AstraKit: Customizable, reproducible workflows for biomedical research and precision medicine

Nadine S. Kurz[1,2,*], Kevin Kornrumpf[1], Martin Stoves[1], and Jürgen Dönitz[1,2,3]

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, 37075 Göttingen, Niedersachsen, Germany
[2]Comprehensive Cancer Center Göttingen (G-CCC), Göttingen, 37075, Germany
[3]Campus-Institut Data Science (CIDAS), Section Medical Data Science (MeDaS), 37075 Göttingen, Niedersachsen, Germany

Abstract

Motivation: Fragmented bioinformatics tools compels researchers and clinicians to resort to error-prone manual pipelines. The success of precision medicine depends largely on the efficient interpretation of genetic variants and the selection of highly effective targeted therapies. However, the success of precision medicine and biomedical research depends on the availability of efficient software solutions for processing and interpreting genetic variants, interpreting multi-omics data, and integrating drug screen analyses. **Results:** We present AstraKit, a unified KNIME workflow suite enabling end-to-end precision medicine analytics. AstraKit introduces three transformative innovations: 1) Dynamic variant interpretation with customizable annotation and filtering for disease-specific genomic contexts; 2) Multi-layered omics analyses integrating genomic, transcriptomic, and epigenetic data; and 3) Translational drug matching that correlates in vitro drug screens with clinical outcomes. Validated across oncology cohorts, AstraKit demonstrates concordance between experimental drug sensitivity and clinical trial responses, resolving discordances to uncover resistance mechanisms. By unifying variant analysis, multi-omics, and drug response modeling on a single customizable platform, AstraKit eliminates siloed workflows, accelerating biomarker validation and enabling clinicians to directly link molecular profiles to therapeutic decisions. As all AstraKit workflows are open-source and platform-independent, we provide a versatile comprehensive software suite for a multitude of tasks in bioinformatics and precision medicine.
Availability and implementation: The KNIME workflows are available at KNIME Hub https://hub.knime.com/bioinf_goe/spaces/Public/AstraKit~lfVsGBY2HnPYc1h1/. The source code is available at https://gitlab.gwdg.de/MedBioinf/mtb/astrakit.
Keywords: Variant interpretation, Multi-Omics integration, Drug screen analysis, Precision medicine, Automated workflows

1 Introduction

Precision medicine demands robust computational frameworks to transform heterogeneous molecular data into clinically actionable insights. While next-generation sequencing (NGS) and high-throughput drug screening have exponentially expanded our capacity to generate genomic, transcriptomic, epigenomic, and functional drug response data, the fragmentation of computational tools and data formats remains a critical barrier to reproducible analysis. Software packages typically operate in silos: variant calling (GATK [2]), genomic operations (samtools [56], bedtools [54]), annotation (ANNOVAR [10], VEP [12], Onkopus [46]), and filtering (SNPSift [6], AdaGenes [23]) address isolated steps in the analytical continuum. Consequently, researchers must manually chain tools across incompatible data formats (VCF [24], MAF, BED), perform auxiliary tasks like genome assembly conversion (LiftOver), and implement custom filtering criteria—compromising traceability, reproducibility, and accessibility for non-

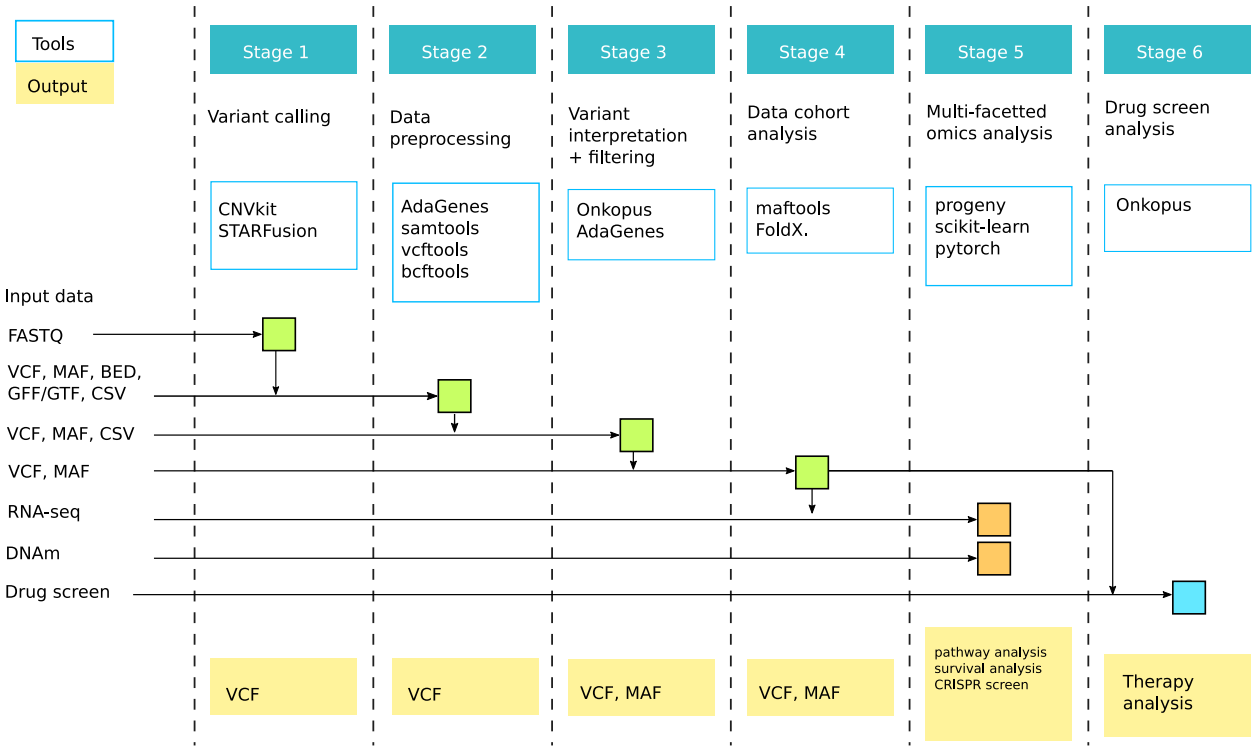


Figure 1: Overview of the AstraKit scope. AstraKit is able to preprocess mutational data in various formats, including VCF, MAD, BED, GFF/GTV, and CSV, for file format conversions or genome assembly LiftOver. The variant interpretation workflows then allow to create customized variant annotation and interpretation workflows. Variant cohort analysis is enabled by tools like maftools, and the analysis of specific variant features. Further omics data can be included in multi-layered omics analysis workflows, and the validation of experiment drug sensitivity data through drug screen analysis workflows.

programmers. This fragmentation is particularly acute in two translational domains: First, cohort-level variant analysis, which is essential for identifying recurrent mutations and clinical associations, requires specialized tools like maftools, yet lacks seamless integration with upstream variant processing pipelines. The joint analysis of various omics data, including transcriptomics and epigenetics, places further demands on data integration. Second, while clinical interpretation tools (e.g., Onkopus [46], Pandrugs [15]) predict targeted therapies from mutational profiles, and drug screen platforms (e.g., Breeze 2.0 [31]) enable in vitro data exploration, no existing framework bridges clinical genomic classifications with functional drug response validation. Current solutions fail to systematically correlate patient-derived biomarkers (e.g., mutation status) with high-throughput drug sensitivity metrics (IC_{50}, growth inhibition), hindering biomarker discovery and therapeutic personalization.

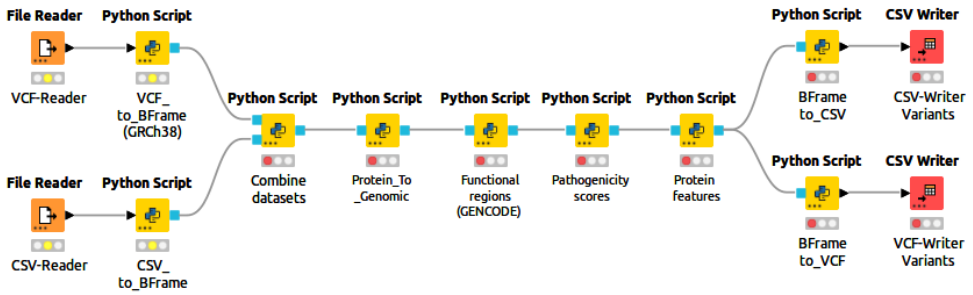
Workflow systems like Snakemake [45] and Nextflow [44] improve reproducibility, but require coding expertise, while cloud-based platforms like Galaxy [43] may raise privacy concerns with sensitive clinical data. Although KNIME [60] offers a local, GUI-based alternative for building integrative workflows, prior KNIME implementations focused narrowly on variant calling or annotation [34, 35].

Here we present AstraKit, a comprehensive suite of scalable, reproducible workflows supporting a multitude of processing steps in bioinformatics and precision medicine. The workflows encompass the entire data processing pipeline in precision medicine, including variant calling, data preprocessing, customizable variant annotation, filtering and interpretation. Critically AstraKit addresses significant challenges in biomedical research, including the generation of multi-layered omics workflows for the cohort analyses of mutational profiles, and the integration of CRISPR screens. We provide customizable workflows for the cohort analysis of molecular profiles, and for integrating various omics data for analyses, includign pathway score calculation, survival analysis, and the integration of CRISPR screens. By directly comparing

A

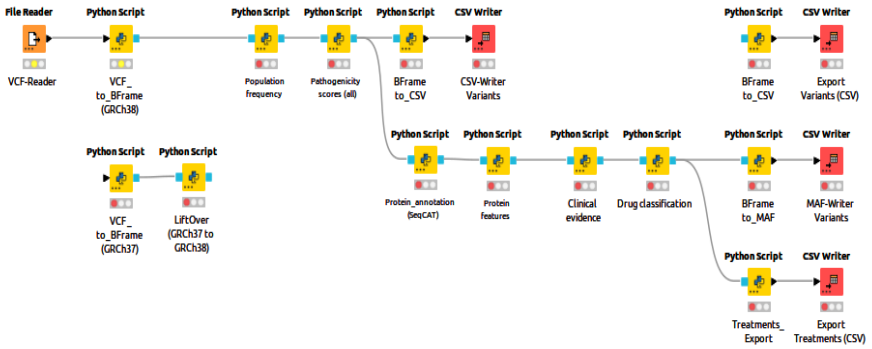
Gene	AA Exchange
TP53	E2K
TP53	E2Q
TP53	E2A

Gene	AA Exchange
KRAS	G12C
BRAF	V600E



B

CHROM	POS	ID	REF	ALT	....
1	46278251	.	G	A	.
1	158627717	.	G	C	.
2	61499712	.	C	T	.



C

Genomic location	Gene symbol	AA exchange	transcript	Protein ID	Uniprot ID	type	1000Genomes_AF	AlphaMissense_Score	CIViC_exact
chr12:25245351C>A	KRAS	G12C	NM_004985.5:c.34G>T	NP_004976.2	P01116	protein_coding	.	0.998	34

D

Associated biomarker	Match type	Disease	Drug class	Drug	Evidence level	Response	Source	Citation ID
KRAS G12C	Exact match	Lung Non-small cell carcinoma	KRAS Inhibitor	Sotorasib	1	Sensitive	CIViC	34096690

Figure 2: Customizable variant interpretation workflows in AstraKit. **A** A KNIME variant annotation workflow start with reading in one or multiple data files. Users can then dynamically align variant processing nodes to annotate the variants with selected features. The annotated variants can be written out again in different data formats. **B** For clinical interpretation, AstraKit provides the functionality to add clinical evidence and drug classification annotation, and the option to export therapeutic options for a patient. **C** AstraKit can export the functional annotations of variants, as well as **D** the retrieved evidence-based biomarkers for selected targeted therapies.

clinical outcomes with high-throughput drug screen data, we bridge observational clinical data with functional cellular biology.

2 Materials and Methods

2.1 Datasets and analytical workflows

To test and validate the workflows for data preprocessing and variant interpretation, we used multiple publicly available datasets, including somatic variants in VCF format [52, 51, 50]. For the variant data cohort analysis, we downloaded somatic variant data, gene expression, and DNA methylation files

from 32 cancer types from The Cancer Genome Atlas (TCGA) using the TCGAbiolinks package [55]. Drug sensitivity and CRISPR screen data was retrieved from cancer cell line data from GDSC [20] (release 8.5), which includes IC_{50} and AUC values from drug sensitivity assays for over 200.000 drug-cancer cell lines.

The AstraKit workflows were constructed using the KNIME analytics platform (v5.8.0) using the Python and R extensions. In order to transfer non-tabular data between nodes, such as JSON structures, the data was packed into objects using the Python pickle package.

2.2 Variant annotation and interpretation

Nodes for reading and writing different bioinformatic data formats, including VCF, MAF, FASTA, CSV and plain text, as well as the LiftOver between different genome assemblies, were implemented using AdaGenes [23]. For the variant annotation and interpretation steps, we created Python nodes using Onkopus [46]. Annotated variant data was passed through the workflows by creating pickled data objects. Genetic variants identified at transcript and protein level were mapped to the genomic level using SeqCAT [59].

Targeted therapies based on genetic alterations were retrieved from CIViC [27], MetaKB [26], OncoKB [25]). To normalize cancer types, we manually created a mapping between DepMap and TCGA cancer types, where only Depmap data that could be mapped to TCGA cancer types was considered. The disease specification of the entries from CIViC, OncoKB and MetaKB were converted to lowercase and common synonyms were normalized. We then applied a combination of exact and fuzzy matching to map the cancer entitigy to the OncoTree [1] cancer types. The method first checks for an exact match between the source data and the standardized cancer types. If no exact match is found, the method applies fuzzy matching using the token sort ratio algorithm, which performs a fuzzy comparison to measure similarity, and selects the highest scored cancer type term. The token sort ratio R is computed as the Levenshtein distance $D(S1, S2)$ between the token strings $S1$ and $S2$

$$R = \frac{D(S1, S2)}{max(len(S1), len(S2))} \cdot 100 \qquad (1)$$

To normalize drug names, we queried drug name synonyms from the ChEMBL database and mapped drug names to ChEMBL IDs [30].

2.3 Molecular and protein structure analysis

Protein structural features were derived using the Onkopus protein analysis Application Programming Interface (API) endpoints. Biochemical features of variants were computed using BioPython [36]. The change in protein stability upon mutation ($\Delta\Delta G$) was predictes using FoldX (v5.0) [14]. $\Delta\Delta G$ values were calculated as the difference in folding free energy between mutant and wild-type structure:

$$\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt}$$

All calculations were performed using the BuildModel command with default parameters.

2.4 Cohort analysis of molecular profiles

Cohort-specific analysis and visualization of somatic variants was implemented using MAFtools [57]. Intersection analysis was implemented using bedtools [54]. Additional data analysis and visualizations were implemented with scikit-learn [39] and matplotlib [40] and seaborn [41].

To calculate the similarity of a variant file and a cell line, we use a weighted Jaccard index that incorporates variant-specific weights to prioritize clinically significant variants. For two mutational profiles A and B, we calculate the Jaccard index J_w with a weight function $w(v)$, assigning importance to a variant v:

$$J_w(A, B) = \begin{cases} \dfrac{\sum\limits_{v \in A \cap B} w(v)}{\sum\limits_{v \in A \cup B} w(v)} & \text{if } A \cup B \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

2.5 Drug screen analysis

To analyze drug sensitivity in relation to clinical outcomes, IC_{50} values were first converted to molar units based on user-specified input units. IC_{50} values were then transformed to pIC_{50} values:

$$pIC_{50} = -log_{10}(IC_{50})$$

This transformation linearizes the relationship between drug concentration and biological effect, with higher pIC_{50} values indicating greater drug sensitivity. For normalization, we implemented a percentile-based approach within each sample (grouped by the sample ID). First, pIC_{50} values were converted to molar units and then to pIC_{50} values. Within each sample group, pIC_{50} values were scaled using the 5th and

95th percentiles of the distribution:

$$\text{pIC}_{50}\text{-norm} = \frac{\text{pIC}_{50} - P_5}{P_{95} - P_5}$$

where P_5 and P_{95} represent the 5th and 95th percentiles of pIC_{50} values within each sample group. Values were clipped to the [0,1] range to maintain interpretability, with values below the 5th percentile mapped to 0 and values above the 95th percentile mapped to 1.

For binary classification, normalized values were categorized using a user-defined threshold, per default set to 0.5:

$$\text{drug sensitivity}(\text{IC}_{50}) = \begin{cases} 1 & \text{if pIC}_{50}\text{-norm} > 0.5 \\ 0 & \text{if pIC}_{50}\text{-norm} \leq 0.5 \end{cases}$$

This classification scheme assigns samples with normalized pIC_{50} values in the upper half of the distribution (greater than the median of the normalized scale) as sensitive (1), and those in the lower half as resistant (0). The in vitro response based on drug sensitivity measurements was then compared against the observed clinical outcomes r, where:

$$\text{Clinical Response} = \begin{cases} 1 & \text{if } r_i = \text{sensitive} \\ 0 & \text{if } r_i = \text{resistant} \end{cases}$$

The agreement between predicted response (based on IC_{50}) and clinical response was evaluated using appropriate statistical measures for binary classification, including sensitivity, specificity, and area under the receiver operating characteristic curve (ROC-AUC).

3 Results

3.1 The AstraKit workflow suite

AstraKit delivers a comprehensive suite of modular, automated workflows for a multitude of tasks in precision medicine and biomedical research, including genomic data processing, multi-omics integration, and drug screen analysis (Figure 1). These pipelines cover end-to-end genetic data processing—from variant calling and molecular data preprocessing to variant annotation, filtering, and interpretation. They further enable comprehensive workflows to analyze patient cohorts, and adaptive pipelines to integrating different omics data for analyses. Critically, AstraKit introduces innovative tools for preclinical model assessment and clinical trial evaluation by integrating molecular data with experimental drug screen results. Implemented in KNIME, AstraKit provides a graphical user interface that allows researchers to customize, modify, and execute workflows without coding expertise. Built-in workflow saving ensures reproducibility and seamless sharing of research results.

3.2 Customizable mutational profile processing

AstraKit's variant interpretation framework empowers users to construct annotation and filtering pipelines through a unique modular architecture. Each processing step (e.g., LiftOver, annotation, filtering) is encapsulated in standalone, interoperable nodes. This modularity is enabled by a standardized branched JSON data structure that serves as the universal input/output format across all nodes. By enforcing consistent data handling, nodes can be freely combined without manual data reformatting into reproducible, context-specific workflows. The suite provides extensive adaptable nodes for annotation (e.g., pathogenicity predictors, protein features, clinical evidence), filtering, LiftOver, and protein-to-genome conversion. Workflows dynamically begin with format-specific readers (VCF, MAF, CSV) and conclude with export nodes, supporting flexible pipeline design. For instance, a workflow (Figure 2A) may begin by annotating with population allele frequencies (gnomAD), then filtering common variants before layering pathogenicity predictions (e.g., SIFT, PolyPhen), protein structural data, and ClinVar classifications. Benign/non-coding variants are iteratively excluded, streamlining identification of clinically actionable alterations. Another example would be the functional annotation of variant cohorts (Figure 2B), where protein-level variants are mapped to genomic coordinates, merged, and annotated with functional context—including transcript isoforms, protein surface accessibility, alpha-carbon distances, and secondary structures—to reveal mutation-driven mechanistic insights. Final outputs include annotated VCFs, column-wise variant reports (Figure 2C), and evidence-based therapy recommendations (Figure 2D), facilitating seamless transition from analysis to clinical decision support.

3.3 Molecular profile cohort analysis

To establish the analytical robustness of AstraKit for somatic mutation profiling, we created workflows for analyzing variants and mutated genes, mutual exclusivity, and affected pathways (Figure 3A). These analyses may be a starting point for identifying frequently altered genes and co-occurring mutations.

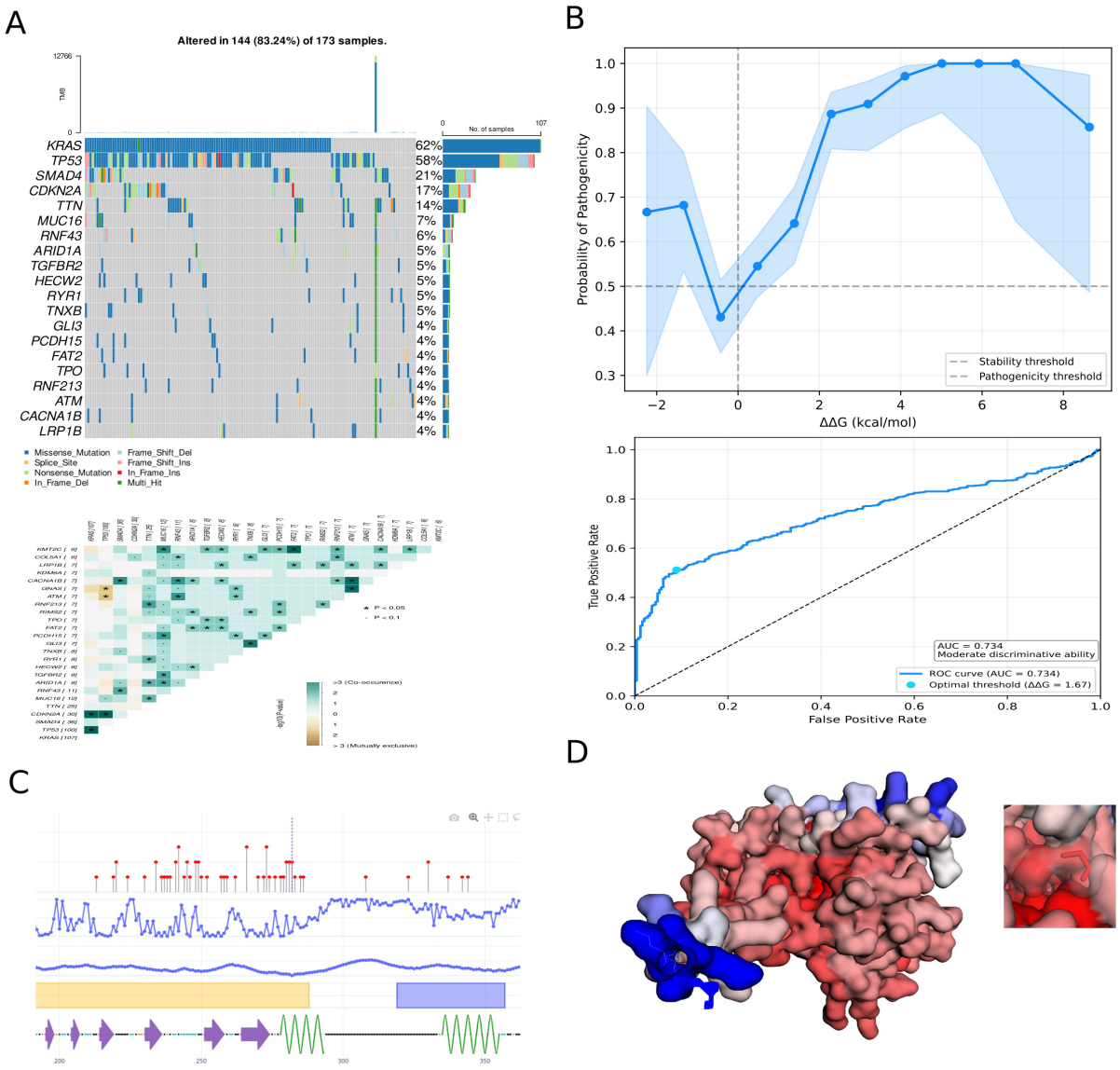


Figure 3: AstraKit variant cohort analysis overview. **A** Oncoprint and mutual exclusivity analysis of the most 20 recurrently altered genes in TCGA PAAD patients, which can be directly generated using the AstraKit variant cohort analysis workflows. **B** Analysis of the correlation between protein stability change predictions and AlphaMissense-predicted pathogenicity of TP53 variants. **C** Additionally, AstraKit enables the dynamic exploration of the protein of variant cohorts, including a protein analysis, and 3D visualizations of the predicted protein structure according to different features, e.g. **D** the predicted binding sites.

In addition, AstraKit enables the analysis of specific annotated features of variant cohorts. For instance, the protein stability change upon mutations can be computed automatically and can be directly compared to the predicted pathogenicity of variants. When comparing FoldX-predicted protein stability changes and AlphaMissense-predicted pathogenicity scores for TP52 variants, we found an increased probability of pathogenicity for both stabilizing variants (positive $\Delta\Delta G$) as well as destabilizing variants (negative $\Delta\Delta G$) (Figure 3B). A direct comparison via receiver operating characteristic (ROC) exhibited a area under the curve (AUC) of 0.73, indicating a significant correlation between protein stability change and pathogenicity. To further analyze the biological context of variant cohorts, AstraKit is able to gen-

erate aligned protein plots with differing tracks for visualizing a variety of annotations (Figure 3C), as well as dynamic 3D protein visualizations according to pathogenicity, predicted binding sites, protein domains, or changes in biophysical characteristics like charge, aromaticity, and hydropathy (Figure 3D).

3.4 Multi-faceted omics analyses

Building on genomic variant interpretation, we developed complementary omics workflows to functionally validate variant impact and prioritize therapeutic targets. For instance, users are able to split datasets according to mutation status by selecting a gene and variant (e.g. *KRAS* amplification), which may then be analyzed separately. Second, patient mutation profiles are matched to DepMap cell lines via Jaccard similarity (threshold≥0.6), and precomputed Chronos scores for top matches identify genes with mutation-essentiality correlations.

3.5 Personalized therapy optimization by integrating clinical biomarkers with experimental drug screens

Complementary to the molecular profiles analyis, AstraKit provides straight-forward workflows to analyze the targeted therapies for patient cohorts, including the most sensitive/resistant drug classes (Figure 5A), drugs (Figure 5B), biomarkers (Figure 5C), and tumor types (Figure 5D). A straightforward approach of AstraKit is the possibility to correlate the results of preclinical drug screens with clinical evidence of the underlying cell lines' molecular alterations. The associated workflow provides the functionality of loading multiple genetic profiles in VCF format, as well as a table-formatted file from a drug screen containing IC_{50} or GR_{max} values. In addition, various parameters can be manually adjusted, such as the normalization of IC50 values, as well as the cancer type and the type of clinical evidence. To analyze the correlations between experimental drug sensitivity and clinical evidence, we examined the results of in vitro drug screens of the GDSC dataset, as well as the mutational profiles of the underlying cell lines. From these profiles, we generated VCF files from retrieved clinical evidence data for the biomarkers included in the cell lines' mutational profiles by searching for exact matches using Onkopus and mapping them to therapeutic options. The direct comparison of clinical evidence records and drug screen results allows to reveal drugs that turned out to be effective (high pIC_{50}), but that have not been explored in clinical studies yet (lowe number of clinical records) (Figure 5E). Complementary, comparing the pIC50 with clinical outcome per cell line reveals promising cell lines for further studies (Figure 5F).

The direct analysis of the retrieved clinical outcome allows a detailed exploration of available therapeutics, which can be explored multifaceted through interactive sunburst visualizations (**??**A). A direct comparison of clinical outcome and drug screen results for TCGA PAAD patients (Figure 5B) through ROC revealed a high concordance between the results of specific cell lines (COSMIC ID 753595: AUC 0.89; 907285: AUC 0.88). However, significant discrepancies were observed depending on the cell line.

4 Discussion

We presented AstraKit, a set of novel workflows for a multitude of use cases in bioinformatics and precision medicine, including genetic variant interpretation, multi-layered omics analyses, and drug screen analysis. As AstraKit is based on the KNIME framework, all workflows provide graphical, easily adaptable nodes that require no prior coding knowledge. While other pipeline software solutions, including Snakemake and Nextflow, excel at scalable, command-line-driven pipeline execution, and Galaxy provides a cloud-based workflow environment, KNIME stands out as a superior solution for unique visual, node-based interfaces combined with deep analytical flexibility. Given that all workflows can be executed locally, KNIME is therefore ideal for processing sensitive clinical data.

AstraKit eliminates manual tool chaining through end-to-end genomic processing, including fully automated variant calling, annotation, filtering, and LiftOver. Furthermore, the workflows provide scalable cohort analysis for mutational landscape analysis and stratified omics investigations. Finally, AstraKit presents a novel approach for translational drug-response modeling by analyzing efficient personalized therapies for precision medicine, and by stratifying the results of experimental drug screens though directly linking experimental drug sensitivity to clinical evidence retrieved for the patients' genomic biomarkers.

An significant novel feature of AstraKit is its ability to create modular, dynamically configurable workflows for processing genetic variant data. Prior approaches [34] presented workflows that were primarily focused on the variant calling process, with limited possibilities to customize the variant processing and annotation process. AstraKit provides workflows in which each processing step is represented by its

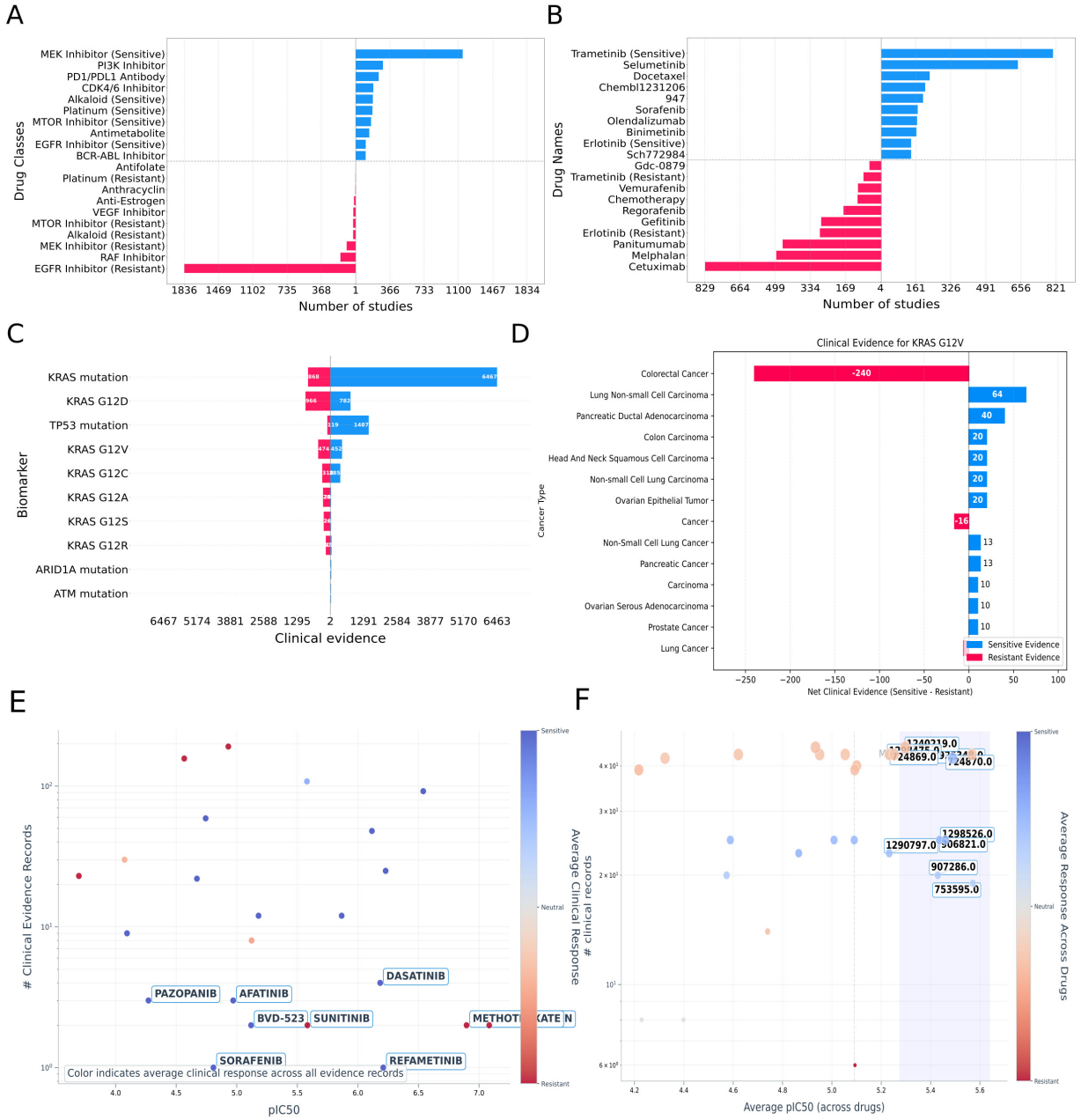


Figure 4: Clinical evidence analysis for patient cohorts and drug screen analysis. **A** Top 10 sensitive and top 10 resistant drug classifications for pancreatic cancer patients. Clinical studies were retrieved using Onkopus exact matches, which retrieves only evidence for the exact biomarker found in patient's mutational profile. **B** Top 10 sensitive and resistant therapeutics, and **C** number of studies per biomarker with sensitive (blue) and resistant (red) outcome. **D** Top sensitive and resistant clinical outcome per cancer type, based on PAAD patient biomarkers. **E** Comparison of clinical evidence records and drug screen results (pIC50) per therapeutic option. Drugs with low clinical evidence, but high pIC50 values are highlighted, indicating potentially underexplored, sensitive therapeutics. **F** Comparison of clinical evidence and drug screen results based on cell lines.

own workflow node, whereby nodes for annotation, LiftOver, and filtering can be freely combined.

In addition to the detailed analysis of individual molecular profiles, AstraKit provides workflows for

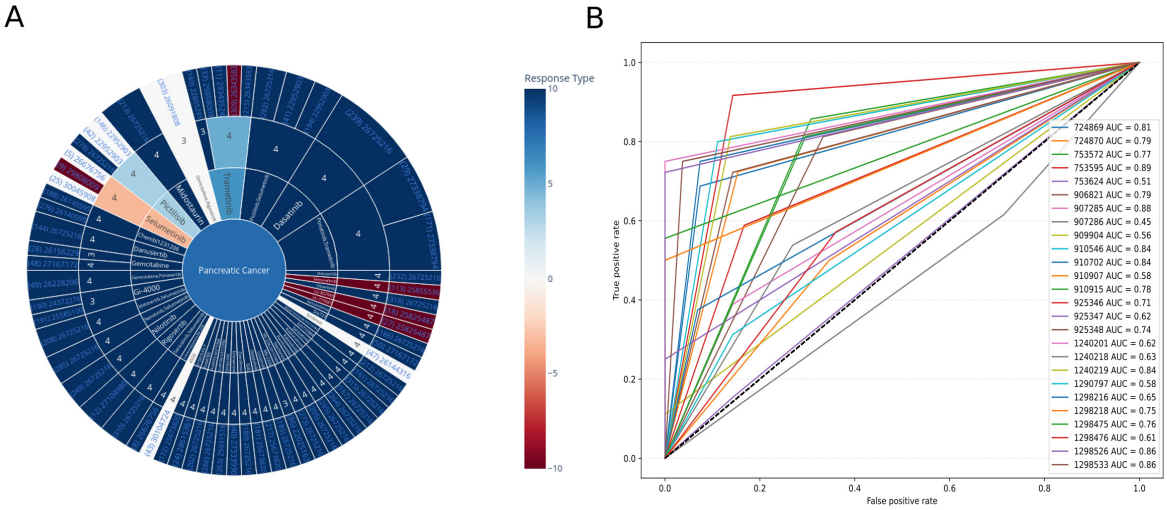


Figure 5: Drug screen analysis: Comparison between experimental drug sensitivity and clinical outcome. **A** Clinical studies for KRAS:p.G12D, filtered to clinical evidence for pancreatic cancer, visualized in the Onkopus sunburst visualization. Blue results indicate sensitive, red resistant outcome. **B** Receiver operating characteristic of clinical outcome of TCGA PAAD patient biomarkers predicting experimental drug sensitivity (IC_{50}) per cell line. The correspondence between drug screen and clinical studies' results varies significantly across cell lines, ranging from an area under the ROC curve (AUC) from 0.45 up to 0.89.

analyzing data cohorts, identifying mutational patterns as well as mutual exclusivity. However, it has to be pointed out that license-protected software, including FoldX, is required to be already installed on the user's host system.

These functionalities are enhanced by workflows for multi-layered omics data analyses. By integrating heterogeneous multi-omics data into biologically interpretable pathway scores, researchers can transcend single-layer molecular noise to pinpoint dysregulated disease mechanisms and prioritize actionable therapeutic targets, thereby accelerating the translation of complex omics signatures into clinically relevant insights for precision medicine. Additionally, we present a novel approach for linking mutational profiles to CRISPR screens across cancer types, identifying gene vulnerabilities specific for a patient's molecular profile.

We introduced an integrated approach for bridging the gap between clinical evidence and preclinical drug response data through systematic correlation of molecular profiles with in vitro drug sensitivity. Our methodology offers complementary value for evaluating clinical evidence: The clinical annotation workflow processes genetic alterations to identify potential therapeutic options supported by existing clinical evidence, while another workflow correlates molecular profiles with drug screening data to reveal associations between genomic alterations and drug response that may not yet be documented in clinical literature. This second capability represents a particularly powerful application, as it enables the identification of sensitive cell lines and potential biomarkers that lack prior clinical validation, a critical gap in precision oncology where promising therapeutic opportunities might otherwise remain undiscovered.

At the methodological core of this integration lies the systematic correlation of quantitative drug response metrics (IC50 and GR_{max} values) with clinical evidence classifications. This correlation requires careful normalization of IC50 values, while the determination of a robust classification thresholds further presents a challenge, as the continuum of drug response data has to be discretized into meaningful "sensitive" versus "resistance" categories that align with clinical observations.

A similar approach was demonstrated by [3], who applied a similar methodology to mature T-cell leukemias and lymphomas, successfully identifying DDR pathway alterations as novel biomarkers for WEE1 inhibitor sensitivity that were beyond currently available evidence in databases. This exemplifies how the method of comparing preclinical drug screen results with existing clinical outcome can uncover previously unrecognized therapeutic vulnerabilities by connecting molecular character-

istics with drug response patterns in cell line models, potentially accelerating the identification of novel biomarker-drug relationships for future clinical investigation. The validation of this workflow revealed significant discrepancies in the correlation between drug screening results and clinical outcomes across different cell lines. This finding suggests the hypothesis that biomarkers that differ between cell lines play a crucial role in drug sensitivity.

In summary, AstraKit provides dynamic, state-of-the-art workflows for automated data processing in biomedical research and precision medicine. It encompasses the entire precision oncology workflow, including data format conversion, variant calling, variant annotation and interpretation, multi-omics analysis, and drug screen analysis. usable out-of-the-box, featuring KNIME's automated detection and installation of missing plugins and packages. Featuring cross-platform compatibility and extensive customization options, we are confident that it will be a tool of great value for researchers seeking a tailored, high-performance solution for processing sequence variant data.

Executed locally via KNIME's intuitive interface, AstraKit ensures data privacy, full auditability, and accessibility for wet-lab scientists and clinicians. By resolving tool fragmentation across the precision medicine continuum, AstraKit accelerates biomarker discovery, therapeutic validation, and the translation of genomic insights into personalized treatment strategies.

5 Data availability

The KNIME workflows are publicly available at KNIME HUB at https://hub.knime.com/bioinf_goe/spaces/Public/~Zd6-FraRjnV7MMpu/. The source code of the KNIME workflows and the data preprocessing scripts are available at https://gitlab.gwdg.de/MedBioinf/mtb/astrakit. The source code of all Onkopus modules is available at https://gitlab.gwdg.de/MedBioinf/mtb/onkopus, including instructions on how to install the Onkopus modules locally. All modules have been generated from freely available data sources.

6 Funding

This work was supported by the Gemeinsamer Bundesauschuss (01NVF20006), the Volkswagen Foundation (11-76251-12-1/19), the Deutsche Krebshilfe (70114018), the Deutsche Forschungsgemeinschaft (KFO5002) and the Bundesministerium für Bildung und Forschung (BMBF) (01KD2437, 01KD2401B, 01KD2208A, 01KD2414A).

7 Competing interests

No competing interest is declared.

8 Author contributions

J.D. and N.S.K. developed the conceptualization and methodology. N.S.K., K.K. and M.S. implemented the KNIME workflows. N.S.K. wrote the manuscript. All authors approved the manuscript.

9 Acknowledgments

The authors thank the Molecular Tumor Board of the University Medical Center Göttingen (UMG), namely Alexander König, Li Beißbarth and Kirsten Reuter-Jessen, for providing constant feedback and discussions on clinical needs and use cases. The authors also thank the International Max Planck Research School for Genome Science (IMPRS-GS).

References

[1] Kundra, R., Zhang, H., Sheridan, R., Sirintrapun, S., Wang, A., Ochoa, A., Wilson, M., Gross, B., Sun, Y., Madupuri, R. & Others OncoTree: a cancer classification system for precision oncology. *JCO Clinical Cancer Informatics*. **5** pp. 221-230 (2021)

[2] Babadi, M., Fu, J., Lee, S., Smirnov, A., Gauthier, L., Walker, M., Benjamin, D., Zhao, X., Karczewski, K., Wong, I. & Others GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nature Genetics*. **55**, 1589-1597 (2023)

[3] Schmidt, N., Kornrumpf, K., Ries, L., Markus, K., Ksionsko, N., Mazzeo, P., Lohrberg, M., Sitte, M., Salinas, G., Braun, T. & Others Discovery of molecularly-informed therapeutic strategies for mature T-cell leukemias and lymphomas. *Blood Advances*. **9**, 5327-5340 (2025)

[4] Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. **38**, e164-e164 (2010)

[5] McLaren, W., Gil, L., Hunt, S., Riat, H., Ritchie, G., Thormann, A., Flicek, P. & Cunningham, F. The ensembl variant effect predictor. *Genome Biology.* **17**, 122 (2016)

[6] Cingolani, P., Patel, V., Coon, M., Nguyen, T., Land, S., Ruden, D. & Lu, X. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers In Genetics.* **3** pp. 35 (2012)

[7] Beg, M., Taka, J., Kluyver, T., Konovalov, A., Ragan-Kelley, M., Thiéry, N., Fangohr, H. Using Jupyter for reproducible scientific workflows. *Computing In Science & Engineering.* **23**, 36-46 (2021)

[8] Gorshkov K, Chen C, Marshall R et al. Advancing precision medicine with personalized drug screening. *Drug Discovery Today* 24, 272-278 (2019).

[9] Buel G, Walters K. Can AlphaFold2 predict the impact of missense mutations on structure?. *Nat Struct Mol Bio* **29**, 1-2 (2022)

[10] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38, e164-e164 (2010).

[11] Cingolani P, Platts A, Wang L et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6, 80-92 (2012).

[12] McLaren W, Gil L, Hunt S et al. The ensembl variant effect predictor. *Genome Biol* **17**, 122 (2016).

[13] Jones P, Binns D, Chang H et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).

[14] Delgado J, Radusky L, Cianferoni D, Serrano L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168-4169 (2019).

[15] Jiménez-Santos MJ, Nogueira-Rodríguez A, Piñeiro-Yáñez E et al. PanDrugs2: prioritizing cancer therapies using integrated individual multi-omics data. *Nucleic Acids Res* **51**, W411-W418 (2023).

[16] Hafner M, Niepel M, Chung M, Sorger P. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat Methods* **13**, 521-527 (2016).

[17] Lappalainen I, Lopez J, Skipper L et al. DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res* **41**, D936-D941 (2012).

[18] Perera-Bel J, Hutter B, Heining C et al. From somatic variants towards precision oncology: evidence-driven reporting of treatment options in molecular tumor boards. *Genome Med* **10**, 18 (2018).

[19] Li M, Datto M, Duncavage E 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 Diagnostics.* **19**, 4-23 (2017)

[20] Yang W, Soares J, Greninger P et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* **41**, D955-D961 (2012).

[21] Tapia Contreras C, Falke J, Seifert D et al. KRASG 12 C-inhibitor-based combination therapies for pancreatic cancer: insights from drug screening. *Mol Oncol* **19**, 295-310 (2025).

[22] Lu C, Wu C, Tsai M et al. MutScape: an analytical toolkit for probing the mutational landscape in cancer genomics. *NAR Genomics Bioinf* **3**, lqab099 (2021).

[23] Kurz NS, Drofenik K, Kornrumpf K et al. AdaGenes: A streaming processor for high-throughput annotation and filtering of sequence variant data. *BioRxiv* pp. 2025-06 (2025).

[24] Danecek, P., Auton, A., Abecasis, G., Albers, C., Banks, E., DePristo, M., Handsaker, R., Lunter, G., Marth, G., Sherry, S. & Others The variant call format and VCFtools. *Bioinformatics.* **27**, 2156-2158 (2011)

[25] Chakravarty D, Gao J, Phillips S et al. OncoKB: a precision oncology knowledge base. *JCO Precis Oncol* **1** pp. 1-16 (2017).

[26] Wagner A, Walsh B, Mayfield G et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. *Nat Genetics* **52**, 448-457 (2020).

[27] Griffith M, Spies N, Krysiak K et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genetics.* **49**, 170-174 (2017).

[28] Malani D, Kumar A, Brück O et al. Implementing a functional precision medicine tumor board for acute myeloid leukemia. *Cancer Discovery* **12**, 388-401 (2022).

[29] Kim S, Chen J, Cheng T et al. PubChem 2023 update. *Nucleic Acids Res* **51**, D1373-D1380 (2023).

[30] Gaulton A, Hersey A, Nowotka M et al. The ChEMBL database in 2017. *Nucleic Acids Res* **45**, D945-D954 (2017).

[31] Potdar S, Ianevski F, Ianevski A et al. Breeze 2.0: an interactive web-tool for visual analysis and comparison of drug response data. *Nucleic Acids Res* **51**, W57-W61 (2023).

[32] Ardin M, Cahais V, Castells X et al. MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics* **17** pp. 1-10 (2016).

[33] Hiltemann S, Mei H, Hollander M et al. CGtag: complete genomics toolkit and annotation in a cloud-based Galaxy. *GigaScience* **3**, 2047-217X (2014).

[34] Herrick N, Walsh S. ILIAD: a suite of automated Snakemake workflows for processing genomic data for downstream applications. *BMC Bioinformatics* **24**, 424 (2023).

[35] Yang J, Beißbarth T, Dönitz J. Onkopipe: A Snakemake Based DNA-Sequencing Pipeline for Clinical Variant Analysis in Precision Medicine. *Stud Health Technol Inf* **307** pp. 60-68 (2023).

[36] Cock P, Antao T, Chang J et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422 (2009).

[37] Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nat Methods* **18**, 1161-1168 (2021).

[38] Fjukstad B, Dumeaux V, Hallett M, Bongo L. Reproducible data analysis pipelines for precision medicine. *2019 27th Euromicro International Conference On Parallel, Distributed And Network-Based Processing (PDP)* pp. 299-306 (2019).

[39] Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res* **12** pp. 2825-2830 (2011).

[40] Barrett P, Hunter J, Miller J, Hsu J, Greenfield P. matplotlib–A Portable Python Plotting Package. *Astronomical Data Analysis Software And Systems XIV* **347** pp. 91 (2005).

[41] Waskom M. Seaborn: statistical data visualization. *J Open Source Software* **6**, 3021 (2021).

[42] Larivière D, Abueg L, Brajuka N et al. Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy. *Nat Biotechnology* **42**, 367-370 (2024).

[43] Community, T. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Research.* **52**, W83 (2024)

[44] Di Tommaso P, Chatzou M, Floden E et al. Nextflow enables reproducible computational workflows. *Nat Biotechnology* **35**, 316-319 (2017).

[45] Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520-2522 (2012).

[46] Kurz NS, Kornrumpf K, Tucholski T et al. Onkopus: precise interpretation and prioritization of sequence variants for biomedical research and precision medicine. *Nucleic Acids Res* pp. gkaf376 (2025).

[47] Reel P, Reel S, Pearson E et al. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol Adv* **49** pp. 107739 (2021).

[48] Ewald J, Zhou G, Lu Y et al. Web-based multi-omics integration using the Analyst software suite. *Nat Protocols* **19**, 1467-1497 (2024).

[49] Chen C, Wang J, Pan D et al. Applications of multi-omics analysis in human diseases. *MedComm* **4**, e315 (2023).

[50] Zook J, Catoe D, McDaniel J et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3** pp. 160025 (2016).

[51] Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res* **48**, D941-D947 (2020).

[52] Pereira B, Chin S, Rueda O et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun* **7** pp. 11479 (2016).

[53] Kurz NS, Perera-Bel J, Höltermann C et al. Identifying Actionable Variants in Cancer-The Dual Web and Batch Processing Tool MTB-Report. *Stud Health Technol Inf* **296** pp. 73-80 (2022).

[54] Quinlan A, Hall I. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

[55] Colaprico A, Silva T, Olsen C et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **44**, e71-e71 (2016).

[56] Danecek P, Bonfield J, Liddle J et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

[57] Mayakonda A, Lin D, Assenov Y et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res* **28**, 1747 (2018).

[58] Gao J, Aksoy B, Dogrusoz U et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci Signaling* **6**, pl1 (2013).

[59] Kornrumpf K, Kurz NS, Drofenik K et al. SeqCAT: Sequence Conversion and Analysis Toolbox. *Nucleic Acids Res* **52**, W116-W120 (2024).

[60] Fillbrunn A, Dietz C, Pfeuffer J et al. KNIME for reproducible cross-domain analysis of life science data. *J Biotechnol* **261** pp. 149-156 (2017).