Antiviral Research 98 (2013) 273–283

Contents lists available at SciVerse ScienceDirect

Antiviral Research

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

A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model

Clara Ramírez [a], Josep Gregori [b], Maria Buti [c,d], David Tabernero [a,d], Sílvia Camós [a,d], Rosario Casillas [a,b], Josep Quer [b,d], Rafael Esteban [c,d], Maria Homs [a,d], Francisco Rodriguez-Frías [a,d,*]

[a] Biochemistry Department, Hospital Vall d'Hebron, Universitat Autònoma de Barcelona, Spain
[b] Liver Unit, Research Institute Vall d'Hebron, Universitat Autònoma de Barcelona, Spain
[c] Hepatology Department, Hospital Vall d'Hebron, Universitat Autònoma de Barcelona, Spain
[d] Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Instituto Carlos III, Spain

ARTICLE INFO

Article history:
Received 16 December 2012
Revised 5 March 2013
Accepted 11 March 2013
Available online 20 March 2013

Keywords:
Quasispecies
HBV
Ultra-deep pyrosequencing
Cloning
Polymorphic positions
Haplotypes

ABSTRACT

In this study, the reliability and reproducibility of viral quasispecies quantification by three ultra-deep pyrosequencing (UDPS) methods (FLX+, FLX, and Junior) were investigated and results compared with the conventional cloning technique. Hepatitis B virus (HBV) infection was selected as the model. The pre-Core/Core region, the least overlapped HBV region, was analyzed in samples from a chronic hepatitis B patient by cloning and by UDPS.

After computation filtering of the UDPS results, samples A1 and A2 (FLX+) and sample B (FLX) yielded the same 20 polymorphic positions. Junior yielded 18 polymorphic positions that coincided with the FLX results. In contrast, 50 polymorphic positions were detected by cloning. Quasispecies complexity plotted on graphs showed superimposed patterns and the quantitative parameters were similar between FLX+, FLX, Junior, and the cloning sequences. Twenty-two haplotypes were detected by Junior, and 37, 40, and 39 were detected by FLX A1, A2, and B, respectively. These differences may be attributable to methodological differences between FLX and Junior. By cloning, 47 haplotypes were detected. Eight clones with insertions and deletions that induced *de novo* stop codons were not observed by UDPS because the UDPS filter discarded them.

Our results indicate that UDPS is an optimal alternative to molecular cloning for quantitative study of the viral quasispecies. Nonetheless, specific mutations, such as insertions and deletions, were only detected by cloning. A filter should be designed to analyze cloning sequences, and UDPS filters should be improved to include the specific mutations.

1. Introduction

The evolutionary dynamics of many viruses, including most RNA viruses is characterized by high turnover, high mutation rates and large population sizes, yielding a large number of viral variants and highly heterogeneous populations, which are often referred to as viral quasispecies (Eigen, 1971; Domingo and Holland, 1997; Nowak et al., 1996). Quasispecies infection is commonly defined as an infection caused by a complex distribution of variants that are genetically distinct, but closely related. This swarm of variants is dynamic, and those viruses with the highest fitness will be selected when environmental changes occur. RNA viruses such as human immunodeficiency virus (HIV) and hepatitis C virus (HCV), and some DNA viruses, such as hepatitis B virus (HBV), which present a quasispecies distribution, account for millions of infections worldwide (Domingo and Gomez, 2007). Study of quasispecies evolution is essential for optimal management of these infections. Genetic diversity is considered an important factor in HIV disease progression (Shankarappa et al., 1999), pathogenesis (Vignuzzi et al., 2006), immune escape (Nowak et al., 1991), vaccine design (Gaschen et al., 2002), and the development of drug resistance (Johnson et al., 2008; Tsibris et al., 2009).

Until recently, the genetic diversity of a viral population could only be assessed by cloning individual viruses. Among other applications, this method has been used to study the diversity of viral quasispecies (Martell et al., 1992) and to detect mutations conferring resistance to antiviral treatment (Cubero et al., 2008). Cloning of viral fragments has been widely used as the reference method to

* Corresponding author. Address: Liver Pathology Unit, Department of Biochemistry, Hospital Valld'Hebron, Valld'Hebron Avenue, 119-129, 08035, Barcelona, Spain. Tel.: +34 93 274 6991; fax: +34 93 274 68 31.
E-mail address: frarodri@gmail.com (F. Rodriguez-Frías).

0166-3542/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.antiviral.2013.03.007

study viral populations, but it has notable drawbacks: it is a time-consuming technique and its sensitivity is limited by the number of clones that can be feasibly sequenced. Thus, to survey overall viral populations in appreciable detail, cloning is a labor-intensive technique that is costly when obtaining a significant number of clones. In addition, errors can be introduced by the polymerases or during the sequencing reaction, and these errors can mistakenly contribute to the observed variability.

Next-generation sequencing (NGS) (Margulies et al., 2005) has implied an important step ahead in genome analysis, as in complete human genome sequencing (Porreca et al., 2007), identification of new microorganisms (Hormozdiari et al., 2009), cancer research (Campbell et al., 2008), and the study of viral quasispecies (Hoffmann et al., 2007; Swenson et al., 2011; Bull et al., 2011; Solmone et al., 2009; Margeridon-Thermet et al., 2009; Homs et al., 2011a; Rodriguez-Frias et al., 2012; Homs et al., 2012; Ghedin et al., 2009; Wright et al., 2011). NGS can overcome the limitations of classic clonal Sanger sequencing by direct parallel clonal sequencing of mixed samples, resulting in more than 10 000 reads per base. In the study of viral quasispecies, ultra-deep pyrosequencing (UDPS) was initially used in HIV infection for detecting low-frequency resistance mutations to antiviral treatments (Hoffmann et al., 2007) and studying HIV tropism (Swenson et al., 2011). Application of UDPS to other viral infections has increased rapidly, including the study of HCV transmission bottle-necks (Bull et al., 2011), HBV diversity, HBV low-frequency drug resistance mutations (Rodriguez-Frias et al., 2012; Solmone et al., 2009; Homs et al., 2011a; Homs et al., 2012; Margeridon-Thermet et al., 2009), and mixed influenza infections (Ghedin et al., 2009), and foot-and-mouth disease virus diversity (Wright et al., 2011). As compared to cloning, UDPS has the advantages of being a simpler, less time-consuming technique, with considerable sensitivity for detecting minor populations. UDPS has provided substantial information on the structure and dynamics of viral populations by setting the types and frequencies of various low-frequency variants (Wang et al., 2007).

NGS experiments, such as UDPS, generate high volumes of data that require a powerful, complex pipeline system for storage, management, and processing of the data. In addition, as has been described, PCRs and sequencing can induce errors (Beerenwinkel and Zagordi, 2011; Gorzer et al., 2010) that must be detected and discarded; hence, accurate filtering criteria are needed (Wang et al., 2007). The complexity of NGS raw data analysis may be the reason why these technologies have mainly been used in qualitative approaches, and there are no studies studies validating their reproducibility. Among the more consolidated NGS platforms, UDPS is the one most often used to investigate viral quasispecies because it enables study of the longest fragments (400 bp).

The GS 454 platform includes three different methods, FLX+, FLX, and Junior, which are all based on UDPS and show minor differences in the preparation of the sample for sequencing (the protocol in Junior simpler than in FLX or FLX+). Briefly, incorporation of a nucleotide by the DNA polymerase results in pyrophosphate release, which initiates a series of downstream reactions that ultimately produce light by the firefly enzyme luciferase (Margulies et al., 2005). The 454 sequencing methods applied in this study use GS titanium chemistry. In all three cases, the complete sequencing workflow comprises four main steps that start with purified DNA and end with long and highly accurate reads: (1) generation of a single-stranded template DNA library from amplicons, paired ends or shotgun samples, (2) emulsion-based clonal amplification of the library, (3) data generation via sequencing-by-synthesis, and (4) data analysis using different bioinformatic tools. Using GS titanium chemistry, the read length is 400–500 bp for single reads, with 70–100 000 reads per run in the Junior method and 700 000–1 000 000 reads in the FLX and FLX+ methods.

The aim of this study was to compare UDPS with the reference cloning method to evaluate the adequacy of UDPS for studying viral quasispecies. The main aspects to analyze were the quantitative results and the reproducibility (duplicates analyzed in different laboratories and with different UDPS methods). HBV infection was selected as the model for this purpose. Although HBV is a DNA virus, its replication involves retrotranscription of a viral RNA intermediate by the viral polymerase. This enzyme does not carry proofreading activity and mutations can accumulate throughout the HBV genome. This fact, in addition to the high replication rate of HBV, results in a high variability of this DNA virus (around 100 times higher that other DNA viruses) (Mizokami and Orito, 1999). The single constraint for HBV to accumulate variability is the great overlapping of the viral genes. Among them, the preCore/Core region is the least overlapped (Osiowy et al., 2006), and it contains most of the naturally occurring changes in the genome, such as mutations in the main epitopic regions related to the host immune response (Carman et al., 1997) and mutations associated with the lack of hepatitis B "e" antigen (HBeAg) expression (Carman et al., 1989). For this reason, the preCore/Core region was selected for comparative analysis by cloning and UDPS in the present study.

2. Materials and methods

2.1. Patient studied

The patient selected for this study was a woman with chronic HBV infection, born in 1938, and diagnosed in our hospital in 1991 after an acute episode of hepatitis B. The patient tested negative for HIV, HCV, and HDV co-infection. She was monitored in our hospital and her blood samples were stored in a serum bank at −20 °C. The patient presented HBV genotype D and was HBeAg-negative due to selection of the main preCore mutation, G1896A. The evolution of the infection is presented in Fig. 1. In 1993, she was treated with interferon (IFN) for 6 months, with no significant response. In 2001, she was treated with lamivudine (LVD) and after 14 months, treatment resistant variants were selected (mutations rtL180M, rtM204IV). In August 2002, adefovir was added to her treatment and combined with lamivudine for different periods (Fig. 1). In August 2005, several resistant variants against lamivudine (rtL180M and rtM204IV) and adefovir (rtA181AV and rtN236NT) were detected, and in October 2005, the patient started tenofovir treatment. Since then, she has presented undetectable HBV DNA levels, and normal laboratory results for transaminases, bilirubin, and GGT.

For the present study, a serum sample from 1998 that presented an HBV DNA titer of 7.3 $\log_{10}$ IU/mL was selected. This sample coincided with the period after IFN treatment, which has been associated with selection of changes in the Core gene, due to the immunomodulator activity of the drug (Radecke et al., 2000).

2.2. preCore/Core amplification

HBV DNA was extracted from 200 μL of serum with the QIAamp DNA MiniKit (QIAGEN, Hilden, Germany). The amplicon library corresponding to the HBV preCore/Core region was obtained after two PCR runs (nested). The nested PCR was performed in duplicate: one HBV-DNA-pcCore nested PCR product was cloned and the other HBV-DNA-pcCore nested PCR was UDPS-analyzed on the Genome Sequencer (GS, Roche) FLX+, FLX, and Junior systems (Fig. 2). The nested PCR products were directly sequenced prior to cloning and UDPS.

The first PCR primers were as follows: sense (position 1721-43) 5′ GTTTAA$^{A}/_{G}$GACTGGGAGGAG$^{C}/_{T}$TGG 3′ and antisense (position

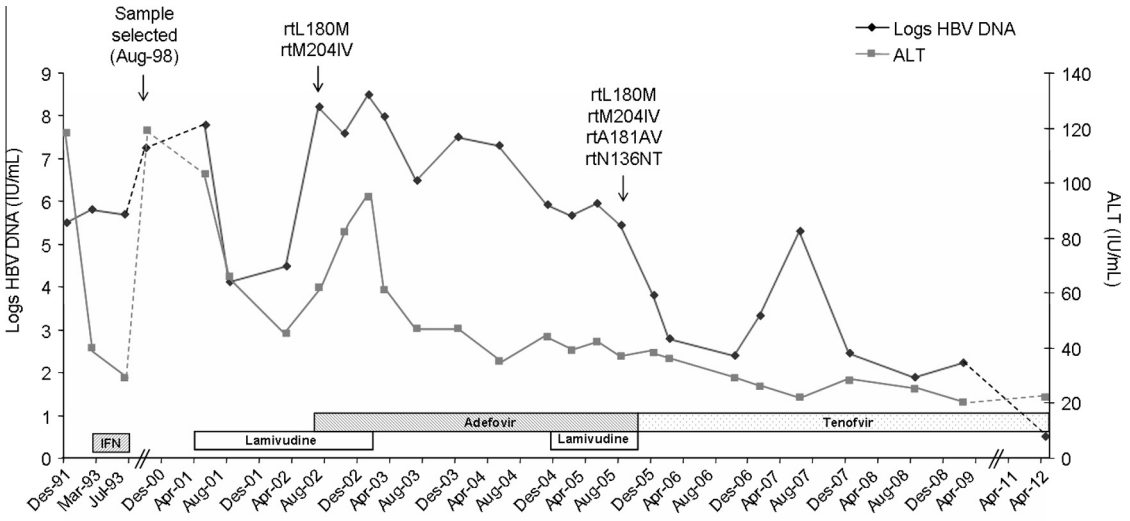

Fig. 1. Evolution of the patient selected for the study.

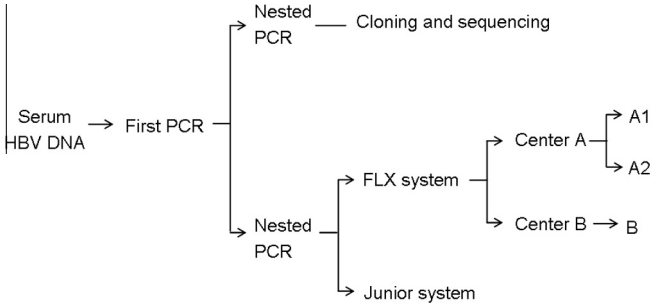

Fig. 2. Treatment of the sample and distribution of the PCR products among the different methodologies to study the viral quasispecies.

2823-04) 5′ TGTTCCCA$^{A}/_{G}$GAATA$^{A}/_{T}$GGTGA 3′. The nested PCR primers included the recognition site for UDPS, shown in italics. The sequence of the sense primer (position 1737-56) was 5′ *CGTATCGCCTCCCTCGCGCCATCAG*GAG$^{C}/_{T}$TGGGGGAGGAGA$^{C}/_{T}$TAG 3′ and the antisense primer (position 2152-30) was 5′*CTAAT GCGCCTTGCCAGCCCGCTCAG*CCAT$^{A}/_{G}$TTAGT$^{A}/_{G}$TT$^{A}/_{G}$ACATAA$^{C}/_{T}$T$^{A}/_{C}/_{G}$ ACTAC 3′. To minimize the error rate of the PCR process, high fidelity polymerase (Pfu Ultra-II, Stratagene, La Jolla, USA) was used. All the nested PCR products had a length of 494 bp and were isolated from 0.9% agarose gel with the QIAquick Extraction Kit (QIAquick Spin Handbook, QIAGEN, Hilden, Germany) and quantified using Quan-iTPicogreens DNA reagent (Invitrogen).

2.3. Cloning and sequencing

One of the HBV-DNA-pcCore nested PCR products was used for cloning by ligation with a vector (pCR™4Blunt-TOPO®) from the Zero Blunt TOPO PCR cloning Kit (Invitrogen), following the manufacturer's instructions. *Escherichia coli* competent cells were chemically transformed with the plasmid using the heat shot method. The transformed bacteria were incubated on a Luria broth (LB) agar plate overnight at 37 °C.

A total of 142 clones were selected for growth in LB medium (24 h at 37 °C, and 200 rpm). The bacterial pellets obtained after centrifugation (13 000 rpm, 5 min, 25 °C) were frozen at −20 °C. Plasmid pCRBlunt4-TOPO was purified with the QIAprep Miniprep kit (QIAGEN, Hilden, Germany), following the manufacturer's instructions.

More than 100 ng of each plasmid was used for sequencing the 142 clones (ABI Prism BigDye Terminator v3.1 Cycle Sequencing Kits, Applied Biosystems). The primers used for sequencing were M13FW primer 5′ GTAAAACGACGGCCAG 3′ and TOPO RV 5′ GAATTGAATTTAGCGGCCGCGAATTC 3′, both of which were provided in the cloning kit. The sequenced DNA was purified with the BigDye X-Terminator Purification Kit (Applied Biosystems) and the product was directly sequenced on the ABI 310 (Applied Biosystems).

2.4. UDPS sequencing

The other HBV-DNA-pcCore nested PCR products were sequenced in our research institute as described (Homs et al., 2011a; Rodriguez-Frias et al., 2012; Homs et al., 2012), once by GS-FLX and once by GS-Junior; and the same HBV-DNA-pcCore nested PCR products were sequenced twice by GS-FLX+ in an external laboratory (CRAG lab, Universitat Autònoma de Barcelona) to obtain an estimation of UDPS variability with different methods and in different labs (Fig. 2). The amplicon libraries were pooled to obtain a concentration of 4×10^6 molecules of the HBV pre-Core/Core region and were UDPS processed, as described previously (Homs et al., 2011a; Rodriguez-Frias et al., 2012; Homs et al., 2012).

2.5. UPDS, filtering of the sequences and study of the polymorphic positions and haplotypes

Data processing was performed on the open source R environment (R Development Core Team, 2012), using the Bioconductor (Gentleman et al., 2004) and Biostrings libraries (Pages et al., 2011) for pattern matching and sequence alignment, and developing R functions.

FLX+, FLX, and Junior generated a fasta file in which the reads were sequenced for each Pico-Titer plate lane. The same data treatment was applied in all three methods. The files used for the analysis were those obtained from the 454 GS system's software, which applies stringent quality controls on each sequenced nucleotide to guarantee the integrity of the full length of the amplicon. Data accuracy was validated according to reported procedures (Huse et al., 2007; Zagordi et al., 2010) and previous unpublished experience from our group. A data processing workflow was developed (Fig. 3).

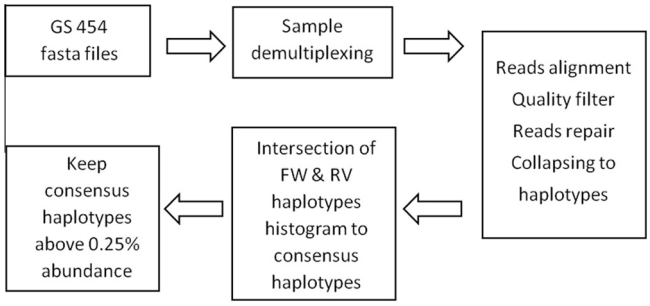

Fig. 3. UDPS data processing workflow to obtain HBV error-free haplotypes, as a sampling of the actual HBV viral quasispecies.

The first step consisted in demultiplexing the sequenced reads to obtain a separate fasta file with the reads corresponding to each sample and amplicon. Demultiplexing was carried out by matching the corresponding primer by alignment with each sequenced read. Primer length and primer degeneracy were the parameters that set the number of mismatches in primer alignment.

The second step assessed the quality of the demultiplexed reads according to previously reported procedures (Huse et al., 2007; Zagordi et al., 2010; Campbell et al., 2008; Homs et al., 2011a; Rodriguez-Frias et al., 2012), attending to the number of Ns and gaps detected from alignment of the reference sequence. This first filter involved appropriate recognition of the primer in each sequence, allowing a maximum of three mismatches and a maximum offset of five positions for the first nucleotide. The primers were trimmed, and all reads that did not cover the full amplicon were excluded. In contrast, the second filter was a repair step that consisted in pair-wise alignment of all the previously accepted reads to the reference sequence. After pair-wise alignment, all nucleotides representing insertions in the reference were removed, leaving a multiple alignment without these insertions. Subsequent collapsing of the reads into haplotypes with corresponding frequencies notably simplified the next steps. All haplotypes with only one read, representing roughly 0.01% in the population, were excluded without further consideration. All haplotypes with more than two Ns, or three gaps, or 99 differences in comparison with the reference sequence were excluded. These values are established by the program's parameters. The accepted Ns and gaps were repaired as per the reference sequence contents in order to preserve the maximum of available information. After the corrections, the haplotypes were recollapsed and their frequencies were updated. The initially identified master sequence was changed if, after repairing and recollapsing, a different haplotype emerged as the most abundant one. Haplotypes were labeled according to the number of mutations in comparison with the master sequence, and with an additional integer identifier as an index of abundance within each number of mutations. At this point, a fasta file with haplotypes and frequencies was obtained for each sample and strand.

The third step corresponded to an abundance filter that removed all haplotypes below an abundance threshold that had been previously established by a clonal sequence processed in parallel by UDPS. The purpose of this step was to remove most of the background noise while preserving the minority variants present in both strands.

The fourth step confronted the haplotypes in the two strands for each amplicon. Only haplotypes that appeared in both strands were preserved. These haplotypes were named consensus haplotypes and their frequencies were taken from the intersection of the haplotype abundances of both strands and further renormalized to one. The purpose of this step was to remove the strand-specific errors, while accepting that some true haplotypes might be lost. This step removed most of the errors, and it was possible to control the percent overlap of the two distributions as an assessment of the quality of the previous steps. Poor overlaps were usually a consequence of repair errors due to differential gaps in one of the strands.

For HBV amplicons of 400–500 nt in length, with a sequencing depth of not less than 10 000 reads per strand, a level of 0.25% abundance assured with very high confidence that no erroneous haplotype would be analyzed. The raw reads, the filtered reads, and the reads >0.25% obtained for each experiment (FLX A1, A2 and B, Junior and cloning) are indicated in Table 1.

2.6. Point mutations

Point mutations were studied in two ways. The first was by direct analysis of the forward (Fw) and reverse (Rv) consensus haplotypes, which is termed *analysis by rows*. The second, called *analysis by columns*, confronts the mutations observed on the FW and RV strands, obtained after applying the frequency threshold filter by homogeneity tests. Mutations unique to Fw or Rv were excluded. The frequencies of mutations common to the two strands were compared using a chi-square test. When significant differences were detected, the minimum frequency was taken, and when non-significant differences were observed, the marginal frequency was taken. The nucleotide frequencies in each segregating site were renormalized when required. This method may be more sensitive than analysis by consensus haplotypes, but it lacks the high confidence of the latter.

2.7. Characterization of HBV quasispecies complexity

The equations and definitions of the different quasispecies complexity parameters are described in Section 6 (Glossary of Terms). All calculations were done on the open source R environment (R Development Core Team, 2012).

Complexity of the viral quasispecies was quantified by three different variables: The *Shannon entropy* (S_n) measures haplotype diversity regardless of the number of mutations implicated, the *mutation frequency* (*Mf*) measures the diversity with respect to the most represented sequence, and the *nucleotide diversity* (*Pi*) takes into account the average number of mutations between each pair of individuals in the viral population. Each of these parameters explains a different part of the mutation space occupied by a quasispecies, and all are relevant when considering mutation barriers to resistance.

Mutation density plots have been used as a graphic tool to visualize the population diversity in comparison with the dominant haplotype in terms of the number of mutations. To visualize population diversity, new plots were designed, which we call *Montserrat plots,* because their shape is similar to that of distinctive a mountain ridge in Catalonia known as "Montserrat". Thus, the mutation frequency (Mf) is a summary of the information contained in this plot: taking the dominant (master) sequence in the population of

Table 1
Number of sequences obtained in each experiment after filtering and after selecting percentages higher than 0.25%.

	UDPS				Clones
	FLX A1	FLX A2	FLX B	Junior	
Raw reads fw	8917	8030	10591	34162	142
Raw reads rv	17377	13662	22497	33930	142
Filtered reads fw	5945	5314	5886	22641	142
Filtered reads rv	12099	9088	15651	19469	142
Reads fw at 0.25%	4715	4387	5193	17058	–
Reads rv at 0.25%	9412	7277	13170	14055	–

each amplicon as reference, the Montserrat plot depicts the density of the number of mutations in the viral population with respect to this reference. The usual histogram is transformed into a density plot using a Gaussian kernel with a bandwidth of 0.3 (Venables and Ripley, 2002), which makes the plot continuous, as informative as the histogram, and visually closer to the dynamics of a quasispecies. As the plot spreads in the number of mutations, it takes a lower profile and points to a more diverse population. The Poisson distribution is shown as an overlay whose parameter is the mean number of mutations with respect to the master in the whole population, and it is smoothed in the same way, giving a density plot. Thus, two different distributions are represented, the empirical distribution and the random distribution. The empirical plot displays the expected distribution of mutants in the population according to Poisson distribution, using the observed average number of mutations in comparison with the dominant haplotype. The empirical distribution is a comparison of the theoretical and experimental plots, which reflect deviations relative to the expected generation of mutant genomes due to fitness effects or random variations in mutant generation. When the two curves coincide, a homogeneous population is indicated. In contrast, notable divergence of the curve indicates a diverse population, with two or more subpopulations separated by a few mutations (bi- or poly-modal plots). Therefore, spreading of the plots indicates a larger number of mutations and a more diverse population.

2.8. Statistics and computation environment

The list of haplotypes with corresponding frequencies resulting from UDPS data treatment was used to estimate the viral population distribution of each sample and replicate. When only two replicates were available, the percentage of deviation between replicates was computed as the difference between the abundance of replicates for each haplotype and its corresponding mean abundance, as an approximation of the coefficient of variation (CV). Likewise, the relative errors of the quasispecies complexity measurements were computed as the difference between replicates and the mean value.

All computations and graphics were done under the environment and language R, an open-source environment for statistical computations (R Development Core Team, 2012).

3. Results

3.1. Establishment of the cut-off value

To establish a threshold for analyzing sequences from UDPS and cloning, the main criterion to determine the cut-off value was the minimal frequency at which no haplotypes different from the expected one were observed in the clonal sequences. This value was 0.25%, and the reads analyzed at percentages above 0.25% are indicated in Table 1. All the polymorphic sites present at percentages higher than 0.25% were present in the three FLX samples (A1, A2 and B), in the sample processed by GS Junior and in the 142 clones (Table 2).

3.2. Polymorphic sites and mutations

After the initial analysis of fasta files obtained from the three 454 methods, we observed that most PCR and sequencing errors were massive deletions on homopolymeric tracts, mainly larger than five nucleotides, but in some cases only 4 nucleotides in size. In addition, the deletions and substitutions detected were strongly dependent on the current nucleotide neighborhood, and were possibly associated with their secondary structure and their location in one or the other (forward/reverse) strand. Based on these observations, two levels of errors were considered. The first was a background level of noise that affected both strands in the same way. The second (and higher level) was the neighborhood, which is structure-dependent and therefore, strand-specific.

Taking into account these considerations, and after application of the described workflow, 14127 reads were analyzed from A1-FLX; 11664 from A2-FLX; 18363 from B-FLX and 31113 from Junior sequencing. The 142 sequences from the different clones were also analyzed (Table 1).

Setting the threshold to 0.25%, 20 polymorphic sites were detected in the three FLX samples (A1, A2 and B), 18 polymorphic sites in the Junior experiments, and 50 in the 142 clones. These polymorphic sites presented one mutation in each position, except in the clone results, in which one polymorphic site presented two different mutations. Thus, a total of 20 mutations were detected in the FLX experiments, 18 mutations in Junior, and 51 mutations in cloning.

Some polymorphic sites detected with cloning were only present in 1, 2 or 3 clones, representing frequencies of 0.7%, 1.4% and 2.1% (confirmed by repeat sequencing). However, despite these high frequencies, most of these polymorphic positions were not observed in any of the duplicate experiments on UDPS analysis.

At the quantitative level, the distribution of the common polymorphic sites and the percentages of variability are presented in Table 2, which shows high similarity between the different UDPS results and cloning.

The nucleotide positions and amino acid substitutions are indicated in Table 2. Of note, high variability was observed in nucleotides 1762 and 1764, which, in the major haplotype, correspond to the mutated version of the main basic Core promoter variants G1762A and A1764T (associated with HBeAg negative status), but with a high percentage of the wild-type version of these variants. The high variability of multiple codons in the Core region was also relevant, and was mainly located in the main epitopic regions of this protein, such as codons 55 and 64, mapped in the Th 50–69 epitope, and codons 74, 77, 80, and 81, located in the B 74–84 epitope (Homs et al., 2011b). The sequence obtained by direct sequencing of the PCR-UDPS product was the same as the dominant haplotype obtained from UDPS and cloning. However, some nucleotide mixtures were detected as minor variants in positions 1976, 2018, 2063, 2092, 2138, and 2139.

To compare the results of variability in the polymorphic sites, the coefficient of variation (CV) of variability was calculated between the 20 common polymorphic positions from the different experiments (FLX A1, A2 and B, Junior, and clones). Two kinds of CV values were established, attending to the type of PCR product. It must be kept in mind that the two different nested PCRs were performed from a single first PCR product (Fig. 2) and this yielded two different amplicon libraries: one for cloning and the other for the UDPS experiments. This different origin might also have induced variability, so different CV values were determined: two intra-library CVs and three inter-library CVs (Table 3). From one site performing the analyses, the intra-library CV was established between the results from the FLX experiments (A1, A2 and B) (FLX column, Table 3) and between FLX and Junior (FLX Junior column, Table 3). And from the other site, the inter-library CV was established between the results from the different nested PCRs; that is, the results obtained between the three FLX and clones, between Junior and cloning, and between the overall UDPS results (mean variability of the three FLX and Junior experiments) and clones.

The intra-library CV values obtained in the comparison of the three FLX experiments were very similar to those obtained when FLX was compared with Junior: median CV between the FLX experiments was 7.95% (range, 2.83–24.3%) and median CV in the comparison of FLX with Junior was 15.20% (range, 4.39–25.53%). As to

Table 2
Common polymorphic sites and their percentage of variability, detected by the three UDPS methods, [FLX+ (A1, A2), FLX (B) and Junior], and by conventional cloning. The nucleotide corresponding to the major haplotype (Ntm), the nucleotide change in relation to the major haplotype (Ntch), the aminoacid corresponding to the preCore/Core ORF in the major haplotype (AAm), and the aminoacid change corresponding to the major haplotype (AAch) are shown.

NT Position	Ntm	Ntch	AA position	AAm	AAch	UDPS				Mean UDPS	Clones
						FLX+ A1	FLX+ A2	FLX B	Junior*		
1757^	A	G	x128	Arg	Arg	4.39	5.16	4.9	2.27	4.29	4.23
1762^	T	A	x130	Met	Lys	20.11	19.85	20.95	13.14	18.51	9.86
1764^	A	G	x131	Ile	Val	20.11	19.85	20.95	13.14	18.51	9.86
1806^	C	T	x144	Ala	Val	1.08	1	1.06	1.37	1.13	0.7
1810#	C	T	x145	Pro	Pro	0.52	0.44	0.56	−0.52	0.51	0.7
1899$	A	G	pc29	Asp	Gly	0.86	0.78	0.96	1.21	0.95	1.41
1961&	T	G	c21	Ser	Ala	0.86	0.78	0.96	1.21	0.95	1.41
1976&	A	T	c26	Thr	Ser	44.86	45.8	41.03	43.18	43.72	41.55
1978&	A	C	c26	Thr	Thr	8.28	8.53	9.87	5.79	8.12	9.86
1996&	T	C	c32	Asp	Asp	2.03	2.45	2.49	1.63	2.15	2.11
2018&	C	G	c40	Gln	Glu	34.3	33.02	38.09	22.57	32	30.99
2063&	C	A	c55	Leu	Ile	44.78	45.74	41.36	44.94	44.21	42.25
2092&	A	C	c64	Glu	Asp	36.56	35.21	40.75	27.89	35.1	33.1
2093&	C	T	c65	Leu	Leu	1.29	1.58	0.96	1.21	1.26	1.41
2119&	T	C	c73	Gly	Gly	1.64	1.63	1.96	−1.83	1.74	0.7
2121&	T	G	c74	Val	Gly	18.25	19.41	17.19	18.36	18.3	8.45
2131&	A	C	c77	Glu	Asp	0.86	1.12	0.96	1.21	1.04	1.41
2138&	A	G	c80	Thr	Ala	44.52	44.55	40.88	48.28	44.56	42.96
2139&	C	G	c80	Thr	Arg	44.52	44.55	40.88	48.28	44.56	42.96
2143&	T	C	c81	Ser	Ser	3.04	2.99	2.5	2.63	2.79	2.11

Lower case letters before the amino acid position indicate the HBV protein affected: x (HBx), pc (preCore), c (Core).
The dark background highlights cases of amino acid changes.
Symbols in the position indicate the affected region:
^ Basic Core promotor (BCP).
# Kozac region.
$ preCore region.
& Core region.
* After applying the abundance filter of 0.25%, Junior sequencing yielded 18 polymorphic positions. Positions 1810 and 2119 were only observed after the abundance filter of 0.1%.

Table 3
Coefficient of variation (CV) of the 20 common polymorphic positions detected in FLX, Junior, and cloning experiments. Intra-library CV refers to amplicons from the same library and inter-library CV to amplicons from different nested PCRs.

	Intra-library CV		Inter-library CV		
Position	FLX	FLX junior	FLX clones	Junior clones	Mean UDPS clones
1757	8.13	25.53	9.3	39.25	22.19
1762	2.83	19.51	29.6	17.65	29.65
1764	2.83	19.51	29.6	17.65	29.65
1806	3.98	14.65	18.4	34.58	22.92
1810	12.06	9.87	19.6	20.87	17.43
1899	10.41	19.61	28.1	11.69	24.98
1961	10.41	19.61	28.1	11.69	24.98
1976	5.76	4.79	5.5	2.67	4.75
1978	9.61	20.96	9.3	49.71	19.69
1996	10.97	18.81	10.3	20.82	16.37
2018	7.50	20.76	8.8	26.38	18.15
2063	5.24	4.39	4.7	4.23	4.33
2092	7.70	15.26	8.9	13.21	13.61
2093	24.30	20.28	20.0	11.69	17.92
2119	10.77	9.02	36.7	43.66	31.95
2121	6.07	4.96	31.6	38.17	27.40
2131	13.38	15.15	22.1	11.69	19.34
2138	4.87	6.78	4.0	7.79	6.13
2139	4.87	6.78	4.0	7.79	6.13
2143	10.49	9.53	16.6	13.98	14.37

FLX: includes the three experiments, A1, A2 and B.
Mean UDPS: mean of variability of the three FLX results and the junior results.

the CV values for inter-library analysis, a slight increase in CVs in comparison with the intra-library values was observed: median CV 17.47% (range, 4.01–36.68%) between FLX and cloning, median CV 15.82% (range, 2.67–49.71%) between Junior and cloning, and median CV 18.75% (range, 4.33–31.95%) between all UDPS experiments and cloning. The highest variability was observed for position 1757, the first nucleotide in the fragments sequenced. Overall, a tendency to an increase in CV was observed at the extremes of the region sequenced. Of note, positions described to include variability (e.g., 1762 and 1764, which are associated with

changes in the basic Core promoter of HBV) seemed to be underestimated by Junior (13.14%) relative to the FLX experiments (all around 20%) (Table 2).

3.3. HBV quasispecies complexity

To evaluate the variability of the quasispecies in each sample, the Shannon entropy, mutation frequency, and nucleotide diversity were calculated (Table 4). The values for these parameters were very similar between the different methods, a fact suggesting that there were no differences between the methods for determining viral variability. Montserrat plots were created to visualize the density of mutations in the viral population. As is seen in Fig. 4, the main population in all the samples analyzed consisted of genomes without mutations; however, all of them also showed important percentages of populations with 6 and 7 mutations. Lastly, a minor population of nine mutations was observed in all the analyses performed. These graphic representations were quite similar for each library, and showed that the quasispecies distribution, obtained by the different UDPS experiments and by cloning were comparable and clearly different from a random distribution.

3.4. Haplotype distribution

A total of 47 haplotypes were detected by cloning, 22 haplotypes by Junior and 37, 40, and 39 by FLX A1, A2, and B, respectively. In the FLX studies, 37 haplotypes were common to both A1 and A2, and 36 haplotypes were common to B and A1, and to B and A2 (Table 5) (higher than 98% identity). All 22 haplotypes detected by Junior were also observed in the FLX replicates (A1, A2 and B), yielding an overall similarity of 90%. This finding may reflect differences in the number of polymorphic positions and their variability between the FLX and Junior methods. In contrast to these comparable results, when haplotype distribution was compared with the conventional cloning results, a low number of common haplotypes was observed, 11 with FLX and 10 with Junior, which yields and overall similarity of only 71%.

A detailed description of the haplotypes in common, by sequences from left (forward sequences) and right (reverse sequences) is also presented in Table 5. Most of the differences between FLX and Junior were due to left sequences. The same was observed in the sequences found in common by cloning and UDPS.

3.5. Specific haplotypes detected in the study by clones

Cloning and Sanger sequencing yielded 142 sequences that contained 50 polymorphic sites and 51 mutations, distributed in a population of 47 different haplotypes. In addition to these 142 clones, eight clones showing deletions that induce a change at the HBV core open reading frame were detected (Fig. 5). Specifically, a deletion of two thymines at positions 1825–1826 was detected in three clones. This deletion induced an in-frame *de novo* stop codon in the precore ORF that codes for HBeAg and results in abolishment of this viral antigen. However, as this patient was HBeAg negative, the presence of these three clones with the deletion may not have had direct implications on HBV replication.

In five other clones, a guanine deletion was observed at position 2091. This mutation directly affects the synthesis of Core antigen, and also induces three different in-frame *de novo* stop codons (Fig. 5). These *de novo* stop codons resulted in truncated forms of HBeAg and, what is particularly relevant, truncated Core antigens, yielding capsid-defective genomes. However, two of these five clones showed insertion of a cytosine at position 2093, which might act as a compensatory mutation.

Table 4
Shannon entropy (S_n), mutation frequency (*Mf*) and nucleotide diversity (*Pi*) from the results of FLX, junior, and cloning.

	FLX A1	FLX A2	FLX B	Junior	Clones
S_n	0.7145	0.6999	0.7295	0.7002	0.6935
Mf	0.00841	0.0084	0.00833	0.00754	0.0079
Pi	0.01088	0.01094	0.01101	0.00975	0.0112

4. Discussion

Since the description of next-generation sequencing, and especially after the development of pyrosequencing (Margulies et al., 2005), multiple applications have been described for this method. This massive parallel sequencing can detect low-frequency nucleotide substitutions in highly complex mixed genomic populations, such as viral quasispecies, thereby providing a snapshot of the entire viral population. UDPS has been particularly productive in the study of highly prevalent viral infections such as HIV, HBV, and HCV (Hoffmann et al., 2007; Swenson et al., 2011; Bull et al., 2011; Solmone et al., 2009; Margeridon-Thermet et al., 2009; Homs et al., 2011a; Rodriguez-Frias et al., 2012; Homs et al., 2012). Until recently, study of these viruses has been based on targeting specific mutations using techniques such as reverse hybridization, which give qualitative results. Quantitative study has been based on cloning and sequencing, but use of this method is limited by the number of clones that can be sequenced: around 100 clones in the best situation, which poorly represents an overall population containing millions or even billions of particles. Due to the limitations of cloning, UDPS has been proposed as an alternative for quantitative study of viral populations. Recent studies applying UDPS have analyzed the viral quasispecies qualitatively, mainly because of incompletely defined sources of errors in the method related to PCR and sequencing. Currently, computational and statistical advances, such as error corrections, haplotype reconstruction, and haplotype frequency estimation, have rationally minimized some of the confounding factors associated with UDPS.

The present study was designed to test the accuracy of UDPS technology for quantitative description of HBV viral quasispecies. Three UDPS methods, the high throughput FLX/FLX+ techniques and the low throughput (and more amenable to routine application) Junior technique, were compared, and each was compared with classic clonal study. In addition, between-laboratory variability of the UDPS method was examined. To our knowledge, this is the first study analyzing the reproducibility of quantitative estimation of UDPS results. The filtering methods applied in the present study gave a cut-off value of 0.25%, which is higher than the 0.05% previously reported by our group (Homs et al., 2012; Homs et al., 2011a; Rodriguez-Frias et al., 2012). This lower sensitivity was expected because our previous report used UDPS technology that sequenced fragments 200 bp in size, whereas with the present method, 494 bp fragments were analyzed. Longer fragments may be more prone to accumulate PCR and sequencing errors than shorter amplicons, and they may include structural differences between forward and reverse sequences. It is interesting that most of the differences between UDPS results obtained by FLX and Junior were observed in reverse sequences, and that the same was seen in the common sequences between clones and UDPS reads; this likely reflects differences in the secondary structures in the two strands. Therefore, it seems that possible structural differences between complementary forward and reverse strands should be taken into account in further refinements of the UDPS sequencing protocol.

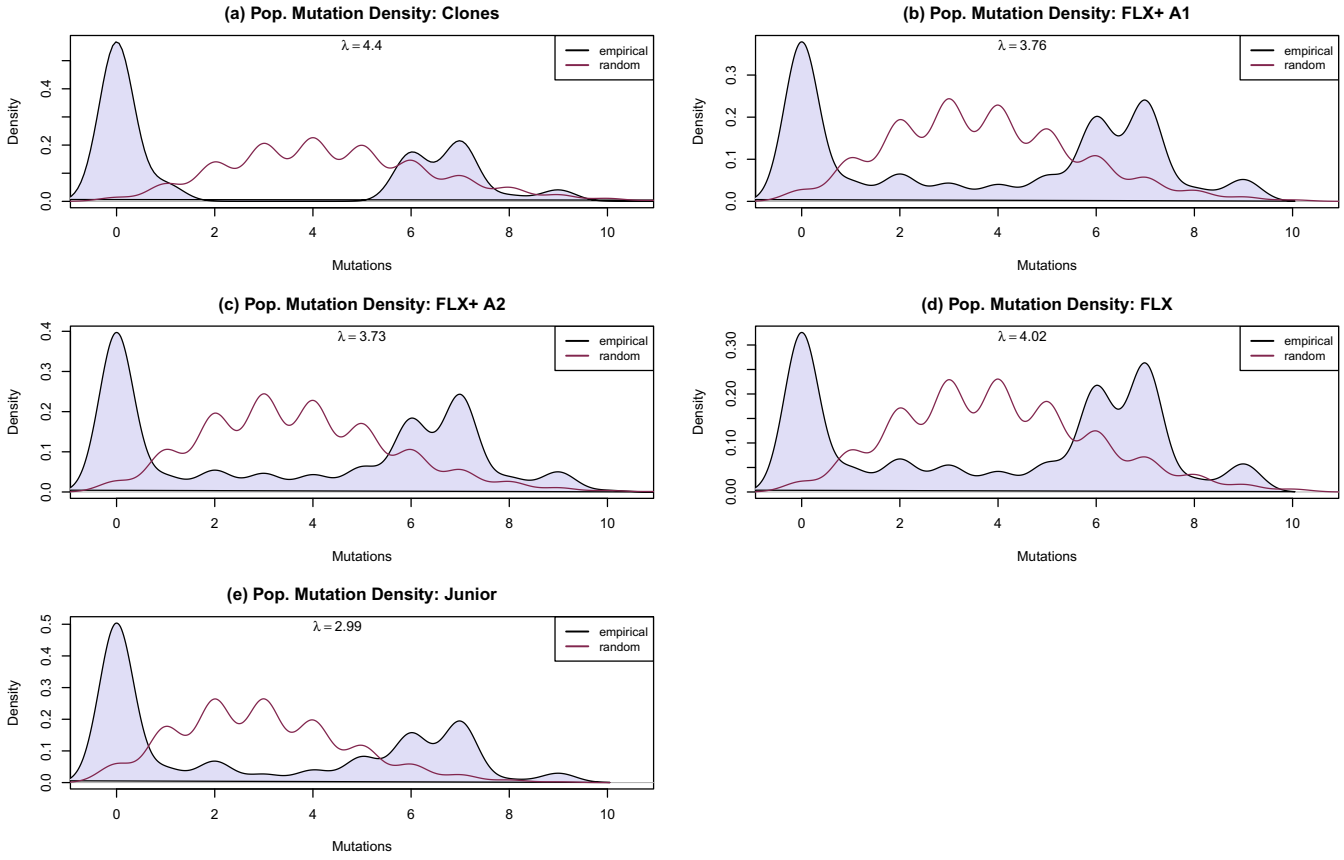

Fig. 4. Montserrat plots of the viral quasispecies analyzed by cloning (a), FLX+ A1, (b) and A2, (c), FLX (d), Junior (e).

Table 5
Number of common haplotypes described in the three FLX experiments, junior, and cloning.

	Common haplotypes	% Reads in common (left)	% Reads in common (right)
Intra-library			
FLX A1 vs. FLX A2	37	98.4	99.4
FLX A1 vs. FLX B	36	99.6	98.2
FLX A2 vs. FLX B	36	99.5	99.0
FLX A1 vs. Junior	22	90.1	99.9
FLX 2 vs. Junior	22	90.0	99.9
FLX B vs. Junior	22	91.0	99.9
Inter-library			
Clones vs. FLX A1	11	72.4	75.3
Clones vs. FLX A2	11	72.4	74.2
Clones vs. FLX B	11	72.4	75.3
Clones vs. Junior	10	71.7	79.7

Once the limitation in sensitivity was set at 0.25%, we found that the different UDPS systems applied in this study yielded comparable results when describing the viral quasispecies at an overall qualitative level (by Montserrat plots) and also at a quantitative level (by S_n, *Mf*, and *Pi*). Furthermore, the reproducibility and accuracy of the UDPS method was mainly reflected by the median intra-library CV values (most of them under 10%), which are similar to those reported for other biological constituents commonly analyzed in clinical laboratories (Westgard et al., 1994; Westgard et al., 1996). Similar conclusions were deduced from the comparison between UDPS and cloning, but, as was expected, the median inter-library CV values were higher (nearly 20%) than the CVs between UDPS experiments. Despite the remarkable reproducibility, the CV ranges in certain polymorphic positions suggested minor sources of variability that that might indicate differing performance between the FLX and Junior sequencing platforms, particularly at the ends of the amplicons. These differences warrant future analysis.

Globally, the relevant similarity in the frequencies of polymorphic sites observed between the duplicates and the intra- and inter-library CVs clearly reinforce the idea that FLX and Junior UDPS might be as useful as cloning. However, the differing number and distribution of haplotypes might indicate a possible source of variability originating in the processes used to create the nested PCR library. This distribution was highly similar between FLX and Junior duplicates (higher than 90%), but the number and distribution of haplotypes between classic cloning and UDPS experiments showed a similarity of only 71%. Furthermore, the number of clones analyzed could also cause differences in the inter-library CVs. Despite the high number of clones processed in the present study relative to previous reports (Cheng et al., 2012), the number is far from truly representing the viral population and the selection of clones is random, fact that can also skew the final results.

Moreover, in contrast to the high restrictive filtering used in UDPS, no rational filtering was applied to cloning. Haplotypes detected by cloning and not by UDPS were observed in only one or two sequences, and these are at a high risk of being artifacts produced during construction of the amplicon library. Extension to hundreds (or thousands) of clones and repetition of the cloning experiments would yield more conclusive results, but such experiments would involve a huge cost in time and funding.

Another factor that must be kept in mind is the possibility of recombination events during PCR amplification (Gorzer et al., 2010). These can occur by the presence of short, incomplete amplicons that act as primers for different haplotypes, or by crossing of amplified sequences, similar to chromosome crossing. These phe-

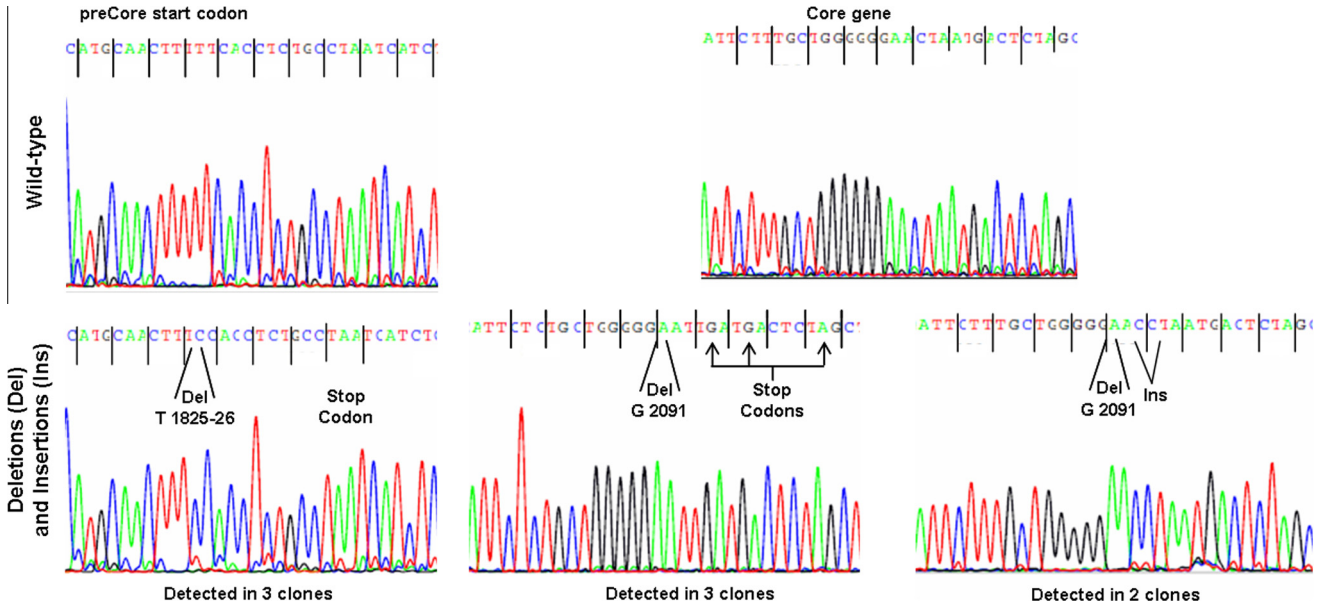

Fig. 5. Sequences obtained by Sanger sequencing of one wild-type clone and clones carrying mutations not detected by UDPS (insertions and deletions).

nomena may be a limitation of classic cloning and UDPS, but they can also occur in other techniques including PCR amplification, which would question their reliable use to analyze haplotypes in the quasispecies. Similarly, the use of high fidelity polymerase from different manufacturers might have an impact on the results. Therefore, the choice of a polymerase with a low probability of recombination and high proofreading activity should be contemplated, or should be standardized, in order to compare the results from different studies. The true impact of these recombination events should be carefully evaluated in the future by UDPS analysis of in vitro mixtures from different clonal sequences. It would be interesting to know the impact of recombination in order to derive conclusions from linkage analyses of different relevant nucleotide substitutions, as have been reported in the polymerase and preCore HBV regions by classic cloning (Cheng et al., 2012) and by UDPS (Homs et al., 2011a; Rodriguez-Frias et al., 2012; Homs et al., 2012). Despite this possible, but still unproven event, the frequencies of polymorphic sites ("column analysis") observed in all the experiments were mainly very similar, even between UDPS and cloning, in which different nested PCRs were used, in contrast to what occurred with haplotype analysis ("raw analysis"). The similarities in the polymorphic sites seem to validate UDPS as a quantitative tool for viral quasispecies analysis and show that it is a highly sensitive method (cut-off of 0.25%). Therefore, UDPS represents an extremely practical alternative to conventional cloning, up to recently considered the reference standard for quasispecies analysis.

Although UDPS seemed useful for quantitative study of viral quasispecies, the classic cloning approach detected haplotypes with deletions in a significant percentage that were not detected by UDPS. For example, five clones presented a guanine deletion in position 2091 (three of five had a compensatory mutation) and three clones showed deletion of two thymines at positions 1825 and 1826. These mutations were located in homopolymeric regions that are highly prone to include artifacts, such as deletions or insertions, in UDPS processing (Wang et al., 2007). The filter applied to the UDPS data corrects some reads with insertions and deletions, and therefore, these mutations cannot be studied by the present UDPS workflow. These results indicate that classic clonal analysis remains useful to study these variants, and that computational analysis of UDPS reads must be improved to detect them. For example, the filter should accept sequences with known deletions and insertions (such as those described in this study).

The insertions and deletions detected by cloning might yield potentially defective particles, especially the uncompensated G2091 deletion, which results in a "capsid-defective genome". The presence of such "defective genomes" in the HBV quasispecies has been previously described by UDPS as a result of changes in non-homopolymeric regions, mainly due to change of a TGG codon into TAG or TGA stop codons (Homs et al., 2011a; Rodriguez-Frias et al., 2012). This "defective genome" provides evidence of possible trans-complementation mechanisms in HBV infection to tolerate defective particles, and could even point to a kind of natural tendency to maintain these defective particles due to a possible beneficial effect for viral evolution, according to games theory strategies (Neumann, 2005).

5. Conclusions

Our results show that UDPS is a useful, highly accurate alternative to molecular cloning for quantitative study of viral quasispecies, which can be extended to other types of complex genomic mixtures (e.g., cancer and mitochondrial studies). The UDPS experiments yielded reproducibility levels similar to those of other complex quantitative determinations. Nonetheless, additional studies must be performed to establish and minimize differences in the structural constraints imposed by forward and reverse strands and the impact of certain confounding factors, such as recombination during PCR steps. Moreover, the complex manual steps in UDPS protocols might increase variability in the results, and for this reason, it is essential to automate the method. Lastly, our study highlights the importance of establishing a reference method to study the viral quasispecies, such as selection of specific polymerases that are optimal for standardization of the overall UDPS process, and highlights the urgency of a clear definition of reliable values for determining the variability of quasispecies composition.

6. Glossary

(Nei, 1987; Domingo et al., 2004).

Amplicon: RNA or DNA fragment amplified with specific primers by PCR.

Read: synthesis of an amplicon during the UDPS process (a read does not always cover the full amplicon length).

Mutation: substitution of one nucleotide for another nucleotide. The number of mutations refers to the total number of different mutations (regardless of their frequencies). Insertions or deletions are excluded from the analysis of UDPS reads.

Polymorphic position: a single position in the RNA or DNA sequence that can include one or several mutations. In cases of single mutations per site, the number of polymorphic sites coincides with the number of mutations.

Haplotype: each of the different sequences identified from amplicon analysis by UDPS. The haplotype is represented by a number of reads that roughly corresponds to its frequency in the quasispecies population, and is obtained by sampling. The haplotypes are, therefore, the set of unique sequences representing the whole viral population.

Haplotype diversity (*Hd*): Also known as gene diversity, it is the probability that two random sequences (reads) are different.

$$H_d = \frac{n}{n-1}\left(1 - \sum_{i=1}^{h} p_i^2\right) \quad \text{(I)}$$

where n is the number of reads, h is the number of haplotypes, and p_i is the relative frequency of the i-th haplotype.

Mutation frequency(*Mf*): Obtained as the ratio of the number of observed mutations relative to the master sequence divided by the total number of nucleotides sequenced. The number of nucleotides sequenced here is the product of the read length multiplied by the number of reads. Computation of the observed mutations takes into account the haplotype frequencies within the viral population, and the number of mutations in each haplotype.

Mean number of pairwise differences(*K*): The differences between each pair of haplotypes are weighted by the respective haplotype frequencies; they are then added up and lastly, divided by the number of different pairs to get the average. K is indicative of the observed population diversity:

$$K = \frac{2}{n(n-1)} \sum_{i<j} w_i w_j d_{i,j} \quad \text{(II)}$$

with:

$$\sum_{i}^{h} w_i = n$$

w_i being the population frequency of the i-th haplotype, h the number of haplotypes, and n the number of reads.

Nucleotide diversity(*pi*, π): The nucleotide diversity takes this into account normalizing K to the reads length m:

$$\pi = K/m \quad \text{(III)}$$

The longer reads, the higher the number of mean differences obtained.

Quasispecies Shannon entropy (*H*): Given the population frequencies of each haplotype, Shannon entropy measures the variability within the quasispecies:

$$H = -\sum_i p_i \ln(p_i) \quad \text{(IV)}$$

where p_i is the relative frequency of the i-th haplotype within the viral population, and the sum extends over all haplotypes. A high value of H implies high variability. The minimum entropy value, corresponding to a maximum order or information, is 0, when we have only one haplotype, as $\ln(1) = 0$.

Normalized quasispecies Shannon entropy (S_n): This normalization attends to the number of haplotypes and the relative weight of each of them, with the highest diversity being all haplotypes with equal frequencies, giving a Shannon entropy of $\ln(h)$:

$$S_n = -\sum_i p_i \ln(p_i) = -h(1/h)\ln(1/h) = ln(h) \quad \text{(V)}$$

The normalized Shannon entropy takes this into account, by dividing H between *ln(h)*, the maximum attainable entropy:

$$S_n = -\sum_i p_i \ln(p_i)/\ln(h) \quad \text{(VI)}$$

S_n ranges from 0 to 1, and a high value is nearer 1 (maximum entropy and maximum diversity).

Information contents at a site(ic_i): Given the relative frequency of each nucleotide in the population, p_i, at a specified site, the information contents, in bit units, is obtained by:

$$ic = \log_2(4) + \sum_i p_i \log_2(pi) = 2 + \sum_i p_i \log_2(pi) \quad \text{(VII)}$$

Entropy is a measure of diversity, but it lacks information. The ic_i is a reversed measure to give the information contents at a site within a sequence in a given population. Both variables are inversely related: S_n is measured by rows (sequences) and the ic_i is an analysis of entropy by columns (sites).

A 100% conserved site has an ic_i, of 2 bits whereas a site with 25% of each nucleotide shows an ic_i of 0 bits. The values for all the positions along the amplicon are not usually given explicitly, but instead are represented on a graph.

Acknowledgments

This work was funded by Instituto de Salud Carlos III, grants FIS PI09/0899 and FIS PI11/1973 cofinanced by the European Regional Development Fund (ERDF).

References

Beerenwinkel, N., Zagordi, O., 2011. Ultra-deep sequencing for the analysis of viral populations. Curr. Opin. Virol. 1, 413–418. http://dx.doi.org/10.1016/j.coviro.2011.07.008.

Bull, R.A., Luciani, F., McElroy, K., Gaudieri, S., Pham, S.T., Chopra, A., Cameron, B., Maher, L., Dore, G.J., White, P.A., Lloyd, A.R., 2011. Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. PLoS Pathog. 7, e1002243. http://dx.doi.org/10.1371/journal.ppat.1002243.

Campbell, P.J., Pleasance, E.D., Stephens, P.J., Dicks, E., Rance, R., Goodhead, I., Follows, G.A., Green, A.R., Futreal, P.A., Stratton, M.R., 2008. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc. Natl. Acad. Sci. U. S. A. 105, 13081–13086. http://dx.doi.org/10.1073/pnas.0801523105.

Carman, W.F., Jacyna, M.R., Hadziyannis, S., Karayiannis, P., McGarvey, M.J., Makris, A., Thomas, H.C., 1989. Mutation preventing formation of hepatitis B e antigen in patients with chronic hepatitis B infection. Lancet 2, 588–591.

Carman, W.F., Boner, W., Fattovich, G., Colman, K., Dornan, E.S., Thursz, M., Hadziyannis, S., 1997. Hepatitis B virus core protein mutations are concentrated in B cell epitopes in progressive disease and in T helper cell epitopes during clinical remission. J. Infect. Dis. 175, 1093–1100.

Cheng, Y., Guindon, S., Rodrigo, A., Lim, S.G., 2012. Increased viral quasispecies evolution in Hbeag seroconverter patients treated with oral nucleoside therapy. J. Hepatol.. http://dx.doi.org/10.1016/j.jhep.2012.09.017; 10.1016/j.jhep.2012.09.017.

Cubero, M., Esteban, J.I., Otero, T., Sauleda, S., Bes, M., Esteban, R., Guardia, J., Quer, J., 2008. Naturally occurring NS3-protease-inhibitor resistant mutant A156T in the liver of an untreated chronic hepatitis C patient. Virology 370, 237–245. http://dx.doi.org/10.1016/j.virol.2007.10.006.

Domingo, E., Gomez, J., 2007. Quasispecies and its impact on viral hepatitis. Virus Res. 127, 131–150. http://dx.doi.org/10.1016/j.virusres.2007.02.001.

Domingo, E., Holland, J.J., 1997. RNA virus mutations and fitness for survival. Annu. Rev. Microbiol. 51, 151–178. http://dx.doi.org/10.1146/annurev.micro.51.1.151.

Domingo, E., Escarmís, C., Lazaro, E., Manrubia, S.C., 2004. Quasispecies dynamics and RNA virus extinction. Virus Res. 107, 129–139. http://dx.doi.org/10.1016/j.virusres.2004.11.003.

Eigen, M., 1971. Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58, 465–523.

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B.H., Bhattacharya, T., Korber, B., 2002. Diversity considerations in HIV-1

vaccine selection. Science 296, 2354–2360. http://dx.doi.org/10.1126/science.1070441.
Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J., 2004. Bioconductor: Open software development for computational biology and bioinformatics R. Genome Biol. 5, R80. http://dx.doi.org/10.1186/gb-2004-5-10-r80.
Ghedin, E., Fitch, A., Boyne, A., Griesemer, S., DePasse, J., Bera, J., Zhang, X., Halpin, R.A., Smit, M., Jennings, L., St George, K., Holmes, E.C., Spiro, D.J., 2009. Mixed infection and the genesis of influenza virus diversity. J. Virol. 83, 8832–8841. http://dx.doi.org/10.1128/JVI.00773-09.
Gorzer, I., Guelly, C., Trajanoski, S., Puchhammer-Stockl, E., 2010. The impact of PCR-generated recombination on diversity estimation of mixed viral populations by deep sequencing. J. Virol. Methods 169, 248–252. http://dx.doi.org/10.1016/j.jviromet.2010.07.040.
Hoffmann, C., Minkah, N., Leipzig, J., Wang, G., Arens, M.Q., Tebas, P., Bushman, F.D., 2007. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. Nucleic Acids Res. 35, e91. http://dx.doi.org/10.1093/nar/gkm435.
Homs, M., Buti, M., Quer, J., Jardi, R., Schaper, M., Tabernero, D., Ortega, I., Sanchez, A., Esteban, R., Rodriguez-Frias, F., 2011a. Ultra-deep pyrosequencing analysis of the hepatitis B virus preCore region and main catalytic motif of the viral polymerase in the same viral genome. Nucleic Acids Res.. http://dx.doi.org/10.1093/nar/gkr451.
Homs, M., Jardi, R., Buti, M., Schaper, M., Tabernero, D., Fernandez-Fernandez, P., Quer, J., Esteban, R., Rodriguez-Frias, R., 2011b. HBV core region variability: effect of antiviral treatments on main epitopic regions. Antivir. Ther. 16 (1), 37–49. http://dx.doi.org/10.3851/IMP1701.
Homs, M., Buti, M., Tabernero, D., Quer, J., Sanchez, A., Corral, N., Esteban, R., Rodriguez-Frias, F., 2012. Quasispecies dynamics in main core epitopes of hepatitis B virus by ultra-deep-pyrosequencing. World J. Gastroenterol. 18, 6096–6105. http://dx.doi.org/10.3748/wjg.v18.i42.6096; 10.3748/wjg.v18.i42.6096.
Hormozdiari, F., Alkan, C., Eichler, E.E., Sahinalp, S.C., 2009. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res. 19, 1270–1278. http://dx.doi.org/10.1101/gr.088633.108.
Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., Welch, D.M., 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 8, R143. http://dx.doi.org/10.1186/gb-2007-8-7-r143.
Johnson, J.A., Li, J.F., Wei, X., Lipscomb, J., Irlbeck, D., Craig, C., Smith, A., Bennett, D.E., Monsour, M., Sandstrom, P., Lanier, E.R., Heneine, W., 2008. Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naive populations and associate with reduced treatment efficacy. PLoS Med. 5, e158. http://dx.doi.org/10.1371/journal.pmed.0050158.
Margeridon-Thermet, S., Shulman, N.S., Ahmed, A., Shahriar, R., Liu, T., Wang, C., Holmes, S.P., Babrzadeh, F., Gharizadeh, B., Hanczaruk, B., Simen, B.B., Egholm, M., Shafer, R.W., 2009. Ultra-deep pyrosequencing of hepatitis B virus quasispecies from nucleoside and nucleotide reverse-transcriptase inhibitor (NRTI)-treated patients and NRTI-naive patients. J. Infect. Dis. 199, 1275–1285. http://dx.doi.org/10.1086/597808.
Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380. http://dx.doi.org/10.1038/nature03959.
Martell, M., Esteban, J.I., Quer, J., Genesca, J., Weiner, A., Esteban, R., Guardia, J., Gomez, J., 1992. Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. J. Virol. 66, 3225–3229.
Mizokami, M., Orito, E., 1999. Molecular evolution of hepatitis viruses. Intervirology 42, 159–165. http://dx.doi.org/10.1159/000024975.
Nei, M., 1987. Molecular Evolutionary Genetics, first ed. Columbia University Press, New York.
Neumann, A.U., 2005. Hepatitis B viral kinetics: a dynamic puzzle still to be resolved. Hepatology 42, 249–254. http://dx.doi.org/10.1002/hep.20831.
Nowak, M.A., Anderson, R.M., McLean, A.R., Wolfs, T.F., Goudsmit, J., May, R.M., 1991. Antigenic diversity thresholds and the development of AIDS. Science 254, 963–969.
Nowak, M.A., Bonhoeffer, S., Hill, A.M., Boehme, R., Thomas, H.C., McDade, H., 1996. Viral dynamics in hepatitis B virus infection. Proc. Natl. Acad. Sci. USA 93, 4398–4402.
Osiowy, C., Giles, E., Tanaka, Y., Mizokami, M., Minuk, G.Y., 2006. Molecular evolution of hepatitis B virus over 25 years. J. Virol. 80, 10307–10314. http://dx.doi.org/10.1128/JVI.00996-06.
Pages, H., Aboyoun, P., Gentleman, R., DebRoy, S., 2011. Biostrings: String objects representing biological sequences, and matching algorithms. R package version 2.13.2.
Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F., Gao, Y., Church, G.M., Shendure, J., 2007. Multiplex amplification of large sets of human exons. Nat. Methods 4, 931–936. http://dx.doi.org/10.1038/nmeth1110.
R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 3-900051-07-0http://www.R-project.org/.
Radecke, K., Protzer, U., Trippler, M., Meyer Zum Buschenfelde, K.H., Gerken, G., 2000. Selection of hepatitis B virus variants with aminoacid substitutions inside the core antigen during interferon-alpha therapy. J. Med. Virol. 62, 479–486.
Rodriguez-Frias, F., Tabernero, D., Quer, J., Esteban, J.I., Ortega, I., Domingo, E., Cubero, M., Camos, S., Ferrer-Costa, C., Sanchez, A., Jardi, R., Schaper, M., Homs, M., Garcia-Cehic, D., Guardia, J., Esteban, R., Buti, M., 2012. Ultra-deep pyrosequencing detects conserved genomic sites and quantifies linkage of drug-resistant amino acid changes in the hepatitis B virus genome. PLoS One 7, e37874. http://dx.doi.org/10.1371/journal.pone.0037874; 10.1371/journal.pone.0037874.
Shankarappa, R., Margolick, J.B., Gange, S.J., Rodrigo, A.G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C.R., Learn, G.H., He, X., Huang, X.L., Mullins, J.I., 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. 73, 10489–10502.
Solmone, M., Vincenti, D., Prosperi, M.C., Bruselles, A., Ippolito, G., Capobianchi, M.R., 2009. Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen. J. Virol. 83, 1718–1726. http://dx.doi.org/10.1128/JVI.02011-08.
Swenson, L.C., Mo, T., Dong, W.W., Zhong, X., Woods, C.K., Thielen, A., Jensen, M.A., Knapp, D.J., Chapman, D., Portsmouth, S., Lewis, M., James, I., Heera, J., Valdez, H., Harrigan, P.R., 2011. Deep V3 sequencing for HIV type 1 tropism in treatment-naive patients: a reanalysis of the MERIT trial of maraviroc. Clin. Infect. Dis. 53, 732–742. http://dx.doi.org/10.1093/cid/cir493.
Tsibris, A.M., Korber, B., Arnaout, R., Russ, C., Lo, C.C., Leitner, T., Gaschen, B., Theiler, J., Paredes, R., Su, Z., Hughes, M.D., Gulick, R.M., Greaves, W., Coakley, E., Flexner, C., Nusbaum, C., Kuritzkes, D.R., 2009. Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. PLoS One 4, e5683. http://dx.doi.org/10.1371/journal.pone.0005683.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, fourth ed. Springer, New York.
Vignuzzi, M., Stone, J.K., Arnold, J.J., Cameron, C.E., Andino, R., 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439, 344–348. http://dx.doi.org/10.1038/nature04388.
Wang, C., Mitsuya, Y., Gharizadeh, B., Ronaghi, M., Shafer, R.W., 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res. 17, 1195–1201. http://dx.doi.org/10.1101/gr.6468307.
Westgard, J.O., Seehafer, J.J., Barry, P.L., 1994. European specifications for imprecision and inaccuracy compared with operating specifications that assure the quality required by US CLIA proficiency-testing criteria. Clin. Chem. 40, 1228–1232.
Westgard, J.O., Bawa, N., Ross, J.W., Lawson, N.S., 1996. Laboratory precision performance. State of the art versus operating specifications that assure the analytical quality required by clinical laboratory improvement amendments proficiency testing. Arch. Pathol. Lab. Med. 120, 621–625.
Wright, C.F., Morelli, M.J., Thebaud, G., Knowles, N.J., Herzyk, P., Paton, D.J., Haydon, D.T., King, D.P., 2011. Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. J. Virol. 85, 2266–2275. http://dx.doi.org/10.1128/JVI.01396-10.
Zagordi, O., Klein, R., Daumer, M., Beerenwinkel, N., 2010. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res. 38, 7400–7409. http://dx.doi.org/10.1093/nar/gkq655.