
Downloaded from: http://researchonline.lshtm.ac.uk/3548904/

DOI: https://doi.org/10.1093/bioinformatics/btp134

Usage Guidelines:

Please refer to usage guidelines at https://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by/2.5/
Antiretroviral (ARV) therapy has greatly advanced the management of human immunodeficiency virus type 1 (HIV-1) infection. When used in combination, ARV drugs targeting the viral reverse transcriptase (RT) and protease (PR) activities can suppress HIV-1 replication to undetectable levels, leading to significant clinical benefit. However, a number of factors can lead to the emergence of drug-resistant virus strains. Once drug-resistant strains have emerged, they are ‘archived’ in resting white blood cells and can rapidly re-emerge if therapeutic regimens using drugs to which they are resistant are restarted. Expert panels recommend that, where possible, selection of ARV drug regimens should be guided by ‘genotypic’ screening in which viral drug resistance mutations (DRMs) are identified by population sequencing of the dominant HIV-1 strain in plasma. A number of software and web resources have been developed to support this procedure (Liu and Shafer, 2006). Drug-resistant viruses selected by treatment can be transmitted, potentially compromising options for first line therapy in untreated individuals (Kuritzkes et al., 2008). Surveillance of HIV-1 drug resistance (HIVDR) is therefore crucial to maintain the success of HIV-1 prevention efforts. However, variations in the methodologies used for surveillance of transmitted HIVDR, such as the specific DRMs taken as indicating transmitted resistance, have so far limited the potential to draw general conclusions from these studies. There is a widely recognized requirement for standardized protocols in this area, so that trends in HIVDR can be investigated through comparison between studies performed in distinct geographic regions and over time (Pillay, 2004; van de Vijver et al., 2007).

We recently published a list of standard surveillance DRMs (SDRMs), endorsed by the World Health Organization (WHO) for epidemiological surveillance of transmitted HIVDR (Bennett et al., 2008b; Shafer et al., 2007, 2008). Here we describe an online program, the calibrated population resistance (CPR) tool, providing a standardized framework for estimating transmitted HIVDR from population-sampled HIV-1 PR and RT sequence sets.

### 2 FUNCTIONALITY
The CPR program accepts FASTA-formatted HIV-1 PR and/or RT sequence data. Options to carry out genotyping (subtyping) and to estimate genotypic resistance to specific ARV drugs are provided. A profile alignment of the submitted sequence set is created by aligning each nucleotide sequence to a polypeptide reference sequence for the region of the HIV-1 genome encoding PR and RT (by default, a subtype B consensus sequence, available from http://hivdb.stanford.edu/). Mutations, deletions and insertions (defined as changes relative to the reference sequence) are recorded for each submitted sequence. The prevalence of individual mutations is calculated by dividing mutation frequency by the number of valid codons at the corresponding position in the alignment. CPR implements a standard approach to handling contingencies such as missing data (i.e. incomplete sequences) and the nucleotide ambiguities common in HIV-1 sequence data obtained through population sequencing of viral RNA. These procedures are described in the program release notes.

A list of DRMs (by default the most recent version of the SDRM list) is used to compute the prevalence of resistance to each of the three main classes of ARV drug: protease inhibitors (PIs), nucleoside RT inhibitors (NRTIs) and non-nucleoside RT inhibitors (NNRTIs). The prevalence of transmitted HIVDR to each drug class is estimated as the number of sequences containing any DRM specific to that drug class relative to the number of times the target gene is represented in the alignment.

Analysis generates a report that summarizes the input dataset in terms of drug resistance, genetic diversity and sequence quality. The CPR report includes a graphical overview of DRMs and resistance-associated mutations present in the input dataset, and a plot showing coverage across the target region (i.e. the PR and RT genes). If the option to perform genotypic estimation of resistance is selected, resistance scores (ranging from 1 (susceptible) to 5 (highly resistant)) are calculated.
The HIV drug resistance database (HIVDB) is used to develop a list of PR and RT amino acid variants that have been observed at a prevalence of $\geq 0.1\%$ in a database containing sequences from about 20,000 individuals (Rhee et al., 2003). Atypical mutations that have been reported less frequently and that are not known drug-resistant variants are highlighted in the report. APoBEC3G-mediated sequence editing (see, Holmes et al. (2007) for review) is detected using a subset of atypical mutations that typically occur in edited sequences (Gifford et al., 2008). Other sequence quality indicators, such as stop codons and frameshifts are also identified and listed in a quality analysis section of the report. A number of mutations have been described that are marginal with respect to their inclusion on the SDRM list, and an option is provided to highlight these ‘borderline/suspicious’ mutations in the report in addition to SDRMs. The CPR report shows the prevalence of individual mutations in the query dataset alongside their corresponding prevalence (stabilized by subtype) in sequences from untreated patients in HIVDB. This allows investigators to rapidly identify sequence polymorphisms that are disproportionally represented in query datasets, and to discriminate between subtype-specific polymorphisms, sequence quality problems and mutational markers of prior-drug selection pressure. Mutation lists used within the program are standardized and version-tracked, as it is expected that changes may occur as new information about drug resistance and viral polymerase becomes available.

The CPR tool is written in PERL and can readily be installed on computers running UNIX or LINUX operating systems. Alignments are constructed using LAP (Huang and Zhang, 1996). Viral subtypes are assigned using STAR (Myers et al., 2005). Genotypic estimation of resistance is performed using the Stanford SIERRA web service.

3 DISCUSSION

The CPR tool aims to promote consistency between epidemiological studies by providing investigators worldwide with ready access to a simple, standard protocol for genotypic estimation of transmitted HIVDR. Because the CPR tool is closely linked to HIVDB, it allows investigators to leverage the power of large quantities of published HIV-1 sequence data within their analyses. Additionally, by standardizing protocols for genotypic estimation of transmitted HIVDR, the CPR program can facilitate comparison between sequence datasets that cannot be shared due to legal or proprietary constraints. These include datasets collated by some of the largest national and international surveillance programs (Little et al., 2002; SPREAD programme, 2008; UK Collaborative Group on HIV Drug Resistance, 2007; Yerly et al., 2007).

In regions of the world with minimal health infrastructure and large numbers of HIV-1 infected individuals, management of ART is necessarily based on simplified, standard treatment protocols (Bennett et al., 2008a; Gilks et al., 2006). The WHO has developed a minimum-resource approach for surveillance of transmitted HIVDR to accompany the expansion of access to ART in these regions, based on routine genotypic screening in a representative subset of the HIV-infected, untreated population (Bennett et al., 2008b). Due to resource constraints, the number of individuals surveyed is likely to be small ($\leq 47$), and surveillance is likely to rely partly on archived and convenience samples (Bertagnolio et al., 2007). Since the threshold for implementing changes in ART policy is low (1–5% prevalence of transmitted HIVDR (Bennett et al., 2008b)) it is crucial that protocols deal accurately and consistently with the contingencies of sequence analysis. The quality control measures implemented in CPR will help investigators to identify artifacts in sequence datasets collected for surveillance purposes [e.g. spurious ‘drug resistance mutations’ introduced by APoBEC-mediated sequence editing (Gifford et al., 2008)] so that expensive and unnecessary changes in health policy may be avoided.

Although designed specifically for surveillance of HIVDR, we propose that the framework implemented in the CPR program represents a prototype for other areas of molecular epidemiology—in particular studies of microbial drug resistance—in which the primary unit of analysis is a population-sampled set of sequences.

**Conflict of Interest:** none declared.

**REFERENCES**


