

SOFTWARE

Open Access



# AMLVaran: a software approach to implement variant analysis of targeted NGS sequencing data in an oncological care setting

Christian Wünsch<sup>1\*</sup> , Henrik Banck<sup>1</sup>, Carsten Müller-Tidow<sup>2</sup> and Martin Dugas<sup>1</sup>

## Abstract

**Background:** Next-Generation Sequencing (NGS) enables large-scale and cost-effective sequencing of genetic samples in order to detect genetic variants. After successful use in research-oriented projects, NGS is now entering clinical practice. Consequently, variant analysis is increasingly important to facilitate a better understanding of disease entities and prognoses. Furthermore, variant calling allows to adapt and optimize specific treatments of individual patients, and thus is an integral part of personalized medicine.

However, the analysis of NGS data typically requires a number of complex bioinformatics processing steps. A flexible and reliable software that combines the variant analysis process with a simple, user-friendly interface is therefore highly desirable, but still lacking.

**Results:** With AMLVaran (AML Variant Analyzer), we present a web-based software, that covers the complete variant analysis workflow of targeted NGS samples. The software provides a generic pipeline that allows free choice of variant calling tools and a flexible language (SSDL) for filtering variant lists. AMLVaran's interactive website presents comprehensive annotation data and includes curated information on relevant hotspot regions and driver mutations. A concise clinical report with rule-based diagnostic recommendations is generated.

An AMLVaran configuration with eight variant calling tools and a complex scoring scheme, based on the somatic variant calling pipeline *appreci8*, was used to analyze three datasets from AML and MDS studies with 402 samples in total. Maximum sensitivity and positive predictive values were 1.0 and 0.96, respectively. The tool's usability was found to be satisfactory by medical professionals.

**Conclusion:** Coverage analysis, reproducible variant filtering and software usability are important for clinical assessment of variants. AMLVaran performs reliable NGS variant analyses and generates reports fulfilling the requirements of a clinical setting. Due to its generic design, the software can easily be adapted for use with different targeted panels for other tumor entities, or even for whole-exome data. AMLVaran has been deployed to a public web server and is distributed with Docker scripts for local use.

**Keywords:** Genomics, NGS sequencing, Variant calling, Variant annotation, Variant filtering, Variant interpretation, Mutation database, Diagnostic report

\*Correspondence: [c.wuensch@uni-muenster.de](mailto:c.wuensch@uni-muenster.de)

<sup>1</sup>Institute for Medical Informatics, University of Münster, Albert-Schweitzer-Campus 1, Building A11, Münster, Germany  
Full list of author information is available at the end of the article



## Background

Next-Generation Sequencing (NGS) enables large-scale and very cost-effective sequencing of genetic samples for the detection of mutations. Already being used in research worldwide, sequencing data is now starting to enter also routine care settings [1]. It can be a valuable instrument for a better understanding of the emergence and prognosis of a disease [2]. Also, it can help to optimize the type of treatment for personalized medicine [3].

Implementing NGS into clinical settings imposes a lot of challenges [4]. Most noticeably, analysis of raw sequencing data is complex due to the large variety of available bioinformatics pipelines and associated configuration parameters. With rapid development of new analysis pipelines that tend to produce highly differing result lists (“moving target”), bioinformatics expertise is required for an optimized analysis. Even sophisticated pipelines suffer from artifacts caused by immanent weaknesses of the sequencing technology. Discrimination between polymorphisms and pathogenic variants is not trivial, either. The type of tissue to be analyzed is important as well, since relevant variants tend to occur in low frequencies for tumor probes due to heterogeneous cell mixtures. The variety of available clinical databases complicates the annotation of identified variants, and their incompleteness leads to a high number of variants of unknown significance, calling for further (manual) inspection.

Clinical use requires software that is able to perform a complete bioinformatics analysis, while offering a user interface that can be operated by medical or laboratory personnel without a dedicated bioinformatics background. For clinical usage, it is furthermore critical to guarantee reproducibility and traceability of the generated results over long time periods. With ongoing research to clarify the clinical significance of unknown variants and increasing development of related systems in the clinical routine, software tools for the automated detection of disease-specific variants are increasingly relevant for routine clinical practice.

Acute Myeloid Leukemia (AML) is a myeloid disease that is particularly well characterized regarding genetic aberrations. The prognosis for affected patients is significantly influenced by certain driver mutations [5, 6]. With NGS-based techniques already being used in AML routine care, this leukemia entity is a particularly promising target for the design and evaluation of a flexible, automated and user-friendly bioinformatics software.

## Related work

Available solutions for variant analysis from NGS sequencing data include commercial products (e.g. the Qiagen CLC Workbench), which are closed-source and therefore provide only limited insights into variant calling and filter algorithms and only limited options for

customization. Even more critical, commercial web-based tools (like VarSome [7]) typically require the sequencing data to be uploaded to the manufacturer's servers. Open-source command line tools are available for a multitude of specialized subtasks of NGS analyses, but require more in-depth bioinformatics knowledge (e.g. BWA [8] for sequence alignment, GATK [9], FreeBayes [10], SamTools [11] for variant calling or SNPeff [12] or AnnoVar [13] for biological annotations, etc.). While McKerrel et al. present a complete pipeline optimized for AML specific source data [14], their software does not provide a graphical user interface, and therefore requires time-consuming configuration and adaptation by bioinformaticians. Furthermore, a number of generic graphical tools exist, which integrate different variant calling pipelines and aim to automate the variant calling and annotation process (e.g. Galaxy [15], Chipster [16], CoVaCS [17]). However, since the main focus of these tools is variant calling, their output is typically presented as basic lists of detected variants, which have to be further inspected and annotated to yield applicable information. Annotation tools like VariantDB [18] or VIS [19] start from a list of detected variants and can accumulate and manage annotation information from various sources, but are limited to basic research contexts. Neither of these tools addresses specific clinical requirements like reproducibility, long-term data storage or usability for non-bioinformaticians, and are therefore of limited use in a clinical setting. The combination of a complete variant calling pipeline, a management system for annotation databases and pre-curated literature data, detailed clinical reports and elaborated, interactive visualizations is desirable for use in routine patient care, but not primarily provided by any of these tools.

## Objectives

In this article, we present AMLVaran, the AML *Variant Analyzer*. AMLVaran is a web-based software platform for analysis of somatic variants on targeted NGS data, addressing the requirements of a clinical setting. This platform shall cover the complete workflow from raw sequencing data to interactive clinical reports. It shall provide a flexible, modular analysis pipeline that can combine an arbitrary number of variant calling tools, and offer a generic model for variant filtering through a customizable scoring scheme. AMLVaran shall include a user-friendly interface that presents results in form of a structured clinical report with interactive features, which support further research. Furthermore, comprehensive curated data on therapy-relevant hotspot regions shall be incorporated, and presence, absence, or coverage of known driver mutations related to a chosen disease entity shall be provided. AMLVaran shall be tested for use with AML data, but is intended as a generic system adaptable to other cancer types. Since the software is designed for clinical

application, AMLVaran's focus is the generation of accurate and reliable results. We therefore apply the program to three AML datasets, and test its recall and precision to evaluate the software's performance, 0.95 is targeted for recall and precision. Results of two user-studies shall provide evidence for AMLVaran's practical usability by medical professionals.

## Implementation

### System architecture

AMLVaran is composed of three main components. First, a flexible, generic variant calling pipeline that generates variant lists from raw sequencing data. Second, an interactive website, which presents the analysis results together with interactive filter settings and creates a standardized, rule-based clinical report. A relational database storing both, results and annotation data, is the third integral part of AMLVaran.

The workflow consists of the following steps: A web-form manages the upload of one or more NGS raw sequencing files for analysis. Acceptable formats include aligned data in the Binary Alignment Map format (.bam, [20]) or raw data in fastq format [21]. The input files are subsequently processed by the variant calling pipeline, which invokes an adaptable number of variant calling tools and combines their results according to a customizable scheme. When the analysis is finished, the output can be inspected on AMLVaran's web interface.

AMLVaran's web interface provides four main functions for presenting its results, grouped into separate views. The first view lists basic information concerning the patient, and metadata for the sample that is currently being examined. The second one presents an overview of the mutation status for the predefined driver mutation sites. For each mutation site, additional information can be retrieved, such as manually curated information regarding the significance of the respective variant for the disease in question. A coverage plot and details about the detected mutations in the chosen region are integrated as well. Additionally, predefined diagnostic recommendations are included, which are derived from the mutation status of known driver mutation sites. The third view provides an interactive browser, which displays all detected variants. Results can be dynamically filtered and sorted by different criteria, e.g. variant allele frequency or read coverage. The last view features a genome browser, which allows to inspect the mutation of interest on the basis of aligned reads from the original sequence data.

The most important pieces of information from the dynamic website are composed into a static, clearly structured clinical report containing the diagnostic and therapeutically relevant details. A tracking system is provided, ensuring, that in case of updated annotation databases,

the previously generated reports remain accessible. The web-interface is described in detail in the AMLVaran QuickStart Guide (see Additional file 2).

The architecture of the main system components is shown in Fig. 1.

Data security aspects have been taken into account throughout development, details on which are provided in Additional file 1: Section 2.

### Implementation details

#### Generic variant calling pipeline

A central challenge for AMLVaran as well as related tools is the generation of reliable and clinically usable variant lists from a sequenced sample. Different variant calling tools generate highly differing variant lists [22], and several studies indicate, that none of the available open-source variant calling algorithms can obtain the sensitivity and specificity required for clinical usage [17, 23, 24].

It has been shown, that combining several variant calling tools can improve the quality of the resulting variant lists, but an optimized filtering strategy to remove false positive calls is equally important for satisfying results [25]. Chiara et al. [17] showed, that the use of three calling tools with a majority vote consensus strategy can lead to a valuable increase in sensitivity. By combining eight variant callers with a sophisticated score calculation strategy, Sandmann et al. achieved an even greater increase in sensitivity of up to 0.99 [24]. These improvements in detection of low frequency mutations, which are clinically relevant, in particular to detect minimal residual disease, now allow for reliable variant calling, especially with tumor samples.

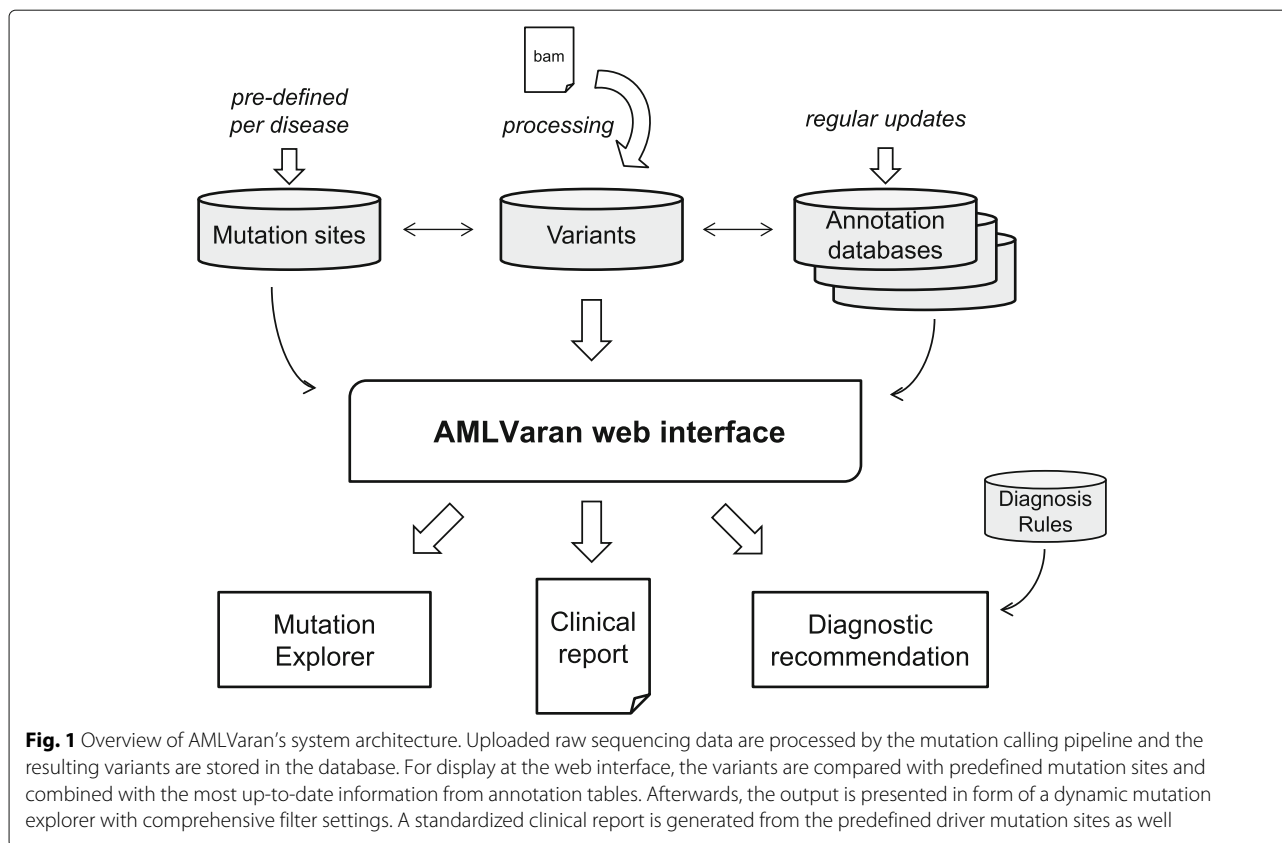
Hence, we implemented a generic variant calling pipeline to support the combination of arbitrary variant calling tools. It was initialized with the tools used in CoVaCS [17] and appreci8 [24], but can be equipped with any other combination of variant callers as well. The results of the individual tools are then integrated via Variant Tools [26] and are subsequently filtered using a flexible, generic score calculation scheme.

The workflow of the generic pipeline is shown in Fig. 2. A detailed documentation of the employed tools is provided in Additional file 3.

#### Storing variants

The Variant Call Format (vcf) [29] defines a standard for storage of genetic variants in a text file. However, the vcf definition allows equivalent variants to be represented in an ambiguous way. Even widely-used databases like dbSNP contain up to 14% non-normalized variants, and 4.6% unrecognized duplicates [30].

In order to get a unified, unique representation of all variants, a number of preprocessing steps is applied to AMLVaran's variant lists, as well as the annotation



databases. The preprocessing is described in Additional file 1: Section 1.

### Variant filtering

The combination of variant callers increases sensitivity compared to only one tool, but at the expense of a higher false positive rate, if no additional filters are employed. Sandmann et al. [24] showed, that a simple consensus strategy as proposed before [17] cannot provide satisfactory results with regards to sensitivity and specificity. Sandmann et al.'s *appreci8* therefore employs a more sophisticated strategy. The output from eight variant callers is combined, and thoroughly filtered with a complex score calculation system. An Artifact and Polymorphism score (A/P score) is calculated from various criteria [24], which can then be used to classify the variants. In principle, however, any rule-based filtering process of variant lists can be described by a criteria-based scoring scheme.

To further improve the filtering process, a formal language for the definition of variant filtering scores was developed and implemented. This language is flexible enough to implement complex scoring algorithms like *appreci8*, easily adaptable, and customizable by a graphical interface [31]. A scoring scheme is specified in JSON and consists of a list of criteria which contribute to the

score by a defined weight. Individual criteria can be linked by logical operators AND and OR, and can include variable thresholds. The software is able to derive a graphical interface from the abstract scheme. This enables the end user to activate and deactivate individual criteria, and to adjust threshold values and weightings as needed.

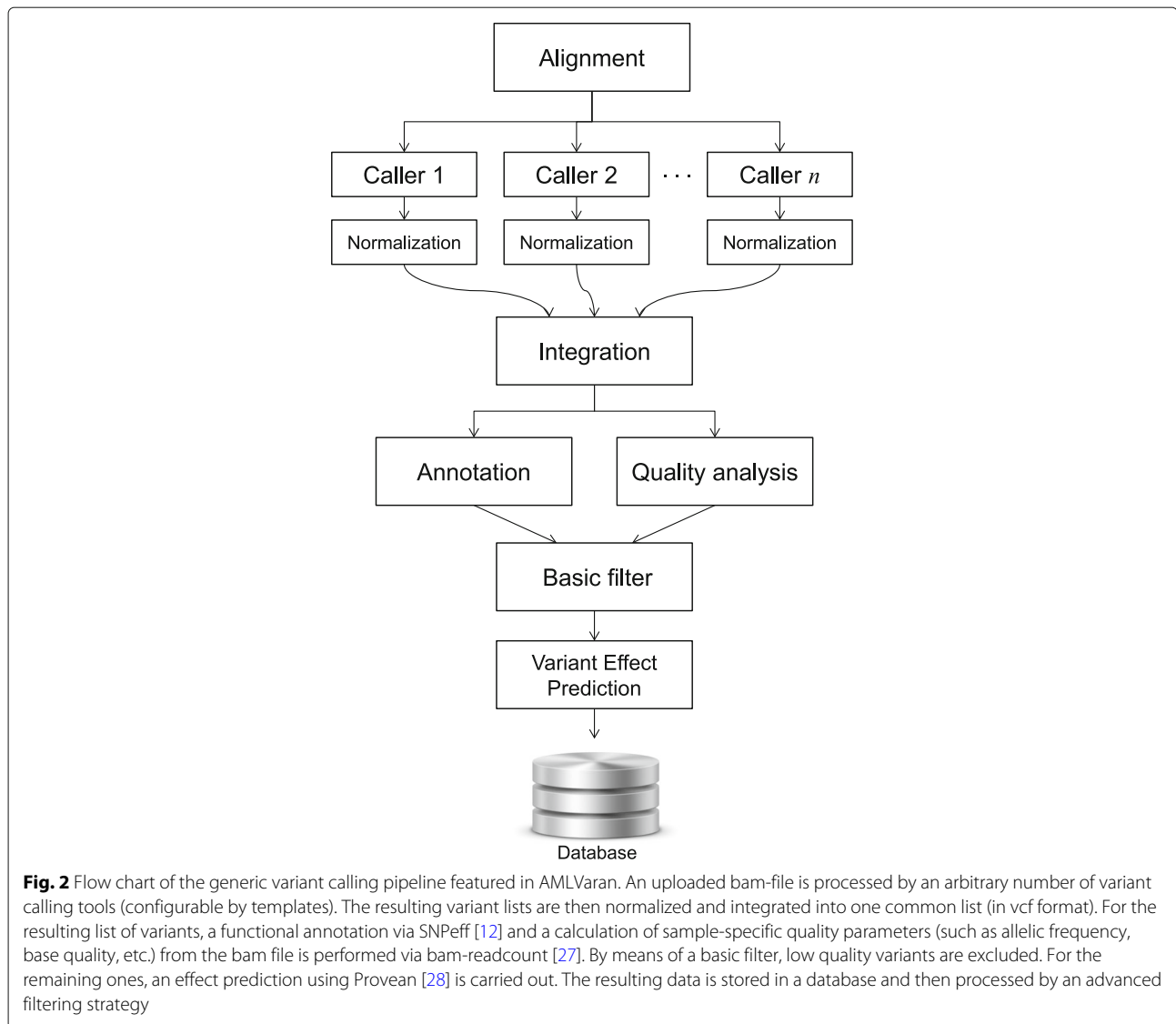
The novel Scoring Scheme Definition Language (SSDL) is universally applicable in the sense that, once created, scoring schemes can be used not only within the AMLVaran platform, but also in standalone applications, websites, and in standard applications such as R or Microsoft Excel. Implementations in JavaScript, R, and Excel VBA are already publicly available [32]. A formal definition of the SSDL's grammar in Backus-Naur-Form is provided in Fig. 3.

A subset of the features of SSDL, as well as the resulting graphical user interface, is shown in Fig. 4.

Furthermore, a default scoring scheme for the *appreci8* filtering process is available from our website[33] and from the AMLVaran package.

### Gathering annotation data

For adequate biological and clinical interpretation of variants, annotation with comprehensive and up-to-date biological information is a necessity. However, clinical annotation databases are often updated on a regular basis,



and there are several options to address these changes [34].

On the one hand, there is a strong medical motivation to always consider the most recent biological data regarding the detected mutations for both old and new samples. On the other hand, it is required to reproduce previous reports based on annotation from the past, for practical as well as legal reasons.

To meet these requirements, an interactive online report equipped with the latest available biological annotations is presented to the user. Simultaneously, an archive of all previously generated reports is stored on AMLVaran's server. A report is automatically saved in PDF format each time the results page is accessed. As proposed by Cutting et al. [34], references to a set of online annotation sources are integrated, allowing for easy access to most recent annotation sources.

The technical implementation features stored variant lists and annotation data in separate tables within the database. Whenever a sample is retrieved, relevant data are merged.

Currently, AMLVaran uses dbSNP [35], COSMIC<sup>1</sup> [36], ClinVar [37], dbNSFP [38], 1000Genomes [39], ESP6500 [40], ExAC [41], as well as functional prediction data from PROVEAN [28] and SIFT [42] for variant annotation.

#### **Identifying diagnostically relevant mutations**

For certain cancer types such as AML, specific driver mutations are known to influence the prognosis of the patient, as well as the choice of therapy. Several of these are summarized in the WHO Guidelines of 2017 [43] for AML. Precise mapping of the detected mutations to the

<sup>1</sup>COSMIC is not available for public usage due to license restrictions.



```

Scheme      ::= [ <Categories> ]
Categories  ::= <Category> | <Category> , <Categories>
Category    ::= { "Type":"Cat", "Caption":<String> , "Color":<Color> , "Entries": [ <Entries> ] }
Entries     ::= <Entry> | <Entry> , <Entries>
SubEntries  ::= <SimpleEntry> | <SimpleEntry> , <SubEntries>
Entry       ::= <ANDNode> | <ORNode> | <SimpleEntry>
SimpleEntry ::= { "Type":"Entry", <Attributes> }
ANDNode     ::= { "Type":"AND-Node", <Attributes> , "Entries": [ <SubEntries> ] }
ORNode      ::= { "Type":"OR-Node", <Attributes> , "Entries": [ <SubEntries> ] }
Attributes  ::= <Attribute> | <Attribute> , <Attributes>
Attribute   ::= <Checkbox> | <Caption> | <Protocol> | <Default> | <ArtiScore> | <PolyScore> | <Condition>
Checkbox    ::= "Checkbox": <Boolean>
Caption     ::= "Caption": <String>
Protocol    ::= "Protocol": <String>
Default     ::= "Default": <Float>
ArtiScore   ::= "ArtiScore": <Integer>
PolyScore   ::= "PolyScore": <Integer>
Condition   ::= "Condition": <Function>
    
```

**Fig. 3** Formal definition of the SSDL in Backus-Naur-Form. A filter scheme consists of one or more categories, which can hold one or more entries, each of which is a rule that contributes to the artifact and/or polymorphism score. Entries can be combined by AND- or OR-nodes. Customizable thresholds can be included as well. The primitive data types (<String>, <Integer>, etc.) are defined canonically. <Function> is an external boolean function specifying the condition that has to be fulfilled for the rule to be applied. The function can utilize the syntax of the underlying R/JavaScript programming language. Access to the columns of the variant table is provided by the custom function "x(columnName)". Other provided functions like "isEmpty(string)" or "stringContains(haystack, needle)" can be used in order to define language-independent filter schemes

(a) Visualization of the generic, system-independent description in SSDL format. The image shows a tree structure of categories and entries. The first category is "Base Quality" with color "#96B4D7". It contains an OR-Node with two sub-entries. The first sub-entry has a checkbox set to "false", a caption "One of the following conditions fulfilled", a default value of 4, and a condition  $x('BQalt') \leq t$ . The second sub-entry has a checkbox set to "true", a caption "BaseQuality of reference by %d higher than of alternative", a default value of 7, and a condition  $x('BQref') \geq x('BQalt') + t$ . The second category is "Occurance in other samples" with color "#96B4D7". It contains an AND-Node with a checkbox set to "true", a caption "Nr of samples with same variant > %d", a default value of 3, an artiScore of 2, a polyScore of 1, and a condition  $x('NrSamples') > t$ . It also has a sub-entry with a checkbox set to "true", a caption "and 90% of these samples have VAF > %d", and a default value of 0.85.

(b) The graphical panel for selecting and adapting the parameters for calculation of artifact and polymorphism scores. It shows a user interface for the "Base Quality" category. The "Arti" score is 4 and the "Poly" score is 0. The conditions are:
 

- One of the following conditions fulfilled: 4
  - Low BaseQuality of alternative (threshold: 15)
    - AND: (reads with reference base present) OR (BaseQuality of reference >thr)
  - BaseQuality of reference by 7 higher than of alternative
  - Low BaseQuality of reference ( $\leq 15$ )
    - AND (reads with other alternative nucleotides present)
- Occurance in other samples: 2
  - Nr of samples with same variant > 3
  - and 90% of these samples have VAF > 0.85
  - (no Hotspot) AND (same variant in > 50% of all samples)
  - No occurrence in any other sample (NrSamples = 1)
- Insertions / Deletions: 1
  - Variant is an insertion / deletion
    - Different variants at same locus found in other samples?
    - VAF < 0.05
- StrandBias: 1
  - Large StrandBias ( $\geq 0.001$ ):
    - alternative on Forward strand  $\leq 2$  and Nr\_ref  $\geq \frac{1}{4}$  Nr non-alternative

**Fig. 4** Generic model for A/P-score calculation. **a** Visualization of the generic, system-independent description in SSDL format. **b** The graphical panel for selecting and adapting the parameters for calculation of artifact and polymorphism scores is derived from the SSDL code shown in **a**. Elements such as categories, simple combinations like AND and OR, checkboxes and adjustable thresholds can be freely combined. The thresholds can be adapted, and the score calculation strategy can be fully customized. By default, a score calculation scheme based on appreci8's complex filtering strategy is provided (see Additional file 4)

listed reference mutations is necessary for an automated application of these guidelines. This mapping process is aggravated by differing levels of accuracy regarding mutation positions in the literature. While a point mutation in DNMT3A at position R882 that is associated with a bad prognosis in AML [44] is well defined, a 4 bp frameshift insertion in NPM1 around position 960 (chr5:170837544) has rather positive effect on therapy response [45], however its actual position can vary. Similarly, c-KIT insertions/deletions anywhere in exon 8 are associated with an increased risk of relapse [46].

For the hotspot overview, a clear, unambiguous and reproducible identification of a mutation site is therefore needed, regardless of the different kinds of mutation descriptions. To achieve this, all mutation sites are stored in AMLVaran’s databases with precise genomic coordinates. As set of genomic ranges (from 1 bp to complete exons) can be associated with each mutation. Additionally, restrictions to certain mutation types like frameshift insertions can be specified. Any matching, non-synonymous, not benign variant entry located in the specified areas is then regarded as driver mutation.

However, it is equally important to reliably rule out the presence of relevant driver mutations. This requires, that no relevant mutation was found in the corresponding region, but also that the coverage of the entire region is sufficiently high. Only then a site is classified as “wild type” by AMLVaran.

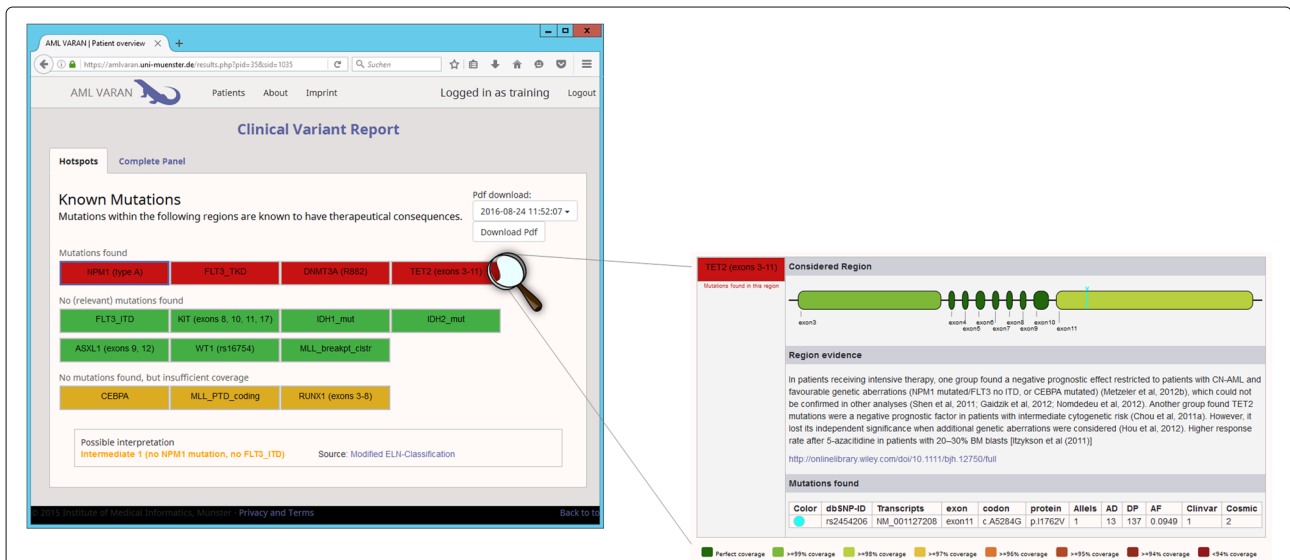
Implications defined in any disease specific treatment guidelines can be automatically derived on the basis of the detection or exclusion of these defined driver mutations, and are displayed as diagnostic recommendations. The rule-based system is stored within the database, and can be edited by the system’s administrator.

**Clinical reports**

In interviews with clinical users, a need for a quick and standardized way of transferring AMLVaran’s results into electronic healthcare or archiving systems was expressed. To fulfill this demand, a standardized PDF report as shown in Fig. 5) is created for each sample, containing the most relevant information for clinical diagnostics. In particular, it includes the detected or excluded driver mutations, the coverage distribution in the stored hotspot areas, as well as predefined curated information about clinical implications with references to corresponding studies. These reports are structured from coarse to fine [34], starting with an overview of the predefined hotspot mutation sites and their status, as well as displaying more detailed information on each of the hotspot areas. In order to ensure reproducibility of medical results, further details are provided, e.g. type and version of the analysis.

**Interactive variant explorer**

In addition to already known clinically relevant driver mutations, assessment of further mutations can be rele-



**Fig. 5** Visualization of a sample clinical report. AMLVaran’s clinical reports consist of an overview of the predefined mutation sites. For each site, a color-coded area shows one of three states. Red indicates that a relevant mutation was found in the given area, yellow intervals have no associated mutation, but the coverage is not sufficient to allow a safe exclusion of variants. Finally, the presence of mutations in green areas can reliably be excluded due to good coverage, and lack of mutation evidence. Based on the mutation status and the stored diagnostic rules, the applicable diagnostic information is displayed. Furthermore, an assessment summary can be added by the evaluating pathologist. The report can be augmented with more detailed information for each mutation site, including details on the detected mutations, the coverage in the examined area, and general information on the function of the considered gene

vant for certain cancer types [47]. Clinical users expressed interest in comprehensive and up-to-date information on these variants, and for interactive overviews in order to enable quick and accurate interpretation. While AML-Varan retrieves known clinical information from annotation databases like ClinVar or COSMIC, additional resources are required for those mutations. Hence, functional prediction tools, conservation scores and mutation frequency in sequencing projects, such as 1000 genomes, were integrated into the platform, and provide information on the potential pathogenicity of variants. The AML-Varan software integrates available information for each detected variant within an interactive variant explorer, together with adjustable filtering and sorting options (Fig. 6a). All outputs of the variant explorer, with applied scoring, filtering and sorting settings, can be exported for further processing in tabular form.

For each variant, a detailed view with all available annotation can be selected, including eight functional prediction scores. For each functional prediction tool, the Converted Rank Score from dbNSFP [48] is displayed to support assessment of harmfulness of the current mutation, compared to predictions of all other non-synonymous coding SNVs (Fig. 6b).

In addition, it is possible to inspect aligned reads for each mutation site with the integrated genome browser (IGV [50]).

## Results

AMLVaran was installed on a virtual server with high performance hardware, and on a standard desktop workstation. Detailed information regarding hardware configuration and runtime is provided in Additional file 1: Section 5. The software's performance was evaluated on AML and MDS datasets, and tested with regards to its sensitivity, positive predictive value, and general usability.

Three datasets were analyzed with the generic pipeline in its default appreci8-based configuration, with the following eight variant calling tools: GATK 3.3<sup>2</sup> [9], FreeBayes 1.0.2 [10], SamTools 1.3 [11], LoFreq 2.1.2 [51], Platypus 0.8.1 [52], SNVer 0.5.3 [53], VarScan 2.4.0 [54], and VarDict (Java) 1.5.5 [55].

For annotation and variant filtering process, the following databases were used: dbSNP v150 (2017-04-03) [35], ClinVar (2017-12-31) [37], COSMIC v83<sup>3</sup> (2017-11-07) [36], 1000Genomes v5b (all, Aug 2015) [39], ExAC 65000 v0.3 (all, 2015-11-29) [41], and ESP6500 (2014-12-22) [40].

Pre-calculated prediction scores were downloaded from PROVEAN v1.1 [28].

The test system with results on the described datasets are available at <https://amlvaran.uni-muenster.de> [56].

<sup>2</sup>GATK is only available for public usage from version 4 and higher due to license restrictions.

<sup>3</sup>COSMIC is not available for public usage due to license restrictions by the authors.

## Validation of the generic variant calling pipeline

### Validation strategy and source data

To validate the generic AMLVaran pipeline in its appreci8-like configuration, its output was compared to the original appreci8's output.

AMLVaran was used to analyze a set of three published datasets. Two of them were well characterized datasets from myelo-dysplastic diseases, for which biological validation data was available (MDS-1 and MDS-2). The third dataset (AML-1) was taken from routine AML diagnostics, which resembled the intended use case, but only limited validation data was available.

MDS-1 and MDS-2 were collected and sequenced within the European MDS Triage project and focused on patients with MDS. The latter one was provided by the University Hospital of Halle, and contained a mixed set of samples (67x AML, 1x AL, 2x t-AML, 2x s-AML after MDS, 1x MDS/AML unsure diagnosis, 17x MDS, 19x MPN and 10x unknown diagnosis).

Source data is available for download from the NCBI Sequence Read Archive (BioProjectID: PRJNA388411 [57]). The raw variant result tables are provided in Additional files 5, 6, 7 and 8.

### Evaluation on hotspot regions

For clinical use, it is crucial, that existing variants are detected with very high reliability. No less important, however, is the pipeline's ability to safely exclude the presence of mutations for a chosen position. This is of special importance for hotspot regions, which are typically designed to contain mutations relevant for treatment.

First, we evaluated results within these defined, therapy-relevant areas for all three datasets. For both MDS and AML datasets, there was a high overlap between the two pipelines and the biological truth. In total, 415 out of 419 variant calls matched for AMLVaran and appreci8 and the ground truth, leading to sensitivity  $\geq 0.99$ , and PPV  $\geq 0.99$ . The only two cases of discrepancy between AMLVaran and appreci8 were borderline cases that were reported by only two or three out of eight variant callers, respectively. Differences between dbSNP versions changed the ratings of those variants in AMLVaran's output.

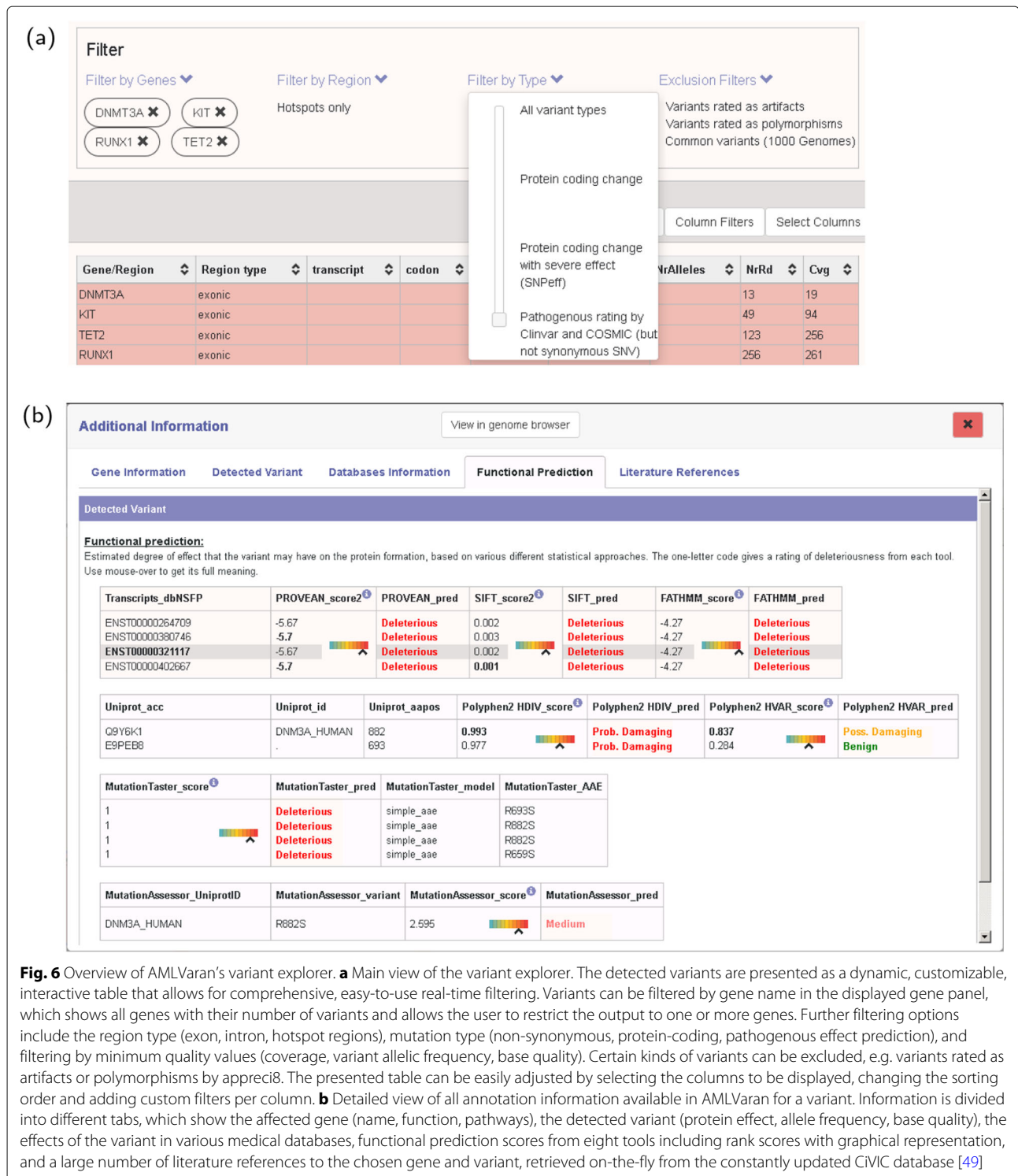
A detailed comparison of the detected variants within these regions is provided in Table 1.

### Evaluation on the complete target region

Next, we considered the whole sequenced target region of 42 kbp length, which includes segments of 19 genes. The target region's average coverage was 1450.

Almost all true positive calls from appreci8's reference implementation were confirmed by AMLVaran. For dataset MDS-1, 395 of 397 true variants were identified, and all 89 variants of the MDS-2 dataset overlapped in





both outputs. Likewise, 16 of appreci8's 17 false positive calls for dataset MDS-1 and all three false positive calls for dataset MDS-2 were matching in AMLVaran's results.

The observed differences can be explained by an improved normalization procedure and updated reference

databases. Due to these improvements, the sensitivity of AMLVaran's pipeline (0.98) was higher than of the reference implementation (0.95) for MDS-1. For MDS-1, 12 previously filtered true positive variants were correctly classified by AMLVaran, and another six new high-quality

**Table 1** Accumulated numbers of variants in hotspot regions detected by AMLVaran in comparison with appreci8 reference implementation

	MDS-1 (n=237, #Mut=195)		MDS-2 (n=46, #Mut=52)		AML-1 (n=119, #Mut=172)	
	appreci8	AMLVaran	appreci8	AMLVaran	appreci8	AMLVaran
Called variants	193	194	52	52	172	171
True positives	193	193	52	52	172	171
False positives	0	1	0	0	0	0
Sensitivity	0.9897	0.9897	1.0000	1.0000	1.0000	0.9942
PPV	1.0000	0.9948	1.0000	1.0000	1.0000	1.0000

candidate variants were identified, for which validation is pending.

In general, AMLVaran's pipeline was slightly more sensitive than the original appreci8 implementation, resulting in 12 additional false positive calls that were reported by AMLVaran. These differences were caused by changes in the updated dbSNP version (10 variants), and a corrected bug in appreci8's method for the counting of samples (2x).

For dataset MDS-2, AMLVaran recognized all 89 validated mutations. Another promising candidate mutation as well as five additional false positive calls were reported. Reasons for the differences were differing PM flag due to newer dbSNP version (three variants), an update to the integrated COSMIC version (2x), and improved InDel normalization (1x).

In total, AMLVaran's variant lists showed high concordance with the reference results for both MDS datasets. A detailed comparison of the results of both pipelines and the biological ground truth is shown in Fig. 7.

#### Evaluation of dataset AML-1

For dataset AML-1, the investigated target regions were much larger. In total, 523 kbp in 112 genes were analyzed, with an average coverage of 570 and over 90% of the unfiltered calls were artifacts/polymorphisms. Since this dataset originated from routine diagnostics, biological validation data was limited to a part of the samples and to the genes NPM1 ( $n = 75$ ), JAK2 ( $n = 10$ ), FLT3\_TKD ( $n = 50$ ), and CEBPA ( $n = 30$ ).

For these biologically validated sites, AMLVaran detected 51 out of 55 variants. In addition, AMLVaran did not report a single false positive call for any of the 110 validated wild-type alleles. While the program did report two unexpected calls for FLT3-TKD, these two variants were also reported by appreci8 and validated by expert inspection afterwards.

Compared to the biological truth, AMLVaran achieved a high number of overlapping variants, resulting in high sensitivities of 0.95 resp. 1.00 for the NPM1 and JAK2 variants. For FLT3-TKD, one variant was missed. A sensitivity of 0.67 or higher was achieved for CEBPA, a gene that is typically correlated to coverage issues and high numbers

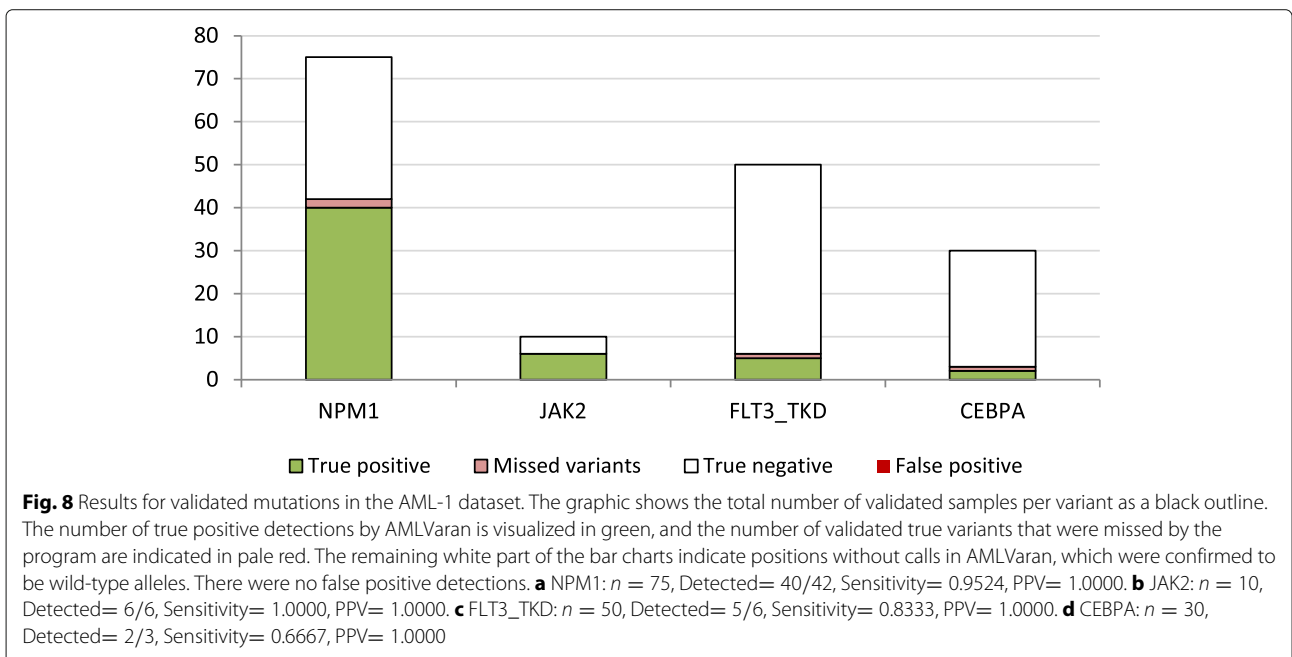
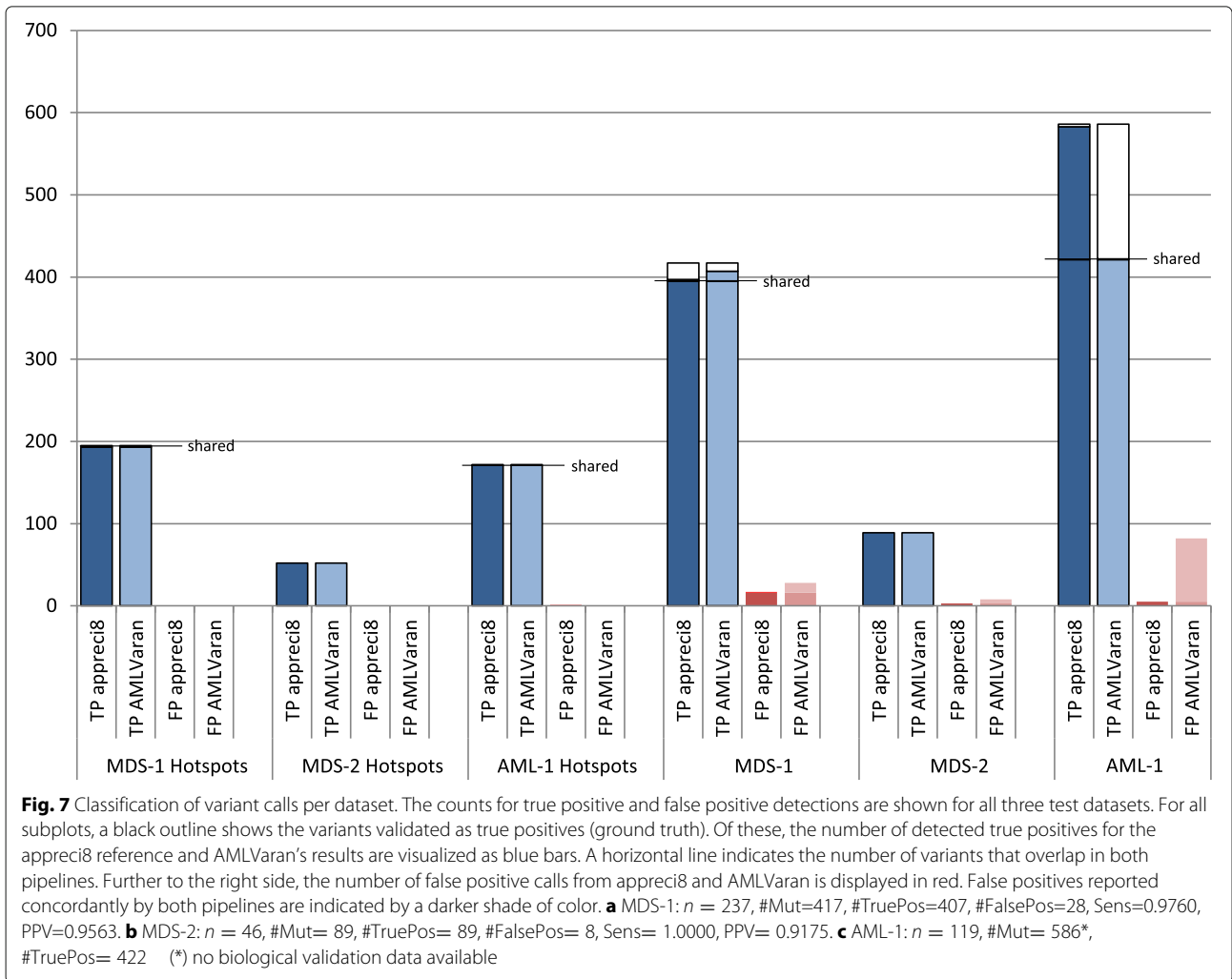
of false positive and false negative variants [58]. Detailed results for all groups of validated mutations are shown in Fig. 8.

In total, AMLVaran reported 504 variants in 119 samples for dataset AML-1, ranging from 0 to 13 variants per sample, and with an average of 4.2 calls per sample. In contrast to datasets MDS-1 and MDS-2, more differences between AMLVaran and appreci8 were noticeable. A set of 426 variants matched between AMLVaran and appreci8, however 78 variants were called by appreci8 only, and 162 additional variants were specific for AMLVaran. It stands to reason, that these differences are caused by the same effects that were discovered for datasets MDS-1 and MDS-2 (database updates and improvements to normalization technique). Results for validated variants indicated that deviations preferentially affected cases of high ambiguity. Hotspot mutations with high coverage or other unambiguous variants were rarely affected.

#### Evaluation of the system's usability

The development of AMLVaran was driven by a focus group, consisting of computer scientists as well as several clinical domain experts and medical professionals. The involvement of clinical experts was especially valuable to identify and prioritize user requirements. Two user studies were conducted during the development process to ensure that the software met critical requirements and offered good usability. In the first user study, a prototype of the software [59] was evaluated, while the final AMLVaran program was examined in the second study. 9 and 14 independent medical professionals were asked to fulfill application-related tasks using AMLVaran, respectively. Afterwards, participants were asked to rate their experience with regards to the software's general performance, functionality, and user-friendliness.

During the first test of the prototype, all 9 users had difficulties with the interpretation of the presence or severity of given hotspot mutations. On average only 4.3 of 8 test cases were solved correctly. None of the participants was able to answer all 8 cases correctly. Three persons achieved 6 of 8 points, and three persons 3 points. From these test results and unstructured interviews with the



test candidates, six major usability issues were identified, which are described in Additional file 1: Section 3.

Detailed feedback from the study was used to address the mentioned key factors and to optimize a number of features during the continued development of the software. Data presentation was improved by a new, clear structuring of related views. A detailed view for each variant was composed and grouped by categories (Fig. 6b). The use of more intuitive headings, tool-tips, additional explanatory texts and improved predefined filtering options helped to increase the interpretability of AMLVaran's results. A major improvement was the introduction of the central overview of the mutation status for known driver mutations (Fig. 5). In this overview, both detection and exclusion of these mutations were represented as intuitive traffic light color codes.

After these improvements, a second user study was carried out with a different user group of 14 experts, and updated tasks. Regarding the interpretation of hotspot mutations, 5 out of 6 test cases were now correctly interpreted by all 14 participants. The sixth test case was excluded due to ambiguity of the data. In addition to the improved performance, the second group of users also reported a higher degree of confidence while using the software during the interviews.

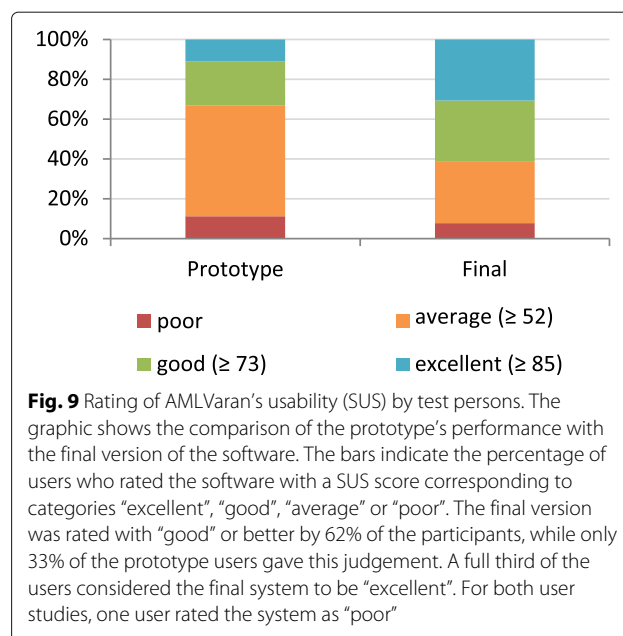
Finally, the perceived usability of the program was rated by participants of both studies with the help of a standardized questionnaire (System Usability Scale [60]), which provides a score between 0 and 100 (100 = best usability). AMLVaran's prototype achieved an average score of 69 (median: 67, interquartile range: 14). The final version of the software reached an average score of 74 (median: 80, IQR: 28), which indicated a notable increase in user-friendliness. For the second study, 2 out of 14 test persons had to be excluded from the SUS rating, because they did not fully complete the test.

According to a study of Bangor et al. [61], scores between 60 and 80 are typical results for average software systems. Systems with scores of 73 or above are considered as "good", 85 or greater signals "excellent" usability. It should be noted, that users received no training before carrying out the study. Figure 9 shows the distribution of scores assigned by the participants of both user studies.

While the increase of the usability score was only moderate, the overall complexity of the software and the difficulty of tasks increased in the final version. Since these tasks required a higher degree of coordination from the users, the results indicate, that the second system version was more usable.

## Discussion

High sensitivity and precision of variant calling are basic, but critical requirements for a successful application to clinical diagnostics. In the presented setting, sensitivity



$\geq 0.98$  and  $PPV \geq 0.92$  ( $\geq 0.99$  for hotspot regions) were reached for validated datasets, which should be generally sufficient for clinical use. However, the signal quality and the incidence of mutations per gene needs to be considered for a thorough interpretation.

Considerable differences between AMLVaran and the appreci8 reference implementation were limited to dataset AML-1, and were largely caused by an improved normalization method and different versions of annotation databases. For updated versions of appreci8, we therefore expect smaller differences between this pipeline and AMLVaran.

Normalization of the different callers' variant lists before combination, as well as the use of newer annotation databases, had a relevant influence on the outcome of the filtering process. In particular, the PM flag in the dbSNP database, which plays a crucial role in the appreci8 filtering scheme, is subjected to frequent changes, making the pipeline sensitive to database updates. Although biological validation was lacking for the AML-1 dataset, we assume, that the pipeline optimizations led to improved precision in comparison to the original appreci8 output due to the apparent improvements in the newer dbSNP version.

The deviations on dataset AML-1 show, that the output of individual callers and different pipelines can differ greatly, even for low-volume targeted sequencing data. We demonstrated, that existing variant calling pipelines are very sensitive to small adjustments of analysis parameters, an effect that is most notable in regions with lower data quality. It is therefore important, that these parameters are chosen based on a sufficient amount of valida-

tion data for each dataset to account for differences in sequencing depth, differences in sample preparation, and other parameters of data generation. This is especially important for the design of a complete variant calling pipeline, which combines several variant calling tools to achieve optimal results.

In general, the performance of the presented pipeline can be improved through the incorporation of additional information regarding disease type and hotspot data. AMLVaran's default appreci8-like analysis pipeline takes such information into account. In our datasets, the sensitivity and precision of variant calling was higher for hotspots than on the whole target region. Even with dataset AML-1, 171 of 172 calls in hotspot areas were matching between both pipelines. This might be explained by the fact, that some regions of the genome are difficult to sequence in general and therefore the knowledge about these regions has limitations.

For clinical use, not all detectable mutations are of the same relevance. In this scenario, it is often more important to identify or exclude therapy-relevant mutations than finding variants of unknown significance. Whole-genome sequencing with lower coverage and without special consideration of hotspots is therefore expected to yield data that is less suited for highly sensitive variant calling in regions of therapeutic interest. Consequently, at present, targeted sequencing appears to be the best choice for clinical diagnostics.

### Strengths

To the best of our knowledge, AMLVaran is the first freely available web-based software that combines a sophisticated but generic variant calling pipeline, a complex filter strategy, and disease-specific evaluation of predefined mutations in a form that is suitable for clinical application.

The novel Scoring Scheme Definition Language (SSDL) allows a standardized, reproducible, cross-platform definition of complex variant filtering systems.

Both, sensitivity and precision of the variant calling reach the specified goal of 0.95 or higher for hotspot regions in all three evaluated datasets.

The software not only supports confirmation, but also exclusion of predefined driver mutations by means of an integrated coverage analysis. This aspect is often lacking in common variant calling pipelines.

Since personalized medicine and genetic diagnostics is a moving target, frequent changes and updates of medical knowledge have to be taken into account to guarantee optimal variant filtering and interpretation. AMLVaran stores variants and annotation data separately. This flexible annotation approach enables to add, update or remove complete annotation databases at any time. There is no need to explicitly re-annotate previously analyzed samples after a database update.

The databases stored on AMLVaran's server can also be used for independent real-time annotation without a full AMLVaran analysis [32].

In addition to the functionality, usability of a software system is a key success factor with regards to its practical use. We therefore assessed the usability of AMLVaran with medical experts, and refined the software iteratively. The proportion of users which rated the software with "good" or better increased from 33% to 62%. This clearly demonstrates the need for usability studies in clinical NGS software.

It stands to reason, whether a probabilistic approach of variant prediction could be employed instead of binary classification, in order to further improve prediction quality, especially for borderline variants. A detailed discussion of this idea is included in Additional file 1: Section 4.

### Limitations

The distinction between somatic mutations and artifacts or polymorphisms is a complex challenge, especially for tumor samples with potential subclones and lower expected variant allele frequencies. We chose a combination approach of several variant callers and demonstrated, that AMLVaran's results had a high overlap with the reference pipeline's variant lists. However, a productive implementation of AMLVaran into a clinical routine setting requires further validation through established sequencing methods (e.g. Sanger sequencing [62]).

Currently, AMLVaran can detect only SNVs and small InDels at the required precision rate for clinical use. Gene fusions or larger tandem duplications, such as FLT3-ITD, are much more difficult to detect by targeted data. However, McKerrel et al. have shown, that it is technically possible to detect structural variants from targeted NGS data, as long as a suitable sequencing panel is used [14]. With appropriate data and choice of precise detection tools, AMLVaran can be extended to detect those variants as well.

AMLVaran has been designed with a clear focus on analyzing single samples for certain diagnostically relevant mutations. Family-based analyses or investigation of new mutations in a cohort of samples are not in focus of this tool. The variant calling algorithms have been optimized for and tested with tumor data only, which requires special care due to mixture of cell material and high intra-tumor heterogeneity. Especially, somatic mutations may be observed at arbitrary low frequencies in tumor probes [63–65]. This is, why a highly optimized cancer-pipeline (appreci8) has been selected for AMLVaran, that utilizes eight callers to provide best sensitivity at the cost of a higher runtime.

AMLVaran might be applied for germline studies as well, but the variant calling and filtering algorithms should be adapted for this purpose.



### Transferability / universality

AMLVaran is a generic application for NGS variant analysis, as it offers a freely customizable scoring scheme interface, which can be used to define a multitude of score calculation and filter schemes. Furthermore, it can be equipped with any variant calling tools, and different targeted sequencing panels are available for selection. Custom panels can be imported via the web interface as well.

To adapt AMLVaran for other diseases, predefined driver mutation sites have to be changed and the rule system for treatment guidelines is to be adapted in the database.

Although AMLVaran has been developed and optimized for targeted NGS sequencing, it may be used for whole-exome (WES) data as well. Some notes on recommended adjustments, as well as a comparison of runtimes and number of detections when used with a WES germline sample can be found in Additional file 1: Section 6.

### Future work

In a pre-clinical user study, AMLVaran was well accepted by medical experts. The next step towards clinical use would be prospective evaluation and validation in a suitable clinical setting. At a later stage, integration in the Electronic Health Record system would be advantageous.

### Conclusions

With AMLVaran, we have developed a software system that enables variant analysis on NGS sequencing data and fulfills central requirements for the utilization in clinical routine care.

The presented software platform performs highly reliable variant calling, and achieves sensitivities of at least 0.95 on several datasets, as well as precision scores of at least 0.92, with a minimum precision of 0.99 on hotspot regions. None of the eight variant callers was able to achieve such high values on the tested datasets, when applied alone. This underlines the need for validation of present NGS pipelines before use.

AMLVaran is flexible enough to employ and combine arbitrary variant calling tools, and provides a user-friendly interface that has been evaluated in order to facilitate use by medical staff. The system's performance is sufficient to perform analyses on standard desktop personal computers within less than one hour for targeted data, and within 24 hours for WES data.

From our perspective, the approach of AMLVaran can be a valuable contribution for use of NGS data in a clinical context. We demonstrated that fine-tuning of NGS pipelines with reproducible filtering of NGS results and usability analysis is needed to provide valid variant calling results which are comprehensible for medical users.

### Availability and requirements

**Project Name:** AMLVaran

**Project home page:** <https://amlvaran.uni-muenster.de> [56]

**Source code repository:** <https://github.com/cwuensch/AMLVaran> [66]

**Operating system(s):** Linux

**Programming language:** PHP, JavaScript, Python

**Other requirements:** Apache web server, MySQL, Java 1.8, Python 2.7, VariantTools 2.7.0, samtools 1.3, vcftools 0.1.15, SNPeff 4.2, bam-readcount 0.8.0, reference genomes and the variant calling tools

**License:** GNU GPL v3

**Any restrictions to use by non-academics:** none

The presented web platform AMLVaran is available for demonstration purposes at <https://amlvaran.uni-muenster.de> [56].

The following credentials can be used to view the published test datasets:

MDS-1: user: "test1", password: "Sweden2017".

MDS-2: user: "test2", password: "Sweden2017".

AML-1: user: "test3", password: "Halle2015".

(Upload and analysis of own samples is purely for research purposes and only possible after contact with the corresponding author.)

The source code is freely available from GitHub at <https://github.com/cwuensch/AMLVaran> [66], and can be adapted for new projects under the provided license.

A Docker installation script is available that guides a bioinformatician through the process of assembling and configuration of the software components.

A QuickStart Guide that helps with setup and usage of the software is provided in Additional file 2.

**Notice:** This is open source test code. The software components need to be adapted to local requirements. Especially the variant calling parameters are to be adapted for the type of data to be used. Importantly, the system needs to be assembled and validated locally before use! This code is provided "AS IS" and any express or implied warranties, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose are disclaimed.

**Datasets:** The datasets used for validation of the software are available for download from the NCBI Sequence Read Archive (BioProjectID: PRJNA388411; <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA388411> [57]).

### Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s12920-020-0668-3>.

**Additional file 1:** A PDF document with additional detailed information.

**Additional file 2:** A PDF document, giving basic installation and usage instructions for the software.

**Additional file 3:** The variant calling tools and parameters, as well as filter scheme used to generate the described results.

**Additional file 4:** AMLVaran's default variant scoring and filter scheme which resembles the appreci8 algorithm.

**Additional file 5:** Variant calling results from dataset MDS-1 (MS Excel)

**Additional file 6:** Variant calling results from dataset MDS-2 (MS Excel)

**Additional file 7:** Variant calling results from dataset AML-1 (MS Excel)

**Additional file 8:** Variants and validation for dataset AML-1 (MS Excel)

## Abbreviations

#Mut: Number of mutations; #TruePos: Number of true positive calls; A/P-score: Artifact / Polymorphism score; AF: Allele frequency; AL: Acute leukemia; AML: Acute myeloid leukemia; AMLVaran: AML variant analyzer; bam: Binary aligned map; bed: Browser extensible data (a list of genomic regions); bp: base pairs; CPU: Central processing unit; DNA: Deoxyribonucleic acid; e.g.: exempli gratia - for example; EHR: Electronic health record; GB: Gigabyte; GHz: Gigahertz; InDel: Insertion / Deletion; IQR: Interquartile range; JSON: JavaScript object notation; kbp: kilo base pairs; MDS: Myelo-dysplastic syndrome; MPN: Myeloproliferative neoplasm; NGS: Next-Generation sequencing; PDF: Portable document format; PM: PubMed; PPV: Positive predictive value; RAM: Random access memory; resp: respective; SNP: Single nucleotide polymorphism; SNV: Single nucleotide variant; SUS: System usability scale; SSDL: Scoring scheme definition language; UTR: Untranslated region; VBA: Visual basic for applications; vcf: Variant call format; WES: Whole-exome sequencing; WGS: Whole-genome sequencing; WHO: World Health Organization

## Acknowledgements

Jan Stenner, Sebastian Windau and Henning Ridder helped with coding parts of the website. Sarah Sandmann, developed appreci8, on which our default score calculation scheme is based. Bo Peng, developed VariantTools, which is used for normalization and integration of the results of different caller tools, and gave a lot of support with usage. Carolin Walter helped with proof-reading and language editing of the manuscript. V. Gradenwitz, L. Möller, B. Wünsch also helped with proof-reading the manuscript.

## Authors' contributions

[CW] designed the software and database architecture, implemented the generic variant calling pipeline and generic scoring scheme format as well as the appreci8 filtering scheme, and performed the analysis of the described datasets as well as the user studies. [HB] did the medical research, identified the known driver mutations and guidelines for AML. [CMT] provided the dataset AML-1 with 119 patient samples from routine care. [MD] coordinated the project, supervised the work and revised the manuscript. All authors read and approved the final manuscript.

## Funding

This work was supported by European Union grants Triage-MDS [ERA-Net TRANSCAN BMBF 01KT1401] and -Horizon2020 MDS-RIGHT [634789]; Deutsche Krebshilfe grant 'Verbesserung der Diagnostik von Tumorerkrankungen durch neue DNA-Sequenzierverfahren und Algorithmen' [110495]. There was no influence of the funding agency on the design of the study or on collection, analysis, and interpretation of data or in writing the manuscript.

## Availability of data and materials

The datasets supporting the conclusions of this article are available from the NCBI Sequence Read Archive (BioProjectID: 388411 [57]).

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Author details

<sup>1</sup>Institute for Medical Informatics, University of Münster, Albert-Schweitzer-Campus 1, Building A11, Münster, Germany. <sup>2</sup>Department

of Medicine V, Hematology, Oncology and Rheumatology, University of Heidelberg, Heidelberg, Germany.

Received: 19 August 2019 Accepted: 21 January 2020

Published online: 04 February 2020

## References

- Maes B, Willemse J, Broekmans A, Smets R, Cruys B, Put N, Madoe V, Janssen M, Soepenbergh O, Bries G, Vrelust I, Achten R, van Pelt K, Buvé K, Theunissen K, Peeters V, Froyen G. Targeted next-generation sequencing using a multigene panel in myeloid neoplasms: Implementation in clinical diagnostics. *Int J Lab Hematol*. 2017;39(6):604–12. <https://doi.org/10.1111/ijlh.12709>.
- Duncavage EJ, Tandon B. The utility of next-generation sequencing in diagnosis and monitoring of acute myeloid leukemia and myelodysplastic syndromes. *Int J Lab Hematol*. 2015;37 Suppl 1:115–21. <https://doi.org/10.1111/ijlh.12361>.
- Ruffalo M, Husseinzadeh H, Makishima H, Przychodzen B, Ashkar M, Koyutürk M, Maciejewski JP, LaFramboise T. Whole-exome sequencing enhances prognostic classification of myeloid malignancies. *J Biomed Informa*. 2015;58:104–113. <https://doi.org/10.1016/j.jbi.2015.10.003>.
- Bacher U, Shumilov E, Flach J, Porret N, Joncourt R, Wiedemann G, Fiedler M, Novak U, Amstutz U, Pabst T. Challenges in the introduction of next-generation sequencing (NGS) for diagnostics of myeloid malignancies into clinical routine use. *Blood Cancer J*. 2018;8(11):113. <https://doi.org/10.1038/s41408-018-0148-6>.
- Martelli MP, Sportoletti P, Tiacci E, Martelli MF, Falini B. Mutational landscape of AML with normal cytogenetics: Biological and clinical implications. *Blood Reviews*. 2013;27(1):13–22. <https://doi.org/10.1016/j.blre.2012.11.001>.
- Liersch R, Müller-Tidow C, Berdel WE, Krug U. Prognostic factors for acute myeloid leukaemia in adults—biological significance and clinical use. *British J Haematol*. 2014;165(1):17–38. <https://doi.org/10.1111/bjh.12750>.
- Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, Massouras A. VarSome: The human genomic variant search engine. *Bioinformatics (Oxford, England)*. 2019;35(11):1978–80. <https://doi.org/10.1093/bioinformatics/bty897>.
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*. 2009;25(14):1754–60. <https://doi.org/10.1093/bioinformatics/btp324>.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–303. <https://doi.org/10.1101/gr.107524.110>.
- Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. <http://arxiv.org/pdf/1207.3907v2>.
- Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*. 2011;27(21):2987–93. <https://doi.org/10.1093/bioinformatics/btr509>.
- Cingolani P, Platts A, Le Wang L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 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*. 2012;6(2):80–92. <https://doi.org/10.4161/fly.19695>.
- Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):164. <https://doi.org/10.1093/nar/gkq603>.
- McKerrell T, Moreno T, Ponstingl H, Bolli N, Dias JML, Tischler G, Colonna V, Manasse B, Bench A, Bloxham D, Herman B, Fletcher D, Park N, Quail MA, Manes N, Hodkinson C, Baxter J, Sierra J, Foukaneli T, Warren AJ, Chi J, Costeas P, Rad R, Huntly B, Grove C, Ning Z, Tyler-Smith C, Varela I, Scott M, Nomdedeu J, Mustonen V, Vassiliou GS. Development and validation of a comprehensive genomic diagnostic tool for myeloid malignancies. *Blood*. 2016;128(1):1–9. <https://doi.org/10.1182/blood-2015-11-683334>.
- Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, Grünig B, Guerler A, Hillman-Jackson J, von Kuster G, Rasche E, Soranzo N, Turaga N, Taylor J, Nekrutenko A, Goecks J. The Galaxy platform for accessible,

- reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 2016;44(W1):3–10. <https://doi.org/10.1093/nar/gkw343>.
16. Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, Koski M, Käki J, Korpelainen E. Chipster: User-friendly analysis software for microarray and other high-throughput data. *BMC genomics.* 2011;12:507. <https://doi.org/10.1186/1471-2164-12-507>.
  17. Chiara M, Gioiosa S, Chillemi G, D'Antonio M, Flati T, Picardi E, Zambelli F, Horner DS, Pesole G, Castrignanò T. CoVaCS: A consensus variant calling system. *BMC genomics.* 2018;19(1):120. <https://doi.org/10.1186/s12864-018-4508-1>.
  18. Vandeweyer G, van Laer L, Loeys B, van den Bulcke T, Kooy RF. VariantDB: A flexible annotation and filtering portal for next generation sequencing data. *Genome Med.* 2014;6(10):74. <https://doi.org/10.1186/s13073-014-0074-6>.
  19. Starlinger J, Pallaraz S, Ševa J, Rieke D, Sers C, Keilholz U, Leser U. Variant information systems for precision oncology. *BMC Med Informa Decision Making.* 2018;18(1):107. <https://doi.org/10.1186/s12911-018-0665-z>.
  20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England).* 2009;25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.
  21. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 2010;38(6):1767–71. <https://doi.org/10.1093/nar/gkp1137>.
  22. Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, Dugas M. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Sci Rep.* 2017;7:43169. <https://doi.org/10.1038/srep43169>.
  23. Trubetskoy V, Rodriguez A, Dave U, Campbell N, Crawford EL, Cook EH, Sutcliffe JS, Foster I, Madduri R, Cox NJ, Davis LK. Consensus Genotyper for Exome Sequencing (CGES): Improving the quality of exome variant genotypes. *Bioinformatics (Oxford, England).* 2015;31(2):187–93. <https://doi.org/10.1093/bioinformatics/btu591>.
  24. Sandmann S, Karimi M, de Graaf AO, Rohde C, Göllner S, Varghese J, Ernsting J, Walldin G, van der Reijden BA, Müller-Tidow C, Malcovati L, Hellström-Lindberg E, Jansen JH, Dugas M. appreci8: A Pipeline for Precise Variant Calling Integrating 8 Tools. *Bioinformatics (Oxford, England).* 2018. <https://doi.org/10.1093/bioinformatics/bty518>.
  25. Field MA, Cho V, Andrews TD, Goodnow CC. Reliably Detecting Clinically Important Variants Requires Both Combined Variant Calls and Optimized Filtering Strategies. *PLoS one.* 2015;10(11):0143199. <https://doi.org/10.1371/journal.pone.0143199>.
  26. San Lucas FA, Wang G, Scheet P, Peng B. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics (Oxford, England).* 2012;28(3):421–2. <https://doi.org/10.1093/bioinformatics/btr667>.
  27. The McDonnell Genome Institute. bam-readcount. (Mar. 9, 2017). <https://github.com/genome/bam-readcount>. Accessed 15 Jul. 2019.
  28. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS one.* 2012;7(10):46688. <https://doi.org/10.1371/journal.pone.0046688>.
  29. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. The variant call format and VCFtools. *Bioinformatics (Oxford, England).* 2011;27(15):2156–8. <https://doi.org/10.1093/bioinformatics/btr330>.
  30. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. *Bioinformatics (Oxford, England).* 2015;31(13):2202–4. <https://doi.org/10.1093/bioinformatics/btv112>.
  31. Wünsch C, Sandmann S, Windau S, Dugas M. AML-Varan – a generic approach to integrate a multi-tool-combination based NGS variant calling pipeline into a web-based diagnostics platform. <https://doi.org/10.7490/F1000RESEARCH.1114839.1>.
  32. Wünsch C, Dugas M. REST-based annotation and filtering of Next-Generation Sequencing variant lists with integration into standard office tools. <https://doi.org/10.7490/F1000RESEARCH.1116119.1>.
  33. Annoserv. <http://annoserv.uni-muenster.de>. Accessed 05 Dec. 2019.
  34. Cutting E, Banchemo M, Beitelshes AL, Cimino JJ, Fiol GD, Gurses AP, Hoffman MA, Jeng LJB, Kawamoto K, Kelemen M, Pincus HA, Shuldiner AR, Williams MS, Pollin TI, Overby CL. User-centered design of multi-gene sequencing panel reports for clinicians. *J Biomed Informa.* 2016;63:1–10. <https://doi.org/10.1016/j.jbi.2016.07.014>.
  35. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29(1):308–11.
  36. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, Stefancsik R, Harsha B, Kok CY, Jia M, Jubb H, Sondka Z, Thompson S, De T, Campbell PJ. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 2017;45(D1):777–83. <https://doi.org/10.1093/nar/gkw1121>.
  37. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018;46(D1):1062–7. <https://doi.org/10.1093/nar/gkx1153>.
  38. Liu X, Jian X, Boerwinkle E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation.* 2011;32(8):894–9. <https://doi.org/10.1002/humu.21517>.
  39. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature.* 2015;526(7571):68–74. <https://doi.org/10.1038/nature15393>.
  40. Exome Variant Server: NHLBI GO Exome Sequencing Project (ESP). Seattle, WA. (Nov. 3, 2014). <http://evs.gs.washington.edu/EVS/>. Accessed 15 Jul. 2019.
  41. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Beohrne M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536(7616):285–91. <https://doi.org/10.1038/nature19057>.
  42. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003;31(13):3812–4.
  43. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, Dombret H, Ebert BL, Fenaux P, Larson RA, Levine RL, Lo-Coco F, Naoe T, Niederwieser D, Ossenkoppele GJ, Sanz M, Sierra J, Tallman MS, Tien H-F, Wei AH, Löwenberg B, Bloomfield CD. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood.* 2017;129(4):424–47. <https://doi.org/10.1182/blood-2016-08-733196>.
  44. El Ghannam D, Taalab MM, Ghazy HF, Eneen AF. DNMT3A R882 mutations in patients with cytogenetically normal acute myeloid leukemia and myelodysplastic syndrome. *Blood cells, molecules & diseases.* 2014;53(1-2):61–6. <https://doi.org/10.1016/j.bcmed.2014.01.004>.
  45. Falini B, Mecucci C, Tiacci E, Alcalay M, Rosati R, Pasqualucci L, La Starza R, Diverio D, Colombo E, Santucci A, Bigerna B, Pacini R, Pucciarini A, Liso A, Vignetti M, Fazi P, Meani N, Pettirossi V, Saglio G, Mandelli F, Lo-Coco F, Pelicci P-G, Martelli MF. Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *New Engl J Med.* 2005;352(3):254–66. <https://doi.org/10.1056/NEJMoa041974>.
  46. Care RS, Valk PJM, Goodeve AC, Abu-Duhier FM, Geertsma-Kleinkoort WMC, Wilson GA, Gari MA, Peake IR, Lowenberg B, Reilly JT. Incidence and prognosis of c-KIT and FLT3 mutations in core binding factor (CBF) acute myeloid leukaemias. *Brit J Hematol.* 2003;121(5):775–7.
  47. Chern JY, Lee SS, Frey MK, Lee J, Blank SV. The influence of BRCA variants of unknown significance on cancer risk management decision-making. *J Gynecol Oncol.* 2019;30(4):60. <https://doi.org/10.3802/jgo.2019.30.e60>.
  48. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human mutation.* 2016;37(3):235–41. <https://doi.org/10.1002/humu.22932>.
  49. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, Ainscough BJ, Ramirez CA, Rieke DT, Kujan L, Barnell EK, Wagner AH,

- Skidmore ZL, Wollam A, Liu CJ, Jones MR, Bilski RL, Lesurf R, Feng Y-Y, Shah NM, Bonakdar M, Trani L, Matlock M, Ramu A, Campbell KM, Spies GC, Graubert AP, Gangavarapu K, Eldred JM, Larson DE, Walker JR, Good BM, Wu C, Su AI, Dienstmann R, Margolin AA, Tamborero D, Lopez-Bigas N, Jones SJM, Bose R, Spencer DH, Wartman LD, Wilson RK, Mardis ER, Griffith OL. CIVIC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genet.* 2017;49(2):170–4. <https://doi.org/10.1038/ng.3774>.
50. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinforma.* 2013;14(2):178–92. <https://doi.org/10.1093/bib/bbs017>.
51. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* 2012;40(22):11189–201. <https://doi.org/10.1093/nar/gks918>.
52. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, McVean G, Lunter G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genet.* 2014;46(8):912–8. <https://doi.org/10.1038/ng.3036>.
53. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. SNVer: A statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.* 2011;39(19):132. <https://doi.org/10.1093/nar/gkr599>.
54. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics (Oxford, England).* 2009;25(17):2283–5. <https://doi.org/10.1093/bioinformatics/btp373>.
55. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, Dry JR. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* 2016;44(11):108. <https://doi.org/10.1093/nar/gkw227>.
56. AMLVaran. <https://amlvaran.uni-muenster.de>. Accessed 05 Dec. 2019.
57. NCBI Sequence Read Archive: BioProjectID: 388411. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA388411>. Accessed 10 Dec. 2019.
58. Yan B, Hu Y, Ng C, Ban KHK, Tan TW, Huan PT, Lee P-L, Chiu L, Seah E, Ng CH, Koay ES-C, Chng W-J. Coverage analysis in a targeted amplicon-based next-generation sequencing panel for myeloid neoplasms. *J Clin Pathol.* 2016;69(9):801–4. <https://doi.org/10.1136/jclinpath-2015-203580>.
59. Wünsch C, Banck H, Stenner J, Dugas M. AML-Varan – a web-based platform to display and analyze genomic variants from targeted NGS sequencing data. <https://doi.org/10.7490/F1000RESEARCH.1113145.1>.
60. Brooke J. Usability Evaluation in Industry. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. London: Taylor and Francis; 1996. p. 189–94.
61. Bangor A, Kortum PT, Miller JT. An Empirical Evaluation of the System Usability Scale. *Int J Human-Comput Interaction.* 2008;24(6):574–94. <https://doi.org/10.1080/10447310802205776>.
62. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc National Acad Sci USA.* 1977;74(12):5463–7.
63. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput Struct Biotechnol J.* 2018;16:15–24. <https://doi.org/10.1016/j.csbj.2018.01.003>.
64. Josephidou M, Lynch AG, Tavaré S. multiSNV: A probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples. *Nucleic Acids Res.* 2015;43(9):61. <https://doi.org/10.1093/nar/gkv135>.
65. Shiraiishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y, Sanada M, Ogawa S, Miyano S. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* 2013;41(7):89. <https://doi.org/10.1093/nar/gkt126>.
66. GitHub: Project AMLVaran. <https://github.com/cwuensch/AMLVaran>. Accessed 05 Dec. 2019.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

