ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(3): 214-218 (2022)


BIOINFORMATION
Discovery at the interface of physical and biological sciences

OPEN ACCESS GOLD

BIOMEDICAL INFORMATICS

www.bioinformation.net
**Volume 18(3)**

**Research Article**

Received February 25, 2022; Revised March 24, 2022; Accepted March 31, 2022, Published March 31, 2022

DOI: 10.6026/97320630018214

**Declaration on Publication Ethics:**
The author's state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

**Declaration on official E-mail:**
The corresponding author declares that lifetime official e-mail from their institution is not available for all authors



**Comments from readers:**
Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article without open access charges. Comments should be concise, coherent and critical in less than 1000 words.

**Edited by P Kangueane**
**Citation**: Kasaragod *et al.* Bioinformation 18(3): 214-218 (2022)

# A computational workflow for predicting cancer neo-antigens

**Sandeep Kasaragod[1], Chinmaya Narayana Kotimoole[1], Sumrati Gurtoo[1], Thottethodi Subrahmanya Keshava Prasad[1], Harsha Gowda[1, 2,*] & Prashant Kumar Modi[1,*]**

[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; *Corresponding authors - Harsha Gowda & Prashant Kumar Modi

**Author contacts:**
Sandeep Kasaragod - E-mail: sandeepk@yenepoya.edu.in
Chinmaya Narayana Kotimoole - E-mail: chinmaya_k@yenepoya.edu.in
Sumrati Gurtoo - E-mail: sumrati@yenepoya.edu.in
Thottethodi Subrahmanya Keshava Prasad - E-mail: keshav@yenepoya.edu.in
Harsha Gowda - E-mail: harshahc@gmail.com
Prashant Kumar Modi - E-mail: prashantmodi@yenepoya.edu.in

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(3): 214-218 (2022)


**Abstract:**
Neo-antigens presented on cell surface play a pivotal role in the success of immunotherapies. Peptides derived from mutant proteins are thought to be the primary source of neo-antigens presented on the surface of cancer cells. Mutation data from cancer genome sequencing is often used to predict cancer neo-antigens. However, this strategy is associated with significant false positives as many coding mutations may not be expressed at the protein level. Hence, we describe a computational workflow to integrate genomic and proteomic data to predictpotential neo-antigens.

**Keywords:** Neoantigens, proteogenomics, cancer proteogenomics, multi-omics.

**Background:**
Cancer is the second leading cause of morbidity and mortality worldwide. As per a survey conducted by Cancer Research UK in 2018, there are 17 million new cases worldwide, and cancer-related death has risen to 9.6 million **[1]**. Cancer is heterogeneous. As cancer cells proliferate, they create tumors with genetically heterogeneous cells making it challenging to treat **[2], [3]**. Chemotherapy and targeted therapies are effective only in select cancer types. The advent of immunotherapy has revolutionized cancer treatment in the last decade **[4]**. For example, checkpoint blockade-based treatments have significantly improved cancer survival **[5], [6]**. Other immunotherapy treatments such as adoptive cell transfer therapy and small molecule inhibitors are widely used to treat various cancer types **[7], [8]**. The success of immunotherapy strategies is dependent on presentation of neo-antigens on cancer cell surface. These neo-antigens are presented by MHC complex on the cell surface, which are recognized by T cells **[9]**. Cancer genome sequencing has revealed thousands of mutations associated with various cancers **[10], [11], [12]**. Mutation data from cancer genome sequencing is often used to predict cancer neo-antigens. However, this approach can result in false positives as many mutations may not be expressed at the protein level **[13], [14]**. We previously developed a computational workflow to integrate genomic and proteomic data to identify coding variations **[15]**. Therefore, we describe a workflow to predict cancer neo-antigens.

**Methods:**
***Genomics and proteomics data analysis:***
Genomics datasets **[18], [19], [20], [21]** were analyzed using the CusVarDB tool. A custom protein database was developed by incorporating coding mutations that was used to carry out proteomics searches. Proteomeics searches were carried out using Proteome Discoverer 2.3 (Thermo Fisher Scientific, Bremen, and Germany). The cancer type-specific raw files were searched against the corresponding customized variant protein database using Sequest-HT search engine **[22]**. The search parameters were set as reported in the original studies **[23], [24], [25]**. False discovery rate (FDR) was set to 1% at PSM, peptide, and protein levels. (**Figure 1-a**) describes the proteogenomics workflow used in our study.

**Workflow development:**
The workflow is created using snakemake version 6.12.3 **[16]**. All the supporting scripts for the workflow are written in Python 3.9. This snakemake workflow requires variant annotation results from ANNOVAR **[17]** and proteomics search results. Proteomics data is searched against a custom database that incorporates coding mutations identified in genomics data. Peptides that do not have sequence variations are filtered by mathching sequences to reference protein sequence database. Variant peptides are assigned unique accessions and are provided as a tab-delimited or comma-separated file that can be queried using SQL.

**Prediction of neoantigens:**
Neoantigen prediction was performed using offline version of net MHCpan 4.1 **[26]**. We kept a window of ± 15 amino acid sequence from the variant amino acid. It created an overall sequence length of 30 amino acids. These sequences were stored in FASTA format to perform predictions. HLA allele information for corresponding cell lines was taken from the literature **[27]**, **[28]**, and Expasy (https://web.expasy.org/cellosaurus/).

**Code availability:**
Workflow is available at Github (https://github.com/sandeepkasaragod/Proteogenomics_workflow)

**Table 1**: List of proteomics and genomics datasets used in the present study.

| Sl. No | Cell lines | Cancer type | Genomics | Proteomics |
|---|---|---|---|---|
| 1 | BT20 | TNBC | SRR925751 | PXD008222 |
| 2 | BT474 | TNBC | SRR925752 | PXD008222 |
| 3 | BT549 | TNBC | SRR925754 | PXD008222 |
| 4 | HCC1143 | TNBC | SRR925765 | PXD008222 |
| 5 | HCC1806 | TNBC | SRR925771 | PXD008222 |
| 6 | HCC1937 | TNBC | SRR925772 | PXD008222 |
| 7 | HCC38 | TNBC | SRR925778 | PXD008222 |
| 8 | HCC70 | TNBC | SRR925780 | PXD005295 |
| 9 | MDAMB157 | TNBC | SRR925788 | PXD008222 |
| 10 | MDAMB231 | TNBC | SRR925790 | PXD008222 |
| 11 | MDAMB468 | TNBC | SRR925794 | PXD008222 |
| 12 | SKBR3 | TNBC | SRR925800 | PXD008222 |
| 13 | SUM229 | TNBC | SRR925807 | PXD005295 |
| 14 | T47D | TNBC | SRR925811 | PXD005390 |
| 15 | COLO-205 | Colon | SRR7366613 | PXD005946 |
| 16 | HCT-116 | Colon | SRR7366622 | PXD005946 |
| 17 | HCT-15 | Colon | SRR7366619 | PXD005946 |
| 18 | HT29 | Colon | SRR1232556 | PXD005946 |
| 19 | KM-12 | Colon | SRR7366594 | PXD005946 |
| 20 | SW-620 | Colon | SRR7366632 | PXD005946 |
| 21 | OVCAR-3 | Ovarian | SRR7366635 | PXD005946 |
| 22 | OVCAR-4 | Ovarian | SRR8657373 | PXD005946 |
| 23 | OVCAR-5 | Ovarian | SRR7366581 | PXD005946 |
| 24 | OVCAR-8 | Ovarian | SRR7366617 | PXD005946 |
| 25 | SK-OV-3 | Ovarian | SRR8657598 | PXD005946 |

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(3): 214-218 (2022)


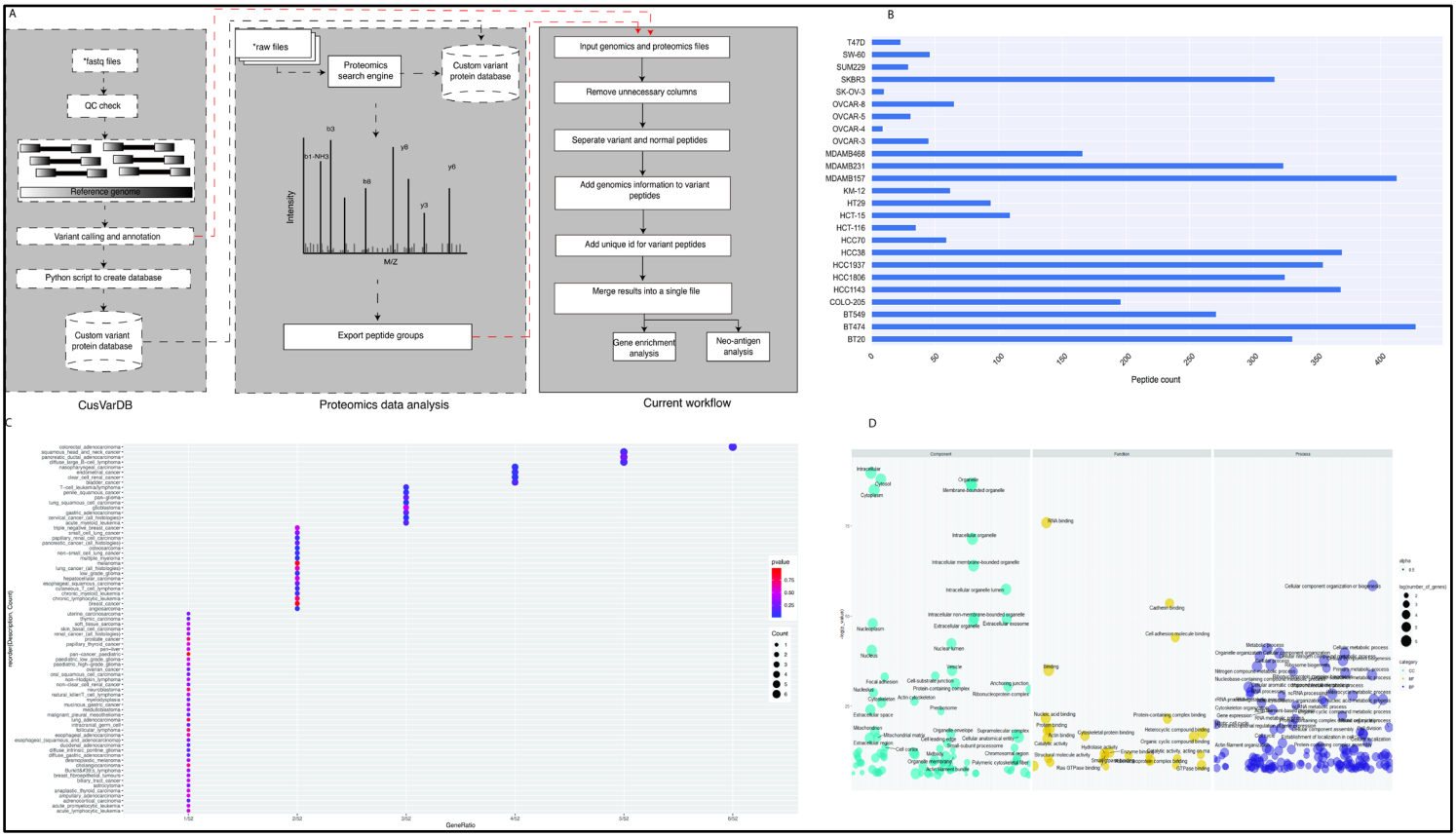


**Figure 1:** Schematic representation of the execution workflow (A). Number of variant peptides identified in each sample (B). Cluster profiler analysis on neoantigen peptides reveals their involvement in various cancer types (C). Gene enrichment analysis results in identification of key cellular components and processes.

**Results and Discussion:**

Our study was carried out using datasets from twenty-five cancer cell lines. Fourteen datasets were from TNBC, five from ovarian cancer, and six from colon cancer. The exome datasets were subjected to variant analysis. We identified 125,687 unique non-synonymous variants from 25 datasets. Non-synonymous variants were incorporated into protein sequences from RefSeq database to create a custom variant protein database to perform proteomics data analysis. Proteomics searches identified 231,886 unique peptides from 25 datasets. Overall, we identified 4,673 variant peptides corresponding to 1,249 genes (**Figure 1b**). We also identified 1,297 variants that correspond to 295 genes reported in COSMIC **[29]** and ClinVar **[30]**. These include well-known cancer-related genes such as TP53, KRAS, EGFR, AARS, ACTN4, SAMHD1, and many other genes. Enrichment analysis showed genes involved in important functions including cell division, cellular metabolic process, cellular localization, and other events. (**Figure 1d**). The same set of peptides was run on Net MHCpan for neoantigen prediction. We identified a total of 5,865 neoantigens with strong binding affinity. We also identified corresponding wild type peptides for 1,915 variant peptides. We predicted binding affinity for corresponding wild type peptides. Of these, 707 variant peptides had a stronger binding affinity when compared to their wild type (supplementary available at GitHub). Cluster Profiler analysis of these variant proteins showed their involvement in various cancers including breast cancer, colorectal adenocarcinoma and cervical cancer (**Figure 1c**). In this study, we utilized the power of multi-omics datasets to predict potential cancer neo-antigens. We developed a proteogenomics data analysis workflow using snakemake package. The workflow is highly customizable and efficiently executed in a condo environment.

**Conclusions:**

Identification of cancer neoantigens is important to develop effective immunotherapy strategies. Predicting cancer neoantigens using genomic data alone can result in several false positives. In this study, we present an integrated approach combining genomic and proteomic data to predict cancer neoantigens. This computational workflow can be used on any dataset where both genomic and proteomic data is available.

**Conflict of interest:**

None declared

**Acknowledgment:**

SK was a recipient of the Indian Council of Medical Research (ICMR) Senior Research Fellow (SRF) application number

[ISRM/11(27)/2017]. SG is a recipient of the Indian Council of Medical Research (ICMR) Senior Research Fellow (SRF).

**References:**

**[1]** Bray F *et al*. *CA Cancer J Clin*. 2018 **68**:394. [PMID: 30207593]
**[2]** Bedard PL *et al*. *Nature*. 2013 **501**:355. [PMID: 24048068]
**[3]** Pucci C *et al*. *Ecancermedicalscience*. 2019 **13**:961. [PMID: 31537986]
**[4]** Han X-J *et al*. *Front Cell Dev Biol*. 2020 **8**:728. [PMID: 32850843]
**[5]** Sharma P & Allison JP. *Science* 2015 **348**:56. [PMID: 25838373]
**[6]** Brahmer JR *et al*. *N Engl J* Med. 2012 **366**:2455. [PMID: 22658128]
**[7]** Efremova M *et al*. *Front Immunol*. 2017 **8**:1679. [PMID: 29234329]
**[8]** Wu W *et al*. *Oncol Lett*. 2020 **20**:123 [PMID: 32934692]
**[9]** Roudko V *et al*. *Front Immunol*. 2020 **11**:27 [PMID: 32117226]
**[10]** ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium *Nature*. 2020 **578**:82. [PMID: 32025007]
**[11]** Alexandrov LB *et al*. *Nature*. 2020 **578**:94. [PMID: 32025018]
**[12]** Yuan Y *et al*. *Nat Genet*. 2020 **52**:342. [PMID: 32024997]
**[13]** Geiger T *et al*. *PLoS Genet*. 2010 **6**:e1001090. [PMID: 20824076]
**[14]** Zhang B *et al*. *Nature*. 2014 **513**:382. [PMID: 25043054]
**[15]** Kasaragod S *et al*. *F1000Res*. 2020 **9**:344. [PMID: 33274046]
**[16]** Molder F *et al*. *F1000Res*.2021 **10**:33 [PMID: 34035898]
**[17]** Wang K *et al*. *Nucleic Acids Res*. 2010 **38:**e164. [PMID: 20601685]
**[18]** Abaan OD *et al*. *Cancer Res*. 2013 **73**:4372. [PMID: 23856246]
**[19]** Daemen A *et al*. *Genome Biol*. 2013 **14**:R110. [PMID: 24176112]
**[20]** Barretina J *et al*. *Nature*. 2012 **483**:603. [PMID: 22460905]
**[21]** Ghandi M *et al*. *Nature*. 2019 **569**:503. [PMID: 31068700]
**[22]** Eng JK *et al*. *J Proteome Res*. 2008 **7:**4598. [PMID: 18774840]
**[23]** Lawrence RT *et al*. *Cell Rep*. 2015 **11**:990. [PMID: 28843283]
**[24]** Yen TY *et al*. *J Proteome Res*. 2017 **16**:1391. [PMID: 28287265]
**[25]** Gholami AM *et al*. *Cell Rep*. 2013 **4**:609. [PMID: 23933261]
**[26]** Reynisson B *et al*. *Nucleic Acids Res*. 2020 **48**:W449. [PMID: 32406916]
**[27]** Adams S *et al*. *J Transl Med*. 2005 **3**:11 [PMID: 15748285]
**[28]** Robinson J *et al*. *Nucleic Acids Res*. 2020 **48:**D948. [PMID: 31667505]
**[29]** Tate JG *et al*. *Nucleic Acids Res*. 2019 **47**:D941. [PMID: 30371878]
**[30]** Landrum MJ *et al*. *Nucleic Acids Res*. 2018 **46**:D1062. [PMID: 29165669]

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(3): 214-218 (2022)