Literature search: Tools to assess the pathogenicity of Nontuberculous Mycobacteria

# Background information – Mechanisms of bacterial pathogenicity

The general steps of pathogenesis are: exposure, adhesion, invasion, infection and transmission. Bacteria use a plethora of mechanisms to do this (3).

Virulence factors contribute to execution of the previously mentioned steps. These virulence factors can be toxins, surface coats/capsules, surface receptors that bind to host cells, … (1). The presence of the virulence genes are of interest in these NTM genomes. That’s why virulence finder tools are needed.

The virulence factors can be encoded on plasmids and bacteriophage DNA, as well as the chromosomal DNA (1). A plasmid is a small circular, dsDNA molecule that is distinct from a cell’s chromosomal DNA (2). Plasmids can contain genes that give the bacteria a genetic advantage thus can help overcome stressful situations (5). So when researching the pathogenicity, plasmids provide a lot of useful information about it. That is why plasmid finder tools are of interest.

Bacteriophages or phages are viruses that infect and replicate only in bacterial cells. There are two replication strategies that can happen: lytic or lysogenic. Lytic means it will introduce its genome in the host cell and use the host bacterium to assemble multiple copies of the phage. The bacterium dies to release the assembled phages. Lysogenic means that the genome will also be introduced in the bacterial cell genome. However, the bacterium will not die, but pass the incorporated genome of the phage on to daughter cells without killing them. The integrated phage genomes are called prophages (6). This is why (pro)phage finder tools are of interest.

One way of bacteria to spread virulence factors is by horizontal gene transfer (HGT). It is defined by the movement of genetic information between organisms, but not in a parent-offspring relationship (8). When the genetic information is incorporated into the genome of the recipient organism, it forms genomic islands. These islands are blocks of DNA that contain mobile genetic elements. The genomic islands are also referred to as pathogenicity islands since they often contain large blocks of virulence factors (4). This is why genomic/pathogenicity island finder tools are of interest.

Transposon was mentioned regularly in literature when searching for bacterial pathogenicity.

From 7: “Transposons are a group of mobile genetic elements that are defined as a DNA sequence. Transposons can jump into different places of the genome; for this reason, they are called jumping genes.”

And

“Transposons are divided into two main groups: retrotransposons (class І) and DNA transposons (class ІІ). Retrotransposons are often found in eukaryotes. DNA transposons can be found in both eukaryotes and prokaryotes. The bacterial transposons belong to the DNA transposons and the Tn family, which are usually the carrier of additional genes for antibiotic resistance. Transposons can transfer from a plasmid to other plasmids or from a DNA chromosome to plasmid and vice versa that cause the transmission of antibiotic resistance genes in bacteria.”

Reference

1 - [Bacterial Pathogenesis - Medical Microbiology - NCBI Bookshelf (nih.gov)](https://www.ncbi.nlm.nih.gov/books/NBK8526/#:~:text=Pathogenic%20Mechanisms-,Bacterial%20Infectivity,that%20bind%20to%20host%20cells.)

2 - [plasmid / plasmids | Learn Science at Scitable (nature.com)](https://www.nature.com/scitable/definition/plasmid-plasmids-28/#:~:text=A%20plasmid%20is%20a%20small,advantages%2C%20such%20as%20antibiotic%20resistance.)

3 - [Summary of Microbial Mechanisms of Pathogenicity - LabXchange](https://www.labxchange.org/library/items/lb:LabXchange:c0dcc582-a76c-33a4-9d40-d2f7dd23200c:html:1?source=%2Flibrary%2Fbooks%2Fd16810cd-172e-4c2d-b82a-ba21e5dfbe0f#:~:text=Pathogens%20enter%20the%20body%20through,invasion%2C%20infection%2C%20and%20transmission.)

4 - [v078p00216.pdf (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1742320/pdf/v078p00216.pdf)

5 - [Bacterial DNA – the role of plasmids — Science Learning Hub](https://www.sciencelearn.org.nz/resources/1900-bacterial-dna-the-role-of-plasmids)

6 - [Bacteriophages - StatPearls - NCBI Bookshelf (nih.gov)](https://www.ncbi.nlm.nih.gov/books/NBK493185/)

7 - [Transposons: the agents of antibiotic resistance in bacteria - PubMed (nih.gov)](https://pubmed.ncbi.nlm.nih.gov/30113080/)

8 - [Horizontal Gene Transfer - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4536854/#:~:text=Horizontal%20gene%20transfer%20(HGT)%20is,offspring)%2C%20fueling%20pathogen%20evolution.)

# 2. Tools

## 2.1 Tools for detection of virulence factors

**VFDB**

* Description

The virulence factor database (VFDB) is an integrated and comprehensive online resource for curating information about virulence factors of bacterial pathogens.

* URL

Website: [VFDB - pathogenesis of Mycobacterium (mgc.ac.cn)](http://www.mgc.ac.cn/cgi-bin/VFs/genus.cgi?Genus=Mycobacterium)

Paper: [VFDB 2022: a general classification scheme for bacterial virulence factors | Nucleic Acids Research | Oxford Academic (oup.com)](https://academic.oup.com/nar/article/50/D1/D912/6446532?login=false)

* Notes

We can just BLAST our sequences against the VFDB.

**Virulence Searcher 🡪 was mentioned in project description**

* Description

a tool for searching raw genome sequences from bacterial genomes for putative virulence factors

* URL

Paper: [Virulence Searcher: a tool for searching raw genome sequences from bacterial genomes for putative virulence factors - Clinical Microbiology and Infection](https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X(14)62387-9/fulltext)

Website: /

* Notes

I think it doesn’t exist anymore, because I can’t find anything about it online.

**VirulenceFinder**

* Description

VirulenceFinder identifies viruelnce genes in total or partial sequenced isolates of bacteria

* URL

GitHub-like page: <https://bitbucket.org/genomicepidemiology/virulencefinder/src/master/>

Website: <https://cge.food.dtu.dk/services/VirulenceFinder/>

* Notes

On the GitHub-like page, it says that it is currently only useable for E. coli, Enterococcus, S. aureus and Listeria.

**PATRIC (now implemented in BV-BRC) 🡪 maybe interesting db to remember**

* Description

PATRIC is the Bacterial Bioinformatics Resource Center, an information system designed to support the biomedical research community’s work on bacterial infectious diseases via integration of vital pathogen information with rich data and analysis tools.

* URL

Website: <https://www.bv-brc.org/>

* Notes

The Metagenomic Read Mapping is the online tool that you can use to find virulence factors. It uses KMA (K-mer alignment) to align your input against VFDB db when searching for virulence factors.

But you have to sign in, don’t know if it’s free.

**PathoFact**

* Description

A pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data.

* URL

Paper: <https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00993-9>

GitLab: <https://git-r3lab.uni.lu/laura.denies/PathoFact>

* Notes

This tool predicts virulence factors and other things. You can specify (in the config.yaml file) what workflow it has to use, in this case vir for virulence prediction. When using this workflow, the prediction tool consists of two parts: (1) a db consisting of virulence factor HMM profiles (HMMER3) and (2) a random forest model. The training set consists of known virulence factor sequences retrieved from VFDB. The sequences were all associated with experimentally verified virulence factors. Then for the construction of the virulence HMM db, HMM profiles were annotated for the training set using HMMER3 against multiple pre-compiled and in-house annotation dbs (PFAM-A, TIGR, KEGG, …) The best hit in each HMM set was assigned to each gene in the training set if the HMM score was higher than the binary logarithm of the number of target genes (?). Then the random forest model was trained using these five features: amino acid composition (aac), dipeptide composition (DPC), composition (CTDC), transition (CTDT) and distribution (CTDD). Results from both parts are combined into a final prediction.

Input files are .fna files. Output files will be .faa files (containing translated gene sequences) and .contig files (TAB-delimited file containing a mapping from contig ID).

Accuracy for prediction of virulence factors is 0.921 (in 92% of the cases the prediction is correct) and specificity is 0.957 (in a seq with no vf, 95% will be predicted as not a virulence factor, 5% false positive).

**Overview**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tool** | **Input format** | **Advantage** | **Disadvantage** |
| VFDB | Fasta file (.fna) | Identification | Automation (?)  No novel plasmids (but not the goal here) |
| PathoFact | Fasta file (.fna) |  | Prediction |

I would choose VFDB.

## 2.2 Tools for detection of plasmids

**PlasmidFinder 🡪 web application didn’t seem to work**

* Description

PlasmidFinder identifies plasmids in total or partial sequenced isolates of bacteria.

* URL

Paper: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4068535/>

Web application: <https://cge.food.dtu.dk/services/PlasmidFinder/>

Conda install link: <https://anaconda.org/bioconda/plasmidfinder>

GitHub-like page: <https://bitbucket.org/genomicepidemiology/plasmidfinder/src/master/>

* Notes

You upload your assembled bacterial genomes or plasmids and they are turned into a BLAST database. The BLASTn algorithm is used to look for DNA homologies. For a hit to be reported, your input sequence has to cover at least 60% (default) of the length of the replicon sequence in the database.

Seems to be referenced regularly. There is a command line option and a web application option. With the web application you can specify the database (enterobacteriales or gram positive), minimum % identity, minimum % coverage and type of reads (raw sequence data or assembled/draft genome/contig).

When trying it, none of the input sequences had a plasmid. In the information received from NCBI, some actually should have a plasmid. Database also doesn’t seem updated since 2020. Literature says this tool was originally made for the enterobacteriales, but there is an option for gram positive bacteria. So it should normally work on Mycobacteria, but it seems like it doesn’t work.

**plaSquid**

* Description

plaSquid is a Nextflow pipeline for plasmid detection and classification in genomic and metagenomic data. This pipeline accepts either genomic or metagenomic assemblies as input (.fasta). It uses two different approaches to detect plasmids sequences: alignment with minimap2 against a plasmidic database (minidist) and HMM dependent search of plasmid specific genes (repsearch). plaSquid also classifies plasmids into replicon types and MOB (mobility/mobilization genes) groups by comparing RIPs (Repeat-Induced Point Mutation) or Relaxases against custom HMMs. plaSquid can extract plasmids RIP or MOB sequences in order to further analyze these proteins. plaSquid summarises the information gathered by the two complementary approaches in a single output table and allows further analysis as it outputs plasmidic contigs in a single multifasta file ("Result.fasta")

* URL

Paper: [Improved detection and classification of plasmids from circularized and fragmented assemblies | bioRxiv](https://www.biorxiv.org/content/10.1101/2022.08.04.502827v1.full)

GitHub: [mgimenez720/plaSquid: Nextflow pipeline for plasmid detection and classification from metagenomic data (github.com)](https://github.com/mgimenez720/plaSquid)

* Notes

It seems like a fairly new/not well known tool. It does kind of look interesting. However, it was published in bioRxiv (which is a journal where people can submit their not peer-reviewed, edited or typeset articles) which makes the possible results generated with the tool not really reliable, in my opinion.

**plasmidID**

* Description

PlasmidID is a mapping-based, assembly-assisted plasmid identification tool that analyzes and gives graphic solution for plasmid identification.

* URL

Paper: /

GitHub: [BU-ISCIII/plasmidID: PlasmidID is a mapping-based, assembly-assisted plasmid identification tool that analyzes and gives graphic solution for plasmid identification. (github.com)](https://github.com/BU-ISCIII/plasmidID)

* Notes

Input type is Illumina paired-end reads or SMRT sequencing (only contigs) data. Here you have to specify the plasmid database (which you downloaded) and the contigs file (if this is supplied, you don’t need the paired-end reads files). Your input will get mapped against the plasmid db you downloaded. The output file will be a table with id, length, species and description of reference plasmid, also contig name and some images.

Afbeelding met tekst, diagram, schermopname, Lettertype

Automatisch gegenereerde beschrijvingIn the picture below and when reading the information, it looks like only paired-end reads get plasmid detection. If you use contigs, where does the plasmid identification happen? It’s not very clear if you can use this tool only for plasmid identification.

**PlasmidSPAdes**

* Description

plasmidSPAdes is software tool for assembling plasmids from whole genome sequencing data

* URL

Paper: [Plasmid detection and assembly in genomic and metagenomic data sets - PubMed (nih.gov)](https://pubmed.ncbi.nlm.nih.gov/31048319/)

<https://academic.oup.com/bioinformatics/article/32/22/3380/2525610>

GitHub: [ablab/spades: SPAdes Genome Assembler (github.com)](https://github.com/ablab/spades#plasmid)

* Notes

I think this is only an assembler and not a plasmid finder. It will assemble the plasmid, if there is one present, but not tell you what plasmid it is. Don’t think this will be useful.

**PlasForest**

* Description

A random forest classifier to identify contigs of plasmid origin in contig and scaffold genomes.

* URL

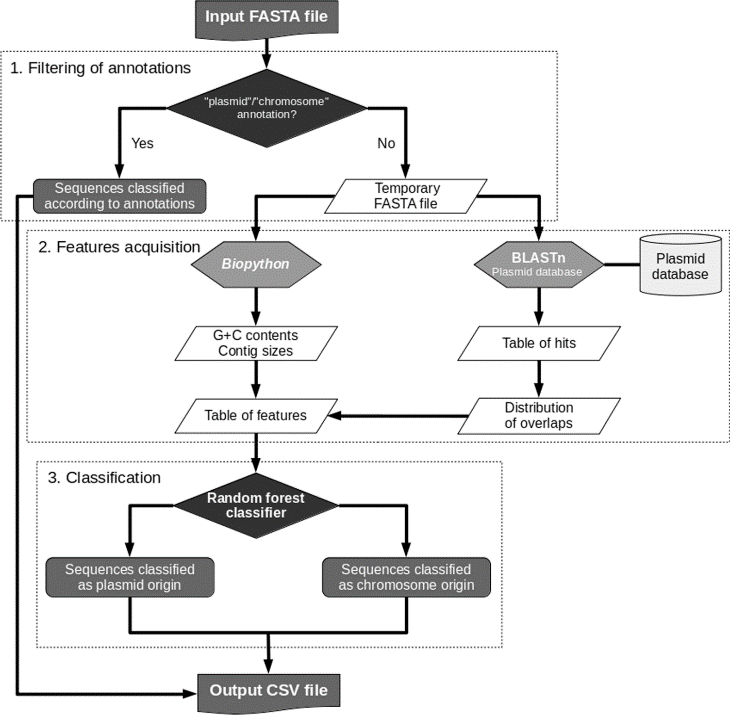
Paper: [PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets | BMC Bioinformatics | Full Text (biomedcentral.com)](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04270-w)

GitHub: [leaemiliepradier/PlasForest: A random forest classifier to identify contigs of plasmid origin in contig and scaffold genomes (github.com)](https://github.com/leaemiliepradier/PlasForest)

* Notes

PlasForest uses a Random Forest classifier to assign contigs in genomic datasets to a plasmid or a chromosome. First the type of input file is identified. Files that are already annotated, will not get annotated again. Files containing only the fasta sequence will undergo following steps. The first step is features acquisition. This consists of submitting the filtered sequences to BLASTn against a local copy of the plasmid db. Also the overlap between query and subject sequences get calculated. Seven features are then computed for each query contig. The features are then passed to the random forest classifier which outputs the predicted identification for each query contig.

It is able to predict 92,7% of plasmid contigs.



**PlasmidSeeker**

* Description

A k-mer based program for the identification of known plasmids from whole-genome sequencing reads

* URL

Paper: <https://peerj.com/articles/4588/>

GitHub: <https://github.com/bioinfo-ut/PlasmidSeeker>

* Notes

The input sample file should be fastq file of raw WGS reads. We have already assembled genomes. You also need an assembled genome of a reference bacterial strain related to the isolate.

**PlasFlow**

* Description

Software for prediction of plasmid sequences in metagenomic assemblies

* URL

Paper: [PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5887522/)

GitHub: <https://github.com/smaegol/PlasFlow>

* Notes

In README files it says that the tools is not maintained. Maybe it is not really reliable anymore.

This tools also uses a machine learning approach. It trains the dataset using neural networks.

It doesn’t do well on short seq (< 1000 bp). Has an accuracy of 96%.

It requires a fasta file (with assembly contigs) as input. Output is tabular file with contig id, contig name, contig length, id and classification label.

**BLAST database of plasmids from ncbi**

* Notes

Just BLAST the sequences against a plasmid database. That’s what some other tools use. The only difference is that they will provide a results/overview folder or file, which is easy to use.

You can make your own plasmid db by downloading the plasmids.txt file from <https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/plasmids.txt> . Maybe we can only filter out the group: Terrabacteria or subgroup: Actinomycetota. I would do this since the downloaded genomes are all classified under Terrabacteria and Actinomycetota. Then the RefSeq id can be searched for and the fasta file for each RefSeq id can be downloaded all into 1 big fasta file and db is made. (This is actually what plasmidID does, kind of).

**Overview**

PlasmidID (maybe PlaSquid) are based on homology search. PlasForest and PlasFlow are both based on machine learning where the tool was trained to detect plasmids. My BLAST option is doing what PlasmidID does but without all the raw seq construction things.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tool** | **Input format** | **Advantage** | **Disadvantage** |
| PlasmiID | Fasta | Identification | Possible only for plasmid identification (?)  Documentation isn’t clear |
| PlasForest | Fasta |  | Prediction |
| PlasFlow | Fasta |  | Not maintained  Prediction |
| BLAST db |  | Identification |  |

I would choose plasForest or Blast db.

Use plasForest and plaSquid.

## 2.3 Tools for detection of prophages

**PHASTER 🡪 mentioned in project description**

* Description

PHASTER (PHAge Search Tool Enhanced Release) is a significant upgrade to the popular PHAST web server for the rapid identification and annotation of prophage sequences within bacterial genomes and plasmids

* URL

Paper: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4987931/>

Paper of PHAST: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125810/>

Website: <https://phaster.ca/>

* Notes

The input file for this tool can be GenBank (annotated genomes) or fasta formatted genomic sequence data (raw sequencing data). It will perform a BLAST search (when a genbank file is given, when giving a fasta file some extra steps are performed) against a custom prophage/phage database. The e-value cutoff is 10e-4. This custom db combines protein sequences from NCBI phage db and a prophage db developed by Srividhya et al (A). Phage-like genes are then clustered into prophage regions using DBSCAN (2 parameters were decided: n = cluster size (6) and e = distance (3000)). After all prophage regions have been detected, a completeness score is assigned to each identified prophage. There are 3 possible scenarios: (1) region only contains genes/proteins of a known phage 🡪 region gets completeness score of 150 (max), (2) >50% of the genes/proteins in the region are related to a known phage 🡪 completeness score is calculated as the sum of the scores corresponding to the regions’ size and number of genes and (3) <50% of the genes/proteins in the region are are related to a known phage 🡪 same calculation as (2) + counts the number of “cornerstone” genes as well as the density of phage-like genes in the region.

A - <https://link.springer.com/chapter/10.1007/978-3-540-37256-1_110>

We will have to download the GenBank files from NCBI. In this link, looks like there is an example of script to automate this. <https://widdowquinn.github.io/2018-03-06-ibioic/01-introduction/02-annotation.html#:~:text=GenBank%20format%20is%20intended%20to,same%20INSDC%20feature%20table%20design>. Doesn’t necessarily need to happen since giving the .fna files also worked.

This tool detects prophages (instead of predicting it). It detects them based on blasting it against a db of prophages. So the result will be prophages that are present in the db and are found in our sequences.

Try docker

**Prophage Hunter**

* Description

Integrative tool that employs similarity matching within our customized phage parts library and machine learning of prophage genetic features, to score the probability of a prophage being active.

* URL

Paper: <https://academic.oup.com/nar/article/47/W1/W74/5494712?login=false>

Website: <https://pro-hunter.genomics.cn/>

* Notes

Cannot access the website, Gateway time-out 🡪 maybe bcs website from china?

**VirSoter2**

* Description

a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses

* URL

Paper: <https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00990-y>

GitHub: <https://github.com/jiarong/VirSorter2>

* Notes

First of all, the input sequences are automatically annotated and relevant features are extracted. Prodigal is used for the identification of coding sequences (CDS). Then annotation of predicted CDS is done using HMMER3 against Pfam and a custom comprehensive viral HMM db. The extracted features are used as input for five distinct random forest classifiers, each associated with a different major type of viral group. Each classifier yields a “viralness” score. This score can be used to determine the likelihood of the input sequence to represent a partial or complete genome from the corresponding viral group. Lastly, the scores are aggregated into a single prediction provided to the user. The F1-score (accuracy) is more than 0.8.

Input file is fasta file (no annotations). You can specify the viral groups you want.

This is not particularly for prophages, but for viral sequences in general. It also predicts rather than detect viral sequences. Both are disadvantages in my opinion.

Little bit confused, do viral sequences/genomes in a bacterial genome always originate from a bacteriophage? So if viral sequences are found in our bacterial genomes of interest, are they then always prophages? Prophages = viral genomes integrated in a microbial genome. Proviruses also exist = virus genome integrated into the DNA of a host cell (basically same as prophage), but it doesn’t cut itself out when the host cell + host cell is eukaryotic.

**PhiSpy**

* Description

Prediction of prophages from bacterial genomes

* URL

Paper: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3439882/>

Extra doc: [supp\_gks406\_nar-03036-met-n-2011-File011.pdf](file:///C:\Users\dedec\Downloads\supp_gks406_nar-03036-met-n-2011-File011.pdf)

GitHub: <https://github.com/linsalrob/PhiSpy>

* Notes

The input format for this tool is GenBank format. The first step of this tool is the calculation of different characteristics for the whole genome using a sliding window of n genes. The calculated characteristics are the customized AT/GC skew, difference in median protein length, transcription strand orientation, abundance of phage words and homology. The second step is to classify a window as a bacterial or a prophage window using random forests. If there is a closely related training genome present, the random forest is executed using that closely related training genome. If there is no closely related training genome present, a generic training set is used. The random forest will produces a rank for each sliding window and it produces a rank (0: non-prophage genes or 1: prophage genes) for each gene by taking the average rank of the window in which the gene participated. The next step is to define the att sites for the predicted prophages (rank 1). This is done by identifying a repeated short DNA sequence which has minimum distance from integrase. Then the predicted prophage region is verified. A region is considered as prophage if there are >5 unknown/phage like proteins and if the number of phage like/unknown proteins >= half of the total amount of proteins in the predicted region. Lastly, the whole genome is traversed to see if there is a group of phage like genes that was not considered in the initial prediction. The output will be a list of potential prophages and some more files/data.

This tool also predicts rather than detect prophages. The input file is also in a GenBank format, which we need to download again. These are disadvantages. It is however specifically designed to predict prophages which is an advantage.

**VirFinder**

* Description

a novel k-mer based tool for identifying viral sequences from assembled metagenomic data

* URL

Paper: [VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data (springer.com)](https://link.springer.com/epdf/10.1186/s40168-017-0283-5?author_access_token=YQgkTWibFIFPtRICkTjZF2_BpE1tBhCbnbw3BuzI2RMCpVMGldKV8DA9scozc7Z-db3ufPFz9-pswHsYVHyEsCrziBuECllLPOgZ6ANHsMeKF5KejrdDKdeASyDkxB5wfFDq523QSd01cnqxCLqCiQ%3D%3D)

GitHub: <https://github.com/jessieren/VirFinder>

* Notes

VirFinder identifies viral sequences based on the empirical observation that viruses and hosts have discernibly different k-mer signatures.

Also based on machine learning techniques so it predicts viral sequences. Also not specifically for prophages.

**Overview**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tool** | **Input type** | **Advantage** | **Disadvantage** |
| PHASTER | GenBank format  Fasta file (.fna) | For prophages  Identification of prophages | Multi-fasta file with contigs =< 2000 bp won’t be processed  Automation (?) |
| VirSorter2 | Fasta file (.fna, .fa …) |  | Prediction  For viral sequences (general) |
| PhiSpy | GenBank format | For prophages | Prediction |
| VirFinder | Fasta file (.fna, .fa …) |  | Prediction  For viral sequences (general) |

I would choose PHASTER or PhiSpy.

## 2.4 Tools for detection of genomic islands

**IslandViewer 4 🡪 mentioned in project description**

* Description

Prediction and analysis of genomic islands in bacterial and archael genomes

* URL

Paper: [IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5570257/)

Paper Islandpath-DIMOB: <https://academic.oup.com/bioinformatics/article/34/13/2161/4904263>

Paper SIGI-HMM: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1489950/>

Paper IslandPick: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2518932/>

Website: [Islandviewer 4 - Genomic Island Prediction and Genome Visualization Tool (sfu.ca)](https://www.pathogenomics.sfu.ca/islandviewer/)

* Notes

Each month all sequenced genomes are downloaded from NCBI FTP server and loaded into a local MySQL database. For the prediction of GIs, the tools IslandPath-DIMOB, SIGI-HMM and IslandPick were used.

**IslandPath-DIMOB** predicts GIs based on the detection of dinucleotide biases and the identification of mobility genes in the same region. Using dinucleotide biases, instead of the conventional GC content analysis, makes the detection of GIs more sensitive. The presence of a mobility gene reduces the false positive predictions (such as highly expressed genes that also exhibit abnormal sequence composition). The identification is performed in two parallel steps: (1) identification of known Pfam domains in proteins and (2) identification of keywords in protein function annotation.

**SIGI-HMM** is based on codon usage bias with a HMM approach.

**IslandPick** is based on a comparative genomics approach. With the use of stringent but potentially flexible criteria, and with distance cutoffs, query genomes that gave a sufficient number of suitably related species or strains are selected to conduct the analysis of GIs. With IslandPick also regions that are not likely to contain GIs are identified.

# 3. Extra

## 3.1 Tools for detection of transposons

Might be good paper: [A benchmark of transposon insertion detection tools using real data | Mobile DNA | Full Text (biomedcentral.com)](https://mobilednajournal.biomedcentral.com/articles/10.1186/s13100-019-0197-9)

## 3.2 Tools for general pathogenicity prediction

PathogenFinder 1.1 (prediction of a bacteria’s pathogenicity towards human hosts)

Paper: [PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data | PLOS ONE](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077302)

Website: [CGE Server (dtu.dk)](https://cge.food.dtu.dk/services/PathogenFinder/)

This website seems to list a lot of useful tools: <http://www.genomicepidemiology.org/services/>