SOFTWARE TOOL ARTICLE

REVISED CusVarDB: A tool for building customized sample-specific variant protein database from next-generation sequencing datasets [version 2; peer review: 2 approved, 1 not approved]

Sandeep Kasaragod[1], Varshasnata Mohanty[1], Ankur Tyagi[1], Santosh Kumar Behera[1], Arun H. Patil[1], Sneha M. Pinto[1], T. S. Keshava Prasad[1], Prashant Kumar Modi[1], Harsha Gowda[1,2]

[1]Center for Systems Biology and Molecular Medicine, Yenepoya Research Centre, Yenepoya (Deemed to be University), Mangalore, 575018, India
[2]Institute of Bioinformatics, International Technology Park, Bangalore, 560066, India

v2 **First published:** 11 May 2020, **9**:344
https://doi.org/10.12688/f1000research.23214.1
Latest published: 16 Nov 2020, **9**:344
https://doi.org/10.12688/f1000research.23214.2

Abstract

Cancer genome sequencing studies have revealed a number of variants in coding regions of several genes. Some of these coding variants play an important role in activating specific pathways that drive proliferation. Coding variants present on cancer cell surfaces by the major histocompatibility complex serve as neo-antigens and result in immune activation. The success of immune therapy in patients is attributed to neo-antigen load on cancer cell surfaces. However, which coding variants are expressed at the protein level can't be predicted based on genomic data. Complementing genomic data with proteomic data can potentially reveal coding variants that are expressed at the protein level. However, identification of variant peptides using mass spectrometry data is still a challenging task due to the lack of an appropriate tool that integrates genomic and proteomic data analysis pipelines. To overcome this problem, and for the ease of the biologists, we have developed a graphical user interface (GUI)-based tool called CusVarDB. We integrated variant calling pipeline to generate sample-specific variant protein database from next-generation sequencing datasets. We validated the tool with triple negative breast cancer cell line datasets and identified 423, 408, 386 and 361 variant peptides from BT474, MDMAB157, MFM223 and HCC38 datasets, respectively.

Keywords

Next-generation sequencing, Variant protein database, NGS-pipeline

Open Peer Review

Reviewer Status ✓ ✓ ✗

	Invited Reviewers		
	1	**2**	**3**
version 2 (revision) 16 Nov 2020	✓ report	✓ report	✗ report
version 1 11 May 2020	? report	? report	

1. **Suresh Mathivanan**, La Trobe University, Melbourne, Australia
2. **Richard Kumaran Kandasamy**, Norwegian University of Science and Technology, Trondheim, Norway
3. **Jorge Duitama**, Universidad de Los Andes, Bogotá, Colombia
   Daniel Mahecha, Universidad de Los Andes, Bogotá, Colombia

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Prashant Kumar Modi (prashantmodi@yenepoya.edu.in), Harsha Gowda (harsha@ibioinformatics.org)

Author roles: Kasaragod S: Conceptualization, Methodology, Project Administration, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Mohanty V**: Methodology, Validation, Writing – Review & Editing; **Tyagi A**: Methodology, Validation, Writing – Review & Editing; **Behera SK**: Conceptualization, Project Administration, Software, Writing – Review & Editing; **Patil AH**: Project Administration, Software, Writing – Review & Editing; **Pinto SM**: Supervision, Writing – Review & Editing; **Prasad TSK**: Conceptualization, Supervision, Writing – Review & Editing; **Modi PK**: Conceptualization, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Gowda H**: Conceptualization, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: SK is a recipient of the Indian Council of Medical Research (ICMR) Senior Research Fellow (SRF) application number [ISRM/11(27)/2017]. VM is a recipient of Women Scientist-A award from the Department of Science and Technology (DST). SKB is a recipient of DBT-BINC Junior Research Fellow. SMP is a recipient of INSPIRE Faculty Award from the Department of Science and Technology (DST), Government of India. We thank Karnataka Biotechnology and Information Technology Services (KBITS), Government of Karnataka, for the support to the Center for Systems Biology and Molecular Medicine at Yenepoya (Deemed to be University) under the Biotechnology Skill Enhancement Programme in Multiomics Technology (BiSEP GO ITD 02 MDA 2017).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Kasaragod S, Mohanty V, Tyagi A *et al.* **CusVarDB: A tool for building customized sample-specific variant protein database from next-generation sequencing datasets [version 2; peer review: 2 approved, 1 not approved]** F1000Research 2020, **9**:344 https://doi.org/10.12688/f1000research.23214.2

First published: 11 May 2020, **9**:344 https://doi.org/10.12688/f1000research.23214.1

REVISED **Amendments from Version 1**

The new modifications for this article are as follows:

1. We removed some sentences in articles to avoid exceeding the word limit

2. As per the reviewer 1 and 2 comments, we modified the contents in the Introduction and Use case section

3. Supplementary Table 3 have been added to address the reviewer-2 comments

4. References 26–30 have been added

Any further responses from the reviewers can be found at the end of the article

Introduction

Cancer genome sequencing projects have revealed thousands of genomic variations in cancers (Forbes *et al.*, 2010; Tate *et al.*, 2019; Tomczak *et al.*, 2015; Zhang *et al.*, 2011). Some mutants are oncogenic and drive proliferation in various cancers. For example, amino acid substitution of arginine to leucine at position 858 (L858R) in the epidermal growth factor receptor (EGFR) gene has been observed in 17% of pulmonary adenocarcinoma patients (Morgensztern *et al.*, 2015). In addition, a mutation in gene BRAF V600E is known to drive some melanomas (Chapman *et al.*, 2011). Some of these mutant proteins are proteolytically processed in cancer cells, resulting in major histocompatibility complex (MHC) presentation of mutant peptides. These mutant peptides serve as neo-antigens that recruit T cells and result in immune activation (Kreiter *et al.*, 2015). However, not all mutants are encoded at the protein level. Therefore, it is important to identify mutant proteins expressed by cancer cells. However, there are no easy-to-use pipelines for biologists to identify such coding variants, which alter the protein sequences and may play an important role in tumorigenesis. Detection of cancer-specific proteoforms has been studied by several research teams using proteogenomics methods. This approach integrates proteomics data with genome sequence data to identify protein complement of genomic variants (Menschaert & Fenyö, 2017; Nesvizhskii, 2014; Ruggles & Fenyö, 2016). Detection of coding variants can provide candidate molecules for novel therapeutic interventions (Subbannayya *et al.*, 2016).

Several tools have been developed in the last decade to carry out onco-proteogenomics. CPTAC (Ellis *et al.*, 2013) program was initiated to understand the complexity of cancer and its sub types using multi omics approach. Various tools and approaches have been employed to identify mutant peptides in cancers (Mathivanan *et al.*, 2012) (Yeom *et al.*, 2016) (Nagaraj *et al.*, 2015). Several qualitative and quantitative proteomics studies have been reported, which identify altered expression of proteins in cancers (Subbannayya *et al.*, 2015). Often, conventional workflows were used in such investigations, wherein a reference protein database was used to search tandem mass spectrometry data for identification and quantification of proteins (Kelkar *et al.*, 2014). However, such a reference database is usually devoid of sample-specific amino acid variations resulting from genomic alterations. The publicly available databases such as dbSNP (Sherry *et al.*, 2001), COSMIC (Forbes *et al.*, 2015) and UniProt (Apweiler *et al.*, 2004) can be used to identify the variant peptides (Alfaro *et al.*, 2017) but millions of protein-variants from these databases might increase the probability of identifying false positives. Therefore, there is a need for an improved method to identify sample specific sequence variations at the proteomic level. Tools such as CustomProDB (Wang & Zhang, 2013) and MZVar (https://bitbucket.org/sib-pig/mzvar-public/src/master/) have been developed for generating variant protein database. These tools require processed files such as VCF or BED file as an input. Executing the preprocessing steps requires knowledge of computation and also requires multiple tools to generate the output. Hence, we developed CusVarDB with an in-built pipeline for genomics suite to identify variants and create custom variant protein database.

Methods

Implementation

CusVarDB is available at https://sourceforge.net/projects/cusvardb/ and http://bioinfo-tools.com/Downloads/CusVarDB/V1.0.0/. Our tool supports graphical user interface (GUI) for easy execution of next-generation sequencing (NGS) pipelines. The GUI was developed using Microsoft Visual Studio Community edition 2017. Linux commands were executed through the Python program (terminal.py) developed by Linwei available on GitHub. Portable version of Perl was used to execute the Annovar scripts. A python script was used to generate the custom variant database; the scripts were made portable using PyInstaller. The dry-run concept is implemented in the tool to customize commands according to user's need and run in batches.

CusVarDB inherits different NGS pipelines for genomics, RNA-Seq and exome-seq datasets. The tool performs the following steps: i) alignment of genomic data to a reference genome; ii) variant calling; iii) variant annotation; and iv) generate variant protein database (Figure 1). Burrows-Wheeler Aligner (BWA) is executed for alignment of genome and exome (Li & Durbin, 2009) data, while HISAT2 is used for RNA-Seq data (Kim *et al.*, 2015). Subsequent steps involving sample labelling, variant calling and annotation are performed using Picard, GATK (McKenna *et al.*, 2010) and ANNOVAR (Wang *et al.*, 2010), respectively. CusVarDB substitutes the amino acid variations from the previous steps to generate a custom variant protein database.

Operation

CusVarDB runs on Windows 10 system. It requires Linux Bash Shell and ANNOVAR tool to be enabled and installed by the user. Minimum system requirements include Intel i5 or i7, having at least four cores with 8 GB of RAM and 1 TB hard drive. High performance processors such as Intel i9 or Xeon and large quantity of RAM can enable faster execution of tasks.

The tool requires FASTA, FASTQ and dbSNP database as input files. FASTA file will be loaded at configuration panel and indexed using bwa index or hisat2-build command. The quality control panel loads the FASTQ file and produces a quality check report using the FastQC tool. The third panel asks the user to

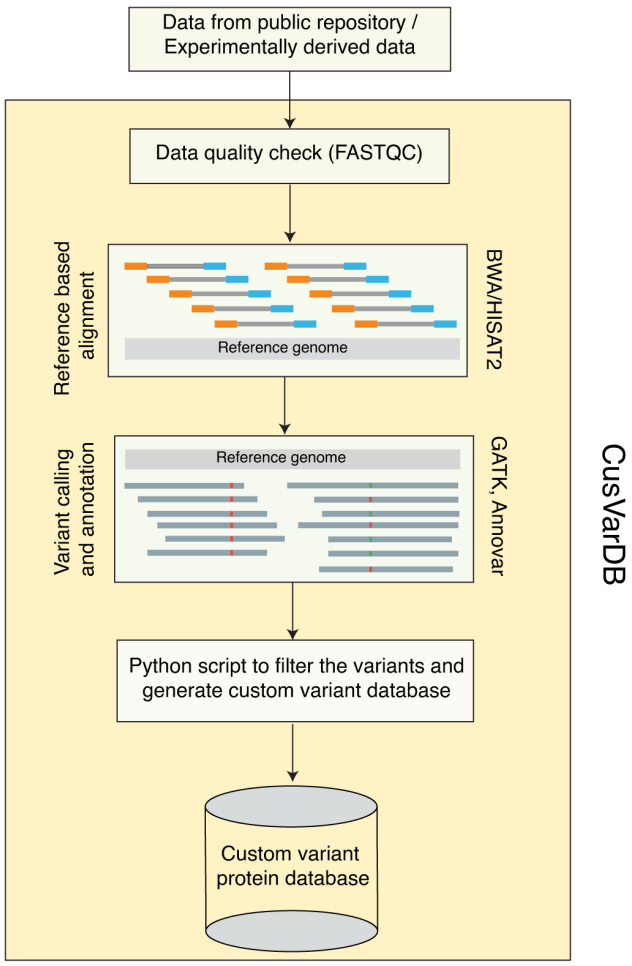

Figure 1. Schematic representation of the workflow.

load the FASTQ file(s), set the number of threads and RAM to perform the alignment and variant calling steps. This step produces the raw VCF file, which will be used for annotation in annotation tab. The annotation panel annotates the variants and incorporates those variants to the protein database to create a custom variant protein database. Detailed information of tool installation and execution of a test dataset are available in the manual (Manual.pdf) available to download here and on Zenodo (Kasaragod *et al.*, 2020b).

Use cases methods

The genomics datasets downloaded from the SRA repository were subjected to a quality check using FastQC tool. Poor quality reads with Phred score less than twenty were trimmed using fastx_trimmer. Trimmed reads were mapped to human reference genome version 19 (hg19). AddOrReplaceReadGroup and Markduplicates operations from GATK were performed to add label information and to remove polymerase chain reaction (PCR) artefacts. Further, post alignment processing steps such as IndelRealigner and BaseRecalibrator were performed to correct the mapping errors made by alignment tool to increase variant calling accuracy. Variant calling was performed using HaplotypeCaller. Finally, raw variants were annotated using ANNOVAR and the annotated variants were incorporated to the protein database to create custom variant protein database.

Proteomic searches were carried out using Proteome Discoverer 2.3 (Thermo Scientific, Bremen, Germany). Searches could also be carried out using freely available open source alternatives such as MaxQuant, MsFragger, MS-GF+, MyriMatch or OMSSA. Mass spectrometry raw files were obtained from the PRIDE archive (PXD008222) and searched against the customized variant protein database using SEQUEST-HT search algorithm. Trypsin and LysC were set as proteolytic enzymes with a maximum missed cleavage of one. Carbamidomethylation of cysteine was set as a fixed modification, and acetylation of protein N-terminus and oxidation of methionine were set as variable modifications with a minimum peptide length of seven amino acids. The precursor mass tolerance was fixed as 10 ppm, and 0.05 Da for fragment ions. Mass spectrometry data were searched against the decoy database with 1% false discovery rate cut-off at the peptide level.

We have provided a dataset for users to test the tool. The test dataset was taken from the study SRR7418758 archived on the NCBI Sequence Read Repository (SRA) and aligned using human reference genome version 19 (hg19). Using samtools view command, reads mapped to chromosome 22 were extracted and converted to FASTQ files (paired-end). We have also provided chromosome 22 nucleotide sequences from hg19 and corresponding variant information from dbSNP database.

Use cases

As a test case, we analyzed exome (Daemen *et al.*, 2013) and proteome datasets (Lawrence *et al.*, 2015) from breast cancer cell lines. We incorporated 12,429; 13,923; 12,386 and 11,600 non-synonymous SNPs from BT474 (accession number SRR925752), MDMAB157 (accession number SRR925788), MFM223 (accession number SRR925796) and HCC38 (accession number SRR925778) cells, respectively, to the protein database (Figure 2A). These non-synonymous SNPs were incorporated to the reference protein database (Human RefSeq release 93) to create customized variant protein database. Mass spectrometry-based raw files were searched against their respective custom variant protein database, which resulted in identification of 423, 408, 386 and 361 variant peptides from BT474, MDMAB157, MFM223 and HCC38, cell line datasets (Figure 2B). Interestingly, we observed mutant protein expression of Replication Timing Regulatory Factor 1 (RIF1) and Torsin-1A-interacting protein 1 (TOR1AIP1) across all four breast cancer cell lines. Mutant plectin (PLEC), marker Of Proliferation Ki-67 (MKI67), HEAT Repeat Containing 1 (HEATR1) and AHNAK nucleoprotein (AHNAK) were detected in three of the four breast cancer cell lines. These coding mutations have also been reported in other cancers The resultant variant peptide lists are available as *Underlying data* (Kasaragod *et al.*, 2020a).

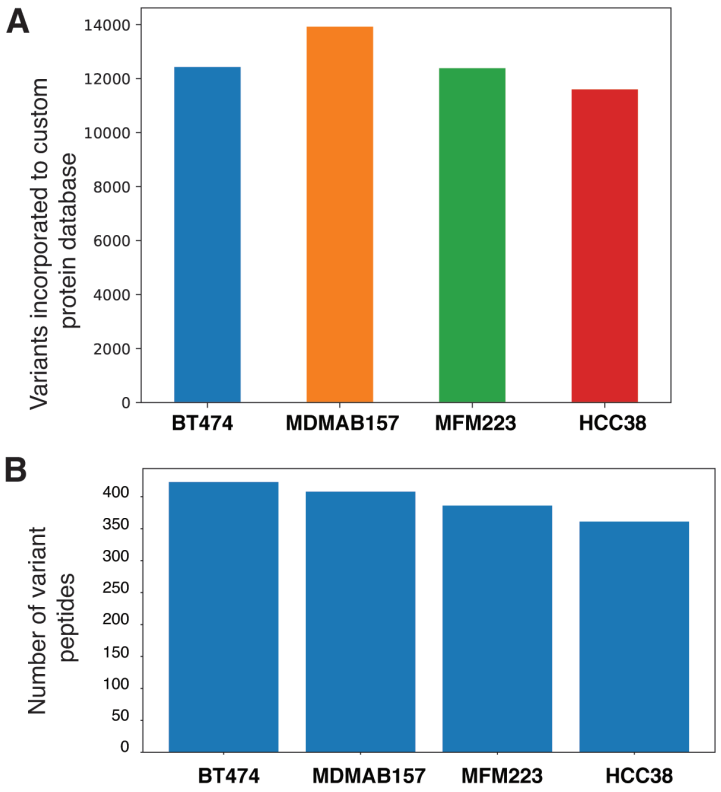

Figure 2. A) Number of nonsynonymous variants incorporated to protein database. **B**) Variant peptides identified across four cell lines (BT474, MDMAB157, MFM223 and HCC38) by searching mass spectrometry data against sample specific protein database.

The data generated from the tool demonstrates the usefulness and the ease of detection of variant peptides in an integrated omics analysis.

Conclusions

CusVarDB generates customized variant protein database from NGS datasets. To our knowledge, this is the first tool which uses Linux Bash Shell to execute the NGS tools on a Windows OS. The tool provides additional feature of dry-run, making the commands highly customizable. The variant protein database generated by this tool is highly compatible for use as a reference protein database for analysis of variants at the proteomic level. CusVarDB currently allows incorporation of non-synonymous mutations. It does not allow incorporation of protein variations resulting from indels and frame-shifts. We developed a flexible tool with the intension of updating it in the future.

Data availability

Source data

Whole exome data from Daemen *et al.* (2013) on BioProject, Accession number PRJNA210427

Mass spectrometry proteome data from Lawrence *et al.* (2015) on PRIDE, Accession number PXD008222: https://identifiers.org/pride.project:PXD008222

Test dataset on SRA, Accession number SRR7418758: https://identifiers.org/insdc.sra:SRR7418758

Underlying data

Zenodo: CusVarDB: A tool for building customized sample-specific variant protein database from Next-generation sequencing datasets. http://doi.org/10.5281/zenodo.4018694 (Kasaragod *et al.*, 2020a)

This project contains the following underlying data:

- Gowda_CusVarDB_Supplementary_table1.xlsx (This table contains the resultant variant peptides along with the wild-type peptides from BT474, MDMAB157, MFM223, and HCC38 datasets. Along with mutant peptides, this section also provides additional information such as peptide-spectrum match [PSM], Protein accession, cross-correlation value from the search [Xcorr] and retention time [RT])

- Gowda_CusVarDB_Supplementary_table2.xlsx (This table provides the complete details of the resultant peptides. Here the mutant and corresponding wild-type peptides are mentioned in different sheets. For a given mutant peptide its wild-type peptide and corresponding information can be mapped using the VLOOKUP function in Excel by keeping column A [Sl.No] as lookup parameter.)

- Gowda_CusVarDB_Supplementary_table3.xlsx (This table briefs about the variants which are already reported in other cancers.)

Software availability

Software available from: https://sourceforge.net/projects/cusvardb/ and http://bioinfo-tools.com/Downloads/CusVarDB/V1.0.0/

Source code available from: https://github.com/sandeepkasaragod/CusVarDB

Archived source code at time of publication: https://doi.org/10.5281/zenodo.3780645 (Kasaragod *et al.*, 2020b)

Acknowledgements

We thank Mr. Hitesh Ugaram Kore, Ph.D. student, Cancer Precision Medicine Group, QIMR Berghofer Medical Research Institute, Australia for his assistance with testing the tool and fixing the bugs which has greatly improved the performance of the tool.

References

Alfaro JA, Ignatchenko A, Ignatchenko V, *et al.*: **Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines.** *Genome Med.* 2017; **9**(1): 62.
PubMed Abstract | Publisher Full Text | Free Full Text

Apweiler R, Bairoch A, Wu CH, *et al.*: **UniProt: the Universal Protein knowledgebase.** *Nucleic Acids Res.* 2004; **32**(Database issue): D115–9.
PubMed Abstract | Publisher Full Text | Free Full Text

Chapman PB, Hauschild A, Robert C, *et al.*: **Improved survival with vemurafenib in melanoma with BRAF V600E mutation.** *N Engl J Med.* 2011; **364**(26): 2507–16.
PubMed Abstract | Publisher Full Text | Free Full Text

Daemen A, Griffith OL, Heiser LM, *et al.*: **Modeling precision treatment of breast cancer.** *Genome Biol.* 2013; **14**(10): R110.
PubMed Abstract | Publisher Full Text | Free Full Text

Ellis MJ, Gillette M, Carr SA, *et al.*: **Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium.** *Cancer Discov.* 2013; **3**(10): 1108–12.
PubMed Abstract | Publisher Full Text | Free Full Text

Forbes SA, Tang G, Bindal N, *et al.*: **COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer.** *Nucleic Acids Res.* 2010; **38**(Database issue): D652–D657.
PubMed Abstract | Publisher Full Text | Free Full Text

Forbes SA, Beare D, Gunasekaran P, *et al.*: **COSMIC: exploring the world's knowledge of somatic mutations in human cancer.** *Nucleic Acids Res.* 2015; **43**(Database issue): D805–11.
PubMed Abstract | Publisher Full Text | Free Full Text

Kasaragod S, Mohanty V, Tyagi A, *et al.*: **CusVarDB: A tool for building customized sample-specific variant protein database from Next-generation sequencing datasets [Data set].** *Zenodo.* 2020a.
http://www.doi.org/10.5281/zenodo.4018694

Kasaragod S, Mohanty V, Tyagi A, *et al.*: **CusVarDB: A tool for building customized sample-specific variant protein database from Next-generation sequencing datasets: First release (Version 1.0.0).** *Zenodo.* 2020b.
http://www.doi.org/10.5281/zenodo.3780645

Kelkar DS, Provost E, Chaerkady R, *et al.*: **Annotation of the zebrafish genome through an integrated transcriptomic and proteomic analysis.** *Mol Cell Proteomics.* 2014; **13**(11): 3184–98.
PubMed Abstract | Publisher Full Text | Free Full Text

Kim D, Langmead B, Salzberg SL, *et al.*: **HISAT: a fast spliced aligner with low memory requirements.** *Nat Methods.* 2015; **12**(4): 357–60.
PubMed Abstract | Publisher Full Text | Free Full Text

Kreiter S, Vormehr M, van de Roemer N, *et al.*: **Mutant MHC class II epitopes drive therapeutic immune responses to cancer.** *Nature.* 2015; **520**(7549): 692–6.
PubMed Abstract | Publisher Full Text | Free Full Text

Lawrence RT, Perez EM, Hernández D, *et al.*: **The proteomic landscape of triple-negative breast cancer.** *Cell Rep.* 2015; **11**(4): 630–44.
PubMed Abstract | Publisher Full Text | Free Full Text

Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2009; **25**(14): 1754–60.
PubMed Abstract | Publisher Full Text | Free Full Text

Mathivanan S, Ji H, Tauro BJ, *et al.*: **Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry.** *J Proteomics.* 2012; **76**: 141–9.
PubMed Abstract | Publisher Full Text

McKenna A, Hanna M, Banks E, *et al.*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res.* 2010; **20**(9): 1297–303.
PubMed Abstract | Publisher Full Text | Free Full Text

Menschaert G, Fenyö D: **Proteogenomics from a bioinformatics angle: A growing field.** *Mass Spectrom Rev.* 2017; **36**(5): 584–599.
PubMed Abstract | Publisher Full Text | Free Full Text

Morgensztern D, Politi K, Herbst RS: ***EGFR* Mutations in Non-Small-Cell Lung Cancer: Find, Divide, and Conquer.** *JAMA Oncol.* 2015; **1**(2): 146–8.
PubMed Abstract | Publisher Full Text

Nagaraj SH, Waddell N, Madugundu AK, *et al.*: **PGTools: a software suite for proteogenomic data analysis and visualization.** *J Proteome Res.* 2015; **14**(5): 2255–66.
PubMed Abstract | Publisher Full Text

Nesvizhskii AI: **Proteogenomics: concepts, applications and computational strategies.** *Nat Methods.* 2014; **11**(11): 1114–25.
PubMed Abstract | Publisher Full Text | Free Full Text

Ruggles KV, Fenyö D: **Next Generation Sequencing Data and Proteogenomics.** *Adv Exp Med Biol.* 2016; **926**: 11–19.
PubMed Abstract | Publisher Full Text

Sherry ST, Ward MH, Kholodov M, *et al.*: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res.* 2001; **29**(1): 308–311.
PubMed Abstract | Publisher Full Text | Free Full Text

Subbannayya Y, Mir SA, Renuse S, *et al.*: **Identification of differentially expressed serum proteins in gastric adenocarcinoma.** *J Proteomics.* 2015; **127**(Pt A): 80–8.
PubMed Abstract | Publisher Full Text | Free Full Text

Subbannayya Y, Pinto SM, Gowda H, *et al.*: **Proteogenomics for understanding oncology: recent advances and future prospects.** *Expert Rev Proteomics.* 2016; **13**(3): 297–308.
PubMed Abstract | Publisher Full Text

Tate JG, Bamford S, Jubb HC, *et al.*: **COSMIC: the Catalogue Of Somatic Mutations In Cancer.** *Nucleic Acids Res.* 2019; **47**(D1): D941–D947.
PubMed Abstract | Publisher Full Text | Free Full Text

Tomczak K, Czerwińska P, Wiznerowicz M, *et al.*: **The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.** *Contemp Oncol (Pozn).* 2015; **19**(1A): A68–77.
PubMed Abstract | Publisher Full Text | Free Full Text

Wang K, Li M, Hakonarson H, *et al.*: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Res.* 2010; **38**(16): e164.
PubMed Abstract | Publisher Full Text | Free Full Text

Wang X, Zhang B: ***customProDB*: an R package to generate customized protein databases from RNA-Seq data for proteomics search.** *Bioinformatics.* 2013; **29**(24): 3235–7.
PubMed Abstract | Publisher Full Text | Free Full Text

Yeom J, Kabir MH, Lim B, *et al.*: **A proteogenomic approach for protein-level evidence of genomic variants in cancer cells.** *Sci Rep.* 2016; **6**: 35305.
PubMed Abstract | Publisher Full Text | Free Full Text

Zhang J, Baran J, Cros A, *et al.*: **International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data.** *Database (Oxford).* 2011; **2011**: bar026.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔ ✘

Version 2

Reviewer Report 02 December 2020

https://doi.org/10.5256/f1000research.30424.r74962

✘ **Jorge Duitama**
Systems and Computing Department, Universidad de Los Andes, Bogotá, Colombia
Daniel Mahecha
Systems and Computing Department, Universidad de Los Andes, Bogotá, Colombia

This manuscript by Kasagarod *et al.*, describes a GUI-based tool that executes a pipeline of Linux based NGS tools on a Windows OS, to create a variant protein database specific for a sample. This variant protein database can be further used to identify mutant proteins on mass spectrometry data from proteomics studies.

The manuscript claims that CusVarDB is a useful tool on the onco proteo-genomics and different from other tools that require BED or VCF files as input, while it works with FASTA and FASTQ files as an input. As a use case, it provides examples of its use in genomics and proteomics data from four cancer cell lines.

Unfortunately, in my opinion there are several technical and conceptual issues related to this work:

1. Regarding decisions on technological architecture, it is not clear which is the advantage of building an application on technology that only runs on windows to run programs that run mostly on Linux. As a consequence, I could not install and execute the software successfully. The manuscript claims that VCF files are difficult to generate because: “Executing the preprocessing steps requires knowledge of computation and also requires multiple tools to generate the output.”. However, there are different commercial and open source solutions to analyze sequencing reads and generate VCF files. As an example, my research group maintains the Next-Generation Sequencing Experience Platform, which is able to make this analysis with a graphical interface on any operative system. Other well known web based solutions are Galaxy or CyVerse. For variants interpretation, functionalities of CusVarDB should be compared with those offered by well known platforms such as VEP.

2. One common issue with wrapping tools for other software such as GATK and bwa is that the software tools are evolving and hence the wrapping software quickly becomes outdated.

Regarding this issue, the manuscript does not describe if a maintenance plan is planned for this tool, including the way different enclosed tools will be updated.

3. In the section “use cases”, the manuscript describes one single use case in which alternative proteins are identified for four samples. It is not clear which parameters were used to run the mapping and variant calling tools, and especially to filter variants. It is not clear if the reported variants are the complete set of variants after this process. The false positive rate of the proposed analysis pipeline should be estimated and compared with other pipelines to identify variants in cancer samples. Moreover, it is not clear if tumor and normal tissue was considered to increase the reliability of the reported variants. If the section is called “use cases”, then more use cases should be included.

4. From the description of the use case, it looks like both genomics and proteomics data is required to operate CusVarDB. Obtaining both data types can be too expensive for real clinical settings. The manuscript should clearly state if both data types are required or what can be done if only DNA sequencing data is available.

5. The conclusions state that the variant protein database generated by this tool is highly compatible for use as a reference protein database for analysis of variants at the proteomic level. However, from the manuscript and the manual it is not clear what format the output variant protein database has and what information does it include, and whether the output format is compatible with commonly used proteomics tools. If the database is relational, an entity relationship model should be provided and instructions on how to install and use the database should be provided and maintained as part of the documentation.

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: My research group maintains the open source software tool Next-Generation Sequencing Experience Platform, which currently has functionalities to analyze DNA

sequencing reads and generate VCF files. NGSEP also has a graphical interface and can run on any operative system.

Reviewer Expertise: Bioinformatics, applied algorithms, software development, population genomics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 23 November 2020

https://doi.org/10.5256/f1000research.30424.r74866

Suresh Mathivanan
Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Vic, Australia

All issues addressed.

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cancer proteomics and bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 16 November 2020

https://doi.org/10.5256/f1000research.30424.r74867

Richard Kumaran Kandasamy
Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

The authors have addressed the minor issues and I recommend accepting the article for indexing.

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research expertise include proteomics, proteogenomics, bioinformatics and systems immunology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 26 August 2020

https://doi.org/10.5256/f1000research.25627.r63338

?

Richard Kumaran Kandasamy
Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

The manuscript by Kasaragod *et al.,* describes a software application for developing sample-specific protein sequence database that takes into account the protein-coding variants, for interpretation of mass spectrometry data to see what portion of the variants are seen at the protein expression level. This is an important aspect in exemplifying the use of genomic data on proteomics datasets associated with cancer samples.

The software is made available for download and enough documentation has been provided for potential users.

The authors use this data set to identify protein-level evidence for expression of variants in four different cancel cell lines.

The manuscript can be accepted for indexing. I have one important point that I would like the authors to address:

While the findings are potentially interesting, it would be nice if the authors pick a few important examples specific to these four cell lines.

This would give the users a better appreciation of the biological insights given by this application.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the

findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research expertise include proteomics, proteogenomics, bioinformatics and systems immunology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 23 July 2020

https://doi.org/10.5256/f1000research.25627.r63336

?

Suresh Mathivanan
Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Vic, Australia

This is a manuscript that describes an important tool.

1. The authors may need to discuss other proteogenomic tools similar to this and provide the advantages. If such tools don't exist, perhaps make a strong case. Zhang lab tools can also be discussed with references.

2. Can the authors also discuss and refer previous attempts on onco-proteogenomics? One of the first paper on onco-proteogenomics came in 2012: Mathivanan, S., Ji, H., Tauro, B.J., Chen, Y.S. and Simpson, R.J. (2012) Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry. ***Journal of Proteomics***. 76, 141-9[1].

3. Does CurVardb provide only non-synonymous point mutations? What about frame shit mutations? Is that also provided or perhaps will be added in next version.

References

1. Mathivanan S, Ji H, Tauro BJ, Chen YS, et al.: Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry.*J Proteomics*. 2012; **76 Spec No.**: 141-9 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cancer proteomics and bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.