Guide d'Annotation d'articles décrivant des workflows d'analyse bioinformatique

Clémence Sebe

Juillet 2022 - V3

Sommaire

1 Introduction

Ce guide d'annotation est le mode d'emploi pour annoter des articles décrivant des workflows d'analyse bioinformatique.

L'annotation des articles a été réalisée à l'aide de l'outil BRAT. La figure 4 représente l'annotation d'un mot sous Brat et toutes les différentes catégories existantes.

Pour annoter les différents articles, j'ai fait le choix d'annoter toutes les occurences d'un élément (càd dès que le mot est mentionné, je l'annote) et pour chaque élément de l'annoter sous toutes ses formes (singulier, pluriel, majuscule, minuscule, conjugué ...). D'autre part, lors de l'annotation, j'ai choisi de ne pas retenir les *"premiers"* pronoms, déterminants, articles et mots de liaison. (*exemple : pour "the Ramachandran plot", on ne sélectionnera que "Ramachandran plot"*).

Lors de l'annotation, les entités imbriquées et discontinues sont accéptées.

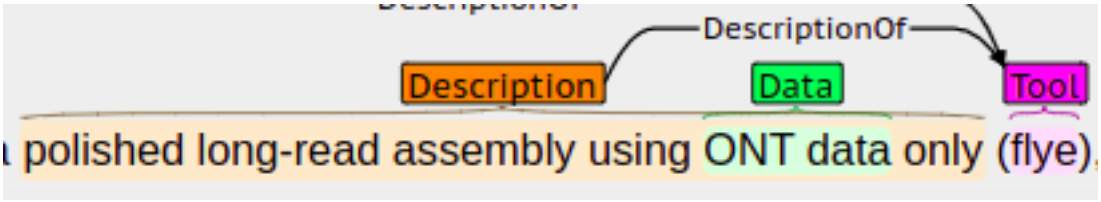


Figure 1: Entité imbriquée

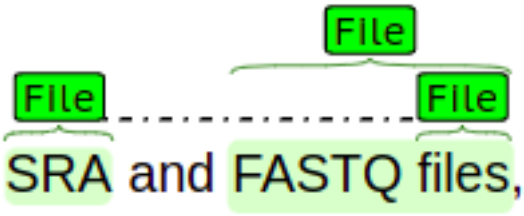


Figure 2: Entité discontinue : deux entités distinctes : SRA files *et* FASTQ files

2 Catégories

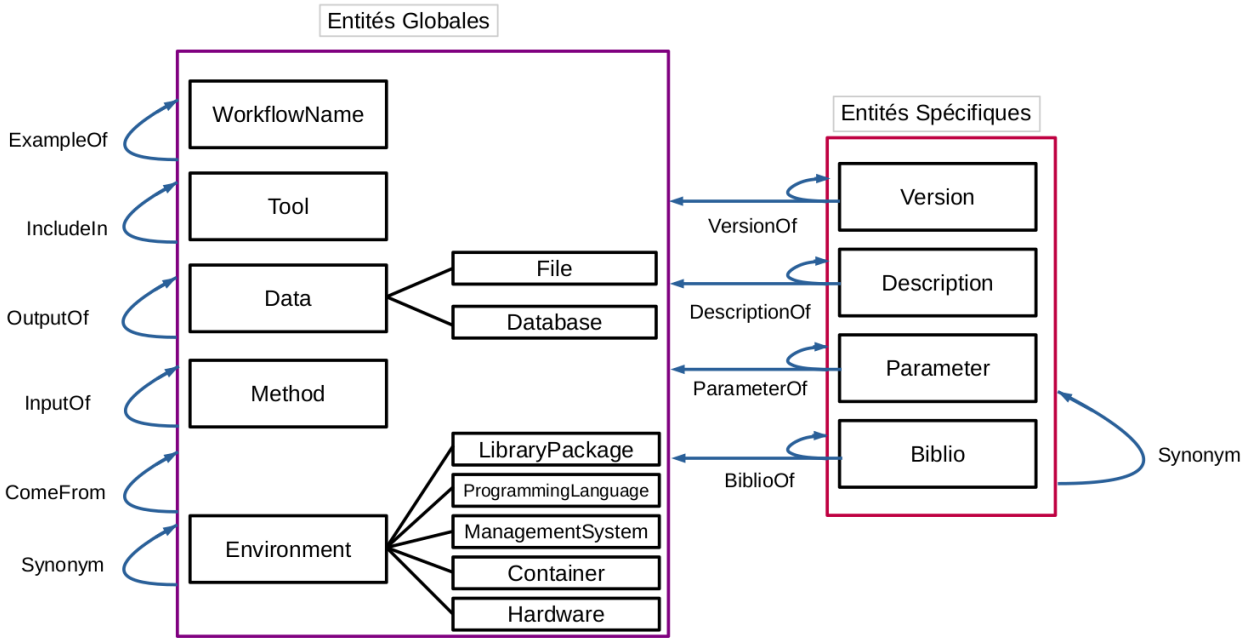


Figure 3: Organisation des entités et des relations

Les informations présentes dans les articles peuvent être réparties en cinq grandes catégories :

- Une catégorie spécifique pour **le nom du workflow** décrit dans l'article,
- Une catégorie pour les **Outils bioinformatiques**,
- Une catégorie pour les **données**. Dans certains cas, des détails supplémentaires peuvent être ajoutés comme le type de fichiers ou le nom d'une base de données utilisée,

- Une catégorie **Method** aux méthodes provenant d'une library ou d'un package déjà programmé dans un langage de programmation et non à une étape par un outil bioinformatique
- Une catégorie sur **l'environnement de programmation du workflow**, c'est à dire le système de gestion du workflow, la possibilité de le lancer dans un container, le langage des scripts utilisés dans le workflow et les librairies, packages utilisés.

Chaque élément peut être complété par des informations plus spécifiques tels la version utilisée, des paramètres particuliers, une description et une bibliographie associés.

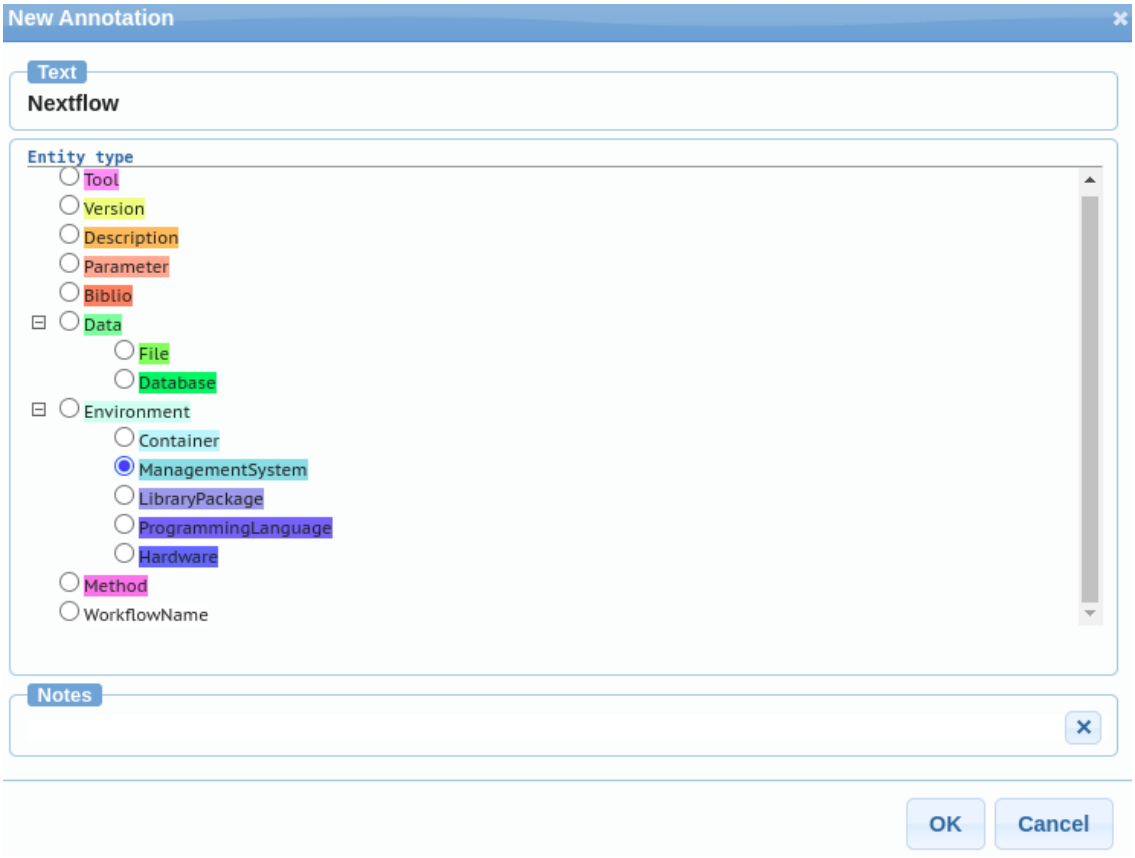


Figure 4: Interface des entités dans BRAT (code couleur)

2.1 Entités globales

2.1.1 Workflow

Cette entité correspond au **nom du workflow**. On le retrouve majoritairement dans le titre de l'article. Quelques exemples :

1. **GEMmaker**: process massive RNA-seq datasets on heterogeneous computational infrastructure
2. **SWAAT** Bioinformatics Workflow for Protein Structure-Based Annotation of ADME Gene Variants
3. Large-scale quality assessment of prokaryotic genomes with **metashot/prok-quality**
4. **nextNEOpi**: a comprehensive pipeline for computational neoantigen prediction

2.1.2 Tool

On annote sous l'attribut Tool tous les outils bioinformatiques mentionnés dans l'article. Sous cette entité, on annote seulement les **noms d'outils** et non des morceaux de phrases faisant références à des outils.

1. There are four primary paths for gene expression quantification within GEMmaker: STAR, HISAT2, Salmon and kallisto.
2. Fastp v0.20.1 (Chen et al. 2018) was used to trim adapter and low-quality Illumina sequences. We then constructed three types of assemblies: a polished long-read assembly using ONT data only (flye), one with short-read correction of the ONT long-read assembly (pilon) and one that first assembles short reads and scaffolds the assembly with long reads.
3. Then, we used MODELLER [26] homology modeling software to predict the 3D coordinates of missing segments and atoms.
4. GUNC v1.0.1. SCG-based tools like CheckM can have very low sensitivity towards contamination by fragments from unrelated organisms (non-redundant contamination).[6] In order to circumvent this problem, the recent GUNC[14] tool was added to the pipeline. GUNC quantifies the lineage homogeneity of contigs with respect to the full gene complement, accurately detecting chimerism induced by both redundant and non-redundant contamination.
5. MiXCR is used to predict B-cell receptor and T-cell receptor (BCR and TCR) repertoires (Bolotin et al., 2015).

2.1.3 Data

Dans les articles, nous trouvons aussi des références sur les données manipulées par les workflows. Certaines données sont globales et d'autres sont décrites plus spécifiquement.

1. Fastp v0.20.1 (Chen et al. 2018) was used to trim adapter and low-quality Illumina sequences.
2. Reference Illumina assemblies were generated with the pipeline Shovill v1.1.0 (https://github.com/tseemann/shovill)
3. Training and validation sets for the classifiers were split into 60% training and 40% validation data.
4. nextNEOpi takes as input raw WES or WGS data from matched tumor-normal samples and, optionally, bulk-tumor RNA-seq data (Fig. 1 andSupplementary Fig. S1).

Dans certains cas, les articles sont plus précis sur le "type" de données :

File l'article précise le type de fichier (souvent présence du mot file ou d'un synonyme de ce mot)

1. To ensure storage requirements are not exceeded, GEMmaker moves input FASTQ files between three folders: "stage", "processing" and "done".
2. As the workflow progresses for each sample, GEMmaker will cleanup unwanted intermediate files.
3. Users can keep downloaded SRA and FASTQ files, trimmed FASTQ files, SAM and BAM alignment files, and kallisto and Salmon pseudoalignment files.
4. Then, stereochemical quality was verified by establishing the Ramachandran plot.
5. The core functionality of SWAAT includes the main workflow (Figure 1B) that can process a list of variants, annotate them according to sequence-based and biophysical-based properties, and return a detailed report in HTML and CSV format.

Database les données proviennent-elles d'une base de données ?

1. The following is an example command-line for execution of GEMmaker on a local machine using Singularity (for containerization), quantification using Salmon, and a file containing a list of SRA run IDs for Arabidopsis thaliana Illumina datasets:

2. The 32 core ADME genes according to PharmaADME [5,25] (Supplementary Materials: data file 1) and the 23 genes from PharmVar (www.pharmvar.org, accessed on the 23 January 2020) were screened to identify available 3D structures in the Protein Data Bank (PDB).

2.1.4 Environment

La partie Environnement regroupe toutes les informations relatives aux workflows en général:

1. By leveraging conda environment and singularity container capabilities, the installation demands for nextNEOpi are kept on a minimal level facilitating its usage by users with limited bioinformatics expertise.

Container Les auteurs ont-ils donné des informations sur sa containérisation càd le workflow peut-il être lancé à l'aide d'un container comprenant dejà tous les éléments nécessaires pour l'exécuter (outils, packages et langages utilisés).

1. Alternatively, the Singularity container engine can be used in place of Docker.

ProgrammingLanguage certains workflows peuvent avoir des commandes écrites dans un autre langage de programmation que le Bash.

1. GEMmaker uses Nextflow and is a combination of Groovy scripts for interfacing with Nextflow, Python scripts for wrangling intermediate data, and Bash scripts for execution of each software tool in the workflow.

2. Results were summarized using the bdskytools package in R,

ManagementSystem système de gestion de workflow utilisé (Nextflow, Snakemake, ...).

1. SWAAT was built and tested on version 20.10.0 of Nextflow [41].

LibraryPackage les librairies et les packages qui sont implémentées dans les différents langages.

1. The user must have version 3 of Python with installed modules Pandas, Numpy Matplotlib, Bokeh, and biopython [42].

2. The predictive model was built using the Python library scikit-learn [33].

3. The scripts make use of NumPy,[17] Pandas and scikit-learn libraries.

4. Results were summarized using the bdskytools package in R,

Hardware Des informations sur le matériel informatique utilisé pour pouvoir faire tourner le workflow.

1. The workflow can run in a workstation with 16 GB of RAM

2.1.5 Method

Method correspond à une étape d'un workflow réalisé par une autre méthode, qui est différente d'un outil bioinformatique.

1. We implemented the feature extraction and random forest design from Sanderson and colleagues (2020) who use the RandomForest classifier from scikit-learn (Pedregosa et al. 2011) with default hyperparameter settings and feature extraction with pysamstats.
2. PRODRES pipline (https://github.com/ElofssonLab/PRODRES, accessed on the 2 February 2022) is required to calculate the PSSMs of the protein sequences to annotate during the preparation process used by the auxiliary workflow.
3. Unlike other methods, ENCoM accounts for the diversity of amino acids thanks to a specific set of coarse grain parameters, thus allowing to study the effect of mutations. The eigenvectors calculated by ENCoM are used to compute △△Svib.

2.2 Entités plus spécifiques

Chaque entité citée dans la section précédente peut posséder des informations complémentaires et ainsi apporter de nouvelles informations.

2.2.1 Version

La version utilisée d'un logiciel, d'un package, d'une librairie et du workflow peut être spécifiée dans la publication.

1. Fastp v0.20.1 (Chen et al. 2018) was used to trim adapter and low-quality Illumina sequences.
2. Sequencing runs were managed on two MinIONs and monitored in MinKNOW > v20.3.1.
3. The workflow includes three Python3 custom scripts
4. A database for the 36 ADME genes was prepared and made available in the main repository of SWAAT (https://github.com/hothman/SWAAT/tree/master/database, accessed on the 2 February 2022)

2.2.2 Parameter

Certaines entités peuvent être accompagnées de paramètres pour lancer les différents outils/programmes comme par exemple des modeles de calculs.

1. The presence of 5S, 23S and 16S rRNA genes is predicted by the BAsic Rapid Ribosomal RNA Predictor (Barrnap) using Hidden Markov models (HMM).
2. tRNA genes are searched using tRNAscan-SE,[15] using bacteria and archaea covariance models.

2.2.3 Description

Une description de quelques mots peut etre ajoutée pour décrire les entités. On ne garde que les descriptions "plutôt courtes" et les mots clés, par exemple quantification et correction.

1. To account for the large-scale conformational movements, SWAAT integrates the calculation of the protein normal modes using ENCoM [45].
2. nextNEOpi: a comprehensive pipeline for computational neoantigen prediction
3. The following is an example command-line for execution of GEMmaker on a local machine using Singularity (for containerization), quantification using Salmon

2.2.4 Biblio

Dans certains articles, la bibliographie est donnée et on l'annote. Une bibliographie peut être un article, un lien internet, une référence sous la forme [*num*].

1. Reference Illumina assemblies were generated with the pipeline Shovill v1.1.0 (https://github.com/tseemann/shovill) using Skesa v2.4.0 and genotyped with Mykrobe v0.9.0 (Hunt et al. 2019) (from reads) and SCCion v0.4.0 (https://github.com/esteinig/sccion), a wrapper around common assembly-based genotyping tools and databases (Zankari et al. 2012; Chen et al. 2016;Kaya et al. 2018) for S. aureus.

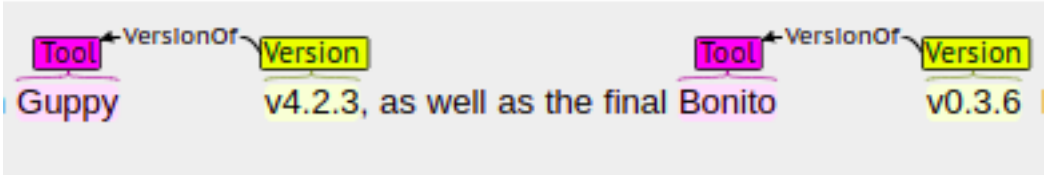
2. dRep.[18]

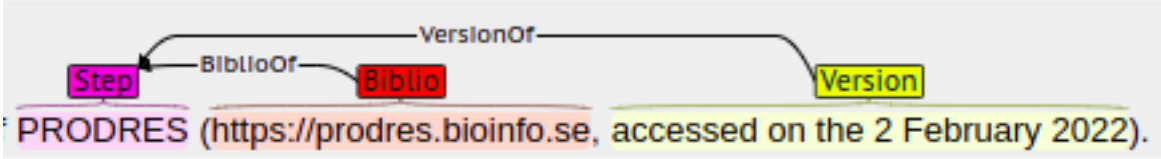
3 Relations

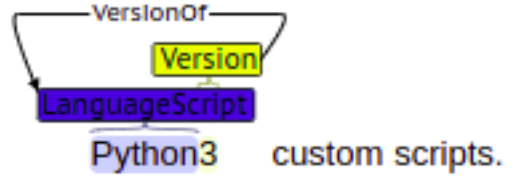
Les entités globales peuvent être reliées à des entités ou reliées entre elles. Il existe différentes relations:

3.1 VersionOf

Cette relation met en lien une entité version vers une entité globale.

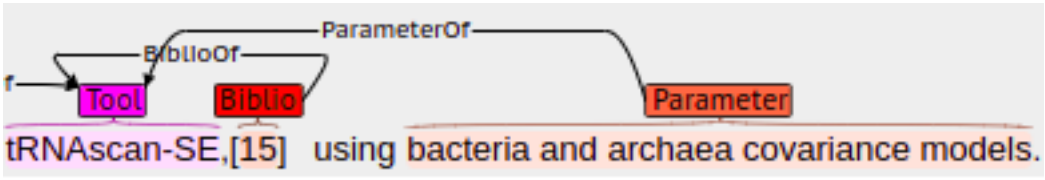


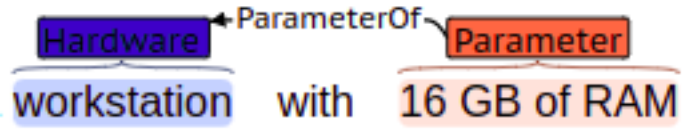




3.2 ParameterOf

Dans certains articles, les auteurs peuvent avoir ajouté à certaines entités globales des informations sur les paramètres à prendre en compte. Cette relation relie un paramètre vers l'entité qu'il complète.



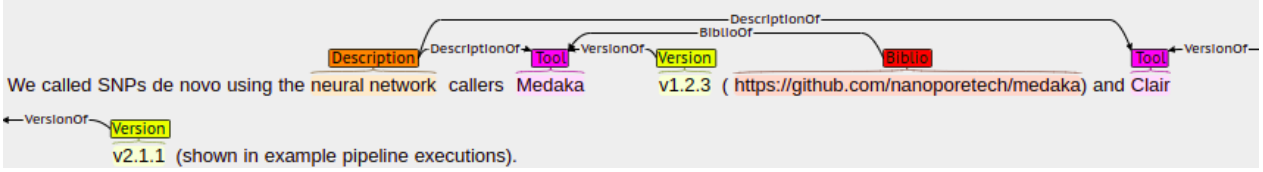


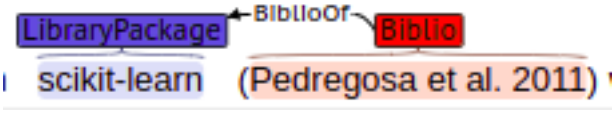
3.3 DescriptionOf

De même, cette relation relie une description vers une entité globale.

3.4 BiblioOf

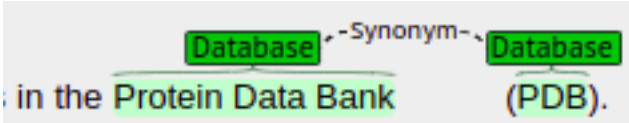
Relation allant d'une bibliographie vers l'élément qu'il décrit





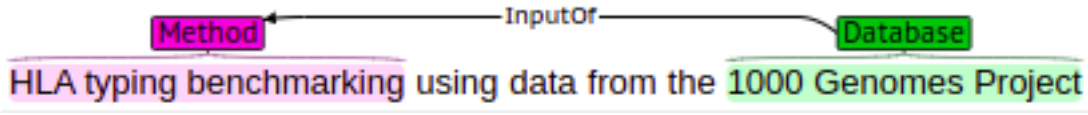
3.5 Synonym

Dans certains articles, des noms "longs" peuvent être abrégés ou posséder des synomynes.



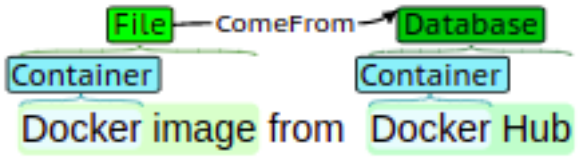
3.6 InputOf - OutputOf

Des paramètres et des données peuvent être utilisés comme input ou output d'une méthode, d'un outil ...



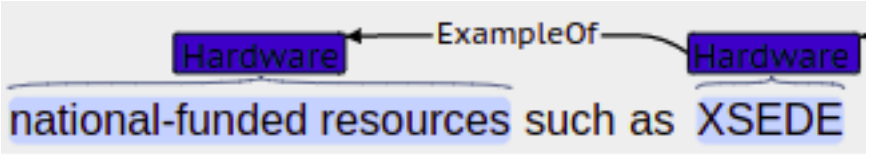
3.7 ComeFrom

Certaines informations peuvent être incluses dans une entité plus grandes (par exemple, certaines données proviennent d'une base de données)



3.8 ExampleOf

Des exemples peuvent être cités



3.9 IncludeIn

Certaines données peuvent être incluses dans un ensemble plus grand ou une méthode peut être incluse dans un outil.

References

[1] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In Proceedings of the Demonstrations Session at EACL 2012.

[2] Hadish JA, Biggs TD, Shealy BT, Bender MR, McKnight CB, Wytko C, Smith MC, Feltus FA, Honaas L, Ficklin SP. GEMmaker: process massive RNA-seq datasets on heterogeneous computational infrastructure. BMC Bioinformatics. 2022 May 2;23(1):156. doi: 10.1186/s12859-022-04629-7. PMID: 35501696; PMCID: PMC9063052.

[3] Steinig E, Duchêne S, Aglua I, Greenhill A, Ford R, Yoannes M, Jaworski J, Drekore J, Urakoko B, Poka H, Wurr C, Ebos E, Nangen D, Manning L, Laman M, Firth C, Smith S, Pomat W, Tong SYC, Coin L, McBryde E, Horwood P. Phylodynamic Inference of Bacterial Outbreak Parameters Using Nanopore Sequencing. Mol Biol Evol. 2022 Mar 2;39(3):msac040. doi: 10.1093/molbev/msac040. PMID: 35171290; PMCID: PMC8963328.

[4] Othman H, Jemimah S, da Rocha JEB. SWAAT Bioinformatics Workflow for Protein Structure-Based Annotation of ADME Gene Variants. J Pers Med. 2022 Feb 11;12(2):263. doi: 10.3390/jpm12020263. PMID: 35207751; PMCID: PMC8875676.

[5] Albanese D, Donati C. Large-scale quality assessment of prokaryotic genomes with metashot/prok-quality. F1000Res. 2021 Aug 17;10:822. doi: 10.12688/f1000research.54418.1. PMID: 35136576; PMCID: PMC8804904.

[6] Rieder D, Fotakis G, Ausserhofer M, René G, Paster W, Trajanoski Z, Finotello F. nextNEOpi: a comprehensive pipeline for computational neoantigen prediction. Bioinformatics. 2021 Nov 12;38(4):1131–2. doi: 10.1093/bioinformatics/btab759. Epub ahead of print. PMID: 34788790; PMCID: PMC8796378.