*W634–W642* *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 29 May 2009*
*doi:10.1093/nar/gkp455*

# A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences

**Luiz Carlos Junior Alcantara[1], Sharon Cassol[2], Pieter Libin[3], Koen Deforche[4], Oliver G. Pybus[5], Marc Van Ranst[3], Bernardo Galvão-Castro[1], Anne-Mieke Vandamme[3] and Tulio de Oliveira[6,]***

[1]Laboratório Avançado de Saúde Publica, Centro de Pesquisas Gonçalo Moniz, Fundação Oswaldo Cruz, Brazil, [2]MRC Unit for Inflammation and Immunity, Department of Immunology, University of Pretoria and Tshwane Academic Division of the National Health Laboratory Service, Pretoria, South Africa, [3]Laboratory for Clinical and Epidemiological Virology, Katholieke Universiteit Leuven, BG, [4]MyBioData bvba, Rotselaar, Belgium, [5]Department of Zoology, University of Oxford, UK and [6]Africa 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 $10^9$ virions per day) and, in the case of HIV-1, a strong

*To whom correspondence should be addressed. Tel: +27 35 5650 7531; Fax: +27 35 550 7565; Email: tuliodna@gmail.com; tdeoliveira@africacentre.ac.za

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

propensity to undergo intra- and inter-subtype recombination (2). These properties, combined with the long-term persistence of these viruses in large numbers of people, provide tremendous scope for the evolution and spread of an extraordinary range of genetic variants. As a result, HCV has evolved into six major genotypes (numbered 1–6) and multiple subtypes (designated a, b, c, etc.) that differ in their distribution, transmission route and response to therapy (3). The main (M) group of HIV-1 has evolved into nine major subtypes (A–D, F–H, J and K), four sub-subtypes (A1 and A2, F1 and F2) and over 40 circulating inter-subtype recombinant forms (CRFs) that differ in their prevalence and global distribution (4). Whether these subtypes also differ in their transmissibility and sensitivity to antiretroviral therapy has not been clearly established. Recent studies suggest that HIV-1 C viruses may be more transmissible than other subtypes due to increased shedding and replication in the female genital tract (5,6). There is also evidence that drug-resistance pathways may differ between HIV-1 subtypes (7). An understanding of how HIV-1 and HCV evolve in relationship to the human immune system and in response to therapy is critical to the effective control of these viruses, both at individual and population levels. However, Both HIV and HCV exist as mixed populations in a patient and no genotyping tool will ever be able to type or subtype the full extent of sequence variation within a single patient.

Studies of the human T-lymphotropic virus type-1 (HTLV-1) from different regions and ethnic groups have revealed that this epidemic may be more homogeneous than that of HIV-1 and HCV. Six different genetic subtypes of HTLV-1 have been proposed, based on phylogenetic analyses of the viral LTR and *env* regions: a, or Cosmopolitan, which is distributed worldwide (8); b, from Central Africa (9); c, a highly divergent Melanesian strain (10); d, isolated from the Central African Republic, Cameroon and Gabon (11); e, isolated in a sample from an Efe pygmy in the Democratic Republic of Congo and f, from Gabon (12). When analyses are based solely on the LTR region, the most variable and appropriate genetic region for HTLV-1 subtyping, the Cosmopolitan subtype can be divided into five subgroups based on geographical distribution: Transcontinental (A), Japanese (B), West African/Caribbean (C), North African (D) and Black Peruvian (E) (7). HTLV-1 infection is endemic in Japan, the Caribbean Basin, as well as some South American and African regions, and only 2–5% of those infected develop a disease associated with HTLV-1 (13).

The relationship between genetic diversity and the epidemic spread of different viral subtypes (or genotypes) is not fully understood. In a few cases, genetic variation has been linked to differences in disease severity and (or) treatment outcome. Emerging evidence suggests that patients infected with the C genotype of Hepatitis B Virus (HBV) exhibit a poor response to interferon and lamuvidine when compared to patients infected with genotype B (14). It is also well known that each of the 15 genotypes of Human Papilloma Virus (HPV) act as independent infections that differ in their potential to cause cervical carcinoma (15).

Tools for studying the impact of genetic diversity on the biological properties, therapeutic response and epidemic potential of HIV-1/HCV, HTLV-1 and other viruses, remain a major challenge. Most methods used in the classification and genetic profiling of viral subtypes, including the Stanford HIV-Seq Program (http://hivdb.Stanford.edu), the Los Alamos Recombinant Identification Program (http://hivweb.lanl.gov/RIP/RIPsubmit.html) and the NCBI-genotyping Program (http://www.ncbi.nih.gov/projects/genotyping/) employ a similarity search tool to determine the genotype of a new query sequence. Similarity-based methods allow for the identification of recombinant viruses using similarity-scanning but they all require further confirmation using proper phylogenic methods. In contrast, the new genotyping tools described in this study and in a previous report (16) utilize a sliding window to generate multiple overlapping segments of a query sequence and its reference dataset. Separate phylogenetic trees are reconstructed for each segment, and the reference sequence with the highest bootstrap value is assigned to that segment of the query sequence. Processing of the genome in multiple segments along the length of the virus increases the accuracy and reliability of the results, especially when analyzing complex recombinants. In addition, in the case of recombinant viruses, separate phylogenetic confirmation of non-recombinant fragments will be required.

To date, we have constructed reference datasets for HIV-1, HIV-2, HBV, HCV, HTLV-1, HHV-8 and HPV. In contrast to other commonly used methods such as SIMPLOT, RIP and the NCBI-Genotyping Tool, which utilize a single reference sequence or a consensus reference sequence, our new genotyping tools use a set of carefully selected full-length reference genomes to represent each individual genotype. The use of multiple reference sequences enhances the consistency and reproducibility of the data and ensures that the phylogenies are not limited by a small number of inappropriate, or uninformative, reference strains.

## METHODS

### Selections of viral strains

Initial studies were designed to assess the ability of the reference strains (initially selected from strains curated at the RNA Virus Database (17) (http://virus.zoo.ox.ac.uk/rnavirusdb/), Los Alamos HIV (http://www.hiv.lanl.gov) and HCV Sequence Databases (http://www. hcv.lanl.gov) to accurately classify a set of well-classified (gold standard) genomic sequences. Individual NJ trees were constructed for each test genome together with its appropriate reference set. Phylogenetic analyses were performed separately on each complete HIV-1, HCV and HBV genome, as well as on sub-genomic regions of HTLV-1, HPV and HHV8. Test sequences in the 'gold standard' dataset were considered to be accurately classified if they clustered within a known genotype, or sub-genotype, with a bootstrap value >70%. Fragments as large as 1000 nt in length were successfully genotyped using our genotyping tools. Reference alignments of complete and sub-genomic gold standard sequences that gave a bootstrap value of >95% were deemed suitable for routine use (16).

As with all genotyping tools, the accuracy and consistency of the data is dependent on the selection of appropriate reference sequences. To overcome the limitations of other commonly used methods that employ a single reference sequence or a consensus reference sequence (SIMPLOT, RIP and NCBI-genotyping tools), we used sets of carefully selected, full-length viral genomes to represent each individual subtype and recombinant virus. The initial step in the selection of reference strains involved the screening of published data to identify highly divergent, but equidistant, genomes that were representative of the diversity within a given subtype or CRF. The selected sequences were then aligned, edited and subjected to phylogenetic analysis using NJ, Bayesian and ML methods (18–20). Sequences that gave similar topologies using all three tree construction methods were retained for further analysis of their sub-genomic regions. In this phase of the evaluation, the sub-genomic regions were assessed using consecutive windows of fixed, but increasing, sizes, ranging from 200 to 2000 nt. The process began with an initial window size of 200 nt and was repeated with subsequent windows until all segments of the genome were classified with a bootstrap value of ≥70%.

### Subtyping analyses

The subtyping tools described in this study were developed using Java programming and PHP scripts. These tools accept up to 1000 sequences at a time. In the first step of the analysis, the genomic region of each reference sequence (HXB2 for HIV-1, NC_003977 for HBV, 1a.COLONEL for HCV, ATK1 for HTLV-1, GK18 for HHV8 and NC_001356 for HPV) is identified using BLAST software. The second step involves the alignment of the query sequence with a complete reference dataset composed of all subtypes. Depending on the virus being analyzed, this alignment can include reference sequences for genotypes, subgroups, CRFs and (or) genera/species, (information available in Table 1). The final step involves the construction of a phylogenetic tree using Tamura-Nei or HKY distance methods with a gamma distribution of among site rate heterogeneity, as implement in PAUP* software (19).

The query sequence is then divided into small segments and a sliding window of 400 nt is moved along the sequence in 20 nt increments. Each segment of the query sequence and the reference datasets is subsequently analyzed for recombination using bootscanning methods. This involves the construction of a phylogenetic tree with bootstrapping, as implemented in PAUP*. A series of JAVA programming and PHP scripts are used to interpret the bootstrapping analysis and graphically plot the results using R software (http://www.r-project.org/). The evaluation of the phylogenetic signal (for sequences smaller than 800 bp) is performed using TreePuzzle software (16,21).

A series of PHP scripts are used to read the XML output format produced by the JAVA program and create HTML report pages. The batch report contains information on the sequence name, length, assigned subtype and subgroup, and an illustration of the virus' genome. Accessing the report link will take the user to a report generated for each submitted sequence. This report is composed of three areas, named 'sequence assignment', 'analysis details' and 'phylogenetic analyses'. The sequence assignment contains information on the sequence submitted (name and length), the classification assignment (genotype, subtype, or subgroup and bootstrap support >70%), a graphical representation of the viral genome showing the genomic region of the query sequence with the start and end positions related to the reference strain, and the motivation of the classification, based on the decision tree. The phylogenetic analysis section contains the tree in PDF and Nexus formats, the log file generated by PAUP with information on the model of evolution and its parameters, as well as the alignment used.

## RESULTS

### Strain selection

To select suitable reference strains and ensure that they cover the full spectrum of genetic variation, we: (i) screened published databases (such as the RNA Virus database, Los Alamos HIV and HCV Sequence databases) to identify highly divergent but equidistant genomes that were representative of the medium diversity within a given subtype, genotype or circular recombinant form (CRF); (ii) determined whether the selected reference sequences were recombinant or non-recombinant [reference sequences were classified as 'pure' genotypes/subtypes if the same genotype was assigned to all, or most, of its sub-genomic fragments in neighbor joining (NJ) and maximum likelihood (ML) phylogenies] (iii) Determined the sub-genomic regions that were adequate for virus subtyping/genotyping with a scanning technique (this process identified gene segments that were suitable for phylogenetic analysis and provided information on the minimal window size needed for accurate classification) and (iv) assessed the ability of the selected reference dataset to generate reliable and reproducible phylogenetic trees and to correctly classify 'gold standard' sequences.

The selected the reference strains for each virus are summarized in Table 1. The Table includes information on the reference datasets used for classification including the total number of reference sequences selected for each viral subtype/genotype, the average number of sequences selected per subtype/genotype, the genetic region most suitable for viral subtyping/genotyping and the minimum size required to obtain consistent results for HIV-1, HIV-2, HBV, HCV, HTLV-1, HHV-8 and HPV. This table shows, for example, that the HTLV-1 dataset is composed of 42 sequences, which represent six subtypes and five subgroups whereas the HCV dataset is composed of 57 sequences representing six genotypes and 30 subtypes. HTLV-1 genotyping should be done exclusively with the LTR region whereas HCV genotyping can be performed on the complete viral genome, excluding the UTR and NS4A region. More information on the

**Table 1.** Evaluation process and summary of reference datasets chosen for the virus-genotyping tools

| Organism | HIV-1 | HIV-2 | HTLV-1 | HBV | HCV | HPV | HHV8 |
|---|---|---|---|---|---|---|---|
| Number of sequences | 48 | 11 | 42 | 21 | 57 | 92 | 23 |
| Number of subtypes/genotypes | 9 subtypes (A–D,F–H, J–K)<br>4 sub-subtypes (A1–A2, F1–F2)<br>13 CRFs (01–08, 10–14) | 2 subtypes (A & B)<br>2 outgroups (SMM & RCM) | 6 subtypes (a–f)<br>5 subgroups (aA–aE) | 8 subtypes (A to H) | 6 genotypes (A to E)<br>30 subtypes (1a to 6p) | 20 generas (alpha to sigma)<br>48 species (1 to 15 in the generas) | 6 subtypes (A, A5, B to E) |
| Average sequences per subtype/genotype | 2.4 per subtype<br>2 per sub-subtype<br>2 per CRF | 2.5 per subtype<br>3 per outgroup | 6 per subtypes<br>5 per subgroups | 2.6 per subtype | 9.5 per genotype<br>1.9 per subtype | 4.6 per genera<br>1.9 per species | 3.8 per subtype |
| Complete genome/genetic region | CG | CG | LTR | CG | CG | L1 | K1 |
| Size | 9208 bp | 9421 bp | 725 bp | 3257 bp | 9525 bp | 1071 bp | 821 bp |
| Min. size of query sequence. | 500 bp | 600bp | 200 bp | 600 bp | 600 bp | 400 bp | 400 bp |
| Genetic sub-region best suited for genotyping | Gag, Pol, Env, Nef, Tat (with intron), Rev (with intron), Vpr, Vpu, Nef | Gag, Pol, Env, Nef, Tat (with intron), Rev (with intron), Vif, Vpx, Nef | LTR | E2,E4,E6,L1,L2 | C, E1, E2, P7, NS2, NS3, NS4B, NS5A, NS5B | L1 | K1 |
| Genetic sub-region not suitable for genotyping | LTR, Vif (why is that?) | LTR | *gag, pol, env, px* | E1, E7 | UTRs, NS4A | N/A | N/A |

This table shows the number of sequences used in the reference datasets, the number of subtypes/genotypes represented in the reference sequences, and the average number of sequences per subtypes/genotypes. In addition the table also give information on the genetic region used in the reference datasets (CG = complete genome), the size of the reference alignment and the minimum size of a query sequence to be subtyped/genotyped. Information is also given on which genetic sub-regions are most suitable for the classification of subtype/genotype and which ones should be avoided.

reference datasets can be requested from the authors. A list of appropriate reference sequences for each virus can be found on the bioafrica website (http://www.bioafrica.net/rega-genotype/html). It is important to note, that reference datasets may change over time with every update of the tool when new divergent sequences and/or subtypes/genotypes are identified.

## Accuracy of the gold reference datasets

The Virus-Genotyping Tools described in this report showed a high level of accuracy (>90%) when used to analyze gold standard datasets (i.e. published datasets that have been extensively classified by phylogenetic methods) constructed from complete (HIV-1, HIV-2, HCV and HBV) or full sub-genomic regions (HTLV-1, HPV, HHV8) [Table 2; ref. (16) for HIV-1]. The Tool had an accuracy of >90% when applied to sub-genomic segments of HIV-1, HIV-2 and HCV that were >500 bp in length. However, the Tool's accuracy decreased significantly (<50%) when applied to HIV-1, HIV-2, HCV, HBV sequences that were smaller than 300 bp, or to HPV, HHV8 and HTLV-1 sequences <200 bp (data not shown). Thus, as with all phylogentic methods, the accuracy of the Virus-Genotyping Tool is dependent on the size of the sequences that are to be subtyped/genotyped.

## Identification of recombinant viruses

Recombinant genomes were detected among our HIV-1, HIV-2, HCV and HBV datasets. The level of recombination was higher in HIV-1 and HBV genomes. In a previous published manuscript, we analyzed 3201 HIV-1 pol sequences (length ≈ 1000 bp) with the REGA HIV subtyping tool and estimated that up to 15% of HIV-1 infections in the UK were caused by CRFs or unique recombinant forms (22). In this study, we determined the percentage of recombinant viruses among 1454 publicly available complete HBV genomes (kindly compiled by Ted M.H. Mess from Erasmus Institute, Netherlands). We found evidence of recombination in 410 (28.19%) of these HBV sequences. The remaining 1044 (71.81%) sequences were clearly classifiable as one of the eight HBV subtypes (A–H) with an accuracy of 90.1% (Table 2). HBV was the most difficult virus to genotype due to its extremely high level of genetic variation together with its propensity to recombine and generate new genotypes have not yet been described. Caution should be used when interpreting HBV-genotyping results.

## Bootscanning support and graphical interpretation of recombinant data

Recombination analyses were performed using a combined scanning window and phylogenetic approach. Recombinant viruses are identified based on a variable referred to as 'bootscanning support'. A sequence is considered to be a potential recombinant if <90% of the windows analyzed do not represent a single dominant subtype. As an example, a sequence of 1000 bp will generate 13 windows of 400 bp with a step of 50 bp (w1 = 1–400, w2 = 50–450, w3 = 100–500, etc.). Each of these windows is then used for the construction of a NJ tree with

**Table 2.** Results of the virus'-genotyping tool

| Organism | HIV-1 | HIV-2 | HTLV-1 | HBV | HCV | HPV | HHV8 |
|---|---|---|---|---|---|---|---|
| Number of sequences | 108 | 28 | 678 | 1044 | 61 | 121 | 86 |
| Method subtyped | Los Alamos | Los Alamos | All database sequences | Manual phylogenetic | Los Alamos HCV database | Manual phylogenetic | Manual phylogenetic |
| Match with virus subtyping tool | 100% | 96.4% | 98.5% | 90.1% | 100% | 96.7% | 98.8% |
| Genetic region | Complete Genome | Complete genome | LTR | Complete genome | Complete Genome | L1 | K1 |
| Size | ≈9000 bp | ≈9000 bp | 152–725 bp | ≈3100 bp | ≈9500 bp | ≈1000 bp | ≈800 bp |

These results are related to the usage of gold standard reference databases (which have been well classified by a specialized sequence database or by detailed and manual phylogenetic analysis). This table displays the number of sequences used in each gold standard dataset, the method subtyped, accuracy (match with our tools), the genetic region and size of the query sequences. For HIV-1, see also ref. (13).

bootstrapping. If 100% of the windows (13/13) support a given subtype, independent of the value of the bootstrap, the bootscanning support is 100% and the sequence is considered to be a 'pure' subtype. On the other hand, if 10% of the windows differ from the most dominant (prevalent) subtype, the sequence is given a bootscan support is 90% and is assessed for recombination. Regions of the virus that differ from the dominant subtype are considered recombinants only if they are supported by bootstrap values >70.

Data generated from the analysis of a recombinant full-length HBV genome (Genbank acc. AM494716.1) is shown in Figure 1. The position of the query sequence is depicted in a genome picture and a color-coded figure is then constructed containing the results of each independent phylogenetic analysis. As shown, the HBV Subtyping Tool reported that sequence was unclassifiable (and flagged it with the text, 'check the bootscan') since only 78% of the windows supported the dominant subtype—subtype D. The graphic and genome images in Figure 1 clearly indicate that the sequence is a D/E recombinant, but they do not reveal whether the sequence is the result of a single or multiple independent recombination events. To obtain additional information and assess the limitations of our recombination identification methods, the results are presented in three different formats: (i) graphical interpretation of the recombination bootscanning analysis, (ii) a figure containing all of the most prevalent subtype/genotype groupings in a given window, and (iii) a second figure showing only those sequences with bootstrap values greater than 70% (bootscanning support with confidence). The latter figure (Figure 1c) provides a conservative estimation of the recombination events between HBV subtypes D and E (in this case one recombination event). Figure 1c does not show the exact location of recombination breakpoints or reveal the genomic regions where the classification is not supported by bootstrap values >70%. We recommend using the results from this figure to perform more advanced recombination analyses aimed at defining the exact breakpoint and number of recombination events.

To date, the tools described in this report have only been used to genotype recombinants of HIV-1. To avoid false positive results and over-interpretation of the data we restricted our classification to previously characterized recombinants strains described in established sequence databases (such as circular recombinant forms—CRFs—described in the Los Alamos HIV sequence database). To further ensure the accuracy of our recombinant data, we have applied two additional steps into our analyses: (i) NJ bootstrap tree with subtypes and CRFs; and (ii) bootscanning with the identified CRF. The cut-offs for classification of a query sequence remain the same, a bootstrap >70% and a bootscanning value > 90% for CRF reference sequences [additional info can be found at ref. (22) and at http://bioafrica.mrc.ac.za/rega-genotype/html/subtypedecisiontree.html].

## Resolution of a difficult HCV sequence

Application of the HCV-genotyping tool to a problematic genome from Equatorial Guinea (Genbank accession number AJ851228) confirmed that the sequence in question was not a recombinant but a potential new subtype within HCV genotype 1, as previously described by Bracho and colleagues (23). The sequence consistently clustered inside genotype 1 with a bootstrap of 100%. The position of the sequence in the tree was intermediary among the three known subtypes of HCV genotype 1 (a, b and c) (Figure 2). Recombination analysis with a window size of 500 bp and a step of 100 bp was performed using bootscanning (http://bioafrica.mrc.ac.za/rega-genotype/html/subtypinghcvSUB.html). Support for the analysis was 76.3% when no bootstrap confidence level was taken into account and 40.1% when a bootstrap support of 70% was used (bootscan support was calculated as the percentage of the most prevalent subtype, in this case, subtype 1b). Figure 2 shows a graphical representation of the results without bootscanning support. The data shows that 10.8% of the windows support subtype 1a, 76.3% support subtype 1b and 12.9% support subtype 1c. When the bootscan confidence was taken into account, none of the segments could be classified as subtype 1a, 40.1% were classifiable as 1b, 10% as 1c, while 58% remained unclassifiable, suggesting that >50% of the sequence was from an unknown HCV subtype. This information, in addition to the position of the sequence in the phylogenetic tree (Figure 2A), suggests that this sequence represents a new HCV subtype rather than a recombinant sequence composed of subtypes 1a, 1b and 1c.

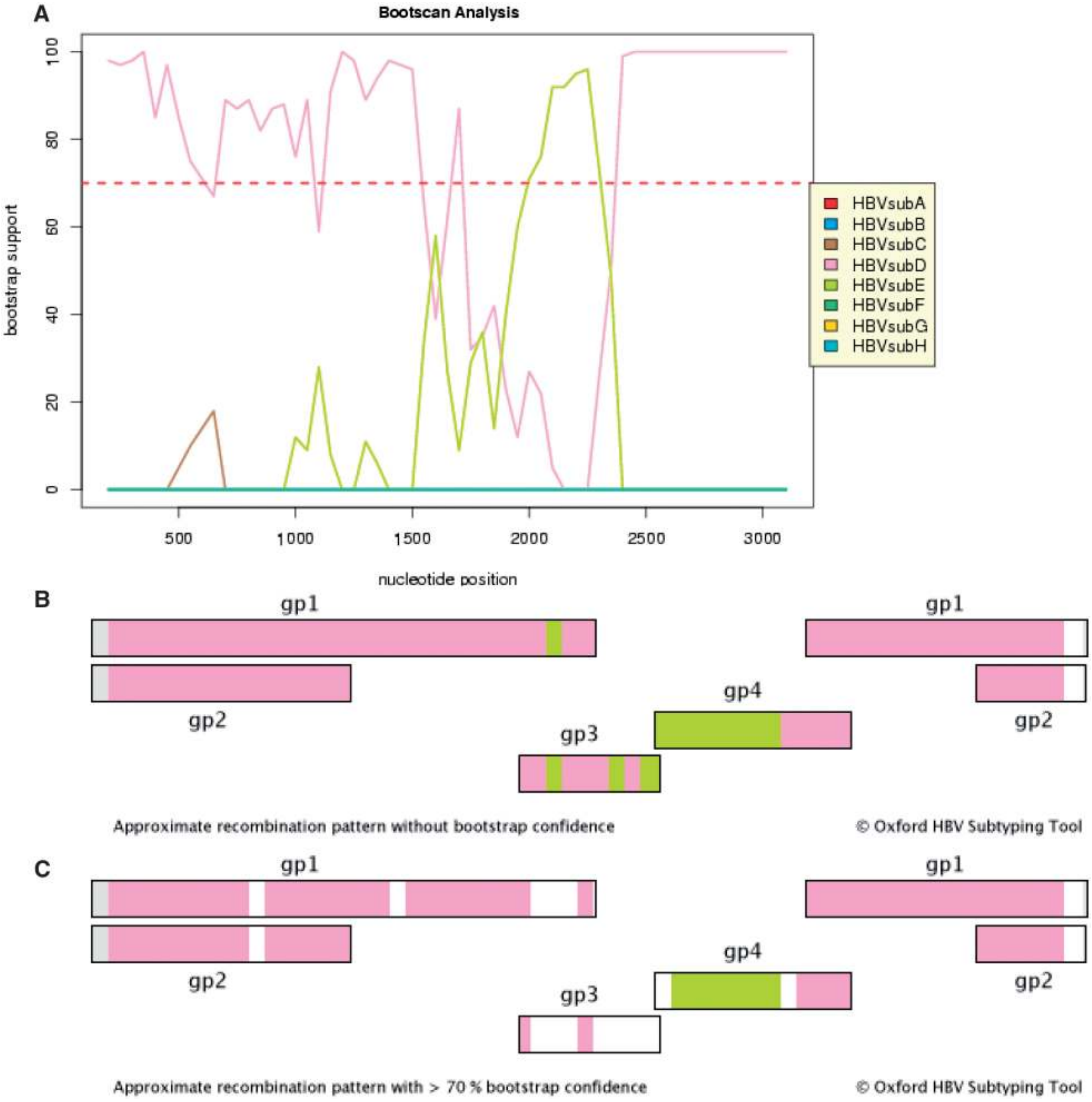

**Figure 1.** (**A**) Bootscanning results of an HBV complete genome recombinant sequence (acc. Number AM4947161). The *X*-axis represents the length of the sequence that is being analysed. The *Y*-axis represents the bootstrap support o the query sequence with subtype reference datasets. The color represents the different symbols. (**B**) Recombination profile not considering bootstrap confidence. (**C**) Recombination profile with >70% bootstrap confidence.

## DISCUSSION

The bioinformatics tools introduced in this report (i.e. the REGA HIV-1, HIV-2 and HPV; Oxford HCV and HBV; LASP HTLV-1 and BioAfrica HHV-8-genotyping tools) provide an accurate and robust framework for the classification of a wide range of viral pathogens. By analyzing sequential overlapping segments of a query sequence and its reference alignment, it is possible to construct phylogenetic trees representing each of the segments, conduct bootscanning analyses and draw statistically supported conclusions relating to a virus genotype and its recombinant or non-recombinant nature. The stringent cut-off for bootstrap (>70%) and bootscan analyses (>90%) greatly reduces the risk of misclassification and obtaining false-positive results. This enhances the quality of the data and makes it possible to screen large databases and select viruses of known genotypes with a high degree of confidence. For example, the REGA HIV-1-genotyping tool successfully classified 92.0% of 35 282 *pol* sequences collected as part of the UK and SPREAD drug resistance programs. This set of correctly identified sequences

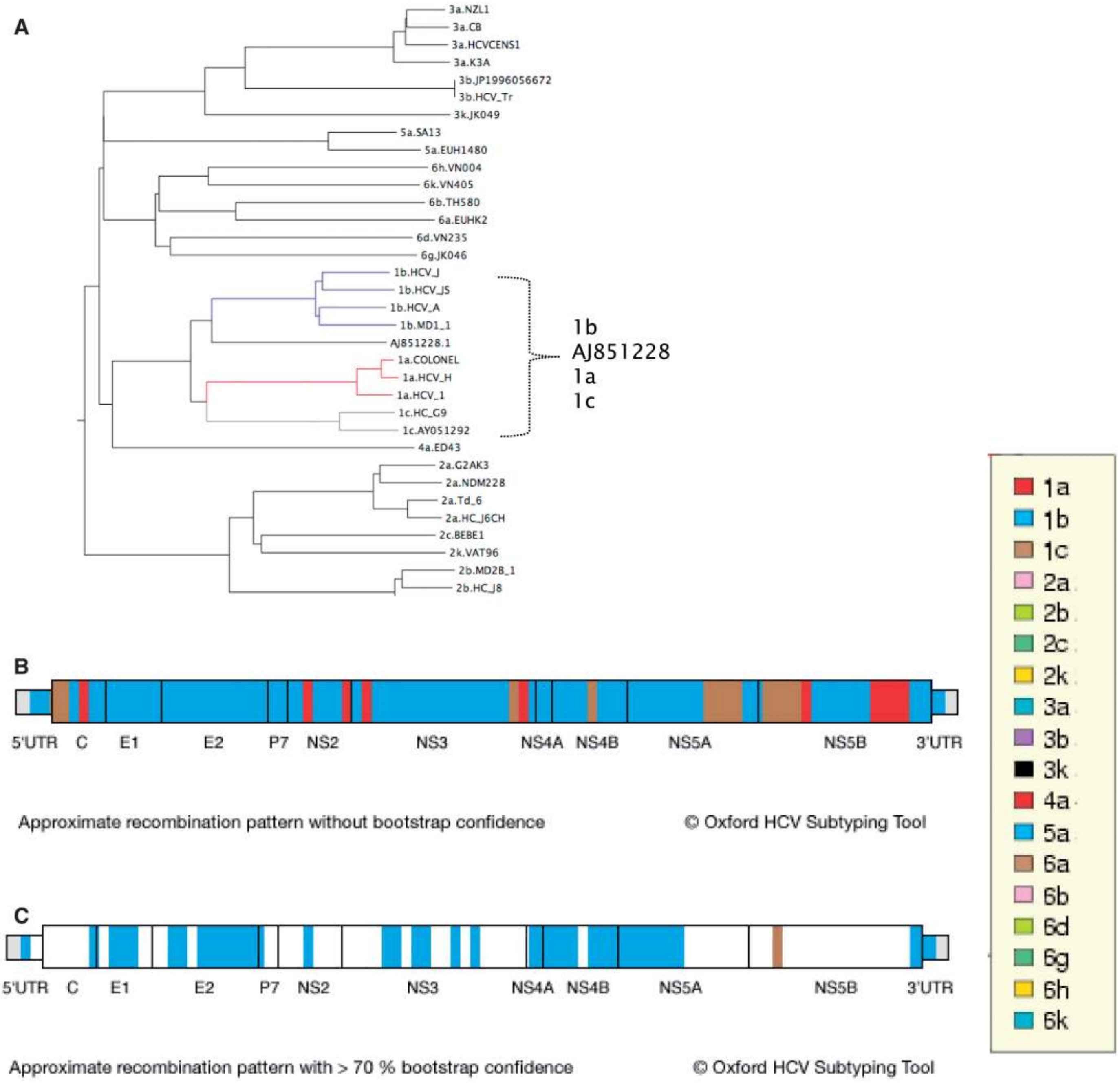

**Figure 2.** (**A**) Phylogenetic tree showing the location of the AJ851228 sequence. (**B**) Recombination profile without bootstrap confidence (top panel); recombination profile with >70% bootstrap confidence.

provided a more accurate estimate of drug resistance among different subtypes and CRFs compared to less stringent subtyping methods (22,24).

The failure of the REGA tool to classify 8.0% of HIV-1 sequences in the above-mentioned study may represent an advantage, rather than a disadvantage, since the majority of unclassified sequences represent highly divergent, newly recognized viral variants. The ability to rapidly screen large databases and identify new viral variants is important, not only for vaccine development, but also for the design of sensitive and specific diagnostic assays that detect all variants of a given pathogen. Once identified, these variants can be flagged for further study using more advanced and time-consuming models, such as the ones applied in the GARD and RDP software applications (25,26). The methods are designed to specifically discriminate between recombinant and non-recombinant viruses and to search for, and accurately localize, recombination breakpoints. These considerations are of paramount importance for the diagnosis of HIV-1 non-B subtypes in Africa, Europe and other regions of the world where non-B subtypes predominate, or are increasing in prevalence (24,27). Sequence diversity appears to emanate out of Africa, with recent data suggesting that up 18% of HIV-1 infections are caused by recombinant viruses (28). Diversity considerations are also of fundamental importance in the therapeutic setting, where subtype-specific resistance pathways have been described (7). Indeed, an increasing number of publications discussing HIV-1 drug resistance refer to this tool for subtyping. This discussion applies, not only to HIV-1, but also to HCV, HPV and other pathogens that exhibit subtype-specific variations in pathogenicity, virulence and (or) response to therapy.

Accurate genotyping, combined with carefully designed longitudinal studies, is also needed to better understand the epidemiological behaviour and epidemic potential of different viral variants, predict the future direction of a given pandemic and elucidate relationships between diversity, disease progression and escape from host immunity. An improved understanding of these factors is the key, not only to the development of effective vaccine and treatment strategies but also for long-term planning and policy-making, the efficient utilization of financial resources and the targeting of education and prevention programs to high-risk populations. In this context it is important to note that we are committed to update the genotyping/subtyping tools so that it accurately reflects the changing epidemic.

Phylogenetic analysis using overlapping gene segments and multiple reference strains provides a robust framework for the genotyping of a wide range of recombinant and non-recombinant viruses. The method is particularly well suited to the genetic classification of HIV-1, HCV and other pathogens with extensive and rapidly expanding databases. The reliability and consistency of the genotyping data is superior to similarity-based methods, especially with respect to the genotyping of complex recombinants. Thus, the virus-genotyping tools presented in this report represent a technological advance with widespread applications. The tools allow for the classification of up to 1000 sequences in a single analysis and can be used to address basic science, as well as epidemiologic and clinical research questions.

## FUNDING

Katholieke Universiteit Leuven (grant OT/04/43); Belgian Fonds voor Wetenschappelijk Onderzoek (F.W.O. nr G.0266.04); Virolab project (EU IST STREP Project 027446). T.O. is funded by the Wellcome Trust (082384/Z/07/Z) and European Union (SANTE 2007 147-790) and the Africa Centre for Health and Population Studies; S.C. is funded by the European Union (SANTE 2007 147-790) and NRF, South Africa. The Royal Society (to O.G.P.). Funding for open access charge: European Union (SANTE 2007 147-790).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Pybus,O.G., Charleston,M.A., Gupta,S., Rambaut,A., Holmes,E.C. and Harvey,P.H. (2001) The epidemic behavior of the hepatitis C virus. *Science*, **22**, 2323–2325.
2. Rambaut,A., Posada,D., Crandall,K.A. and Holmes,E.C. (2004) The causes and consequences of HIV evolution. *Nat. Rev. Genet.*, **5**, 52–61.
3. Simmonds,P., Bukh,J., Combet,C., Deléage,G., Enomoto,N., Feinstone,S., Halfon,P., Inchauspé,G., Kuiken,C., Maertens,G. *et al.* (2005) Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology*, **42**, 962–973.
4. Robertson,D.L., Anderson,J.P., Bradac,J.A., Carr,J.K., Foley,B., Funkhouser,R.K., Gao,F., Hahn,B.H., Kalish,M.L., Kuiken,C. *et al.* (2000) HIV-1 nomenclature proposal. *Science*, **7**, 55–56.
5. Walter,B.L., Armitage,A.E., Graham,S.C., de Oliveira,T., Skinhøj,P., Jones,E.Y., Stuart,D.I., McMichael,A.J., Chesebro,B. and Iversen,A.K. (2009) Functional characteristics of HIV-1 subtype C compatible with increased heterosexual transmissibility. *AIDS*, **23**, 1047–1057.
6. John-Stewart,G.C., Nduati,R.W., Rousseau,C.M., Mbori-Ngacha,D.A., Richardson,B.A., Rainwater,S., Panteleeff,D. and Overbaugh,J. (2005) Subtype C Is associated with increased vaginal shedding of HIV-1. *J. Infect. Dis.*, **192**, 492–496.
7. Camacho,R.J. and Vandamme,A.M. (2007) Antiretroviral resistance in different HIV-1 subtypes: impact on therapy outcomes and resistance testing interpretation. *Curr. Opin. HIV AIDS*, **7**, 123–129.
8. Alcantara,L.C.J., de Oliveira,T., Gordon,M., Pybus,O., Mascarenhas,R.E., Seixas,M.O., Gonçalves,M., Hlela,C., Cassol,S. and Galvão-Castro,B. (2006) Tracing the origin of Brazilian HTLV-1 as determined by analysis of host and viral genes. *AIDS*, **20**, 780–782.
9. Hahn,B.H, Shaw,G.M., Popovic,M., Lo Monico,A., Gallo,R.C. and Wong-Staal,F. (1984) Molecular cloning and analysis of a new variant of human T-cell leukemia virus (HTLV-Ib) from an African patient with adult T-cell leukemia-lymphoma. *Int. J. Cancer*, **34**, 613–618.
10. Gessain,A., Boeri,E., Yanagihara,R., Gallo,R.C. and Franchini,G. (1993) Complete nucleotide sequence of a highly divergent human T-cell leukemia (lymphotropic) virus type I (HTLV-I) variant from Melanesia: genetic and phylogenetic relationship to HTLV-I strains from other geographical regions. *J. Virol.*, **67**, 1015–1023.
11. Mahieux,R, Chappey,C., Georges-Courbot,M.C., Dubreuil,G., Mauclere,P., Georges,A. and Gessain,A. (1998) Simian T-cell lymphotropic virus type I from Mandrillus sphinx as a simian counterpart of human T-cell lymphotropic virus type I subtype d. *J. Virol.*, **72**, 10316–10322.
12. Salemi,M., Van Dooren,S., Audenaert,E., Delaporte,E., Goubau,P., Desmyter,J. and Vandamme,A.M. (1998) Two new human T-lymphotropic virus type I phylogenetic subtypes in seroindeterminates, a Mbuti pygmy and a Gabonese, have closes relatives among African STLV-I strains. *Virology*, **246**, 277–287.
13. Verdonck,K., Gonzalez,E., Van Dooren,S., Vandamme,AM., Vanham,G. and Gotuzzo,E. (2007) Human T-lymphotropic virus 1: recent knowledge about an ancient infection. *Lancet Infect. Dis.*, **7**, 266–281.
14. Akuta,N and Kumada,H. (2005) Influence of hepatitis B virus genotypes on the response to antiviral therapies. *J. Antimicrob. Chemother.*, **55**, 139–142.
15. Schiffman,M., Castle,P.E., Jeronimo,J., Rodriguez,A.C. and Wacholder,S. (2007) Human papillomavirus and cervical cancer. *Lancet*, **370**, 890–907.
16. 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. *et al.* (2005) An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, **21**, 3797–3800.
17. Belshaw,R., de Oliveira,T., Markowitz,S. and Rambaut,A. (2009) The RNA virus database. *Nucleic Acids Res.*, **37**, D431–D435.
18. Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
19. Wilgenbusch,J.C. and Swofford,D. (2003) Inferring evolutionary trees with PAUP*. *Curr. Protoc. Bioinformatics*, **6**, unit 6.4.
20. Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

21. Strimmer,K. and von Haeseler,A. (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl Acad. Sci. USA*, **94**, 6815–6919.
22. Gifford,R., de Oliveira,T., Rambaut,A., Myers,R.E., Gale,C.V., Dunn,D., Shafer,R., Vandamme,A.M., Kellam,P., Pillay,D. *et al.* (2006) Assessment of automated genotyping protocols as tools for surveillance of HIV-1 genetic diversity. *AIDS*, **20**, 1521–1529.
23. Bracho,M.A., Carrillo-Cruz,F.Y, Ortega,E., Moya,A. and González-Candelas,F. (2006) A new subtype of hepatitis C virus genotype 1: complete genome and phylogenetic relationships of an Equatorial Guinea isolate. *J. Gen. Virol.*, **87**, 1697–1702.
24. Gifford,R.J., de Oliveira,T., Rambaut,A., Pybus,O.G., Dunn,D., Vandamme,A.M., Kellam,P., Pillay,D. and UK Collaborative Group on HIV Drug Resistance., (2007) Phylogenetic surveillance of viral genetic diversity and the evolving molecular epidemiology of human immunodeficiency virus type 1. *J. Virol.*, **81**, 13050–13056.
25. Kosakovsky Pond,S.L., Posada,D., Gravenor,M.B., Woelk,C.H. and Frost,S.D. (2006) GARD: a genetic algorithm for recombination detection. *Bioinformatics*, **22**, 3096–3098.
26. Martin,D.P., Williamson,C. and Posada,D. (2005) RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics*, **21**, 260–262.
27. Salemi,M., Goodenow,M.M., Montieri,S., de Oliveira,T., Santoro,M.M., Beshkov,D., Alexiev,I., Elenkov,I., Yakimova,T., Varleva,T. *et al.* (2008) The HIV Type 1 Epidemic in Bulgaria Involves Multiple Subtypes and Is Sustained by Continuous Viral Inflow from West and East European Countries. *AIDS Res. Hum. Retroviruses*, **24**, 771–779.
28. Hemelaar,J., Gouws,E., Ghys,P.D. and Osmanov,S. (2006) Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS*, **20**, W13–W23.