CANCER GENOMICS & PROTEOMICS *18*: 157-166 (2021)
doi:10.21873/cgp.20250

Comparison of Open-access Databases for Clinical Variant Interpretation in Cancer: A Case Study of MDS/AML

HENRIK BANCK[1], MARTIN DUGAS[1], CARSTEN MÜLLER-TIDOW[2] and SARAH SANDMANN[1]

[1]Institute of Medical Informatics, University of Münster, Münster, Germany;
[2]Medizinische Klinik, Abteilung Innere Medizin V, University Hospital Heidelberg, Heidelberg, Germany

Abstract. *Background: Recently, guidelines for variant interpretation in cancer have been established. However, these guidelines do not mention which databases are most suited to performing this task. Materials and Methods: We give an overview of existing databases and evaluate their benefit in practical application. We compared three meta-databases and 12 databases for a dataset of patients with myelodysplastic syndrome or acute myeloid leukemia. Results: Clinical implications were found for 13% of all variants. One-third of variants with therapeutic implications were uniquely contained in one database. The VICC meta-database was the most extensive source of information, featuring 92% of variants with a drug association. However, a comparison of meta-databases and original sources indicated that some variants are missing in those meta-databases. Conclusion: Public databases provide decision support for interpreting variants but there is still need for manual curation. Meta-databases facilitate the use of multiple resources but should be interpreted with caution.*

With the development of next-generation sequencing, a growing amount of data is generated. Analysis of these data results in a long list of variants for each patient. However, the clinical significance of these variants needs to be assessed. In order to cope with this amount of data, there is a need for knowledge bases to filter, annotate and foremost interpret clinically relevant variants (1-3).

A vast repertoire of existing databases were mostly used for research on hereditary diseases in the beginning. These databases come with certain limitations, *e.g.* they lack standardized interpretation and nomenclature of variants (4). Recently, new databases have been emerging with a focus on clinical interpretation. In order to establish generally acknowledged standards for variant reporting, the Joint Consensus Recommendation of the Association for Molecular Pathology, the American Society of Clinical Oncology and the College of American Pathologists (CAP) have published guidelines for sequence variant interpretation (5).

On this basis, there have been recent activities in harmonizing the classification of clinical variants. The Global Alliance for Genomics and Health's Variant Interpretation for Cancer Consortium (VICC) has published standards for genomic data sharing and provided and a classification system called Harmonized Evidence Levels shown in Table I (6).

These classifications establish a guideline for the interpretation of variants. However, it remains unclear which resources are the most suited for performing this task.

In this article, we evaluated open-access databases for interpreting variants in cancer regarding their clinical significance. We reviewed literature to give an overview of the existing variant databases and selected databases meeting the requirements for annotation of single nucleotide variants and insertions or deletions (indels). Furthermore, we assessed their quality for clinical decision support by analyzing an exemplary next-generation sequencing dataset of patients with myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML). For this purpose, we adapted the classifications of the Joint Consensus Recommendation and Harmonized Evidence Levels and developed a modified classification system for the interpretation of clinically relevant variants, adjusted to the dimensions and possibilities of the available databases. More specifically, we distinguish variants with therapeutic implications from disease-associated variants. Furthermore, we propose a novel subdivision of variants with unclear significance.

The aim of our study was to identify the most useful databases for variant interpretation by evaluating available information on variants with therapeutic or diagnostic implications. Furthermore, we aimed to identify typical

This article is freely accessible online.

Correspondence to: Sarah Sandmann, Institute of Medical Informatics, University of Münster, Albert-Schweitzer-Campus 1, Building A11, Münster, Germany. E-mail: sarah.sandmann@uni-muenster.de

Key Words: Variant interpretation, mutation databases, somatic variants in cancer.

Table I. *Variant classification by the Joint Consensus Recommendation (JCR) and by the Harmonized Evidence Levels (HEL). These classifications distinguish between variants with clinical actionability (tier I and II, graded by evidence level A-D), variants of unknown clinical significance (tier III) and benign variants (tier IV).*

Category by JCR	Definition	Category by HEL	Definition
Tier I	Variants with strong clinical significance	Level A	Evidence from professional guidelines or Food and Drug Administration-approved therapies relating to a biomarker and disease.
		Level B	Evidence from clinical trials or other well-powered studies in clinical populations, with expert consensus.
Tier II	Variants with potential clinical significance	Level C	Evidence for therapeutic predictive markers from case studies, or other biomarkers from several small studies. Also, evidence for biomarker therapeutic predictions for established drugs for different indications.
		Level D	Preclinical findings or case studies of prognostic or diagnostic biomarkers. Also includes indirect findings.
Tier III	Variants of unknown clinical significance		
Tier IV	Benign/likely benign variants		

hindrances and pitfalls when annotating and interpreting variants, considering different databases.

Materials and Methods

Search strategy. We searched PubMed (7) for reviews about knowledge bases using combinations of the following key words: Clinical actionability, clinical relevance, databases, next-generation sequencing, somatic variant classification, variant interpretation, pathogenicity prediction. In addition, we screened subject-specific forums and websites for databases about variants with clinical actionability *e.g.* omictools website (8), biostars forum (9), bioconductor.org website (10, 11), National Center for Biotechnology Information (NCBI) web page (12), My Cancer Genome website (13) and the web page of the VICC (14). Furthermore, we collected information on databases used by annotation tools or pipelines such as Annovar (15), Ensembl Variant Effect Predictor (16), varianttools (17), and SnpEff (18), as well as meta-databases, which incorporate other databases, *e.g.* myvariant.info (19) or the VICC's meta-knowledge base (6).

Initially, we selected databases containing several other databases, *i.e.* meta-databases. Subsequently, we included further databases not contained in the meta-databases.

Filtration approach. Generally, we tried to integrate as many databases as possible. This approach minimizes the number of variants without annotation and allowed us to compare and validate different databases. Nonetheless, we decided to define inclusion and exclusion criteria and filtered the available databases accordingly.

For the integration of a database, access (*e.g.* web query or bulk download) is needed as well as unambiguous location information on the variant level. Thus, databases without variant location data or databases with only gene associated information were excluded (in-/exclusion criterion: availability of variant location data).

Additionally, variant databases, which are used for clinical decision support, have to fulfill certain requirements as described in the above-mentioned CAP guidelines and Global Alliance for Genomics and Health's consensus recommendations (5). Key attributes are adequate curation, standardized nomenclature, unambiguous representation of data, and – most important – clinically relevant information on variants regarding therapeutic, prognostic or diagnostic implications (3). Therefore, databases limited to hereditary mutations without information on somatic mutations in cancer were excluded, while databases with information on drug sensitivity or disease association were included (in-/exclusion criterion: focus on clinical variants).

Other important properties were the level of detail regarding interpretation, availability of an evidence rating and adequate source information. Additionally, we considered up-to-dateness of a database as a key attribute. New information emerges daily and former interpretations may become outdated or disproved. Therefore, we only included database which are updated regularly, at least quarterly (in-/exclusion criterion: regular updates).

By these criteria we selected 12 databases and three meta-databases out of 39 databases. The full table of clinical variant databases included in and excluded from this study can be found in the Supplementary Material at: https://doi.org/10.5281/zenodo.4477193. The selected databases were: Jackson Laboratory Clinical Knowledgebase (JAX) (20), Oncology Knowledge Base (OncoKB) (21), Cancer Genome Interpreter's variants database (CGI) (22), Clinical Interpretations of Variants in Cancer (CIViC) (23), MolecularMatch (24), Precision Medicine Knowledgebase (PMKB) (25), NCBI's Clinical Variants (ClinVar) (26), Catalogue Of Somatic Mutations In Cancer (COSMIC) (27), Database of Curated Mutation (DOCM) (28), Human Gene Mutation Database (HGMD) (29), Phenotype for ENCODE (PhenCode) (30) and Pharmacogenomics Knowledgebase (PharmGKB) (31). When possible, these databases were downloaded. In addition, we accessed annotation through three meta-databases: VICC meta-database (6), myvariant.info annotation service (19) and Ensembl phenotype associated variants (32). These databases and their intersections are presented in Figure 1.

With the help of the VICC meta-database we were able to access six clinical variant databases (JAX, OncoKB, CGI, CIViC,

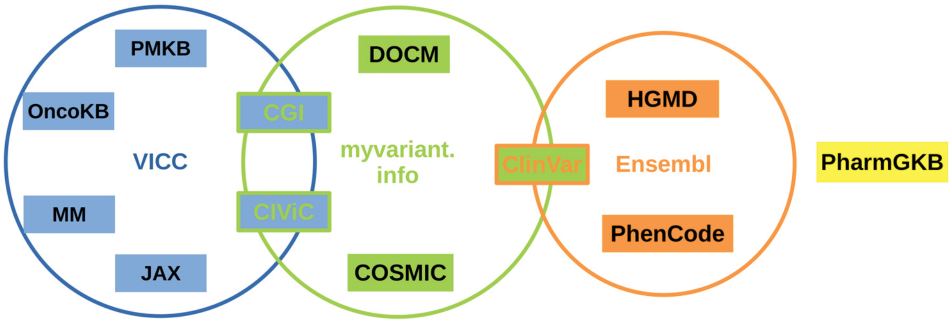

Figure 1. *Selected databases for clinical variant interpretation in cancer. Meta-databases are presented as circles, individual databases as rectangles. The intersection between the circles represents databases which are available within several meta-databases. CGI: Cancer Genome Interpreter's variants database; CIViC: Clinical Interpretations of Variants in Cancer; ClinVar: Clinical Variants of the National Center for Biotechnology Information; COSMIC: Catalogue of Somatic Mutations in Cancer; DOCM: Database of Curated Mutation; HGMD: Human Gene Mutation Database; JAX: Jackson Laboratory Clinical Knowledgebase; MM: MolecularMatch; OncoKB: Oncology Knowledge Base; PharmGKB: Pharmacogenomics Knowledgebase; PhenCode: Phenotype for ENCODE; PMKB: Precision Medicine Knowledgebase; VICC: meta-database of the Variant Interpretation for Cancer Consortium.*

Table II. *Downloadable databases for annotation of variants with clinical implications (tier I variants). For databases with available downloadable datasets we provide the respective accession link. Web links were last accessed on November 7, 2020.*

Database	Accession link
VICC	https://drive.google.com/drive/folders/1ZY6o3uaLOZSjOQPpXSMsnMbXmFWpb58d
Ensembl phenotype associated variants	ftp://ftp.ensembl.org/pub/grch37/current/variation/gvf/ homo_sapiens/homo_sapiens_phenotype_associated.gvf.gz
CGI	https://www.cancergenomeinterpreter.org/data/cgi_biomarkers_latest.zip
ClinVar	https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/
CIViC	https://civicdb.org/releases
PharmGKB	https://www.pharmgkb.org/downloads

CGI: Cancer Genome Interpreter's variants database: CIViC: Clinical Interpretations of Variants in Cancer: ClinVar: Clinical Variants database of the National Center for Biotechnology Information: PharmGKB: Pharmacogenomics Knowledgebase: VICC: meta-database of the Variant Interpretation for Cancer Consortium.

MolecularMatch and PMKB) providing information on clinical actionability in a uniform, standardized way. Via the myvariant.info annotation service, data were obtained from ClinVar, COSMIC, DOCM, CIViC and CGI. As both CIViC and CGI are also included in the VICC meta-database, we were able to assess differences in variant annotation dependent on the meta-database. ClinVar mainly provides information regarding phenotype and rarely on therapeutic implications. COSMIC and DOCM, however, merely contain information on disease associations.

The Ensembl database contains phenotype associated variants from three sources, namely from HGMD, PhenCode and NCBI's dbSNP/ClinVar variants. HGMD and PhenCode provide only information on database availability but no context. Furthermore, we included PharmGKB, which includes variants regarding drug sensitivity or adverse drug response.

Annotation of variants. In order to compare the different databases, we annotated and analyzed validated variant calling data from seven published datasets (33). These datasets consist in total of 678 samples from patients with myelodysplastic syndrome (MDS) or acute myeloid leukemia (AML). Sequencing data were generated on Illumina platforms (HiSeq, NextSeq and HiScanSQ) and Roche 454. From those seven validated datasets, all unique variants were extracted for further interpretation.

All mutations with information on the above-mentioned databases were annotated using R version 3.6.3 (see Supplementary Material at: https://doi.org/10.5281/zenodo.4477083) (34). The myvariant.info database was accessed for variant annotation by means of the R package 'myvariant' version 1.16.0 (35), the other databases were collected *via* downloadable datasets as described in Table II (as of September 7, 2020).

Classification of variants. To classify clinically relevant variants, we adopted VICC Harmonized Evidence Levels, which depend on the current classification system from the Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology and College of American Pathologists. We mapped it to the contents of available databases according to clinical relevance for MDS and AML (see Table III).

Table III. *Mapping of variants with clinical implications. We refined the classification from the Joint Consensus Recommendation and the Harmonized Evidence Levels, in order to differentiate between variants with therapeutic implications (tier I) versus mere disease association, diagnostic or prognostic implications (tier II). In tier III (variants of unclear significance), we distinguish variants without any annotation (tier III D) from variants, which can be found in the respective databases (tier III A-C).*

Tier	Evidence level	Explanation
I		Variants with therapeutic implications
	A	According to professional guidelines/Food and Drug Administration
	B	based on well-powered studies with expert consensus (phase III trials)
	C	Phase I-II trials/inclusion criteria for clinical trials
	D	Based on preclinical studies/therapeutic implications (IA-C) in different/not specified cancer type (other than AML, MDS, myeloproliferative disease, hematological malignancies)
II		Variants with disease association, diagnostic or prognostic implications
	A	Diagnostic or prognostic implications and/or mentioned in guidelines
	B	Diagnostic or prognostic implications and/or mentioned in well-powered studies
	C	Interpretation: pathogenic/likely pathogenic
	D	Disease association/diagnostic or prognostic implications in different/not specified cancer type (other than AML, MDS, myeloproliferative disease, hematological malignancies)
IIIA-C		VUS
	A	Variant with ClinVar Info: Uncertain significance/variant mentioned in database without interpretation
	B	Benign/ likely benign in different or not specified cancer type (other than AML, MDS, myeloproliferative disease, hematological malignancies)/variants with conflicting interpretations present (benign and pathogenic)
	C	Inferred association (VUS with COSMIC allelic frequency ≥0.01)
IIID		VUS without any annotation in knowledge bases
IV		Benign variants/polymorphisms
	A	Benign in specified cancer type
	B	Likely benign in specified cancer type
	C	Inferred benign (VUS with minor allelic frequency >1% in population databases)

AML: Acute myeloid leukemia; ClinVar: Clinical Variants database of the National Center for Biotechnology Information; COSMIC: Catalogue of Somatic Mutations in Cancer; MDS: myelodysplastic syndrome; VUS: variants of unclear significance.

The existing classification was enhanced to distinguish between variants with therapeutic (tier I) and diagnostic (tier II) implications in order to amplify the value of information of drug sensitivity over mere disease association. Furthermore, VICC Harmonized Evidence Levels are designed for tier I variants and do not mention how to handle variants from other tiers. The current version of VICC also does not clarify how to handle conflicting level of evidences from multiple databases. For the overall classification of each variant, the highest ranked contents (tier I > tier II > tier IV > tier III) among the databases was pivotal. Of note, tier IV (known benign variants) is ranked higher than tier III (variants of unclear significance) due to a higher underlying level of evidence.

Availability of data and materials. All datasets analyzed during this study are openly available online and URLs are provided in this published article. The datasets analyzed during the current study are available in the NCBI Sequence Read Archive (BioProjectID: PRJNA388411; https://www.ncbi.nlm.nih.gov/bioproject/PRJNA388411).

Results

In total, we analyzed 990 unique validated variants – containing pathogenic as well as benign variants – from patients with MDS/AML detected in seven public datasets. The full table of annotated and classified variants can be found in the Supplementary Material at: https://doi.org/10.5281/zenodo.4477289.

Variant classification. An overview of all variants and their classification according to tiers I to IV is presented in Figure 2.

Out of 990 unique variants, 64 were classified as tier I (6.5% of all variants), 81 as tier II (7.3%), 312 variants as tier III A-C (31.5%), and none as tier IV. For the majority of variants (533; 53.8%) we were unable to find any annotations in the selected databases. These variants were classified as tier III D.

Database coverage. Figure 3 presents the distribution of variants for the different tiers among the databases as well as the overlap between different data sources (for upset-plots visualizing the overlap between the different databases see Supplementary Material at: https://doi.org/10.5281/zenodo.4477397).

Most tier I variants were found in JAX (46 variants, 72.0% of all tier I variants), followed by CIViC (36 variants, 56.3%), CGI (24 variants, 37.5%) and OncoKB (21 variants, 32.8%

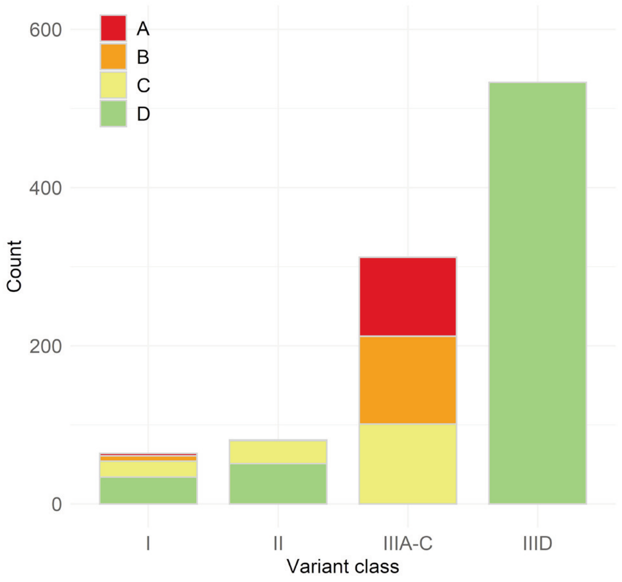

Figure 2. *Variants assigned according to our adjusted classification. The plot visualizes the distribution among the different tiers with help of a stack bar plot. Most variants were classified as tier III (85.3%), i.e. there were no annotations available in the selected databases (tier III D, 53.8%) or variants were of unclear significance (tier III A-C, 31.5%). Only few variants were assigned to tiers I (therapeutic implications, 6.5%) or II (other clinical implications, 7.3%). Highest evidence levels (tier IA+B) were only assigned to 0.8% of all variants.*

variants). Only few variants were represented in MolecularMatch (five variants, 7.8%), PharmGKB (five variants, 7.8%) and ClinVar (three variants, 4.7%). None of the variants were present in PMKB. The VICC meta-database featured 92.2% of all tier I variants, while myvariant.info only contained 65.6%. The other databases did not contain any information on variants with therapeutic implications (=tier I variants).

One third of all tier I variants (21 variants) were uniquely represented in one database. These variants were present in CGI (eight variants, 12.5% of all tier I variants), JAX (six variants, 9.4%), CIViC (five variants, 7.8%) and PharmGKB (two variants, 3.1%). No variants were uniquely present in OncoKB, ClinVar or MolecularMatch. Two-thirds of all tier I variants (43 out of 64) were included in more than one database, 55.0% in two or three databases. Only one variant (KRAS proto-oncogene p.G13D) was present in five databases; not a single variant was present in more than five.

In all of these cases, databases contained unique information on therapeutic agents. Regarding specifically mentioned drugs, more than half (58%) of the recommendations overlapped. Conflicting indications of sensitivity or resistance were not found. For instance, regarding the variant KIT proto-oncogene (*KIT*) p.D816V, JAX, CGI and CIViC databases concordantly reported sensitivity to dasatinib, while information on resistance to ponatinib was solely reported by JAX.

Among tier I variants available in multiple databases, only five out of 43 variants were discordantly classified. In all five cases, a different interpretation was found in ClinVar. We always opted against the ClinVar annotation. Three variants (ABL proto-oncogene 1 (*ABL1*) p.K247R, ATM serine/threonine kinase (*ATM*) p.D1853N and *KIT* p.M541L) were classified as benign in a different disease or cancer type (tier III B): ClinVar classified the variant *KIT* p.M541L as benign in a not-specified disease, while CGI noted its drug resistance to imatinib. Variant *ABL1* p.K247R was classified as benign in mastocytosis, partial albinism, not-specified disease and gastrointestinal stroma tumor by ClinVar, while JAX, OncoKB and CIViC report their drug sensitivity to imatinib. ClinVar classified *ATM* p.D1853N as benign in hereditary cancer-predisposing syndrome, ataxia-telangiectasia syndrome and not-specified disease but PharmGKB noted an adverse drug response to cyclophosphamide, doxorubicin and fluorouracil.

Two variants [tumor protein p53 (*TP53*) p.S127Y and fms related receptor tyrosine kinase 3 (*FLT3*) p.T227M] were labeled as variants with uncertain significance (tier III A) by ClinVar. However, in the case of *TP53* p.S127Y, JAX reported sensitivity to dasatinib. For *FLT3* p.T227M, CIViC and PharmGKB noted an adverse drug response to sunitinib.

The most extensive source for tier II and tier III variants was ClinVar (91.5% for tier II, 62.4% for tier III A-C). A majority of variants classified as tier III A-C were also found in COSMIC (50.2%).

Database-dependent annotation. In order to further analyze the performance of the annotation *via* meta-databases, we compared the concordance and discordance between the original source and the derived meta-system shown in Figure 4.

For variants present in CIViC (Figure 4A), the majority of all variants (78%, 31 out of 40) were characterized by concordant information. With the help of myvariant.info, all variants found in the original source were annotated, while VICC missed nine variants.

Regarding CGI (Figure 4B), concordant information was only observed for 37.5% of all variants (nine out of 24). These represent all variants that were found on the basis of the original source. The VICC database provided information from CGI for 15 additional variants not found in the downloadable dataset. In all these cases, matching variant annotation was available with a web query at https://www.cancergenomeinterpreter.org/analysis.

Concerning ClinVar (Figure 4C), myvariant.info overlapped in 94% of all cases with the original source but missed 19 variants found with the help of the downloadable dataset from ClinVar. These variants were, in 14 out of 19 cases, indels. Querying Ensembl, only 35% of all variants found in the original source were annotated.

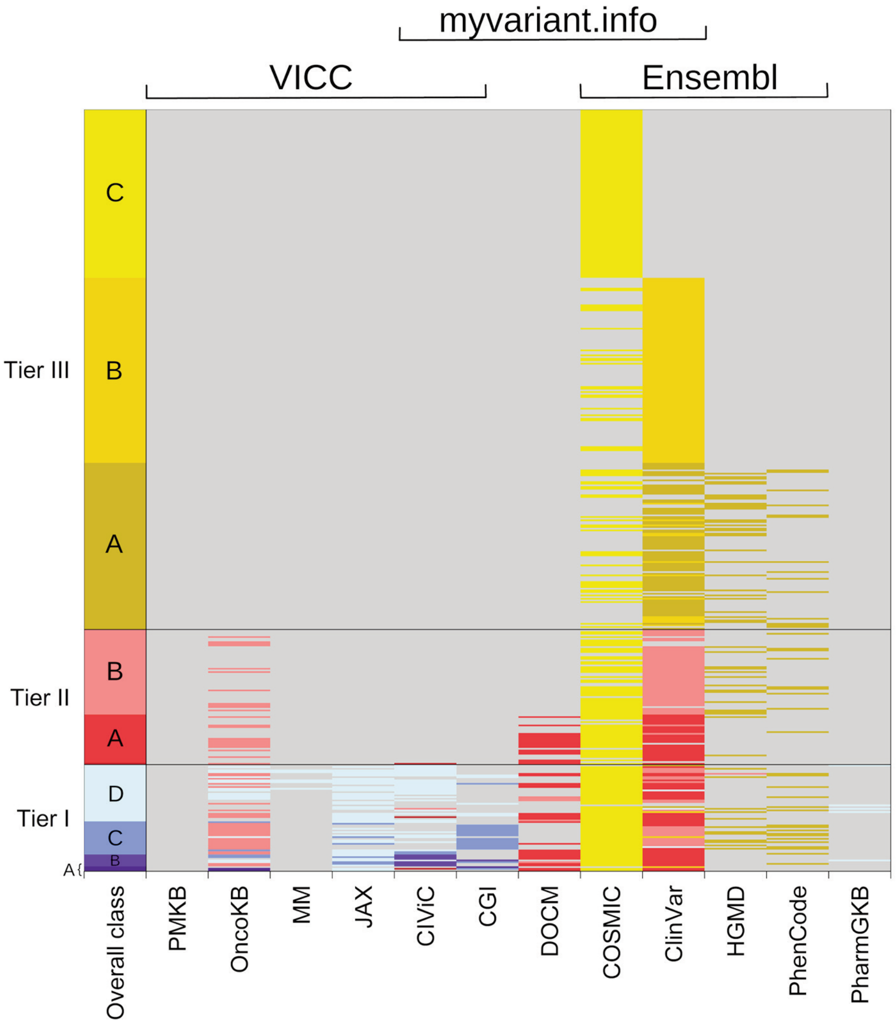

Figure 3. *Distribution of variants among databases. The plot shows the distribution of variants among the different databases (tier I: blue, tier II: red, tier III: yellow). Most variants with therapeutic implications were present in databases included in the meta-database of the Variant Interpretation for Cancer Consortium (VICC). The Jackson Laboratory Clinical Knowledgebase (JAX), Clinical Interpretations of Variants in Cancer database (CIViC), Cancer Genome Interpreter's variants database (CGI), Oncology Knowledge Base (OncoKB) and MolecularMatch (MM) contained 92.2% of all tier I variants. Tier II variants were mainly present in the Clinical Variants (ClinVar) of the National Center for Biotechnology Information (NCBI) (91.5% of all tier II variants). Regarding variants with unclear significance, ClinVar (containing 62.4% of all tier III A-C variants) and the Catalogue of Somatic Mutations in Cancer (COSMIC) (50.2%) were the major resources. DOCM: Database of Curated Mutation; HGMD: Human Gene Mutation Database; PharmGKB: Pharmacogenomics Knowledgebase); PhenCode: Phenotype for ENCODE; PMKB: Precision Medicine Knowledgebase.*

Discussion

Due to the increasing number of databases available for the interpretation of clinical variants, there is a growing need for instruments to facilitate further evaluation and decision-making. The distribution of variants with therapeutic implications among the databases (Figure 3) shows that a relevant proportion of variants was only present in a single database. Thus, all sources containing information on therapeutic implications – small as well as large databases – should be considered. On the one hand, there are reports about discordances between annotation services regarding pathogenicity and clinical actionability of variants (36, 37). In this study, on the other hand, only few discordant interpretations were observed. Even more, we emphasize that competing services and databases should not be seen as a hindrance but as possibility for sharing information. The more sources of information that are included, the more

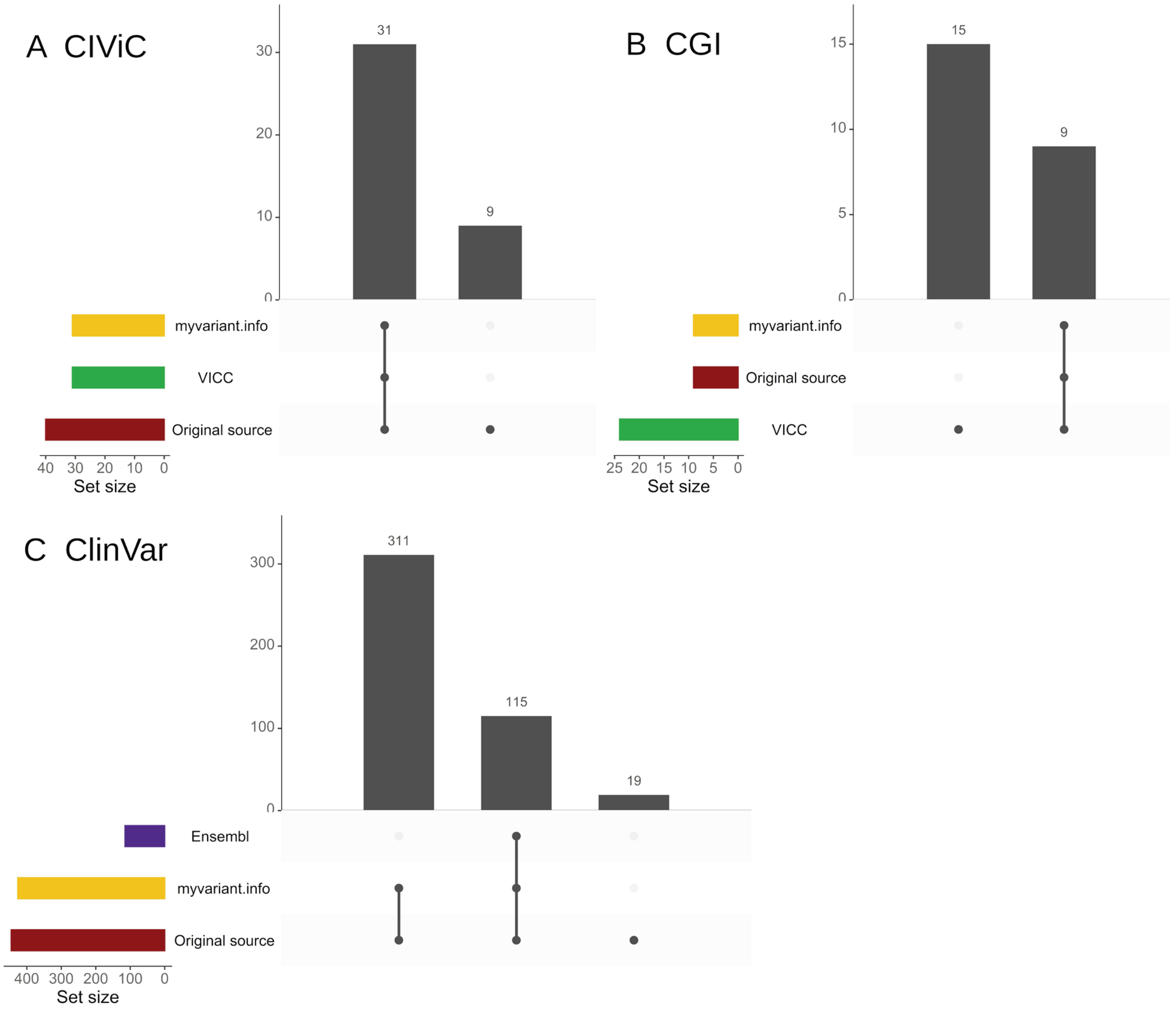

Figure 4. *Intersections between databases. The upset plots show the concordance of variants between the original source and meta-databases. A: Comparing Clinical Interpretations of Variants in Cancer (CIViC) to myvariant.info and the meta-database of the Variant Interpretation for Cancer Consortium (VICC): While myvariant.info included all variants from the original source, VICC missed nine. B: Comparing Cancer Genome Interpreter's variants database (CGI) to myvariant.info and VICC: All variants were present in both meta-databases. VICC even included 15 additional variants, which were only available through a web query on https://www.cancergenomeinterpreter.org/analysis. C: Comparing the Clinical Variants database of the National Center for Biotechnology Information (ClinVar) to myvariant.info and Ensembl: myvariant.info missed 19 variants (6%) from the original source, while Ensembl covered only 35% of all variants found in ClinVar.*

relevant variants can be found. However, from a practical perspective, a low number of high-quality databases is preferable. Thus, meta-databases are expected to be the most useful databases for interpretation of somatic variants.

Based on our case study, evaluating variants from MDS/AML, the VICC meta-database, featuring almost all databases of variants with therapeutic implications (tier I variants), was the most extensive source. Only PharmGKB and ClinVar were not included by VICC at the time of this study.

Regarding variants with disease association, the distribution of tier II and III variants shows that ClinVar and COSMIC – neither included in VICC but both featured in myvariant.info – are currently the most comprehensive resources. However, ClinVar contained in our case study all five variants with discordance in variant classification and only 4.7% of all tier I variants (three out of 64 variants). Therefore, this database seems to have limitations regarding reporting of therapeutic implications.

Meta-databases such as myvariant.info or the VICC meta-database facilitate the use of multiple data sources but they still come with certain limitations. The analysis of overlapping information (Figure 4) indicates incomplete data present in the meta-databases. This was most obvious in the case of the phenotype-associated variants derived from the Ensembl meta-database. The use of this dataset was associated with little added value. In the case of the myvariant.info meta-database, the CGI and CIViC databases were completely mapped, while variants from ClinVar were missing, including important hotspot variants such as the nucleophosmin 1 (*NPM1*) codon 288 frameshift mutation (38-40). Even VICC missed, with regards to CIViC, variants from the original source. However, in the case of CGI, VICC contained all variants from the downloadable dataset and even several additional variants that were otherwise only available *via* web query. Therefore, VICC seems to have contained a more recent version of CGI than the downloadable catalog, which was last updated January 2018. Moreover, VICC is the only available public resource to access bulk data from OncoKB, JAX, MolecularMatch and PMKB. However, it should be noted that our results represent only a snap-shot in time, all databases are subject to regular updates.

When annotating variants, two different ways have to be considered. If a unique identifier is available for a specific variant, *e.g.* a genomic position and a nucleotide change, we refer to this approach as the variant-centered approach. On the contrary, if information on a specified (partly large) genomic region is provided, it is considered a region-based approach. As precise information on the relevant variants is, in this case, not available, the user has to determine the pathogenic variants by himself. Due to applicability, we only worked with the variant-centered approach in our case study. Due to the fact that PMKB predominantly utilizes the region-based approach, no annotations were available from this database. In order to make full use of all available databases, it might also be beneficial to additionally consider a region-based approach for the interpretation of variants.

Furthermore, each database has a unique design as well as unique inclusion criteria for variants. Therefore, differences in coverage of the different types of mutations are expected. Knowledge bases like JAX, CGI, MolecularMatch, PharmGKB, PMKB, CIViC and OncoKB only contain variants with clinical implications (tier I and II variants), while ClinVar is designed in a more universal way, featuring somatic mutations, germline mutations and polymorphisms. COSMIC, on the contrary, only contains somatic variants which are found in cancer, without giving information on clinical implications.

It should be mentioned that our analysis did not classify any variant as benign (tier IV). However, this observation was due to the databases used. We did not apply an inferred classification as polymorphism by allele frequency (tier IV C). When evaluating variants as polymorphism by minor allele frequency, the use of population databases can lead to incorrect classification, especially in the case of interpreting somatic variants in hematological malignancies (5). Furthermore, due to their intrinsic database design, the focus of most knowledge bases is on therapeutic and other clinical implications. Therefore, benign variants are not represented. Only ClinVar contains information about benign variants but mostly for germline variants in hereditary diseases. Variants annotated as benign regarding a different entity (not MDS or AML) are classified as variants of unclear significance (tier III B) and not as tier IV variants as they may possibly have therapeutic implications for patients with MDS or AML. In our case study, we observed three examples: ABL1 p.K247R, KIT p.M541L and ATM p.D1853N.

Conclusion

Our case study revealed that a single public variant database is currently not sufficient to interpret clinical variants. Recent activities in the harmonization of clinical variants by means of the VICC or the myvariant.info meta-database show a promising development in the direction of a 'one-stop tool'. However, at present the use of knowledge bases can just be considered as a supportive instrument. Manual interpretation of variants by experts and tumor boards remains obligatory. Caution should be exercised when using meta-databases and, if possible, the original source should also be considered.

Conflicts of Interest

The Authors declare that they have no competing interests.

Authors' Contributions

H.B. performed the analysis and wrote the article. M.D. and S.S. reviewed the analysis and provided feedback. C.M.T. supervised the study design. All Authors read and approved the final version of the article.

Acknowledgements

The Authors thank Christian Wünsch for his support, especially for analyzing the JSON formatted VICC meta-database. We are also grateful to Albina Caetano da Silva and Nicolette Banck for moral support and proof-reading of the article.

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 article.

References

1 Gao P, Zhang R and Li J: Comprehensive elaboration of database resources utilized in next-generation sequencing-based tumor somatic mutation detection. Biochim Biophys Acta BBA – Rev Cancer *1872(1)*: 122-137, 2019. PMID: 31265877. DOI: 10.1016/j.bbcan.2019.06.004

2 Prawira A, Pugh TJ, Stockley TL and Siu LL: Data resources for the identification and interpretation of actionable mutations by clinicians. Ann Oncol *28(5)*: 946-957, 2017. PMID: 28327901. DOI: 10.1093/annonc/mdx023

3 Van Allen EM, Wagle N and Levy MA: Clinical analysis and interpretation of cancer genome data. J Clin Oncol *31(15)*: 1825-1833, 2013. PMID: 23589549. DOI: 10.1200/JCO.2013.48.7215

4 Yen JL, Garcia S, Montana A, Harris J, Chervitz S, Morra M, West J, Chen R and Church DM: A variant by any name: quantifying annotation discordance across tools and clinical databases. Genome Med *9*: 7, 2017. DOI: 10.1186/s13073-016-0396-7

5 Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, Tsimberidou AM, Vnencak-Jones CL, Wolff DJ, Younes A and Nikiforova MN: Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn *19*: 4-23, 2017. PMID: 27993330. DOI: 10.1016/j.jmoldx.2016.10.002

6. Wagner AH, Walsh B, Mayfield G, Tamborero D, Sonkin D, Krysiak K, Deu-Pons J, Duren RP, Gao J, McMurry J, Patterson S, Fitz C del V, Pitel BA, Sezerman OU, Ellrott K, Warner JL, Rieke DT, Aittokallio T, Cerami E, Ritter DI, Schriml LM, Freimuth RR, Haendel M, Raca G, Madhavan S, Baudis M, Beckmann JS, Dienstmann R, Chakravarty D, Li XS, Mockus S, Elemento O, Schultz N, Lopez-Bigas N, Lawler M, Goecks J, Griffith M, Griffith OL and Margolin AA: VICC: A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. Nat Genet *52*: 448-457, 2020. PMID: 32246132. DOI: 10.1038/s41588-020-0603-8

7 PubMed. Available at: https://pubmed.ncbi.nlm.nih.gov/ [Last accessed on November 1 2020]

8 Perrin H, Denorme M, Grosjean J, Community Omic, Dynomant E, Henry VJ, Pichon F, Darmoni S, Desfeux A and Gonzalez BJ: OMICtools: a community-driven search engine for biological data analysis. arXiv: 170703659 [cs, q-bio], 2017.

9 Griffith M: biostars forum: Database Of Tumor Suppressors And/Or Oncogenes. Available at: https://www.biostars.org/p/15890/ [Last accessed on August 1, 2020]

10 RJ Carlson M: Bioconductor Package annotation: Genomic Annotation Resources. R package version 1.10.0. Bioconductor. Available at: http://bioconductor.org/packages/annotation/ [Last accessed on April 1, 2020]

11 Bioconductor: variants: Annotating Genomic Variants. R package version 1.10.0. Bioconductor. Available at: http://bioconductor.org/packages/variants/ [Last accessed on April 1, 2020]

12 NCBI Variation Guide. Available at: https://www.ncbi.nlm.nih.gov/guide/variation/ [Last accessed on April 1, 2020]

13 My Cancer Genome Data Sources – My Cancer Genome. Available at: https://www.mycancergenome.org/content/page/my-cancer-genome-data-sources/ [Last accessed on April 1, 2020]

14 VICC|Research. VICC. Available at: https://cancervariants.org/research/ [Last accessed on January 27, 2021]

15 Wang K, Li M and Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res *38*: e164, 2010. PMID: 20601685. DOI: 10.1093/nar/gkq603

16 McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P and Cunningham F: The Ensembl Variant Effect Predictor. Genome Biol *17*: 122, 2016. PMID: 27268795. DOI: 10.1186/s13059-016-0974-4

17 Lawrence M and Gentleman R: VariantTools: an extensible framework for developing and testing variant callers. Bioinformatics *33*: 3311-3313, 2017. PMID: 29028267. DOI: 10.1093/bioinformatics/btx450

18 Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X and Ruden DM: SnpEff: 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 (Austin) *6*: 80-92, 2012. PMID: 22728672. DOI: 10.4161/fly.19695

19 Xin J, Mark A, Afrasiabi C, Tsueng G, Juchler M, Gopal N, Stupp GS, Putman TE, Ainscough BJ, Griffith OL, Torkamani A, Whetzel PL, Mungall CJ, Mooney SD, Su AI and Wu C: High-performance web services for querying gene and variant annotation. Genome Biol *17*: 91, 2016. PMID: 27154141. DOI: 10.1186/s13059-016-0953-9

20 Patterson SE, Liu R, Statz CM, Durkin D, Lakshminarayana A and Mockus SM: The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. Hum Genomics *10*: 4, 2016. PMID: 26772741. DOI: 10.1186/s40246-016-0061-7

21 Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, Chang MT, Chandarlapaty S, Traina TA, Paik PK, Ho AL, Hantash FM, Grupe A, Baxi SS, Callahan MK, Snyder A, Chi P, Danila DC, Gounder M, Harding JJ, Hellmann MD, Iyer G, Janjigian YY, Kaley T, Levine DA, Lowery M, Omuro A, Postow MA, Rathkopf D, Shoushtari AN, Shukla N, Voss MH, Paraiso E, Zehir A, Berger MF, Taylor BS, Saltz LB, Riely GJ, Ladanyi M, Hyman DM, Baselga J, Sabbatini P, Solit DB and Schultz N: OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol *2017*: PO.17.00011, 2017. PMID: 28890946. DOI: 10.1200/PO.17.00011

22 Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, Tusquets I, Albanell J, Rodon J, Tabernero J, Torres C de, Dienstmann R, Gonzalez-Perez A and Lopez-Bigas N: Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genom Med *10*: 25, 2018. DOI: 10.1186/s13073-018-0531-8

23 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 and Griffith OL: CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet *49*: 170-174, 2017. PMID: 28138153. DOI: 10.1038/ng.3774

24 molecularmatch website. Available at: https://www.molecularmatch.com/ [Last accessed on June 28, 2020]

25 Huang L, Fernandes H, Zia H, Tavassoli P, Rennert H, Pisapia D, Imielinski M, Sboner A, Rubin MA, Kluk M and Elemento O: The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. J Am Med Inform Assoc *24*: 513-519, 2017. PMID: 27789569. DOI: 10.1093/jamia/ocw148

26. 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 and Maglott DR: ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res *46*: D1062-D1067, 2018. PMID: 29165669. DOI: 10.1093/nar/gkx1153

27 Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ and Forbes SA: COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res *47*: D941-D947, 2019. PMID: 30371878. DOI: 10.1093/nar/gky1015

28 Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Choudhary MN, McMichael JF, Fulton RS, Wilson RK, Griffith OL and Mardis ER: DoCM: a database of curated mutations in cancer. Nat Methods *13*: 806, 2016. PMID: 27684579. DOI: 10.1038/nmeth.4000

29 Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD and Cooper DN: The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet *133*: 1-9, 2014. PMID: 24077912. DOI: 10.1007/s00439-013-1358-4.

30 Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H, Kern A, Kuhn R, Patrinos GP, Hughes J, Higgs D, Chui D, Scriver C, Phommarinh M, Patnaik SK, Blumenfeld O, Gottlieb B, Vihinen M, Väliaho J, Kent J, Miller W and Hardison RC: PhenCode: connecting ENCODE data with mutations and phenotype. Hum Mutat *28*: 554-562, 2007. PMID: 17326095. DOI: 10.1002/humu.20484

31 Thorn CF, Klein TE and Altman RB: PharmGKB: The Pharmacogenomics Knowledge Base. Methods Mol Biol *1015*: 311-320, 2013. PMID: 23824865. DOI: 10.1007/978-1-62703-435-7_20

32 Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Marugán JC, Cummins C, Davidson C, Dodiya K, Fatima R, Gall A, Giron CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Lavidas I, Le T, Lemos D, Martinez JG, Maurel T, McDowall M, McMahon A, Mohanan S, Moore B, Nuhn M, Oheh DN, Parker A, Parton A, Patricio M, Sakthivel MP, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sycheva M, Szuba M, Taylor K, Thormann A, Threadgold G, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Flint B, Frankish A, Hunt SE, IIsley G, Kostadima M, Langridge N, Loveland JE, Martin FJ, Morales J, Mudge JM, Muffato M, Perry E, Ruffier M, Trevanion SJ, Cunningham F, Howe KL, Zerbino DR and Flicek P: Ensembl 2020. Nucleic Acids Res *48*: D682-D688, 2020. PMID: 31691826. DOI: 10.1093/nar/gkz966

33 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 and Dugas M: appreci8: a pipeline for precise variant calling integrating 8 tools. Bioinformatics *34*: 4205-4212, 2018. PMID: 29945233. DOI: 10.1093/bioinformatics/bty518

34 R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/ [Last accessed on June 27, 2020]

35 Mark A (2020). myvariant: Accesses MyVariant.info variant query and annotation services. R package version 1.20.0. Available at: https://bioconductor.org/packages/myvariant/ [Last accessed on Apr 8, 2020]

36 Katsoulakis E, Duffy JE, Hintze B, Spector NL and Kelley MJ: Comparison of annotation services for next-generation sequencing in a large-scale precision oncology program. JCO Precis Oncol *4*: PO.19.00118, 2020. PMID: 32923873. DOI: 10.1200/PO.19.00118

37 Gradishar W, Johnson K, Brown K, Mundt E and Manley S: Clinical variant classification: A comparison of public databases and a commercial testing laboratory. Oncologist *22*: 797-803, 2017. PMID: 28408614. DOI: 10.1634/theoncologist.2016-0431

38 Suzuki T, Kiyoi H, Ozeki K, Tomita A, Yamaji S, Suzuki R, Kodera Y, Miyawaki S, Asou N, Kuriyama K, Yagasaki F, Shimazaki C, Akiyama H, Nishimura M, Motoji T, Shinagawa K, Takeshita A, Ueda R, Kinoshita T, Emi N and Naoe T: Clinical characteristics and prognostic implications of NPM1 mutations in acute myeloid leukemia. Blood *106*: 2854-2861, 2005. PMID: 15994285. DOI: 10.1182/blood-2005-04-1733

39 Kihara R, Nagata Y, Kiyoi H, Kato T, Yamamoto E, Suzuki K, Chen F, Asou N, Ohtake S, Miyawaki S, Miyazaki Y, Sakura T, Ozawa Y, Usui N, Kanamori H, Kiguchi T, Imai K, Uike N, Kimura F, Kitamura K, Nakaseko C, Onizuka M, Takeshita A, Ishida F, Suzushima H, Kato Y, Miwa H, Shiraishi Y, Chiba K, Tanaka H, Miyano S, Ogawa S and Naoe T: Comprehensive analysis of genetic alterations and their prognostic impacts in adult acute myeloid leukemia patients. Leukemia *28*: 1586-1595, 2014. PMID: 24487413. DOI: 10.1038/leu.2014.55

40 Patel JP, Gönen M, Figueroa ME, Fernandez H, Sun Z, Racevskis J, Van Vlierberghe P, Dolgalev I, Thomas S, Aminova O, Huberman K, Cheng J, Viale A, Socci ND, Heguy A, Cherry A, Vance G, Higgins RR, Ketterling RP, Gallagher RE, Litzow M, van den Brink MRM, Lazarus HM, Rowe JM, Luger S, Ferrando A, Paietta E, Tallman MS, Melnick A, Abdel-Wahab O and Levine RL: Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. N Engl J Med *366*: 1079-1089, 2012. PMID: 22417203. DOI: 10.1056/NEJMoa1112304

Received December 29, 2020
Revised January 23, 2021
Accepted January 29, 2021