An automated method for identification of Dengue, Zika, Chikungunya and Yellow Fever virus species and genotypes

Dengue, Zika, Chikungunya & Yellow Fever Viruses Typing Tool Version 1.0

This tool is designed to use Blast and phylogenetic methods in order to identify the Dengue, Zika, Chikungunya & Yellow Fever Viruses serotypes and genotypes of a nucleotide sequence. **Note for batch analysis:** The tool accepts up to 2000 sequences at a time.

You may either:

A. paste one or more sequences in FASTA format in the input field.
B. upload a FASTA file.
C. revisit results of a previous run

**A) Paste nucleotide sequence(s) in FASTA format:**

```
>21_AF100409, Mexico, 1992
TCCAGGCTTTACCATTAGGCACATTGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAATGGGACATAGCTGGCTTCAAT

B) Or, upload a FASTA with nucleotide sequences:

Luiz Carlos Alcantara
Oswaldo Cruz Foundation, Salvador-Bahia, Brazil
A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences

Luiz Carlos Junior Alcantara¹, Sharon Cassol², Pieter Libin³, Koen Deforce⁴, Oliver G. Pybus⁵, Marc Van Ranst³, Bernardo Galvão-Castro¹, Anne-Mieke Vandamme³ and Tulio de Oliveira⁶,*

¹Laboratório Avançado de Saúde Publica, Centro de Pesquisas Gonçalo Moniz, Fundação Oswaldo Cruz, Brazil, ²MRC Unit for Inflammation and Immunity, Department of Immunology, University of Pretoria and Tshwane Academic Division of the National Health Laboratory Service, Pretoria, South Africa, ³Laboratory for Clinical and Epidemiological Virology, Katholieke Universiteit Leuven, Belgium, ⁴MyBioData bvba, Rotselaar, Belgium, ⁵Department of Zoology, University of Oxford, UK and ⁶Africa Centre for Health and Population Studies, University of KwaZulu-Natal, South Africa

Received February 11, 2009; Revised May 13, 2009; Accepted May 14, 2009

ABSTRACT

Human immunodeficiency virus type-1 (HIV-1), hepatitis B and C and other rapidly evolving viruses are characterized by extremely high levels of genetic diversity. To facilitate diagnosis and the development of prevention and treatment strategies that efficiently target the diversity of these viruses, and other pathogens such as human T-lymphotropic virus type-1 (HTLV-1), human herpes virus type-8 (HHV8) and human papillomavirus (HPV), we developed a rapid high-throughput-genotyping system. The method involves the alignment of a query sequence with a carefully selected set of pre-defined reference strains, followed by phylogenetic analysis of multiple overlapping segments of the alignment using a sliding window. Each segment of the query sequence is assigned the genotype and sub-genotype of the reference strain with the highest bootstrap (>70%) and bootscanning (>90%) scores. Results from all windows are combined and displayed graphically using color-coded genotypes. The new Virus-Genotyping Tools provide accurate classification of recombinant and non-recombinant viruses and are currently being assessed for their diagnostic utility. They have incorporated into several HIV drug resistance algorithms including the Stanford (http://hivdb.stanford.edu) and two European databases (http://www.umcutrecht.nl/subsite/spread-programme and http://www.hivrdb.org.uk/) and have been successfully used to genotype a large number of sequences in these and other databases. The tools are a PHP/JAVA web application and are freely accessible on a number of servers including: http://bioafrica.mrc.ac.za/rega-genotype/html/ http://lasp.cpqgm.fiocruz.br/virus-genotype/html/ http://jose.med.kuleuven.be/genotypetool/html/.

INTRODUCTION

Human immunodeficiency virus type 1 (HIV-1) and hepatitis C virus (HCV) are two of the most serious infectious diseases to have affected humankind. HCV has infected an estimated 170 million people worldwide and is the leading cause of chronic liver disease and hepatocellular carcinoma (1). HIV-1/AIDS, the most widespread pandemic in recorded human history, has already infected an estimated 42 million people and has claimed the lives of 22 million people with the majority of deaths (70%) occurring in sub-Saharan Africa (http://www.unaids.org/en/KnowledgeCentre/HIVData/GlobalReport/2008/). Both pathogens are small, rapidly evolving RNA viruses with high mutation rates, high production rates (in excess of 10⁹ virions per day) and, in the case of HIV-1, a strong
Global-Complete Genomes (GenBank)

- **Alignment**: MAFFT software
- **Phylogenetic analysis**: PhyML (ML, 1000 bootstrap) and MrBayes (Bayesian)
- **4 DENV serotypes and 18 genotypes**: 1I, 1II, 1III, 1V, 2I, 2II, 2III, 2IV, 2V, 2VI, 3I, 3II, 3III, 3V, 4I, 4II, 4III, 4IV
- **3 CHIKV genotypes**: ECSA, Asian-Caribbean, West African
- **3 ZIKV genotypes**: African, Asian, West African
Selecting the Reference Strains

• Identification of highly divergent but equidistant whole-genome sequences that are representative for the diversity within the different DENV, CHIKV and ZIKV genotypes

• 190 sequences: 146 DENV, 26 CHIKV, 18 ZIKV
• **Alignment**: MAFFT and Clustal softwares
• **Phylogenetic analysis**: PAUP* (NJ), PhyML (ML) (1000 bootstrap support) and MrBayes (Bayesian)
Maximum likelihood phylogenetic tree of the DENV reference strains. Outliers are in blue.
Using the Bootscanning method: to recombination analysis of the 10 DENV whole genome (9 outliers and 1 recombinant from Philippines-AY496879)

RDP software: to be sure that non of the selected strains (180 seqs) were recombinants
Outline of the Classification Procedure:

Evolutionary model: HKY + gamma among site rate variation

If the bootstrap support is < 70%, the genotype is reported to be unassigned.
All source code is written in the Java 188 programming language. The software framework is freely available on GitHub 189 (http://github.com/rega-cev/rega-genotype) under the GNU General Public License v2. Also available at http://www.bioafrica.net/software.php
### Dengue, Zika & Chikungunya Viruses Typing Tool Results

You may bookmark this page to revisit results of this job (1013889988) later.

A whole genome typing will take approximately 30-60 seconds, please be patient.

Viral species can be identified with sequences longer than 200bp

#### Rega Assignment

<table>
<thead>
<tr>
<th>Rega Assignment</th>
<th>sequences</th>
<th>Percentage</th>
<th>Source</th>
<th>Legend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zika Virus</td>
<td>1</td>
<td>33,333</td>
<td>NCBI</td>
<td></td>
</tr>
<tr>
<td>Chikungunya</td>
<td>1</td>
<td>33,333</td>
<td>NCBI</td>
<td></td>
</tr>
<tr>
<td>Dengue Virus</td>
<td>1</td>
<td>33,333</td>
<td>NCBI</td>
<td></td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Download results: [Table (Excel format)](#) [Table (CSV format)](#) [Sequences (Fasta format)](#)
**Zika Virus Typing Tool Results**

You may bookmark this page to revisit results of this job (1481235664) later.

A whole genome typing will take approximately 60 seconds, please be patient.

<table>
<thead>
<tr>
<th>Name</th>
<th>Length</th>
<th>Virus</th>
<th>Serotype/Clade</th>
<th>Report</th>
<th>Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR815989_Brazil_2015</td>
<td>330</td>
<td>Zika Virus</td>
<td>Asian</td>
<td>Report</td>
<td></td>
</tr>
<tr>
<td>KU955594_Uganda_1947</td>
<td>630</td>
<td>Zika Virus</td>
<td>African</td>
<td>Report</td>
<td></td>
</tr>
</tbody>
</table>

Download results: [XML File](#) [Table (Excel format)](#) [Table (CSV format)](#) [Sequences (Fasta format)](#)

Developed by: **FIOCRUZ/Bahia, Brazil** (Vagner Fonseca, Marta Giovanetti, Maria Inês Restovic, Murilo Fraira, Luiz Alcantara), **KU Leuven, Belgium** (Kristof Theys, Pieter Libin, Liza Cuypers, Ana Abecasis, Anne-Mieke Vandamme), **University of Oxford, U.K.** (Nuno Faria and Oliver Pybus), **Evandro Chagas Institute, Pará, Brazil** (Marcio Roberto Teixeira Nunes), **CDC/OID/NCEZID** (Gilberto A. Santiago), **Emweb bvba, Belgium** (Koen Defoerche) and **Africa Centre/UKZN, South Africa** (Tulio de Oliveira).

**Contact:** Vagner Fonseca, Dr. Luiz Carlos Junior Alcantara and/or Prof. Tulio de Oliveira
Zika Virus Genotyping Details

Sequence Assignment

Name: KR815989_Brazil_2015  Length: 330

Virus assignment

Virus assignment: Zika Virus

Serotype/Clade and Genotype result

Serotype/Clade assignment: Asian

Supported with phylogenetic analysis and bootstrap 1000.0 (>= 700.0)

Genome region

Dengue and Zika Genomes

Your sequence starts at position 1560 and finishes at position 1890 relative to the KJ776791 reference sequence.
# Zika Virus Typing Tool Results

You may bookmark this page to revisit results of this job (1481235664) later.

A whole genome typing will take approximately 60 seconds, please be patient.

<table>
<thead>
<tr>
<th>Name</th>
<th>Length</th>
<th>Virus</th>
<th>Serotype/Clade</th>
<th>Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR815989_Brazil_2015</td>
<td>330</td>
<td>Zika Virus</td>
<td>Asian</td>
<td>Report</td>
</tr>
<tr>
<td>KU955594_Uganda_1947</td>
<td>630</td>
<td>Zika Virus</td>
<td>African</td>
<td>Report</td>
</tr>
</tbody>
</table>

Download results: [XML File](#) [Table (Excel format)](#) [Table (CSV format)](#) [Sequences (Fasta format)](#)

Developed by: FIOCRUZ/Bahia, Brazil (Vagner Fonseca, Marta Giovanetti, Maria Inês Restovic, Murilo Freire, Luiz Alcantara), KU Leuven, Belgium (Kristof Theys, Pieter Libin, Lize Cuypers, Ana Abecasis, Anne-Mieke Vandamme), University of Oxford, U.K. (Nuno Faria and Oliver Pybus), Evandro Chagas Institute, Pará, Brazil (Marcio Roberto Teixeira Nunes), CDC/OID/NCEZID (Gilberto A. Santiago), Emweb bvba, Belgium (Koen Deforce) and Africa Centre/UKZN, South Africa (Tulio de Oliveira).

Contact: Vagner Fonseca, Dr. Luiz Carlos Junior Alcantara and/or Prof. Tulio de Oliveira
Phylogenetic Analysis Details (Serotype)

- Assignment: Asian
- Bootstrap support: 1000.0, bootstrap inside 2888.0, bootstrap outside 0.0
- Download the alignment (NEXUS format, FASTA format)
- Phylogenetic Tree (export as PDF, NEXUS Format):

```
  KR815989 Brazil 2015
   |  JN860885 Cambodia 20
    |   KJ776791 FrenchPolyn
     |      KF993678 Canada 2013
      |          EU545988 Micronesia
       |               HQ234499 Malaysia 19
        |                   KF383118 Senegal 200
         |                       KF383117 Senegal 199
          |                           HQ234501 Senegal 198
           |                               KF383116 Senegal 196
            |                                 HQ234500 Nigeria 196
             |                                    AY632535 Uganda 1947
              |                                        KF383119 Senegal 200
               |                                             KF383121 SENEGAL
                |                                                 KF268950 Cent Afr Re
                 |                                                        KF268949 Cent Afr Re
                  |                                                            KF383115 Cent Afr Re
                   |                                                                KF383120 Senegal Fay
                    |                                                                      Spondweni virus SM6
```

- View the PAUP* Log file (Contains bootstrap values)
Sensitivity and Specificity of the Tool

To assess the accuracy of our automated method, we compared the genotype/serotype classification results:

• of all complete genome sequences and the subgenomic regions, to the golden standard ML and Bayesian classification methods (considering a bootstrap value >70% as a cut-off and a posterior probability > 0.9) → 99-100%.

• The products obtained by using non-specific sequencing methods (i.e. short sequence reads produced by Illumina or RNA-seq) were also classified according to genotype/serotype.
Using the Bootscanning method: to investigate the suitability of sub-genomic regions for genotyping

Sliding window = 1500 bp, moving steps = 100 bp. The light-grey rectangular area marks the env region.
Using the TreePuzzle software: to define the phylogenetic signal by likelihood mapping for sub-genomic regions

<table>
<thead>
<tr>
<th></th>
<th>DENV-sero1</th>
<th></th>
<th>DENV-sero2</th>
<th></th>
<th>DENV-sero3</th>
<th></th>
<th>DENV-sero4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Resolved</td>
<td>Unresolved</td>
<td>Resolved</td>
<td>Unresolved</td>
<td>Resolved</td>
<td>Unresolved</td>
<td>Resolved</td>
<td>Unresolved</td>
</tr>
<tr>
<td>C</td>
<td>54.2%</td>
<td>42.8%</td>
<td>68.2%</td>
<td>25.9%</td>
<td>86.2%</td>
<td>9.7%</td>
<td>78.5%</td>
<td>15.3%</td>
</tr>
<tr>
<td>E</td>
<td>93.1%</td>
<td>3.5%</td>
<td>95.4%</td>
<td>2.4%</td>
<td>97.0%</td>
<td>1.5%</td>
<td>93.8%</td>
<td>3.8%</td>
</tr>
<tr>
<td>M</td>
<td>73.3%</td>
<td>22.2%</td>
<td>87.4%</td>
<td>9.0%</td>
<td>88.3%</td>
<td>7.5%</td>
<td>84.2%</td>
<td>11.4%</td>
</tr>
<tr>
<td>NS1</td>
<td>92.3%</td>
<td>4.4%</td>
<td>89.5%</td>
<td>8.0%</td>
<td>96.6%</td>
<td>1.1%</td>
<td>95.6%</td>
<td>2.1%</td>
</tr>
<tr>
<td>NS2A</td>
<td>82.2%</td>
<td>15.8%</td>
<td>87.8%</td>
<td>9.3%</td>
<td>88.8%</td>
<td>8.0%</td>
<td>87.8%</td>
<td>8.4%</td>
</tr>
<tr>
<td>NS2B</td>
<td>79.0%</td>
<td>19.2%</td>
<td>83.1%</td>
<td>14.2%</td>
<td>89.3%</td>
<td>6.2%</td>
<td>82.6%</td>
<td>12.6%</td>
</tr>
<tr>
<td>NS3</td>
<td>94.6%</td>
<td>2.8%</td>
<td>94.4%</td>
<td>2.7%</td>
<td>97.4%</td>
<td>1.1%</td>
<td>95.4%</td>
<td>2.3%</td>
</tr>
<tr>
<td>NS4A</td>
<td>75.8%</td>
<td>18.1%</td>
<td>78.0%</td>
<td>19.9%</td>
<td>88.8%</td>
<td>8.3%</td>
<td>84.8%</td>
<td>9.8%</td>
</tr>
<tr>
<td>NS4B</td>
<td>87.2%</td>
<td>10.7%</td>
<td>90.3%</td>
<td>5.6%</td>
<td>92.0%</td>
<td>4.3%</td>
<td>83.6%</td>
<td>10.7%</td>
</tr>
<tr>
<td>NS5</td>
<td>95.1%</td>
<td>2.2%</td>
<td>94.1%</td>
<td>3.1%</td>
<td>98.5%</td>
<td>0.3%</td>
<td>97.1%</td>
<td>1.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ZIKV</th>
<th></th>
<th>CHIKV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Resolved</td>
<td>Unresolved</td>
<td>Resolved</td>
<td>Unresolved</td>
</tr>
<tr>
<td>C</td>
<td>78.0%</td>
<td>3.8%</td>
<td>NSP1</td>
<td>52.2%</td>
</tr>
<tr>
<td>E</td>
<td>94.9%</td>
<td>1.2%</td>
<td>NSP2</td>
<td>52.1%</td>
</tr>
<tr>
<td>M</td>
<td>81.2%</td>
<td>4.9%</td>
<td>NSP3</td>
<td>62.3%</td>
</tr>
<tr>
<td>NS1</td>
<td>91.9%</td>
<td>1.9%</td>
<td>NSP4</td>
<td>52.7%</td>
</tr>
<tr>
<td>NS2A</td>
<td>93.5%</td>
<td>2.3%</td>
<td>CAP</td>
<td>56.9%</td>
</tr>
<tr>
<td>NS2B</td>
<td>85.4%</td>
<td>1.1%</td>
<td>E3</td>
<td>58.2%</td>
</tr>
<tr>
<td>NS3</td>
<td>94.0%</td>
<td>1.7%</td>
<td>E2*</td>
<td>77.1%</td>
</tr>
<tr>
<td>NS4A</td>
<td>90.8%</td>
<td>3.7%</td>
<td>E1</td>
<td>53.1%</td>
</tr>
<tr>
<td>NS4B</td>
<td>93.2%</td>
<td>2.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS5</td>
<td>96.3%</td>
<td>2.4%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Whole-genome sequences → ART software (NGS reads simulator) → many Illumina NGS reads of 150bp (coverage of 20 reads/position). On the X-axis the nucleotide position in the genome is represented, while on the Y-axis the frequency of mapped reads is visualized. Two overlaid histograms are shown: the white histogram shows the number of reads that were available at a particular position, the grey histogram show the number of reads that could be assigned at that position. *Paired Illumina NGS pseudo reads were simulated from the Los Alamos HIV and HCV database, in order to verify that such reads would not produce false positive results.
http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/aedesviruses
http://www.bioafrica.net/software.php

Luiz C. J. Alcântara¹, Nuno R. Faria², Pieter J. K. Libin³,⁴, Marcio R. T. Nunes⁵, Vagner Fonseca¹, Maria I. Restovic¹, Murilo Freire¹, Marta Giovanetti¹,⁶, Kristof Theys⁴, Lize Cuypers⁴, Ann Nové⁴, Ewout Vanden Eynden⁴, Ana Abecasis⁷, Koen Deforce⁸, Gilberto A. Santiago⁹, Isadora C. de Siqueira¹, Janaina M. Vasconcelos³, Rivaldo Venâncio da Cunha¹⁰, Oliver G. Pybus², Anne-Mieke Vandamme⁴,⁷, Tulio de Oliveira¹,¹¹

1 Department?, Oswaldo Cruz Foundation, Salvador, Bahia, Brazil
2 Department of Zoology, Oxford University, Oxford, United Kingdom
3 Artificial Intelligence Lab, Department of Computer Science, Vrije Universiteit Brussel, Brussels, Belgium
4 KU Leuven - University of Leuven, Department of Microbiology and Immunology, Rega Institute for Medical Research, Clinical and Epidemiological Virology, Leuven, Belgium
5 Department?, Instituto Evandro Chagas, Ananindeua, Para, Brazil
6 Department?, University of Rome Tor Vergata, Rome, Italy
7 Center for Global Health and Tropical Medicine, Unidade de Microbiologia, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisbon, Portugal
8 EMWEB (private company), Herent, Belgium
9 Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, San Juan, Puerto Rico, United states of America
10 Oswaldo Cruz Foundation, Campo Grande, Mato Grosso do Sul, Brazil
11 Africa Centre for Population Health, Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa