Instructions for Protein and Nucleic Acid Target Identification: making a Hit List

David Condon

November 25, 2024

Abstract

HitListis a suite of scripts to enable searching of entire genomes, using subtractive genomics in massively parallel fashion to find protein and/or RNA possible targets for antimicrobials.

Throughout this document, `typewriter font` will be used to indicate typing at the command line.

1 Dependencies

HitList was written to minimize dependencies and maximize portability, and a Linux or Mac operating system is assumed. The suite of scripts needs perl 5.12 or greater, python3, and python3's matplotlib. Perl modules are in the same directory as HitList scripts. HitList needs BLAST installed in `$PATH`.

In the case of missing Perl modules, install Data::Printer, NBCI BLAST, and Devel::Confess thus on an Ubuntu system:
`sudo apt install libdata-printer-perl libdevel-confess-perl ncbi-blast+`
If `blast` commands are in `PATH`, then ncbi-blast doesn't need to be installed.

2 BLAST Databases

2.1 Download Fasta

Input fasta, whether proteome, genome, or transcriptome, should be downloaded, usually from NCBI, or another source.

2.2 Make BLAST Databases

Blast databases should be created using `makeblastdb` and the fasta files downloaded in section 2.1.

3 Creation of Input Files

Species and the input files they use are represented in tab-delimited files, where the species is in the first column, and the fasta file for that species is in the right column. HitList assumes that `makeblastdb` has been done, and that for a given fasta input, the file suffix, e.g. `.faa` has `.ntf`, `.ntq` etc. files available.

1. Below can be seen an example for hosts (`hosts.tsv`):

```
H. Sapiens   /home/con/bio.data/blastdb/human.protein/GCF_000001405.40_GRCh38.p14_protein.faa
```

2. and an example pathogen file (`pathogens.tsv`):

```
C.auris   /home/con/bio.data/blastdb/candida.auris/GCF_003013715.1_ASM301371v2_protein.faa.gz
```

   which may have as many species in either file as desired,

3. a fasta-format file of essential genes, e.g. `gene.list.fa`, for example, should be created. Due diligence should be done to ensure that appropriate genes/proteins are included.

4 Selection of Protein/RNA List

The list of proteins/RNA to be entered into the pipeline is at the discretion of the end user. The first list of proteins when using this project was from *Saccharomyces cerevisiae* at the Database for Essential Genes. However, that essential list contained genes/proteins that were clearly targetable by anti-fungals and were not listed, such as Fks1, so proper discernment must be used when selecting genes/proteins. This list will be further referred to as `gene.list.fa`

5 Running Scripts

All scripts reference the HitList source directory, where all scripts are kept: `$dir`, which will vary depending on your directory names.

Run `perl $dir/scripts/hitlist.pl --test` from the `test` directory to ensure that all sub-scripts are capable of running on your local system.

The list of options is:

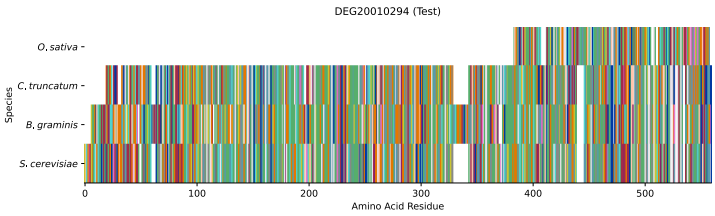
Option	Description
hosts	The tab-delimited file of species name and source file for hosts
output-svg	Output scatterplot in SVG format
pathogens	The tab-delimited file of species name and source file for pathogens
plot-title	Title on plot for resulting scatterplot
query-list	The fasta file of protein/RNA sequences
query-species	Source species for gene/RNA-list; shouldn't have any spaces
test	Run test to ensure that HitList works; every other option is ignored

When you're ready to run your own data, run

```
perl $dir/scripts/hitlist.pl --query-list test.progen.fa --hosts hosts.tsv --pathogens pathogens.tsv --output-svg scatterplot.svg --plot-title 'Test' --query-species 'S.cerevisiae'
```

6 Output

1. MSA images in SVG format, for example:



2. a list of target proteins/RNAs, and the length of the targetable region in xlsx format,
3. a scatterplot showing which proteins will be most likely targetable