APPLICATION NOTE Genomics & Informatics Vol. 6(1) 54-56, March 2008

GTVseq: A Web-based Genotyping Tool for Viral Sequences

Jae-Min Shin[1,2]*, Ho Eun Park[1], Yong-Ju Ahn[2], Doo-Ho Cho[1], Ji Han Kim[2], Mee Kyung Kee[2,3], Sung-Soon Kim[3], Joo-Shil Lee[3] and Sangsoo Kim[2]*

[1]SBscience Inc., Seongnam, [2]Department of Bioinformatics, Soongsil University, Seoul 156-743, Korea, [3]Department of Immunology and Pathology, National Institute of Health, Seoul 122-701, Korea Centers for Disease Control and Prevention, Seoul 122-701, Korea

Abstract

Genotyping Tool for Viral SEQuences (GTVseq) provides scientists with the genotype information on the viral genome sequences including HIV-1, HIV-2, HBV, HCV, HTLV-1, HTLV-2, poliovirus, enterovirus, flavivirus, Hantavirus, and rotavirus. GTVseq produces alternative and additive genotype information for the query viral sequences based on two different, but related, scoring methods. The genotype information produced is reported in a graphical manner for the reference genotype matches and each graphical output is linked to the detailed sequence alignments between the query and the matched reference sequences. GTVseq also reports the potential 'repeats' and/or 'recombination' sequence region in a separated window. GTVseq does not replace completely other well-known genotyping tools such as NCBI's virus sequence genotyping tool (http://www.ncbi.nlm.nih.gov/projects/genotyping/formpage.cgi), but provides additional information useful in the confirmation or for further investigation of the genotype(s) for the newly isolated viral sequences.

Keywords: Genotyping, Recombination, Subtyping, Viral Genome, Viral Sequence, Virus database, Visualization, Web-Server

Introduction

Identification of the genotype (or sub-genotype) of viral genome sequences or viral sequence fragments is important both for studying epidemiology and for developing vaccines. Recent advances in genome analysis and accumulation in sequence databases enable scientists to obtain detailed sequence information for viral genotype through database searching. For example, phylogenetic analysis has been used to identify the viral genotype of newly isolated viral sequences by comparing them to existing alignments and trees. There are several useful web-based tools and database resources for the genotyping analysis of the viral sequences, based on phylogenetic trees (Kuiken *et al*., 2005; de Oliveira *et al*., 2005), sequence similarities of whole genome and/or local sequences (Rozanov *et al*., 2004; Bao *et al*., 2004; Tcherepanov *et al*., 2006), or position-specific scoring matrices (Myers *et al*., 2005). However, accurate determination of the genotype of viral sequences based on known viral genome databases, is often very difficult or even impossible because many newly sequenced viral sequences are determined as 'new-subtype' in existing phylogeny or as 'recombination' of known genotypes.

Genotyping Tool for Viral SEQuences (GTVseq) does not determine the single most probable genotype of the given viral genome sequences, but rather provides useful suggestions for genotype information. Compared to the current genotyping tools, GTVseq has several unique and useful features in the following aspects:

* GTVseq uses two different scoring schemes and the results are reported separately. One of the scoring schemes is similar to that of NCBI, while the other is particularly useful for viral sequences with new or complicated genotypes (*vide infra*).

* GTVseq offers an easy and interactive web-based user interface, with intuitive reports for genotyping results.

* GTVseq can be used for genotyping many important viruses such as HIV-1, HIV-2, HBV, HCV, HTLV-1, HTLV-2, poliovirus, enterovirus, flavivirus, Hantavirus, and rotavirus, thus permitting the most comprehensive genotyping of viral genomes to date.

Methods

For genotyping of viral genome sequences, we need to establish 'reference sequences' for each genotype. We have downloaded the reference sequence database collections from NCBI (http://www.ncbi.nlm.nih.gov/projects/genotyping), for HIV-1, HIV-2, HBV, HCV, HTLV-1, HTLV-2, and poliovirus. For HIV-1 reference sequences, GTVseq also provides several different collections of reference databases such as HIV-1 (2004) & CRF, HIV-1 (2005), HIV-1 (2005) & CRF. For enterovirus, flavivirus, Hantavirus, and rotavirus, the reference sequences were

*Corresponding author: J.M. Shin, E-mail sbscience@gmail.com, Tel +82-31-719-7937, Fax +82-31-719-7938, S.S. Kim, E-mail sskimb@ssu.ac.kr, Tel +82-2-820-0457, Fax +82-2-824-4383
Accepted 14 March 2008

obtained through personal communications with the corresponding research communities.

Based on these reference sequence databases, GTVseq provides two different, but related, scoring strategies based on the BLAST (Altschul *et al*., 1997) results: 'local-identity' and 'HSPs-bit' scores. In the local-identity scoring method, GTVseq adopts a sliding-sequence-window scoring method, which is conceptually similar to NCBI genotyping methods (Rozanov *et al*., 2004). But for a given sequence window, GTVseq averages the local-identity scores of the overlapping windows. Thus, the genotyping output based on local-identity-score may give results similar to NCBI's genotyping tool (Rozanov *et al*., 2004), where the most similar match to the known reference genotype in every local sequence window is always ranked at the top of the list. In most cases, the local-identity score is adequate for searching all the potential "circulating recombinant forms (CRFs)" or intra- or inter-recombinant forms between two or more virus strains.

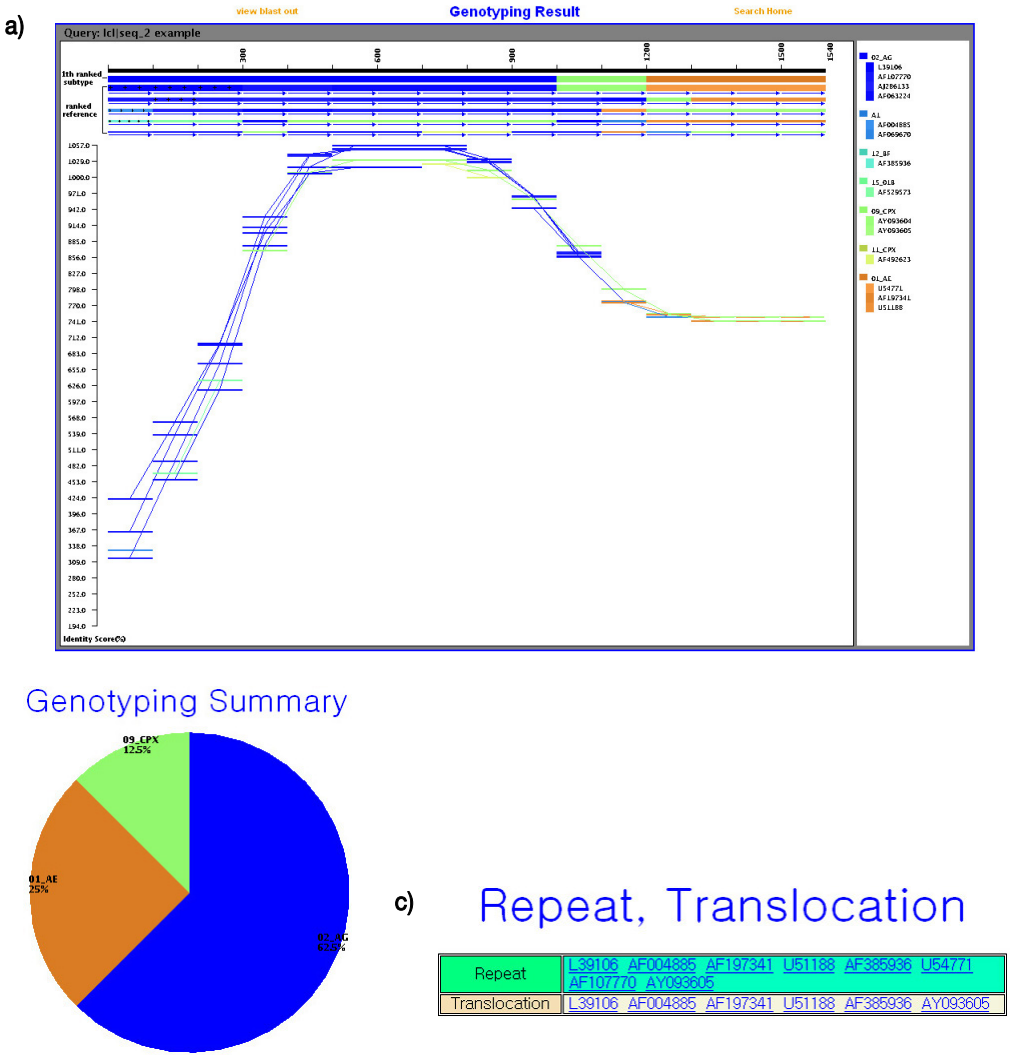

Fig. 1. An example web-page of GTVseq results using an artificial HIV-1 sequence that was a mosaic composite of one of the GenBank sequences (Acc. M62320). (a) The main genotyping output window, where the top-five hits for the local sequence window of the query are displayed with different colors. The genotypes corresponding to the matching colors are displayed on the right-hand side. The line-graph in the bottom represents the 'score', where all the scores for local sequences are separated. (b) The pop-up window for the summary of the genotypes for the query. The percentage refers to the genotype composition for the top ranked hit. The color schemes are the same as in (a). (c) Another pop-up window for potential 'repeat' or 'translocation' sequence regions, based on the sequence alignments between the query and each reference genome. The detailed 'repeas' or 'translocation' regions are displayed when each link is clicked.

In the HSPs-bit scoring method, all HSP (High-scoring Segment Pair) matches between query and reference sequences are collected. All the HSPs spanning a given sliding window are considered and the subtype of the HSP with the highest bit score is then assigned to the region. As HSPs are often much longer than the window, this procedure has the effect of considering the alignment over wider range including neighboring windows and consequently being more robust to the local variations of the query sequences with complex recombinations of subtypes or divergent from the references. In these cases, those methods relying on only 'local identity' only typically produce unnecessarily complicated subtype patterns.

Results and Discussion

Our system was tested with an artificial sequence that was a mosaic composite of one of the HIV-1 sequence (GenBank Acc. M62320). As shown in Fig. 1(a), GTVseq reports the top five hits for each local sequence window among the known reference genotypes. For a recombinant sequences, it conveniently shows with a graphic user interface the pie-chart of the genotype composition [Fig. 1(b)]. One of the most intriguing features of GTVseq is its ability to detect 'repeats' or 'translocation' in the query sequence [Figure 1(c)]: A segment in the reference sequence may match multiple regions of the query (repeat) or a segment in the query may be out of order compared to the reference (translocation). These 'repeats' or 'translocations' in retrovirus such as HIV-1 can be interesting and should be considered in the interpretation of the genotyping based on sequence alignments. To our knowledge, no other genotyping/ Subtyping tool specifically addresses these issues. More detailed information on the scoring can be obtained from our GTVseq web-site (http://vsd.ssu.ac.kr:8080/GTVseq/Help.htm).

Acknowledgements

This study was supported by a grant of the Korea Centers for Disease Control and Prevention.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 26, 3986-3991.

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., and Tatusova, T. (2007). FLAN: a web server for influenza virus genome annotation. *Nucleic Acids Res.* 35, W280-284.

de Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., van Rensburg, E.J., Wensing, A.M., van de Vijver, D.A., Boucher, C.A., Camacho, R., and Vandamme, A.M. (2005). An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics* 21, 3797-3800.

Kuiken, C., Yusim, K., Boykin, L., and Richardson, R. (2005). The Los Alamos hepatitis C sequence database. *Bioinformatics* 21, 379-384.

Myers, R.E., Gale, C.V., Harrison, A., Takeuchi, Y., and Kellam, P. (2005). A statistical model for HIV-1 sequence classification using the subtype analyser (STAR). *Bioinformatics* 21, 3535-3540.

Rozanov, M., Plikat, U., Chappey, C., Kochergin, A., and Tatusova, T. (2004). A web-based genotyping resource for viral sequences. *Nucleic Acids Res.* 32, W654-W659.

Tcherepanov, V., Ehlers, A., and Upton, C. (2006). Genome Annotation Transfer Utility (GATU): rapid annotation of viral genomes using a closely related reference genome. *BMC Genomics* 7, 150.