Journal of Clinical Virology 65 (2015) 68–73

Contents lists available at ScienceDirect

Journal of Clinical Virology

journal homepage: www.elsevier.com/locate/jcv

Effects of sequence alterations on results from genotypic tropism testing

Alejandro Pironti[a,*], Saleta Sierra[b], Rolf Kaiser[b], Thomas Lengauer[a], Nico Pfeifer[a]

[a] Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Campus E 14, 66123 Saarbrücken,Germany
[b] Institute of Virology, University of Cologne, Furst- Puckler-Str. 56, 50935 Cologne, Germany

ARTICLE INFO

Article history:
Received 3 November 2014
Received in revised form 28 January 2015
Accepted 6 February 2015

Keywords:
HIV-1
Tropism
Genotypic tropism determination
Geno2pheno
Triplicate
Simulation of Sanger bulk-sequencing errors

ABSTRACT

Background: geno2pheno$_{[coreceptor]}$ is a bioinformatic method for genotypic tropism determination (GTD) which has been extensively validated.
Objectives: GTD can be affected by sequencing/base-calling variability and unreliable representation of minority populations in Sanger bulk sequencing. This study aims at quantifying the robustness of geno2pheno$_{[coreceptor]}$ with respect to these issues. GTD with a single amplification or in triplicate (henceforth singleton/triplicate) is considered.
Study Design: From a dataset containing 67,997HIV-1 V3 nucleotide sequences, two datasets simulating sequencing variability were created. Further two datasets were created to simulate unreliable representation of minority variants. After interpretation of all sequences with geno2pheno$_{[coreceptor]}$, probabilities of change of predicted tropism were calculated.
Results: geno2pheno$_{[coreceptor]}$ tends to report reduced false-positive rates (FPRs) when sequence alterations are present. Triplicate FPRs tend to be lower than singleton FPRs, resulting in a bias towards classifying viruses as X4-capable. Alterations introduced into nucleotide sequences by simulation change singleton predicted tropism with a probability $\leq 2\%$. Triplicate prediction lowers this probability for predicted X4 tropism, but raises it for predicted R5 tropism $\leq 6\%$. Simulated limited detection of minority variants in X4 sequences resulted in unchanged predicted tropism with probability above 90% as compared to probability above 98% with triplicate FPRs.
Conclusions: geno2pheno$_{[coreceptor]}$ proved to be robust when sequence alterations are present and when detectable minorities are missed by bulk sequencing. Changes in tropism prediction due to sequence alterations as well as triplicate prediction are much more likely to result in false X4-capable predictions than in false R5 predictions.

1. Background

The Human Immunodeficiency Virus type 1 (HIV-1) employs two host molecules in order to enter the host cell: the CD4 receptor and a coreceptor. In vivo, either of two coreceptors can be: CCR5 and CXCR4. The capability to use a certain coreceptor is called viral tropism. HIV-1 so-called R5 strains can only use CCR5. CXCR4-capable strains can use either CXCR4, exclusively (X4 strains), or both coreceptors (dual/mixed tropic viruses) [1]. Maraviroc (MVC) is an antiretroviral that inhibits HIV-1 entry into the cell by binding to CCR5, and is thus ineffective against X4-capable strains. Therefore, viral tropism determination must precede MVC prescription [2].

Tropism can be determined phenotypically or genotypically [3–6]. Phenotypic determination in cell cultures is expensive, time-consuming, and requires specialized labs. Furthermore, samples with viral loads up to 1000 cp/ml often yield indefinite results, although proviral DNA testing is performed also [7]. Genotypic tropism determination (GTD) requires genotyping the third hypervariable loop of the HIV-1 env gene (V3) with subsequent computer-based interpretation. Several methods for interpreting sequences in order to determine HIV-1 tropism have been

Abbreviations: HIV-1, human immunodeficiency virus type 1; MVC, maraviroc; GTD, genotypic tropism determination; FPR, false-positive rate; V3, third hypervariable loop; EP, electropherogram dataset; ME, manually-edited sequence; AS, automatic sequence; SEP, single-electropherogram sequence; indel, insertion or deletion; LA, Los-Alamos dataset; SA, Sanger-alteration dataset; SE, single-error dataset; M, mixture dataset; MS, mixture-sampling dataset; CPT, change in predicted tropism; S, singleton; T, triplicate; SD, standard deviation.

* Corresponding author. Tel.: +49 681 9325 3122; fax: +49 681 9325 3099.
E-mail address: apironti@mpi-inf.mpg.de (A. Pironti).

http://dx.doi.org/10.1016/j.jcv.2015.02.006
1386-6532/© 2015 Published by Elsevier B.V.

developed [6]. geno2pheno$_{[coreceptor]}$ [8] is an extensively validated bioinformatic method for GTD [3–6,9]; its use as an alternative to phenotypic determination is recommended by the European and the Austrian–German HIV-treatment guidelines [12–14].

geno2pheno$_{[coreceptor]}$ interprets V3 with a Support Vector Machine (SVM) trained on genotype–phenotype pairs. Geno2pheno$_{[coreceptor]}$ outputs the false-positive rate (FPR)[1] with X4-capable being defined as positive [8]. FPR dichotomization yields a (predicted) viral classification into X4-capable or R5. When the FPR is in a range where MVC antiviral action is considered possible, yet uncertain, the virus is classified as X4-capable. Alternatively, this intermediate FPR range can be explicitly labeled, as is customary for interpretation of drug resistance to other antiretroviral drugs. Thus, MVC administration with an FPR in the intermediate range could be made dependent on whether other therapy options co-exist, rather than excluding it altogether. Furthermore, the use of an intermediate category when determining tropism can also be appropriate for predictions with FPR-decreasing sequence alterations, especially when determining tropism in triplicate.

Establishment of the most suitable cutoff for FPR dichotomization has been a matter of substantial debate. Currently, there is no universally accepted cutoff.

2. Objectives

The input to geno2pheno$_{[coreceptor]}$ is aV3 sequence. Therefore, the quality of the predictions depends on the quality of these sequences. With Sanger bulk sequencing, the measured sequence is a consensus of the dominating strains in the viral population. Here, minorities comprising less than 10%-20% of the viral population are unreliably represented, due to the limits of the experimental technology. X4-capable minorities may render MVC ineffective. Therefore, some labs perform the amplification step of the sequencing procedure in triplicate [15] to increase the chances of detecting minorities. This analysis addresses two related, unresolved questions: (1) How robust is geno2pheno$_{[coreceptor]}$ with respect to sequencing/base-calling variability in terms of change of predicted tropism (CPT)? (2) What is the influence of undetected minority populations on the predictions of geno2pheno$_{[coreceptor]}$? Both issues are of critical importance for assessing the reliability of geno2pheno$_{[coreceptor]}$ for clinical purposes. This article aims at providing answers to both questions.

3. Study design

A dataset of 163,958HIV-1 V3 nucleotide sequences was downloaded from the Los Alamos National Laboratory (http://www.hiv.lanl.gov/) on September, 19th 2013. Nucleotide sequences with duplicate V3 regions were discarded, resulting in the Los-Alamos dataset (LA) comprising 67,997 nucleotide sequences. Subtypes in LA were determined with Comet [16]. LA was used to create further datasets by altering its sequences *in silico*.

The probability of each type of sequence variation (base exchanges, ambiguities, insertions/deletions [indels]) at each sequence position resulting from differential primer specificity, and misincorporation/indel events during the RT-PCR amplification and sequencing reactions was determined using 164 clinical blood samples obtained from the Department of Clinical Virology of the University of Cologne, as described in Section Supplementary methods. This knowledge was used to generate the Sanger-alteration dataset (SA). The dataset contains 10 *in-silico* generated variants of each LA sequence. Alterations in these *in-silico* variants may be present at any of the 105 nucleotide positions.

The single-error dataset (SE) was created by generating sequences from each LA sequence by systematically exchanging every nucleotide in V3 by each of the 15 possible definite and ambiguous bases, independently of their probability of occurrence. Thus, from each sequence in LA, all possible sequences diverging by one definite or ambiguous base were generated.

The mixture dataset (M) was created from sequences in LA containing ambiguities (excluding N). Ambiguities in each of these sequences were combinatorially resolved into all possible sequence alternatives without ambiguities. To avoid combinatorial explosion, sequences that would result in more than 20,000 derived sequences were excluded from this procedure.

The mixture-sampling dataset (MS) was created to simulate a scenario in which sequencing depth is insufficient to resolve all sequence variants in the sample. From each sequence group in M derived from the same LA sequence, a certain proportion of sequences was extracted at random by uniform sampling without replacement, and a new sequence was created by retaining positions that are identical among the sequences in the subset and representing differential positions with the corresponding ambiguities. Proportions represent sequencing depth and ranged from 1% to 100% in steps of 2%; each sequence group was sampled 3 × 100 times (100 repetitions that allow for triplicate FPRs).

Sequences in LA, SA, SE, M, and MS were interpreted with geno2pheno$_{[coreceptor]}$. For each sequence, the FPR shift was calculated as the difference between the FPR of the altered sequence and that of its unaltered counterpart in LA. When we consider *n* FPRs from variability-simulation replicates on the same sequence, we call *singleton FPR* the FPR obtained with the first sequence. For all further sequences, we take the minimum FPR among the first n and call it nth replicate FPR. The 3rd replicate FPR is also called triplicate FPR. geno2pheno$_{[coreceptor]}$‘s FPR was used to determine coreceptor tropism as X4-capbable or R5. Four different FPR cutoff sets were used for tropism determination:

- {5, 10}: FPR < 5 ⇒ X4-capable, 5 ≤ FPR < 10 ⇒ Intermediate, FPR ≥ 10 ⇒ R5
- {5, 15}: FPR < 5 ⇒ X4-capable, 5 ≤ FPR < 15 ⇒ Intermediate, FPR ≥ 15 ⇒ R5 [14]
- {10}: FPR < 10 ⇒ X4-capable, FPR ≥ 10 ⇒ R5 [13]
- {20}: FPR < 20 ⇒ X4-capable, FPR ≥ 20 ⇒ R5 [13]

According to Austrian–German treatment guidelines, MVC can be effective when a tropism prediction is labeled intermediate albeit with much less certainty than for R5 variants [14].

The probability the tropism predicted by geno2pheno$_{[coreceptor]}$ changed due to the introduced sequence alterations was calculated by sample counting as $P(T_{A,C} \mid T_{U,C})$, with $T_{A,C}$ denoting the tropism of the altered sequences as determined with cutoff set *C*, and $T_{U,C}$ denoting the tropism of the unaltered sequences as determined with cutoff set *C*. The reference sequence used to number V3 nucleotide positions is consensus B (105 nucleotides), which is the reference used by geno2pheno$_{[coreceptor]}$.

4. Results

Alteration rates estimated for generation of SA (Section Supplementary methods) are shown in Supplementary Tables 1 and 2.

The FPR distribution in LA is illustrated in Supplementary Fig. 1. The numbers of strains by subtype are tabulated in Table 1. In LA, 0.24% of the bases are ambiguous, while 99.76% of the bases are definite.

[1] More accurately: it outputs the smallest FPR, of an SVM-classifier that classifies the sequence under inspection as X4-capable.