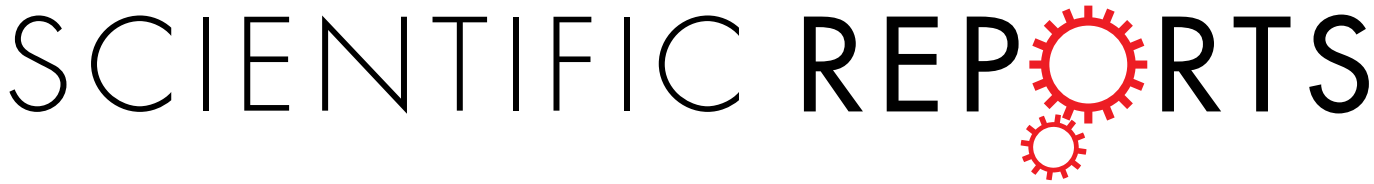


OPEN

Position-specific automated processing of V3 *env* ultra-deep pyrosequencing data for predicting HIV-1 tropism

Received: 17 February 2015
Accepted: 22 October 2015
Published: 20 November 2015

Nicolas Jeanne[1,*], Adrien Saliou[1,*], Romain Carcenac[1], Caroline Lefebvre[1], Martine Dubois[1,2], Michelle Cazabat[1,2], Florence Nicot[1,2], Claire Loiseau[2], Stéphanie Raymond[1,2,3], Jacques Izopet[1,2,3] & Pierre Delobel[2,3,4]

HIV-1 coreceptor usage must be accurately determined before starting CCR5 antagonist-based treatment as the presence of undetected minor CXCR4-using variants can cause subsequent virological failure. Ultra-deep pyrosequencing of HIV-1 V3 *env* allows to detect low levels of CXCR4-using variants that current genotypic approaches miss. However, the computation of the mass of sequence data and the need to identify true minor variants while excluding artifactual sequences generated during amplification and ultra-deep pyrosequencing is rate-limiting. Arbitrary fixed cut-offs below which minor variants are discarded are currently used but the errors generated during ultra-deep pyrosequencing are sequence-dependant rather than random. We have developed an automated processing of HIV-1 V3 *env* ultra-deep pyrosequencing data that uses biological filters to discard artifactual or non-functional V3 sequences followed by statistical filters to determine position-specific sensitivity thresholds, rather than arbitrary fixed cut-offs. It allows to retain authentic sequences with point mutations at V3 positions of interest and discard artifactual ones with accurate sensitivity thresholds.

Human immunodeficiency virus type 1 enters CD4-expressing cells using one or both of the host cell coreceptors, CCR5 and CXCR4[1–3]. Virus strains that specifically use CCR5 or CXCR4 are termed R5 or X4 variants, while those that use both coreceptors are termed dual/mixed variants (D/M)[4].

Maraviroc is the first CCR5 antagonist approved for treating HIV-1 infections[5]. But the HIV-1 coreceptor usage must be determined to establish that a patient is not harboring CXCR4-using viruses and is thus eligible for CCR5 antagonist treatment[6,7].

Recombinant virus phenotypic entry assays are now considered to be the gold-standard for determining HIV-1 tropism[8–13]. These assays can detect minor CXCR4-using variants down to 0.3–0.5% of the virus population[8,14]. However, their routine use is hampered by technical and cost limitations. Simple alternative genotypic approaches have been developed to infer virus tropism from the V3 *env* amino acid sequence[15–19]. Particularly, the presence of basic residues at V3 positions 11 and/or 25 and an increased net electrostatic charge of V3 have been associated with CXCR4 usage[15,20,21]. Genotypic algorithms based on the V3 *env* sequence perform well for predicting virus tropism when they are used at a clonal level[21]. However, direct sequencing of bulk PCR products of V3 *env* at a population level cannot detect minor CXCR4-using viruses that account for less than about 20% of the quasispecies[21–23]. Failure to detect CXCR4-using variants initially present at low frequencies in the virus population may lead to their

[1]Laboratoire de Virologie, Hôpital Purpan, Toulouse, F-31300 France. [2]INSERM, UMR1043, Toulouse, F-31300 France. [3]Université Toulouse III Paul Sabatier, Toulouse, F-31000 France. [4]Service des Maladies Infectieuses et Tropicales, Hôpital Purpan, Toulouse, F-31300 France. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to P.D. (email: delobel.p@chu-toulouse.fr)

X4 input (%)	X4 quantification in X4:R5 virus mixtures (%)								
	LAI:JR-CSF*			AFG4:AFG1*			CHS2:CHS11*		
0.5	0.5	0.0	0.6	0.7	0.7	0.0	0.0	0.0	0.0
1.0	1.7	0.0	1.0	0.9	1.1	1.0	0.6	0.5	0.0
5.0	2.3	2.6	2.0	5.3	5.4	7.1	5.0	4.2	5.3
20.0	9.2	1.8	2.5	21.4	23.4	—	30.0	25.2	—
50.0	57.8	43.9	40.7	53.2	56.2	—	56.2	53.1	—
75.0	73.6	78.3	79.3	74.3	70.8	—	80.6	81.6	—
100.0	100.0	100.0	100.0	100.0	100.0	—	100.0	100.0	—

Table 1. Quantifying X4 variants in HIV-1 quasispecies by ultra-deep pyrosequencing. *Each virus mixture was assessed in triplicate or duplicate beyond 5%

subsequent selection under CCR5 antagonist-based treatment[24,25]. Thus, there is a need for new genotypic techniques for determining tropism that are sensitive enough to detect minor CXCR4-using variants.

The sequencing of V3 *env* amplicons at high coverage with long read lengths has made massive parallel amplicon pyrosequencing using the 454 technology a promising tool for studying the virus diversity in clinical samples. It competes with ultra-sensitive phenotypic approaches for detecting low levels of CXCR4-using variants that current genotypic approaches miss, while being able to quantify the proportion of each variant in the virus quasispecies[24,26–36].

However, processing the mass of sequence data and the need to identify true minor variants while excluding artifactual sequences is the rate-limiting step in the process. Arbitrary fixed cut-offs (1–2%) are currently used, below which minor variants are discarded, but the errors generated during ultra-deep pyrosequencing are sequence-dependant rather than random, notably in homopolymeric regions[37].

We have developed an automated position-specific processing of V3 *env* ultra-deep pyrosequencing data for rapidly inferring HIV-1 tropism with improved detection of minor variants (PyroVir software). It uses a sequence of logic rules based on the V3 sequence to discard artifactual or non-functional sequences with frame shifts or stop codons (biological filters), followed by a position-specific matrix based on Poisson distribution (statistical filter) to discard sequences with artifactual point mutations at V3 positions of interest. A particular attention had also been paid to provide a representative description of the virus quasispecies by limiting sampling and amplification bias prior to ultra-deep pyrosequencing.

Results

Optimized amplification steps before ultra-deep pyrosequencing for accurate representation of HIV-1 quasispecies. We determined experimentally the number of PCR cycles for which the amplification of a given input of virus copies remains linear without distorting the proportions of minor and major variants in the quasispecies. We found that 34 cycles of RT-PCR for an input of 2,000–3,000 copies followed by 25 cycles of nested PCR were adequate to get high sensitivity without biasing the proportions in the virus population (Supplementary Fig. S1).

Performances of ultra-deep pyrosequencing for detecting CXCR4-using variants in HIV-1 quasispecies compared to an ultrasensitive phenotypic assay. Three artificial mixtures of culture supernatants of pure X4 and R5 clones (LAI:JR-CSF; AFG4:AFG1, CHS2:CHS11) with defined proportions of X4:R5 viruses (0:100; 0.5:99.5; 1:99; 5:95; 20:80; 50:50; 75:25; and 100:0) were submitted in parallel to ultra-deep pyrosequencing and phenotyping. The TTT phenotypic assay detected 0.5% of X4 viruses in the LAI:JR-CSF mixture (1/3 replicates), 0.5% of X4 viruses in the AFG4:AFG1 mixture (1/3 replicates), and 0.5% of X4 viruses in the CHS2:CHS11 mixture (3/3 replicates). Ultra-deep pyrosequencing of the same mixtures detected 0.5% of X4 viruses in the LAI:JR-CSF mixture (2/3 replicates), 0.5% of X4 viruses in the AFG4:AFG1 mixture (2/3 replicates), and 1% of X4 viruses in the CHS2:CHS11 mixture (2/3 replicates). Our optimized process of amplification before ultra-deep pyrosequencing thus accurately described the HIV-1 quasispecies, with a 0.5–1% sensitivity for detecting CXCR4-using variants without distorting the proportions in the virus population (Table 1).

Automated data cleaning of errors occurring during the ultra-deep pyrosequencing process. Ultra-deep pyrosequencing can generate errors that are sequence-dependant, notably in homopolymeric regions, rather than random. Our automated approach distinguishes authentic variants from artifactual ones resulting from errors arising during PCR amplification and ultra-deep pyrosequencing.

Biological filters to discard non-functional V3 sequences. The sequences from the 454 ultra-deep pyrosequencing data were first processed with GS Amplicon Variant Analyzer (AVA) software (Roche). We did not use the AVA software cleaning filters that discard sequences under a fixed cut-off. Instead, the AVA

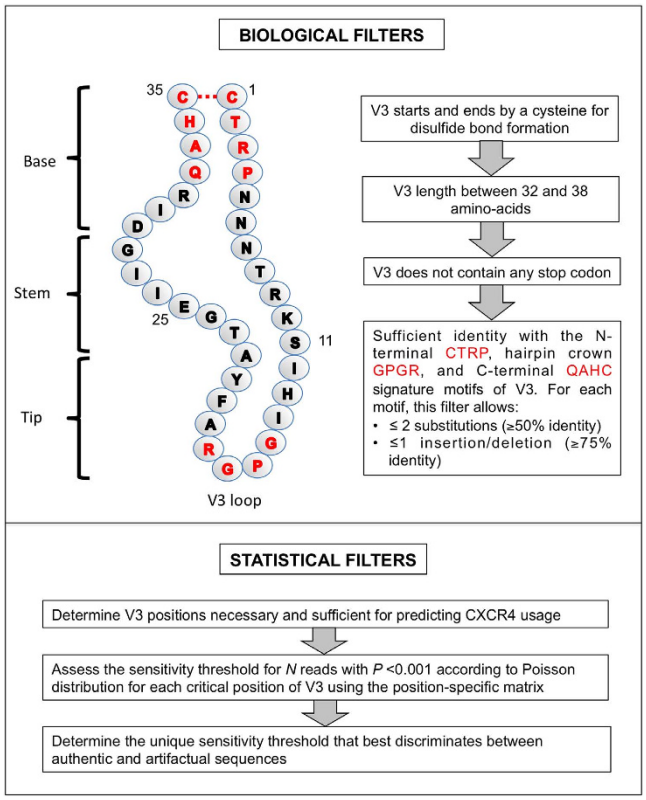


Figure 1. PyroVir flowchart.

alignments were extracted, truncated to the V3 *env* region, and gaps were removed. Reads with undetermined bases were discarded. Reads were then temporary translated into amino acid sequences in the right open reading frame in order to discard V3 sequences considered as non-functional if (i) they did no start and end with the cysteine required for the disulfide bound maintaining the V3 loop; (ii) they were not 32–38 amino-acids long; (iii) they contained a stop codon; (iv) they were not sufficiently identical with the V3 consensus at three typical motifs: a "CTRP"-like signature at the N-terminus (V3 residues 1–4), a "GPGR"-like signature at the hairpin crown (V3 residues 15–18), or a "QAHC"-like signature at the C-terminus (V3 residues 32–35). An identity of 0.5 and 0.75 was allowed for substitution and insertion/deletion of an amino-acid at any of the three signature motifs (Fig. 1). The biological filters mainly discard sequences with frame shifts due to insertions/deletions of nucleotides in homopolymeric regions or stop codons. This step removed 5.7% (mean) of V3 reads.

Statistical filters to discard sequences with artifactual point mutations. The statistical filters assessed the probability that a sequence with a point mutation at a position of interest for predicting coreceptor usage would be artifactual or authentic. We first determined the frequency of artifactual V3 variants among the reads of 20 virus clones whose Sanger sequences were used as reference. The mean frequency of artifactual V3 variants was 0.646% [exact Poisson 99% confidence interval (CI), 0.00560–0.00742] of the reads. We defined the global error rate as the upper 99% confidence interval limit of this mean frequency of artifactual V3 variants. Based on this global error rate of 0.742% (μ), we estimated the expected number of artifactual sequences (λ) for N reads ($\lambda = N * \mu$). We then used Poisson distribution to determine the minimum threshold above which a minor variant could be considered authentic for a given number of reads with $P < 0.001$. This fixed cut-off provided sensitivity thresholds from 1.7% for 1,000 reads to 1.14% for 5,000 reads.

However, only a few key V3 amino-acid residues significantly influence the prediction of CXCR4 coreceptor usage by genotypic algorithms. As the errors generated during ultra-deep pyrosequencing are sequence-dependant, arbitrary fixed cut-offs are not ideal for distinguishing authentic variants from artifactual ones. We have determined position-specific error rates along the V3 sequence, defined as the upper 99% confidence interval limit (Poisson statistics) of the mean frequency of artifactual codons at each V3 position among the 20 virus clones. The error rate varied greatly along the V3 sequence (Fig. 2). V3 position 20 had the highest error rate, followed by positions 19, 22, 21, 26, 18, and 11. We determined the percentage at which a given V3 position contributed to the mean error rate along V3. We then attributed a weighted error rate (ratio to 0.02857 – value if errors occurred constantly along the 35 positions of V3 - multiplied by the global error rate of 0.00742) at each position. These weighted error rates were

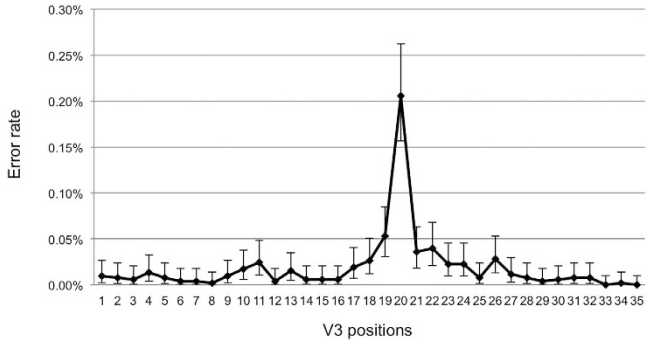


Figure 2. Error rate of amplification and ultra-deep pyrosequencing at each position of the V3 sequence. The mean error rate of amplification and pyrosequencing was estimated at each position of V3 by comparing the pyrosequencing reads to the Sanger sequences of 20 clones. The mean error rate is shown with exact Poisson 99% confidence interval at each position of the V3 sequence.

used to construct a sensitivity threshold matrix for each position of V3 to retain a minor virus variants harboring a point mutations as authentic for a given number of reads with $P < 0.001$ (Table 2).

Genotypic prediction of HIV-1 tropism. The genotypic prediction of CXCR4-tropism for a given V3 sequence must take into account several amino-acids of interest. We identified the positions at which amino acid K, R, D and E were present within a given sequence. The detection thresholds at these particular positions were defined for a given number of reads. A single threshold was then defined for each putative CXCR4-using variant, based on the criteria of the combined 11/25 and net charge rule which are necessary and sufficient to predict CXCR4 tropism (Table 3). This was then used to retain a sequence predicted to be CXCR4-tropic whose frequency was above the specific threshold defined for this particular sequence at a given number of reads. The statistical filter for the sensitivity threshold of sequences predicted to be CCR5-tropic (*i.e.* those having no criteria of the combined 11/25 and net charge rule) was based on the global error rate of the whole V3 region ($\mu = 0.00742$), as defined above.

PyroVir software. This biological and statistical data cleaning strategy has been integrated in a program (PyroVir, IDDN FR.001.160011.000.S.P.2012.000.31230, Inserm-Transfert) that provides a fast, automated position-specific process for inferring HIV-1 tropism from V3 *env* 454 ultra-deep pyrosequencing data with improved detection of minor variants. The genotypic rule used to predict CXCR4-usage can be changed depending on the virus subtype, particularly for subtypes D and CRF01-AE for which we have developed specific algorithms[38,39]. An example of HIV-1 quasispecies coreceptor usage prediction by PyroVir is shown in Fig. 3. PyroVir is accessible at http://diag.ablsa.com/pyrovir/submit.php.

Discussion

HIV-1 quasispecies coreceptor usage must be accurately determined before starting CCR5 antagonist-based antiretroviral therapy, as the presence of undetected minor CXCR4-using variants can lead to subsequent virological failure[24,25]. The development of gene therapy targeting CCR5 on hematopoietic stem cells in the quest for HIV cure would also require sensitive detection of CXCR4-using variants. Recombinant virus phenotypic entry assays are sensitive (0.3–0.5%) and considered to be the gold standard, but these assays are labour-intensive and expensive[8,14]. Genotypic methods based on bulk sequencing of V3 *env* combined with bioinformatics tools for inferring HIV-1 tropism are more rapid and more economical than phenotypic tests. But these simple genotypic assays are not sensitive enough to detect below 20% of minor CXCR4-using variants[21–23]. Ultra-deep pyrosequencing provides genotypic sensitivities similar to those of the current phenotypic assays, and also quantifies the proportion of each variant in the virus quasispecies. Ultra-deep pyrosequencing also has the advantage of using genotypic algorithms at a clonal level where genotype-phenotype correlations are better than for virus populations[21]. The PCR amplification step is an important potential source of artifacts, such as substitutions and recombinations, that can be minimized by an optimized amplification[40–42]. The *Taq* polymerase enzyme used has an important impact on the proportion of correct reads after sequencing[43]. But adequate representation of the virus quasispecies should also be preserved and this requires the use of a reduced number of PCR amplification cycles with a normalized virus input to ensure that the amplifications of both major and minor variants in the quasispecies are still in the logarithmic phase when the reaction is stopped.

Ultra-deep pyrosequencing of HIV-1 V3 *env* allows to detect low levels of CXCR4-using variants that current genotypic approaches miss. However, the extremely large data sets produced pose challenging computational problems, particularly the need to clean up the sequences by removing artifactual errors generated during amplification and pyrosequencing. Our PyroVir software rapidly and reliably predicts HIV-1 coreceptor usage from 454 ultra-deep pyrosequencing data. It has two modules. The first, biological

Number of reads	V3 positions																																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
0	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
5	20	20	20	20	20	20	20	20	20	20	40	20	20	20	20	20	20	40	40	60	40	40	20	20	20	40	20	20	20	20	20	20	20	20	20
10	20	20	10	20	20	10	10	10	20	20	20	10	20	10	10	10	20	20	20	40	20	20	20	20	20	20	20	20	10	10	20	20	10	10	10
15	13	13	10	13	13	10	10	6.7	13	13	13	10	13	10	10	10	13	13	20	33	13	20	13	13	13	13	13	13	10	10	13	13	6.7	6.7	6.7
20	10	10	10	10	10	10	10	6.7	10	10	10	10	10	10	10	10	10	13	15	25	13	15	10	10	10	13	10	10	10	10	10	10	5	6.7	5
25	8	8	8	8	8	8	8	6.7	8	8	10	8	8	8	8	8	10	12	12	24	12	12	10	10	8	12	8	8	8	8	8	8	5	6.7	5
30	6.7	6.7	6.7	8	6.7	6.7	6.7	6.7	6.7	8	10	6.7	8	6.7	6.7	6.7	10	10	12	23	10	10	10	10	6.7	10	6.7	6.7	6.7	6.7	6.7	6.7	5	6.7	5
35	5.7	5.7	5.7	8	5.7	5.7	5.7	5.7	5.7	8	8.6	5.7	8	5.7	5.7	5.7	8.6	8.6	11	20	10	10	8.6	8.6	5.7	8.6	6.7	5.7	5.7	5.7	5.7	5.7	5	5.7	5
40	5.7	5	5	7.5	5	5	5	5	5.7	7.5	7.5	5	7.5	5	5	5	7.5	7.5	10	20	10	10	7.5	7.5	5	7.5	6.7	5	5	5	5	5	5	5	5
45	5.7	5	4.4	6.7	5	4.4	4.4	4.4	5.7	6.7	6.7	4.4	6.7	4.4	4.4	4.4	6.7	7.5	10	18	8.9	8.9	6.7	6.7	5	7.5	6.7	5	4.4	4.4	5	5	4.4	4.4	4.4
50	5.7	5	4.4	6	5	4	4	4	5.7	6	6.7	4	6	4.4	4.4	4.4	6	7.5	10	18	8	8	6.7	6.7	5	7.5	6	5	4	4.4	5	5	4	4	4
55	5.5	5	4.4	5.5	5	3.6	3.6	3.6	5.5	5.5	6.7	3.6	5.5	4.4	4.4	4.4	6	7.3	9.1	16	7.3	7.3	6.7	6.7	5	7.3	5.5	5	3.6	4.4	5	5	3.6	3.6	3.6
60	5	5	4.4	5	5	3.6	3.6	3.3	5	5.5	6.7	3.6	5	4.4	4.4	4.4	6	6.7	8.3	16	7.3	7.3	6.7	6.7	5	6.7	5	5	3.6	4.4	5	5	3.3	3.3	3.3
65	4.6	4.6	4.4	4.6	4.6	3.6	3.6	3.1	4.6	5.5	6.2	3.6	5	4.4	4.4	4.4	6	6.2	7.7	15	7.3	7.3	6.2	6.2	4.6	6.2	4.6	4.6	3.6	4.4	4.6	4.6	3.1	3.1	3.1
70	4.3	4.3	4.3	4.6	4.3	3.6	3.6	3.1	4.3	5.5	5.7	3.6	5	4.3	4.3	4.3	5.7	5.7	7.7	15	7.1	7.1	5.7	5.7	4.3	5.7	4.3	4.3	3.6	4.3	4.3	4.3	2.9	3.1	2.9
75	4	4	4	4.6	4	3.6	3.6	3.1	4	5.3	5.3	3.6	5	4	4	4	5.3	5.7	7.7	15	6.7	6.7	5.3	5.3	4	5.7	4.3	4	3.6	4	4	4	2.7	3.1	2.7
80	3.8	3.8	3.8	4.6	3.8	3.6	3.6	3.1	3.8	5	5.3	3.6	5	3.8	3.8	3.8	5	5.7	7.5	15	6.3	6.3	5	5	3.8	5.7	4.3	3.8	3.6	3.8	3.8	3.8	2.5	3.1	2.5
85	3.8	3.5	3.5	4.6	3.5	3.5	3.5	3.1	3.8	4.7	5.3	3.5	4.7	3.5	3.5	3.5	4.7	5.7	7.1	14	5.9	6.3	5	5	3.5	5.7	4.3	3.5	3.5	3.5	3.5	3.5	2.4	3.1	2.4
90	3.8	3.3	3.3	4.4	3.3	3.3	3.3	3.1	3.8	4.4	5.3	3.3	4.4	3.3	3.3	3.3	4.4	5.6	6.7	14	5.9	6.3	5	5	3.3	5.6	4.3	3.3	3.3	3.3	3.3	3.3	2.2	3.1	2.2
95	3.8	3.3	3.2	4.2	3.3	3.2	3.2	3.1	3.8	4.2	5.3	3.2	4.2	3.2	3.2	3.2	4.4	5.3	6.7	14	5.9	6.3	5	5	3.3	5.3	4.2	3.3	3.2	3.2	3.3	3.3	2.1	3.1	2.1
100	3.8	3.3	3	4	3.3	3	3	3	3.8	4.2	5	3	4	3	3	3	4.4	5	6.7	13	5.9	6	5	5	3.3	5	4	3.3	3	3	3.3	3.3	2.1	3	2.1
200	2.5	2.5	2.5	3	2.5	2	2	2	2.5	3	3.5	2	3	2.5	2.5	2.5	3.5	4	5	11	4.5	4.5	3.5	3.5	2.5	4	3	2.5	2	2.5	2.5	2.5	1.5	2	1.5
300	2.3	2	2	2.3	2	1.7	1.7	1.7	2.3	2.7	3	1.7	2.7	2	2	2	2.7	3.3	4.3	9.7	3.7	3.7	3	3	2	3.3	2.3	2	1.7	2	2	2	1.3	1.7	1.3
400	2	1.8	1.8	2.3	1.8	1.5	1.5	1.5	2	2.5	2.8	1.5	2.3	1.8	1.8	1.8	2.5	2.8	4	9	3.3	3.5	2.8	2.8	1.8	3	2	1.8	1.5	1.8	1.8	1.8	1.3	1.5	1.3
500	1.8	1.6	1.6	2	1.6	1.4	1.4	1.2	1.8	2.2	2.6	1.4	2	1.6	1.6	1.6	2.2	2.6	3.8	8.6	3	3.2	2.4	2.4	1.6	2.8	1.8	1.6	1.4	1.6	1.6	1.6	1	1.2	1
600	1.7	1.5	1.5	1.8	1.5	1.3	1.3	1.2	1.7	2	2.3	1.3	2	1.5	1.5	1.5	2.2	2.5	3.5	8.3	2.8	3	2.3	2.3	1.5	2.5	1.8	1.5	1.3	1.5	1.5	1.5	1	1.2	1
700	1.6	1.4	1.3	1.7	1.4	1.1	1.1	1	1.6	2	2.3	1.1	1.9	1.3	1.3	1.3	2	2.3	3.4	8	2.7	2.9	2.1	2.1	1.4	2.4	1.7	1.4	1.1	1.3	1.4	1.4	0.9	1	0.9
800	1.5	1.4	1.3	1.6	1.4	1.1	1.1	1	1.5	1.9	2.1	1.1	1.8	1.3	1.3	1.3	1.9	2.3	3.3	7.9	2.6	2.8	2.1	2.1	1.4	2.4	1.6	1.4	1.1	1.3	1.4	1.4	0.9	1	0.9
900	1.4	1.3	1.2	1.6	1.3	1.1	1.1	1	1.4	1.8	2.1	1.1	1.7	1.2	1.2	1.2	1.9	2.2	3.1	7.7	2.6	2.7	2	2	1.3	2.2	1.6	1.3	1.1	1.2	1.3	1.3	0.8	1	0.8
1000	1.4	1.3	1.2	1.5	1.3	1	1	0.9	1.4	1.7	2	1	1.6	1.2	1.2	1.2	1.8	2.1	3.1	7.5	2.5	2.6	2	2	1.3	2.2	1.5	1.3	1	1.2	1.3	1.3	0.8	0.9	0.8
1100	1.4	1.2	1.1	1.5	1.2	1	1	0.9	1.4	1.6	2	1	1.6	1.1	1.1	1.1	1.7	2.1	3	7.5	2.4	2.6	1.9	1.9	1.2	2.1	1.5	1.2	1	1.1	1.2	1.2	0.7	0.9	0.7
1200	1.3	1.2	1.1	1.5	1.2	1	1	0.8	1.3	1.6	1.9	1	1.6	1.1	1.1	1.1	1.7	2	2.9	7.3	2.3	2.5	1.8	1.8	1.2	2.1	1.3	1.2	1	1.1	1.2	1.2	0.7	0.8	0.7
1300	1.2	1.2	1.1	1.5	1.2	0.9	0.9	0.8	1.2	1.6	1.9	0.9	1.5	1.1	1.1	1.1	1.7	1.9	2.9	7.2	2.3	2.5	1.8	1.8	1.2	2	1.3	1.2	0.9	1.1	1.2	1.2	0.7	0.8	0.7
1400	1.2	1.1	1	1.4	1.1	0.9	0.9	0.8	1.2	1.6	1.9	0.9	1.5	1	1	1	1.6	1.9	2.9	7.1	2.3	2.4	1.8	1.8	1.1	2	1.3	1.1	0.9	1	1.1	1.1	0.6	0.8	0.6
1500	1.2	1.1	1	1.3	1.1	0.9	0.9	0.8	1.2	1.5	1.8	0.9	1.5	1	1	1	1.6	1.9	2.8	7.1	2.2	2.3	1.7	1.7	1.1	1.9	1.3	1.1	0.9	1	1.1	1.1	0.6	0.8	0.6
1600	1.2	1.1	1	1.3	1.1	0.9	0.9	0.8	1.2	1.5	1.8	0.9	1.4	1	1	1	1.6	1.9	2.8	7	2.2	2.3	1.7	1.7	1.1	1.9	1.3	1.1	0.9	1	1.1	1.1	0.6	0.8	0.6
1700	1.2	1.1	0.9	1.3	1.1	0.9	0.9	0.8	1.2	1.5	1.8	0.9	1.4	0.9	0.9	0.9	1.5	1.8	2.7	6.9	2.2	2.3	1.7	1.7	1.1	1.9	1.2	1.1	0.9	0.9	1.1	1.1	0.6	0.8	0.6
1800	1.1	1.1	0.9	1.3	1.1	0.8	0.8	0.7	1.1	1.4	1.7	0.8	1.4	0.9	0.9	0.9	1.5	1.8	2.7	6.9	2.1	2.3	1.7	1.7	1.1	1.9	1.2	1.1	0.8	0.9	1.1	1.1	0.6	0.7	0.6
1900	1.1	1	0.9	1.3	1	0.8	0.8	0.7	1.1	1.4	1.7	0.8	1.4	0.9	0.9	0.9	1.5	1.8	2.6	6.8	2.1	2.2	1.6	1.6	1	1.8	1.2	1	0.8	0.9	1	1	0.6	0.7	0.6
2000	1.1	1	0.9	1.3	1	0.8	0.8	0.7	1.1	1.4	1.7	0.8	1.4	0.9	0.9	0.9	1.5	1.8	2.6	6.8	2.1	2.2	1.6	1.6	1	1.8	1.2	1	0.8	0.9	1	1	0.6	0.7	0.6
2100	1.1	1	0.9	1.2	1	0.8	0.8	0.7	1.1	1.4	1.7	0.8	1.3	0.9	0.9	0.9	1.5	1.8	2.6	6.8	2.1	2.2	1.6	1.6	1	1.8	1.1	1	0.8	0.9	1	1	0.6	0.7	0.6
2200	1.1	1	0.9	1.2	1	0.8	0.8	0.7	1.1	1.4	1.6	0.8	1.3	0.9	0.9	0.9	1.5	1.7	2.6	6.7	2.1	2.2	1.6	1.6	1	1.8	1.1	1	0.8	0.9	1	1	0.6	0.7	0.6
2300	1	1	0.9	1.2	1	0.8	0.8	0.7	1	1.4	1.6	0.8	1.3	0.9	0.9	0.9	1.4	1.7	2.6	6.7	2	2.1	1.6	1.6	1	1.8	1.1	1	0.8	0.9	1	1	0.5	0.7	0.5
2400	1	1	0.9	1.2	1	0.8	0.8	0.7	1	1.3	1.6	0.8	1.3	0.9	0.9	0.9	1.4	1.7	2.5	6.7	2	2.1	1.5	1.5	1	1.8	1.1	1	0.8	0.9	1	1	0.5	0.7	0.5
2500	1	1	0.8	1.2	1	0.8	0.8	0.6	1	1.3	1.6	0.8	1.3	0.8	0.8	0.8	1.4	1.7	2.5	6.6	2	2.1	1.5	1.5	1	1.7	1.1	1	0.8	0.8	1	1	0.5	0.6	0.5
2600	1	0.9	0.8	1.2	0.9	0.8	0.8	0.6	1	1.3	1.6	0.8	1.2	0.8	0.8	0.8	1.4	1.7	2.5	6.6	2	2.1	1.5	1.5	0.9	1.7	1.1	0.9	0.8	0.8	0.9	0.9	0.5	0.6	0.5
2700	1	0.9	0.8	1.2	0.9	0.7	0.7	0.6	1	1.3	1.6	0.7	1.2	0.8	0.8	0.8	1.4	1.6	2.5	6.6	2	2.1	1.5	1.5	0.9	1.7	1.1	0.9	0.7	0.8	0.9	0.9	0.5	0.6	0.5
2800	1	0.9	0.8	1.1	0.9	0.7	0.7	0.6	1	1.3	1.6	0.7	1.2	0.8	0.8	0.8	1.4	1.6	2.5	6.6	1.9	2.1	1.5	1.5	0.9	1.7	1.1	0.9	0.7	0.8	0.9	0.9	0.5	0.6	0.5
2900	1	0.9	0.8	1.1	0.9	0.7	0.7	0.6	1	1.3	1.6	0.7	1.2	0.8	0.8	0.8	1.3	1.6	2.5	6.6	1.9	2.1	1.5	1.5	0.9	1.7	1.1	0.9	0.7	0.8	0.9	0.9	0.5	0.6	0.5
3000	1	0.9	0.8	1.1	0.9	0.7	0.7	0.6	1	1.3	1.5	0.7	1.2	0.8	0.8	0.8	1.3	1.6	2.4	6.5	1.9	2	1.5	1.5	0.9	1.7	1.1	0.9	0.7	0.8	0.9	0.9	0.5	0.6	0.5
3100	1	0.9	0.8	1.1	0.9	0.7	0.7	0.6	1	1.3	1.5	0.7	1.2	0.8	0.8	0.8	1.3	1.6	2.4	6.5	1.9	2	1.5	1.5	0.9	1.7	1.1	0.9	0.7	0.8	0.9	0.9	0.5	0.6	0.5
3200	1	0.9	0.8	1.1	0.9	0.7	0.7	0.6	1	1.3	1.5	0.7	1.2	0.8	0.8	0.8	1.3	1.6	2.4	6.5	1.9	2	1.5	1.5	0.9	1.7	1	0.9	0.7	0.8	0.9	0.9	0.5	0.6	0.5
3300	1	0.9	0.8	1.1	0.9	0.7	0.7	0.6	1	1.2	1.5	0.7	1.2	0.8	0.8	0.8	1.3	1.6	2.4	6.5	1.9	2	1.5	1.5	0.9	1.6	1	0.9	0.7	0.8	0.9	0.9	0.5	0.6	0.5
3400	0.9	0.9	0.8	1.1	0.9	0.7	0.7	0.6	0.9	1.2	1.5	0.7	1.2	0.8	0.8	0.8	1.3	1.6	2.4	6.4	1.9	2	1.4	1.4	0.9	1.6	1	0.9	0.7	0.8	0.9	0.9	0.5	0.6	0.5
3500	0.9	0.9	0.8	1.1	0.9	0.7	0.7	0.6	0.9	1.2	1.5	0.7	1.2	0.8	0.8	0.8	1.3	1.6	2.4	6.4	1.9	2	1.4	1.4	0.9	1.6	1	0.9	0.7	0.8	0.9	0.9	0.5	0.6	0.5
3600	0.9	0.9	0.8	1.1	0.9	0.7	0.7	0.6	0.9	1.2	1.5	0.7	1.2	0.8	0.8	0.8	1.3	1.6	2.4	6.4	1.9	2	1.4	1.4	0.9	1.6	1	0.9	0.7	0.8	0.9	0.9	0.5	0.6	0.5
3700	0.9	0.9	0.8	1.1	0.9	0.7	0.7	0.6	0.9	1.2	1.5	0.7	1.2	0.8	0.8	0.8	1.3	1.5	2.4	6.4	1.8	2	1.4	1.4	0.9	1.6	1	0.9	0.7	0.8	0.9	0.9	0.5	0.6	0.5
3800	0.9	0.8	0.8	1.1	0.8	0.7	0.7	0.6	0.9	1.2	1.5	0.7	1.2	0.8	0.8	0.8	1.3	1.5	2.3	6.4	1.8	2	1.4	1.4	0.8	1.6	1	0.8	0.7	0.8	0.8	0.8	0.5	0.6	0.5
3900	0.9	0.8	0.8	1.1	0.8	0.7	0.7	0.6	0.9	1.2	1.5	0.7	1.1	0.8	0.8	0.8	1.3	1.5	2.3	6.3	1.8	2	1.4	1.4	0.8	1.6	1	0.8	0.7	0.8	0.8	0.8	0.5	0.6	0.5
4000	0.9	0.8	0.8	1.1	0.8	0.7	0.7	0.6	0.9	1.2	1.5	0.7	1.1	0.8	0.8	0.8	1.3	1.5	2.3	6.3	1.8	2	1.4	1.4	0.8	1.6	1	0.8	0.7	0.8	0.8	0.8	0.5	0.6	0.5
4100	0.9	0.8	0.8	1.1	0.8	0.7	0.7	0.6	0.9	1.2	1.5	0.7	1.1	0.8	0.8	0.8	1.3	1.5	2.3	6.3	1.8	1.9	1.4	1.4	0.8	1.6	1	0.8	0.7	0.8	0.8	0.8	0.4	0.6	0.4
4200	0.9	0.8	0.7	1.1	0.8	0.7	0.7	0.6	0.9	1.2	1.5	0.7	1.1	0.7	0.7	0.7	1.3	1.5	2.3	6.3	1.8	1.9	1.4	1.4	0.8	1.6	1	0.8	0.7	0.7	0.8	0.8	0.4	0.6	0.4
4300	0.9	0.8	0.7	1.1	0.8	0.7	0.7	0.6	0.9	1.2	1.4	0.7	1.1	0.7	0.7	0.7	1.3	1.5	2.3	6.3	1.8	1.9	1.4	1.4	0.8	1.6	1	0.8	0.7	0.7	0.8	0.8	0.4	0.6	0.4
4400	0.9	0.8	0.7	1.1	0.8	0.7	0.7	0.6	0.9	1.2	1.4	0.7	1.1	0.7	0.7	0.7	1.3	1.5	2.3	6.3	1.8	1.9	1.4	1.4	0.8	1.6	1	0.8	0.7	0.7	0.8	0.8	0.4	0.6	0.4
4500	0.9	0.8	0.7	1	0.8	0.6	0.6	0.6	0.9	1.2	1.4	0.6	1.1	0.7	0.7	0.7	1.2	1.5	2.3	6.3	1.8	1.9	1.4	1.4	0.8	1.6	1	0.8	0.6	0.7	0.8	0.8	0.4	0.6	0.4
4600	0.9	0.8	0.7	1	0.8	0.6	0.6	0.5	0.9	1.2	1.4	0.6	1.1	0.7	0.7	0.7	1.2	1.5	2.3	6.2	1.8	1.9	1.4	1.4	0.8	1.5	1	0.8	0.6	0.7	0.8	0.8	0.4	0.5	0.4
4700	0.9	0.8	0.7	1	0.8	0.6	0.6	0.5	0.9	1.2	1.4	0.6	1.1	0.7	0.7	0.7	1.2	1.5	2.3	6.2	1.8	1.9	1.4	1.4	0.8	1.5	1	0.8	0.6	0.7	0.8	0.8	0.4	0.5	0.4
4800	0.9	0.8	0.7	1	0.8	0.6	0.6	0.5	0.9	1.2	1.4	0.6	1.1	0.7	0.7	0.7	1.2	1.5	2.3	6.2	1.8	1.9	1.4	1.4	0.8	1.5	1	0.8	0.6	0.7	0.8	0.8	0.4	0.5	0.4
4900	0.9	0.8	0.7	1	0.8	0.6	0.6	0.5	0.9	1.2	1.4	0.6	1.1	0.7	0.7	0.7	1.2	1.5	2.3	6.2	1.8	1.9	1.4	1.4	0.8	1.5	1	0.8	0.6	0.7	0.8	0.8	0.4	0.5	0.4
5000	0.9	0.8	0.7	1	0.8	0.6	0.6	0.5	0.9	1.2	1.4	0.6	1.1	0.7	0.7	0.7	1.2	1.5	2.3	6.2	1.8	1.9	1.3	1.3	0.8	1.5	0.9	0.8	0.6	0.7	0.8	0.8	0.4	0.5	0.4

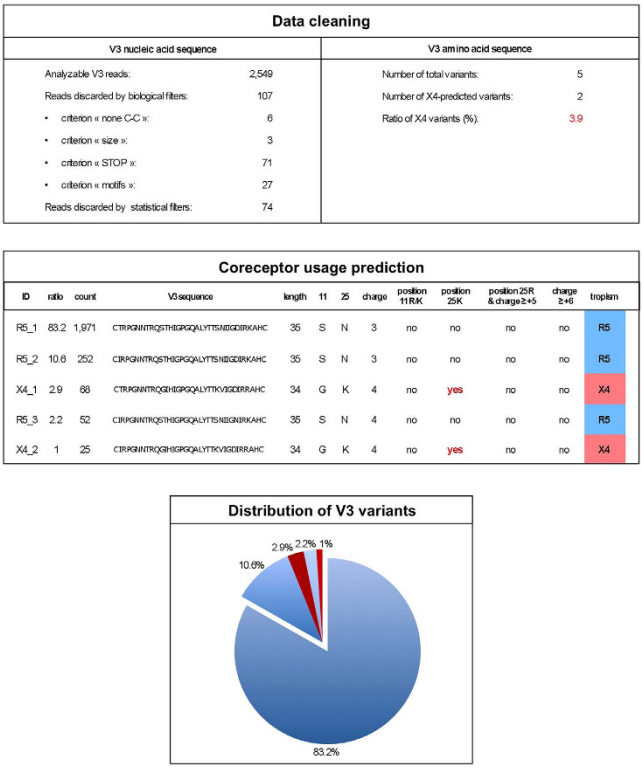
Table 2. V3 position-specific matrix of sensitivity thresholds.

case	Criteria of the combined 11/25 and net charge rule for HIV-1 subtype B						retained threshold
			25R and net charge* $\geq +5$		net charge* $\geq +6$		
	11R or 11K	25K	25R and net charge $= +5$	25R and net charge $> +5$†	net charge $= +6$	net charge $> +6$‡	
1	x						position 11
2		x					position 25
3			x				maximum§
4				x			maximum†,§
5					x		maximum¶
6						x	maximum‡,¶
7	x	x					minimum#
8	x		x				minimum§,‖
9	x			x			minimum†,§,‖
10	x				x		minimum¶,**
11	x					x	minimum‡,¶,**
12		x			x		minimum¶,††
13		x				x	minimum‡,¶,††
14	x	x			x		minimum¶,‡‡
15	x	x				x	minimum‡,¶,‡‡

Table 3. Determining a single sensitivity threshold necessary and sufficient for predicting CXCR4-usage according to the combined 11/25 and net charge rule. *The V3 net charge was calculated by subtracting the number of negatively charged amino acids (D and E) from the number of positively charged ones (K and R). †The most pejorative positions harboring a positively charged residue were eliminated to reach the criterion << 25R and net charge $= +5$ >>. Alternatively, position 25 can be eliminated to arrive at the criterion << net charge $= +6$ >> if more favorable. ‡The most pejorative positions harboring a positively charged residue were eliminated to arrive at the criterion << net charge $= +6$ >>. §The most pejorative threshold between that of position 25 and those required for a net charge of $+5$ is used. ¶The most pejorative threshold of the positions harboring a positively charged residue was used. #The least pejorative threshold between those of positions 11 and 25 is used. ‖The least pejorative threshold between that of position 11 and those required for the << 25R and net charge $= +5$ >> criterion is used. **The least pejorative threshold between that of position 11 and those required for a net charge of $+6$ is used. ††The least pejorative threshold between that of position 25 and those required for a net charge of $+6$ is used. ‡‡The least pejorative threshold between those of positions 11, and 25, and those required for a net charge of $+6$ is used.

filters, discard artifactual and non-functional sequences, particularly those due to frame-shifts generated by insertions or deletions of nucleotides in homopolymeric regions or stop codons. The second, statistical filters based on Poisson distribution, discard artifactual point mutations. This method is position-specific and does not use arbitrary fixed cut-offs to discard sequences with artifactual point mutations at V3 positions of interest as the errors generated during ultra-deep pyrosequencing are sequence-dependant. We found that the error rate of ultra-deep pyrosequencing varied along the V3 sequence, being maximum around V3 position 20. PyroVir automatically determines the sensitivity threshold for a given number of reads at each of the V3 positions involved in predicting CXCR4-usage, and then retains the highest threshold of the critical positions necessary and sufficient for predicting that a sequence is CXCR4-using. Subtype-specific algorithms could be used, especially for non-B subtypes such as subtypes D and CRF01-AE[38,39,44]. Arbitrary fixed cut-offs are currently used for cleaning up ultra-deep pyrosequencing data, usually 1 to 2%, below which sequences are discarded. These cut-offs have been determined based on rough error rates of ultra-deep pyrosequencing, and virological response rates in clinical studies with a limited number of patients[34]. Our results show that position-specific thresholds must be used to reliably detect minor variants. The sensitivity threshold varies greatly for a given number of reads (0.4 to 6.2% for 5,000 reads), depending on the V3 positions critical for CXCR4 usage. Therefore, using a fixed cut-off could result in a lack of sensitivity for some variants if the V3 positions involved have low error rates or even to false positives if the error rate is high.

The clinical relevance of minor variants is a matter of debate but should be distinguished from the analytical sensitivity of the method used. Previous studies reported cases of virological failure under CCR5 antagonists due to minor variants $<1\%$[24], while other found that a 2% cut-off optimally predicted the clinical response[34]. Most of the data on the relevance of minor variants in HIV drug resistance have been reported in studies of non-nucleoside reverse transcriptase inhibitors (NNRTIs). Minor variants at



Data cleaning

V3 nucleic acid sequence		V3 amino acid sequence	
Analyzable V3 reads:	2,549	Number of total variants:	5
Reads discarded by biological filters:	107	Number of X4-predicted variants:	2
• criterion « none C-C »:	6	Ratio of X4 variants (%):	3.9
• criterion « size »:	3		
• criterion « STOP »:	71		
• criterion « motifs »:	27		
Reads discarded by statistical filters:	74		

Coreceptor usage prediction

ID	ratio	count	V3 sequence	length	11	25	charge	position 11R,K	position 25K	position 25R & charge ≥+5	charge ≥+6	tropism
R5_1	83.2	1,971	CTRPGNNTRQSTHIGPGQALYTTSNIIGDIRKAHC	35	S	N	3	no	no	no	no	R5
R5_2	10.6	252	CIRPGNNTRQSTHIGPGQALYTTSNIIGDIRKAHC	35	S	N	3	no	no	no	no	R5
X4_1	2.9	68	CTRPGNNTRQGHIGPGQALYTTKVIGDIRRAHC	34	G	K	4	no	yes	no	no	X4
R5_3	2.2	52	CIRPGNNTRQSTHIGPGQALYTTSNIIGNIRKAHC	35	S	N	4	no	no	no	no	R5
X4_2	1	25	CTRPGNNTRQGHIGPGQALYTTKVIGDIRRAHC	34	G	K	4	no	yes	no	no	X4

Figure 3. Example of PyroVir analysis for predicting HIV-1 quasispecies coreceptor usage. RNA was extracted from a plasma sample from an HIV-infected subject and submitted to nested RT-PCR amplification of V3 *env*, ultra-deep pyrosequencing, and PyroVir analysis. Ultra-deep pyrosequencing provided 2,549 analyzable V3 reads, of which 107 were discarded by the biological filter and and 74 by the statistical filter. 2 variants, one accounting for 2.9% and the other for 1% of the quasispecies, were predicted to use CXCR4.

frequencies of $<0.5\%$ have been demonstrated to have a clinical impact on the virological response[45]. It has also been suggested that absolute numbers of resistant viruses are more clinically relevant than their frequencies for assessing the risk of subsequent virological failure. The measured frequency of viruses harboring a mutation associated with drug resistance should thus be multiplied by the plasma virus load to determine the absolute numbers of resistant viruses per mL of plasma. Minor resistant viruses in concentrations of 10–99 copies/mL were found to have a statistically significant impact on the virological response to NNRTIs, while concentrations of 1–9 copies/mL did not[45]. Interpretations of the impact of minor resistant viruses on the virological response are subject to additional caveats, notably the fitness and infectivity of the minor resistant viruses, and the effectiveness of the other molecules included in the combined antiretroviral regimen given to the subject. Analysis of the response to CCR5 antagonists is further complicated by the antiviral effect of CCR5 antagonists on R5 × 4 dualtropic variants in which CCR5 usage is more important than that of CXCR4 (<< dualR5 >> variants)[46].

To summarize, we have developed an optimized process for the undistorted amplification and ultra-sensitive characterization of the coreceptor usage of HIV-1 quasispecies using ultra-deep pyrosequencing. This automated approach uses biological filters to discard artifactual or non-functional V3 sequences followed by statistical filters to determine position-specific sensitivity thresholds to identify authentic sequences with point mutations at V3 positions of interest.

Methods

Sample processing. The HIV-1 RNA in plasma samples was quantified with COBAS Ampliprep/COBAS TaqMan HIV-1 test version 2.0 (Roche). Plasma samples with a virus load of $<10{,}000$ copies/ml were ultracentrifuged at 20,000 g for 2h to concentrate the virus and RNA was extracted using the QIAamp Viral RNA Mini Kit (Qiagen). The initial input was adjusted to 2,000–3,000 copies of virus per PCR reaction, performed in duplicate, to avoid sampling bias. A lower input (300 copies) results in greater variability in the initial RT-PCR amplification for viruses at low frequencies and a risk of resampling (Supplementary Fig. S2).

Amplification steps. A 1009-bp nucleotide fragment encompassing the V1-V3 *env* region of HIV-1 RNA was amplified by RT-PCR. The linearity of the PCR amplification process was checked by

comparing the proportions of X4 (LAI, GenBank accession no. K02013.1) and R5 (JR-CSF, GenBank accession no. M38429.1) clones in the pyrosequencing output with the input of X4:R5 virus clones mixed in proportions of 0:100, 0.5:99.5, 1:99, 5:95, 20:80, 50:50, 75:25, and 100:0, adjusted to a total input of 2,000–3,000 copies of RNA, and submitted to multiple parallel PCR amplifications with various numbers of cycles. The resulting optimized process used the SuperScript III One-Step RT-PCR System (Invitrogen) for RT-PCR with the following conditions: 60 min at 55 °C; 2 min at 94 °C; 30 s at 94 °C, 30 s at 55 °C, and 1 min 30 s at 68 °C for 10 cycles; the annealing temperature was then increased to 58 °C for the next 24 cycles, without a final extension step. The following primers were used: forward 5′- CCACCACTCTATTTTGTGCATCA-3′; reverse 5′- CAGTAGAAAAATTCCCCTCCACA-3′. The nested PCR of V3 *env* was performed on pooled products of the first amplification with the Phusion High-Fidelity DNA Polymerase (Thermo Scientific) in the presence of DMSO (3%) as follows: 30 s at 98 °C; 10 s at 98 °C, 30 s at 55 °C, and 20 s at 72 °C for 10 cycles; the annealing temperature was then increased to 60 °C for the next 15 cycles without a final extension step. The number of PCR cycles was limited to 25 to ensure that the amplification remained linear, as described above. The nested primers were specific fusion primers needed to fuse to the emulsion PCR beads required by the 454 technology. They also included a 4-nucleotide sequencing key "TCAG" to identify the DNA library, 10-nucleotide multiplex identifiers (MIDs) used as a DNA barcode to identify samples after sequencing was complete, and a V3-spanning degenerate sequence (forward 5′- ACAATGYACACATGGAATTARGCCA -3′; reverse 5′- AGAAAAATTCYCCTCYACAATTAAA -3′). The amplified PCR products were analyzed using a LabChip GX (Caliper) and then purified using Agencourt Ampure PCR Purification beads (Beckman Coulter) to remove small (<300 bp) fragments. The purified PCR products were the quantified using a Quant-iT Picogreen dsDNA Assay Kit (Invitrogen) on a LightCycler 480 (Roche) and diluted to a concentration of 1×109 molecules/μl.

V3 *env* ultra-deep pyrosequencing. Ultra-deep pyrosequencing was performed on a 454 GS Junior. PCR amplicons were combined and clonally amplified on DNA capture beads in water-in-oil emulsion micro-reactors at a ratio of 0.4 copies per capture bead. A total of 500,000 enriched-DNA beads were thus deposited in the wells of a full GS Junior Titanium PicoTiterPlate device and pyrosequenced in both forward and reverse directions. Bases were flowed sequentially and always in the same order (TCAG) across the wells of the PicoTiterPlate device during a 10-hour sequencing run generating long (500 bp) sequences.

Genotypic prediction of HIV-1 coreceptor usage from V3 ultra-deep pyrosequencing data. The sequences of the V3 *env* regions were first processed using GS Amplicon Variant Analyzer (AVA) software, version 2.5 p1 (Roche). This software extracts sequences from the standard flowgram format (SFF) files generated after pyrosequencing and automatically assigns each read to the proper sample by looking for the MIDs located at both ends of V3. Only sequences with an average phred equivalent quality score >Q30 were conserved. Moreover, only sequences that had been read in both senses were used for further analyses. The MIDs and primer sequences within the read have also to be complete without mismatch. Moreover, the reads have to match the full-length amplicon The sequence reads were aligned with the BaL consensus sequence (GenBank accession no. AY426110.1) and processed using an in-house automated data cleaning strategy (see Results) rather than the AVA filters. We used the combined 11/25 and net charge rule to infer the tropism of each virus clone from the V3 amino acid sequence. It requires one of the following criteria for predicting the CXCR4 coreceptor usage of HIV-1 subtype B[21,23]: (i) an R or K at position 11 of V3 and/or a K at position 25; (ii) an R at position 25 of V3 and a net charge of at least +5; and (iii) a net charge of at least +6. The V3 net charge was calculated by subtracting the number of negatively charged amino acids (D and E) from the number of positively charged ones (K and R). Subtype-specific algorithms derived from the combined 11/25 and net charge rule have been developed for subtypes D, and CRF01-AE[38,39].

Determining the global error rate of ultra-deep pyrosequencing of V3. We estimated the frequency of errors introduced during V3 amplification and GS Junior pyrosequencing by comparing the pyrosequencing reads to the Sanger sequences of 20 plasmid clones of *env* obtained from HIV-1 subtype B primary isolates. We first determined the frequency of artifactual V3 variants among the reads of each virus clone. The global error rate of ultra-deep pyrosequencing was then defined as the upper limit of the 99% confidence interval (Poisson statistics) of the mean frequency of artifactual V3 variants among the reads of the 20 clones.

Poisson distribution was applied to this global error rate to assess the risk of an artifactual V3 sequence being an authentic variant. We calculated the probability that a minor variant with n occurrences in N reads would occur n or more times if it was an error, using the following formula:

$$P = 1 - \sum_{k=1}^{n-1} \frac{e^{-\lambda}}{k!} \times \lambda^{k} \quad (1)$$

Here, λ is the expected number of artifactual sequences given N reads and is calculated by $\lambda = N^* \mu$, with μ being the global error rate (defined above). Only those variants whose frequency of occurrence yielded a P value of <0.001 according to the Poisson model were considered authentic.

Determining the position-specific error rates of ultra-deep pyrosequencing along the V3 sequence. As the errors generated during ultra-deep pyrosequencing are sequence-dependant, we determined specific error rates at each position in V3. We measured the mean codon error rate among the 20 clones at each V3 position. The position-specific error rates were then defined as the upper limit of the 99% confidence interval (Poisson statistics) of the mean frequency of artifactual codons among the 20 clones at each position of V3. We then determined weighted error rates to construct a sensitivity threshold matrix at each position of V3 to identify authentic virus variants harboring a point mutations for a given number of reads with $P < 0.001$.

Sensitivities of phenotyping and ultra-deep pyrosequencing for detecting and quantifying minor CXCR4-using variants. We assessed the capacity of ultra-deep pyrosequencing to detect and correctly quantify minor CXCR4-using variants in a virus population of CCR5-using variants using artificial mixtures of X4 and R5 virus clones that were phenotyped in parallel using the ultrasensitive TTT phenotypic assay[8]. Three artificial mixtures of X4 and R5 virus clones were used: LAI (GenBank accession no. K02013.1, X4 phenotype) and JR-CSF (GenBank accession no., R5 phenotype); AFG04 (GenBank accession no. DQ136796.1, X4 phenotype) and AFG01 (GenBank accession no. DQ136807.1, R5 phenotype); CHS02 (GenBank accession no. DQ136867.1, X4 phenotype) and CHS11 (GenBank accession no. DQ136859.1, R5 phenotype). AFG and CHS are primary HIV-1 isolates that had previously been cloned and phenotyped for CCR5 anf CXCR4 coreceptor usage[47]. The HIV-1 RNA in culture supernatants of the pure R5 and X4 clones was quantified using the COBAS Ampliprep/COBAS TaqMan HIV-1 test version 2.0 (Roche), and then mixed in defined proportions of X4:R5 viruses (0:100; 0.5:99.5; 1:99; 5:95; 20:80; 50:50; 75:25; and 100:0, each with 2–3 replicates). RNA was then extracted, adjusted to a total of 3,000 virus copies/reaction in triplicate, and submitted to ultra-deep pyrosequencing and phenotyping in parallel.

Statistics. Poisson statistics were calculated using R version 3.0.0.

Langage programming. PyroVir was written in the Java programming language and run with the Java 6.25 software.

References

1. Deng, H. *et al.* Identification of a major co-receptor for primary isolates of HIV-1. *Nature* **381,** 661–666 (1996).
2. Dragic, T. *et al.* HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. *Nature* **381,** 667–673 (1996).
3. Alkhatib, G. *et al.* CC CKR5: a RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1. *Science* **272,** 1955–1958 (1996).
4. Berger, E. A. *et al.* A new classification for HIV-1. *Nature* **391,** 240 (1998).
5. Dorr, P. *et al.* Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. *Antimicrobial agents and chemotherapy* **49,** 4721–4732 (2005).
6. Fatkenheuer, G. *et al.* Subgroup analyses of maraviroc in previously treated R5 HIV-1 infection. *N Engl J Med* **359,** 1442–1455 (2008).
7. Gulick, R. M. *et al.* Maraviroc for previously treated patients with R5 HIV-1 infection. *N Engl J Med* **359,** 1429–1441 (2008).
8. Raymond, S. *et al.* Development and performance of a new recombinant virus phenotypic entry assay to determine HIV-1 coreceptor usage. *J Clin Virol* **47,** 126–130 (2010).
9. Trouplin, V. *et al.* Determination of coreceptor usage of human immunodeficiency virus type 1 from patient plasma samples by using a recombinant phenotypic assay. *J Virol* **75,** 251–259 (2001).
10. Whitcomb, J. M. *et al.* Development and characterization of a novel single-cycle recombinant-virus assay to determine human immunodeficiency virus type 1 coreceptor tropism. *Antimicrobial agents and chemotherapy* **51,** 566–575 (2007).
11. Gonzalez, N. *et al.* A sensitive phenotypic assay for the determination of human immunodeficiency virus type 1 tropism. *J Antimicrob Chemother* **65,** 2493–2501 (2010).
12. Lin, N. H. *et al.* The design and validation of a novel phenotypic assay to determine HIV-1 coreceptor usage of clinical isolates. *J Virol Methods* **169,** 39–46 (2010).
13. Raymond, S., Delobel, P. & Izopet, J. Phenotyping methods for determining HIV tropism and applications in clinical settings. *Curr Opin HIV AIDS* **7,** 463–469 (2012).
14. Su, Z. *et al.* Response to vicriviroc in treatment-experienced subjects, as determined by an enhanced-sensitivity coreceptor tropism assay: reanalysis of AIDS clinical trials group A5211. *J Infect Dis* **200,** 1724–1728 (2009).
15. Fouchier, R. A. *et al.* Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J Virol* **66,** 3183–3187 (1992).
16. Hwang, S. S., Boyle, T. J., Lyerly, H. K. & Cullen, B. R. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science* **253,** 71–74 (1991).
17. Jensen, M. A. *et al.* Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *J Virol* **77,** 13376–13388 (2003).
18. Lengauer, T., Sander, O., Sierra, S., Thielen, A. & Kaiser, R. Bioinformatics prediction of HIV coreceptor usage. *Nat Biotechnol* **25,** 1407–1410 (2007).
19. Sing, T. *et al.* Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. *Antivir Ther* **12,** 1097–1106 (2007).

20. De Jong, J. J., De Ronde, A., Keulen, W., Tersmette, M. & Goudsmit, J. Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. *J Virol* **66,** 6777–6780 (1992).
21. Delobel, P. *et al.* Population-based sequencing of the V3 region of env for predicting the coreceptor usage of human immunodeficiency virus type 1 quasispecies. *J Clin Microbiol* **45,** 1572–1580 (2007).
22. Low, A. J. *et al.* Current V3 genotyping algorithms are inadequate for predicting X4 co-receptor usage in clinical isolates. *AIDS* **21,** F17–24 (2007).
23. Raymond, S. *et al.* Correlation between genotypic predictions based on V3 sequences and phenotypic determination of HIV-1 tropism. *AIDS* **22,** F11–16 (2008).
24. Archer, J. *et al.* Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4 (CXCR4)-using HIV-1 with ultra-deep pyrosequencing. *AIDS* **23,** 1209–1218 (2009).
25. Cooper, D. A. *et al.* Maraviroc versus efavirenz, both in combination with zidovudine-lamivudine, for the treatment of antiretroviral-naive subjects with CCR5-tropic HIV-1 infection. *J Infect Dis* **201,** 803–813 (2010).
26. Saliou, A. *et al.* Concordance between two phenotypic assays and ultradeep pyrosequencing for determining HIV-1 tropism. *Antimicrobial agents and chemotherapy* **55,** 2831–2836 (2011).
27. Abbate, I. *et al.* Detection of quasispecies variants predicted to use CXCR4 by ultra-deep pyrosequencing during early HIV infection. *AIDS* **25,** 611–617 (2011).
28. Vandenbroucke, I. *et al.* HIV-1 V3 envelope deep sequencing for clinical plasma specimens failing in phenotypic tropism assays. *AIDS research and therapy* **7,** 4 (2010).
29. Rozera, G. *et al.* Archived HIV-1 minority variants detected by ultra-deep pyrosequencing in provirus may be fully replication competent. *AIDS* **23,** 2541–2543 (2009).
30. Abbate, I. *et al.* Analysis of co-receptor usage of circulating viral and proviral HIV genome quasispecies by ultra-deep pyrosequencing in patients who are candidates for CCR5 antagonist treatment. *Clinical microbiology and infection* **17,** 725–731 (2011).
31. Dybowski, J. N., Heider, D. & Hoffmann, D. Structure of HIV-1 quasi-species as early indicator for switches of co-receptor tropism. *AIDS research and therapy* **7,** 41 (2010).
32. Tsibris, A. M. *et al.* Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy *in vivo*. *PLoS One* **4,** e5683 (2009).
33. Bunnik, E. M. *et al.* Detection of inferred CCR5- and CXCR4-using HIV-1 variants and evolutionary intermediates using ultra-deep pyrosequencing. *PLoS Pathog* **7,** e1002106 (2011).
34. Swenson, L. C. *et al.* Improved detection of CXCR4-using HIV by V3 genotyping: application of population-based and "deep" sequencing to plasma RNA and proviral DNA. *J Acquir Immune Defic Syndr* **54,** 506–510 (2010).
35. Kagan, R. M. *et al.* A genotypic test for HIV-1 tropism combining Sanger sequencing with ultradeep sequencing predicts virologic response in treatment-experienced patients. *PloS one* **7,** e46334 (2012).
36. Swenson, L. C., Daumer, M. & Paredes, R. Next-generation sequencing to assess HIV tropism. *Current opinion in HIV and AIDS* **7,** 478–485 (2012).
37. Gilles, A. *et al.* Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC genomics* **12,** 245 (2011).
38. Raymond, S. *et al.* Genotypic prediction of HIV-1 subtype D tropism. *Retrovirology* **8,** 56 (2011).
39. Raymond, S. *et al.* Genotypic prediction of HIV-1 CRF01-AE tropism. *J Clin Microbiol* **51,** 564–570 (2013).
40. Larsen, B. B. *et al.* Improved detection of rare HIV-1 variants using 454 pyrosequencing. *PloS one* **8,** e76502 (2013).
41. Di Giallonardo, F. *et al.* Next-generation sequencing of HIV-1 RNA genomes: determination of error rates and minimizing artificial recombination. *PloS one* **8,** e74249 (2013).
42. Brodin, J. *et al.* PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. *PloS one* **8,** e70388 (2013).
43. Brandariz-Fontes, C. *et al.* Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. *Scientific reports* **5,** 8056 (2015).
44. Lee, G. Q. *et al.* Comparison of population and 454 "deep" sequence analysis for HIV type 1 tropism versus the original trofile assay in non-B subtypes. *AIDS research and human retroviruses* **29,** 979–984 (2013).
45. Li, J. Z. *et al.* Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. *Jama* **305,** 1327–1335 (2011).
46. Symons, J. *et al.* Maraviroc is able to inhibit dual-R5 viruses in a dual/mixed HIV-1-infected patient. *J Antimicrob Chemother* **66,** 890–895 (2011).
47. Delobel, P. *et al.* Naive T-cell depletion related to infection by X4 human immunodeficiency virus type 1 in poor immunological responders to highly active antiretroviral therapy. *J Virol* **80,** 10229–10236 (2006).

Acknowledgements

French National Institute for Health and Medical Research-French National Agency for Aids and Viral Hepatitis Research (Inserm-ANRS). The English text was checked by Dr Owen Parkes.

Author Contributions

P.D. designed the project. N.J., A.S., J.I. and P.D. analyzed results and wrote the manuscript. N.J. wrote the PyroVir software. R.C., C.L., M.D., M.C., F.N. and S.R. performed the experimental work.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Jeanne, N. *et al.* Position-specific automated processing of V3 *env* ultra-deep pyrosequencing data for predicting HIV-1 tropism. *Sci. Rep.* **5**, 16944; doi: 10.1038/srep16944 (2015).