

# Minimizing Next-Generation Sequencing Errors for HIV Drug Resistance Testing

José A. Fernández-Caballero<sup>1</sup>, Natalia Chueca<sup>1</sup>, Eva Poveda<sup>2</sup> and Federico García<sup>1</sup>

<sup>1</sup>Department of Clinical Microbiology, Hospital Universitario San Cecilio de Granada, Instituto de Investigación Biosanitaria (IBIS), Granada, Spain;

<sup>2</sup>Division of Clinical Virology, Instituto de Investigación Biomédica de A Coruña (INIBIC)-Complejo Hospitalario Universitario de A Coruña (CHUAC), Sergas, Universidade da Coruña (UDC), La Coruña, Spain

## Abstract

**Next-generation sequencing prototypes for the routine diagnosis of resistance to antiretrovirals approved for the treatment of HIV infection are now being used in many clinical diagnostic laboratories. As some of the next-generation sequencing platforms may be a source of errors, it is necessary to improve the currently available protocols and implement bioinformatic tools that may help to correctly identify the presence of resistance mutations with clinical impact. Several studies have addressed these issues in recent years. Some of them are mainly focused on improving protocols for decreasing the magnitude of errors during the polymerase chain reaction. Other studies propose specific bioinformatic tools, able to reach both a 93-98% reduction of indels (insertions/deletions) and a sensitivity and specificity close to 100% in single nucleotide polymorphism variant calling. The implementation of new protocols and bioinformatic tools improving the accuracy of next-generation sequencing results must be considered for a correct analysis of HIV resistance mutations for making clinical decisions. This review summarizes the most relevant data available for the optimization of next-generation sequencing applied to HIV resistance testing. (AIDS Rev. 2017;19:231-8)**

Corresponding author: José Ángel Fernández-Caballero, jose.angel.fernandez.caballero@gmail.com

## Key words

**HIV. NGS. 454. Denoising. Filter. Indel.**

## Introduction

Next-generation sequencing (NGS) has revolutionized the studies of genomics characterization applied to virology<sup>1-4</sup>. This technology allows generating a large

amount of information in short periods of time and at a relatively low cost<sup>5</sup>. In RNA viruses, such as HIV, high rates of error-prone mutations during viral replication generate a multitude of genetically diverse but similar variants known as quasispecies<sup>6</sup>. The NGS technologies are very useful for the genetic characterization of HIV infection, allowing an in depth analysis of the spectrum of variants present in an HIV-infected patient<sup>7-9</sup>. NGS has a technical limit of detection of 0.01% and a clinical threshold above 1% for mutations present on the viral population compared with the 15-20% obtained using Sanger (population) sequencing. Therefore<sup>10-12</sup>, NGS provides both technical and clinical improvements for the detection of resistance to anti-retroviral drugs against HIV infection<sup>13</sup>. More specifically, NGS improvements are especially for choosing

### Correspondence to:

José Ángel Fernández-Caballero  
Department of Clinical Microbiology  
Hospital Universitario San Cecilio de Granada  
Instituto de Investigación Biosanitaria (IBIS)  
Av. Del Conocimiento, s/n  
18016 Granada, Spain  
E-mail: jose.angel.fernandez.caballero@gmail.com

first-line antiretroviral regimens including drugs with low genetic barrier for resistance<sup>14-16</sup>. This is the case for the non-nucleoside reverse transcriptase inhibitors (NNRTI) (i.e. efavirenz)<sup>17</sup>, and potentially also for integrase inhibitors (INI) in the next future. In addition, NGS has also proven to be of value for salvage therapy<sup>14</sup> and as a surrogate of the cumulative resistance in the failing patient.

The 454 GS Junior platform (454 Life Sciences/Roche Diagnostics) has been widely used across some countries for testing HIV resistance mutations in the clinical setting<sup>18-20</sup> in the last two years. Roche Diagnostics has distributed a prototype allowing sequencing reverse transcriptase (RT), protease and integrase fragments, with an appropriate amplicon length of 300-500 base pairs (bp), which include all resistance associated codons described to date<sup>21</sup>.

Although technically sound, 454 technologies have several handicaps. As the method is based on pyrosequencing, homopolymeric regions serve as an error source leading frequently to insertions and deletions (indels)<sup>22</sup>. In addition, the quality of the reads decreases at the end of the reaction<sup>23</sup>. Finally, an additional source of error for HIV resistance analysis is that the protocol is based on a reverse transcribed polymerase chain reaction (RT-PCR). These errors may highly impact the quality of the sequences, with a significant impact on sequence assembly, polymorphism detection, HIV subtype assignment, and resistance analysis<sup>24</sup>.

Most NGS platforms, including 454, incorporate quality control (QC) pipelines in their sequencing protocol to filter the final results. However, as these QC pipelines may be insufficient, it is reasonable to run an own user level additional QC based on high-quality control filtering. This software has been widely used for microbiome analysis based on the 454 platforms, aiming to improve the quality of results<sup>25-28</sup>.

To date, there is no systematic review on the use of methods that can correct sequencing errors on the 454 platforms with protocols used for the characterization of HIV resistance in the clinical routine. In this study, we have reviewed all studies dealing with software or methods aiming to decrease these errors, which have been published during the period 2006-2016.

## Selection criteria and systematic review

We considered as bioinformatic strategies, software aiming to delete or detect sequencing errors, and as

protocol improvements those changes in PCR temperature profiles and/or reagent concentration aiming to minimize sequencing errors.

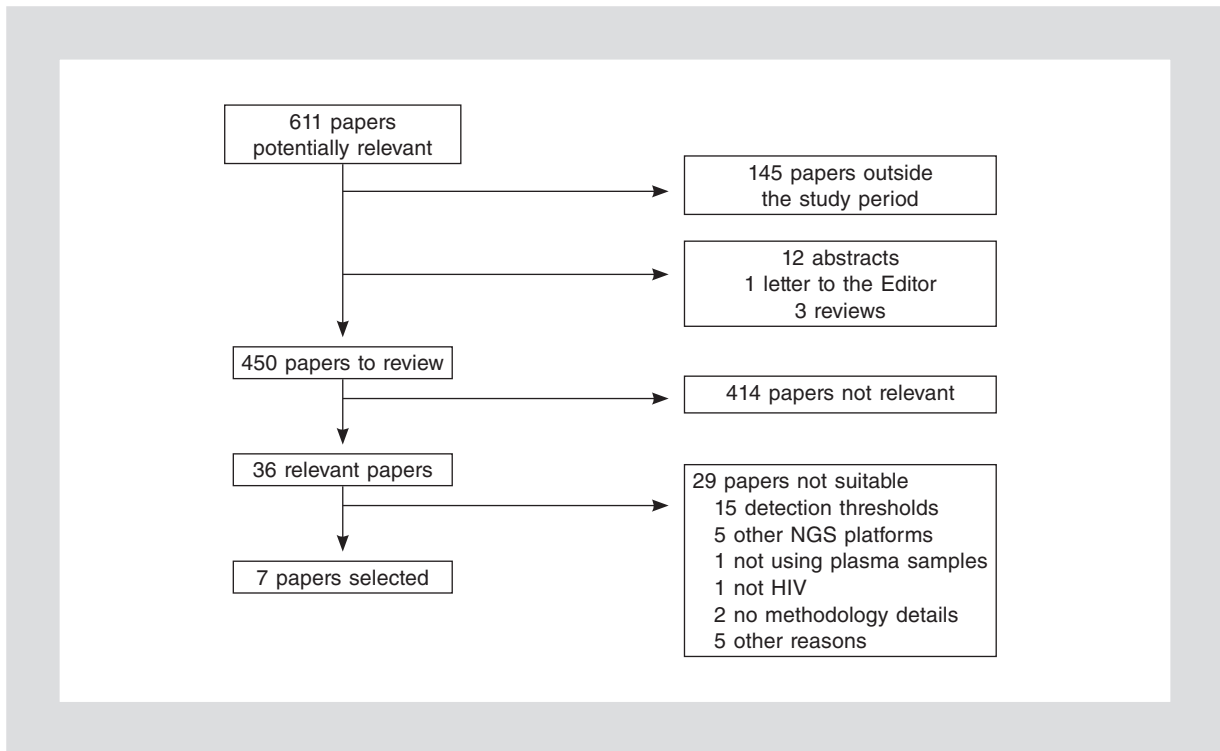
Our systematic review was performed according to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidance<sup>29</sup>. We used a combination of non-MeSH and MeSH terms related with error correction and NGS sequence filtering. We searched PubMed, Medline (Ovid) and Embase (Ovid). The term "HIV" was combined with "NGS" and "Next generation sequencing" and "Error rates" and "454" and "Artifact recombination" and "Denoising" and "Filter sequence" and "Insertion-deletion" and "Carry forward correction".

All abstracts of papers available through January 2006 and June 2016 were reviewed. We used an iterative search strategy, as all references in the papers that were selected for reviewing were also studied. We limited our review to original articles that evaluated error correction of HIV NGS sequences obtained on the 454 platforms. Studies had to provide enough information on how sequence filtering was performed, and which modification before/after processing was performed. Abstracts, comments, and letters to the Editor were excluded as they lacked information on sequence processing.

## Data analysis and risk of bias

JA.F reviewed the studies, excluding those with irrelevant titles and references; N.C and JA.F independently read studies that were selected, paying attention to abstracts, descriptive terms, and titles, and identified potentially eligible papers. Both authors finally reviewed the full text of the papers, and applied inclusion criteria. There were no discrepancies among both authors' selection. JA.F and N.C independently extracted the data from the selected studies, providing them in a separate standardized file. Again, no discrepancies between authors were found. When provided, crude 95% confidence intervals (95% CI) were extracted; they were calculated when not available and enough data were provided in the papers. However, due to the high heterogeneity of most studies and different methodologies, meta-analysis could not be performed.

Seven studies were selected, and all were estimated to have no overall risk of bias (supplementary data). Only one study<sup>30</sup> had a high risk of bias, due to the scarce information provided by the authors for all the parameters we evaluated.



**Figure 1.** Flowchart for study selection. NGS: next-generation sequencing.

JA.F and N.C evaluated the overall quality of the studies, pointing out their strengths and weakness, according to STROBE recommendations<sup>31</sup>. Three categories were given for the quality parameter (high, medium, low). Here, several discrepancies were found between the two reviewers. When this happened, F.G was asked to re-evaluate and discrepancies were solved accordingly.

### Selection of the studies

Our search selected 611 studies. After removing all the studies that were not research studies and by the date in which NGS was introduced, 161 papers were removed. After irrelevant studies were removed, only 36 studies remained eligible. From these, 29 papers were removed, 15 because they focused only on detection cut-offs, five because they did not deal with 454 NGS errors, two due to a poor description of the methodology used, and seven for some other reasons (non-plasma samples, non-HIV studies, and not showing the results). We finally selected seven papers that met all the eligibility criteria (Fig. 1). Three of them dealt with protocol modifications to avoid or diminish sequencing errors<sup>32-34</sup>. These studies reported data on the

number of reads, the percentage of recombination, and error rates (Table 1). The other four studies dealt with bioinformatics aiming to eliminate errors<sup>30,35-37</sup>. In this case, the studies provided data on the number of reads, error rate, indels, single nucleotide polymorphism (SNP) variant calling, and software characteristics (Table 1).

### Modifications to the amplification protocol

Some previous reports have shown how error rates, as well as chimera and indel formation, increase after 30 cycles of amplification, whatever the fidelity of the polymerase may be<sup>38-40</sup>. Three of the studies we have included in this systematic review aimed to minimize sequencing errors by modifying the amplification protocol<sup>32-34</sup> (Table 2). Two of the studies we have evaluated optimized the PCR by lowering the number of PCR cycles and doubling PCR reagent concentrations, with no variation on RNA input. One of the studies estimated that median (95% CI) recombination was 48.7% (53.6-43.9) for standard PCR, compared to 0.8% (0.07-2.1) after PCR optimization<sup>32</sup>. The error rate for both, standard and optimized PCR, was also analyzed; a drop from 23.2% (95% CI: 18.8-27.6) for standard PCR

**Table 1. Studies based on modifications to the amplification protocol or bioinformatics solutions**

Study	Year	NGS platform	Region	Design	Procedure
Di Giallonardo, et al. <sup>32</sup>	2013	454 FLX	Protease (271 bp)	Protocol modification; Use of several HIV-1 strains to allow recombination. PCR reagent optimization	1) Standard PCR: 40 cycles (94°C 15"/55°C 30"/72°C 30") plus 72°/8'. Reagents: Primers (uM) 0.4; dNTPs (mM) 0.2; FastStart High Fidelity DNA polymerase (U) 1.25. 2) Optimized PCR: 35 cycles (94°C 30"/55°C 60"/72°C 60"). Primers (uM) 1; dNTPs (mM) 0.4; FastStart High Fidelity DNA polymerase (U) 3
Shao, et al. <sup>33</sup>	2013	454 FLX	RT (pol)	Protocol modification; Use of several HIV-1 strains to allow recombination. PCR reagent optimization	1) Standard PCR: 95°C 2'; 45 cycles (95°C 30"/50°C 30"/72°C 30"). Reagents; Primers (Nm) 400; dNTPs (uM) 200, MgSO <sub>4</sub> (mM) 4; Hi Fidelity Platinum Taq (U) 2.5. 2) Optimized PCR: 95°C 15'; 25 cycles (95°C 15"/51°C 30"/68°C 1'30"). Primers (uM) 1; dNTPs (uM) 200; MgCl <sub>2</sub> (mM) 2.3; Taq Gold (U) 5
Waugh, et al. <sup>34</sup>	2015	454 FLX	Gag	Protocol modification; Use of several HIV-1 strains to allow recombination. Changes in PCR cycle number and RNA input	Two PCRs in parallel for each of 3 RNA inputs (ng): 160, 1600, 3990. 1 <sup>st</sup> PCR: 27 cycles; 2 <sup>nd</sup> PCR: 35 cycles; 98°C 30"/98°C 10"/72°C 1'. Primers (Nm) 400; dNTP (uM) 200; Phusion DNA polymerase (U) 0.3
Iyer, et al. <sup>35</sup>	2013	454 FLX	gag/env/nef (1,500 bp/2550/681)	Software: "CorQ". Data learning for massive sequencing data generated from a mixture of strains, and from artificial SNP errors	CorQ that utilizes a multiple sequence alignment to map base qualities to the positions within the alignment. Reads bases according to coverage, quality between adjacent bases, and the base in question
Deng, et al. <sup>36</sup>	2013	454 FLX	gag and pol (269 bp-443 bp)	Software: "Indel and Carryforward Correction (ICC)" Data learning for massive sequencing data generated from a mixture of strains	Unique sequences ranking by their abundance, and align the sequences taking into account the abundances of the aligned unique sequences and scoring parameters; match, mismatch, gap, penalty. Reads are filtered based on parameters; ambiguous bases, length, and average quality
Brodin, et al. <sup>37</sup>	2013	454 FLX	pol (167 bp)	Software: "BioPerl" using error correction scripts. Data learning from env SG3Δ plasmid subjected to nested PCR (30+30 cycles) using Faststart high fidelity	Filtering reads with: less than 80% similarity to a user reference sequence, taking ambiguous nucleotide calls, indels and stop codons into consideration
Kijak, et al. <sup>30</sup>	2013	454 FLX	Pol	Software: "Nautilus"	Using as an input an alignment determines the nucleotide base frequency and read depth at each position and computes the haplotype frequencies based on the linkage among polymorphisms. Also computes the frequency of the variants in the setting of their sequence context and mapping orientation

Bp: base pairs; dNTP: deoxynucleotide; NGS: next-generation sequencing; PCR: polymerase chain reaction; SNP: single nucleoside polymorphism.

**Table 2. Standard vs. optimized polymerase chain reaction: differences in the number of reads and other quality variables**

Study	Reads median (95% CI)	Recombination: % (95% CI)	Error rate: % (95% CI)
Di Giallonardo, et al. <sup>32*</sup>	S; 52,389 (84,645-20,133) O; 7,827.5(23,15.6-14,548.7)	S; 48.7(53.6-43.9) O; 0.8 (0.07-2.1)	S; 23.2 (18.8-27.6) O; 2 (0.6-3.1)
Shao, et al. <sup>33</sup>	S; 12,2327 O; 62,437	S; 12 O; 0.8	
Waugh, et al. <sup>34</sup>		27 cycles [ RNA 160 ng 0.006 (0-0.01) RNA 1600 ng 0.1 (0.08-0.1) RNA 3,990 ng 0.9 (0.4-1.8)	
		35 cycles [ RNA 160 ng 2.9 (2.8-3.2) RNA 1,600 ng 4.6 (4.5-4.8) RNA 3,990 ng 1.1 (0.8-1.4)	

\*Not provided in the study (calculated).  
O: optimized; PCR: polymerase chain reaction; S: standard.

to 2% (95% CI: 0.6-3.1) was described. A second paper also observed a reduction in median values of recombination from 12% (standard PCR) to 0.8% (optimized PCR)<sup>33</sup>. Finally, the last study used different RNA inputs, changing PCR conditions, but with the same reagent concentrations<sup>34</sup>. A higher number of PCR cycles always resulted in a higher degree of recombination. This study demonstrates that a higher RNA input at first stages of amplification, together with

a reduction in the number of cycles, minimizes sequencing errors.

### Bioinformatics-based solutions

Four studies evaluated the use of bioinformatics to locate and diminish/eliminate sequencing errors<sup>30,35-37</sup> (Table 3). Two of them tested for different variables before and after they were used, and describe less

**Table 3. Main characteristics of bioinformatics tools available to eliminate next-generation sequencing errors**

Study	Reads median (95% CI)	Error rate: % (95% CI)		Indel	SNP variant calling	Software characteristics					
		Prior to filtering	Post filtering			Chim	Trim	Var L (n)	Qual.	k	
Iyer, et al. <sup>35*</sup>	26,620 gag 48,927 env 21,963 nef			Reduction 93-97%	Sensitivity: 99% Specificity: 88%	✓	?	✓	✓	✓	✓
Deng, et al. <sup>36*</sup>	12,617 (8,247-16,987) gag 17,228.5 (12,001-22,456) pol	0.3 (0-1)	gag: 0.02 (0-0.04) pol: 0.01 (0-0.04)	Reduction 98-99%	Sensitivity: 100% Specificity: 98%	✓	✓	✓	✓	✓	✓
Brodin, et al. <sup>37*</sup>	Prior to filtering; 8,749.5 (4,137.1-11,760.5) Post filtering; 5,394 (1,958.4-8,951.6)	0.2 (0.08-0.4)	0.06 (0.05-0.08)			?	?	?	?	?	?
Kijak, et al. <sup>30</sup>						?	?	?	?	?	?

\*Not provided in the study (calculated).  
Chim: chimeric; k: k-mer; n: number of reads; trim: trimming; qual: quality scores; SNP: single-nucleotide polymorphism; var L: variation lengths.

**Table 4. GS Reference Mapper characteristics**

Region	Reads median (95% CI)		Error rate: % (95% CI)		Indel	SNP variant calling	Software characteristics					
	Prior to filtering	Post filtering	Prior to filtering	Post filtering			Chim	Trim	Var L	(n)	Qual	k
pol (533 bp)	7,214 (5,321-9,106)	7,184 (5,299-9,069)	0.3 (0.18-0.42)	0.03 (0.02-0.05)	Reduction 99%	Sensitivity: 99% Specificity: 98%	✓	✓	✓	✓	✓	✓

Bp: base pair; Chim: chimeric; k: k-mer; n: number of reads; trim: trimming; qual: quality scores; SNP: single-nucleotide polymorphism; var L: variation lengths.

numbers of errors after their usage<sup>36,37</sup>. Deng, et al., used the indel and Carryforward Correction (ICC) software, finding that error rates decreased from a median value (95% CI) of 0.3% (0-1) for both gag and pol regions, to 0.02% (0-0.04) for gag and 0.01% (0-0.04), after using ICC<sup>36</sup>. Indels were reduced in 98-99%, and sensitivity and specificity for SNP variant calling was 100 and 98%, respectively. Brodin, et al., used BioPerl for sequence processing: error rates decreased from a median value (95% CI) of 0.2% (0.008-0.4) before processing to 0.06% (0.05-0.08); using the software resulted in the filtering of approximately 2,000 reads per sample (from 8,749.5 reads [95% CI: 4,137.1-11,760.5], to 5,394 reads (95% CI: 1,958.4-8,951.6))<sup>37</sup>. Iyer, et al. used CorQ that resulted in a reduction on indels of 93-97% for gag/env/nef regions, with a sensitivity and specificity of 99 and 88%, respectively, for SNP variant calling<sup>35</sup>. The last study, performed by Kijak, et al., used Nautilus software, which is an observational tool that helps to discern between errors; in contrast to the other three, this study did not show any data on the parameters that were analyzed<sup>30</sup>.

Overall, bioinformatics software have shown to be highly efficacious to remove sequences that contained sequencing errors, with a 93-99% reduction in indels, which is highly relevant, specially for 454-based NGS<sup>35-37</sup>. In a similar way, using these programs also decreased error rate, allowing a more accurate estimation of variants, especially when they are present at a low relative proportion (1-5%). Filtering may result in a significant loss in the number of reads, and may enable NGS to go down to 1% for minor variant detection. All the software that have been reviewed in this paper did not incur a high loss in the number of reads; therefore, this was not seen to be a problem.

All bioinformatics software also showed excellent results for SNP variant calling, with sensitivity and specificity values near 100%.

The GS Reference Mapper in another tool for bioinformatic analysis. This software filters the sequences based on quality parameters, length of the sequences, expected deep and variant coverage. The GS Reference Mapper allowed improving the quality of sequences, especially at their terminal end, improving the Q values from 10 to 25 (Table 4). The Q value is an index of the probability of a given base of being a sequencing error; Q values > 30 are optimal, as the probability error is between 1 and 1000<sup>41</sup>. This software has been used for the analysis of 73 samples from HIV-1-infected patients (84.2% subtype B) that were sequenced (RT and protease) on a 454 GS Junior. After denoising, most of the resistance mutations that were corrected by GS Reference Mapper (43/68) had been detected at very low relative prevalence (1-2%), and all were below 5%. In addition, after filtering, only a small number of sequences are deleted (mean 95% CI: 30; 22-37), and it shows a high sensitivity (99%) and specificity (98%) for SNP variant calling. This software was able to eliminate HIV resistance mutations that were artificially detected at low levels, ranging 1-5% of the whole quasispecies population that was infecting the patients.

## Recommendations and conclusions

This is the first systematic review evaluating the benefits of sequence processing after massive parallel sequencing for the characterization of HIV drug resistance. There are specific technical recommendations and several bioinformatics software that can be very useful to improve the results obtained using 454 GS

**Table 5. Technical recommendations to minimize errors**

What	How
Lower the number of cycles during PCR	Use 5-7 cycles less
Use high fidelity polymerases	Phusion High-Fidelity DNA Polymerase or pfu DNA polymerase
Use bioinformatic software to filter short and low quality sequences	GS Reference Mapper or CorQ
Filter all bases with a Q value < 25, and sequence depth 10	Delete indels that can affect the final results

PCR: polymerase chain reaction.

Junior NGS. The recommendations proposed in this review will certainly help NGS users to be really confident that HIV resistance mutations detected using 454 protocols are true mutations and not test-associated artifacts, especially if they are at low proportions (1-5%) in the whole viral population. Implementing these recommendations will certainly be of benefit and will improve patient care.

Considering all these data, there are some specific technical recommendations that might help for the minimization of the errors generated using the 454 GS Junior NGS platform (Table 5). Some of them include; lowering the number of cycles during PCR, using high-fidelity polymerases (Phusion High-Fidelity DNA Polymerase or pfu DNA polymerase) and using bioinformatic software to filter short and low quality sequences, including all bases with a Q value < 25, and sequence depth 10, might help for the minimization of the errors generated using the 454 GS Junior NGS platform.

## Supplementary Data

Supplementary data is available at AIDS Reviews journal online (<http://www.aidsreviews.com>).

This data is provided by the author and published online to benefit the reader. The contents of all supplementary data are the sole responsibility of the authors.

## Acknowledgements

This review is part of the results of the doctoral thesis of Jose Angel Fernandez-Caballero Rico, enrolled in a doctorate program in clinical medicine and public health at the University of Granada.

## Declaration of interest

This work was supported in part by grants from RD12/0017/006 (Plan Nacional de I+D+I and Fondo Europeo de Desarrollo Regional-FEDER). Federico García has received a research extension grant from the Programa de Intensificación de la Actividad de Investigación del Servicio Andaluz de Salud.

## References

1. Visser M, Bester R, Burger JT, Maree HJ. Next-generation sequencing for virus detection: covering all the bases. *Virology*. 2016;13:85.
2. McGinnis J, Laplante J, Shudt M, George KS. Next generation sequencing for whole genome analysis and surveillance of influenza A viruses. *J Clin Virol*. 2016;79:44-50.
3. Thorburn F, Bennett S, Moshia S, Murdoch D, Gunson R, Murcia PR. The use of next generation sequencing in the diagnosis and typing of respiratory infections. *J Clin Virol*. 2015;69:96-100.
4. Quan PL, Wagner TA, Briesse T, et al. Astrovirus encephalitis in boy with X-linked agammaglobulinemia. *Emerg Infect Dis*. 2010;16:918-25.
5. Dudley DM, Chin EN, Bimber BN, et al. Low-cost ultra-wide genotyping using Roche/454 pyrosequencing for surveillance of HIV drug resistance. *PLoS One*. 2012;7:e36494.
6. Biebricher CK, Eigen M. What is a quasispecies? *Curr Top Microbiol Immunol*. 2006;299:1-31.
7. Zukrov JP, do Nascimento-Brito S, Volpini AC, Oliveira GC, Janini LM, Antoneli F. Estimation of genetic diversity in viral populations from next generation sequencing data with extremely deep coverage. *Algorithms Mol Biol*. 2016;11:2.
8. Nicot F, Sauné K, Raymond S, et al. Minority resistant HIV-1 variants and the response to first-line NNRTI therapy. *J Clin Virol*. 2015;62:20-4.
9. Fisher RG, Smith DM, Murrel B et al. Next generation sequencing improves detection of drug resistance mutations in infants after PMTCT failure. *J Clin Virol*. 2015;62:48-53.
10. Ram D, Leshkowitz D, Gonzalez D, et al. Evaluation of GS Junior and MiSeq next-generation sequencing technologies as an alternative to Trugene population sequencing in the clinical HIV laboratory. *J Virol Methods*. 2015;212:12-16.
11. Mohamed S, Penaranda G, Gonzalez D, et al. Comparison of ultra-deep versus Sanger sequencing detection of minority mutations on the HIV-1 drug resistance interpretations after virological failure. *AIDS*. 2014;28:1315-24.
12. Stelzl E, Pröll J, Bizon B, et al. Human immunodeficiency virus type 1 drug resistance testing: Evaluation of a new ultra-deep sequencing-based protocol and comparison with the TRUGENE HIV-1 Genotyping kit. *J Virol Methods*. 2011;178:94-7.
13. Liang B, Luo M, Scott-Herridge J, et al. A comparison of parallel pyrosequencing and sanger clone-based sequencing and its impact on the characterization of the genetic diversity of HIV-1. *PLoS One*. 2011;6:e26745.
14. Pou C, Noguera-Julian M, Pérez-Álvarez S, et al. Improved prediction of salvage antiretroviral therapy outcomes using ultrasensitive HIV-1 drug resistance testing. *Clin Infect Dis*. 2014;59:578-88.
15. Simen BB, Simons JF, Hullsiek KH, et al. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. *J Infect Dis*. 2009;199:693-701.
16. Messiaen P, Verhofstede C, Vandenbroucke I, et al. Ultra-deep sequencing of HIV-1 reverse transcriptase before start of an NNRTI-based regimen in treatment-naive patients. *Virology*. 2012;426:7-11.
17. Wang J, Zhang G, Bambara RA, et al. Nonnucleoside reverse transcriptase inhibitor-resistant HIV is stimulated by efavirenz during early stages of infection. *J Virol*. 2011;85:10861-73.
18. Chen X, Zou X, He J, Zheng J, Chiarella J, Kozal MJ. HIV drug resistance mutations (DRMs) detected by deep sequencing in virologic failure subjects on therapy from Hunan Province, China. *PLoS One*. 2016;11:e0149215.

19. Fernández-Caballero JA, Chueca N, Alvarez M, et al. Usefulness of Integrase resistance testing in proviral HIV-1 DNA in patients with Raltegravir prior failure. *BMC Infect Dis.* 2016;16:197.
20. Dauwe K, Staelens D, Vancoillie L, Mortier V, Verhofstede C. Deep sequencing of HIV-1 RNA and DNA in newly diagnosed patients with baseline drug resistance showed no indications for hidden resistance and is biased by strong interference of hypermutation. *J Clin Microbiol.* 2016;54:1605-15.
21. Gall A, Ferns B, Morris C, et al. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *J Clin Microbiol.* 2012;50:3838-44.
22. Loman NJ, Misra RV, Dallman TJ, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol.* 2012;30:434-9.
23. Patel RK, Jain M. NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. *Plos One.* 2012;7:e30619.
24. Fernández-Caballero Rico JA, Chueca Porcuna N, Alvarez Estévez M, Mosquera Gutiérrez MD, Marcos Maeso MA, García F. [A safe and easy method for building consensus HIV sequences from 454 massively parallel sequencing data]. *Enferm Infecc Microbiol Clin.* [Epub ahead of print].
25. Reeder J, Knight R. Rapid denoising of pyrosequencing amplicon data: exploiting the rank-abundance distribution. *Nat Methods.* 2010;7:668-9.
26. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D. Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet.* 2012;13:47-58.
27. Gibson J, Shokralla S, Porter TM, et al. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasytematics. *Proc Natl Acad Sci USA.* 2014;111:8007-12.
28. Aagaard K, Riehle K, Ma J, et al. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLoS One.* 2012;7:e36466.
29. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev.* 2015;4:1.
30. Kijak GH, Pham P, Sanders-Buell E, et al. Nautilus: a bioinformatics package for the analysis of HIV type 1 targeted deep sequencing data. *AIDS Res Hum Retroviruses.* 2013;29:1361-4.
31. Von Elm E, Altman DG, Egger M, Pocock SJ, Gotsche PC, Vandenbroucke JP. The Strengthening of Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Int J Surg.* 2014;12:1495-9.
32. Di Giallonardo F, Zagordi O, Duport Y, et al. Next-generation sequencing of HIV-1 RNA genomes: determination of error rates and minimizing artificial recombination. *Plos One.* 2013;8:e74249.
33. Shao W, Boltz VF, Spindler JE, et al. Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of low-frequency drug resistance mutations in HIV-1 DNA. *Retrovirology.* 2013;10:18.
34. Waugh C, Cromer D, Grimm A, et al. A general method to eliminate laboratory induced recombinants during massive, parallel sequencing of cDNA library. *Virology.* 2015;12:55.
35. Iyer S, Bouzek H, Deng W, Larsen B, Casey E, Mullins JI. Quality score based identification and correction of pyrosequencing errors. *Plos One.* 2013;8:e73015.
36. Deng W, Maust BS, Westfall DH, et al. Indel and Carryforward Correction (ICC): a new analysis approach for processing 454 pyrosequencing data. *Bioinformatics.* 2013;29:2402-9.
37. Brodin J, Mild M, Hedskog C, et al. PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. *Plos One.* 2013;8:e70338.
38. Smyth RP, Schlub TE, Grimm A et al. Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene.* 2010;469:45-51.
39. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol.* 2005;71:8966-9.
40. Sipos R, Szekeley AJ, Palatinszky M, Revesz S, Marialigeti K, Nikolausz M. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targeting bacterial community analysis. *FEMS Microbiol Ecol.* 2007;60:341-50.
41. Brockman W, Alvarez P, Young S, et al. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.* 2008;18:763-70.