Virus Genotyping: Standard Operating Procedure

1. Purpose

This document describes the ViPR computational method used to determine virus genotypes. It is a command line, back-end data processing tool, ran from a front end user interface. Currently version 1.3.0 operates on the following Flavivirus species: Dengue, West Nile, Japanese encephalitis, St. Louis encephalitis, Tick-borne encephalitis, Yellow Fever, Murray Valley encephalitis, Bovine viral diarrhea virus, and Zika virus. HCV is computed for the user-input data using another algorithm and is described under:

https://www.viprbrc.org/brcDocs/documents/SOP_HCV_typing.pdf

As of version 1.3.0, for Zika virus, the reference phylogeny tree is from Lanciotti et al. (1). The draft reference phylogeny tree for HCV is from Smith et al. (2). Some components were derived from the work of Los Alamos HCV database team and the VBRC (Virus Bioinformatics Resource Center) project. It is portable to other virus families, provided that reference sequences for the relevant genotypes are available.

2. Method Description

The ViPR genotype/recombination method uses phylogenetic trees to find the closest reference genome, from input genotypes. Recombination is detected by using a sliding window, where each k-mer is aligned against reference genomes, and any changes along the genome are noted. It is a relativistic method, rather than using absolute homology standards such as an *e-value* from *Blastn*.

Operationally it builds a phylogenetic tree from input DNA sequence, and determines the best genotype based on the tree structure. Sensitivity is set by defining Branching Indices (BI) cutoffs derived from the phylogeny trees; hits with BI < 0.714 are deemed unassignable. The package includes a Perl script vipr_genotype.pl, Perl modules Genotype.pm and Draw_graph.pm, and an MSA for each of the viral species.

Dependencies are on these packages in the executable PATH for alignment and phylogeny compute: MUSCLE, dnadist of the Felsenstein “Phylip” package, and FastME (8), a “minimum evolution” tree inference engine. The dnadist/FastME approach is used for speed; other tree-generating programs such as Phylip and PhyML could be used with minimal code change. Pplacer is currently used to identify the proper leaf hit in the phylogeny tree.

Validation and testing of the HCV data set (200k genomes) was performed against 419 sequences from Krishnan et. al. (3) References for these were drawn from the ICTV (4,7). Though the method is not available for users to compute their own HCV genotype, the current results are drawn from those standards.

3. Input Data Preparation

Input for the method is a file in GenBank format (GBK) or FASTA format. The DNA sequence is aligned by MUSCLE with the corresponding reference sequences. The resulting MSA is then used to generate a distance matrix by dnadist, which is then used by FastME to generate the phylogeny tree. The BI value (0-1) for the input genome is calculated as the ratio of distance between grandparent and parent of the target and distance between sibling (a reference) and grandparent. If the BI is above a certain threshold, it is assumed that the target belongs to the sibling node genotype. For HCV the reference threshold was 0.711, from Hraber *et. al* (5, 6).

Command options of the vipr_genotype.pl script:

-t	Taxon of the input genome. Valid options are: DENGUE, STLOUIS, WESTNILE, JAPENCEPH, TKBENCEPH, YELLOWFEVER, BOVDIARRHEA1, MURRAY, ZIKA. If absent, a separate blastn run will be performed to determine the species. However this process isn't as accurate as it should be, so a user-supplied species is preferred.
-d	Directory of the input file(s)
-i	Name of input genome file in GenBank or FASTA format. If omitted, process all GenBank files in directory.
-recomb	Also run the recombination calculation, where a sliding window of 400 nucleotides is moved by 100 nucleotides each step, to produce a genotype profile along the sequence.

Sample commands:

```
$ ./vipr_genotype.pl -t HCV -d ./ -i AB119282.gb
$ ./vipr_genotype.pl -t HCV -d ./ -i AB119282.faa
$ ./vipr_genotype.pl -t HCV -d ./
$ ./vipr_genotype.pl -t HCV -recomb -d ./ -i AB119282.gb
$ ./vipr_genotype.pl -t HCV -recomb -d ./
```

4. Output Data, Processing, and Display

Results of the genotype analysis are output in a .tsv file named *_genotype_summary.tsv with rows of data, one for each input sequence. The genotype analysis generates a single output line, listing the job_name, job_type, start and end of the sequence, BI value, genotype, time stamp, status, and any comment.

```
Job_name  Job_type      start..end Branch_Index    genotype        Date(ymdhms)    Status  Comment
AB119282  genotype       1..9658   BI=0.765676     genotype=1b     20111109104143  Success comment=
```

Results of the recombination analysis are output in a .tsv file named *_recombination_summary.tsv with rows of data, one for each input sequence, delimited in format:

```
AB119282  recomb_summ    1..10262  BI=0.000000     genotype=1b     20111109104326  Success comment=
AB119282  recomb_window  1..400    BI=0.286235     genotype=1a     20111109104144  Success comment=
...
```

For a recombination analysis, a line is first generated to give the summary of all steps. Any subtype with >=5 steps with BI>threshold is listed in the genotype field and separated by a comma. This is followed by output lines labeled by job_type=recomb_window for each step. If the method does not produce an answer, such as when there are no similar strains, a ‘Failed’ error notice is given. This helps distinguish between technical and scientific bugs. Other files include the following formats: Input DNA sequence (FASTA), MSA (FASTA or Phylip), phylogenetic tree (Newick), and a bar graph of the BI profile, in .gif format. This is a stand-alone method.

5. References

1. Lanciotti, R.S., Lambert, A.J., Holodniy, M., Saavedra, S., del Carmen Castillo Signor, L. (2016) Phylogeny of Zika virus in Western Hemisphere. *Emerg Infect Dis*. 22:5 (May 2016) http://dx.doi.org/10.3201/eid2205.160065.
2. Smith, D.B., Bukh, J., Kuiken, C., Muerhoff, A.S., Rice, C.M., Stapleton, J.T., Simmonds, P. (2104) Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. *Hepatology*. 59(1):318-27.
3. Krishnan et. al. *(2016) In Vitro* and *In Vivo* Antiviral Activity and Resistance Profile of Ombitasvir, an Inhibitor of Hepatitis C Virus NS5A. Antimicrob Agents Chemother. 2016; 60: 5368–5378.
4. Donald B. Smith, Jens Bukh, Carla Kuiken, A. Scott Muerhoff, Charles M. Rice, Jack T. Stapleton and Peter Simmonds *HCV Classification*: A web resource to manage the classification and genotype and subtype assignments of hepatitis C virus. URL (Jun 2017): https://talk.ictvonline.org/ictv_wikis/flaviviridae/w/sg_flavi/56/hcv-classification
5. Hraber, P., Kuiken, C., Waugh, M., Geer, S., Bruno, W.J., Leitner, T. (2008) Classification of hepatitis C virus and human immunodeficiency virus-1 sequences with the branching index. *J Gen Virol*. *Sep;89 (Pt 9):2098-107*. PMID: 18753218.
6. Hraber, P.T., Leach, R.W., Reilly, L.P., Thurmond, J., Yusim, K., Kuiken, C; Los Alamos HIV database team. (2007) Los Alamos hepatitis C virus sequence and human immunology databases: an expanding resource for antiviral research. *Antivir Chem Chemother*. 18(3):113-23. Erratum in: *Antivir Chem Chemother*. *18(4):243*. PMID: 17626595.
7. Kuiken, C., Simmonds, P. (2009) Nomenclature and numbering of the hepatitis C virus. *Methods Mol Biol*. *510:33-53*. PMID: 19009252.
8. Desper, R. & Gascuel, O. (2002). Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle. *Journal of Computational Biology 19(5), pp. 687-705*.

