BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 8 2014, pages 1104–1111
doi:10.1093/bioinformatics/btt768

Sequence analysis

Advance Access publication January 2, 2014

Inference with viral quasispecies diversity indices: clonal and NGS approaches

Josep Gregori[1,2,3,*], Miquel Salicrú[3], Esteban Domingo[4,5], Alex Sanchez[3,6], Juan I. Esteban[1,4,7], Francisco Rodríguez-Frías[4,7,8] and Josep Quer[1,4,7]

[1]Liver Unit, Internal Medicine Lab Malalties Hepàtiques, Vall d'Hebron Institut Recerca (VHIR-HUVH), 08035 Barcelona, Spain, [2]Roche Diagnostics SL, 08174, Sant Cugat del Vallès, Spain, [3]Statistics Department, Biology Faculty, Barcelona University, 08028, Barcelona, Spain, [4]CIBER de Enfermedades Hepáticas y Digestivas (CIBERehd) del Instituto de Salud Carlos III, 28029 Madrid, Spain, [5]Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Campus de Cantoblanco, 28049, Madrid, Spain, [6]Bioinformatics and Statistics Unit, Vall d'Hebron Institut Recerca (VHIR-HUVH), 08035, Barcelona, Spain, [7]Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain and [8]Biochemistry Unit. Virology Unit/Microbiology Department, HUVH, 08035 Barcelona, Spain

Associate Editor: Michael Brudno

ABSTRACT

Given the inherent dynamics of a viral quasispecies, we are often interested in the comparison of diversity indices of sequential samples of a patient, or in the comparison of diversity indices of virus in groups of patients in a treated versus control design. It is then important to make sure that the diversity measures from each sample may be compared with no bias and within a consistent statistical framework. In the present report, we review some indices often used as measures for viral quasispecies complexity and provide means for statistical inference, applying procedures taken from the ecology field. In particular, we examine the Shannon entropy and the mutation frequency, and we discuss the appropriateness of different normalization methods of the Shannon entropy found in the literature. By taking amplicons ultra-deep pyrosequencing (UDPS) raw data as a surrogate of a real hepatitis C virus viral population, we study through in-silico sampling the statistical properties of these indices under two methods of viral quasispecies sampling, classical cloning followed by Sanger sequencing (CCSS) and next-generation sequencing (NGS) such as UDPS. We propose solutions specific to each of the two sampling methods—CCSS and NGS—to guarantee statistically conforming conclusions as free of bias as possible.

Contact: josep.gregori@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online

Received on September 5, 2013; revised on December 20, 2013; accepted on December 25, 2013

1 INTRODUCTION

RNA viruses show a high replication error rate due to the lack of proofreading mechanisms, and it is estimated that for viruses with typically high replicative loads every possible point mutation and many double mutations are generated with each viral replication cycle, and may be present within the population at any time (Domingo *et al.*, 2012). In the case of hepatitis C virus (HCV), the viral load—defined as the number of viral particles per milliliter of serum in acutely or chronically infected patients—may reach 10^7 in immunocompetent patients, which roughly means a population of circulating particles of 10^{10}–10^{11} at any given time. This population is highly dynamic, with a viral half-life of a few hours, and with the production and clearance of 10^{10}–10^{12} genomes per day in a patient (Herrmann *et al.*, 2000; Neumann *et al.*, 1998). Given this inherent dynamics, we are often interested in the comparison of diversity indices of sequential samples of a patient or among groups of patients. These comparisons may be informative of the patient evolution or the appropriateness of a given treatment (Supplementary Material).

Next-generation sequencing methods (NGS) will likely be increasingly adopted in clinical diagnostics in the next years. Improvements in costs, protocols and coverage are closing the gap between what was feasible in research and diagnostics. The first diagnostics likely to be moved to NGS will be those currently based on classical molecular cloning and Sanger sequencing (CCSS) because it is labor intensive and has limited sensitivity. In this work we use in-silico sampling from viral reference distributions to study the statistical properties of diversity indices aimed at quantifying RNA virus quasispecies complexity.

Estimates of the species richness and other diversity indices as defined in ecology (Supplementary Material) are challenging when populations are complex in genomic composition (Magurran and McGill, 2010) as is the case with viral quasispecies (Domingo *et al.*, 2012; Perales *et al.*, 2010). The approaches in the ecology domain are extensive and still active (Chao and Shen, 2003; Chao *et al.*, 2009, 2010; Colwell *et al.*, 2012; Heip and Engels, 1974; Hellmann and Fowler, 1999; Hutcheson, 1970; Jost, 2006; Magurran and McGill, 2010; Nemenman *et al.*, 2011; Pardo *et al.*, 1997; Salicru *et al.*, 1993; Tuomisto, 2010; Walther and Moore, 2005) and can be useful for the analysis of viral quasispecies. Although the quasispecies definition as a '*dynamic distributions of non-identical but closely related mutant and recombinant viral genomes subjected to a continuous process of genetic variation, competition and selection, and which act as a unit of selection*' (Domingo *et al.*, 2005) conveys an intuitive image of complexity, no comprehensive and universally admitted

*To whom correspondence should be addressed.

index of quasispecies complexity exists. In a large population in equilibrium or with small perturbations, the genome frequencies are related with their relative fitness. There are a number of useful indices and variables but none of them fully captures that intuitive image. Viral quasispecies complexity may be viewed as a multivariate feature, where the number of haplotypes of polymorphic sites and their relative frequencies are its dimensions. Each of these indices and variables are difficult to estimate given the expected diversity of a quasiespecies from available data and the limited sample size amenable to analysis (Domingo *et al.*, 2012).

The primary indices measure the extend of the viral quasispecies complexity by the number of haplotypes, polymorphic sites and number of different mutations; these may be considered as richness indices. Other indices such as the Shannon entropy (S) (Shannon, 1948) or the Simpson index (Magurran, 2004) measure the diversity, or the evenness when normalized to maximum diversity (Sn), while others such as the mutation frequency (Mf) or the nucleotide diversity (Pi) measure the intrapopulation heterogeneity, that is how different are the members of the population among them. S and Sn, or the Simpson index, are not sensitive to the number of mutations. The Simpson index has been less used with viral quasispecies (Nowak *et al.* 1991, Wolinsky *et al.* 1996), as it provides a more stable, although less sensitive, measure of diversity by downweighting the rare haplotypes. Mf measures the heterogeneity with respect to the most represented (dominant) sequence (Ramirez *et al.*, 2013) or the consensus sequence of the population (Cabot *et al.*, 2000). Pi gives the global population heterogeneity, taking into account the average number of mutations between each pair of individuals in the viral population (Nei, 1987). Each of these variables describes a different part of the mutation space occupied by a quasispecies, and they all provide relevant information regarding mutation barriers to antiviral treatment resistance.

We studied by in-silico sampling the distribution and properties of three of the most common variables used to quantify the viral quasispecies complexity in the literature, the diversity through S and Sn and the heterogeneity through Mf. The quasispecies richness by the number of estimated haplotypes in the population is also studied because of its implications on Sn and Mf. We propose methods for inference for each sampling scheme—CCSS and NGS—with these complexity indices.

2 METHODS

2.1 Basic assumptions

To make simulations of CCSS or NGS sampling experiments, we need the distribution of haplotypes of a viral quasispecies. We can empirically approach a 10^{10} genomes distribution by taking the raw data from high coverage amplicon ultra-deep pyrosequencing (UDPS) experiments of samples of a wide complexity spectrum as reference distributions.

Simulations of measures by CCSS will be obtained by in-silico sampling a given number of particles from the reference distribution, where any particle has the same probability to be sampled. Simulations of measures from NGS data will be obtained by in-silico sampling a number of particles from these distributions and setting an abundance filter, corresponding to RT + PCR + NGS noise levels (Archer *et al.*, 2012; Beerenwinkel and Zagordi, 2011; Beerenwinkel *et al.*, 2012; Flaherty *et al.*, 2012; Gilles *et al.*, 2011; Huse *et al.*, 2007; Loman *et al.*, 2012; Macalalad *et al.*, 2012; Mild *et al.*, 2011; Prosperi and Salemi, 2012; Prosperi *et al.*, 2011; Vandenbroucke *et al.*, 2011; Zagordi *et al.*, 2012).

This study is based on the following set of basic assumptions:

(a) A very high coverage (~50 000 times) UDPS amplicon dataset from patient samples of HCV may be considered as a coarse approximation to the high complexity of RNA virus quasiespecies, and the observed distribution of haplotypes may be used as a viral population reference distribution from which to sample viral particles.

(b) A CCSS in-silico experiment consists in sampling a given number of viral particles from a reference distribution. All obtained sequences are accepted as true members of the population. Measures of viral quasispecies complexity are then computed from the observed haplotypes and frequencies.

(c) The NGS methods have a noise level, due to Reverse transcription (RT) and polymerase chain reaction (PCR) sequencing errors, below which we may not distinguish true from erroneous mutations. Any data treatment of amplicon NGS sequences requires some sort of abundance filter to exclude artifactual haplotypes and point mutations.

(d) As a simple approach, a NGS in-silico experiment consists in sampling a given number of molecules from the reference distribution, followed by an abundance filter to exclude all haplotypes with abundance below the noise level. Measures of viral quasispecies complexity are then computed from the filtered haplotypes and frequencies.

2.2 Indices of diversity, definitions and equations

We give in the Supplementary Material all relevant definitions and equations used throughout this work. That is, the definitions related to viral quasispecies and to diversity indices, and the equations with and without bias corrections.

2.3 Distribution of diversity measures

The distribution of a variable measuring viral quasispecies complexity obtained by a CCSS experiment will be estimated by repeating a number of times (2000) an in-silico sampling of a given number of viral particles, and computing such variable each time. In this study, we repeated a number of times experiments with 20 and 50 clones, covering the most common range of sample sizes in the literature. The distribution of NGS measures were obtained by repeating the same number of times (2000) in-silico samplings of 400 and 1000 reads sampled from the reference populations, filtering at a noise level of 0.5% and computing the complexity variables each time. This is a feasible expected mean coverage in clinical settings with ~50 samples in a 454 Junior plate.

2.4 Shannon entropy normalization

In the ecology literature, the Shannon entropy (Supplementary Equation SI) is normalized to the natural logarithm of the number of estimated species in the population (Supplementary Equation SVI) so that a population where all species are equally represented corresponds to a maximum entropy of 1, whereas a population with a single species is a population of minimum entropy, with $Sn = 0$. In the virology literature, we observe other two normalizations. Either to log(N) (Abbate *et al.*, 2005; Cabot *et al.*, 2000; Grande-Perez *et al.*, 2002; Pawlotsky *et al.*, 1998) or to N (Fishman and Branch, 2009; Nasu *et al.*, 2011; Nishijima *et al.*, 2012), where N is the sample size, that is the number of clones in each sample. The normalization to log(N) is justified by saying that maximum entropy is attained when all observed molecules are different. These two normalizations are sample size-dependent, that is, having the same S for two samples of different size from the same population, we obtain two

different Sn. Normalizing to log(N) may be accepted when the number of clones of all samples to be compared is the same as in (Abbate *et al.*, 2005; Pawlotsky *et al.*, 1998) but lacks justification otherwise.

A different measure of Shannon entropy may be obtained by the average of the per-site S, which would be normalized to log(4) for nucleotide sequences, or to log(20) for amino acid sequences—the natural logarithm of the alphabet size.

In this study, we use the per-haplotype S, with Sn normalized to log(h), where h is the number of estimated haplotypes in the population, according to the definition of Shannon entropy used in ecology. The meaning of the three normalizations is different. Where S/log(N) and S/N are scaled versions of S with equivalent statistical properties, and S/log(h) requires the estimate of h, and is influenced by its distribution.

2.5 Rarefaction

When the expected value of a diversity index depends of the sample size, we render comparable two samples of different size by rarefaction. The process of rarefaction (Magurran and McGill, 2010) is defined as a repeated resampling without replacement from a sample to a smaller sample size. In ecology, it is specifically used to compare species richness values, and to construct rarefaction curves. This is particularly useful for biased estimators where the bias is a function of the sample size, as the number of haplotypes, S and Sn.

2.6 Fringe trimming

When filtering the haplotypes of an NGS experiment above a given noise level, say 0.5% for instance, because of the sampling process there are chances to accept haplotypes with real abundances <0.5% while rejecting haplotypes that are >0.5% in the population. This produces fringes of haplotypes at the lower end of the NGS filtered sample, which could compromise the comparison of samples. A conservative way to make comparable samples of filtered data, eventually of different sizes, is to trim these fringes up to a given confidence level. Fringe trimming and haplotype filtering may be carried out in a single step by excluding all haplotypes with

$$P(n \leq n_i | N, P = 0.005) < 0.9$$

that is, by excluding the haplotypes with n_i reads for which the probability to observe up to n_i counts in a sample of size N, when the haplotype abundance in the population is 0.5%, is <90%. Both the noise level and the confidence level may be modified as required. As examples, 0.5 and 90% are just given, which fit our requirements on HCV NS3 samples, according to previous experience (Gregori *et al.* 2013, Ramirez *et al.* 2013). In the Supplementary Material, we show results filtering at 0.2 and at 1%, and trimming at different confidence levels.

2.7 Software and statistical methods

The in-silico sampling and all the computations and graphics were done on the open source R language and environment (R Core Team, 2013) using default libraries, and libraries in the Bioconductor project (Gentleman *et al.*, 2004) as the Biostrings library (Pages *et al.*, 2012). The R scripts are available upon request. NGS data simulations from a set of haplotypes of the high complexity population were performed by the Grinder program (Angly *et al.* 2012) with the parameters described in the Supplementary Material.

2.8 Data

Samples from two patients, one with an acute HCV infection and another with a chronic HCV infection were used to obtain the reference distributions used in the in-silico sampling. Six amplicons covering the NS3 HCV region were compared. The methods and protocols followed from patient sampling to UDPS sequencing have been described elsewhere (Cubero *et al.*, 2013). A coarse quality filter is used on the raw 454 reads to exclude all haplotypes represented by a single read, or those with more than two indeterminations or three gaps. We took the haplotype distribution of three of these amplicons as reference distributions of examples of quasispecies with low, mid and high complexity. The corresponding fasta files are included in the Supplementary Material, with frequencies (number of reads and percentage) in the header of each haplotype. The characterization of these reference quasispecies, along with the number of reads obtained in sequencing are given in Table 1. Although the reference distributions are based on HCV patient samples, we think that the conclusions are equally extensible to any virus passing through an RNA phase.

2.9 In-silico sampling

The statistical properties of diversity indices are studied by in-silico sampling from the reference distributions described above. The sampling is done by generating n random integers, where n is the sample size, between 1 and N, where N is the number of molecules in the reference population, with replacement, and assigning each random number to the corresponding haplotype by the population cumulative distribution (Fig. 1A).

3 RESULTS

3.1 Data characterization

Three reference distributions of different levels of viral quasispecies complexity—low, mid and high—are used as datasets (Table 1). The profile of these quasispecies populations may be depicted by the cumulative distribution of its haplotype frequencies (Fig. 1A). A complementary plot in Figure 1B gives the haplotype frequencies in descending order, with a dash-dot line at the 0.5% cutoff showing the incidence of filtering on each population. On the other hand, Supplementary Table S1 shows the effect of filtering at different noise levels on the reference populations. The most dramatic change is produced on the number of haplotypes, followed by the number of polymorphic sites. Mf and Pi show a smooth transition, while Sn displays a similar behavior except for the mid-complexity population where larger changes are

Table 1. Characterization of viral quasispecies population distributions used in the simulations

Population	Reads	Haplotypes	Polymorphic sites	Max, difference	Sn	Mf	Mean differences	Pi
Low	42 436	496	300	4	0.2194	5.089E-04	0.39	1.012E-03
Mid	43 300	550	269	6	0.2562	1.449E-03	0.93	2.502E-03
High	52 250	2064	266	14	0.5705	1.198E-02	5.23	1.585E-02

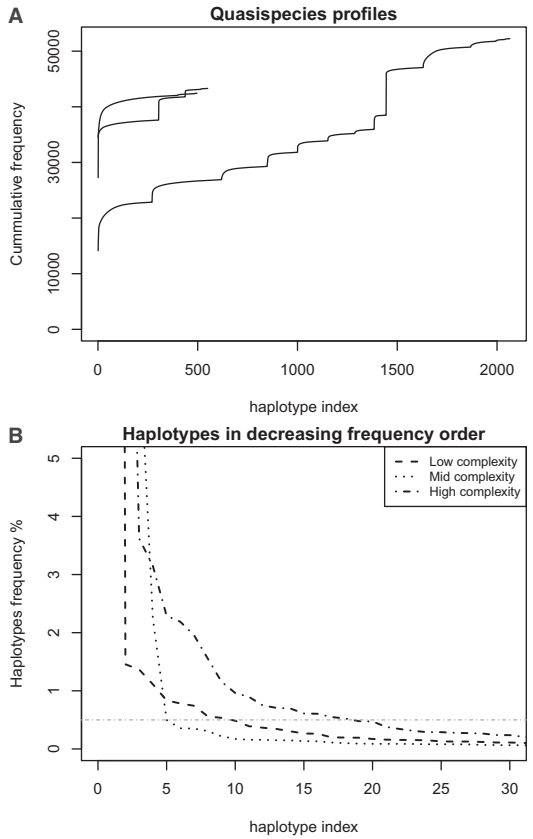

Fig. 1. (**A**) Quasiespecies profile as a cumulated distribution of the three reference populations used in the study. In abscissa the haplotypes are ordered primarily by the Hamming distance to the most frequent haplotype, and ties are determined in descending order of frequency. The first haplotype is the dominant one, while the last is the one showing more differences respect to the dominant and with a lower frequency in the population. The flatter the profile, the less complex is the quasispecies. (**B**) Quasispecies profile as a frequency distribution with the haplotypes ordered by decreasing frequency, the plot shows a detail of the full plot to view the impact of filtering on each of the three viral populations, with a dash-dot line at the 0.5% threshold

observed. Increasing levels of filtering are considered as the gradual elimination of genomes of low replication fitness. The number of reads excluded by these filters is particularly high for the high-complexity population, where filtering at 1% abundance represents the exclusion of 42.5% of the population. This is consistent with the production of tails of low fitness mutants from each of the haplotypes with enough replicating fitness.

3.2 Inference on complexity values in CCSS

We studied the distribution of S, Sn and Mf for CCSS samples of 20 and 50 clones, respectively, by 2000 replicates of in-silico sampling, for each of the three populations. The median of the observed values and the standard deviations are given in Supplementary Table S2A. The corresponding boxplots are shown in Figure 2A. Mf shows no bias with respect to the population value in any of the three populations. For S and Sn, we observe a bias with respect to the population value, which is sample size-dependent. When comparing pairs of samples of size 20 and 50, this differential bias could bring to the wrong conclusion that they come from populations of different diversity.

When applying the bias corrections of Hutcheson (Supplementary Equation SII) (Hutcheson, 1970) and Chao1 (Supplementary Equation SIII) (Chao *et al.*, 2009) the bias is partially corrected but remains sample size dependent (Supplementary Table S2B and Supplementary Fig. S1). While applying the rarefaction of the samples of size 50 to size 20, the median values of both samples are brought to the same level (Supplementary Table S2C and Supplementary Fig. S2) and the samples become comparable despite the different sample size.

In conclusion, the Mf values are not biased and may be directly compared with no further precaution, but the comparison of S or Sn values requires a bias correction. When the sizes of the two samples being compared are unbalanced, the comparison of S or Sn also requires the rarefaction of the big sample to the small sample size (Box 1). Inference is carried out by the t-test (Supplementary Equation SXI) (Hutcheson, 1970).

3.3 Inference on complexity values in NGS

We have taken 400 and 1000 reads as feasible sample sizes in a clinical setting for the determination of the viral complexity by NGS, with the same ratio as the 20 and 50 clones used in CCSS. Now we consider as population diversity values those obtained from the populations filtered at the noise level (Supplementary Table S2D). That is, the values that at best could be obtained by NGS.

While filtering at the noise level most 'rare' haplotypes are removed and the bias correction on S and Sn, as seen under CCSS, has a lower impact. On the distribution of 2000 replicates of samples of 400 and 1000 reads, filtered at 0.5%, we still observe a sample size differential bias, not only for S and Sn, but also for Mf in this sampling scheme (Supplementary Table S2D and Fig. 2B). The bias correction of Hutcheson (Supplementary Equation SII) has a limited impact, as expected.

We observed that the filtering has effects that depend of the sample size, as may be seen in Figure 3A with a scatterplot of the number of haplotypes observed on 2000 replicates of pairs of samples of size 400 and 1000. The small samples are clearly biased toward higher number of haplotypes despite being sampled from the same population, and filtered at the same abundance level. This effect is explained by the lower frequencies at which the same haplotypes are observed when increasing the sample size, particularly for those at the lower frequency end. The number of haplotypes observed before filtering in the big samples is higher than those observed in the small samples. As a consequence, the relative frequencies of the same haplotypes are lower in the big than in the small samples. This is illustrated in Figure 3B, where we show a barplot with the probabilities to observe an haplotype at a frequency of 0.5% in the population with a number of reads up to a given number of counts, both for samples of size 400 and 1000. The probability to observe such haplotype in a sample of size 400 with up to 2 reads is higher than the probability to observe the same haplotype in a sample of size 1000 with up to 5 reads.

According to this observation, after filtering at noise level, the small sample carries more information than the big sample. So

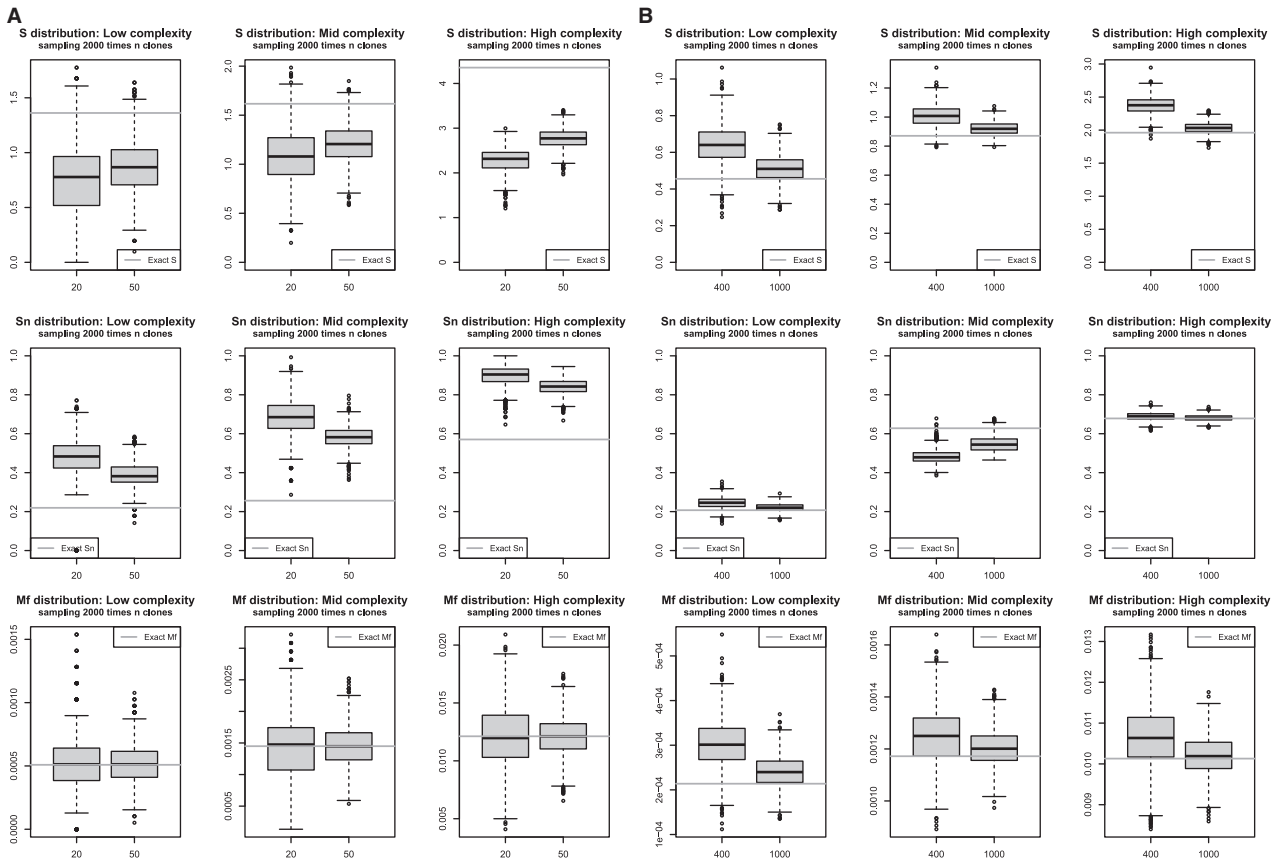

Fig. 2. (**A**) Boxplots with the distribution of the observed values of S, Sn and Mf in 2000 replicates CCSS experiments of size 20 and 50, for each of the three viral populations. (**B**) Boxplots with the distribution of the observed values of S, Sn and Mf in 2000 replicates of NGS experiments of size 400 and 1000, filtering at a noise level of 0.5%, for each of the three viral populations

the clean big sample is not useful for rarefaction. Instead the basis for rarefaction should be the raw big sample, including all rare haplotypes and artifacts. At each rarefaction cycle, the resampled reads should be filtered previously to compute the diversity indices. An alternative strategy could be trimming the haplotype fringes at noise level at a given confidence level. The effect of this additional filtering is seen by comparing Figure 3A and C. The bias of S, Sn and Mf is also greatly reduced (Supplementary Fig. S3). To assess the sensitivity of this method to small changes in the parameters, we explored the results filtering at levels of 0.2 and 1%, and trimming at 80, 90 and 99% confidence. Filtering deeper, into the noise level at 0.2%, the differential bias is exacerbated for S and Sn, and a differential bias is introduced in Mf. In these circumstances, the fringe trimming alleviates both, the absolute and the differential bias, of S, Sn and Mf, but does not completely cancel them. On the other hand, filtering well above noise level, at 1%, the absolute and the differential bias are rather limited and the fringe trimming strategy alone is able to compensate for the differential bias (See Supplementary Materials Parameters Sensitivity.doc).

Finally to assess the generality of the two strategies (rarefy the raw sampled data, and fringe trimming), we performed a prospective simulation study using the Grinder program (Angly *et al*. 2012) to simulate NGS data on the 18 clean haplotypes of the high complexity population, with corresponding frequencies. We used a linear error rate profile with three different mean error rates (0.15, 0.25 and 0.35%) and with three different slopes each. This simulation confirmed that the fringe trimming approach reduces both bias and differential bias, with the rarefaction giving the minimum differential bias, but showing higher absolute bias (See Supplementary Materials Grinder Simulations.doc).

In conclusion, under the NGS sampling scheme the comparison of diversity indices—S, Sn and Mf—requires of rarefaction or haplotypes fringe trimming above noise level. When the sizes of the two samples being compared are markedly unbalanced, the use of rarefaction should be preferred for S and Sn. The fringe trimming suffices for Mf in either case. The use of analytical formulations of rarefaction for S and Sn (Chao *et al*. 2013) is not possible with NGS data as the abundance filter discards singletons, doubletons and rare haplotypes in general. Resampling should be used instead.

4 DISCUSSION

Quasispecies dynamics represents an important challenge for the control of infectious diseases associated with RNA viruses and some DNA viruses. In particular, we are interested in improved molecular diagnosis of B and C hepatitis viruses, which are responsible for >500 million cases of chronic infections worldwide.

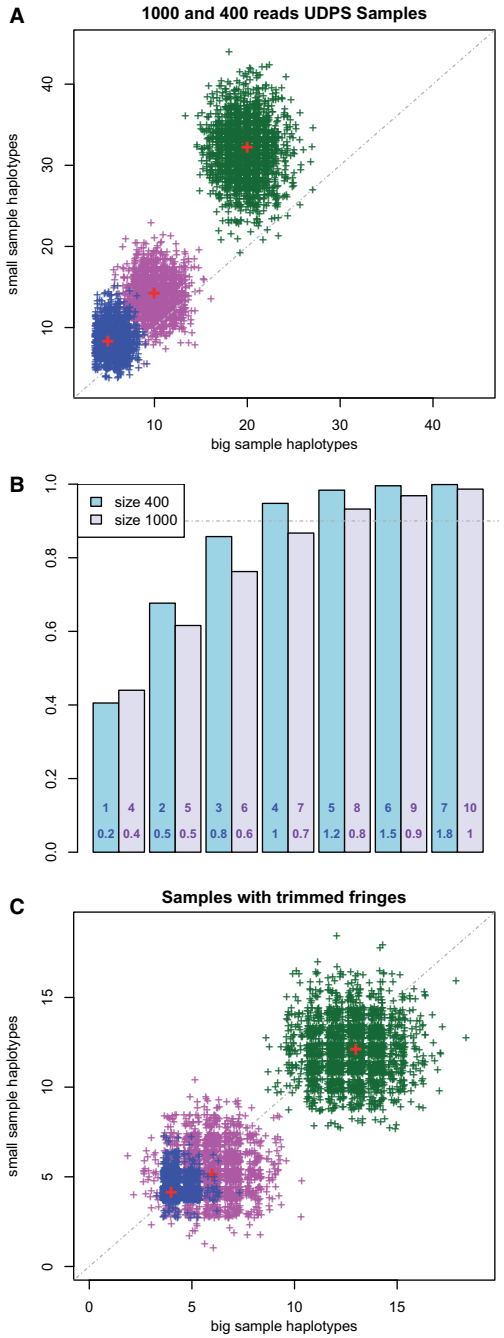

Fig. 3. (**A**) Scatterplot with the number of observed haplotypes in pairs of samples of size 400 and 1000 after filtering the haplotypes below the noise-level. The clouds are biased to higher values for the small samples. (**B**) Plots with the cumulated probabilities to observe an haplotype with an abundance in the population at the noise level (0.5% here) with growing number of reads, for samples of size 400 and 1000. The numbers inside the bars give the number of reads, on top, and the percentage in the sample below. (**C**) Scatterplot with the number of observed haplotypes in pairs of samples of size 400 and 1000 when the haplotype fringes have been trimmed at a 90% confidence level. The clouds are now centered on the diagonal (see Box 2). Use the Z test on S (Supplementary Equation SV), SN (Supplementary Equation SV with SVI and SVII) or MF (Supplementary Equation SX)

As strongly evidenced by recent reports, viral quasispecies complexity, measured by diversity indices, has clear clinical relevance in the course and prognosis of these diseases. Moreover, the adequate diagnosis of quasispecies complexity has direct implications for antiviral treatment failure because of its reflection in genetic barriers to resistance (Cheng *et al.*, 2013; Homs *et al.*, 2011, 2012; Jacobson *et al.*, 2011; Jardim *et al.*, 2013; Liu *et al.*, 2011; Margeridon-Thermet *et al.*, 2009, 2013; Nasu *et al.*, 2011; Nishijima *et al.*, 2012; Perales *et al.*, 2012; Poordad *et al.*, 2011; Powdrill *et al.*, 2011; Sarrazin and Zeuzem, 2010; Solmone *et al.*, 2009). Because of these reasons, it is paramount to establish a standard method of measuring and comparing diversity indices with statistic grounds and fitted to the expected degrees of viral quasispecies complexity.

We argue that the virus field would benefit of implementing solutions already established in ecology to compare diversity indices.

Useful connections between ecology and viral quasispecies have been previously established. Self-organization of subpopulations from a viral quasispecies that exhibited competition-colonization dynamics was approached by applying ecological models of biodiversity in spatially structured habitats (Tilman, 1994). The study revealed that host cell killing by viruses can be modulated by a trade-off between competition and colonization, and suggested a model of virus virulence based on intramutant spectrum interactions (Ojosnegros *et al.*, 2010). Also niche theory of competition communities and the replicator-mutator equation were combined to show that a typical quasispecies profile required both competition and cooperation among variants (Arbiza *et al.*, 2010; Vignuzzi *et al.*, 2006).

By a review of methods used in ecology that could be approached to describe RNA viral quasispecies, and thanks to deep coverage amplicon UDPS data, which has been used as source of in-silico sampling, we have studied the behavior and statistical properties of S, Sn and Mf under the sampling schemes of CCSS and NGS.

By CCSS, we may sample any virion with equal chance, but their estimated frequency will never be lower than 1/N, the granularity or resolution of the device, where N is the number of clones in the experiment. That is, when using 20 clones, no observed haplotype will have an estimated frequency in the population <5%. This granularity together with the high diversity of RNA viruses causes a systematic bias in the estimation of S and Sn. On the other hand, Mf does not suffer of estimation bias. Another consideration for the CCSS method is that we lack any means to control whether any of the observed clones are artifactual or of low abundance. In a recently published study (Ramirez *et al.*, 2013), we compared experimentally a patient sample of HBV sequenced in replicates by UDPS (two 454-FLX+, one 454-FLX and one 454 Junior, in the forward and reverse) and by 150 sequences obtained by CCSS. Among the 36 singleton haplotypes by CCSS, 10 were also identified by UDPS in 5–8 of the UDPS replicates, and 24 could not be identified in any of the replicates with 96 221 quality filtered reads covering the full amplicon. As an example, filtering at 0.5% the high complexity reference population, a 36.2% of the reads is removed (Supplementary Table S1), which means that in CCSS experiments, on this kind of viral populations, roughly one out

Box 1 Quasispecies diversity inference on S, Sn or Mf with CCSS samples

(1) Establish the significance level.
(2) Specify the null and alternative hypotheses.
(3) For Mf compute variance and go to step 7. For S or Sn, follow to next step.
(4) Use Chao1 (Supplementary Equation SIII), or other methods to estimate the number of haplotypes in the population from the distribution of haplotypes in the sample.
(5) Correct the bias in S or Sn by Hutcheson (Supplementary Equation SII), preferably to the third term, using the estimated number of haplotypes and the sample size.
(6) If the samples to be compared are unbalanced rarefy the big sample to the size of the small one to obtain an estimate of S or Sn and its variance. Use the observed value and the computed variance for the small sample, and the rarefied values of S or Sn and variance for the big sample.
(7) Test the null hypothesis by the Welch t-test (Supplementary Equation SXI) and compute the CI.

Box 2 Quasispecies diversity inference on S, Sn or Mf with NGS samples

(1) Establish noise level by controls
(2) Establish the significance level.
(3) Specify the null and alternative hypotheses.
(4) Clean the NGS sequences by the method of choice.
(5) Trim haplotypes at the noise level, at 90% confidence.
(6) Correct the bias in the Shannon entropy by Hutcheson (Supplementary Equation SII), preferably to the third term
(7) Compute variances by the theoretical expression.
(8) If the samples to be compared are unbalanced use rarefaction before filtering, as in Box 1.
(9) Test the null hypothesis by the Z test (Supplementary Equations SVI and SX) and compute the CI.

of each three clones observed will correspond to haplotypes <0.5% in the population.

Under the assumption that all observed clones are true members of the population, we observed by this sampling scheme that S and Sn are biased, and that the bias is sample size-dependent. We also observed that S shows a better behavior to analytical bias correction than Sn, and that Mf is an unbiased estimator. A less biased comparison of S or Sn values between samples requires of intra-sample normalization, composed of terms of correction (Supplementary Equations SII and SIII). When the sample sizes are unbalanced, the normalization requires a rarefaction of the big sample to the small sample size as well.

On the other hand, NGS methods are highly sensitive and reproducible but they are limited by the technical noise level. By discarding observed haplotypes, our diversity estimates are biased with respect to the true population values. We may nevertheless consider the haplotypes below the 0.5% frequency in our example as spurious or of low biological relevancy at the sampling time point. In this case, we used the diversity values of the filtered population as gold standard, being the best we can achieve by NGS. We observed a sample size-dependent bias on S, Sn and Mf.

The minimum differential bias is provided by rarefaction for S and Sn. For Mf, fringe trimming provides an unbiased comparison. When the samples to be compared are not very unbalanced and the abundance filter is above noise level, fringe trimming could give good results for S, Sn and Mf.

The in-silico sampling and simulation allowed us to assess the validity of the estimate and tests used in ecology when dealing with viral quasispecies with S, Sn and Mf, and permitted to identify the key points for less biased comparisons of complexity indices under the same sampling scheme.

In this work, we have empirically studied the statistical properties of S, Sn and Mf while observing the quasispecies viral complexity either by CCSS or by NGS, and through this we assessed the means for less biased comparisons of complexity indices. These methods could allow us to statistically conclude whether a viral quasispecies is expanding or contracting in diversity, independently of the size of the samples being compared.

In the Supplementary Material we give the formulation, and in Boxes 1 and 2 we propose the methods of data treatment for inference for each of the two methodologies, CCSS and NGS.

ACKNOWLEDGEMENTS

We are indebted to Maria Cubero, Celia Perales and Damir Garcia-Cehic for their collaboration in the experimental data that has been used in this manuscript.

Funding: SAF2009-10403 from Spanish Ministry of Economy and Competitiveness (MINECO), FIS PI10/01505, PI12/1893 and PI13/00456 from Health Ministry, ref.IDI-20110115 CDTI (Centro para el Desarrollo Tecnológico Industrial) from MINECO, CIBERehd is funded by the Instituto de Salud Carlos III, Madrid. Work at CBMSO was supported by grant BFU2011-23604 from MINECO, FIPSE and Fundación Ramon Areces.

Conflict of Interest: none declared.

REFERENCES

Abbate,I. *et al.* (2005) Cell membrane proteins and quasispecies compartmentalization of CSF and plasma HIV-1 from aids patients with neurological disorders. *Infect. Genet. Evol.*, **5**, 247–253.

Angly,F.E. *et al.* (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.*, **40**, e94.

Arbiza,J. *et al.* (2010) Viral quasispecies profiles as the result of the interplay of competition and cooperation. *BMC Evol. Biol.*, **10**, 137.

Archer,J. *et al.* (2012) Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. *BMC Bioinformatics*, **13**, 47.

Beerenwinkel,N. and Zagordi,O. (2011) Ultra-deep sequencing for the analysis of viral populations. *Curr. Opin. Virol.*, **1**, 413–418.

Beerenwinkel,N. *et al.* (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front. Microbiol.*, **3**, 329.

Cabot,B. *et al.* (2000) Nucleotide and amino acid complexity of hepatitis C virus quasispecies in serum and liver. *J. Virol.*, **74**, 805–811.
Chao,A. and Shen,T. (2003) Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stats.*, **10**, 429–443.
Chao,A. *et al.* (2009) Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*, **90**, 1125–1133.
Chao,A. *et al.* (2010) Phylogenetic diversity measures based on Hill numbers. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **365**, 3599–3609.
Chao,A. *et al.* (2013) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol. Monogr.*, [Epub ahead of print, doi:10.1890/13-0133.1].
Cheng,Y. *et al.* (2013) Increased viral quasispecies evolution in HBeAg seroconverter patients treated with oral nucleoside therapy. *J. Hepatol.*, **58**, 217–224.
Colwell,R. *et al.* (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J.Plant Ecol.*, **5**, 3–21.
Cubero,M. *et al.* (2013) Identification of host and viral factors involved in a dissimilar resolution of hepatitis C virus infection. *Liver Int.*, [Epub ahead of print, doi: 10.1111/liv.12362].
Domingo,E. *et al.* (2005) Quasispecies dynamics and RNA virus extinction. *Virus Res.*, **107**, 129–139.
Domingo,E. *et al.* (2012) Viral quasispecies evolution. *Microbiol. Mol. Biol. Rev.*, **76**, 159–216.
Fishman,S.L. and Branch,A.D. (2009) The quasispecies nature and biological implications of the hepatitis C virus. *Infect. Genet. Evol.*, **9**, 1158–1167.
Flaherty,P. *et al.* (2012) Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Res.*, **40**, e2.
Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
Gilles,A. *et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, **12**, 245.
Grande-Perez,A. *et al.* (2002) Molecular indetermination in the transition to error catastrophe: systematic elimination of lymphocytic choriomeningitis virus through mutagenesis does not correlate linearly with large increases in mutant spectrum complexity. *Proc. Natl Acad. Sci. USA*, **99**, 12938–12943.
Gregori,J. *et al.* (2013) Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. *Plos One*, [Epub ahead of print, doi: 10.1371/journal.pone.0083361].
Heip,C. and Engels,P. (1974) Comparing species diversity and evenness indices. *J. Mar. Biol. Assoc. U.K.*, **54**, 559–563.
Hellmann,J. and Fowler,G. (1999) Bias, precision, and accuracy of four measures of species richness. *Ecol. Appl.*, **9**, 824–834.
Herrmann,E. *et al.* (2000) Hepatitis C virus kinetics. *Antivir. Ther.*, **5**, 85–90.
Homs,M. *et al.* (2011) 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.*, **39**, 8457–8471.
Homs,M. *et al.* (2012) Quasispecies dynamics in main core epitopes of hepatitis B virus by ultra-deep-pyrosequencing. *World J. Gastroenterol.*, **18**, 6096–6105.
Huse,S.M. *et al.* (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.*, **8**, R143.
Hutcheson,K. (1970) A test comparing diversities based on the Shannon formula. *J. Theor. Biol.*, **29**, 151–154.
Jacobson,I.M. *et al.* (2011) Telaprevir for previously untreated chronic hepatitis C virus infection. *N. Engl. J. Med.*, **364**, 2405–2416.
Jardim,A.C. *et al.* (2013) Analysis of HCV quasispecies dynamic under selective pressure of combined therapy. *BMC Infect. Dis.*, **13**, 61.
Jost,L. (2006) Entropy and diversity. *Oikos*, **113**, 363–375.
Liu,F. *et al.* (2011) Evolutionary patterns of hepatitis B virus quasispecies under different selective pressures: correlation with antiviral efficacy. *Gut*, **60**, 1269–1277.
Loman,N.J. *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.*, **30**, 434–439.
Macalalad,A.R. *et al.* (2012) Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. *PLoS Comput. Biol.*, **8**, e1002417.
Magurran,A. (2004) *Measuring Biological Diversity*. Blackwell Publishing, Oxford.
Magurran,A. and McGill,B.J., eds. (2010) *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press, Oxford, UK.
Margeridon-Thermet,S. *et al.* (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.
Margeridon-Thermet,S. *et al.* (2013) Low-level persistence of drug resistance mutations in hepatitis B virus-infected subjects with a past history of Lamivudine treatment. *Antimicrob. Agents Chemother.*, **57**, 343–349.
Mild,M. *et al.* (2011) Performance of ultra-deep pyrosequencing in analysis of HIV-1 pol gene variation. *PLoS One*, **6**, e22741.
Nasu,A. *et al.* (2011) Genetic heterogeneity of hepatitis C virus in association with antiviral therapy determined by ultra-deep sequencing. *PLoS One*, **6**, e24907.
Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.
Nemenman,I. *et al.* (2011) Entropy and inference, revisited. *arXiv*, physics/0108025v2.
Neumann,A.U. *et al.* (1998) Hepatitis C viral dynamics *in vivo* and the antiviral efficacy of interferon-alpha therapy. *Science*, **282**, 103–107.
Nishijima,N. *et al.* (2012) Dynamics of hepatitis B virus quasispecies in association with nucleos(t)ide analogue treatment determined by ultra-deep sequencing. *PLoS One*, **7**, e35052.
Nowak,M.A. *et al.* (1991) Antigenic diversity thresholds and the development of AIDS. *Science*, **254**, 963–969.
Ojosnegros,S. *et al.* (2010) Competition-colonization dynamics in an RNA virus. *Proc. Natl Acad. Sci. USA*, **107**, 2108–2112.
Pages,H. *et al.* (2012) *Biostrings: String Objects Representing Biological Sequences, and Matching Algorithms*. R package 2.24.1. http://www.bioconductor.org/packages/2.14/bioc/html/Biostrings.html (8 January 2014, date last accessed).
Pardo,L. *et al.* (1997) Large sample behavior of entropy measures when parameters are estimated. *Commun. Stat. Theory Methods*, **26**, 483–501.
Pawlotsky,J.M. *et al.* (1998) Interferon resistance of hepatitis C virus genotype 1b: relationship to nonstructural 5A gene quasispecies mutations. *J. Virol.*, **72**, 2795–2805.
Perales,C. *et al.* (2012) The impact of quasispecies dynamics on the use of therapeutics. *Trends Microbiol.*, **20**, 595–603.
Perales,C. *et al.* (2010) Mutant spectra in virus behavior. *Future Virol.*, **5**, 679–698.
Poordad,F. *et al.* (2011) Boceprevir for untreated chronic HCV genotype 1 infection. *N. Engl. J. Med.*, **364**, 1195–1206.
Powdrill,M.H. *et al.* (2011) Contribution of a mutational bias in hepatitis C virus replication to the genetic barrier in the development of drug resistance. *Proc. Natl Acad. Sci. USA*, **108**, 20509–20513.
Prosperi,M.C. and Salemi,M. (2012) QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics*, **28**, 132–133.
Prosperi,M.C. *et al.* (2011) Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinformatics*, **12**, 5.
R Core Team. (2013) *R: A Language and Environment for Statistical Computing*. Foundation for Statistical Computing, Vienna, Austria.
Ramirez,C. *et al.* (2013) A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. *Antiviral Res.*, **98**, 273–283.
Salicru,M. *et al.* (1993) Asymptotic distributions of (h,Fi)-entropies. *Commun. Stat. Theory Methods*, **22**, 2015–2031.
Sarrazin,C. and Zeuzem,S. (2010) Resistance to direct antiviral agents in patients with hepatitis C virus infection. *Gastroenterology*, **138**, 447–462.
Shannon,C. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.
Solmone,M. *et al.* (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.
Tilman,D. (1994) Competition and biodiversity in spatially structured habitats. *Ecology*, **75**, 2–16.
Tuomisto,H. (2010) A consistent terminology for quantifying species diversity?Yes, it does exist. *Oecologia*, **164**, 853–860.
Vandenbroucke,I. *et al.* (2011) Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications. *Biotechniques*, **51**, 167–177.
Vignuzzi,M. *et al.* (2006) Quasispecies diversity determines pathogenesis through cooperative interactions within a viral population. *Nature*, **439**, 344–348.
Walther,B. and Moore,J. (2005) The concept of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, **28**, 815–829.
Wolinsky,S.M. *et al.* (1996) Adaptive evolution of Human Immunodeficiency Virus-Type 1 during the natural course of infection. *Science*, **272**, 537–542.
Zagordi,O. *et al.* (2012) Read length versus depth of coverage for viral quasispecies reconstruction. *PLoS One*, **7**, e47046.