Geno2pheno[454]: A Web Server for the Prediction of HIV-1 Coreceptor Usage from Next-Generation Sequencing Data

Alexander Thielen^{a, b} Thomas Lengauer^a

^aMax Planck Institute for Informatics, Saarbrücken, and ^bInstitute of Immunology and Genetics, Kaiserslautern, Germany

Key Words

HIV • Coreceptor usage • Tropism • 454 sequencing • Genotype

Abstract

Inferring HIV-1 coreceptor usage from a genotype is becoming more and more important for the appropriate treatment of long-term patients. While results are already encouraging where standard bulk-nucleic acid sequencing methods are used, they are limited with respect to the detection of minor variants. In contrast, next-generation sequencing methods (ultradeep sequencing, pyrosequencing) are capable of sequencing virus quasispecies at very low quantities. However, as well as being very expensive, these methods generate vast amounts of data such that sequence analysis has to be automated by computer assistance. Here, we describe the geno2pheno[454] system which handles all processing and prediction steps involved in the prediction of coreceptor usage from massively parallel sequencing data. The system is split into a JAVA preprocessor which is run locally on the client side and a Web server which generates the prediction results. Predictions are based on the same prediction method as used in the geno2pheno[coreceptor] tool.

Copyright © 2012 S. Karger AG, Basel

Introduction

Attachment and subsequent infection of cells by HIV-1 is mediated by 2 cellular receptors: the CD4 receptor and a so-called coreceptor. In vivo, the coreceptor is one of the 2 chemokine receptors CCR5 and CXCR4. Viruses infecting a cell via the CCR5 receptor are termed R5 viruses, whereas X4 viruses can use only the CXCR4 receptor. Strains that are able to infect cells through both of these receptors are called dual-tropic.

Determination of coreceptor usage is mandatory in clinical routine before the administration of coreceptor antagonists such as maraviroc (Celsentri, ViiV Healthcare). This drug competitively binds to the CCR5 receptor, thereby inhibiting viral entry of R5 viruses; however, the drug is not effective against viruses infecting cells via CXCR4.

Coreceptor usage can be determined with phenotypic assays (e.g. Monogram Trofile tropism assay [1]) as well as genotypic approaches [2]. Prediction of HIV-1 coreceptor usage from genotype has recently become the standard method in most countries in Europe. Such genotypic testing is recommended in the European guidelines on the clinical management of HIV-1 tropism testing before prescribing a chemokine receptor antagonist [3].

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2012 S. Karger AG, Basel
0300–5526/12/0552–0113\$38.00/0

Accessible online at:
www.karger.com/int

Alexander Thielen
Max Planck Institute for Informatics
Campus E1 4
DE–66123 Saarbrücken (Germany)
Tel. +49 681 9325 3008 or +49 631 316 7023, E-Mail athielen@mpi-inf.mpg.de

Nucleic acid sequences used in genotypic approaches are generally determined by applying conventional bulk sequencing methods. However, although the effectiveness of genotypic approaches compared to the phenotypic Trofile assay in predicting clinical outcome has been demonstrated [4], it is also known that these methods lack the power of detecting minor variants [5, 6]. The limit of detection is usually considered to be at 15–20% of the viral population, i.e. variants occurring less frequently will be missed by standard bulk sequencing technologies. In contrast, phenotypic assays can reliably detect minorities of less than 5% [see e.g. 1]. These minor variants need not, but can, have an impact on clinical outcome when a coreceptor antagonist is administered. Another drawback of genotypic methods based on classic Sanger sequencing is that this sequencing method generates a consensus sequence of the whole viral quasispecies, thereby losing haplotype information of the individual strains. Two mutations appearing in the bulk consensus sequence at different sequence positions are therefore not necessarily located on the same viral strain; this can complicate the interpretation of these sequences.

In contrast to standard bulk sequencing, next-generation sequencing methods can reliably determine minorities at very low quantities and are thus sometimes also referred to as ultradeep sequencing. In addition, the sequenced reads are of a clonal nature, such that the haplotype structure of a viral quasispecies is retained.

The drawback of this method is the amount of data generated. With thousands of viruses sequenced in parallel, the process of interpretation has to be automated. We have therefore developed the geno2pheno[454] Web system to predict HIV-1 coreceptor usage from sequence data generated with the 454 pyrosequencing technology (454 Life Sciences/Roche). The 454 technology has been used in several studies [e.g. 5, 6] as it is capable of covering the whole V3 loop within a single read, whereas other next generation sequencing methods still do not reach this read length.

Materials and Methods

Overview

The geno2pheno[454] Web system for predicting HIV-1 coreceptor usage from massively parallel sequencing data is split into 2 parts: a preprocessor that is run locally on the client and the software on the geno2pheno server that calculates the predictions and provides different forms of data output. The reason for this division of the software is that the data generated by the sequencer are too voluminous to be uploaded to the Web server via the Internet. Thus, in the preprocessing step computationally expen-

Fig. 1. Graphical user interface of the preprocessor.

sive tasks on the complete data volume, such as the identification of V3 loops and their alignment against a set of reference sequences, are performed on the client side and the resulting data are compressed. The preprocessor generates a file, much smaller in size than the raw data, which is then sent to the Web server for processing.

Preprocessing

The preprocessor (fig. 1) can be downloaded from the geno2pheno[454] website. It is a JAVA program and therefore independent of the user's operating system. The program is capable of loading input files in the standard flowgram format (.sff) as well as FASTA-like formats (.fna, .fas or .fasta). In addition, if provided with a FASTA file containing the multiplex identifiers (MIDs) used in the sequencing run (FASTA header = sample name, sequence = MID), the program demultiplexes the data as well and groups the sequences into the different samples sequenced. The preprocessor facilitates choosing a name for the output file (default: default.g2p) which is later submitted to the Web server.

Workflow of the Preprocessor

The preprocessing program processes every read as follows: first, the sequence is translated into all 6 reading frames (3 forward and 3 reverse) and checked against an internal lookup database of already-aligned V3 loops. This database is built up during preprocessing. If the lookup database contains a V3 loop that is found in 1 of the 6 reading frames, then the information stored in the database is used and the next read is processed. Otherwise, all 6 frames have to be searched for typical motifs of the V3 loop and subsequently the identified loop is profile-aligned against a set of reference sequences.

A V3 loop has to pass several quality checks: it has to start and end with typical V3 motifs (e.g. CTR and AHC), the length of the sequence has to be between 30 and 40 amino acids and, last but not least, the alignment score has to be above a certain threshold. This threshold can be changed by selecting a 'percentile cutoff' in the preprocessor. The percentile score displays how many reads in a set of about 50,000 sequences found in public databases received an alignment score below the respective threshold when being aligned to the reference sequence profile. Thus, the lower the percentile score, the higher the probability that, in fact, the sequence is a V3 loop.

If a sequence is accepted as a V3 loop it is added to the lookup database, together with its alignment to the reference sequence profile. In contrast, if no frame was identified as a V3 loop, the one with the lowest percentile is stored in a separate database that is not used for lookup. These reads are also stored in the output file, allowing the user to analyze or filter them out later.

Fig. 2. Output page. The upper part lists some general information about the run. The table below displays the results stratified for each multiplexed sample. The right-most column links to additional graphics describing the prediction results.

Finally, the identified V3 loops are encoded in binary and stored in the output file. Each variant is stored only once, together with its percentile score and the multiplicity with which it appears in the sequencing data.

The Website Processor

The website processor can be accessed at www.genafor.org under the menu item 'Services'. The input page offers the download of the preprocessor and related files as well as the submission of the preprocessed data file. For prediction, one can select a percentile cutoff which filters out sequences assumed not to be V3 loops. These reads will still appear in the 'detailed results' section on the output page (fig. 2), but they will be excluded from all additional analyses and plots.

The first part of the output page displays general information on the processed data set such as the name of the uploaded file, the selected percentile cutoff, the version of the preprocessor and how many reads were sequenced in total. The number of reads containing a V3 loop and information on the used MIDe are also presented.

The second part is a table showing the results for the different MIDe individually. The website generates the predictions with the same software routine as used in the `geno2pheno[coreceptor]` tool [7]. It displays the sample name and the used MIDe sequence for each MIDe. Furthermore, the total number of reads for this MIDe, the number of reads that passed the quality control, and the number of different V3 amino acid sequences are listed. Next, the fraction of viruses predicted as X4 is displayed for different false positive rate cutoffs. Finally, more detailed results and additional plots can be found by following the links in the right-hand columns of the table. In figure 3, one of these plots is shown describ-

Fig. 3. Distribution of prediction scores within 1 sample. Every dot in this plot represents the prediction result (y-axis) from 1 individual read in the sequenced sample. Scores were sorted such that the x-axis can be regarded as a measure of how many reads were predicted as 'X4'. FPR = False positive rate.

Fig. 4. Coreceptor usage and homogeneity of a viral quasispecies. Two dominant variants can be seen in this example. The x-axis shows the alignment score, the y-axis describes the predicted false positive rate and the z-axis represents the frequencies of these variants.

ing the distribution of prediction scores. Every dot in this graph describes the prediction result from one individual read in the sequenced sample.

Another representation of the viral quasispecies is depicted in figure 4. One axis describes the predicted false positive rate while the other represents the alignment score reflecting the diversity of the population. The z-axis presents the number of reads found with these properties.

All results are downloadable as png-graphic files or csv files and a link to a specific URL is provided which stores the result page for about a week. Hence, the data do not have to be processed over again.

Discussion

In this report we present the geno2pheno[454] Web system for prediction of coreceptor usage from massively parallel sequencing data. The system combines a client-based data preprocessing pipeline with the well-known geno2pheno[coreceptor] tool on the server side.

This system handles all processing steps starting with the raw data and reaching to the prediction outcomes including demultiplexing, sequence alignment, data compression, prediction of coreceptor usage and reporting the results in different forms. Moreover, it works efficiently as it processes each individual V3 loop variant just once, independently of its multiplicity in the data set. Thus, the diversity of the sequenced reads has a major

impact on the required computing resources. Diverse sequence sets require more computing resources than those with few sequence variants.

A computational bottleneck is the memory requirement for the preprocessor. Dependent on the number of reads generated by the sequencer, the program requires much memory. When tested with GS FLX data containing up to 200,000 reads with an average read length of about 350 bp, the program required 1–2 GB of memory. More complex runs with the Titanium chemistry and several different samples needed up to 4 GB of RAM. When provided with enough memory, preprocessing usually completes in 1–4 h; otherwise the system has to swap memory and therefore runs for much longer.

So far, setting the percentile cutoff is a parameter-sensitive procedure. The parameter should thus only be changed if really necessary. If the cutoff is too low, reads with higher percentile scores will not be added to the lookup database. Hence, when these variants occur again in the data set, they will also have to be reprocessed and realigned. On the other hand, if the cutoff is too high, imperfect V3 loops are added to the database. Nevertheless, changing the parameter may be required occasionally (sometimes with rare subtypes) if the data contain strange sequences that are untypical V3 loops.

Altogether we have developed a tool capable of significantly simplifying the prediction of coreceptor usage

from 454 data. The system is open to everyone and free of charge. It has been thoroughly tested and used in several research settings. As deep-sequencing technology is gaining relevance in the treatment of HIV-immunosuppressed patients we expect this software could be a useful tool for interpreting this new kind of data.

Acknowledgments

The authors would like to thank all members of the EuResist and Arevir consortia, especially André Altmann and Martin Däumer, for critical discussions throughout this work. Furthermore, we would like to thank Conan K. Woods (BC Center for Excellence, Vancouver, B.C., Canada) and Daniel Struck (Laboratory of Retrovirology, CRP-Santé, Luxembourg) for testing the system.

References

- Whitcomb JM, Huang W, Fransen S, Limoli K, Toma J, Wrin T, Chappey C, Kiss LD, Paxinos EE, Petropoulos CJ: Development and characterization of a novel single-cycle recombinant-virus assay to determine human immunodeficiency virus type 1 coreceptor tropism. *Antimicrob Agents Chemother* 2007;51:566–575.
- Saliou A, Delobel P, Dubois M, Nicot F, Raymond S, Calvez V, Masquelier B, Izopet J: Concordance between two phenotypic assays and ultra-deep pyrosequencing for determining HIV-1 tropism. *Antimicrob Agents Chemother* 2011;55:2831–2836.
- Vandekerckhove L, Wensing A, Kaiser R, Brun-Vezinet F, Clotet B, De Luca A, Dressler S, Garcia F, Geretti A, Klimkait T, et al: European guidelines on the clinical management of HIV-1 tropism testing. *Lancet Infect Dis* 2011;11:394–407.
- McGovern RA, Thielen A, Mo T, Dong W, Woods CK, Chapman D, Lewis M, James I, Heera J, Valdez H, et al: Population-based V3 genotypic tropism assay: a retrospective analysis using screening samples from the A4001029 and MOTIVATE studies. *AIDS* 2010;24:2517–2525.
- Däumer M, Kaiser R, Klein R, Lengauer T, Thiele B, Thielen A: Genotypic tropism testing by massively parallel sequencing: qualitative and quantitative analysis. *BMC Med Inform Decis Mak* 2011;11:30.
- Swenson LC, Mo T, Dong WW, Zhong X, Woods CK, Jensen MA, Thielen A, Chapman D, Lewis M, James I, et al: Deep sequencing to infer HIV-1 co-receptor usage: application to three clinical trials of maraviroc in treatment-experienced patients. *J Infect Dis* 2011; 203:237–245.
- Lengauer T, Sander O, Sierra S, Thielen A, Kaiser R: Bioinformatics prediction of HIV coreceptor usage. *Nat Biotechnol* 2007;25: 1407–1410.