinVar) within a containerised, batch-capable workflow. The pipeline supports somatic and germline variants, applies the appropriate classification framework, and accepts outputs from Thermo Fisher Ion Torrent/Oncomine, Illumina DRAGEN, and nf-core Sarek. Consistent consequence reporting is achieved through MANE Select transcripts, while dedicated update modules, custom institutional annotations, and explicit resource versioning support reproducible analyses. A PostgreSQL-backed variant store reuses previously computed annotations, reducing analysis time by an average of 92% when queried variants have already been annotated. Integration of multiple sources provided additional complementary evidence to 47.1% of somatic and 86.9% of germline variants of uncertain significance. In addition, configurable filtering reduced thousands of annotated variants to tens of clinically relevant candidates for interactive review in a Streamlit application. Applied to retrospective cohorts, varianTalker offers a scalable, transparent framework for reproducible variant interpretation.

Background

Molecular profiling has become a central component of modern precision oncology. High-throughput sequencing enables the identification of somatic alterations, germline predisposition variants, copy-number changes, gene fusions and other molecular features that can inform diagnosis, prognosis, therapeutic selection and clinical-trial eligibility. For oncologists and molecular tumour boards, molecular profiles provide a patient-specific framework for interpreting tumour biology and for guiding therapeutic decisions, such as targeted therapies, immunotherapies, off-label treatment options or molecularly guided trials[1]. As sequencing technologies have become increasingly available in routine oncology, the clinical value of a genomic test is no longer determined only by the detection of variants, but also by the ability to transform complex molecular findings into reliable, interpretable and actionable knowledge[2–4].

This translation remains challenging. Variant interpretation requires the integration of heterogeneous evidence, including predicted functional consequence, oncogenicity, germline pathogenicity, therapeutic actionability, pharmacogenomic associations, population frequency, tumour context and evidence from curated knowledgebases and clinical guidelines according to AMP/ASCO/CAP[4] and ACMG/AMP[5] recommendations. State-of-the-art resources such as ANNOVAR[6], VEP[7], ClinVar[8], CIViC[9], OncoKB[10], InterVar[11] and CancerVar[12] have substantially advanced different aspects of variant annotation and clinical interpretation. However, the current landscape remains heterogeneous: tools often focus on specific variant classes, interpretation frameworks, while clinical teams must manually combine outputs from multiple databases and pipelines. Moreover, interpretation practices can vary between institutions and experts, because evidence collection, classification rules, database versions and computational thresholds are not always harmonised. This limits reproducibility and makes it difficult to generate a comprehensive patient-level genomic overview within the short turnaround times required by molecular tumour boards[13].

Several open-source platforms have sought to bridge the gap between variant annotation and clinical interpretation. The Personal Cancer Genome Reporter (PCGR)[14] integrates cancer-relevant knowledge resources to generate tiered, clinically interpretable reports for individual tumour genomes, with emphasis on somatic SNVs/indels, copy-number aberrations, oncogenicity and clinical actionability. OpenCRAVAT[15] provides a flexible and modular framework for variant annotation, scoring and prioritisation, enabling users to assemble customised workflows from a broad catalogue of annotation resources. Both tools, however, leave gaps for precision-oncology use. PCGR is mainly focused on tumour somatic reporting, while germline predisposition interpretation is delegated to a companion tool (Cancer Predisposition Sequencing Reporter, or CPSR). OpenCRAVAT is not cancer-specific and provides no native tier assignment or therapy matching, so its use for precision-oncology requires extensive user-defined configuration and downstream synthesis. What remains missing is a layer that harmonises these annotation workflows into a single coherent representation of a patient's genomic background, reducing the manual burden of evidence gathering and standardising interpretation across reference transcripts and sequencing platforms[13].

Here we introduce varianTalker, a Nextflow DSL2 framework for the clinical interpretation of single-nucleotide variants and short insertions and deletions in cancer patients. varianTalker processes somatic and germline variants in a single execution and classifies each under its established guideline framework, AMP/ASCO/CAP for somatic variants and ACMG/AMP for germline variants. varianTalker aggregates evidence in three layers. Functional and predictive annotation comes from Funcotator and ANNOVAR, complemented by AlphaMissense for missense variants and RENOVO for germline pathogenicity prediction. Guideline-based classification is assigned by InterVar (ACMG/AMP, germline) and CancerVar (AMP/ASCO/CAP, somatic). Curated clinical and pharmacogenomic evidence is drawn from ClinVar, CIViC and PharmGKB, together with an in-house resource mapping gene-drug-tumour-type associations to ESMO ESCAT tiers, which adds therapeutic actionability as an evidence layer rather than a classification framework. Results are returned as a fully annotated variant table together with separate prioritised and excluded variant sets. Both can be explorable through MAFigate, a companion interactive dashboard designed for use by molecular tumour boards during case discussion. We applied varianTalker to retrospective clinical sequencing cohorts from a single institution, spanning 320 analyses of breast, ovarian, colorectal and bone-marrow malignancies generated on three upstream workflows. Integrating complementary evidence supplied an additional classifying call for 86.9% of germline and 47.1% of somatic variants of uncertain significance. Configurable filtering, assessed on 100 of these samples, reduced the primary review set by 97.0% and 99.1% respectively, without excluding a single pathogenic or likely pathogenic variant.

Methods

Implementation

varianTalker is implemented in Nextflow DSL2[16] (requiring version ≥ 22.10.1) and is distributed at https://github.com/zhanyinx/variantalker. Containerized execution is supported through Docker and Singularity/Apptainer, with the execution backend selected at runtime using the Nextflow -profile option. The same workflow definition can therefore be deployed unmodified on local workstations, HPC clusters or

cloud environments. Resumable execution, the full invocation syntax and the runtime parameter schema are described in Supplementary Methods.

Runtime parameters are specified in nextflow.config and formally described in nextflow_schema.json in the repository. Reference genome assemblies GRCh37 (hg19) and GRCh38 (hg38) are both supported by varianTalker.

Input specification

varianTalker accepts a CSV-formatted sample sheet as its primary input. The file requires a mandatory header row and the four columns described in Table 1. Somatic samples require a tumour tissue designation drawn from a controlled vocabulary of 33 terms, listed in Supplementary Methods; germline entries do not require tissue annotation.

Table 1
– varianTalker sample sheet column specification.

Column	Required	Accepted values	Notes
patient	Yes	Any string identifier	Unique per sample; used as output folder name
tumor_tissue	For somatic	33 controlled-vocabulary terms (e.g. Lung, Breast, Colorectal)	Full list given in Input specification
sample_file	Yes	Absolute path to VCF.gz or VCF file	Relative paths are not supported
sample_type	Yes	somatic | germline	Determines annotation tools and filter logic applied

VCF normalisation

Before annotation, all input VCFs are passed through a reference-aware normalization step implemented with bcftools norm[17]. The step splits multiallelic sites and applies reference-based left-alignment with REF allele verification against the selected FASTA resource. This matters because varianTalker accepts VCF files generated by heterogeneous upstream platforms, including Ion Torrent/Oncomine[18], nf-core Sarek[19] and DRAGEN[20], which differ in variant representation, INFO-field structure and genotype formatting. All downstream modules therefore operate on a common, reference-consistent representation.

Integrated annotation resources and MANE transcript standardization

Resources were selected to capture non-redundant dimensions of variant relevance: transcript-level consequence, somatic oncogenicity, germline pathogenicity, therapeutic actionability, pharmacogenomic relevance and population-level evidence. Critically, varianTalker preserves the native evidence model of each resource rather than collapsing all annotations into a single score. Extended descriptions of each resource are given in Supplementary Methods. Table 2 lists all integrated resources with their version information.

Funcotator (GATK)[21] provides transcript-aware functional annotation. It is the functional annotation layer. It accepts custom annotation sources in XSV format, so an institution can add local variant databases or in-house actionability tables without modifying the workflow code. Funcotator is run in canonical transcript-selection mode (--transcript-selection-mode CANONICAL) against a datasource in which MANE[22] Select transcripts are designated canonical, so consequences are reported on MANE Select transcripts. The MANE project pairs matched RefSeq[23] and Ensembl/GENCODE[24].

ANNOVAR[6] provides the annotation infrastructure on which CancerVar[12] and InterVar[11] depend. CancerVar applies the AMP/ASCO/CAP somatic variant classification framework[4], assigning tiers from I (strong clinical significance) to IV (benign or likely benign), while InterVar applies ACMG/AMP 2015 germline classification criteria[5], classifying variants from pathogenic to benign. The complete list of all the databases used by Funcotator and ANNOVAR are reported in Supplementary Table S1.

CIViC[9] contributes manually curated cancer evidence from peer-reviewed literature, each item carrying an evidence level from A (a validated association) to E (an inferential association).

AlphaMissense[25] provides deep-learning-based pathogenicity predictions. It is used as supportive functional evidence, particularly for missense variants lacking strong curated clinical evidence.

ClinVar[8]-based annotation links variants to externally curated clinical significance assertions. ClinVar is therefore treated as curated clinical evidence that complements, rather than replaces, the guideline-based classification frameworks.

ReNOVo[26] provides an independent machine learning based estimate of germline variant pathogenicity. It reports a Pathogenicity Likelihood Score (PLS) and serves as an additional computational evidence layer, particularly for variants with limited, uncertain or conflicting curated evidence

PharmGKB[27] extends varianTalker annotation to pharmacogenomic interpretation by linking germline variants and genes to drug response, toxicity, dosage or metabolism evidence. It is applied to germline samples only, since pharmacogenomic associations reflect inherited rather than tumour-acquired genotypes.

ESCAT[28]-linked resources implement the ESMO Scale for Clinical Actionability of molecular Targets, providing an additional evidence tier for therapeutic actionability in oncology. We curated ESCAT tiering from ESMO guidelines across tumour types and developed software to annotate variants with the corresponding tiers. Tier I alterations are ready for routine clinical implementation, whereas lower tiers represent decreasing levels of clinical evidence (Supplementary Methods). This provides an oncology-specific actionability layer alongside CancerVar and CIViC.

All annotation layers are preserved as separate fields in the unified Mutation Annotation Format (MAF) output, so that downstream filtering and molecular tumour board review can draw on any combination of evidence dimensions while retaining the provenance and interpretation context of each source.

Table 2
– Annotation resources integrated by varianTalker. Versions are the software release or database snapshot used for the analyses reported here. The update_db module pulls newer upstream releases on demand.

Resource	Scope	Version / release	Reference / URL
Funcotator (GATK)	Both	01/2026	gatk.broadinstitute.org
ANNOVAR humandb	Both	01/2026	annovar.openbioinformatics.org
CancerVar database	Somatic	1.1.1	github.com/WGLab/CancerVar
InterVar database	Germline	2.2.2	github.com/WGLab/InterVar
CIViC	Somatic	12/2024	civicdb.org
AlphaMissense	Both	09/2023	zenodo.org/record/8208688
ClinVar VCF	Both	12/2025	ncbi.nlm.nih.gov/clinvar
ReNOVo	Germline	V1	github.com/mazzalab-ieo/renovo
PharmGKB	Germline	3.1.1	pharmgkb.org
ESCAT-linked data	Somatic	09/2025	esmo.org

Variant filtering

The filtering layer applies two classes of filter. Technical and consequence-based filters remove variant classes with no expected coding impact (silent, intronic, untranslated-region, flanking and intergenic) and impose a minimum read depth, a minimum variant allele fraction and, optionally, a gene list (none by default). These are combined with AND logic, with one exception: a pathogenic or likely pathogenic call from any enabled clinical annotation source overrides the consequence, coverage, allele-fraction and gene-list criteria. Annotation-based filters then retain variants according to the classifications and actionability tiers of the integrated resources.

By default, varianTalker applies permissive OR logic to the annotation-based filters: a variant is included in the PASS set if at least one selected annotation source meets or exceeds the configured threshold. This maximises sensitivity for molecular tumour board review while preserving all source-specific annotations.

Default thresholds are listed in Table 3. The same filters can be applied interactively in MAFigate (see below), enabling iterative review of alternative criteria without editing configuration files.

Table 3

– Filter parameters exposed by varianTalker. Numeric thresholds are shown with their default values; for the annotation-based filters the column lists the full set of accepted values rather than a default, since the categories retained are chosen per analysis. All parameters are configurable in nextflow.config or overridable in the Streamlit application. The configuration applied to the IEO cohorts is given in Table 4. VAF: variant allele fraction.

Parameters	Possible Variables	Logic	Notes
Variant type exclusion	Funcotator variant classifications (Silent, Intron, 3'UTR, 5'UTR, IGR, 5'Flank, 3'Flank, RNA and others; see Funcotator documentation)	AND (overridden if any clinical annotation classifies the variant as (likely)pathogenic)	Excluded unless ClinVar/CancerVar/ InterVar classify as Pathogenic or Likely Pathogenic (P/LP)
Minimum coverage	50× (50)	AND (overridden if any clinical annotation classifies the variant as (likely)pathogenic)	Applied to both somatic and germline
Minimum somatic VAF	1% (0.01)	AND (overridden if any clinical annotation classifies the variant as (likely)pathogenic)	Tumor allele fraction
Minimum germline VAF	20% (0.20)	AND (overridden if any clinical annotation classifies the variant as (likely)pathogenic)	Heterozygous germline cutoff
CancerVar tier	Tier_I_strong, Tier_II_potential, Tier_III_Uncertain, Tier_IV_benign	OR	Somatic oncogenicity according to AMP/ASCO/CAP
InterVar class	Pathogenic, Likely pathogenic, Uncertain significance, Likely benign, Benign	OR	Germline ACMG/AMP classification
ClinVar	Pathogenic, Likely_pathogenic, Uncertain_significance, Conflicting_classifications_of_pathogenicity, Likely_benign, Benign, drug_response, not_provided, no_classifications_from_unflagged_records, risk_factor, Affects, no_classification_for_the_single_variant, association, other, Uncertain_risk_allele, Likely_risk_allele, Pathogenic,_low_penetrance, protective, association_not_found, Established_risk_allele, confers_sensitivity	OR	Clinico-pathogenic variant association

Parameters	Possible Variables	Logic	Notes
ReNOVo class	Low, Intermediate or High Probability Pathogenic	OR	Independent pathogenicity score
CIViC evidence	A, B, C, D, E	OR	Clinical actionability level
ESCAT	IA, IB, IC, IIA, IIB, IIIA, IIIB, IV, V	OR	Clinical actionability level
Gene filter	List of genes as csv with one column and no header. The list should be specific for germline or somatic sample type.	AND (overridden if any clinical annotation classifies the variant as (likely)pathogenic)	All genes retained by default

Output structure

For each sample, varianTalker generates a complete annotated MAF file containing all processed variants and two filtered tabular outputs derived from the configurable prioritisation layer: a PASS TSV and a NOPASS TSV. PASS contains variants meeting the active criteria; NOPASS retains the remainder for audit, threshold revision and exploratory review.

PostgreSQL database mode, update_db module, and retrospective re-annotation

To support institutional reuse, varianTalker includes an optional PostgreSQL-backed database (version 17.5) mode. In this mode, variants from a new sample vcf are matched against stored records using canonical variant coordinates, including chromosome, position, reference allele, alternate allele and genome build. When a matching variant is already present in the database, the corresponding annotation record can be retrieved directly instead of re-running the full annotation workflow. This reduces processing time for the recurrent variants typical of cancer genomics. Annotations are reused only when the new sample is processed with an identical analytical configuration, identified by a hash of the annotation tools, database versions, genome build, sequencing technology, germline/somatic mode and tumour tissue; each configuration is mapped to its own schema, so a database or genome-build change never causes outdated annotations to be reused. An optional update_db module maintains the Funcotator and ANNOVAR databases updated on demand. The database schema and the update_db module are described in full in Supplementary Information (Supplementary Methods section).

MAFigate interactive dashboard

Annotated variant tables produced by varianTalker can be reviewed through MAFigate, a companion web application built with Streamlit (https://streamlit.io). It runs entirely on the user's machine, serving only the

local loopback interface, so no patient data leaves the local environment, and is distributed as source code and as self-contained desktop installers for macOS and Windows. It reads the annotated MAF tables produced by varianTalker and operates in somatic or germline mode, which determines the interpretation columns and filters applied; the mode is user-selected, and MAFigate reports any mismatch with the arm inferred from the annotation columns present.

Filtering parameters can be selected from predefined presets, broad or stringent, for somatic and germline analyses, or through fully customised user-defined settings. Configurations persist between sessions and can be exported or re-imported as JSON or YAML, so the same criteria can be applied across a project.

A variant is reported when it satisfies both a set of technical and biological criteria and at least one clinical annotation criterion, or when it is retained by an independent pathogenic-variant rule. The technical and biological criteria follow the pipeline filters described above, with an addition: a variant is retained when its population allele frequency, evaluated across gnomAD[29], ExAC[30], ESP6500[31] and 1000 Genomes[32], is at or below the chosen threshold. Moreover, the following predefined gene list are included: MSK-IMPACT[33], MSK-IMPACT Heme[34], FoundationOne CDx[35], FoundationOne Heme[36], COSMIC Cancer Gene Census[37] or OncoKB[10,38].

Passing and non-passing variants are reported separately, and MAFigate records the settings used and, for each variant, which criterion excluded it.

For interpretation, MAFigate generates a per-variant detail panel. The detail panel reports ACMG PP3/BP4 evidence strengths derived from ClinGen SVI calibrated thresholds for REVEL[39], BayesDel[40], VEST4[41] and CADD[42,43]. The panel also renders the full InterVar ACMG/AMP criteria for germline variants, and CancerVar's twelve AMP/ASCO/CAP evidence categories for somatic variants, graded by evidence strength. Reviewers can record free-text notes and user-defined annotation columns, which are included in the exported tables while the pipeline-generated annotations remain unchanged. Detailed guidelines for operating the application and interpreting its output are provided as Supplementary File 1.

Validation cohorts

varianTalker was evaluated using both ClinVar-derived benchmark datasets and retrospective clinical sequencing cohorts processed at the European Institute of Oncology (IEO). Four aspects were assessed: annotation accuracy, concordance with an independent resource (VarSome[44]), computational performance with and without PostgreSQL-backed retrieval, and clinical relevance of the prioritisation filters.

To evaluate performance under controlled conditions, we generated benchmark datasets using ClinVar database[45]. Random subsets of 5,000, 10,000, 50,000, 100,000, 500,000 and 1,000,000 variants were extracted from ClinVar. One dataset was generated per size and analysed five times to characterise variability in execution time, and the complete ClinVar database was analysed as a full-scale benchmark. A further 1,000 simulated samples of 1,000 random ClinVar variants each were generated to assess scaling with sample number.

ClinVar clinical significance annotations were used as reference labels for annotation accuracy evaluation, with particular attention to variants classified as pathogenic/likely pathogenic and benign/likely benign. Variants annotated in ClinVar as variants of uncertain significance were analysed separately to assess how varianTalker's integrated evidence sources prioritise or reclassify uncertain variants. We examined two reciprocal sets: ClinVar variants of uncertain significance receiving a complementary call from InterVar or CancerVar, and InterVar or CancerVar variants of uncertain significance classified based on ClinVar evidence. Both were submitted to VarSome[44] and agreement assessed at the classification-tier level. VarSome was queried manually (July 2026 release, under institutional licence, v.13.18.1.2).

Computational performance was evaluated using the ClinVar-derived artificial datasets. For each dataset size, we measured the time required to complete varianTalker analyses with and without PostgreSQL database mode. Additional analyses assessed query time as a function of PostgreSQL database size, using progressively larger stored variant collections derived from the same benchmarks.

Clinical validation was performed using retrospective IEO sequencing data processed across the three supported upstream workflows. The Ion Torrent/Oncomine dataset included 78 matched samples (39 germline and 39 somatic cases: 23 breast, 11 ovary and 5 colorectal cancers) from GerSom project[46].

A second dataset, composed of 21 germline and 45 tumour (bone marrow cases) was analysed through both DRAGEN Bio-IT[20] platform (v. 4.4.4) and nf-core/sarek[47] (v. 3.7.1), allowing evaluation of the same clinical samples processed by two independent upstream variant callers.

To assess the filtering and prioritization layer, we used a further 100 samples (50 matched tumor-normal pairs including: 28 breast, 16 ovary and 6 colorectal cases) from GerSom project analysed using Ion Torrent sequencing platform. The clinical relevance of the prioritised variants was assessed against a manually curated review by an internal IEO expert, to determine whether varianTalker retained the variants considered relevant and by how much it reduced the set requiring manual inspection.

Further information related to preprocessing, variant calling and downstream analysis are included in the Supplementary Information.

Results

We evaluated varianTalker across benchmark datasets and retrospective clinical sequencing cohorts to assess its ability to harmonise heterogenous variant annotation, prioritise clinically relevant finding and support scalable institutional review. As an initial qualitative comparison, we examined the documented capabilities of varianTalker alongside widely used variant annotation and interpretation tools, including PCGR, OpenCRAVAT and Ensembl VEP, as described in Supplementary Table S2. Of the four tools, varianTalker is the only one able to classify somatic and germline variants under their respective guideline frameworks within a single execution and to store computed annotations for reuse across samples. We report our results organized into three complementary analyses: evaluation of PostgreSQL-backed annotation retrieval performance using ClinVar-derived datasets; assessment of how integrating different

annotators affects the pathogenicity classification of variants of uncertain significance using the datasets of IEO cohorts; and evaluation of the clinical relevance of the variant-filtering strategies.

PostgreSQL-backed annotation retrieval improves scalability and reduces analysis time

We evaluated the computational scalability of varianTalker using two benchmark sets: single datasets of increasing number of variants randomly drawn from ClinVar, including the complete database, and simulated multi-sample batches of increasing number of samples, each sample containing variants drawn at random from ClinVar. Runtimes were measured both with and without PostgreSQL-backed annotation retrieval, excluding HPC queue waiting time in order to capture the only effective processing time. All analyses were performed on the IEO High-performance computing (HPC) infrastructure, with concurrency capped at 100 processes to match the chunk-based parallelisation strategy implemented in varianTalker. Each dataset or sample batch was analysed five times to assess reproducibility.

In the absence of PostgreSQL-backed retrieval, runtime increased sublinearly with the number of input variants, as shown in Fig. 1A. For datasets containing fewer than 100,000 variants, active workflow runtime (WFR) remained largely stable, indicating that the available parallelization capacity was sufficient to absorb the increase in number of variants. This behaviour reflected the Nextflow-based execution model, in which variants are partitioned into chunks and processed concurrently. Accordingly, datasets below 100,000 variants were annotated in approximately 20 min, with CPU usage below 7 CPU-hours, whereas larger datasets required more than 1 h, with CPU usage exceeding 27 CPU-hours.

We next assessed PostgreSQL-backed retrieval under an ideal reuse scenario in which all queried variants were already present in the local database. Under these conditions, varianTalker bypassed *de novo* execution of annotation tools and directly retrieved previously stored annotation records. Across ClinVar-derived benchmarks, PostgreSQL mode reduced WFR by an average of 92% (range, 90–95%). In this configuration, the 100,000-variant dataset was processed in approximately 2 min, and the complete ClinVar database in less than 1 h. CPU-hours usage showed an even larger reduction, decreasing by an average of 95% (range, 89–98%) across all tests (Fig. 1A).

A similar pattern was observed when increasing the number of simulated samples analysed simultaneously, as shown in Fig. 1B. Although the total number of variants increased with sample number, runtime did not increase proportionally because samples were processed in parallel. varianTalker annotated 10 samples in approximately 6 min and 100 samples in approximately 34 minutes, compared with an estimated 60 min under sequential execution, corresponding to a 43.3% reduction in WFR. Moreover, PostgreSQL-backed retrieval reduced both WFR and CPU-hours usage by an average > 96% across all tested sample batches (range of WFR, 96–97.0%; range of CPU-hours, 92–99%). Additional resource-usage metrics, including memory consumption, disk read and disk write, followed similar trends and are reported in Supplementary Figure S1. PostgreSQL query performance remained stable even as the size of the stored variant collection increased (Supplementary Figure S2), indicating that the database architecture remains scalable for long-term institutional use and repeated retrospective analyses.

Together, these results define the expected upper and lower runtime bounds for institutional use, ranging from complete *de novo* annotation to full reuse of previously annotated variants. They also show that building and maintaining a local PostgreSQL annotation database can substantially accelerate routine sample processing, reduce computational cost and create capacity for additional downstream workflows, such as retrospective re-annotation, advanced filtering or interactive clinical review.

Integrated annotation expands the available evidence for variants annotated as uncertain significance in ClinVar

Having established that varianTalker can process institutional-scale data efficiently, we next evaluated its ability to combine complementary evidence sources into a unified annotation framework. Two distinct situations were considered: absence of annotation, in which a tool provided no information for a variant, and uncertain classification, in which a tool explicitly classified the variant as uncertain significance (VUS). Because a VUS does not provide a definitive pathogenicity assessment, evidence from additional tools provides complementary information to better assess its clinical interpretation.

Analysis of the IEO cohorts showed substantial differences in annotation coverage among the integrated tools (Fig. 2A). In silico pathogenicity predictors returned a score for 41.1% (484,988/1,181,451) of germline variants and 26.5% (24,257/91,569) of somatic variants. Database-derived resources also showed incomplete coverage: ClinVar returned a clinical significance assertion for 30.6% (361,144/1,181,451) of germline and 50.6% (46,307/91,569) of somatic variants, of which 47.4% and 2.7% were benign or likely benign and VUS, respectively.

The guideline or machine learning-based interpretation tools achieved complete coverage: InterVar, ReNOVo and CancerVar returned a classification for every variant. Of these, 99.0% (1,169,723/1,181,451) and 98.1% (1,158,547/1,181,451) of germline variants were classified as benign or likely benign by InterVar and ReNOVo respectively, and 83.3% (76,238/91,569) of somatic variants by CancerVar. Nevertheless, among the variants evaluated by these tools, approximately 1.0% of germline and 16.4% of somatic variants remained classified as uncertain (Fig. 2A).

Complete coverage, therefore, does not mean complete interpretation, and the variants left uncertain are the ones that most need it, because a variant without a definitive call reaches the molecular tumour board with no actionability assessment. We next assessed whether evidence integration could provide an informative classification for variants that remained uncertain according to an individual guideline or machine learning-based tool (Fig. 2B). A variant was considered to have gained additional evidence when at least one integrated source provided a non-VUS classification (which does not per se constitute a validated reclassification). In the germline cohort, the proportion of variants remaining unresolved decreased from 1.0% (according to InterVar) to 0.1%, indicating that complementary evidence was available for 86.9% (9,806/11,284) of the variants initially classified as VUS. In the somatic cohort, the proportion of unresolved variants decreased from 16.4% (according to CancerVar) to 8.6%, corresponding to additional informative evidence for 47.1% (7,055/14,975) of the variants initially classified as uncertain. Thus, the integrated

framework substantially increased the number of uncertain variants for which additional classifying evidence became available, particularly in the germline setting.

To further evaluate the contribution of evidence integration, we examined ClinVar VUS that were assigned an alternative classification by InterVar or CancerVar, as shown in Fig. 3. All ClinVar VUS with pathogenic orthogonal evidence and a similar random subset of variants with benign orthogonal evidence were included in the analysis. These classifications were manually compared with the corresponding VarSome interpretations for germline and somatic variants.

For germline variants, InterVar classifications showed high concordance with VarSome (Fig. 3A). Among variants classified by InterVar as pathogenic or likely pathogenic, 90% (168/186) were also classified as pathogenic or likely pathogenic by VarSome, 8.1% (15/186) remained uncertain, and 2% (3/186), were classified as benign or likely benign. Among the selected variants classified by InterVar as benign, 87.6% (163/186) were classified as benign or likely benign in VarSome, 10.2% (19/186) remained uncertain, and 2.2% (4/186), were classified as likely pathogenic.

For somatic variants, among variants classified by CancerVar as Tier I or Tier II, 30.1% (34/113) were assigned to Tier I or Tier II by VarSome, 15.9% (18/113) were assigned to Tier III, and 54.0% (61/113) were assigned to Tier IV (Fig. 3B). Conversely, among variants classified by CancerVar as Tier IV, 91.2% (103/113) were also classified as Tier IV by VarSome, whereas the remaining 8.8% (10/113) were classified as Tier III. No CancerVar Tier IV variant was assigned to Tier I or Tier II by VarSome.

Agreement with VarSome was even stronger for InterVar and CancerVar VUS that received an alternative classification by ClinVar (Supplementary Figure S3). For germline variants, only 1 of 35 (2.9%) pathogenic or likely pathogenic variants was classified as likely benign by VarSome, and one pathogenic, one likely benign and one benign variant were considered uncertain. For somatic variants, 89,2% (50/56) of ClinVar Tier IV variants agreed with VarSome, whereas 26.8% of Tier I/II variants received the same tier and 69.6% were assigned to Tier III, with 2 of 56 (3.6%) assigned to Tier IV.

Overall, the comparison indicates that evidence from the integrated annotation framework can provide additional information for ClinVar VUS.

Configurable filtering supports clinical prioritization

Unfiltered annotation output is not directly usable in clinical review. A single case yields thousands to tens of thousands of annotated variants. Manual inspection at this scale is not feasible within diagnostic reporting timelines. Clinical review, therefore, depends on a prioritisation step, which must remove variants with no interpretative value while retaining every variant capable of changing patient management. varianTalker implements this step as a configurable filtering layer applied to the annotation fields produced by the upstream modules.

We applied varianTalker to the retrospective IEO cohorts (100 patients from GerSom project) using filter configurations representative of clinical variant review. The configurations were defined in consultation with clinical experts and are reported in Table 4. We used separate configurations for germline and somatic

analyses to reflect their different interpretation requirements. The two configurations differed in their treatment of variants of uncertain significance. The germline configuration retained them, because such variants may warrant segregation analysis, functional investigation, or reinterpretation as evidence accumulates. The somatic configuration excluded variants classified by ClinVar as uncertain significance or uncertain risk allele, and variants assigned to CancerVar Tier IV. It retained variants annotated as conflicting classifications of pathogenicity, since this category reflects disagreement among submitters and may include a pathogenic or likely pathogenic assertion.

We assessed the filtering layer against two criteria: i) reduction of the number of variants presented for review; ii) retention of variants judged clinically relevant by independent expert review of the unfiltered output.

Before prioritisation, individual samples contained a mean of 1,851 annotated germline variants and 1,780 annotated somatic variants requiring potential manual inspection. varianTalker substantially reduced the number of variants retained in the PASS output. A mean of 56 variants per germline sample and 16 variants per somatic sample were retained for downstream review (Fig. 4A).

Table 4

– **Filtering configurations applied to the retrospective IEO cohorts.** The filtering criteria differed between germline and somatic analyses to reflect their distinct interpretation requirements. Pathogenic or likely pathogenic classifications overrode consequence-based and technical thresholds in both settings. Germline filtering also retained variants of uncertain significance for possible clinical assessment or future reinterpretation. In the somatic setting, explicit uncertain ClinVar classifications and CancerVar Tier IV variants were excluded, whereas variants with conflicting ClinVar classifications were retained because they contained discordant submitted interpretations.

Parameters	Filter	Logic
Variant type exclusion	“Silent”, “Intron”, “3’UTR”, “5’UTR”, “IGR”, “5’Flank”, “3’Flank”, “RNA”	AND (overridden if any clinical annotation classifies the variant as (likely)pathogenic)
Minimum coverage	50× (50)	AND (overridden if any clinical annotation classifies the variant as (likely)pathogenic)
Minimum somatic VAF	1% (0.01)	AND (overridden if any clinical annotation classifies the variant as (likely)pathogenic)
Minimum germline VAF	20% (0.20)	AND (overridden if any clinical annotation classifies the variant as (likely)pathogenic)
CancerVar tier	Tier_I_strong, Tier_II_potential	OR (somatic only)
InterVar class	Pathogenic, Likely pathogenic, Uncertain significance	OR (germline only)
ClinVar (for germline)	Pathogenic, Likely_pathogenic, Uncertain_significance, Conflicting_classifications_of_pathogenicity, Uncertain_risk_allele, Likely_risk_allele, Pathogenic,_low_penetrance, Established_risk_allele	OR (germline only)
ClinVar (for somatic)	Pathogenic, Likely_pathogenic, Conflicting_classifications_of_pathogenicity, Likely_risk_allele, Pathogenic,_low_penetrance, Established_risk_allele	OR (somatic only)
ReNOVo class	Intermediate or High Probability Pathogenic (IP, HP)	OR (germline only)
ESCAT	IA, IB, IC, IIA, IIB, IIIA, IIIB, IV	OR (somatic only)
Gene filter	No gene list	AND

The remaining variants were assigned to the NOPASS output, corresponding to a mean of 1,795 germline variants and 1,764 somatic variants per sample. The filtering procedure therefore removed 97.0% of germline variants and 99.1% of somatic variants from the primary review set. In germline samples, pathogenic or likely pathogenic variants represented 17.9% of the PASS output, corresponding to

approximately 10 variants per sample. A further 81.8%, corresponding to approximately 46 variants per sample, were classified as variants of uncertain significance (Fig. 4B). This high representation was expected, since the germline configuration was designed to retain uncertain variants that may be relevant for clinical review, additional investigation, or future reinterpretation. The remaining 0.3% were classified as benign or likely benign by ClinVar or InterVar. varianTalker retained these variants because ReNOVo assigned them a pathogenic prediction. In somatic samples, pathogenic or likely pathogenic variants, including CancerVar Tier I and Tier II variants, accounted for 48.1% of the PASS output, corresponding to approximately eight variants per sample. The remaining 51.9%, also approximately eight variants per sample, carried non-definitive or discordant classifications. This group comprised variants with conflicting ClinVar classifications and variants assigned to CancerVar Tier III (Fig. 4B).

We then assessed whether the filtering layer preserved variants required for clinical reporting (Fig. 4C). Independent expert review of the unfiltered output identified 582 germline and 174 somatic variants as clinically relevant across the cohort. varianTalker retained all these variants in the PASS output, corresponding to a recall of 100% for both sample types. No variant judged reportable by expert review was assigned to the NOPASS output. Filtering also increased the proportion of clinically actionable variants presented for review. In germline samples, pathogenic or likely pathogenic variants represented 0.6% of the unfiltered variant set and 20.7% of the PASS output, an enrichment of 34-fold (Supplementary Figure S4). In somatic samples, the corresponding proportions were 0.2% before filtering and 21.6% after filtering, an enrichment of 108-fold. The filtering layer therefore concentrated the clinically relevant fraction of the variant set without discarding any variant required for the report.

Overall, varianTalker removed 97.0% germline variants and 99.1% somatic variants from the primary review set while retaining every variant identified as clinically relevant by expert review. The distinct germline and somatic configurations preserved different categories of potentially relevant variants, supporting both focused clinical prioritisation and broader exploration of uncertain or discordant findings.

Discussion

Clinical interpretation of genomic variants requires the integration of multiple evidence dimensions, including molecular consequence, germline pathogenicity, somatic clinical significance, therapeutic actionability and pharmacogenomic relevance, which remain distributed across independent tools and databases with different coverage, classification systems and updating schedules[2,3]. varianTalker was developed to address this fragmentation by providing a unified, institution-oriented framework that combines complementary annotation resources, supports continuous knowledgebase updates and facilitates interactive review of genomic findings.

A strength of varianTalker is the integration of heterogeneous evidence sources within a single reproducible workflow. By combining CancerVar, InterVar, ClinVar, CIViC, PharmGKB, AlphaMissense and other functional and clinical annotations, the framework simultaneously integrates evidence spanning somatic interpretation, germline pathogenicity assessment, therapeutic actionability, and pharmacogenomics. The results of the analysis from the IEO cohorts illustrate why such integration is necessary. No individual tool

provided complete information for all variants, and annotation coverage varied substantially according to the type of evidence and the germline or somatic context. Consequently, the absence of information from one source did not necessarily imply that a variant lacked relevant evidence, because another integrated resource could still provide a clinically meaningful or computationally informative annotation.

The implementation of varianTalker in Nextflow DSL2, together with containerised execution, supports portability and reproducibility across computational environments. Its modular architecture also enables the workflow to process samples generated by heterogeneous sequencing and variant-calling pipelines while maintaining a consistent downstream annotation strategy. In addition, MANE-based canonical transcript selection improves the consistency of variant consequence reporting across samples, platforms, and repeated analyses. These characteristics are particularly important in institutional settings, where sequencing data may originate from different technologies and undergo reanalysis as pipelines and knowledgebases evolve.

A further distinctive feature of varianTalker is its support for institutional knowledge management through PostgreSQL-backed annotation storage and explicit database versioning. The analyses of the performance showed that the workflow can parallelise annotation tasks efficiently as the number of variants and samples increases. Reuse of previously stored annotations reduced analysis time by an average of 92% when queried variants had already been annotated, by avoiding repeated execution of the complete annotation workflow for recurrent variants. This architecture is especially relevant in clinical and translational environments, in which the same variants may recur across patients, panels, or retrospective cohorts. Beyond improving computational efficiency, structured storage of annotated variants creates the basis for retrospective identification, systematic re-querying, and comparison of interpretations generated using different database versions.

The evidence-integration analyses also showed that evidence integration increases the proportion of variants for which an informative interpretation can be obtained compared with the use of individual tools in isolation. This result should not be interpreted as evidence that the framework definitively resolves every uncertain variant. Instead, it demonstrates that variants lacking a conclusive classification in one resource may receive additional evidence from another. This distinction was particularly important for variants classified as VUS. In the germline cohort, integration reduced the proportion of variants remaining unresolved from 1.0% to 0.1%, indicating that complementary evidence was available for 86.9% of the variants initially classified as uncertain. In the somatic cohort, the unresolved proportion decreased from 16.4% to 8.6%, corresponding to additional informative evidence for 47.1% of initially uncertain variants.

The comparison with VarSome supported the consistency of the germline classifications by variantTalker. Indeed, most ClinVar variants classified as VUS resolved by InterVar received a concordant call. Agreement was lower for somatic variants. A substantial proportion of CancerVar Tier I and Tier II variants were assigned to Tier IV by VarSome, whereas CancerVar Tier IV variants showed high agreement and were not reassigned to upper tiering. Because somatic tier assignment depends on tumour type, biomarker indication, drug availability and the level of clinical evidence, this asymmetry more likely reflects differences in the knowledgebases and criteria of the two platforms than in an analytical error. In both settings,

integrated evidence can support prioritisation and reinterpretation without substituting for expert assessment.

The configurable filtering layer further reduced the interpretation burden by transforming large annotated variant sets into manageable review lists of tens of variants per sample, corresponding to a reduction of more than 96% in the germline cohort and more than 99% in the somatic cohort. A prioritisation layer is only clinically usable if it loses no relevant variant while reducing the review burden. Expert classification identified 582 pathogenic germline variants and 174 pathogenic somatic variants across the 50 germline and 50 somatic samples assessed. All these variants were present in the PASS output. Filtering raised the pathogenic fraction of the reviewed set from 0.6% to 20.7% in germline samples and from 0.2% to 21.6% in somatic samples, corresponding to enrichments of 34 and 108-fold, respectively. The reduction in review burden was, therefore, achieved without any loss of the findings that determine clinical reporting, condition that any prioritisation strategy intended for diagnostic use needs to satisfy.

The filtering results suggest that varianTalker can substantially reduce the number of variants requiring primary manual review while preserving access to variants that satisfy at least one predefined clinical or research-oriented criterion. Nevertheless, the PASS output should be interpreted as a prioritised review set rather than a collection of definitively clinically relevant variants. Review of the prioritised set is supported by MAFigate, which runs locally so that patient data remain within the institutional environment, and which allows reviewers to adjust filter thresholds interactively and record per-variant notes that are retained in the exported tables. The flexibility of the filtering system allows institutions to modify thresholds, retained classifications, gene lists, and override rules according to the intended use: focused diagnostic review, molecular tumour board preparation, retrospective research, or exploratory analysis. This flexible configurability represents an advantage over fixed prioritisation systems but also requires careful documentation and validation of the selected parameters.

Several limitations should be acknowledged. First, varianTalker currently focuses mainly on small sequence variants and short insertions and deletions, whereas structural variants, and protein fusions remain outside the core workflow. Nevertheless, the modular nature of varianTalker can support a future extension of the pipeline to the annotation of additional variant classes and alterations. Second, annotation quality remains dependent on the coverage, update frequency, and evidence curation practices of the integrated resources. Conflicting or outdated classifications may therefore propagate into the combined output if they are not recognised during expert review. Explicit database versioning and retention of source-specific annotations partially address this limitation by allowing results to be traced and re-evaluated when knowledgebases are updated. Unlike platforms that distribute annotation resources as a fixed bundle tied to each software release, varianTalker decouples resource updates from the pipeline version, so that institutions can refresh knowledgebases as they are published. Supplementary Table S1 compares this and other capabilities with those of established open-source interpretation platforms. However, automated integration cannot replace critical assessment of the underlying evidence. Third, the validation was performed using retrospective cohorts from a single institution. Although the cohorts included different sequencing and upstream analysis platforms, further evaluation of varianTalker in independent institutions, additional tumour types, and larger prospective datasets will be necessary to assess generalisability and clinical impact. The retention analysis

in particular rests on fifty germline and fifty somatic samples, so the recall estimate is subject to the precision limits of a cohort of that size. Future studies should also evaluate whether the reduction in review burden translates into shorter interpretation times, improved inter-reviewer consistency, or changes in clinically reported findings.

Overall, varianTalker provides an open, reproducible, and extensible infrastructure for precision-oncology variant annotation and prioritisation. Rather than replacing expert interpretation, the framework is designed to support molecular tumour boards, diagnostic laboratories, and translational genomics programmes by reducing manual evidence integration, enabling consistent analysis across heterogeneous sequencing workflows, and facilitating retrospective reanalysis as knowledge evolves. The results indicate that combining complementary annotation sources with configurable, context-specific filtering can improve evidence availability and substantially reduce the number of variants requiring primary review, while preserving clinically evaluable findings.

Conclusion

varianTalker provides somatic interpretation, germline pathogenicity assessment, therapeutic actionability and pharmacogenomics into a single containerised Nextflow DSL2 workflow, coupled to versioned institutional storage of annotated variants and a configurable prioritisation layer with local interactive review. In retrospective germline and somatic cohorts, this design increased the proportion of variants for which informative evidence was available, compressed review lists by more than 90% respectively without losing any expert-classified pathogenic variant, and reduced runtime for recurrent variants through annotation reuse. These properties match the practical constraints of institutional precision oncology, where data arrive from heterogeneous pipelines, the same variants recur across patients, and cases are periodically reinterpreted as knowledgebases evolve. varianTalker is an infrastructure that can help the expert the variant assessment reducing time consume and improve the consistence. Its modular design allows future extensions to annotate structural variants and fusions and for prospective, multi-institutional evaluation.

Abbreviations

ACMG
American College of Medical Genetics and Genomics
AMP
Association for Molecular Pathology
ASCO
American Society of Clinical Oncology
CADD
Combined Annotation Dependent Depletion
CAP
College of American Pathologists
CPSR

Cancer Predisposition Sequencing Reporter
CPU
central processing unit
ESCAT
ESMO Scale for Clinical Actionability of molecular Targets
ESMO
European Society for Medical Oncology
ExAC
Exome Aggregation Consortium
GATK
Genome Analysis Toolkit
HPC
high-performance computing
IEO
European Institute of Oncology
IGR
intergenic region
MAF
mutation annotation format
MANE
Matched Annotation from NCBI and EMBL-EBI Select
MSK-IMPACT
Memorial Sloan Kettering–Integrated Mutation Profiling of Actionable Cancer Targets
PCGR
Personal Cancer Genome Reporter
SNV
single nucleotide variant
TSV
tab-separated values
UTR
untranslated region
VEP
Ensembl Variant Effect Predictor
VUS
variant of uncertain significance
WFR
workflow runtime

Declarations

Ethics approval and consent to participate

This study included retrospective analyses of sequencing data from patients treated at the European Institute of Oncology (IEO), Milan, Italy. The use of clinical sequencing data for research purposes was approved by the Ethics Committee with ID 2877 NCT: NCT07555639 Prot. N.0050128/19, 22/11/2019, and R1274/20-IEO 1338, 22/07/2020. All data used in this study were processed and analysed in accordance with institutional policies governing the use of patient-derived genomic data.

Consent for publication

Not applicable. This manuscript does not contain identifiable individual patient data.

Availability of data and materials

varianTalker is available at https://github.com/zhanyinx/variantalker, together with the source code and documentation required for installation and use. The resources bundles used in the analyses are publicly available from https://bioserver.ieo.it/repo/dima/hg38 and https://bioserver.ieo.it/repo/dima/hg19. The ClinVar-derived artificial datasets used for benchmarking were generated from publicly available ClinVar data from https://www.ncbi.nlm.nih.gov/clinvar/ and can be reproduced using the procedures described in the Methods section. The clinical sequencing datasets used for retrospective validation were obtained from patients treated at the European Institute of Oncology (IEO) are not publicly available.

Competing Interests

The authors declare no competing interests.

Funding

V.F. is funded by the SAMET project, realized with the contribution of the Regional Programme of the European Regional Development Fund (PR FESR 2021–2027) of Regione Lombardia, call "COLLABORA E INNOVA" (Project ID: 6042185; CUP: E49I25000990007). M.B. is supported by Alleanza Contro il Cancro (ACC) within the Health Big Data (HBD) project. We thank the IT Department for technical support, including computational resources, server infrastructure, and PostgreSQL database administration.

Author’s contribution

Y.Z. conceived and designed the study. Y.Z., G.A.M. developed varianTalker and implemented the computational workflow. G.A.M., M.B., V.F. performed the bioinformatic and statistical analyses. E.D, C.T., R.D.M, C.N., P.G.P., L.M., G.F., C.R. and E.B. curated and provided the clinical sequencing datasets. Y.Z., G.A.M., M.B. contributed to the validation and interpretation of the results. Y.Z. supervised the study. Y.Z. G.A.M, M.B. wrote the original draft of the manuscript. All authors contributed to the review and editing of the manuscript, read and approved the final version, and agreed to be accountable for their contributions.

Acknowledgements

The authors thank the IT Department of the European Institute of Oncology (IEO) for technical support, including the provision and maintenance of computational resources and server infrastructure. The authors

also thank DIMA group for their contribution to clinical data collection and sequencing workflow. We thank all the supporters of IEO through the 5 × 1000 financing.

References

1. Boos L, Wicki A. The molecular tumor board—a key element of precision oncology. Memo. 2024;17:190–3.
2. Jäger N. Bioinformatics workflows for clinical applications in precision oncology. Semin Cancer Biol. 2022;84:103–12.
3. Krysiak K, et al. A community approach to the cancer-variant-interpretation bottleneck. Nat Cancer. 2022;3:522–5.
4. Li MM, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: A joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn. 2017;19:4–23.
5. Richards S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17:405–24.
6. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
7. McLaren W, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.
8. Landrum MJ, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46:D1062–7.
9. Griffith M, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49:170–4.
10. Chakravarty D et al. OncoKB: A precision oncology knowledge base. *JCO Precis. Oncol.* 2017, 1–16 (2017).
11. Li Q, Wang K, InterVar. Clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. Am J Hum Genet. 2017;100:267–80.
12. Li Q, et al. CancerVar: An artificial intelligence-empowered platform for clinical interpretation of somatic mutations in cancer. Sci Adv. 2022;8:eabj1624.
13. Tsimberidou AM, et al. Molecular tumour boards - current and future considerations for precision oncology. Nat Rev Clin Oncol. 2023;20:843–63.
14. Nakken S, et al. Personal Cancer Genome Reporter: variant interpretation report for precision oncology. Bioinformatics. 2018;34:1778–80.
15. Pagel KA, et al. Integrated informatics analysis of cancer-related variants. JCO Clin Cancer Inf. 2020;4:310–7.
16. Di Tommaso P, et al. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.
17. Danecek P et al. Twelve years of SAMtools and BCFtools. *Gigascience* 10, giab008 (2021).

18. Jantus-Lewintre E, et al. Multicenter in-house evaluation of an amplicon-based next-generation sequencing panel for comprehensive molecular profiling. Mol Diagn Ther. 2025;29:249–61.
19. Hanssen F, et al. Scalable and efficient DNA sequencing analysis on different compute infrastructures aiding variant discovery. NAR Genom Bioinform. 2024;6:lqae031.
20. Behera S, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. Nat Biotechnol. 2025;43:1177–91.
21. Van der Auwera GA, O'Connor B. D. Genomics in the Cloud. O'Reilly Media, Inc.; 2020.
22. Morales J, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022;604:310–5.
23. Goldfarb T, et al. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. Nucleic Acids Res. 2025;53:D243–57.
24. Mudge JM, et al. GENCODE 2025: reference gene annotation for human and mouse. Nucleic Acids Res. 2025;53:D966–75.
25. Cheng J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381:eadg7492.
26. Favalli V, et al. Machine learning-based reclassification of germline variants of unknown significance: The RENOVO algorithm. Am J Hum Genet. 2021;108:682–95.
27. Whirl-Carrillo M, et al. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2021;110:563–72.
28. Mateo J, et al. A framework to rank genomic alterations as targets for cancer precision medicine: the ESMO Scale for Clinical Actionability of molecular Targets (ESCAT). Ann Oncol. 2018;29:1895–902.
29. Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
30. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
31. Fu W, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493:216–20.
32. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.
33. Cheng DT, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn. 2015;17:251–64.
34. Ptashkin RN, et al. Enhanced clinical assessment of hematologic malignancies through routine paired tumor and normal sequencing. Nat Commun. 2023;14:6895.
35. Milbury CA, et al. Clinical and analytical validation of FoundationOne®CDx, a comprehensive genomic profiling assay for solid tumors. PLoS ONE. 2022;17:e0264138.
36. He J, et al. Integrated genomic DNA/RNA profiling of hematologic malignancies in the clinical setting. Blood. 2016;127:3004–14.

37. Sondka Z, et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18:696–705.
38. Suehnholz SP, et al. Quantifying the expanding landscape of clinical actionability for patients with cancer. Cancer Discov. 2024;14:49–65.
39. Ioannidis NM, et al. REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99:877–85.
40. Feng B-J, PERCH. A unified framework for disease gene prioritization: Human mutation. Hum Mutat. 2017;38:243–51.
41. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14(Suppl 3):3.
42. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94.
43. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.
44. Kopanos C, et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35:1978–80.
45. Rehm HL, et al. ClinGen–the clinical genome resource. N Engl J Med. 2015;372:2235–42.
46. Esposito A, et al. The safety and suitability of DNA sequencing of tissue biopsies performed on patients referred to a phase I Unit. Cancers (Basel). 2024;16:4252.
47. Garcia M, et al. A portable workflow for whole-genome sequencing analysis of germline and somatic variants. F1000Res. 2020;9:63. Sarek.

Figures

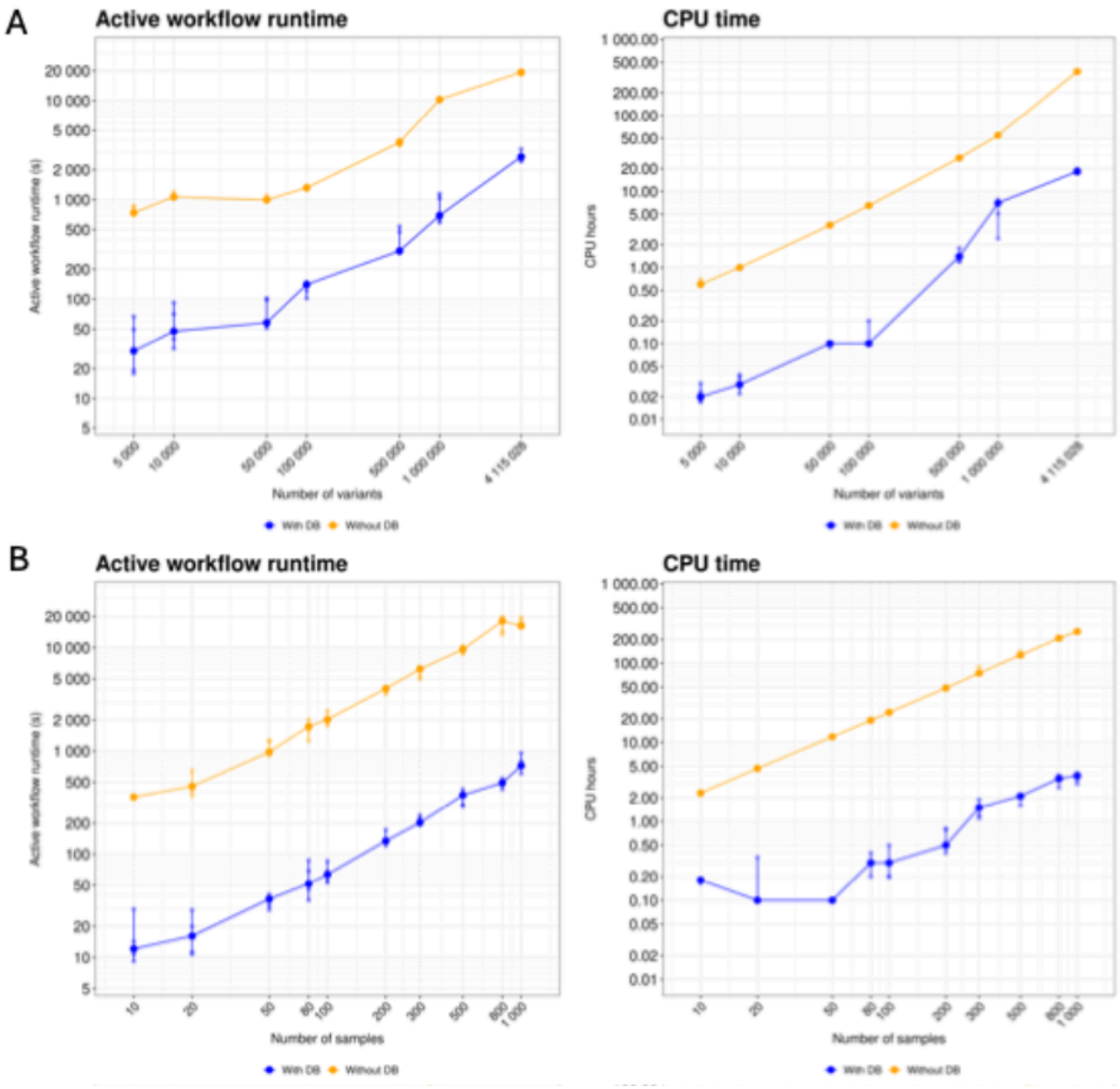


Figure 1

Runtime of varianTalker with and without PostgreSQL-backed retrieval. Line plots show wall-clock time (left panel) and CPU-hours (right panel) for de novo annotation without PostgreSQL retrieval (orange) and PostgreSQL-backed retrieval of previously stored annotations (blue). (A) Performance across ClinVar-derived benchmark datasets of increasing number of variants. (B) Performance across simulated multi-sample analyses of increasing number of samples. Dots represent the median of five independent runs. The error bar represents the range of the values, from minimum to maximum, across the 5 replicates for each run.

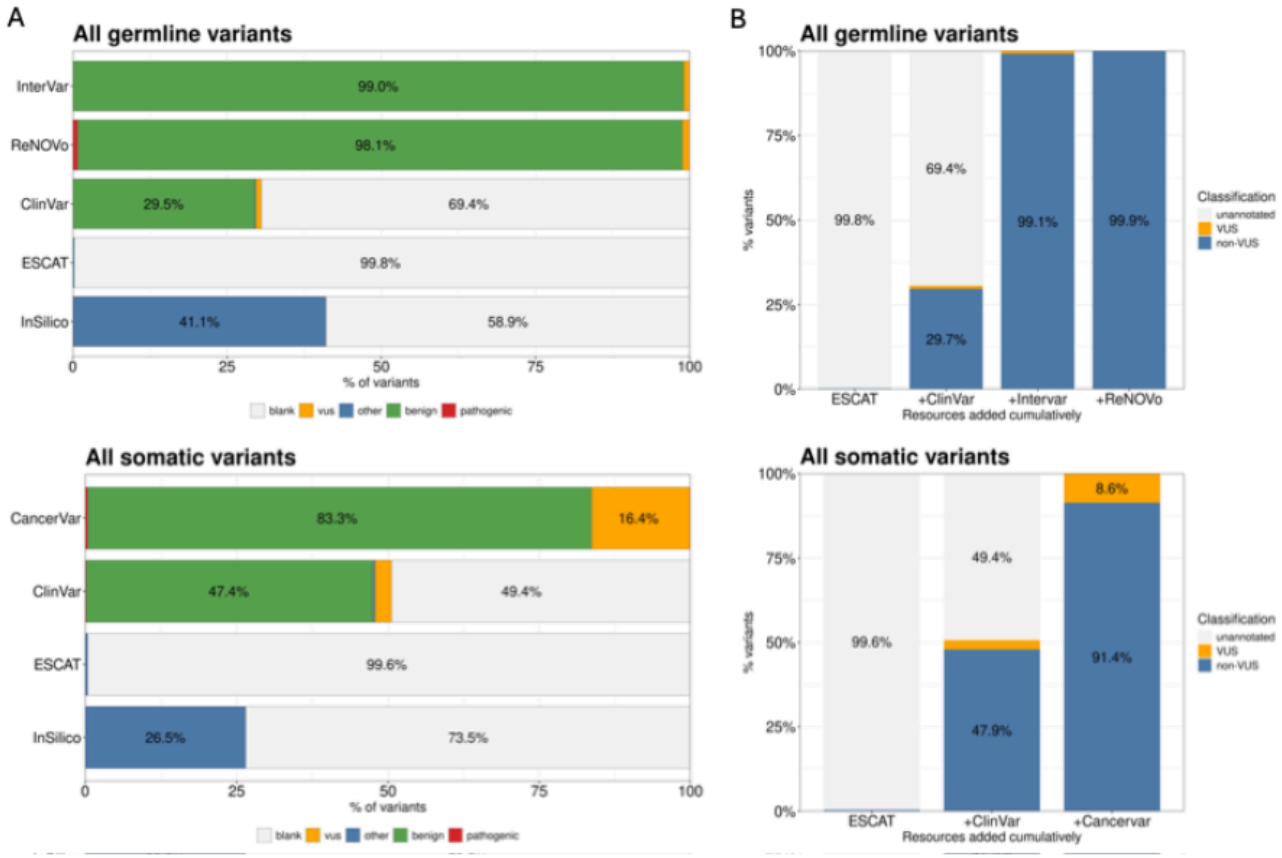


Figure 2

Annotation coverage and evidence integration. (A) Proportion of germline and somatic variants for which each annotation source provided information. Missing annotations (blank) indicate variants for which the corresponding tool returned no result and are distinct from variants explicitly classified as uncertain (VUS, yellow). (**B**) Proportion of variants initially classified as VUS that remained VUS or received a non-VUS classification from at least one additional source integrated by varianTalker. Evidence integration reduced the VUS fraction from 1.0% to 0.1% in germline variants and from 16.4% to 8.6% in somatic variants.

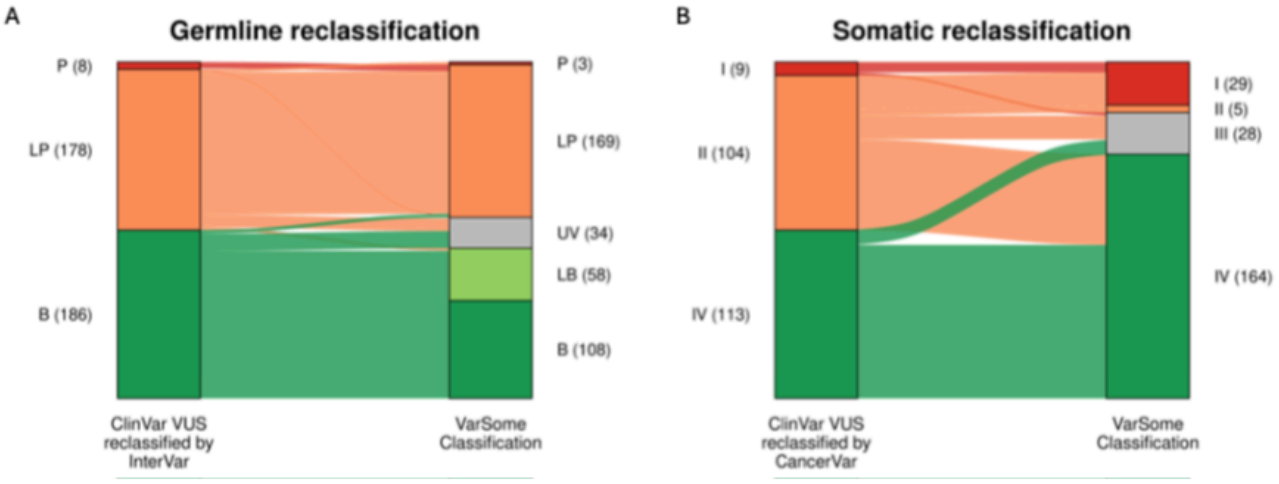


Figure 3

Reclassification of variants by varianTalker. Alluvial plots showing the correspondence between classifications assigned by the rule-based tools integrated into varianTalker and those reported by VarSome. **(A)** Germline ClinVar VUS classified by InterVar as pathogenic/likely pathogenic or benign/likely benign and their corresponding VarSome classifications. **(B)**Somatic ClinVar VUS classified by CancerVar as Tier I/II or Tier IV and their corresponding VarSome tier assignments. Flow widths represent the number of variants in each classification transition.

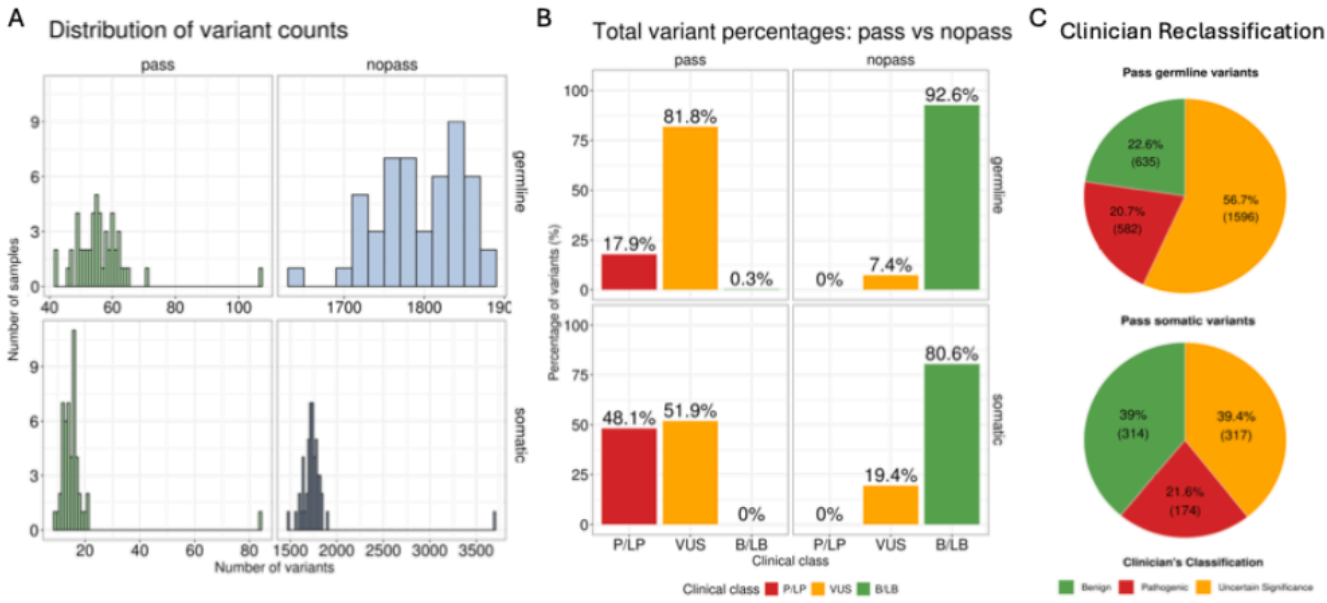


Figure 4

Effect of configurable filtering on variant prioritization in the IEO cohorts. (**A**) Distribution of the number of variants assigned to the PASS and NOPASS outputs in germline (top) and somatic samples (bottom). (B) Distribution of clinical classification categories in the PASS and NOPASS outputs, based on the highest-priority classification provided by ClinVar and InterVar for germline variants or by ClinVar and CancerVar for somatic variants. (C) Distribution of expert classification categories in the PASS output and in the complete annotated variant set, shown for germline and somatic samples. Classifications in these panels were assigned by expert review and are independent from varianTalker calls shown in (B).

Supplementary Files

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

- supplementaryinformationvariantalker.pdf