Epigenetics & Chromatin

METHODOLOGY Open Access

Var|Decrypt: a novel and user-friendly tool to explore and prioritize variants in whole-exome sequencing data

Mohammad Salma[1,2*], Elina Alaterre[4], Jérôme Moreaux[3,4,5] and Eric Soler[1,2*]

Abstract

Background High-throughput sequencing (HTS) offers unprecedented opportunities for the discovery of causative gene variants in multiple human disorders including cancers, and has revolutionized clinical diagnostics. However, despite more than a decade of use of HTS-based assays, extracting relevant functional information from whole-exome sequencing (WES) data remains challenging, especially for non-specialists lacking in-depth bioinformatic skills.

Results To address this limitation, we developed Var|Decrypt, a web-based tool designed to greatly facilitate WES data browsing and analysis. Var|Decrypt offers a wide range of gene and variant filtering possibilities, clustering and enrichment tools, providing an efficient way to derive patient-specific functional information and to prioritize gene variants for functional analyses. We applied Var|Decrypt on WES datasets of 10 acute erythroid leukemia patients, a rare and aggressive form of leukemia, and recovered known disease oncogenes in addition to novel putative drivers. We additionally validated the performance of Var|Decrypt using an independent dataset of ~ 90 multiple myeloma WES, recapitulating the identified deregulated genes and pathways, showing the general applicability and versatility of Var|Decrypt for WES analysis.

Conclusion Despite years of use of WES in human health for diagnosis and discovery of disease drivers, WES data analysis still remains a complex task requiring advanced bioinformatic skills. In that context, there is a need for user-friendly all-in-one dedicated tools for data analysis, to allow biologists and clinicians to extract relevant biological information from patient datasets. Here, we provide Var|Decrypt (trial version accessible here: https://vardecrypt.com/app/vardecrypt), a simple and intuitive Rshiny application created to fill this gap. Source code and detailed user tutorial are available at https://gitlab.com/mohammadsalma/vardecrypt.

Keywords Whole-exome sequencing, Acute leukemia, Disease driver gene, Bioinformatics

Background

Leukemia comprises a heterogeneous group of deadly blood cancers resulting from abnormal or impaired hematopoietic cell differentiation and stem cell function. Many cell intrinsic factors can contribute to leukemia initiation, development and maintenance, including mutations affecting signaling pathways, metabolic genes, splicing components and epigenetic regulators, leading to acquisition of several cancer hallmarks [1]. Although a number of recurrently mutated genes have already been identified in leukemia of both myeloid and lymphoid

*Correspondence:
Mohammad Salma
mohammad.salma@igmm.cnrs.fr
Eric Soler
eric.soler@igmm.cnrs.fr
[1] Institut de Génétique Moléculaire de Montpellier, Univ Montpellier, CNRS, Montpellier, France
[2] Laboratory of Excellence GR-Ex, Université de Paris, Paris, France
[3] Department of Biological Hematology, CHU Montpellier, Montpellier, France
[4] Institute of Human Genetics, UMR 9002 CNRS-UM, Montpellier, France
[5] Institut Universitaire de France (IUF), Paris, France

origins, a number of rare and/or aggressive leukemia subtypes, for which the driving oncogenes are poorly characterized, still require in-depth analyses. This also applies to solid tumors and rare cancers types, for which the mutational landscape remains to be thoroughly characterized. In this context, high-throughput sequencing (HTS) of patient samples is instrumental to unravel the underlying genetic abnormalities.

The era of medicine of precision and customization—i.e., the capacity to provide patient care guided by genetic diagnostic, is at reach since genome-scale sequencing approaches such as whole-exome sequencing (WES) are implemented in routine diagnostics [1, 2]. WES offers a flexible and efficient way to highlight the mutational landscape of hundreds of patients at relatively low cost. Indeed, WES is mostly focused on the gene coding regions of the genome, representing ~ 2% of the total human genome sequence (approximately ~ 30 million base-pairs) [3]. Although WES is unable to highlight complex genomic rearrangements such as chromosomal translocations or inversions, it still provides highly relevant information regarding gene mutations, including splice site mutations. It is, therefore, extensively used in multiple clinical centers and studies over the world (Fig. 1). As a growing number of patient genomes became sequenced, increasing amounts of detected variants were published and added into public databases. The number of known disease genes has, therefore, dramatically increased over the past decade, which reinforces diagnostic test performance [4]. Clinics in the USA [5], France [6] and the Netherlands [7] for instance report WES as a promising tool for systematic use in patients. However, despite its extensive use and significant advantages, extracting relevant information from WES data still represents a challenge (e.g., identifying recurrently mutated genes from cohorts of patients, or the biological pathways which are significantly affected). WES data analysis requires advanced bioinformatics tools and skills, which prevents non-specialists such as wet lab scientists or clinicians from being able to navigate within the datasets and perform custom analyses. A great challenge for scientists and clinicians with limited or no bioinformatics skills is, therefore, to be able to manipulate WES data in order to extract biological meaningful information, and relevant genes or pathways to guide functional testing and therapeutic approaches.

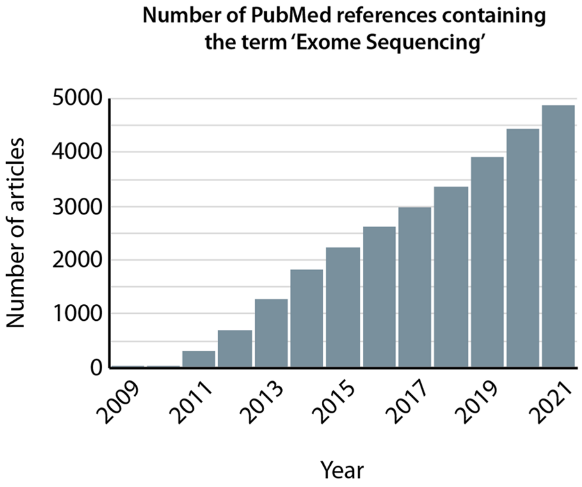

Fig. 1 Widespread use of whole-exome sequencing data. Graph plotting the number of PubMed articles containing the term 'exome sequencing', showing a continuous increase over the 2009–2021 period

WES data analysis consists in two main features: first, the detection of variants or mutations, and second, the post-variant calling analysis for prioritization of genes or variants within patient cohorts. The detection of germinal or somatic variants from matched tumor-control samples is still a complex task [8], despite the availability of dedicated tools. Bertier et al. [3] reported a total of 23 different challenges related mainly to the production, analysis and sharing of WES data [3]. They mentioned that the interpretation of variants of unknown significance was one of the most reported challenges across all articles. Often, the interpretation of these data requires specially trained staff [3]. In order to understand the biological meaning of these variants and/or differences between tumoral and control samples, researchers tend to focus on variant annotation and filtering to eliminate those with low relevance (variant-centered approach). Several tools have been developed in this aspect such as SNPsift [9], GEMINI [10] and VCF tools [11], Ingenuity® Variant Analysis™ Software [12], Golden Helix SNP and Variation Suite [13], BiERapp [14], EVA [15], Exomiser [16], Variant Ranker [17], BrowseVCF [18], TGex [19], and VCF-Miner [20]. Another strategy for prioritizing variants in WES data is attempting to connect the discovered variants with known diseases, biological processes, pathways, etc. using a variety of open-source databases (gene-centered approach). Some available tools such as GeMSTONE [21] and BiERapp [14] perform some enrichment analyses, but either are available online only, or do not support variant and enrichment results visualization. Other web-based tools such as Enrichr [22], GOrilla [23], wKGGSeq [24], geneontology [25] and g:profiler [26] which are often used to analyze transcriptome datasets, can perform some enrichment analyses using gene lists. However, these tools require users who wish to analyze WES data to extract and prepare the input data according to the tool recommendations. In that respect, a flexible and user-friendly all-in-one solution, with built-in functionalities designed to facilitate extraction of relevant meaningful biological data from WES is clearly lacking. Here, we introduce Var|Decrypt, a user-friendly and easy-to-use Rshiny application designed to close this gap.

Results

Exome-seq analysis pipeline

To provide an all-in-one solution, we first implemented an Exome-seq variant analysis pipeline (available as supplementary material, see Additional file 1 and Additional file 8: Fig. S2 for details). This pipeline can be used with raw sequencing data (e.g., FASTQ files) to generate variant calling files (vcf) and input files for downstream processing by Var|Decrypt. For users wishing to use their own vcf files, or vcf files from publicly available repositories as input, we built-in a pre-processing tool called Pre-Var|Decrypt (see https://gitlab.com/mohammadsalma/vardecrypt) allowing to process vcf files to generate Var|Decrypt input files (see below). This step only needs to be performed once for each batch of samples (i.e., patient cohorts), the resulting files can then be stored or directly used in Var|Decrypt for downstream analyses.

WES data processing using Var|Decrypt

In order to facilitate WES data analysis and functional interpretation, we developed Var|Decrypt, an easy-to-use and user-friendly RShiny tool, which can be deployed via Docker on several operating systems (Linux, macOS), downloaded and installed from open-source to run via Rstudio. In addition, we provide a link to an online trial version of Var|Decrypt (https://vardecrypt.com/app/vardecrypt) with access to a test dataset, allowing users to quickly be able to assess the tool and evaluate its capacities. Var|Decrypt includes several R packages to perform different post-VCF downstream analyses, which usually require users with scripting skills to perform tasks such as installing packages, preparing the input data and calling the appropriate function. A detailed tutorial on how to use Var|Decrypt in a simple and intuitive way is available at (https://gitlab.com/mohammadsalma/vardecrypt/-/wikis/Var%7CDecrypt) and as a video tutorial on the front page of Var|Decrypt online version. Var|Decrypt imports the output results from the Exome-seq pipeline or vcf files processed through Pre-Var|Decrypt (Additional file 1) and provides many built-in enrichment analyses options, helping researchers to develop or confirm hypotheses, to easily explore the differences between normal and tumor samples, and to prioritize variants, genes and pathways for functional analyses. Var|Decrypt is a fast-operating tool which provides multiple outputs within short time frames (i.e., seconds to minutes for loading and processing a full dataset, Additional file 6: Table S5). The output results and variables are saved in an Rdata file which lets users to explore Var|Decrypt results subsequently, instead of re-running the analysis. Var|Decrypt allows to explore, filter, sort genes containing variants, or to search for a specific gene through dynamic interfaces (see below).

Overall presentation of Var|Decrypt

The Var|Decrypt interface is composed of several tabs allowing users to get a general overview of the Exome-seq data and to browse the mutated gene lists or focus on single genes, single variant types (e.g., stop gain and frameshift deletions).

The 'Somatic variants explorer' tab provides a gene list and summary table containing all detected mutated genes (we define the somatic variants as being the ones specifically acquired in the tumor sample as compared to the control cells; variants or mutations present in the control cells are considered as germline variants as they are not somatically acquired) (Additional file 8: Fig. S2). For each gene in the table, the total number of variants detected is indicated, together with the different types of variants identified, and the percentage of patients bearing a mutation in a particular gene. The right part of the table shows for each gene which patient sample contains the indicated variants (Additional file 2: Table S1). Instead of focusing on the variants themselves, this dynamic table is gene-centered, and it also provides information on the number of variants detected in the cohort for each gene, the types of variants (e.g., stop gain, frameshift deletions, etc.) and the percentage of patients bearing mutations on a particular gene. When using the 'mutation rate' column, users can sort the entire mutated gene list by mutation frequency (i.e., number of patients showing a mutation or variant within a given gene), which provides an overview of the top mutated genes. All types of variants are shown by default, but users may choose to highlight only a subcategory of variants such as stop gain, frameshift variants (deletions, insertions), etc. Whereas the germline variants from a patient are usually used to filter-out nonspecific variants in cancer samples, Var|Decrypt also allows working on the germline variants ('Germline variant explorer') which is useful for the study of Mendelian genetic disorders or family case studies (not shown here).

The 'General statistics' tab provides information on the frequency of variant types within the cohort using a color code for the different types (e.g., frameshifts, non-sense, missense, etc.), the class of SNV (e.g., C > T, T > G, etc.) which may be useful to check if a particular bias is present in the samples or in the disease under study (Fig. 2). This tab also provides information on the total number of variants per sample, a feature that helps to quickly spot any outlier within the datasets. As exemplified in Fig. 2, sample m_13_D from our cohort contains ~ 30-fold more

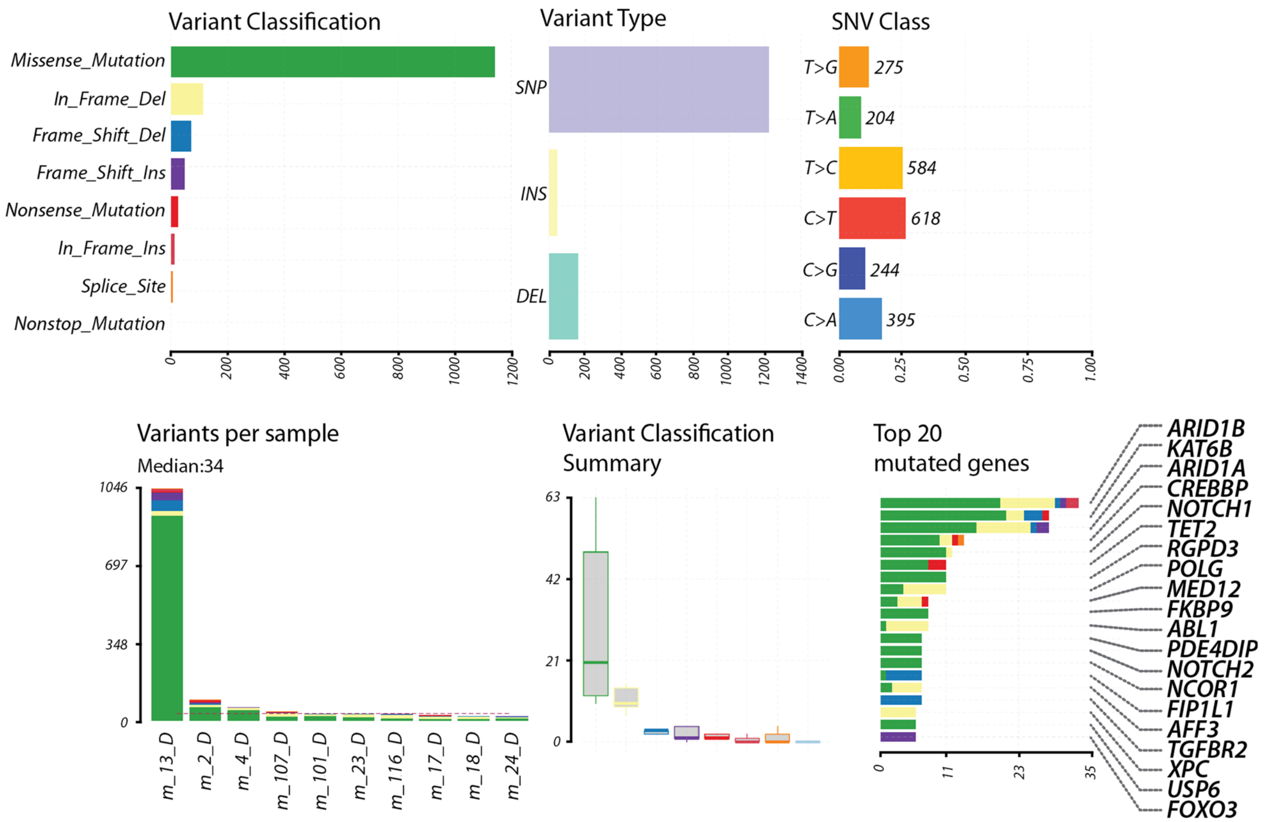

Fig. 2 General overview of the WES datasets. The general features of ten erythroleukemic samples are displayed, showing the variant classification (color-coded as a function of the type of mutation, top left), variant type (single nucleotide polymorphisms (SNP), insertions (INS) and deletions (DEL), top middle), and single nucleotide variant (SNV) class (top right). The bottom panel displays the number of variants per sample (each column represents a unique patient), using the same color code as in the variant classification panel. Note that patient m_13_D is spotted as being an outlier with ~ 30-fold more variants than in the other patients. The dashed red line represents the median number of variants in the cohort. The middle panel shows the variant classification summary in the cohort, using the same mutation-specific color code. Finally, the bottom right panel shows the top 20 mutated genes in the patient cohort (the number of variants/mutations is shown on the horizontal axis), with the percentage of patients bearing a mutation in a given gene indicated

variants than the average of the other samples, likely arising from technical issues during the sequencing or sample handling procedure. Such problematic samples can, therefore, be quickly spotted and excluded from further analyses. Finally, the top 20 mutated genes are shown with the same color code as for the variant types, to get an overview of the recurrently mutated genes (Fig. 2).

Identifying the recurrently mutated gene fraction within a patient cohort

Var|Decrypt offers the opportunity to quickly and easily browse WES data in order to identify recurrently mutated genes. By navigating in the somatic menus, users can in one click access the gene mutations frequencies (i.e., gene mutation percentage within the cohort), an important feature allowing to point at key genes likely involved in the disease phenotype. One key step in the discovery of cancer drivers is to be able to pinpoint the recurrently mutated genes within patient cohorts, as recurrently mutated genes likely represent true oncogenic drivers or genes important to sustain the cells' transformed state. However, despite all the filtering steps applied in various Exome-Seq analysis pipelines, a very large number of variants usually remains, especially in cancer samples. This represents one of the main challenges to prioritize gene mutations when dealing with Exome-seq datasets.

Filtering of putative false-positive gene mutations

A common issue of Exome-Seq data from short reads-associated sequencing platforms (such as Illumina sequencing) is the large fraction of variants called at genes harboring repetitive sequences, such as variable number of tandem repeats (VNTRs). The MUC gene family [27] is a good example of such problematic alignment and variant calling situation, as they contain long polymorphic stretches of ~ 60 bp repeats VNTRs, which is problematic with the current aligners and variant callers. Although some true causative variants may indeed be present within the VNTRs of the MUC gene family [28], we created an empirical filtering option allowing users to define a threshold for the maximum number of variant allowed per gene, in order to 'clean-up' the mutated gene list and get rid of the error-prone VNTR-containing gene sequences in the patient cohort. As a result, by setting a threshold of a maximum of 4 variants per gene in a maximum of 20% of the patients, we could get rid of the apparently highly variable and likely false-positive mutated genes in the final list (Fig. 3).

Another commonly used strategy to enrich for putative causative variants is to filter the mutated gene lists against cancer gene databases such as COSMIC, OncoKB or NCG [29–31]. We also implemented a filtering option allowing to focus on the mutated genes that are tagged as cancer-associated from such databases. The resulting outputs, therefore, are highly enriched in putative oncogenic drivers, allowing to explore the mutational landscape of human cancers. As confirmation, applying such filtering strategy on our AEL WES data produced a mutated gene list enriched for previously reported AEL-associated gene mutations [30, 31] such as the epigenetic modifiers TET2, NCOR1, NCOR2, BCOR, BCORL1, the CBP(*CREBBP*)/p300(*EP300*) co-activators, the polycomb repressive complex proteins EZH2, ASXL1, ASXL2, and the cohesin complex component RAD21 (Additional file 2: Table S1).

Integration of enrichment tools

An important aspect of Var|Decrypt is the access to various types of enrichment analyses thanks to the implementation of dynamic customizable graphical outputs. Var|Decrypt contains different disease ontology, gene ontology (e.g., biological process, molecular function, and cellular component), and Reactome/Kegg pathway enrichment tab offering the opportunity to identify particular pathway of functional alterations in the samples. The 'enrichment' tab offers users to quickly identify enrichments of disease ontology terms, biological pathways (Reactome, KEGG and WIKIpathways), or Gene-Ontology (GO)-terms such as 'Biological Process', Molecular Function', or 'Cellular Component' linked to the mutated gene lists. In addition, searches for gene–disease associations or gene–cancer associations are also available to highlight many known associations with established human disorders and cancers. For

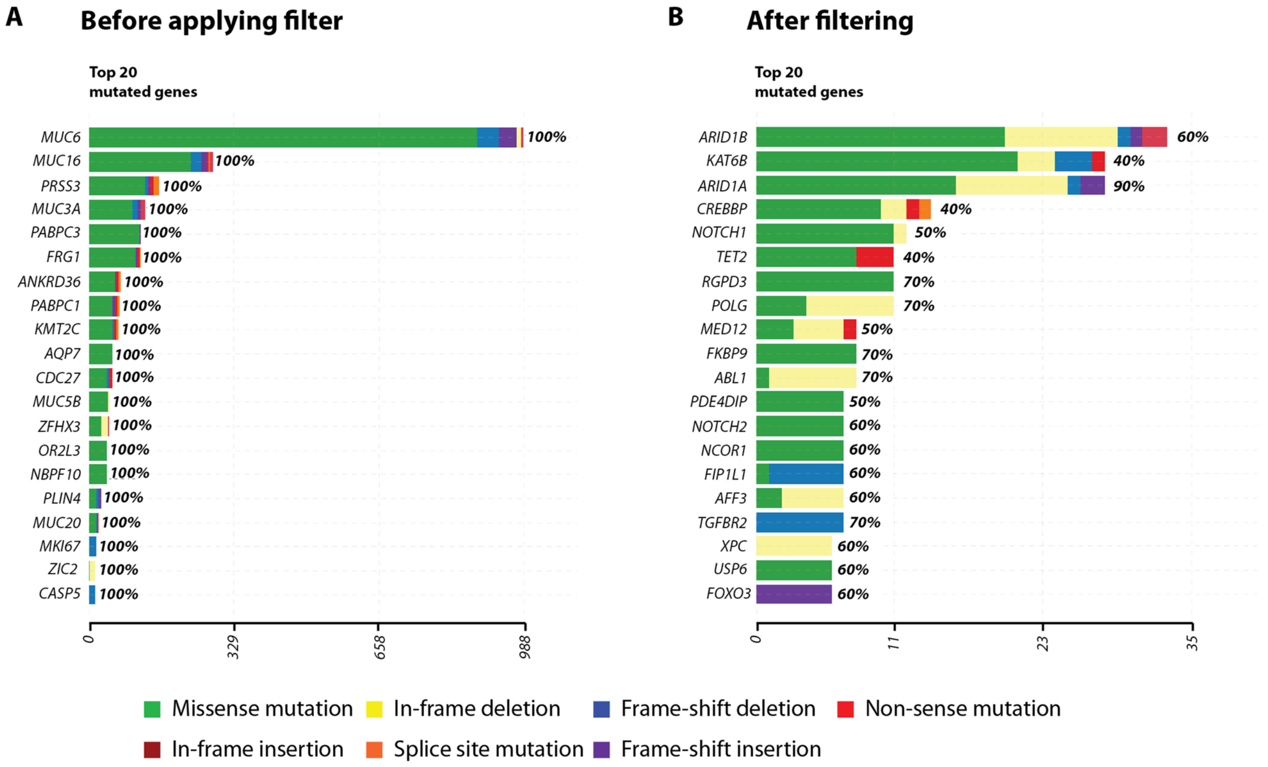

Fig. 3 Custom filtering of putative false-positive mutations. The top 20 mutated genes are show before (**A**) and after (**B**) applying the custom filters. This shows that without this filtering step, a number of genes score positive in 100% of the patients, including gene families containing variable number of tandem repeats (e.g., the *MUC* gene family). After applying a threshold (maximum of 4 variants per gene in a single patient, in a maximum of 20% of the patients) and selecting the option to retain genes present in the COSMIC database, the resulting mutated gene list is highly enriched in known oncogenic drivers and previously reported AEL-mutated genes

each category, users can choose between three different graphical outputs including bar graphs, association or enrichment factor along with color-coded p-value representations (Fig. 4**A**–**C**). These outputs are dynamic and customizable as users can switch from one representation to another or increase/decrease the number of categories to display in one click. Var|Decrypt also provides a somatic interaction view in order to identify which gene mutations tend to co-occur or are mutually exclusive. In the example shown in Fig. 4D, BCOR and XPC mutations seem to be mutually exclusive, suggesting that inhibiting BCOR activity in XPC-mutated AEL cells (and vice versa) may be therapeutically beneficial.

Another useful built-in feature is the enrichment of mutations in genes belonging to known oncogenic signaling pathways. This feature provides a graph representation of the enriched mutated pathways together with the number of patients bearing mutations in the related pathways (Fig. 5A). By simply clicking on a given pathway (right part), users can display a detailed list of the genes contained in the chosen pathway and check which gene and which patient sample harbor the mutation(s). This representation is useful to identify the recurrently mutated genes within a single oncogenic pathway and to check which signaling component is frequently altered in the disease. The example depicted in Fig. 5B shows that the Receptor tyrosine kinase/RAS pathway, and the Notch and TGF-β pathways are frequently altered in AEL patients, and that the ABL Proto-Oncogene 1 (*ABL1*), *NOTCH2*, and TGF-β receptor 2 (*TGFBR2*) receptor genes are among the top mutated genes.

Visualization of mutational hotspots and amino acid changes

Finally, Var|Decrypt provides a visualization tool depicting the localization of mutations on a given gene product (protein). The known protein domains are displayed along with the position of the various mutations or variants detected, with a color code indicating the variant types (STOP gain, frameshifts, non-synonymous SNPs). This feature allows to detect mutational hotspots and preferential localization of mutations in functional protein domains, as shown in Fig. 6A in the succinate dehydrogenase complex flavoprotein subunit A (*SDHA*) gene. In addition, a table provides the identity of amino acid changes along with several variant metrics (Additional file 3: Table S2).

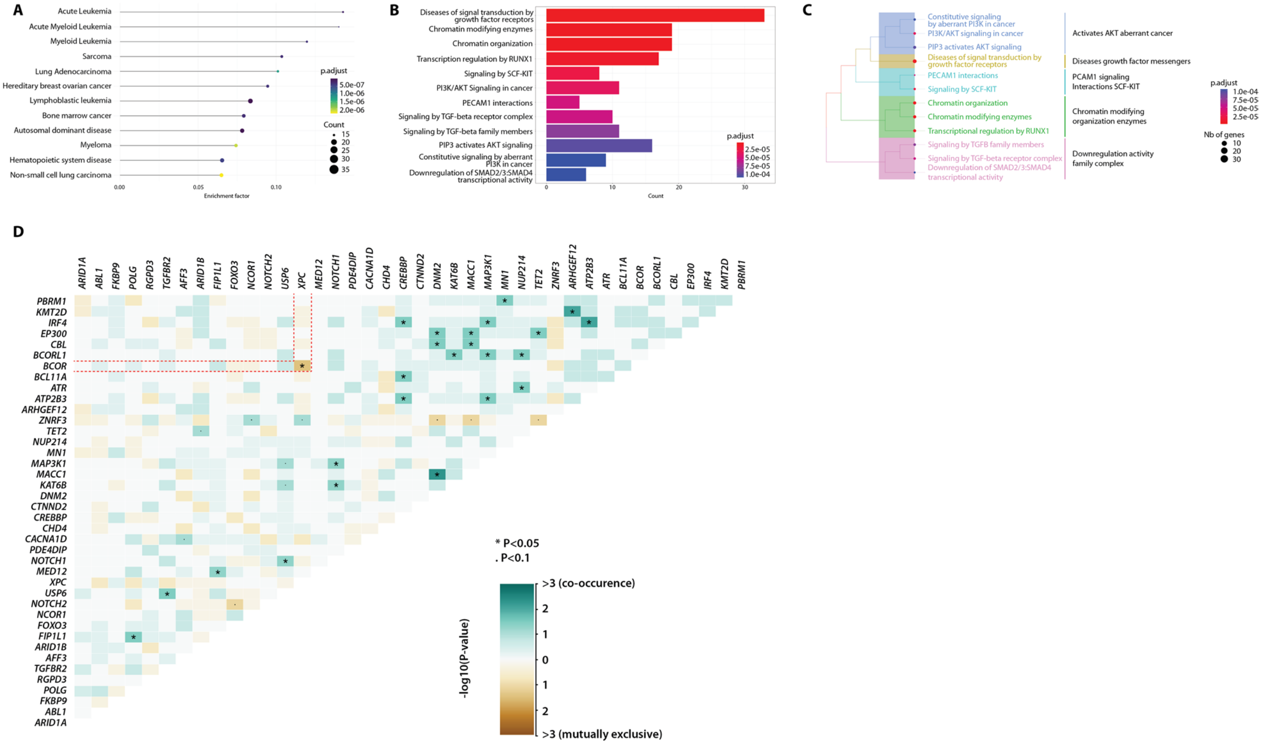

Fig. 4 Disease and pathway enrichment features. Var|Decrypt allows to depict various enrichment plots using enrichment factor (**A**), qValue bar plot (**B**) or cluster tree (**C**) visualization for disease ontology, biological pathways (Reactome, Wiki, KEGG), various gene ontology (GO) categories (biological process, molecular function, cellular component), gene–disease and gene–cancer associations. **D** Matrix showing the mutually exclusive (brown) or cooperating mutations (green) in the AEL patient cohort. Dashed red lines highlight the mutually exclusive *BCOR* and *XPC* mutations

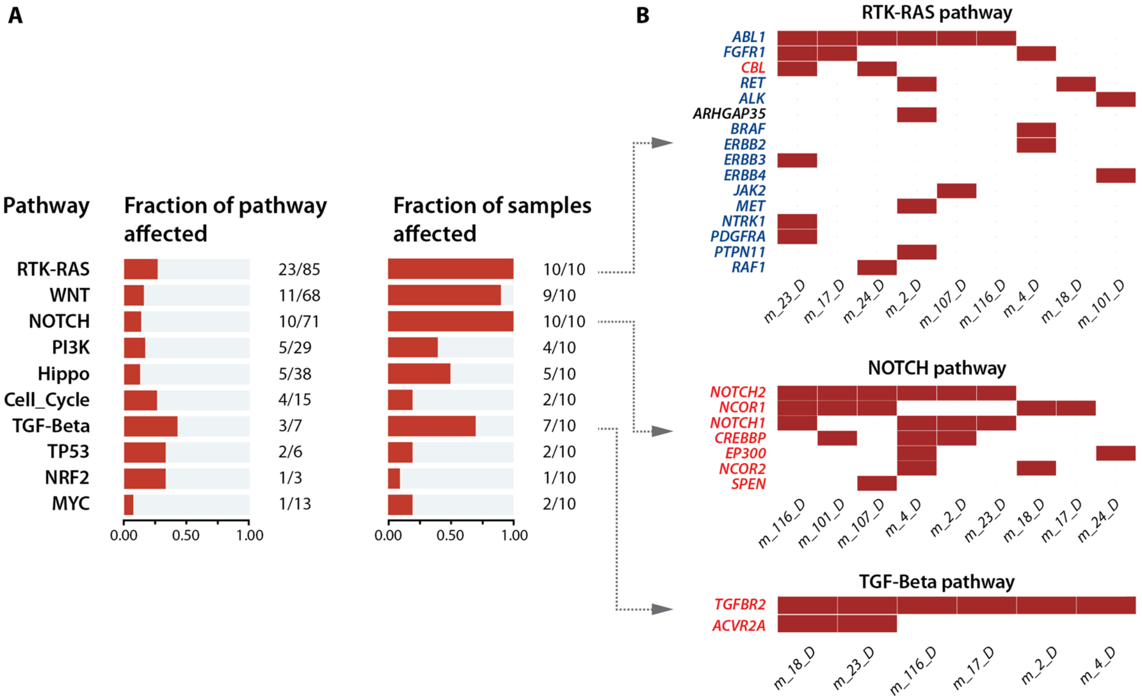

Fig. 5 Overview of oncogenic pathway alterations in the patient cohort. The oncogenic pathways affected in the AEL patients are shown (**A**) with the number of affected genes in relation to the total number of genes linked to each pathway. The second plot displays the fraction of patients bearing mutations in a given pathway. **B** For each pathway, Var|Decrypt allows visualizing the mutated genes for each patient to easily spot the recurrently mutated genes. Oncogenes and tumor suppressor genes are indicated in blue and red, respectively

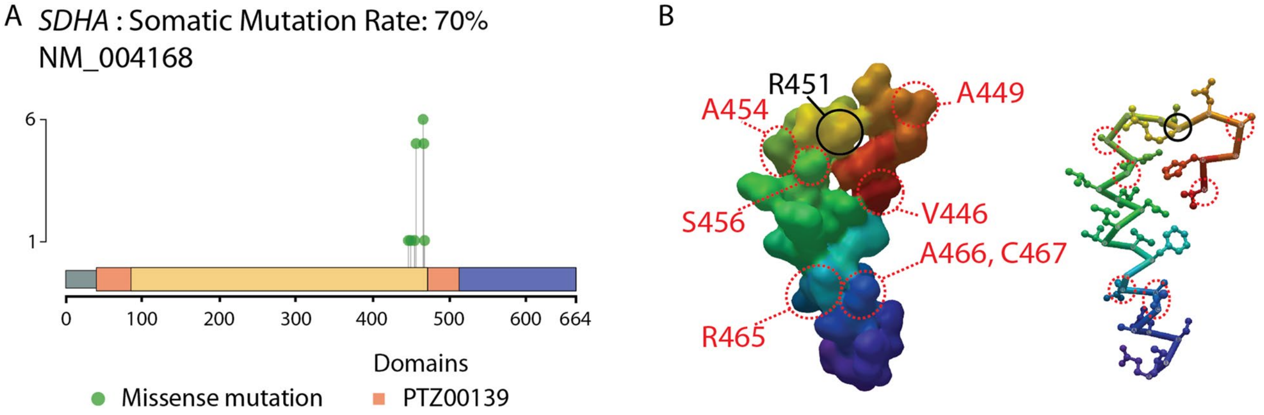

Fig. 6 Visualizing mutational hotspots. **A** The 'amino-acid changes' page displays the protein domains with the localization and type of mutations in the entire cohort. The example of the *SDHA* gene is shown. **B** Structure (accession #6VAX) of the SDHA active site from [44]. Only amino acids 446 to 472 are shown. The key active site residue R451 is indicated in black, the positions of the mutated residues in AEL are shown in red (left). Right, similar representation using the 'schematic view, with amino acid side chains shown as sticks and balls using MMDB viewer [55]

Discovery of putative novel oncogenic mutations in AEL

We applied Var|Decrypt to decipher the mutational landscape of AEL. Besides the known and recently described mutations in *TET2* (40% of patients in our cohort), *TP53* (30% patients), *EZH2* (10%), *NCOR1/2* (50%/20%) or *GATA1* (in 10% of the patients), our tool highlighted mutational hotspots in several additional genes, likely representing important components of the AEL mutational landscape. We identified several mutations within the *SDHA* gene (Fig. 6A), a critical member of the succinate dehydrogenase complex, which were not primarily identified in the previous AEL studies [30–34]. The succinate dehydrogenase (SDH) complex is a mitochondria-localized multiprotein complex involved in cellular respiration through the electron transfer chain (complex II) [35, 36]. The SDH is nuclearly encoded and composed of 4 subunits (SDHA-D). Loss of function of any of SDH subunits may associate with neuroendocrine tumors or neurodegenerative disorders such as Leigh's disease [36]. We identified a mutational hotspot within the *SDHA* gene in 70% of patients (Fig. 6A). Interestingly, all detected missense mutations (V446A, A449V, A454T, S456L, R465Q, A466T, C467S) cluster around the key *SDHA* active site residue SDHA^{R451} [35] (Fig. 6B). Although the precise functional impact of such mutations is currently unknown, some or all may significantly alter SDHA active site spatial conformation and lead to (partial) insufficiency. Measuring complex II activity in AEL cells and its requirement for leukemia development is out of the scope of this study but represents an interesting lead to follow, as complex II alterations may be of importance for the development or maintenance of AEL.

Finally, on a more global scale, Var|Decrypt allowed us to identify enrichment of mutations in oncogenic signaling pathways, in particular, the receptor tyrosine kinase (RTK)/RAS pathway, with a high prevalence of IRS1 mutations (40% of patients, including in-frame deletions, non-synonymous SNVs and a STOP gain), and mutations in *FGFR1*, *RET*, *JAK2* (30, 20, 10% of patients, respectively), or BRAF (10%). Importantly, our tool highlighted the Notch pathway as frequently mutated in AEL with Notch1 and Notch 2 receptor variants found in 40% and 60% of the patients, respectively.

Taken altogether these data highlight putative novel oncogenic processes and pathways in AEL, and underscore the usefulness of Var|Decrypt to provide leads for functional explorations.

Validation of Var|Decrypt using an independent dataset of 90 multiple myeloma samples

We sought to validate Var|Decrypt using an independent dataset. To this aim, we analyzed published WES data from 30 human multiple myeloma cell lines (HMCLs) and primary multiple myeloma (MM) from 59 patients [29]. Previous analysis of these data revealed a prevalent TP53 mutational landscape and altered MAPK pathways. Reanalysing this dataset with Var|Decrypt after filtering out the putative false-positive hits (i.e., highly mutated gene families such as MUC genes, see above) using the filtering options (frequency less than 4 mutations by gene within 20% of the cohort), and after crossing the mutated gene list with cancer gene databases (COSMIC, as in [29]) led to very similar identification of MM mutated hits, with frequent *TP53* (47%), *KRAS* (40%), *NRAS* (30%), *ATM* (33%) alterations, and many epigenetic modifiers (*BRD3*, *BRD4*, *SETD1B*) and DNA repair proteins (*FANCD2*, *RECQL4*) (Additional file 4: Table S3). In particular, we identify the MAPK/RAS pathway as recurrently altered (Additional file 9: Fig. S3 and Additional file 10: Fig. S4) [37, 38], validating the functionality of Var|Decrypt.

Performance

Var|Decrypt inputs are generated by one of the two provided pipelines (Additional file 7: Fig. S1). These pipelines can be used locally or on a cluster. As trimming and alignment steps are resource consuming, the pipeline handling data from fastq format are highly recommended to be used on a cluster. To test the compatibly of our annotation pipeline with large range of aligners and variant calling tool, we have tested it using publicly available VCF files with different file sizes from different variant calling tools [2] using two different methods of deployment (Additional file 5: Table S4). As Var|Decrypt is an RShiny application, it should run without issues on the majority of web browsers. We have measured the performance of Var|Decrypt using public data of primary multiple myeloma (MM) from 59 patients [29]. The time of the processing and the memory resource usage were evaluated during: (1) the processing of new data; (2) the reload of already processed data (Additional file 6: Table S5, Additional file 11: Fig. S5). These measurements indicate that Var|Decrypt is a fast-operating tool, with loading and processing times ranging from seconds to minutes (< 3 min 5 s for the larger datasets on a regular laptop with 8 GB of RAM, < 1 min and 30 s with 16 GB of RAM).

Comparison with other available tools

Other available tools provide some of Var|Decrypt functionalities, but either are (i) only available online, (ii) require bioinformatics expertise to prepare data and export the results in a human readable format, (iii) handle only one type of variants (i.e., somatic/germline), or (iv) do not support variant and enrichment results visualization (Table 1).

Discussion

Although WES is widely used to diagnose human diseases or to discover pathological mutations in Mendelian disorders and cancers, there is a surprising paucity of accessible and easy to use tools for WES analysis. A major hurdle is the presence of hundreds to thousands of variants routinely detected in WES datasets, complicating the identification of causative mutations and prioritization of variants in complex samples such as tumor biopsies. We present here Var|Decrypt, a fully automated dynamic interface for Exome-Seq data analysis. From a

Table 1 Comparison of VarlDecrypt with other available tools

	Variant prioritization tools						Enrichment analysis tools					
	Gene centered tools		Variant centered tools									
Name	Var|Decrypt	maftools	VCF-miner	BrowseVCF	BiERapp	GeMSTONE	clusterProfiler/ DOSE	GOrilla	geneontology	gprofiler/2	wKGGSeq	Enrichr
Variant type												
Germline	Yes	No	-	-	Yes	Yes	-	-	-	-	-	-
Somatic	Yes	Yes	-	-	Yes	No	-	-	-	-	-	-
Enrichment analysis												
Input Files	pipeline output (2 files)	maf / txt files	VCF	VCF	VCF	Internal format	List of genes	List of genes	List of genes	List of genes	VCF	List of genes
Export results	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Explore and filter variants by gene and by pati	Yes	Yes	No	No	No	No	-	-	-	-	-	-
Gene mutation rate in the cohort	Yes	Yes	No	No	No	-	-	-	-	-	-	-
Dynamic exploration	Yes	No	Yes	Yes	Yes	No	No	No	No	No	No	No
Independent output table by enrichment	Yes	No	No	No	No	No	Yes	Yes	Yes	Yes	Pathway only	Yes
All involved genes in the enriched term	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	-	Yes
Compare Tumoral/control	Yes	No	No	Yes	Yes	No	Yes	No[(2)]	No	No	No	No
Visualization												
General statistics	Somatic only	Yes	No	No	Yes	Yes	No	No	No	No	Yes	No
Dynamic enrichment plots	Yes	No	No	No	No	No	No	No	No	No	No	Yes
Customizable	Yes	Yes	No	No	No	No	Yes	No	No	No	No	No
Use												
User-friendly interface	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
programming skills-free	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	No/Yes	Yes	Yes
Run locally (security)	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No/Yes	No	No
Run offline	Yes[(1)]	Yes	Yes	Yes	-	No	Yes	No	No	No	No	No

The table depicts the various advantages of Var I Decrypt over the main online tools for gene ontology and biological feature enrichment analyses. (1) only Kegg enrichment analysis needs an internet connection. (2) Control can be used only as background computational point of view, Var|Decrypt represents a fast, easy to use and flexible tool, which can be installed via Docker on several operating systems (Linux, macOS), downloaded (open-source scripts) to run via Rstudio, or directly used online (https://vardecrypt.com/app/vardecrypt). In addition, Var|Decrypt can be deployed on a Linux sever via the open-source Shiny Server software to be available to a large number of users. It is worth noting that the shiny server solution is suitable for institutions which need to centralize computing resources (Var|Decrypt docker image is configurated already to be deployed on a shiny sever). An appropriate detailed user tutorial is available at https://gitlab.com/mohammadsalma/vardecrypt. Thanks to its user-friendly interface and the variety of incorporated analysis options, Var|Decrypt can easily be used by both bioinformaticians and non-programming experts/wet lab scientists. It, therefore, fills a gap by providing a complete solution for WES data analyses.

The primary goal of our tool is to provide non-specialists with a ready-to-use solution for easy browsing and exploration of Exome-Seq datasets. The rational is that driver mutations or gene mutations that are important for the development or maintenance of a disease state should be overrepresented or enriched in a patient cohort. It is still currently a major challenge to accurately and consistently classify variant pathogenicity [39–41]. Although, unlike previous tools such as VCF-miner, BrowseVCF, BiERapp and GeMSTONE, we do not directly address the pathogenicity of detected variants, we reasoned that focusing on the recurrently mutated genes may provide an increased likelihood of identifying important disease targets. Indeed, finding multiple different variants (possibly being of unknown pathogenicity) affecting the same gene product in the patient cohort would increase the chance to highlight disease drivers. In that respect, rather than focusing on the variants themselves, Var|Decrypt provides a gene-centered analysis, allowing distinct variants affecting the same gene to score similarly. The key benefit of Var|Decrypt over maftools is that the latter requires specialized bioinformatics knowledge to prepare input data and output findings in a manner that can be read by scientists lacking bioinformatic skills. In addition, maftools only handle somatic variants. Hence, Var|Decrypt resulting tables and analyses provide users with a quick overview of the recurrently mutated genes and associated functions (pathways), representing putative disease drivers. The numerous built-in features allow extraction of such meaningful information, e.g., recurrently mutated genes, or recurrently mutated pathway components and highlighting the mechanistic features of human disorders. We provide example of the analysis of AEL, a rare, poorly characterized and particularly aggressive subtype of leukemia. While recent work have started to shed light on the mutational landscape of AEL [30–34, 42], we show here that frequent NOTCH pathway alteration is associated with AEL in our cohort. We also identified mutational hotspots in the mitochondrial complex II component SDHA. This underscores the utility of dedicated analysis tools such as Var|Decrypt to highlight oncogenic signaling pathways and mutated genes that have been overlooked in other studies. Whether NOTCH pathway alteration represents a common feature of AEL and can be exploited as therapeutic vulnerability remains to be tested and is beyond the scope of this article. However, this represents an example of how biological information can be extracted from complex Exome-seq datasets without knowledge in bioinformatics. We further showed that Var|Decrypt could detect known recurrently mutated genes in human multiple myeloma and could highlight signaling pathways known to be important for this disease.

Conclusions

Taken altogether, these data indicate that Var|Decrypt represents a functional and attractive tool allowing efficient analysis of WES data, with the overarching goal to facilitate functional studies and guide therapeutic decisions. We expect that thanks to its ease of use and simple user-friendly interface, Var|Decrypt will help the clinical and biological scientist communities to get critical insight into the molecular mechanisms of human disorders.

Availability and requirements

Project name: Var|Decrypt. Project home page: https://gitlab.com/mohammadsalma/vardecrypt. Operating system(s): Linux, MacOS and Windows. Programming language: R and Python3. Other requirements: R 4.1 or higher, Snakemake 6 or higher and pandas 1.3.5 or higher. License: CeCill-C (http://www.cecill.info/licences/Licence_CeCILL-C_V1-en.html) and GPLv3. Any restrictions to use by non-academics: license needed.

Methods

Data

We used WES data from primary erythroleukemia samples [31]. The data are separated into two types, namely 'Tumoral' samples, being the patient leukemic blasts samples, and 'Normal' representing the non-leukemic matched controls, considered healthy (non-leukemic marrow cells), and used for variant filtering purposes. Therefore, for each patient, matched leukemic sample (tumoral) and a control sample (normal) are used.

Whole-exome sequencing analysis pipeline

The human reference genome hg19 (https://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/) was indexed by BWA. Reads were trimmed with Trimmomatic (version0.36) to eliminate sequencing adapters and low-quality reads. Mapping was performed using BWA-MEM (version 0.7.17) [43] with default parameters. SAM files were converted, sorted and indexed by Samtools (version 1.9) [44]. To improve alignment quality, MarkDuplicates tool from PICARD (version 2.17.11) [45] was used to locate and tag duplicate reads within the BAM. Before preforming base recalibration for BAM files by BaseRecalibrator from GATK (version v3.8-1-0), reads were processed by AddOrReplaceReadGroups from PICARD (version 2.17.11) to define a group for all reads generated from the same run. For the variant calling step, the HaplotypeCaller was used for germline variants and Mutect2 for somatic ones from GATK4 (version 4.0.3.0) [46]. Variant filtering and annotation were performed by ANNOVAR (version Sun, 7 Jun 2020) [47]. First, results from the previous step are filtered by extracting all the known SNV involved in one or more human diseases (regardless of disease using the option: -filter -dbtype clinvar_20200316). Then, all known human variants were ignored (-filter -dbtype 1000g2015aug_all) [47]. Finally, all unknown variants were grouped with the pathogenic variants for each patient/sample data. After the annotation step by ANNOVAR, only exonic polymorphisms were considered and kept in the output file. Using custom Python scripts, synonymous variants were removed from the final output files. The resulting files were used as input files for Var|Decrypt.

Exome sequencing analyses using Var|Decrypt

Downstream analyses and variants prioritization were performed using Var|Decrypt, which mainly uses the following R libraries:

(1) Shiny, which allows to develop a user-friendly graphical interface to visualize the various types of data. This application can be launched in Rstudio or in any modern web browser such as firefox, chrome or safari [48]; (2) DOSE which allows to perform enrichment analyses of a set of genes to discover gene–disease associations. It implements several methods to measure the semantic similarities between the DO (Disease ontology) terms and the different gene products [49]; (3) clusterProfiler, which implements methods for analyzing and visualizing functional profiles from gene clusters [50]; (4) org.Hs.eg.db, which contains annotation of the human genome [51]; (5) ReactomePA, which provides signaling pathway analysis functions based on the REACTOME database with several visualization functions [52]; (6) networkD3, which was used to generate reactive 3D networks [53]. Finally, Var|Decrypt also uses maftools R package to process somatic variants [54].

Abbreviations

AEL	Acute erythroid leukemia
GO	Gene-ontology
HMCL	Human multiple myeloma cell line
MM	Multiple myeloma
SNP	Single-nucleotide polymorphism
SNV	Single-nucleotide variant
VCF	Variant calling file
WES	Whole-exome sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13072-023-00497-4.

Additional file 1: Exome-seq variant analysis pipeline.

Additional file 2: Table S1. Top 100 mutated genes in the AEL cohort.

Additional file 3: Table S2. Variants and protein metrics.

Additional file 4: Table S3. Top 100 mutated genes in HMCL.

Additional file 5: Table S4. Performance of pre Var|Decrypt annotate vcf pipeline using two different methods of deployment. All times are in seconds.

Additional file 6: Table S5. Performance of Var|Decrypt after loading two types of input date in different operating systems with two different methods of deployment. All times are in seconds.

Additional file 7: Figure S1. Bioinformatic pipelines used to process whole exome sequencing data. The WES and Var|Decrypt pipelines are depicted on the left and right, respectively. The various filtering steps and packages used are indicated. Two versions of the WES pipeline are available (The part highlighted in green corresponds to the pipeline to process WES data starting from fastq files; in orange the one allowing users todirectly processing vcf files).

Additional file 8: Figure S2. Example of Var|Decrypt front page showing the mutated gene list together withthe frequencies in the cohort and the mutation types.

Additional file 9: Figure S3. Analysis of WES from 30 human multiple myeloma cell lines. (A) Altered pathways over represented in the HMCL mutated genes. (B) mutations in oncogenic signaling pathways showingthat the RTK-RAS and NOTCH pathways are among the top mutated pathways.

Additional file 10: Figure S4. Detailed view and frequencies of the RTK-RAS pathway mutated genes. Affected genes belonging to the RTK-RAS pathway are shown and highlighted when mutated for each sample. (A)The figure shows prevalent FGFR4 (17 out of 29 samples, 58%), and KRAS (12 out of 29 samples, 41%) mutationsin HMCL, and (B) similar frequencies (FGFR4 38/59, 64%; KRAS 17/59, 28%) were observed in primary humanmultiple myeloma samples.

Additional file 11: Figure S5. Comparison of memory resource usage of Var|Decrypt in different operatingsystems with two different methods of deployment. A) Comparison of memory resource usage during a newdata analysis. B) Comparison of memory resource usage during the reload of already analyzed data. Blue line:Var|Decrypt is installed locally on MacBook pro M1 2020 with 16GB of RAM. Green line: Var|Decrypt is deployed using docker container on a MacBook Pro M1 2020 with 16 GB of RAM. Red line: Var|Decrypt is deployed using docker container on Ubuntu (virtual machines in cloud) with 4 CPU (2GHz AMD) and 8 GB of RAM. Memory usage had been estimated using the command "ps -caxm -o rss, comm" on macOS and "ps -eo rss, comm" on ubuntu.

Acknowledgements
We wish to thank the KIM Interdisciplinary Blood Science network from Montpellier University for their excellent support; Samad El Kaoutari for assistance in the initial steps of Exome-Seq analysis pipeline development, and the GOELAMS for granting access to primary AEL samples.

Author contributions
MS designed the pipeline, wrote the source code and built the tool. EA, JM, ES provided samples and WES data. MS and ES tested and evaluated the tools, and wrote the manuscript. All the authors approved the manuscript. ES supervised the study. All the authors read and approved the final manuscript.

Funding
This work was supported by the Laboratory of excellence Labex EpiGenMed (program "Investissements d'avenir" ANR-10-LABX-12–01); the Laboratory of Excellence GR-Ex (program "Investissements d'avenir" ANR-11-LABX-0051, ANR-18-IDEX-0001); Fondation pour la Recherche Médicale (Equipe FRM DEQ20180339221); Ligue Nationale Contre le Cancer, ANR—AAPG 2022—PRC—ZEBERY; ANR-18-CE15-0010–01 PLASMADIFF-3D, FFRMG (AAP-FFRMG-2021), INSERM PSCI 2020 Smooth-MM, LabUM Metabohox, and Institut Universitaire de France. The funding sources had no involvement in study design, data collection, analysis, interpretation or manuscript writing.

Availability of data and materials
The AEL datasets analyzed during the current study are from [31] and are available in the European Genome-Phenome Archive (EGA) under the accession EGAS00001004203. The AEL, HMCL and MM variant calling files (vcf) were deposited at https://5dkp.short.gy/vardecrypt_data.

Declarations

Ethics approval and consent to participate
All human patient sample data used in this study were obtained from publicly available sources [29, 31] (see "Availability of data and materials"). As stated in [31] and [29], patient samples were obtained with the written informed consent of the patient or their family in accordance with the Declaration of Helsinki, and the study was approved by the Gustave Roussy Institutional Review Board [31] and the institutional research board from Montpellier University hospital [29].

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 13 January 2023 Accepted: 23 May 2023
Published online: 14 June 2023

References

1. Manolio TA, Chisholm RL, Ozenberger B, et al. Implementing genomic medicine in the clinic: the future is here. Genet Med. 2013;15:258–67.
2. Xiao W, Ren L, Chen Z, et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. Nat Biotechnol. 2021;39:1141–50.
3. Bertier G, Hétu M, Joly Y. Unsolved challenges of clinical whole-exome sequencing: a systematic literature review of end-users' views. BMC Med Genomics. 2016;9:52.
4. Smith HS, Swint JM, Lalani SR, et al. Clinical application of genome and exome sequencing as a diagnostic tool for pediatric patients: a scoping review of the literature. Genet Med. 2019;21:3–16.
5. Jacob HJ, Abrams K, Bick DP, et al. Genomics in clinical practice: lessons from the front lines. Sci Transl Med. 2013;5:194cm5-194cm5.
6. Thevenon J, Duffourd Y, Masurel-Paulet A, et al. Diagnostic odyssey in severe neurodevelopmental disorders: toward clinical whole-exome sequencing as a first-line diagnostic test. Clin Genet. 2016;89:700–7.
7. Vrijenhoek T, Kraaijeveld K, Elferink M, et al. Next-generation sequencing-based genome diagnostics across clinical genetics centers: implementation choices and their effects. Eur J Hum Genet. 2015;23:1142–50.
8. Binatti A, Bresolin S, Bortoluzzi S, et al. iWhale: a computational pipeline based on docker and SCons for detection and annotation of somatic variants in cancer WES data. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbaa065.
9. Frontiers | Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift | Genetics. https://www.frontiersin.org/articles/https://doi.org/10.3389/fgene.2012.00035/full. Accessed 4 Dec 2020.
10. Paila U, Chapman BA, Kirchner R, et al. GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput Biol. 2013;9:e1003153.
11. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
12. Home-QIAGEN Digital Insights. Bioinformatics Software and Services | QIAGEN *Digital Insights*. https://digitalinsights.qiagen.com/. Accessed 4 Dec 2020.
13. SNP & Variation Suite (SVS)-Golden Helix. https://www.goldenhelix.com/products/SNP_Variation/index.html. Accessed 4 Dec 2020.
14. Alemán A, García-García F, Salavert F, et al. BiERapp: A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. Nucleic acids research. 2014;42(W1):W88–W93. https://doi.org/10.1093/nar/gku407.
15. Coutant S, Cabot C, Lefebvre A, et al. EVA: exome variation analyzer, an efficient and versatile tool for filtering strategies in medical genomics. BMC Bioinformatics. 2012;13:S9.
16. Smedley D, Jacobsen JOB, Jäger M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015;10:2004–15.
17. Alexander J, Mantzaris D, Georgitsi M, et al. Variant ranker: a web-tool to rank genomic data according to functional significance. BMC Bioinformatics. 2017;18:341.
18. Salatino S, Ramraj V. BrowseVCF: a web-based application and workflow to quickly prioritize disease-causative variants in VCF files. Brief Bioinform. 2017;18:774–9.
19. Dahary D, Golan Y, Mazor Y, et al. Genome analysis and knowledge-driven variant interpretation with TGex. BMC Med Genomics. 2019;12:200.
20. Hart SN, Duffy P, Quest DJ, et al. VCF-Miner: GUI-based application for mining variants and annotations stored in VCF files. Brief Bioinform. 2016;17:346–51.
21. Chen S, Beltrán JF, Esteban-Jurado C, et al. GeMSTONE: orchestrated prioritization of human germline mutations in the cloud. Nucleic Acids Res. 2017;45:W207–14.
22. Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinform. 2013;14:128.
23. Eden E, Navon R, Steinfeld I, et al. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:48.
24. Li MJ, Deng J, Wang P, et al. wKGGSeq: a comprehensive strategy-based and disease-targeted online framework to facilitate exome sequencing studies of inherited disorders. Hum Mutat. 2015;36:496–503.
25. Gene ontology resource. Gene ontology resource. 2020. http://geneontology.org/. Accessed 4 Dec.
26. Raudvere U, Kolberg L, Kuzmin I, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47:W191–8.
27. Debailleul V, Laine A, Huet G, et al. Human mucin genes MUC2, MUC3, MUC4, MUC5AC, MUC5B, and MUC6 express stable and extremely large mRNAs and exhibit a variable length polymorphism. an improved method to analyze large mRNAs. J Biol Chem. 1998;273:881–90.
28. Kirby A, Gnirke A, Jaffe DB, et al. Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. Nat Genet. 2013;45:299–303.
29. Vikova V, Jourdan M, Robert N, et al. Comprehensive characterization of the mutational landscape in multiple myeloma cell lines reveals potential drivers and pathways associated with tumor progression and drug resistance. Theranostics. 2019;9:540–53.
30. Iacobucci I, Wen J, Meggendorfer M, et al. Genomic subtyping and therapeutic targeting of acute erythroleukemia. Nat Genet. 2019;51:694–704.

31. Fagnan A, Bagger FO, Piqué-Borràs M-R, et al. Human erythroleukemia genetics and transcriptomes identify master transcription factors as functional disease drivers. Blood. 2020;136:698–714.
32. Cervera N, Carbuccia N, Garnier S, et al. Molecular characterization of acute erythroid leukemia (M6-AML) using targeted next-generation sequencing. Leukemia. 2016;30:966–70.
33. Cervera N, Carbuccia N, Mozziconacci M-J, et al. Revisiting gene mutations and prognosis of ex-M6a-acute erythroid leukemia with regard to the new WHO classification. Blood Cancer J. 2017;7: e594.
34. Grossmann V, Bacher U, Haferlach C, et al. Acute erythroid leukemia (AEL) can be separated into distinct prognostic subsets based on cytogenetic and molecular genetic characteristics. Leukemia. 2013;27:1940–3.
35. Sharma P, Maklashina E, Cecchini G, et al. The roles of SDHAF2 and dicarboxylate in covalent flavinylation of SDHA, the human complex II flavoprotein. Proc Natl Acad Sci U S A. 2020;117:23548–56.
36. Sharma P, Maklashina E, Cecchini G, et al. Maturation of the respiratory complex II flavoprotein. Curr Opin Struct Biol. 2019;59:38–46.
37. Lohr JG, Stojanov P, Carter SL, et al. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. Cancer Cell. 2014;25:91–101.
38. Bolli N, Avet-Loiseau H, Wedge DC, et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. Nat Commun. 2014;5:2997.
39. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology. Genet Med. 2015;17:405–24.
40. McInnes G, Sharo AG, Koleske ML, et al. Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. Am J Hum Genet. 2021;108:535–48.
41. Nicora G, Zucca S, Limongelli I, et al. A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. Sci Rep. 2022;12:2517.
42. Micci F, Thorsen J, Panagopoulos I, et al. High-throughput sequencing identifies an NFIA/CBFA2T3 fusion gene in acute erythroid leukemia with t (1; 16)(p31;q24). Leukemia. 2013;27:980–2.
43. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio. 2013. http://arxiv.org/abs/1303.3997. Accessed 4 Dec 2020.
44. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008.
45. Picard toolkit. Broad Institute, GitHub repository. (2018). http://broadinstitute.github.io/picard/. Accessed 4 Dec 2020.
46. Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at bioRxiv 2017. https://doi.org/10.1101/201178.
47. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38: e164.
48. Chang W, Cheng J, Allaire JJ, et al. Shiny: Web Application Framework for R. 2020. https://CRAN.R-project.org/package=shiny. Accessed 4 Dec 2020.
49. Yu G, Wang L-G, Yan G-R, et al. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2015;31:608–9.
50. Yu G, Wang L-G, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS J Integr Biol. 2012;16:284–7.
51. Carlson M, org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2. Bioconductor 2019. https://doi.org/10.18129/b9.bioc.org.hs.eg.db.
52. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst. 2016;12:477–9.
53. Allaire JJ, Ellis P, Gandrud C, et al. networkD3: D3 JavaScript Network Graphs from R. 2017. https://CRAN.R-project.org/package=networkD3. Accessed 4 Dec 2020.
54. Mayakonda A, Lin D-C, Assenov Y, et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018;28:1747–56.
55. Madej T, Lanczycki CJ, Zhang D, et al. MMDB and VAST+:tracking structural similarities between macromolecular complexes. Nucleic Acids Res 2014;42:D297–303.

Publisher's Note

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