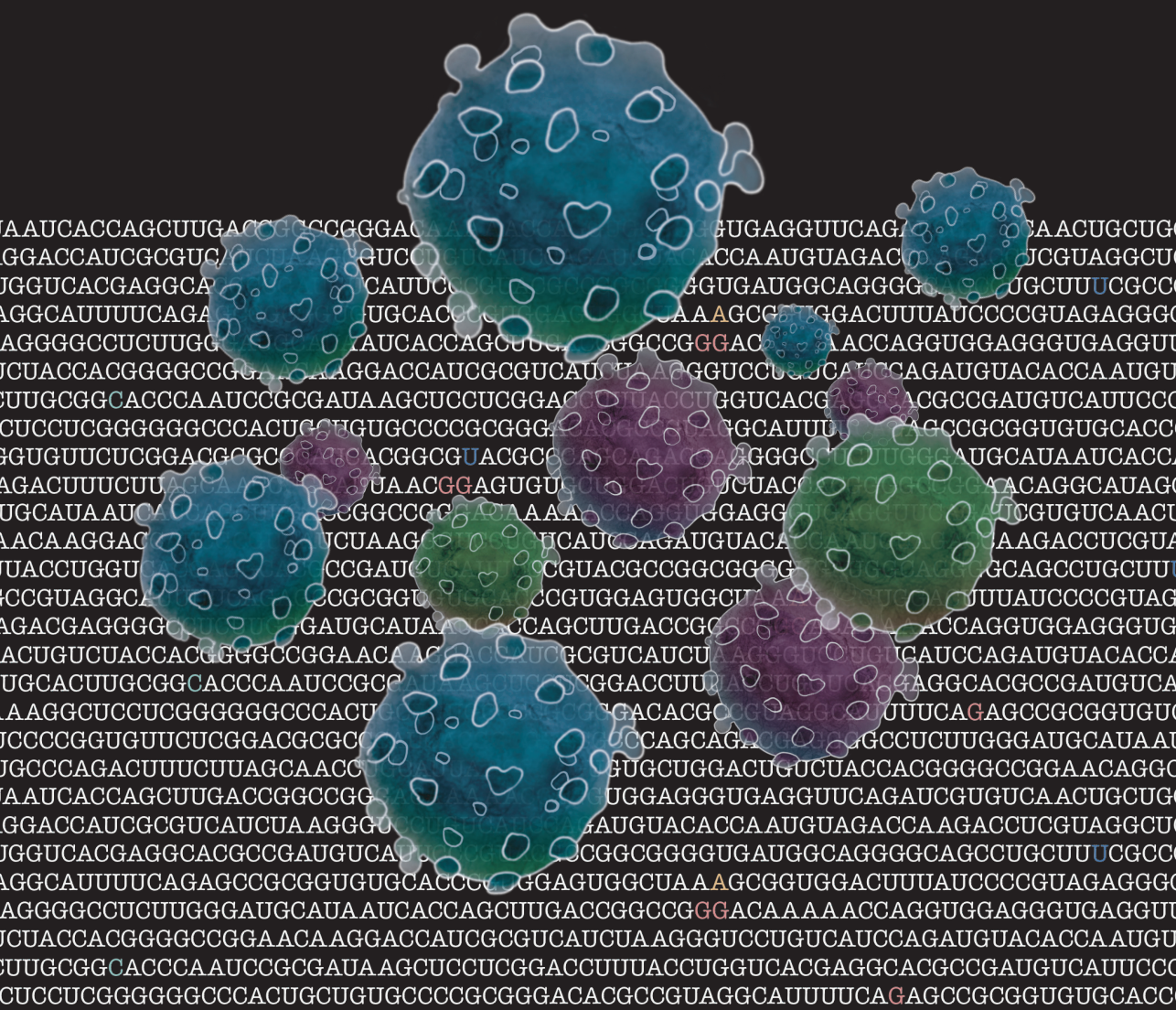
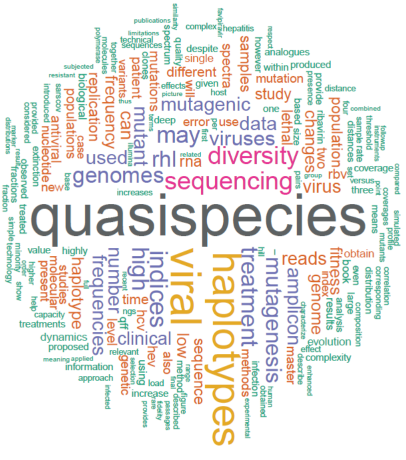
Viral quasispecies diversity and evolution

A BIOINFORMATICS MOLECULAR APPROACH

Josep Gregori, Francisco Rodríguez-Frías, Josep Quer



Viral quasispecies diversity and evolution



Viral quasispecies diversity and evolution

A bioinformatics molecular approach

EDITORS
JOSEP GREGORI, FRANCISCO RODRÍGUEZ-FRÍAS
AND JOSEP QUER

Il Pensiero Scientifico Editore

Edited and published by
Il Pensiero Scientifico Editore, srl
Via San Giovanni Valdarno 8
00138 Rome (Italy)
Ph (+39) 06 862821 Fax (+39) 06 86282250
pensiero@pensiero.it
www.pensiero.it – www.vapensiero.info
www.facebook.com/PensieroScientifico
twitter.com/ilpensiero
www.pinterest.com/ilpensiero

First published in Italy 2023

Typeset by Doppiosegno s.n.c., Rome (Italy)
Printed and bound by Ti Printing S.r.l.
Via delle Case Rosse 23, 00131 Rome (Italy)
Cover design by Susanna Quer
Compiled on 2023-03-02 18:32:13+01:00
Graphic design Antonella Mion
Editorial coordination Silvana Guida
ISBN 978-88-490-0758-9

Authors

EDITORS

Josep Gregori
Liver Diseases-Viral Hepatitis
Liver Unit, Vall d'Hebron
Institut de Recerca (VHIR)
Vall d'Hebron
Barcelona Hospital Campus
(HUVH)
Barcelona, Spain
josep.gregori@gmail.com

Francisco Rodríguez-Frías
Liver Pathology Unit
Clinical Biochemistry and
Microbiology Dept. (HUVH)
Clinical Biochemistry
Research Group (VHIR)
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain
Biochemistry and Molecular
Biology Dept., UAB
Bellaterra, Spain
frarodri@gmail.com

Josep Quer
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain
Biochemistry and Molecular
Biology Dept., UAB
Bellaterra, Spain
josep.quer@vhir.org

AUTHORS

Caroline Melanie Adombie
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain
Institute of Agropastoral
Management
University Peleforo
Gon Coulibaly
Korhogo, Côte d'Ivoire

Albert Bosch
Enteric Virus Laboratory
Section of Microbiology,
Virology and Biotechnology
Dept. of Genetics, Microbiology
and Statistics, School of Biology
University of Barcelona
Barcelona, Spain
Enteric Virus Laboratory, INSA
University of Barcelona
Barcelona, Spain

Maria Buti
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain
Medicine Dept., UAB
Bellaterra, Spain

Carolina Campos
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain
Biochemistry and Molecular
Biology Dept., UAB
Bellaterra, Spain

Sergi Colomer-Castell
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain
Biochemistry and Molecular
Biology Dept., UAB
Bellaterra, Spain

Maria Francesca Cortese
Microbiology Dept., VHIR-HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain

Esteban Domingo
Centro de Biología Molecular "Severo Ochoa" (CBM)
Universidad Autónoma de Madrid (UAM)
Consejo Superior de Investigaciones Científicas (CSIC)
Campus Cantoblanco
Madrid, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain

Juan I. Esteban
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain
Medicine Dept., UAB
Bellaterra, Spain

Isabel Gallego
CBM (CSIC-UAM)
Campus Cantoblanco
Madrid, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain

Damir Garcia-Cehic
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain

Josep Gregori
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain

Mercedes Guerrero-Murillo
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain

Susanna Guix
Enteric Virus Laboratory
Section of Microbiology, Virology and Biotechnology
Dept. of Genetics, Microbiology and Statistics, School of Biology
University of Barcelona
Barcelona, Spain
Enteric Virus Laboratory, INSA
University of Barcelona
Barcelona, Spain

Marta Ibañez-Lligoña
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain

Celia Perales
Liver Diseases-Viral Hepatitis
Liver Unit,VHIR-HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain
CBM (CSIC-UAM)
Campus Cantoblanco
Madrid, Spain

Rosa Pintó
Enteric Virus Laboratory
Section of Microbiology, Virology and Biotechnology
Dept. of Genetics, Microbiology and Statistics, School of Biology
University of Barcelona
Barcelona, Spain
Enteric Virus Laboratory INSA
University of Barcelona
Barcelona, Spain

Josep Quer
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR, HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain
Biochemistry and Molecular Biology Dept., UAB
Bellaterra, Spain

Ariadna Rando-Segura
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain
Microbiology Dept., VHIR-HUVH
Barcelona, Spain

Mar Riveiro-Barciela
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain
Medicine Dept., UAB
Bellaterra, Spain

Francisco Rodríguez-Frías
Liver Pathology Unit
Clinical Biochemistry and Microbiology Dept. (HUVH)
Clinical Biochemistry Research Group (VHIR)
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain
Biochemistry and Molecular Biology Dept., UAB
Bellaterra, Spain

Miquel Salicrú
Statistics Dept., Biology Faculty
University of Barcelona
Barcelona, Spain

Alex Sanchez
Statistics Dept., Biology Faculty
University of Barcelona
Barcelona, Spain
Bioinformatics and Statistics Unit
VHIR-HUVH
Barcelona, Spain

María Eugenia Soria
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain
CBM (CSIC-UAM)
Campus Cantoblanco
Madrid, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain

David Tabernero
Liver Diseases-Viral Hepatitis
Liver Unit, VHIR-HUVH
Barcelona, Spain
CIBERehd
Instituto de Salud Carlos III
Madrid, Spain

Contents

Foreword
Viral quasispecies complexity in a nutshell

The advent of deep nucleotide sequencing has allowed detection of myriads of low frequency genomes that compose viral populations. The presence of mutant spectra that was initially inferred from comparison of biological or molecular viral clones subjected to Sanger sequencing has been fully confirmed with the application of the new sequencing methodologies. If anything, deep sequencing – now with sufficient reliability regarding detection of mutations present in the template molecules, as opposite to sequencing artifacts –, are unveiling even a higher degree of genome population complexity and dynamics than suspected from clonal analyses. Indeed, RNA viruses and many DNA viruses consist of collections of vast mutant clouds in the sense that most individual genomes differ in one or more nucleotides from their companions in the same population.

One may be tempted to dismiss low frequency viral genomes as mere genetic noise resulting from high mutation rates. The stubborn reality is, however, that minority genomes occasionally display a biological behaviour which is different (if not opposite) from that displayed by the ensemble where they are immersed (virulent versus attenuated, antiviral drug-resistant versus antiviral drug-sensitive, antibody-resistant versus antibody-sensitive, interferon responsive versus interferon non-responsive, competent in interferon induction versus defective in interferon-induction, and so forth and so on). Such hidden potential for phenotypic variation is rendered relevant in the context of selection (positive and negative) and random drift, which are the two major forces of genetic diversification. Indeed, such pervasive forces have the power to convert the progeny of those genomes that rank as a minority at a given time point, into those that are dominant at a later time, and vice versa. Such fleeting

dominances may occur within hours, minutes or seconds. We still know little regarding the influence of the time component in viral quasispecies dynamics. Whatever the precise time frame of replacements of minority subpopulations, it is not advisable to sweep minorities under the carpet. We have to confront their quantification, organization, and biological meaning.

What we term the consensus sequence, which represents a weighted average of all the sequences that compose the sample of a population under study, is an abstraction. The consensus sequence may not even coincide with one of the individual genomes in the population it intends to represent. These concepts apply to any RNA virus that has been studied to date, as well as a large number of DNA viruses with a small genome whose replication is catalyzed by low fidelity cellular DNA polymerases. For complex (large genome size) DNA viruses, the level of population complexity and its dynamics within infected organisms is still largely an open question.

The results of deep sequencing pose two major challenges: to find means to organize the astonishingly large amount of information being obtained, and to understand its biological significance. The situation is reminiscent of that of the 1980's when myriads of nucleotide sequences of cellular and viral genomes began to fill data banks. The question was how to translate that information into biological meaning. Struggling with such translation is still ongoing. In viral quasispecies, the first step to approach the input data challenge is the one covered by Josep Gregori, Francisco Rodríguez-Frías and Josep Quer in the present book: to quantify a number of diversity indices that describe complementary features of viral populations. In the commented series of studies, classic diversity indices, together with others imported from ecology to describe biological diversity, are applied to viral quasispecies. Collectively such indices inform of the complexity of viral populations, and mark a way to inquire into functional implications. The book is based on the pioneering contributions of the authors, using deep sequencing to unveil the composition of pathogenic RNA viruses, notably human hepatic viruses such as hepatitis C virus, and to interpret treatment responses and failures in terms of quasispecies dynamics. The book is both informative and tutorial. The authors take advantage of having shared expertise in bioinformatics

and clinical medicine for many years. The collection of articles, followed by concise concluding remarks, will guide the reader into understanding the mathematic formulations conductive to diversity index calculation, and how the results find an application to the clinical setting. The book explains the significance of each individual diversity index, and how it may help in quantifying both, standard viral genome populations, and those subject to antiviral interventions or to lethal mutagenesis. Indeed, lethal mutagenesis, which means the extinction of a virus by an excess of mutations, may be viewed as a fitness-decreasing earthquake of population diversity distortion.

Remarkably, the compilation of studies in the book is timely even for those who have just joined virology for the task of confronting the emerged human coronavirus SARS-CoV-2. The experts anticipated that because of its large genome size, this coronavirus would limit the extent of its variation. The reality is that the virus consists of extremely complex mutant swarms whose characterization will benefit of the present tutorial. Intended for students of an annual Master Degree course that the authors coordinate, the book will also interest other students and professionals of basic and applied virology. They will all have to face viral population complexity with its many clinical implications. Indeed, the book's main topic, that is, complexity indices, is relevant to the new definition of "wild type" virus as a collection of genomes, to the mechanisms of viral adaptability inherent to quasispecies dynamics, and to the planning of antiviral interventions and the understanding of their outcome. To end with an illustrative example, the most dramatic event a virus can undergo – that is, its extinction by lethal mutagenesis – occurs without any variation in its consensus sequence, at least as long as the latter can still be determined.

The drama is in the mutant spectra.

Esteban Domingo

Cantoblanco, Madrid

Preface

This book contains a collation of selected publications by our group dealing with the diversity, complexity, and evolution of viral quasispecies. It describes the developments attained in our laboratory to distill complexity into something simple but still informative. The challenge we faced was to characterize viral quasispecies in terms of their diversity; however, the concepts diversity and complexity have multiple faces in the world of viruses, and a summary value given by a single diversity index can be misleading. The progress in our work includes the use of visual tools to represent viral diversity in simple terms, while retaining, whenever possible, high biological meaning. All the computations imply the use of quasispecies haplotypes and frequencies, with high sequencing coverage. That is, molecular analysis of quasispecies composition.

The book starts with an introduction and a historical note that narrates the transition from molecular cloning to NGS in viral quasispecies studies. It includes development of the software used by our group to obtain amplicon haplotypes with their frequencies from NGS data. Next, the related articles are listed and briefly described, and a section of the book is devoted to each of them. Although most of these articles are published in open access and freely available, they have been gathered together, to provide a complete picture for easy reference. Each article is preceded by an abstract, remarks or conclusions where applicable, and a list of highlights. The book ends with general closing remarks and a note about the meaning and implications of acquiring samples from a dynamic system. To facilitate the reading, each actual article and its supplementary material in the collection are marked with a different colour. In order to keep the book within limits, but at the same time be as informative

as possible, we selected some supplementary materials (see underlined reference) to be added at the end of each article.

This 12-year journey was undertaken with friends and colleagues. Thanks are given to all coauthors in the articles presented and to the corresponding peer reviewers who contributed to enrich the publications. We thank Ms. Celine Cavallo for English language support. A special mention is given to Prof. Esteban Domingo, a colleague and coauthor, who held our hands in the transition from CCSS to NGS and beyond. An important part of the work done in our laboratory in the period that goes from 2011 to 2022 was in collaborative projects where Roche Diagnostics, Spain, was an important partner. We would like to mention in particular Mr. Jaume Vives, General Director, Dr. Artur Palet, Business Development Director, and Dr. Carlos Manchado, Medical and Innovation Manager, and thank them for their continued support and involvement.

Josep Gregori i Font

Francisco Rodríguez-Frías

Josep Quer i Sivila

Introduction

Viruses, and in particular RNA viruses, can quickly adapt to changing environments, thanks to the high error rates of the viral polymerases involved in genome replication. As an example, HCV has an estimated error rate of $1. \times 10^{-4}$ to $1. \times 10^{-3}$ mutations per nucleotide per genomic replication, yielding a natural evolutionary rate of 1.5×10^{-3} base substitutions per site per year. In a genome 9600 base pairs in size, more than 9 errors (substitutions) can be produced every time a virion is replicated. In infections involving viral loads in the order of 10^6 to 10^7 copies/mL of blood (6 to 7 logs) and a replication cycle of just a few hours, an estimated 10^{12} virions are engendered and eliminated daily.

Thus, numerous variants are produced during virion replication, but their viability and abundance in the viral quasispecies population is decided by their replicative capacity or fitness. The variant with greatest fitness will dominate the quasispecies; that is, it will become the most abundant variant in the population. Although a steady state could be expected, the quasispecies theory predicts that steady state may not be reached, as genetic drift, pressure from the host immune system, or the action of treatments can cause changes to occur in the relative fitness of the variants present. New mutants with the capability to predominate may appear, and these, in turn, may be outcompeted by others generated. This continuous flux is known as quasispecies dynamics. Although the quasispecies theory was formulated as an integration of Darwinian evolution and information theory, what we witness in the dynamics of mutant spectra is a sort of genetic motion within a confined space, a potential well that represents the functional genetic space assigned to the corresponding viral subtype. Within these dynamics, Darwinian evolution will occur only when a quasispecies is

able to escape this confinement due to a highly improbable event in which a new genotype or subtype is produced and established.

Viral adaptability makes treatments with direct-acting antivirals (DAAs) that target a single genomic region ineffective, and necessitates the use of multidrug treatments that act against several parts of the genome. This characteristic has been a challenge in the effort to obtain effective therapies against HIV and HCV, for example. Viral adaptability is also a basis to enable zoonotic disease, the spread of these pathogens from animals to humans, as has occurred with SARS-CoV-2, Ebola, MERS, and Zika, to name a few. Zoonosis is of great concern as a likely cause of future pandemics.

Covid-19 infection has been a lesson learned worldwide. Viruses such as HIV, also extending throughout the world, or the lethal Ebola outbreaks in some African countries are real threats, but they only affect a small part of humanity. SARS-CoV-2 has indiscriminately affected us all. Covid-19 infection has shown us that humanity will undoubtedly face new viral threats in this rapidly changing world, where natural environments are being destroyed at an unviable pace and previously remote places are brought closer through increasing travel and commercial exchange. We cannot know what virus will cause the next pandemic; we only know that it is more likely to appear sooner than later during our lives. The reports from the WHO [1], *Horizon Europe* [2], Harvard [3], *Nature* [4], and *The Lancet* [5] provide examples of this concern.

In contrast to bacteria, which have certain features in common that allow the development of effective antibiotics against different species, viruses are very simple entities, lacking sufficient structural similarities to make an overall approach feasible. Furthermore, viral quasispecies causing the same type of infection can differ to some degree in each patient. It is our belief that better understanding of quasispecies dynamics will help us in the development of more effective treatments. This entails not only the detection of mutations potentially resistant to a treatment, but also their effects on quasispecies composition.

This book is a collection of publications by our group, reporting the tools we devised to monitor and quantify the changes in quasispecies composition. The articles corre-

spond to research initiated in 2011 and cover our progress in quasispecies characterization up to 2023. In addition to the studies presented here, we produced a number of other publications in this period: reports on the development of NGS methods to subtype HCV clinical samples ([6],[7],[8]); descriptions of resistance mutations in HCV patients failing DAA treatments ([9],[10],[11]); accurate genotyping of HBV clinical samples [12]; and other basic research involving sequencing of clinical samples from patients infected by various viruses (e.g., [13],[14],[15],[16]), including a few on SARS-CoV-2 (e.g., [17],[18]).

REFERENCES

1. Imagining the future of pandemics and epidemics: a 2022 perspective. Geneva: World Health Organization, 2022. Licence:CC BY-NC-SA 3.0 IGO. https://www.who.int/publications/i/item/9789240052093
2. Smith J. Q&A: future pandemics are inevitable, but we can reduce the risk. Horizon. The EU reasearch & Innovation Magazine. 16th Dec 2021. https://ec.europa.eu/research-and-innovation/en/horizon-magazine/qa-future-pandemics-are-inevitable-we-can-reduce-risk
3. Caruso C. Covid-19's lessons for future pandemics. Harvard Medical School, News & Research, Nov. 17, 2022. https://hms.harvard.edu/news/covid-19s-lessons-future-pandemics
4. Sridhar D. Five ways to prepare for the next pandemic. Nature 2022; 610:S50. https://doi.org/10.1038/d41586-022-03362-8
5. The Lancet Respiratory Medicine. Future pandemics: failing to prepare means preparing to fail. Lancet Respir Med 2022; 10(3):221-2. https://doi.org/10.1016/S2213-2600(22)00056-X
6. Quer J, Gregori J, Rodríguez-Frías F, et al. High-resolution hepatitis C virus subtyping using NS5B deep sequencing and phylogeny, an alternative to current methods. J Clin Microbiol 2015; 53(1):219-26. https://doi.org/10.1128/JCM.02093-14 Erratum in: J Clin Microbiol 2016; 54(7):1933.
7. Rodríguez-Frías F, Nieto-Aponte L, Gregori J, et al. High HCV subtype heterogeneity in a chronically infected general population revealed by high-resolution hepatitis C virus subtyping. Clin Microbiol Infect 2017; 23(10):775.e1-775.e6. https://doi.org/10.1016/j.cmi.2017.02.007
8. Del Campo JA, Parra-Sánchez M, Figueruela B, et al. Hepatitis C virus deep sequencing for sub-genotype identification in mixed infections: a real-life experience. Int J Infect Dis 2018; 67:114-7. https://doi.org/10.1016/j.ijid.2017.12.016
9. Soria ME, Gregori J, Chen Q, et al. Pipeline for specific subtype amplification and drug resistance detection in hepatitis C virus. BMC Infect Dis 2018; 18(1):446. https://doi.org/10.1186/s12879-018-3356-6
10. Perales C, Chen Q, Soria ME, et al. Baseline hepatitis C virus resistance-associated substitutions present at frequencies lower than 15% may be clinically significant. Infect Drug Resist 2018; 11:2207-10. https://doi.org/10.2147/IDR.S172226
11. Chen Q, Perales C, Soria ME, et al. Deep-sequencing reveals broad subtype-specific HCV resistance mutations associated with treatment failure. Antiviral Res 2020; 174:104694. https://doi.org/10.1016/j.antiviral.2019.104694
12. Caballero A, Gregori J, Homs M, et al. Complex genotype mixtures analyzed by deep sequencing in two different regions of hepatitis B virus. PLoS One 2015; 10(12):e0144816. https://doi.org/10.1371/journal.pone.0144816
13. Homs M, Rodríguez-Frías F, Gregori J, et al. Evidence of an exponential decay pattern of the hepatitis delta virus evolution rate and fluctuations in quasispecies complexity in long-term studies of chronic delta infection. PLoS One 2016 Jun 30;

11(6):e0158557. https://doi.org/10.1371/journal.pone.0158557

14. Sopena S, Godoy C, Tabernero D, et al. Quantitative characterization of hepatitis delta virus genome edition by next-generation sequencing. Virus Res 2018; 243:52-9. https://doi.org/doi:10.1016/j.virusres.2017.10.003

15. Sabrià A, Pintó RM, Bosch A, et al.; Working Group for the Study of Outbreaks of Acute Gastroenteritis in Catalonia). Characterization of intra- and inter-host norovirus P2 genetic variability in linked individuals by amplicon sequencing. PLoS One 2018; 13(8):e0201850. https://doi.org/10.1371/journal.pone.0201850 Erratum in: PLoS One 2018 Dec 19; 13(12):e0209714.

16. Sabrià A, Gregori J, Garcia-Cehic D, et al. Evidence for positive selection of hepatitis A virus antigenic variants in vaccinated men-having-sex-with men patients: implications for immunization policies. eBioMedicine 2019; 39:348-57. https://doi.org/10.1016/j.ebiom.2018.11.023

17. Andrés C, Garcia-Cehic D, Gregori J, et al. Naturally occurring SARS-CoV-2 gene deletions close to the spike S1/S2 cleavage site in the viral quasispecies of covid19 patients. Emerg Microbes Infect 2020; 9(1):1900-11. https://doi.org/10.1080/22221751.2020.1806735

18. Gregori J, Cortese MF, Piñana M, et al. Host-dependent editing of SARS-CoV-2 in covid-19 patients. Emerg Microbes Infect 2021; 10(1):1777-89. https://doi.org/10.1080/22221751.2021.1969868

Background: haplotypes, from clones to reads

Before the development of current sequencing methods, viral quasispecies study could only be done by molecular cloning into plasmid vectors using restriction enzymes, followed by Maxam-Gilbert or Sanger sequencing ([1],[2]), with the limitations of this technique in both time and cost ([3],[4]). Most reports based their results on alignment of a very modest number of sequences – 10 to 20 per sample – with only a few reaching more than 50 clones. The clones provided a set of amplicon haplotypes (i.e., the different sequences identified) and their abundance (frequency) in the quasispecies, which served as the basis for further studies and calculations.

By sequencing molecular clones of the hepatitis C virus (HCV), our group described for the first time that HCV has a quasispecies nature [5]. This means that in any single patient, HCV is composed of a complex mixture of different but closely related genomes that undergoes continuous changes due to competitive selection [6] and cooperation [7] between arising mutants. The frequency of an HCV haplotype in the quasispecies depends on its replication efficacy and other known and unknown viral and host factors ([6], [7]). As a consequence of these changes, multiple variants are produced and some may be clinically relevant, with effects on pathogenesis, reduced susceptibility to antiviral therapy, vaccination failure, escape from the immune response, and lack of protection for reinfection. Thus, because of their inherent nature, viruses cannot be studied simply as a single sequence; they must be viewed as a population of sequences. Pioneering studies with the population approach (still using time-consuming, costly, low-throughput techniques) led to important conclusions for managing patients, such as the use of combination therapy to succeed in the treatment of viral infections in humans [8]. These groundbreaking studies on viral quasispecies [9] led to

a demand for high-throughput sequencing methods, as the complex character that was being uncovered could not be encompassed with the available means. The viral particles present in the body of a patient with acute hepatitis infection may outnumber the total human population.

In October 2005, 454 Life Sciences, a member of the Roche group, was the first company to announce development of a powerful sequencing method, called next-generation sequencing (NGS). The company launched the first commercial instrument, the 454 GS20, in 2007. Later, in 2008, the 454 GS-FLX Titanium series was released, with the ability to sequence 400 to 600 million base-pairs (bp) with 400 to 500 bp read lengths. Because of the high accuracy, low cost, and long reads provided by these techniques, many researchers migrated away from traditional Sanger capillary sequencing instruments to NGS platforms.

Nonetheless, NGS posed challenges for quasispecies researchers. Although coverage was much higher than ever before – the 50 clones were suddenly nothing compared to what the new technology offered – there was a cost. The length of the sequenced fragments was shorter than that obtained by Sanger sequencing, the haplotypes the clones provided were not available, and the software required to process the data obtained from the new instruments was inchoate and limited.

In 2010, the Liver Diseases Laboratory at Vall d'Hebron Research Institute in Barcelona, at the time specialized in hepatitis viruses (HBV, HCV, HDV) ([5],[10],[11]), began sequencing clinical samples with a 454 GS-FLX instrument, based on ultra-deep pyrosequencing ([12],[13],[14]). An important problem that had to be resolved was how to deal with GS-FLX data to obtain amplicon haplotypes and estimate their frequencies in the viral quasispecies. Software available to the bioinformatics community at that time had been developed mainly to detect genetic substitutions, mutations and indels, or to estimate differential gene expression in RNA-seq data. The available filters for sequencing errors were based on trimming nucleotides with low sequencing scores, which prevents identification of haplotypes.

Our laboratory then initiated a project in collaboration with Roche Diagnostics, Spain, to develop software that could provide amplicon haplotypes and their frequencies. With

this main objective, a set of clones and clone mixes was sequenced, and the data were examined and analyzed to find a method to obtain amplicon haplotypes that compared well with the original sequences and in the proper proportions. The results were published in two articles: the first with HBV data [15] and the second with HCV data [16]. The method was based on simple, sound principles:

- Respect read integrity. No trimming except for the primers.
- Collapse identical reads to haplotypes and frequencies.
- Reject all haplotypes with a single read and those having more than 2 gaps, 3 Ns, or 99 differences with respect to the reference sequence or with respect to the master sequence.
- Remove all haplotypes with a frequency below a threshold established by comparing with the clones (0.25% or 0.5%).
- Remove all haplotypes not common to both strands (forward and reverse).
- Compute relative frequencies from final reads and haplotypes.

Even with the limited length of the sequenced fragments and the requirement of relatively high viral loads, the GS-FLX and GS-Junior instruments represented a true window into the world of viral quasispecies. The results obtained provided explanations for phenomena such as selection of variants resistant to antiviral treatment in both HBV and HCV, which had only been theoretical hypotheses when investigated with classical methods. It was seen that variants responsible for treatment failure could be present in the viral quasispecies at very low frequencies that were undetectable by molecular cloning.

With Roche's discontinuation of the 454 platform at the end of 2016, our laboratory began sequencing with MiSeq™ Illumina® instruments. The requirement was to obtain amplicon haplotypes in the range of 300 to 550 bp, as was the case with the preceding technology. Again, with the help of clones, the sequencing data analysis was adjusted, respecting the original principles.

- Sequence amplicons using 2x300 bp paired-end reads.
- Obtain amplicon integrity with the help of Flash [17], requiring a minimum overlap of 20 bp between paired-ends, with a maximum of 10% differences.
- Reject all reads in which more than 5% of bases have a

Phred score below 30 (equivalent to 0.001 probability of error).
- Trim primers and collapse identical reads to haplotypes and frequencies.
- Eventually remove all haplotypes with a frequency below a given threshold.
- Remove all haplotypes not common to both strands (forward and reverse).
- Compute relative frequencies from final reads and haplotypes.

Currently, we are only able to obtain high quality amplicon haplotypes of slightly more than 500 bp in size, with coverage in the order of 10^5 reads per amplicon, when sequencing with Illumina instruments. Despite this limitation, quasispecies genomes can be studied amplicon by amplicon. However, when monitoring direct-acting antiviral treatments that target a specific region of the genome, a single amplicon may suffice. Various approaches can be considered, for example, there are a number of inferential methods to estimate full viral haplotypes by reconstructing them from short reads. A recent review has evaluated 12 such methods [18] and some limitations are reported: special computational resources are required for high coverage, poor performance with samples having high genetic diversity, and underestimation of the number of haplotypes.

All the techniques described in the articles collected here use quasispecies haplotypes and their frequencies as the starting point. It is of no concern whether they are amplicon haplotypes of whatever size or full viral haplotypes, and the methods can be used with any sequencing platform. To obtain a comprehensive picture of an infection (viral load $>10^6$ copies/mL of blood), the only requirement is to have a set of high-quality haplotypes with their frequencies, and coverage of more than 10^4 reads (preferably more than 10^5) fully spanning the region of interest.

REFERENCES

1. Molecular cloning (Wikipedia). https://en.wikipedia.org/wiki/Molecular_cloning
2. Domingo E, Sabo D, Taniguchi T, Weissmann C. Nucleotide sequence heterogeneity of an RNA phage population. Cell 1978; 13(4):735-44. https://doi.org/10.1016/0092-8674(78)90223-4
3. Kobayashi Y, Kawamura F. Molecular cloning. Biotechnology 1992; 22:123-41.
4. Sharma K, Mishra AK, Mehraj V, Duraisamy GS. Advances and applications of molecular cloning in clinical microbiology. Biotechnol Genet Eng Rev 2014 Oct; 30(1-2):65-78. https://doi.org/10.1080/02648725.2014.921501
5. Martell M, Esteban JI, Quer J, Genesca J, Weiner A, et al. Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. J Virol 1992; 66: 3225-9. https://doi.org/10.1128/JVI.66.5.3225-3229.1992
6. Duarte EA, Novella IS, Weaver SC, et al. RNA virus quasispecies: significance for viral disease and epidemiology. Infect Agents Dis 1994; 3(4):201-14.
7. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 2006; 439(7074):344-8. https://doi.org/10.1038/nature04388
8. Domingo E. RNA virus evolution and the control of viral disease. Prog Drug Res 1989; 33:93-133. https://doi.org/10.1007/978-3-0348-9146-2_5
9. Holland J, Spindler K, Horodyski F, Grabau E, Nichol S, VandePol S. Rapid evolution of RNA genomes. Science 1982; 215(4540):1577-85. https://doi.org/10.1126/science.7041255
10. Quer J, Murillo P, Martell M, Gómez J, Esteban JI, Esteban R, Guardia J. Subtype mutations in the envelope 2 region including phosphorylation homology domain of hepatitis C virus do not predict effectiveness of antiviral therapy. J Viral Hepat 2004; 11(1):45-54. https://doi.org/10.1046/j.1352-0504.2003.00465.x
11. Quer J, Esteban JI, Cos J, et al. Effect of bottlenecking on evolution of the nonstructural protein 3 gene of hepatitis C virus during sexually transmitted acute resolving infection. J Virol 2005; 79(24):15131-41. https://doi.org/10.1128/JVI.79.24.15131-15141.2005
12. Rothberg J, Leamon J. The development and impact of 454 sequencing. Nat Biotechnol 2008; 26:1117-24. https://doi.org/10.1038/nbt1485
13. Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, et al. Accuracy and quality assessment of 454 GS-FLX titanium pyrosequencing. BMC Genomics 2011; 12: 245. https://doi.org/10.1186/1471-2164-12-245
14. Homs M, Buti M, Quer J, et al. 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 2011; 39(19):8457-71. https://doi.org/10.1093/nar/gkr451
15. Ramírez C, Gregori J, Buti M, et al. 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 2013; 98(2):273-83. https://doi.org/10.1016/j.antiviral.2013.03.007
16. Gregori J, Esteban JI, Cubero M, et al. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One 2013; 8(12):e83361. https://doi.org/10.1371/journal.pone.0083361
17. Magoc T, Salzberg SL. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics 2011; 27:2957-63. https://doi.org/10.1093/bioinformatics/btr507
18. Eliseev A, Gibson KM, Avdeyev P, et al. Evaluation of haplotype callers for next generation sequencing of viruses. Infect Genet Evol 2020; 82:104277. https://doi.org/10.1016/j.meegid.2020.104277

Breakthrough articles

The articles collected in this book are listed here with a brief comment and reference to the section devoted to each of them.

Inference with viral quasispecies diversity indices: clonal and NGS approaches
Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J.
Bioinformatics **2014** Apr 15; 30(8):1104-1111.
https://doi.org/10.1093/bioinformatics/btt768

Research investigating the statistical properties of two basic diversity indices applied to quasispecies study: Shannon entropy and mutation frequency. Comparison of two analytical scenarios, classical cloning followed by Sanger sequencing (CCSS) and next-generation sequencing (NGS). The study aims to provide a means to enable comparisons between quasispecies, controlling bias and sample size dependence, and proposes methods for t-tests in both cases. See Section 1.

Viral quasispecies complexity measures
Gregori J, Perales C, Rodríguez-Frías F, Esteban JI, Quer J, Domingo E.
Virology **2016** Jun; 493:227-37.
https://doi.org/10.1016/j.virol.2016.03.017

Provides a systematized presentation of diversity indices, some of which are taken from the field of biodiversity, to clarify the type of information provided by each index. Diversity profiles are introduced as visual tools to characterize quasispecies. Recommendations are given to select appropriate diversity indices for different settings. See Section 2.

Quasispecies complexity computations: a tutorial
Gregori J, Quer J, Rodríguez-Frías F.
In: Rizzetto M, Smedile A, eds. Hepatitis D. Virology, management and methodology. Rome: Il Pensiero Scientifico Editore, **2019**.

Shows step-by-step computations of the various diversity indices using a simple dataset representing a quasispecies. Illustrates the use of multiple profile plots as a means to visualize several aspects of quasispecies composition. See Section 3.

Rare haplotype load as marker for lethal mutagenesis
Gregori J, Soria ME, Gallego I, Guerrero-Murillo M, Esteban JI, Quer J, Perales C, Domingo E.
PLoS One **2018** Oct 3; 13(10):e0204877.
https://doi.org/10.1371/journal.pone.0204877

Introduces a diversity index specifically designed to quantify mutagenic effects. The statistical properties of the index and correlations between this and other diversity indices are examined. See Section 4.

Quasispecies fitness partition to characterize the molecular status of a viral population. Negative effect of early ribavirin discontinuation in a chronically infected HEV patient
Gregori J, Colomer-Castell S, Campos C, et al.
Int J Mol Sci **2022**; 23:14654.
https://doi.org/10.3390/ijms232314654

Study of a complex clinical case to illustrate a proposed new visual summary of quasispecies composition in the form of a quasispecies fitness partition into four fractions with biological meaning related to quasispecies evolution. See Section 5, with minor corrections to the original.

Quantifying in-host quasispecies evolution
Gregori J, Ibañez-Lligoña M, Quer J.
Int J Mol Sci **2023**; 24(2):1301.
https://doi.org/10.3390/ijms24021301

Borrowed from the ecology or biodiversity field, the similarity or distance between haplotype distributions is proposed as an alternative visual tool to monitor quasispecies evolution. See Section 6.

Section 1

Diversity indices, bias and sample size. NGS vs CCSS

Abstract

Given the inherent dynamics of viral quasispecies, it may be of interest to compare quasispecies diversity indices between sequential samples from a single patient over the course of an infection, or between patient groups in a treated versus control design. Hence, it is important to ensure that the viral diversity measures from each sample can be compared with no bias and within a consistent statistical framework. In the present report, we review various indices used as measures of viral quasispecies complexity and provide the means for statistical inference with them, applying procedures taken from the field of ecology. In particular, we examine the concepts of Shannon entropy and mutation frequency, and we discuss the appropriate use of several normalization methods for Shannon entropy reported in the literature. By taking raw data from amplicons obtained by ultradeep pyrosequencing (UDPS) as a surrogate of a real hepatitis C viral population, we used in-silico sampling to study the statistical properties of these indices under two methods of quasispecies analysis, classical cloning followed by Sanger sequencing (CCSS) and next-generation sequencing (NGS) such as UDPS. We propose specific solutions for each of these methods to guarantee statistically conforming conclusions as free from bias as possible.

Conclusions

In this article, we empirically study the statistical properties of Shannon entropy, normalized Shannon entropy and mutation frequency while observing viral quasispecies complexity by CCSS or NGS, and thereby, assess the means to achieve less biased comparisons of complexity indices. These methods will enable us to statistically conclude whether a viral quasispecies is expanding or diminishing in diversity, regardless of the size of the samples being compared. In the Supplementary Material we provide the formulation, and in Boxes 1 and 2 we propose data treatment methods for inference for CCSS and for NGS.

Highlights

- The Shannon entropy equation is a biased estimator, and the bias is dependent on sample size. This bias can be partially corrected.
- The mutation frequency is an unbiased estimator, moderately sensitive to the sample size.
- The minimum differential bias is provided by repeated resampling to the minimum size.

- Fringe trimming is a good alternative for mutation frequency, and may be considered an approximation for the Shannon entropy.
- Z-test and t-test methods are formulated.

Note

This study was influenced by two previous articles, both written in 2013, involving sequencing of HBV [1] and HCV [2] clones. In these studies, an error threshold was set to exclude haplotypes with technical errors from the analysis of NGS data. As a result, the diversity indices were computed on the haplotypes and frequencies obtained after passing an abundance filter set at a minimum of 0.5%. However, this filter had a side effect, in which small samples showed higher diversity than larger ones, sampled from the same population. This was corrected by the fringe-trimming method.

The next two articles in this book (Sections 2 and 3), continue to use the 0.5% filter followed by the correction. Nevertheless, the study of mutagenesis in HCV-infected cell lines (Section 4) and the later study analyzing samples from a chronically infected HEV patient (Section 5) showed that the level of information lost as a result of this abundance filter was too important to be accepted. At low abundance levels, both real and artefactual haplotypes coexist, and they might not be differentiated. This led us to recommend avoiding abundance filters. Instead, we suggest the use of balanced experimental designs to compensate for any errors in the conditions to be compared, and repeated resampling to the smallest size for sample size correction (Concluding remarks).

REFERENCES

1. Ramírez C, Gregori J, Buti M, et al. 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 2013; 98(2):273-83. https://doi.org/10.1016/j.antiviral.2013.03.007
2. Gregori J, Esteban JI, Cubero M, et al. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One 2013; 8(12):e83361. https://doi.org/10.1371/journal.pone.0083361

1. Inference with viral quasispecies diversity indices: clonal and NGS approaches

JOSEP GREGORI, MIQUEL SALICRÚ, ESTEBAN DOMINGO, ALEX SANCHEZ, JUAN I. ESTEBAN, FRANCISCO RODRÍGUEZ-FRÍAS, JOSEP QUER

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.

Key words

CCSS, NGS, Shannon entropy, mutation frequency, nucleotide diversity, sample size bias, rarefaction, fringe trimming, statistical tests.

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 num-

ber 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.

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; Salicrú 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 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 extent 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 (Ramírez 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 ultradeep 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 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 quasispecies, and the observed distribution of haplotypes may be used as a viral population reference distribution from which to sample viral particles.

- 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.
- 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.
- 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 1) is normalized to the natural logarithm of the number of estimated species in the population (Supplementary Equation 7) 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 on 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, Ramírez 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., 2014). 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.

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

Population	Reads	Haplotypes	Polymorphic sites	Sn	Mf	Pi
Low	42,436	496	300	0.2194	5.089E-04	1.012E-03
Mid	43,300	550	269	0.2562	1.449E-03	2.502E-03
High	52,250	2064	266	0.5705	1.198E-02	1.585E-02

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 1A and 1B 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 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 2A. The corresponding boxplots are shown in Figure 2. *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.

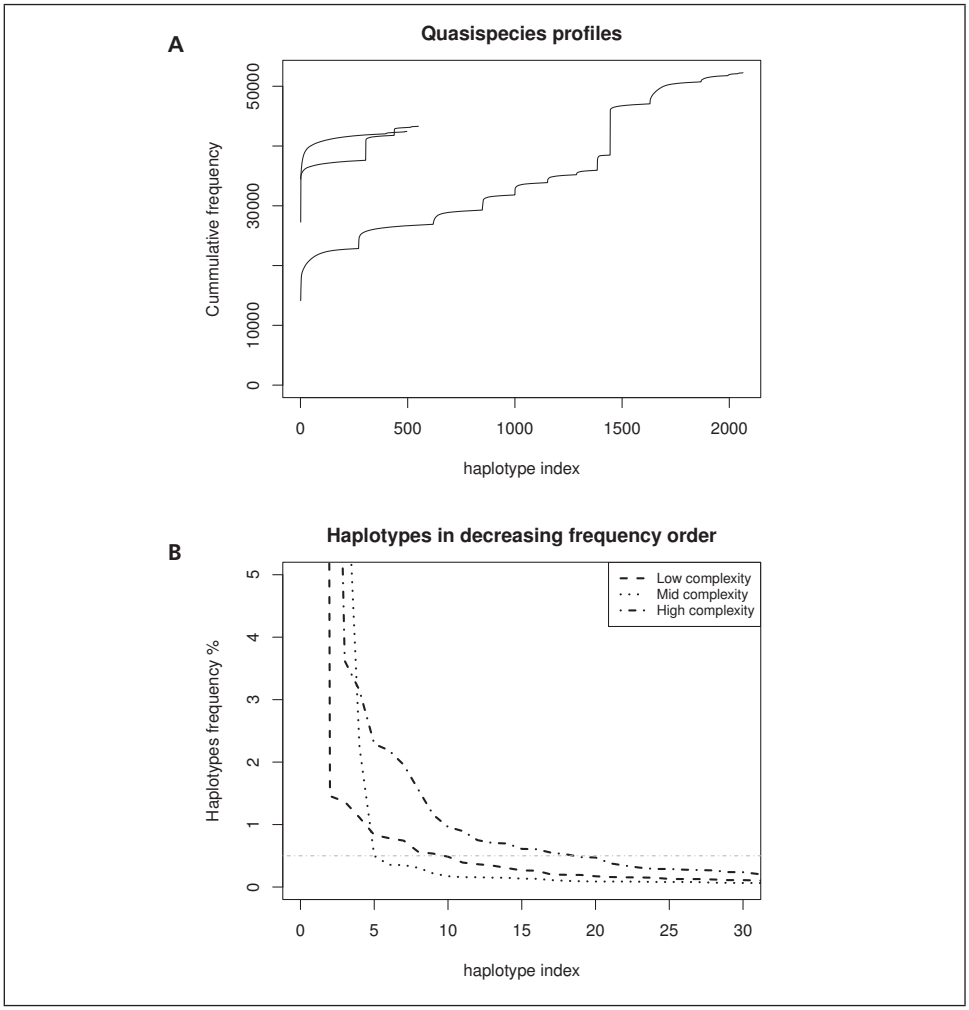


Fig. 1. (**A**) Quasiespecies profile as a cumulated distribution of the three reference populations used in the study. In abscissa the haplotypes are ordered primarily by the Hamming distance to the most frequent haplotype, and ties are determined in descending order of frequency. The first haplotype is the dominant one, while the last is the one showing more differences with 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.

When applying the bias corrections of Hutcheson (Supplementary Equation 2) (Hutcheson, 1970) and Chao 1 (Supplementary Equation 3) (Chao et al., 2009) the bias is partially corrected but remains sample size dependent (Supplementary Table 2B and Supplementary Fig. 1. While applying the rarefaction of the samples of size 50

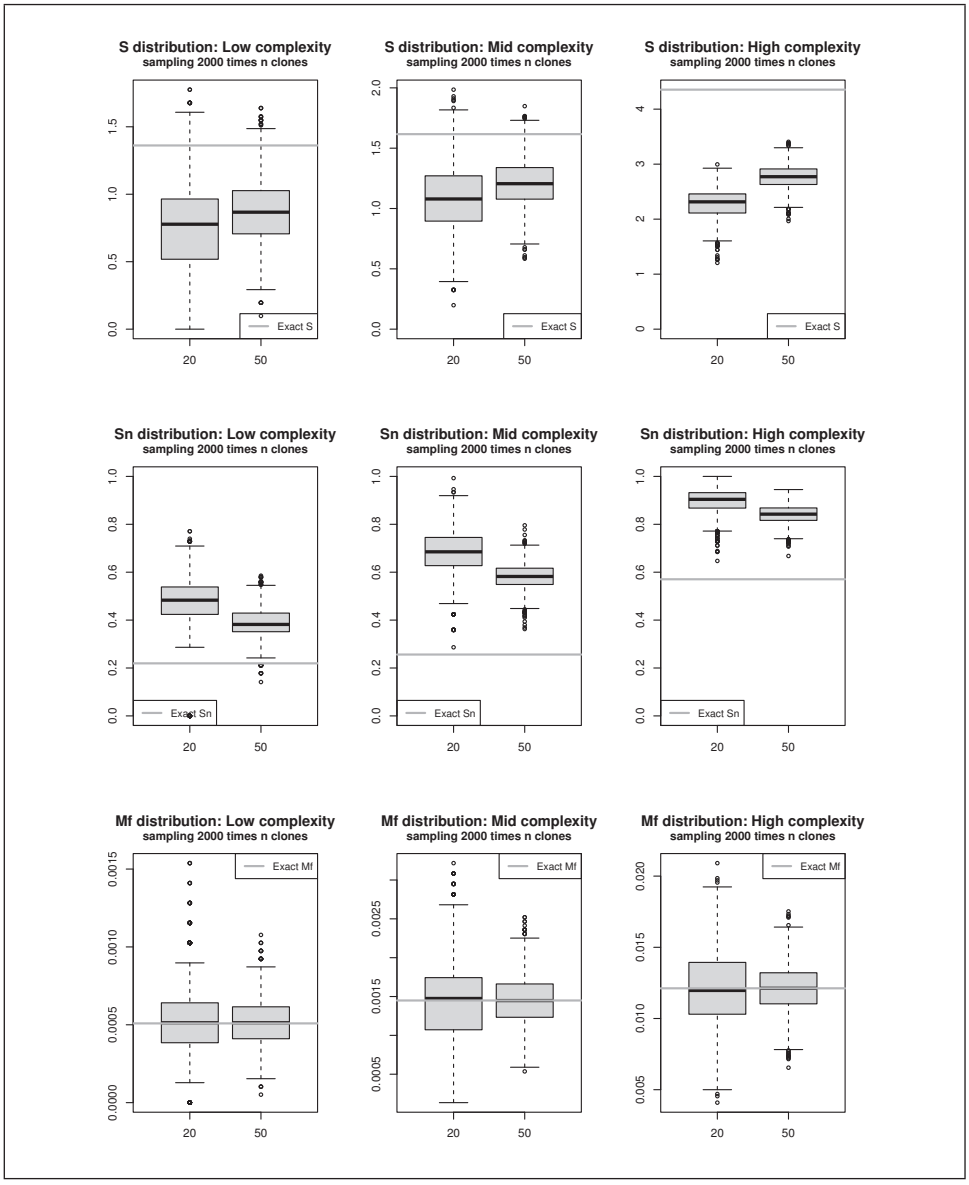


Figure 2. Boxplots with the distribution of the observed values of *S*, *Sn* and *Mf* in 2000 replicates CCSS experiments of size 20 and 50, for each of the three viral populations.

to size 20, the median values of both samples are brought to the same level (Supplementary Table 2C and Supplementary Fig. 2) 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 12) (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 2D). 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. 3). The bias correction of Hutcheson (Supplementary Equation 2) 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 4A 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 4B, where we show a barplot with the probabilities to observe a 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, 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 4A and C. The bias of *S*, *Sn* and *Mf*

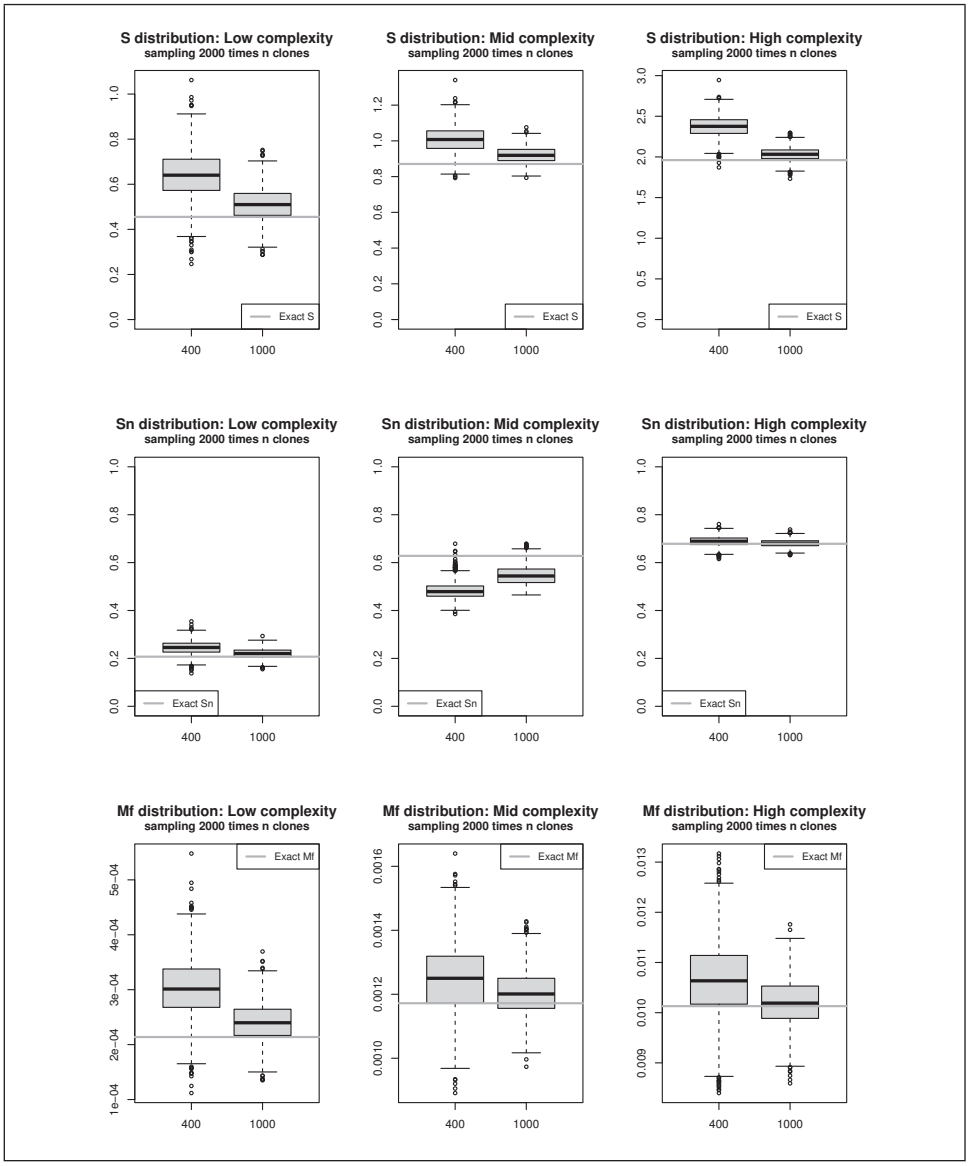


Figure 3. Boxplots with the distribution of the observed values of *S*, *Sn* and *Mf* in 2000 replicates of NGS experiments of size 400 and 1000, filtering at a noise level of 0.5%, for each of the three viral populations.

is also greatly reduced (Supplementary Fig. 3). 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

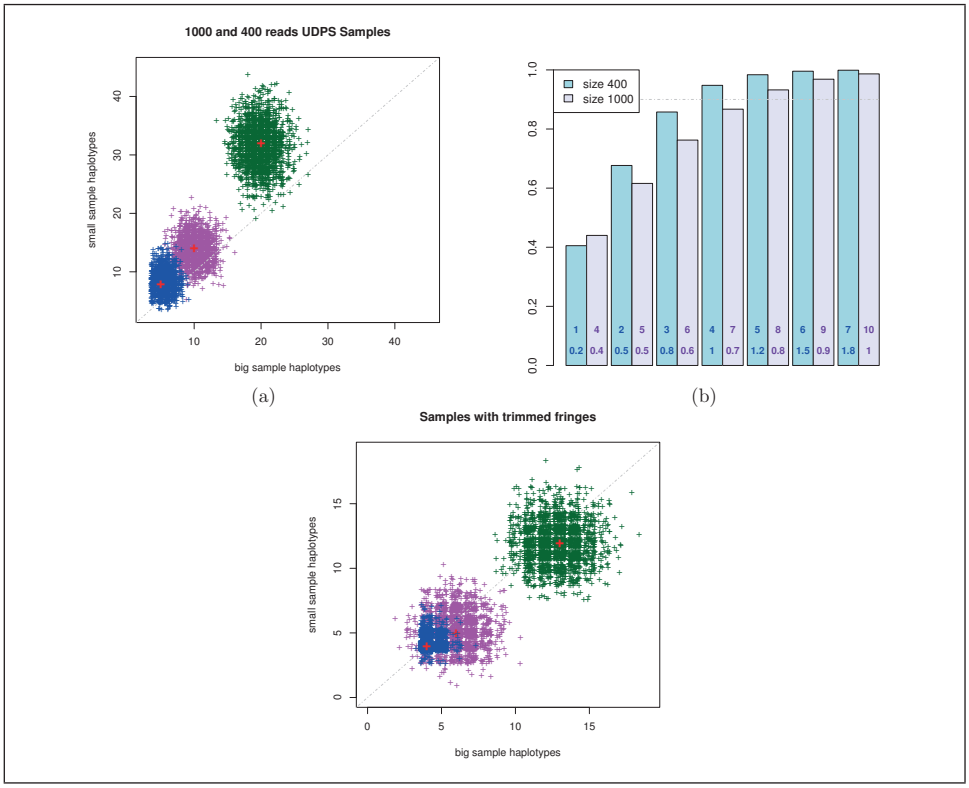


Fig. 4. (**A**) Scatterplot with the number of observed haplotypes in pairs of samples of size 400 and 1000 after filtering the haplotypes below the noise level. The clouds are biased to higher values for the small samples. (**B**) Plots with the cumulated probabilities to observe a haplotype with an 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 5), *Sn* (Supplementary Equation, 5 with 7 and 8) or *Mf* (Supplementary Equation 11).

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 Material 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 lin-

ear 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 Material 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. 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 (Ramírez 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 1 A and B), which means that in CCSS experiments, on this kind of viral populations, roughly one out 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 2 and 3). 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.

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 Chao 1 (Supplementary Equation 3), 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 2), 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 12) 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 2), 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 11) and compute the CI.

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.

Supplementary information

Supplementary data are available at *Bioinformatics* online at https://doi.org/10.1093/bioinformatics/btt768

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 samplebased rarefaction, extrapolation and comparison of assemblages. J Plant Ecol 5, 3-21.

Cubero M. et al. (2014). 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 UK 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. Oxford, UK: Blackwell Publishing.

Magurran A. and McGill B.J., eds. (2010). Biological diversity: frontiers in measurement and assessment. Oxford, UK: Oxford University Press.

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. New York, NY: Columbia University Press.

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. Vienna, Austria: Foundation for Statistical Computing.

Ramírez 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.

Salicrú 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.

Supplementary Material

Glossary

Quasispecies terms [1]

Complexity of a mutant spectrum: number of mutations and genomic sequences in a viral population. It is often quantified by the mutation frequency and the Shannon entropy.

Consensus sequence: in a set of aligned nucleotide or amino acid sequences, the one that results from taking the most common residue at each position.

Genetic distance: usually the Hamming distance between pairs of sequences. Different evolutive models may be used to introduce corrections to the observed number of differences.

Master or dominant sequence: the genomic nucleotide sequence that dominates a mutant spectrum because of its superior fitness. It may or may not be identical to the consensus sequence. The most abundant genome may still be a minority relative to the ensemble of low frequency variants. Owing to the abundance of quasineutral mutations and epistatic interactions in viral genomes, there might be a large ensemble of sequences of almost identical fitness that compose a "master phenotype".

Mutation frequency: the proportion of mutated sites in a population of viral genomes with respect to the dominant haplotype, or to the consensus sequence or to a given reference.

Mutation rate: the frequency of occurrence of a mutation during viral genome replication.

Mutant spectrum: the ensemble of mutant genomes that compose a viral quasispecies. It is also termed mutant swarm or mutant cloud.

Rate of evolution: the frequency of mutations that become dominant (i.e., are represented in the consensus sequence) as a function of time. It may refer to evolution within a host individual or upon epidemic expansion of a virus.

Viral quasispecies: a set of viral genomes that belongs to a replicative unit and subjected to genetic variation, competition, and selection, and which acts as a unit of selection. It has been extended to mean ensembles of similar viral genomes generated by a mutation–selection process.

Biodiversity terms

Complexity: any index or a set of indices quantifying the variability of a viral population in a wide sense, including richness, diversity and heterogeneity.

Diversity index: a measure of compositional complexity expressing the degree of variation of forms in a community. A function of the frequencies of the different species (haplotypes), usually given by an entropy expression [2]. *S* and the Gini-Simpson index are examples of diversity indices.

Effective number of species: number of equally common species, estimated by the Hill numbers of different order ([3], [4]).

Evenness index: the ratio of a diversity index *I* to I_{max}. Where I_{max} is the value that *I* would take if the abundances in the sample were equal. *Sn* is an example of evenness index [5].

Heterogeneity: a measure of diversity taking into account differences among individuals. It is a function of the number of haplotypes, their frequencies, and their differences. The mutation frequency and the nucleotide diversity are examples of heterogeneity measures.

Hill numbers: a function transforming a diversity index into an effective number of species. The richness is the Hill number of order 0. The exponential of the Shannon entropy is the Hill number of order 1, the inverse of the Gini-Simpson index is the Hill number of order 2 ([5], [4]).

Phylogenetic diversity: a measure of heterogeneity using the branches length in a phylogenetic tree as distances [6].

Richness: number of species in a community [2]. In the context of a quasispecies the number of haplotypes, that is, different genomes found in the population. The number of polymorphic sites and the number of mutations may be considered richness indices as well.

True diversity: see effective number of species [3].

Diversity indices and inference

Shannon entropy

The Shannon entropy (*S*) was originally developed in the domain of information exchange [7], and is related to the transmission capacity of a communication channel. It measures the average unpredictability or lack of information contained in a set of items in terms of its alphabet. It found soon its place in ecology [8] and later in virology [9]. In genetics S is used with two different approaches. The first is an analysis of diversity of each sequence position – columns in a multiple alignment – either of nucleotide or amino acid sequences, where the alphabet size – either 4 or 20 – is known [9]. Alternatively, the analysis may be by genomes or haplotypes – rows of the multiple alignment – where the alphabet size is unknown [10], as is the case in ecology. In this work we use the second approach as a measure of the global quasispecies complexity. In this context it is a function of the number of haplotypes in the viral population and their relative frequencies, and its maximum likelihood estimator (MLE) is:

$$S_{MLE} = -\sum_{i=1}^{h} \hat{p}_i \, log(\hat{p}_i) = -\sum_{i=1}^{h} \frac{n_i}{N} log\left(\frac{n_i}{N}\right); \quad N = \sum_{i=1}^{h} n_i \tag{1}$$

with p_i the MLE of the relative frequency of each haplotype, n_i the observed counts of the i-th haplotype, *h* the number of observed haplotypes, and *N* the sample size.

It is known that the Shannon entropy as a function of proportion estimates is negatively biased [11], and Hutcheson [8] provided an approximation to the bias by applying a Taylor series expansion

$$S_{MLE} = S - \frac{H-1}{2N} + \frac{1-\sum \hat{p}_i{}^{-1}}{12N^2} + \frac{\sum(\hat{p}_i{}^{-1} - \hat{p}_i{}^{-2})}{12N^3} + ... \tag{2}$$

were *S* is the exact value, and *H* is the estimate of the number of haplotypes in the population. *H* is also negatively biased and may be corrected, among others, by the Chao 1 method [12]

$$H_{Chao} = h + \frac{f_1(f_1-1)}{2(f_2+1)} \tag{3}$$

where *h* is the number of haplotypes observed in the sample, and *f1* and *f2* are the number of singletons (haplotypes with a single copy) and doubletons (haplotypes with two copies) in the sample. The delta method and the asymptotic normality of the multinomial distribution to the normal provide the means for statistic inference as established by Hutcheson [8] for the Shannon entropy. Salicrú and cols. [13], [14] provided an elegant generalization for inference to a wide fan of entropy indices. The estimated variance of the Shannon entropy is given by:

$$\hat{\sigma}^2_S = \sum_{i=1}^{h} \hat{p}_i \, log(\hat{p}_i)^2 - \left[\sum_{i=1}^{h} \hat{p}_i \, log(\hat{p}_i)\right]^2 \tag{4}$$

with this estimated variance we may compute confidence intervals or test the equality of Shannon entropies between two samples by statistical inference using the *Z*-test

$$P_1 = [p_1^{(1)}, ..., p_{h1}^{(1)}] \qquad P_2 = [p_1^{(2)}, ..., p_{h2}^{(2)}]$$

$$H_0 : S(P_1) = S(P_2) \qquad H_1 : S(P_1) \neq S(P_2)$$

$$Z = \frac{S(P_1) - S(P_2)}{\sqrt{\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2}}; \qquad Z \sim N(0,1) \tag{5}$$

Where P_1 and P_2 are the vectors of observed haplotype frequencies of the two samples to compare, H_o is the null hypothesis H_1 is the alternative hypothesis, and Z is a statistic asymptotically distributed as the standard normal, with N_1 and N_2 the sample sizes.

The normalized Shannon entropy

In the ecology literature the Shannon entropy is normalized to the natural logarithm of the number of estimated species – the size of the alphabet – 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 $Sn = 0$. The proof is as follows:

$$-\sum_{i=1}^{h} \frac{1}{h} log(\frac{1}{h}) \; = \; -\frac{h}{h} log(\frac{1}{h}) \; = \; log(h) \tag{6}$$

where h is the number of species, and $1/h$ is the frequency of each species in the population.

$$Sn \; = \; \frac{S}{log(h)} \tag{7}$$

Then Sn varies from 0 to 1 and is a measure of evenness of the population.By Taylor series expansion we find the estimated variance of Sn, assuming h as a constant ([13], [14]), as:

$$\hat{\sigma}^2_{Sn} \; = \; \left(\frac{1}{log(h)}\right)^2 \left[\sum_{i=1}^{h} \hat{p}_i \, log(\hat{p}_i)^2 \; - \; \left(\sum_{i=1}^{h} \hat{p}_i \, log(\hat{p}_i)\right)^2\right] \tag{8}$$

Sn is also asymptotically normal and the same inference as for S may be used. Sn is not sensitive to heterogeneity, that is i.e. a viral population constituted of two haplotypes at 50%, with just one substitution between them, has the same Sn than a population with 100 haplotypes at 1% each, and a mean of 10 differences among them.

The mutation frequency

Mf is a heterogeneity measure that takes as reference either the most represented haplotype, also known as dominant or master sequence [15], or the consensus sequence [16], and computes the number of observed differences of each individual genome with respect to this reference. The value is normalized to the total number of nucleotides sequenced. The higher its value the more dissimilar are the individuals in the population with respect to the reference.

$$Mf \; = \; \frac{1}{lN} \sum_{i=1}^{h} n_i \, m_{1i} \; = \; \frac{1}{l} \sum_{i=1}^{h} \hat{p}_i \, m_{1i} \; = \; \frac{1}{l} < \hat{P} \mid M_1 > \tag{9}$$

where l is the amplicon sequence length, N the sample size (number of sequences), n_i the observed counts of the i-th haplotype, m_{1i} the number of substitutions between the i-th haplotype and the master sequence, which without loss of generality is taken as the first, and the vector bracket notation has been used in the term of the right.

By the delta method and the asymptotic normality of a multinomial we find the variance of Mf which may be used to obtain confidence intervals, or to compare the mutation frequencies of two samples by the help of the Z-test.

$$\hat{\sigma}^2_{Mf} = \frac{1}{l^2}\left(\sum_{i=1}^{h}\hat{p}_i\, m^2_{1i} - \left[\sum_{i=1}^{h}\hat{p}_i\, m_{1i}\right]^2\right) \tag{10}$$

$$P_1 = [p^{(1)}_1, ..., p^{(1)}_{h1}] \qquad M_1 = [0, m^{(1)}_{12}, ..., m^{(1)}_{1h1}]$$

$$P_2 = [p^{(2)}_1, ..., p^{(2)}_{h2}] \qquad M_2 = [0, m^{(2)}_{12}, ..., m^{(2)}_{1h2}]$$

$$H_0 : Mf(P_1, M_1) = Mf(P_2, M_2) \qquad H_1 : Mf(P_1, M_1) \neq Mf(P_2, M_2)$$

$$Z = \frac{Mf(P_1, M_1) - Mf(P_2, M_2)}{\sqrt{\hat{\sigma}^2_1/N_1 + \hat{\sigma}^2_2/N_2}}; \quad Z \sim N(0, 1) \tag{11}$$

Where P_1 and P_2 are the vectors of observed haplotype frequencies of the two samples to compare, M_1 and M_2 the vectors of Hamming distances of each haplotype with respect to the reference in the respective sample, H_0 is the null hypothesis, H_1 the alternative hypothesis, and Z a statistic asymptotically distributed as the standard normal.

The t-test

When the sample size is mall, as in the CCSS case, the Welch *t*-test [8] should be preferred to the *Z*-test.

$$H_0 : X_1 = X_2 \qquad H_1 : X_1 \neq X_2$$

$$T = \frac{X_1 - X_2}{\sqrt{Var(X_1) + Var(X_2)}}; \quad T \sim t(dof)$$

$$Var(X) = \sigma^2_X/N$$

$$dof = [Var(X_1) + Var(X_2)]/\left[\frac{Var^2(X_1)}{N_1} + \frac{Var^2(X_2)}{N_2}\right] \tag{12}$$

where *X* applies both to *S*, *Sn* and *Mf*, and σ_x is given by 4, 8 or 10.

The nucleotide diversity

The nucleotide diversity [17] considers the differences between each pair of genomes in the population and is a more general measure of heterogeneity than the mutation frequency.

$$\Pi = \sum_{i=1}^{h}\sum_{j=1}^{h}\hat{p}_i\, d_{ij}\, \hat{p}_j = < \hat{P} \mid D \mid \hat{P} > \tag{13}$$

Where d_{ij} is the genetic distance between the i-th and the j-th haplotype, $\hat{p}_i$ and $\hat{p}_j$ are the estimated proportions of the i-th and j-th haplotypes in the quasispecies. The MLE estimator is biased, and the bias is a function of the sample size.

$$E[\hat{\Pi}_{MLE}] - \hat{\Pi}_{MLE} = \frac{1}{N}\hat{\Pi}_{MLE} \tag{14}$$

The variance of π may be found in Nei 1987 [17]. As a quadratic form this index is asymptotically distributed as a sum of a normal and a linear combination of Chi-square distributions [18], and the *Z* or the *t*-test are of no application. A resampling test is a good choice in this case.

Supplementary tables

Table 1A. Observed population complexity at different global haplotype abundance cut-offs.

Complexity	Cut-off	Reads	Excluded	Min reads	Haplotypes	Poly sites
Low	All	42436	0.0%	2	496	300
	0.10%	39608	6.7%	42	32	32
	0.25%	38628	9.0%	106	16	15
	0.50%	37602	11.4%	212	9	8
	1.00%	36136	14.8%	424	4	3
Mid	All	43300	0.0%	2	550	269
	0.10%	40450	6.6%	43	17	15
	0.25%	39853	8.0%	108	8	6
	0.50%	39198	9.5%	216	4	3
	1.00%	39198	9.5%	433	4	3
High	All	52250	0.0%	2	2064	266
	0.10%	39006	25.3%	52	80	38
	0.25%	35088	32.8%	130	28	19
	0.50%	33324	36.2%	261	18	16
	1.00%	30032	42.5%	522	9	12

Excluded: % of excluded reads in the filter; Min reads: minimum number of reads per haplotype; Poly sites: number of polymorphic sites.

Table 1B. Observed population complexity at different global haplotypes abundance cut-offs.

Complexity	Cut-off	Reads	Excluded	Mf	S	Mean diff.	π
Low	All	42436	0.0%	5.089E-04	0.2194	0.39	1.012E-03
	0.10%	39608	6.7%	3.417E-04	0.2267	0.26	6.782E-04
	0.25%	38628	9.0%	2.763E-04	0.2225	0.21	5.473E-04
	0.50%	37602	11.4%	2.138E-04	0.2072	0.16	4.227E-04
	1.00%	36136	14.8%	1.185E-04	0.1715	0.09	2.333E-04
Mid	All	43300	0.0%	1.449E-03	0.2562	0.93	2.502E-03
	0.10%	40450	6.6%	1.257E-03	0.3733	0.79	2.126E-03
	0.25%	39853	8.0%	1.212E-03	0.4630	0.76	2.032E-03
	0.50%	39198	9.5%	1.172E-03	0.6281	0.73	1.958E-03
	1.00%	39198	9.5%	1.172E-03	0.6281	0.73	1.958E-03
High	All	52250	0.0%	1.198E-02	0.5705	5.23	1.585E-02
	0.10%	39006	25.3%	1.067E-02	0.6112	4.60	1.394E-02
	0.25%	35088	32.8%	1.027E-02	0.6531	4.42	1.339E-02
	0.50%	33324	36.2%	1.013E-02	0.6785	4.34	1.316E-02
	1.00%	30032	42.5%	9.964E-03	0.7189	4.27	1.295E-02

Excluded: % of excluded reads in the filter; Mf: mutation frequency; S: Shannon entropy; Mean diff.: mean nucleotide differences between pairs of haplotypes; π: nucleotide diversity.

REFERENCES

1. Perales C. et al. The impact of quasispecies dynamics on the use of therapeutics. Trends Microbiol 2012; 20:595-603.
2. Magurran A. Measuring biological diversity. Oxford, UK: Blackwell Publishing, 2004
3. Jost L. Entropy and diversity. Oikos 2006; 113:363-375.
4. Chao A., Jost L. Diversity measures. In: Hasting A, Gross L. (eds.). Encyclopedia of Theoretical Ecology. Berkeley: University of California Press, 2012.
5. Hill M. Diversity and evenness: a unifying notation and its consequences. Ecology 1973; 54:427-432.
6. Chao A. et al. Phylogenetic diversity measures based on Hill numbers. Philos Trans R Soc Lond B Biol Sci 2010; 365:3599-3609.
7. Shannon C. A mathematical theory of communication. Bell System Technical J 1948; 27:379-423.
8. Hutcheson K. A test comparing diversities based on the Shannon formula. J Theor Biol 1970; 29:151-154.
9. Korber B.T. et al. Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. J Virol 1994; 68:7467-7481.
10. Pawlotsky J.M. et al. Interferon resistance of hepatitis C virus genotype 1b: relationship to nonstructural 5A gene quasispecies mutations. J Virol 1998; 72:2795-2805.
11. Magurran A., McGill B.J. (eds.). Biological diversity: frontiers in measurement and assessment. Oxford, UK: Oxford University Press, 2010.
12. Chao A. Nonparametric estimation of the number of classes in a population. Scand J Statist 1984; 11:265-270.
13. Salicrú M. et al. Asymptotic distributions of (h,Fi)-entropies. Commun Statist Theory Meth 1993; 22:2015-2031.
14. Pardo L. et al. Large sample behavior of entropy measures when parameters are estimated. Commun Statist Theory Meth 1997; 26:483-501.
15. Ramírez C. et al. 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 2013; 98:273-283.
16. Cabot B. et al. Nucleotide and amino acid complexity of hepatitis C virus quasispecies in serum and liver. J Virol 2000; 74:805-811.
17. Nei M. Molecular evolutionary genetics. New York: Columbia University Press, 1987.
18. Dik J., Gunst M. The distribution of general quadratic forms in normal distributions. Statistica Neederlandica 1985; 39:14-26.

Section 2

From ecology to virology

Abstract

In this article, we critically evaluate the information provided by several population diversity indices used to evaluate quasispecies composition, and we propose the introduction of some new ones used in ecology. The indices are separated into three groups: incidence, abundance, and function-related. This classification aims to clarify the type of information provided by each index, imparting context to interpretation of the results. The challenge of quasispecies sampling and the need for rarefaction in fair comparisons is addressed for several indices. We suggest a multidimensional approach, introduce several quasispecies profiles as alternatives to simple indices, propose some guidelines, and illustrate the use of these indices with a simple example, applying them to three hepatitis C virus clinical samples, in which the population heterogeneity differed.

Remarks

It is important to have several available diversity indices and profiles to define viral quasispecies composition at the molecular level. A useful description of a viral quasispecies should result from choosing the most appropriate diversity indices for this purpose. We hope that the systematized organization given here will be of help to better understand the type of information provided by each index, and guide problem-specific selection of those that are most suitable for each scenario under study.

Highlights

- Diversity indices used in ecology are proposed and classified into three groups, according to the information they provide: incidence, abundance, function-related.
- A multidimensional view of diversity is suggested. No single index can provide sufficient descriptive information about a complex system such as a quasispecies.
- Diversity profiles are introduced as graphical representations that provide richer information about quasispecies composition than any single index. In particular, these include Montserrat plots and the Hill numbers profile.
- Recommendations are given to describe quasispecies in several situations.

Availability

In addition to this publication, an R package, QSutils [1], was designed. Qsutils implements functions to compute all the indices mentioned here and includes other functions that are useful for simulations and for manipulating data based on haplotypes and frequencies. Three vignettes illustrate the use of these functions to read and manipulate fasta files with haplotype frequencies in the headings, to compute the corresponding diversity indices, and to simulate quasispecies composition.

REFERENCE

1. Guerrero-Murillo M, Gregori J. QSutils: quasispecies diversity. R package version 1.16.0. 2022.

2.
Viral quasispecies complexity measures

JOSEP GREGORI, CELIA PERALES, FRANCISCO RODRÍGUEZ-FRÍAS, JUAN I. ESTEBAN, JOSEP QUER, ESTEBAN DOMINGO

ABSTRACT

Mutant spectrum dynamics (changes in the related mutants that compose viral populations) has a decisive impact on virus behavior. The several platforms of next generation sequencing (NGS) to study viral quasispecies offer a magnifying glass to study viral quasispecies complexity. Several parameters are available to quantify the complexity of mutant spectra, but they have limitations. Here we critically evaluate the information provided by several population diversity indices, and we propose the introduction of some new ones used in ecology. In particular we make a distinction between incidence, abundance and function measures of viral quasispecies composition. We suggest a multidimensional approach (complementary information contributed by adequately chosen indices), propose some guidelines, and illustrate the use of indices with a simple example. We apply the indices to three clinical samples of hepatitis C virus that display different population heterogeneity. Areas of virus biology in which population complexity plays a role are discussed.

Keywords

Diversity indices, Shannon entropy, Gini-Simpson index, Hill numbers, mutation frequency, nucleotide diversity, quasispecies profiles, Montserrat plots.

1. The complexity challenge

RNA viruses and DNA viruses replicated by low fidelity polymerases (that display mutation rates in the range of 10^{-3}–10^{-5} mutations introduced per nucleotide copied) consist of complex and dynamic mutant spectra. They are termed viral quasispecies because of the resemblance between their population structure and dynamics with the ones proposed in the quasispecies theory of primitive replicons developed by M. Eigen, P. Schuster and colleagues (recent reviews in Andino and Domingo, 2015; Domingo and Schuster, 2016; Domingo et al., 2012; Lauring and Andino, 2010). Mu-

tant spectra, also termed mutant distributions, clouds or swarms, can be regarded as the phenotypic reservoir of asexual populations (Schuster, 2010), thus rendering of interest a quantification of mutant spectrum composition in terms of genome types (haplotypes or genomic sequences with identical mutations) and their frequencies. A dynamic mutant spectrum, with a continuous change of its composition, is tantamount to exploration of sequence space (different related sequences attainable by the viral genome) for virus adaptability (Domingo and Schuster, 2016; Eigen and Biebricher, 1988).

Mutant spectrum-mediated adaptability encompasses at least three relevant parameters: (i) the number of mutants present at a given time in the quasispecies, (ii) the frequency of different haplotypes (set of genomes with the same nucleotide sequence), and (iii) the viral population size in the replicative ensemble (the total number of viral particles in the population under consideration). Parameters (i) and (ii) determine what is frequently referred to as the amplitude of the mutant spectrum. Amplitude may mean either a great diversity of variants with one or two mutations per genome (or genomic region analyzed) with scarcity of sequences with three or more mutations, or a broad distribution of variants with one and multiple mutations.

The virus population size has an obvious participation in adaptive potential that has been expressed with a distinction between extrinsic (dependent on population size) and intrinsic (independent of population size) properties of mutant spectra (Domingo and Perales, 2012). We define complexity of a viral quasispecies as the intrinsic property that quantifies the diversity and frequency of haplotypes, independently of the population size that contains them. The relevance of the complexity level *per se* has been evidenced by the decreased adaptability of viruses whose polymerase displays higher or lower copying fidelity than the corresponding standard viruses, with comparable population size in the same biological context (Bordería et al., 2016; Campagnola et al., 2015; Pfeiffer and Kirkegaard, 2005; Vignuzzi et al., 2006, among other studies). Mutant spectrum complexity can explain or predict virus behavior in the face of specific environmental changes. Variations in population size can render the complexity of a viral population sufficient or insufficient to express an adaptive capacity (Domingo, 2016). For example, the selection of a viral mutant resistant to an antiviral inhibitor, or the capacity of a virus from a reservoir to adapt to a human host may be possible only when the viral population size is sufficiently large to include the mutants that can confer resistance to the inhibitor or capacity to replicate in humans. These possibilities render a quantitative characterization of mutant spectra a key issue of current virology.

Ideally, complexity of a viral quasispecies should be based on knowledge of the complete mutant repertoire in the population under study, an aim which at the moment is unfeasible. Close to being ideal, complexity could be measured with thousands of full length genomes (amounting to about 10% of the total number present in a biological sample), chosen at random from the population, and amplified and sequenced employing low error copying and sequencing procedures. Current trends

in viral genomics render likely the achievement of this goal in a few years. At present, however, we are left with a far less perfect methodology that forces an extrapolation from the analysis of a limited sample to what might be the true complexity of the entire population. Further, as full genome length sequences of viruses are not yet attainable, the study should be limited to some regions of interest (amplicons). The problem is parallel to that encountered in general ecology or paleontology in which procedures for interpolation (rarefaction) and extrapolation (prediction) have been implemented to capture species diversity from samples of natural habitats or fossil records (reviews in Chao et al., 2014b; Hortal et al., 2006). The parallelism between the descriptions of ecological diversity and viral population complexity has encouraged us to comparatively review the state of the art regarding complexity indices used in both disciplines. We propose the application to virology of some indices proven adequate to measure species diversity in ecology, and suggest some guidelines to be considered for the quantification of mutant composition in clinical and experimental virology. Because no single index can capture viral population complexity in full, we suggest a multivariate analysis in the sense of examining the same set of data using different indices.

2. Classical complexity indices used in virology

By classical indices we mean those traditionally used to characterize viral quasispecies through the nucleotide sequencing of a sample of genomes or genomic region from the mutant spectrum of interest. Before the advent of new generation sequencing (NGS) the standard procedure available was based on molecular or biological cloning of genomic RNA or DNA and its amplification, followed by Sanger sequencing. The information to derive viral quasispecies complexity measures is contained in the multiple alignment of all unique sequences (haplotypes) fully covering the region of interest (amplicon) and their observed frequencies (Fig. 1). A multiple alignment of haplotypes displays the entities (haplotypes, polymorphic sites or mutations) that are present in a viral quasispecies. The frequencies inform of the abundances of those entities, and are closely related to the relative fitness of each haplotype. Any individual diversity index provides a partial view of the complete information about the viral quasispecies.

The indices commonly used to compare viral quasispecies diversity are the minimum mutation frequency (*Mf min*), the maximum mutation frequency (*Mf max*), the normalized Shannon entropy (H_{SN}), and the nucleotide diversity (π). They have and are still providing extensive information on mutant spectrum complexity both in natural and laboratory samples of viruses (see Domingo et al., 2012 for a review).

Mf min is the proportion of residues that includes different mutations in the set of sequences (reads), counting only once the mutations repeated at the same genomic position in different clones (Equation I; equations are given in section "Equations and practical examples"). *Mf max* is the proportion of residues that include mutations in the set of sequences (Equation IIb). Mutations are counted by comparison of each individ-

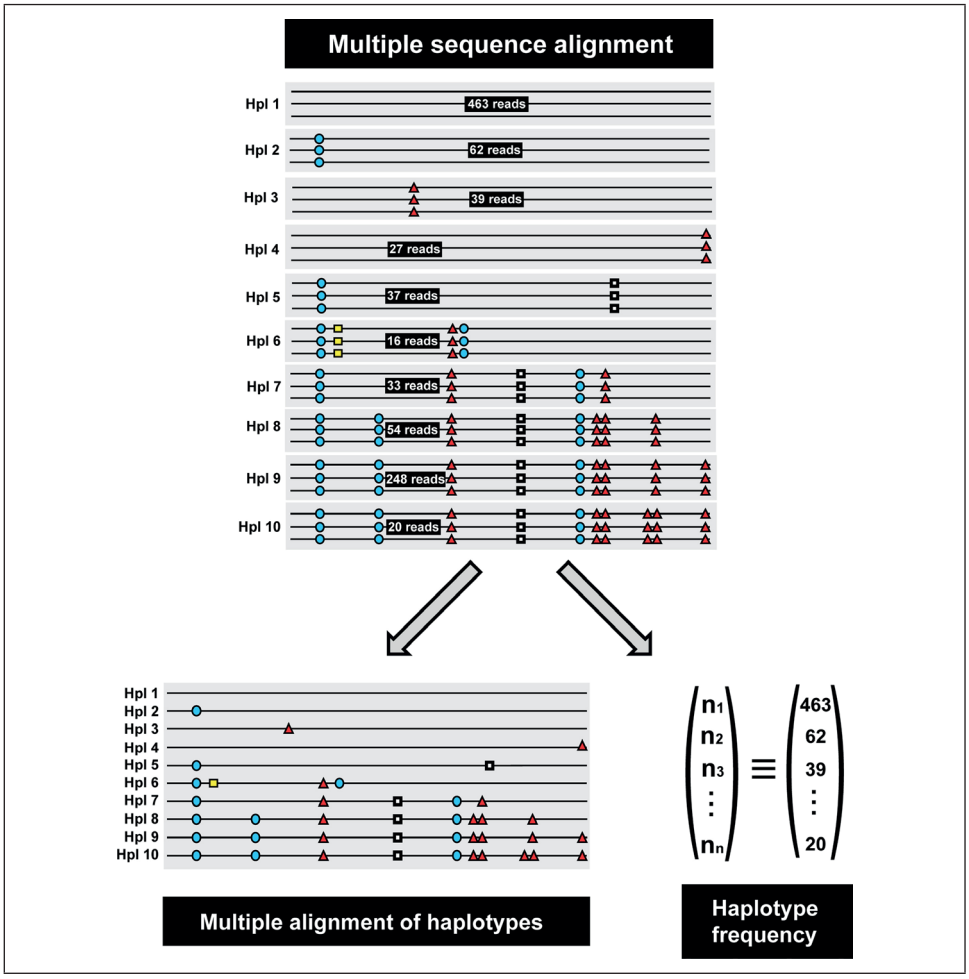


Fig. 1. Schematic representation of the sequences (reads) obtained from a NGS experiment. Viral genomes are represented as horizontal lines and mutations as different colored symbols on the lines. The sequences are clustered by haplotypes (Hpl). The information obtained from the sequence alignment can be divided into a haplotype alignment (each different sequence is counted once), and a vector of frequencies (shows the frequency of different haplotypes).

ual sequence with the consensus sequence of the corresponding population. The latter is either determined experimentally by direct sequencing of a PCR product or built with the most frequent nucleotide at each genomic position in the aligned sequences. Sets of compared sequences from different mutant spectra can have different numbers and lengths since the division by the total number of nucleotides sequenced introduces the required normalization (Equations I and II). It is advisable, however, that a similar and sufficient number of sequences be compared to minimize statistical biases.

The Shannon entropy H_S (Equation IIIa) provides a measure of diversity based on haplotype frequencies. Strictly speaking, H_S is a measure of the uncertainty in assigning a randomly sampled sequence to an haplotype in the viral quasispecies (Gotelli and Chao, 2013). The most common normalization is to *log(N)*, where *N* is the number of clones, to yield H_{SN} (Equation IIIb) which represents the proportion of different sequences in the set of aligned sequences. A maximum value of 1 is obtained when all clones are different, and a minimum of 0 when all clones are identical (Abbate et al., 2005; Cabot et al., 2000; Grande-Pérez et al., 2002; Pawlotsky et al., 1998; Wolinsky et al., 1996). Normalization to *log(N)* demands that the samples under comparison include the same number of clones. Alternative normalizations consist in dividing by *N*, by the length of the sequenced amplicon, *a*, or by *log(H)* (Fishman and Branch, 2009; Larrat et al., 2015; Nasu et al., 2011; Nishijima et al., 2012). The value normalized to *log(H)* is termed H_{SH}, where *H* is the number of estimated haplotypes (Equation IIIc), and corresponds to the normalization used in ecology. H_{SH} has been used with NGS data (Gregori et al., 2014 among other examples).

Population nucleotide diversity or index π measures the average number of nucleotide differences between any two genomes of the quasispecies (Nei, 1987; Nei and Kumar, 2000) (Equation IVb). Pair-wise differences have been traditionally evaluated by the Hamming distance (number of mutations that distinguish a pair of sequences), although any substitution model (JC69, K80, F81, etc.) (Nei and Kumar, 2000 among others) or subsets of differences (transitions or transversions, synonymous or non-synonymous mutations) may be considered. Index π provides more valuable information than *Mf* because it takes into account the differences between any two genomes in the population.

The Simpson index, H_{Si} (Equation V), is the probability that two randomly selected genomes from a viral population belong to the same haplotype (Nowak et al., 1991; Wolinsky et al., 1996). The index was established in ecology to give the probability that two randomly selected individuals from a habitat belong to the same class (Magurran, 2004). In its application to virology, any sampled genome has either the same or a different sequence than a given haplotype. Therefore, if the Simpson index is λ, $1-\lambda$ represents the probability that two genomes taken at random belong to a different haplotype. This transformation is known as the Gini-Simpson index, H_{GS} (Equation VIa). The indices just summarized can be applied to sequences sampled either by the cloning-Sanger procedure or new NGS platforms, using specific genomic regions or entire genome sequences.

3. Application to viruses of diversity measures used in ecology

Introducing the systematics used to classify diversity measures in ecology may help to better understand the quality and extent of information provided by each index. The diversity indices adopted from ecology are classified in three different groups: incidence, abundance and function (Fig. 2).

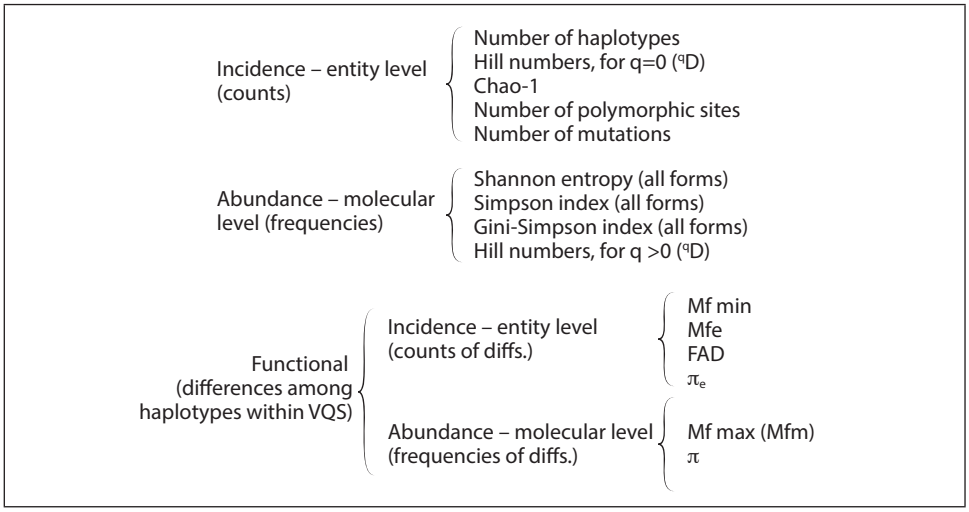


Fig. 2. Hierarchy of diversity indices. This classification contributes to clarify the kind of information provided by each index.

We distinguish between entities in the multiple alignment of haplotypes and instances of those entities in the VQS population. Each entity in a multiple sequence alignment is represented in the viral quasispecies population by a given frequency. Incidence-based diversity indices are those that correspond to counts of entities in the multiple alignment of haplotypes. Abundance-based diversity indices consider both the observed entities and their frequencies in the population; examples are H_S and H_{GS} (Equations IIIa, VIa, VIb). Functional indices are those based on the differences between the observed haplotypes, and may include or not the frequency of each of them in the population. Hence they might be further divided into incidence-based or abundance-based functional indices. *Mf max* and π are examples of abundance-based functional indices (Equations I, IIb, IVb).

The abundance-based indices measure either diversity (number and frequency of different haplotypes) or evenness (uniformity of the haplotype distribution). Evenness ranges between 0 and 1, with 1 meaning that all observed haplotypes have identical abundance. The normalized forms of the Shannon entropy [to *log(H)*] and the Gini-Simpson index constitute measures of evenness.

3.1. Additivity and Hill numbers

Although H_S, H_{GS}, and several of other indices are considered measures of diversity (Magurran, 2004; Magurran and McGill, 2010), their units do not allow an easily interpretable and intuitive appreciation of the authentic diversity. They have important limitations because of their asymptotic behavior (saturability) and lack of linearity re-

garding the addition of new haplotypes; the higher the number of haplotypes the less sensitive these two indices are to frequency changes.

Hill developed a generalization of diversity measures in units of equally abundant species, ${}^{q}D$ (Equation VII), that solves most of the observed inconsistencies, and includes as particular cases a transformation of H_S and H_{Si} indices (Hill, 1973). The q value represents the contribution of the different haplotypes to the diversity: when q is 0 all haplotypes have the same weight and contribute equally to the measurement; with increasing values of q the measure of diversity becomes progressively less sensitive to rare haplotypes, and at infinity only the abundance of the dominant haplotype matters. The Hill number of order q=0, ${}^{0}D$, is the number of haplotypes; ${}^{1}D$ is undefined but its limit as q tends to 1 is the exponential of H_S; ${}^{2}D$ is the inverse of H_{Si}. ${}^{\infty}D$ is the inverse of the relative abundance of the dominant haplotype, while ${}^{-\infty}D$ is the inverse of the relative abundance of the rarest haplotype (see examples below). The Hill numbers obey the replication principle, which means that if we have n equally diverse, equally large viral quasispecies with no haplotype in common, the diversity of the pooled population must be n times the diversity of a single viral quasispecies (Chao et al., 2014a).

Regarding viral quasispecies, the ^{q}D for q=1 and q=2 are the most meaningful and could replace the more classical counterparts, H_S and H_{Si}. The advantage of ${}^{q}D$ for q=1 and q=2 is that they have common units, provide an intuitive interpretation of diversity, and obey also the replication principle. The full Hill numbers profile – the plot of ${}^{q}D$ versus q, as q varies from 0 to infinity – provides a visual method of viral quasispecies comparison in terms of haplotype abundance distribution (see Equation VII for a numerical application of Hill numbers to a set of aligned sequences).

3.2. Prospects of extension to functional diversity

It has long been recognized in ecology that beyond expressions of mere diversity, some additional measurements should reflect the habitat composition, sometimes referred to as functional diversity. Biotype differences among components of a biological community might be computed based on morphologic, taxonomic, genetic, or phylogenetic (tree branch length) differences. One of the major issues in viral quasispecies is how to relate genetic diversity, measured with the indices described here, with phenotypic or functional diversity. In viral populations this connection acquires a different meaning than in ecology, and it has to be simplified to relate functional pluripotency with the representation of different haplotypes (see Concluding remarks).

The functional attribute diversity, *FAD* (Equation VIII), is an incidence-based functional diversity index equal to the sum of the elements in the matrix of dissimilarities (Walker et al., 1999). With viral quasispecies VQS, the matrix of dissimilarities is taken as the matrix of pair-wise genetic distances between haplotypes. More appropiate to quasispecies would be the nucleotide diversity at the entity level, π_e (Equation IVa),

giving the average number of substitutions among pairs of haplotypes in the multiple alignment, which results of dividing the functional attribute diversity by the number of haplotype pairs. By extension one might define the mutation frequency at the entity level, *Mfe* (Equation IIa), a functional incidence-based index that corresponds to the fraction of mutated residues in the multiple alignment with respect to the dominant haplotype. For consistency with the proposed terminology, *Mf max* may be termed *Mfm*, meaning mutation frequency at the molecular or genome level, an abundance-based functional index. Note that in some cases the *Mfe* might be higher than *Mf max*.

The Rao's quadratic entropy is an abundance-based measure of functional diversity based on multiple traits (Rao, 1982; Rao, 2010). It includes a measure of abundances (how frequent is any biotype in the ensemble) in addition to dissimilarities (how different or divergent are the biotypes present). Translated to a viral quasispecies, abundance of biotypes would be equivalent to frequency of different haplotypes, and dissimilarities would be quantified by the dispersion of sequences in terms of the range of the number of mutations per genome; π is a form of Rao entropy.

As observed with abundance-based indices, functional indices lack additive properties (replication principle). In this respect the generalization of Hill numbers has been broadened to include measures of functional and phylogenetic diversity (Chao et al., 2010; Chao et al., 2014a). Due to their novelty it is not yet clear to us how well these indices might be implemented in a VQS scenario, and which of them might be more appropriate.

All functional indices based on genetic distances carry the assumption that genomes which are distant in sequence space are more likely to vary functionally than sequences close in sequence space; this assumption is not fully justified for viral quasispecies because single mutations in viral genomes can affect important traits such as resistance to components of the immune response, or antiviral agents, cell tropism or host range, among others (Domingo, 2016; Domingo et al., 2012).

In summary, diversity indices used in ecology can be applied to the characterization of viral quasispecies complexity, and can complement the information provided by the measurements currently in use. The complexity concept is necessarily multidimensional, and no single index can capture all what the term complexity entails.

4. Sampling a viral quasispecies

Viral populations are analyzed through a sample that is generally of a small size relative to its parental, entire population. The observed values of the indices described here tend to underestimate the true population heterogeneity, and are highly dependent of the sampling effort (see Ovreas and Curtis, 2010 for a discussion of this point in a metagenomics context). The larger the sample size, the larger will be the number of observed haplotypes, polymorphic sites or mutations (Gregori et al., 2014). Howev-

er, under ideal conditions, some predictive estimators can be used to circumvent the problem. For estimating the number of haplotypes, adequate estimators are Chao 1 (Equation IX) and Chao 2, the abundance-base coverage estimator and the incidence coverage estimator respectively (Colwell and Coddington, 1994; Gotelli and Chao, 2013). These estimators are based mainly on the number of observed singletons (*f1*) and doubletons (*f2*) (that is, the haplotypes represented in the sample by one and two clones, respectively), as representatives of rare components. In NGS data, the filtering out of the haplotypes found below the noise threshold level eliminates information on *f1* and *f2*. With the cloning Sanger procedure, the limited sample size precludes the use of this approximation. In these cases, a possible solution is to approximate the population composition by adjusting the abundance data to a population frequency model (Colwell and Coddington, 1994; Magurran and McGill, 2010). However, this proposal may not be always feasible, depending on the amount of excluded information at the lower end of haplotype abundances.

The reconstruction of the composition of viral quasispecies based on the analysis of the types of genomes present in a limited sample belongs to a class of problems that has been extensively addressed in ecology. Two types of predictions are usually sought to determine species richness (number of different species present in a habitat) that here we rephrase in terms of viral quasispecies: (i) an estimate of the number of haplotypes expected in an entire viral population from the data obtained in a sample of the same population, which represents an extrapolation problem. (ii) An estimate of the expected number of haplotypes in a sample which is smaller than the one used to probe the composition of the population. This is a typical interpolation problem that can be addressed by rarefaction (Colwell et al., 2012). Applying their standard representation to viral quasispecies, the interpolation and extrapolation inquiry regions can be readily distinguished by plotting the number of haplotypes as a function of the number of genomes analyzed. Although incidence-based indices will be the more affected by limiting the sample size, the abundance-based indices, and notably all forms of Shannon entropy, are also affected. The estimation of haplotype frequencies by its maximum likelihood estimator $p_i=n_i/N$ causes a bias in the estimated Shannon entropy, H_{SMLE} (Equation IIId). This bias depends on the number of haplotypes in the quasispecies and on the sample size, and may be partially corrected by a Taylor series expansion (Equation IIId). Formally, a more substantial source of error occurs when the number of haplotypes in the population is unknown; an estimator as those described above might be used to minimize the effect. Since the Gini-Simpson index is calculated as a sum of squares of frequencies, its value is influenced by the dominant and common haplotypes, and is less sensitive to the presence of rare (minority) haplotypes, and hence less affected by limited sample size. The unbiased form when using the haplotype frequencies based on the maximum likelihood estimator is given by Equation VIb. Using simulations with NGS data of very high depth, it has been observed that the expected values of Mf and π are minimally affected by the sample size (Gregori et al., 2014).

The comparison of viral quasispecies diversity measures between samples of different size will necessarily entail correction methods of interpolation nature such as rarefaction or down sampling. When dealing with filtered data (NGS), more sophisticated methods such as the fringe trimming might be required (Gregori et al., 2014).

5. Viral quasispecies profiles: example data

As an application of the indices described here, we present an example which consists in an analysis of three clinical samples of hepatitis C virus of different complexity whose origin has been described (Cubero et al., 2014) (Figs. 3 and 4) and which were used in a previous study of simulation with diversity indices (Gregori et al. 2014). They are three HCV NS3 amplicons, sequenced at a high depth and similar coverage (39,006-40,450 quality filtered reads). The three samples are identified as High, Mid and Low complexity populations. High corresponds to a high complexity HCV NS3 amplicon from a chronically infected patient. Mid corresponds to a different and more conserved NS3 amplicon of the same patient. Low is a rather homogeneous HCV NS3 amplicon from an acutely infected patient.

The sequence alignment of the High, Mid and Low complexity populations were used to calculate all the indices discussed in the text. The values calculated are compiled in Table 1 and 2. Both H_S, H_{GS}, *Mf*, π, and Hill numbers with $q \geq 1$ provide the order High > Mid > Low, with the last two indices one order of magnitude apart between samples. The incidence measures (number of haplotypes, polymorphic sites, mutations, *FAD*, and Hill number with $q=0$) give High > Low > Mid (Tables 1 and 2 and Fig. 4).

Indices which result of a sum of terms or that depend on a parameter (such as Hill numbers) may be expanded to produce a diversity profile. The profiles offer a visual representation of the quasispecies complexity in a two dimensional space, instead of a single point on a diversity index axis. They also provide a graphical means to compare samples and to better understand big amounts of genetic data.

The Montserrat plot which is a representation of the terms resulting in *Mf*, provides an example (Fig. 3): Low is a highly homogeneous quasispecies with a dominant haplotype representing 87% of the population, but with a long tail of low abundance single mutants, whereas Mid includes a lower number of mutations and haplotypes with three relatively abundant haplotypes with 1, 2 and 3 differences with respect to the dominant haplotype. The contribution of a high number of single mutants to the complexity of the Low population is similar to the contribution of a lower number of mutants at a higher abundance in the Mid population. The subpopulations that become apparent in the Montserrat plot may be used as a starting point to define a number of clusters in the partition analysis of quasispecies (PAQ), a non-hierarchical clustering method of quasispecies analysis developed by Baccam and colleagues (Baccam et al., 2001). PAQ may provide complementary information on viral population complexity, particularly when used with functional indices.

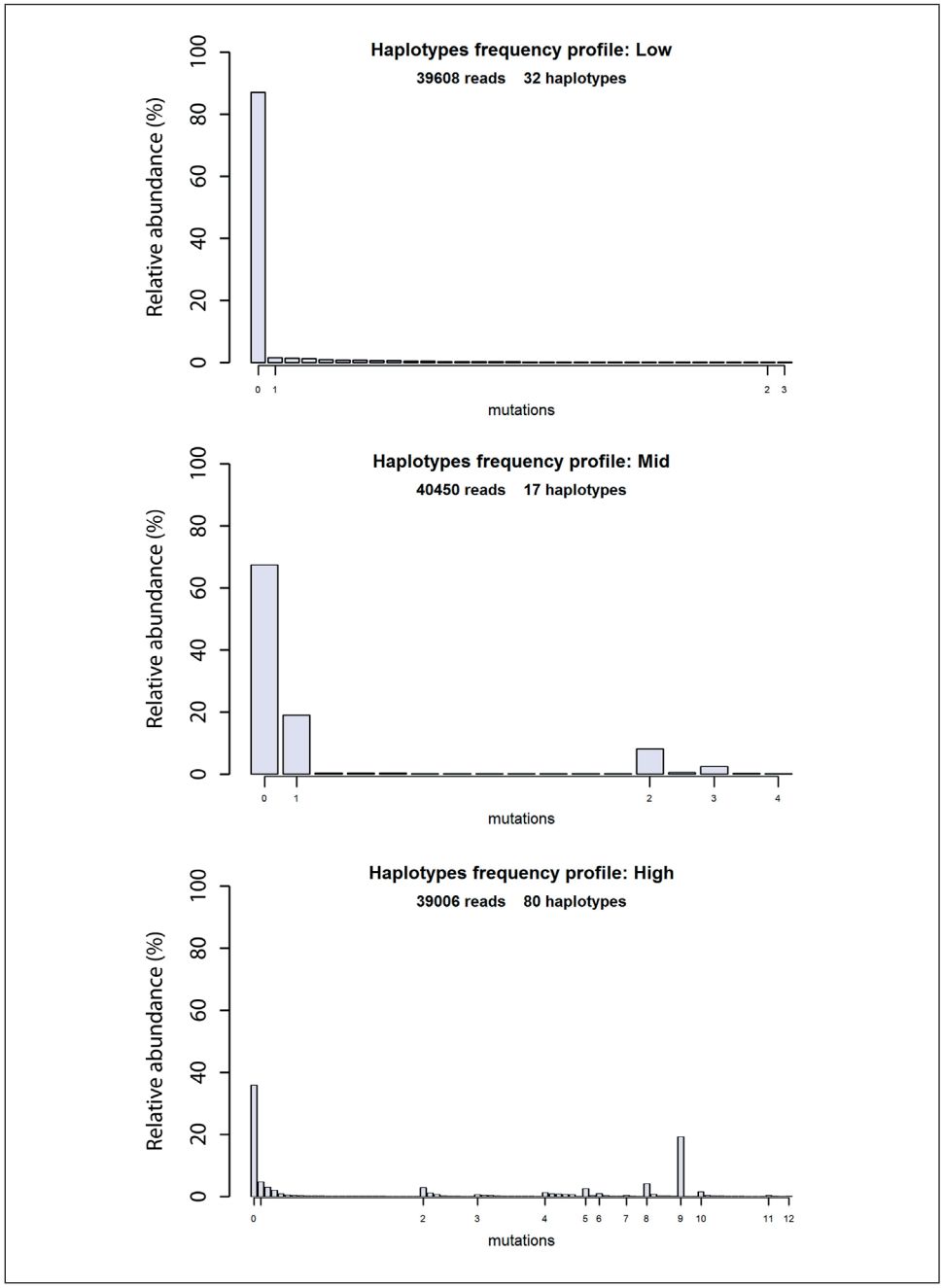


Fig. 3. Montserrat plots as a quasispecies profile with ordered genetic and abundance data. It shows the distribution of haplotypes within the viral quasispecies, where the haplotypes are ordered first by number of mutations with respect to the dominant haplotype, and second by decreasing abundance.

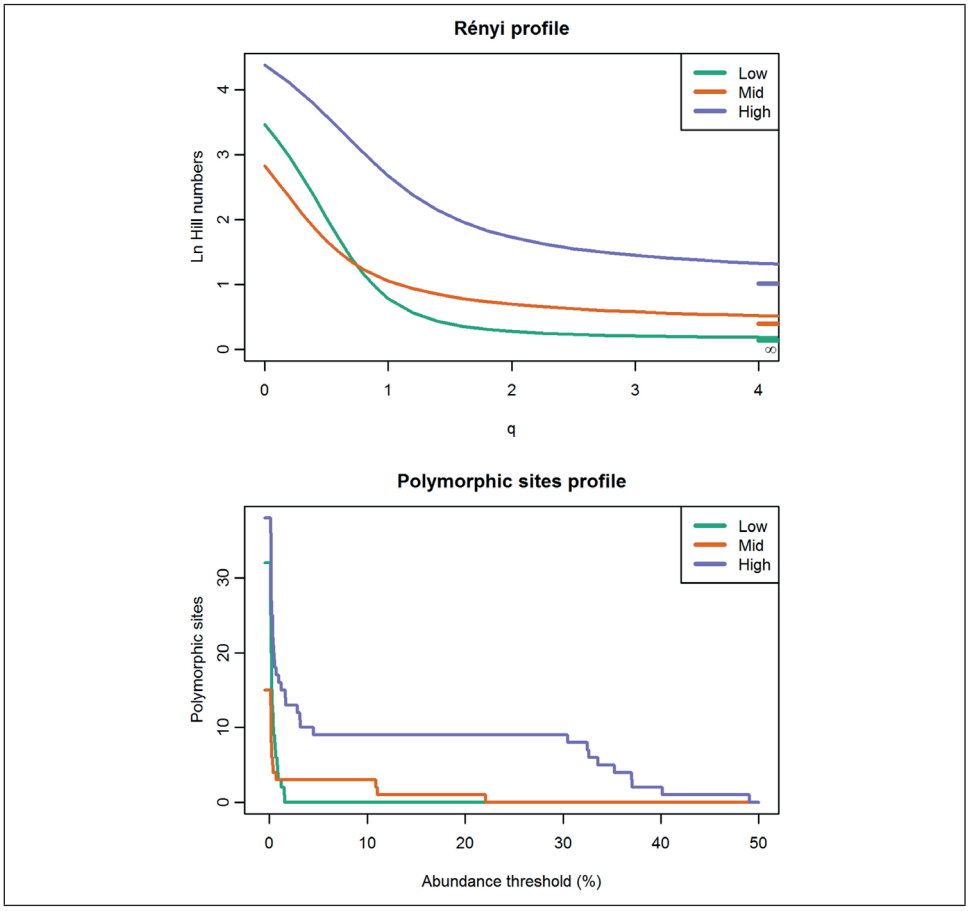


Fig. 4. *Top*: Quasispecies Hill numbers profile, in a log scale, with the profiles of the Low, Mid and High viral quasispecies. It provides a visual way to compare the abundance-based diversity of the viral quasispecies as the order q increases. A flat curve corresponds to a high evenness in viral quasispecies haplotypes distribution, while a steep curve is indicative of highly dominant haplotypes. The three examples yield steep curves, although falling from rather different levels. No flat curve should be expected with a viral quasispecies. The height at increasing values of q is related to the number of relatively abundant haplotypes. The use of a logarithmic scale enhances the visualization of profile crosses at low q values. *Bottom*: Polymorphic sites profile. It shows the relevance of the polymorphic sites as the abundance threshold is increased. A high curve is indicative of a high number of polymorphic sites with a relatively high abundance. A very steep curve is symptomatic of polymorphic sites at low abundances.

The Hill numbers profile is highly informative of the haplotype distribution structure, providing more complete information than any single abundance-based non-functional diversity index (Table 2 and Fig. 4 top). This profile (Fig. 4 top) shows the reversal, discussed above, between q=0 and q=1 due to the presence of a higher number of haplotypes, although at low abundances, in the Low respect with the Mid sample.

Table 1. Characterization of the three Low, Mid and High viral quasispecies amplicons used as example. Hpl: the number of observed haplotypes: PS: the number of polymorphic sites; nM: the number of different mutations; Mfe, π_e, Mfm and π were computed with the Hamming distance.

							Functional		
	Incidence (entity)			Abundance (molecular)		Incidence (entity)		Abundance (molecular)	
VQS	**Hpl**	**PS**	**nM**	H_S	H_{GS}	$Mfe \cdot 10^3$	$\pi_e \cdot 10^3$	$Mfm \cdot 10^4$	$\pi_m \cdot 10^4$
Low	32	32	32	0.786	0.2418	2.724	5.428	3.417	6.782
Mid	17	15	16	1.058	0.5021	3.943	7.195	12.57	21.26
High	80	38	38	2.679	0.8224	13.67	17.78	106.7	139.4

Table 2. Hill numbers of different orders for the three Low, Mid and High quasispecies amplicons used as example.

	Hill numbers of order q=					
VQS	**0**	**1**	**2**	**3**	**4**	**∞**
Low	32	2.195	1.319	1.232	1.204	1.149
Mid	17	2.880	2.009	1.785	1.688	1.483
High	80	14.57	5.630	4.256	3.769	2.760

The polymorphic sites profile (Fig. 4 bottom) shows a higher genetic diversity of the Low sample with almost the double of polymorphic sites, whereas the mutation landscape of the Mid sample is dominated by just two relatively abundant mutations. Also, Fig. 4 bottom contributes to tone down the much higher complexity of the High sample with a polymorphic sites landscape dominated by few highly abundant mutations.

The matrix of haplotype pair-wise distances may be visualized by specific profiles. The mutations profile by haplotypes (Suppl. Fig. 2, page 58) shows the number of haplotypes at each observed Hamming distance with respect to the dominant haplotype by a barplot. The pair-wise distances profile (Suppl. Fig 1, page 57) shows the fraction of pairs of haplotypes at each observed Hamming distance. The functional attribute diversity profile (Suppl. Fig. 1) shows the contribution of each observed Hamming distance to this index. Finally, the nucleotide diversity profile (Suppl. Fig 2) shows the contribution of each observed Hamming distance between pairs of haplotypes to π. Instead of the Hamming distance any suitable genetic distance may be considered.

In the ideal situation in which the diversity indices of all samples are comparable, as in the present example, in which the sequencing coverage is balanced, the con-

clusion would be that despite the differences shown by H_S, H_{GS}, *Mf*, and π, the Low sample represents a situation of richer genetic diversity in terms of polymorphic sites and number of haplotypes which imparts a higher capacity and flexibility to future changes than that observed in the Mid sample. The High sample maintains the highest degree of complexity with all indices and profiles tested; it is characterized by the presence of relatively abundant high order mutants, and by highly abundant mutations at a limited number of polymorphic sites.

This example illustrates how the indices described in the present report can be used to describe viral quasispecies at a fixed time point in a process of evolutionary continuum. We are currently investigating if some indices are particularly suitable to interpret time variations of quasispecies composition in sequential samples (see also Concluding remarks).

6. Recommendation of index choice

In keeping with the comparison of the Low, Mid and High HCV populations, the main recommendations to approach a description of viral complexity are:

- Take a set of diversity indices including incidence, abundance and functional indices, to obtain a multidimensional representation of viral quasispecies complexity.
- Use Hill numbers of order 1 and 2 because they can be more informative than the corresponding H_S and H_{Si}, and are less affected by saturation.
- When comparing samples, all incidence-based indices – including the Hill numbers of order below 2 – all forms of Shannon entropy, and the functional incidence-based indices (*FAD*, *Mfe*, and π_e) should be appropriately corrected, as they are highly sensitive to sample size differences. On the other hand, the Gini-Simpson index, the Hill numbers of order above 2, and the functional abundance-based indices (*Mf* and π) are less sensitive to rare haplotypes and more robust against sample size differences.
- Incidence-based indices are most adequate in a mutagenesis scenario; for example to characterize viral quasispecies subjected to lethal mutagenesis (Perales and Domingo, 2016). Abundance-based indices are strongly correlated with current haplotype fitness, and are appropiate in any evolutionary scenario where fitness is a relevant parameter.
- Compare profile plots as well as the selected indices. A given index may show the same value even with divergent profiles.

A simple sequence alignment and a calculation of the indices in the present article are included in the section “Equations and practical examples”.

7. Concluding remarks

The availability of several diversity indices is important to define viral quasispecies at the molecular level. Common misunderstandings in the field of virology are that the term quasispecies is equivalent to variation, and that each of the genomes in a viral population is a quasispecies. Quasispecies is a dynamic ensemble of related mutants and it is the ensemble that affects virus behavior. Other common misunderstandings affect the uses of some indices, particularly the Shannon entropy (Gregori et al., 2014).

There are at least four main reasons to quantify the complexity of viral quasispecies:

1. Complexity is one of the parameters that predict adaptability of viral quasispecies to complex environments (Domingo and Schuster, 2016; Pfeiffer and Kirkegaard, 2005; Vignuzzi et al., 2006). Modification of polymerase fidelity to deviate the amplitude of the mutant spectrum from an adaptability optimum is a strategy to design attenuated viral vaccines (Vignuzzi et al., 2008). It would be expected that diversity indices and their stability upon vaccine virus passage become part of the quality control of fidelity-based vaccines.
2. Mutant spectrum complexity is one of the factors identified as predictors of viral disease progression and response to treatments (Farci et al., 2000; reviews in Domingo et al., 2012; Farci, 2011).
3. A reduction of mutant spectrum complexity in sequential viral samples alerts of important evolutionary events, particularly the occurrence of a sweeping selection episode or a population bottleneck.
4. An effective antiviral mutagen in a lethal mutagenesis design should produce an increase of mutant spectrum complexity, at least in a transient fashion (Ojosnegros et al., 2008; Perales et al., 2011; review of those concepts in Domingo, 2016).

The evaluation of virus population complexity for biological inferences has as one of its major complications that the diversity profile of a viral quasispecies is not a constant parameter. The generation of a new haplotype is subjected to the uncertainties of mutant generation, and relative abundance of any new haplotype is influenced by past and present fitness levels of the relevant genomes, in interaction with other members of the mutant swarm. Consecutive expansions and contractions of diversity may be observed using standard quantification indices such as *Mf* and H_{SN} (Ojosnegros et al., 2008; Perales et al., 2011). A contraction may occur when a new haplotype with much higher fitness than those currently dominant emerges, eventually resulting in a substantial increase in viral load despite a transient reduction of complexity. A change in relative fitness among haplotypes may be due to a change in the environment, to the production of a new superior mutant, or both. Eventually a nearly stationary state with very high diversity may be reached.

A second important issue concerns the types of mutations. The relative abundance of point mutation classes is particularly relevant when studying lethal mutagenesis

produced by treatment with virus-specific nucleotide analogues, currently an active field of antiviral research. Each mutagenic nucleotide analogue has a preference for specific types of mutations, and a precise computation of mutation types (independently of the abundance of different haplotypes) is indicative of presence or absence of a mutagenic activity, a distinction which is particularly relevant in the clinical setting with antiviral agents that can display different mechanisms of activity (Dietz et al., 2013; Domingo, 2016, Domingo et al., 2012).

A valuable description of a viral quasispecies should come about from the adequate choice of diversity indices. We hope that the systematization introduced (see Fig. 2) could help to better understand the type of information provided by each index, and guide the problem-specific selection of the set of most adequate indices in each case.

8. Equations and practical examples

This section collects all formulae and shows simple examples of computation.

8.1. Notation used in the formulae

Symbol	Description
a	Amplicon length
D	Matrix of haplotypes pair-wise genetic distances (fraction of nt differences)
${}^{q}D$	Hill number of order q
d_{ij}	D element in row i and column j; genetic distance between the i-th and j-th haplotype
FAD	Functional attribute diversity
f_1	Number of singletons (haplotypes with a single clone)
f_2	Number of doubletons (haplotypes with two clones)
H	Number of haplotypes
H_{GS}	Gini-Simpson index
H_S	Shannon entropy
H_{SN}	Shannon entropy normalized to $log(N)$
H_{SH}	Shannon entropy normalized to $log(H)$
H_{Si}	Simpson index
log	All logs are natural (base e) except otherwise expressed
M	Number of mutations
Mfe	Mutation frequency by entity

Mf max	Mutation frequency by molecule
Mf min	Minimum mutation frequency
N	Total number of clones (reads) sampled from the viral quasispecies
n_i	Number of clones (reads) of the i-th haplotype
P	Number of polymorphic sites
p_i	Population frequency of the i-th haplotype in the viral quasispecies
$\hat{p}_i$	Maximum likelihood estimator (MLE) of p_i
π	Nucleotide diversity
q	Order of Hill numbers

8.2. Data description

The data used in the examples that follow is a simplification of the High sample, where most non-polymorphic sites have been removed and the abundances rescaled to a total of 1000 reads or clones. The multiple alignment with observed haplotypes is represented in Fig. 5 with abundances in Table 3, and pair-wise genetic distances in Table 4. The full information about the viral quasispecies genetic complexity is contained in the multiple alignment of all observed haplotypes (Fig. 5) and in the vector of observed frequencies (Table 3). Any diversity index is derived either from the multiple alignment, the vector of abundances, or both, and will inform about one aspect of the viral quasispecies genetic composition. As such no single index is fully informative.

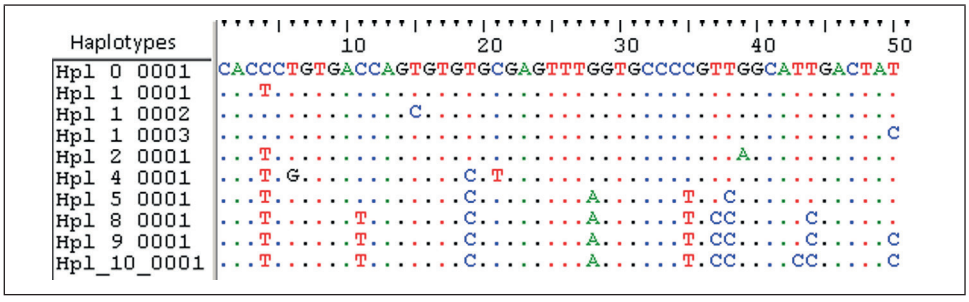


Fig. 5. Model of haplotypes multiple alignment used in the example computations.

Table 3. Haplotype ID and abundances used as model in the example calculations.

ID	Mutations	Reads	Hpl10 Frequency
Hpl_0_0001	-	463	0.463
Hpl_1_ 0001	1	62	0.062
Hpl_1_0002	1	39	0.039
Hpl_1_0003	1	27	0.027
Hpl_2_0001	2	37	0.037
Hpl_4_0001	4	16	0.016
Hpl_5_0001	5	33	0.033
Hpl_8_0001	8	54	0.054
Hpl_9_0001	9	248	0.248
Hpl_10_0001	10	20	0.020
Total		1000	1.000

Table 4. Matrix of pair-wise genetic (Hamming) distances used as model in the example calculations. Divide these numbers by the amplicon length, 50, to obtain the d_{ij} values used below in the formulas.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	1	1	1	2	4	5	8	9	10
2	1	0	2	2	1	3	4	7	8	9
3	1	2	0	2	3	5	6	9	10	11
4	1	2	2	0	3	5	6	9	8	9
5	2	1	3	3	0	4	5	8	9	10
6	4	3	5	5	4	0	5	8	9	10
7	5	4	6	6	5	5	0	3	4	5
8	8	7	9	9	8	8	3	0	1	2
9	9	8	10	8	9	9	4	1	0	1
10	10	9	11	9	10	10	5	2	1	0

8.3. Formulae and examples

Note: The equations are numbered by a combination of a roman number and a letter. The number keeps the order in which each main index is described in the text, and the letter identifies related indices. The aim is to keep together all forms of each index.

Eq I, Mf_{min}, minimum mutation frequency	**Eq IIa**, Mf_e, mutation frequency, entity level
$Mf_{min} = M/(N \times a)$	$Mf_e = \frac{1}{H} \sum_{i=1}^{H} d_{1i}$
Functional / incidence	Functional / incidence
Unique subst. per nucleotide sequenced	Subst. per bp in alignment
$14/(1000 \times 50) = 2.8 \times 10^{-4}$	$\frac{1}{10}\left(\frac{0+1+1+1+2+4+5+8+9+10}{50}\right) = 8.2 \times 10^{-2}$

Eq IIb, Mf_{max}, maximum mutation frequency	**Eq IIIa**, H_S, Shannon entropy
$Mf_{max} = \sum_{i=1}^{H} p_i d_{1i}$	$H_S(p) = \sum_{i=1}^{H} p_i log(p_i)$
Functional / abundance	Abundance
Mean subst. per nucleotide sequenced	nat for log, bit for log_2, ban for log_{10}
$\frac{463\times0+62\times1+...+248\times9+20\times10}{1000\times50} = 6.6 \times 10^{-2}$	$-[0.463 \times log(0.463) + ... + 0.020 \times log(0.020)] = 1.635$

Eq IIIb, H_{SN}, Shannon entropy normalized to log(N)	**Eq IIIc**, H_{SH}, Shannon entropy normalized to log(H)
$H_S(p) = \sum_{i=1}^{H} p_i log(p_i)/log(N)$	$H_S(p) = \sum_{i=1}^{H} p_i log(p_i)/log(H)$
Abundance	Abundance
nat / nat	nat / nat
0.237	0.710

Eq IIId, Hs_{MLE}, Shannon entropy bias correction	**Eq IVa**, π_e, nucleotide diversity, entity level
$Hs_{MLE} = H_S - \frac{\hat{H}-1}{2N} + \frac{1-\sum \hat{p_i}^{-1}}{12N^2} + \frac{\sum(\hat{p_i}^{-1}-\hat{p_i}^{-2})}{12N^3} + \dots$	$\pi_e = \frac{1}{H(H-1)} \sum_{i=1}^{H} \sum_{j=1}^{H} d_{ij}$
Abundance	Functional / incidence
nat	Mean substitutions between haplotypes
1.640	0.1098

Eq IVb, π, population nucleotide diversity	**Eq IVc**, π, sample nucleotide diversity
$\pi = \sum_{i=1}^{H} \sum_{j=1}^{H} p_i d_{ij} p_j$	$\pi = \frac{N}{N-1} \sum_{i=1}^{H} \sum_{j=1}^{H} \hat{p}_i d_{ij} \hat{p}_j$
Functional / abundance	Functional / abundance
Mean substitutions between molecules	Mean substitutions between molecules
0.08626	0.08634

Eq V, H_{Si}, Simpson index	**Eq VIa**, H_{GS}, Gini-Simpson index
$H_{Si}(p) = \sum_{i=1}^{H} p_i^2$	$H_{GS}(p) = 1 - \sum_{i=1}^{H} p_i^2$
Abundance	Abundance
probability	probability
$0.463^2 + 0.062^2 + ... + 0.020^2 = 0.2889$	0.7111

Eq VIb, $\hat{H}_{GS}$ sample-based Gini-Simpson index	**Eq VII**, ${}^qD(p)$, Hill numbers
$H_{GS}(p) = \frac{N}{N-1}\left(1 - \sum_{i=1}^{H} p_i^2\right)$	${}^qD(p) = \left(\sum_{i=1}^{H} p_i^q\right)^{1/(1-q)}$
Abundance	Abundance
probability	effective number of haplotypes
0.7118	${}^0D = 10$, ${}^1D = 5.130$, ${}^2D = 3.461$

Eq VIII, *FAD* Functional Attribute Diversity	**Eq IX**, $\hat{H}_{Chao}$, Chao-1 richness estimate
$FAD = \sum_{i=1}^{H} \sum_{j=1}^{H} d_{ij}$	$\hat{H}_{Chao} = H_{Obs} + \frac{f_1(f_1-1)}{2(f_2+1)}$
Functional / incidence	Incidence
Total substitutions between haplotype pairs	Haplotypes
$494/50 = 9.88$	

9. Note on software

All computations were made in the R language and platform (R Core Team, 2013) with in-house developed scripts and with the help of the packages Biostrings, ape, seqinr and ade4. DNA sequences distances were computed with function dna.dist() in package ape, dN and dS were computed with function kaks() in package seqinr using the method of (Li, 1993). A specific R package collecting all developed functions is under preparation.

Acknowledgements

Work in Barcelona supported by grants PI-12/01893, PI13-00456, PI-15/0856 and PI15-00829 from the Spanish Health Ministery. These grants were funded by Instituto de Salud Carlos III and cofinanced by the European Regional Development Fund (ERDF). Work in Madrid supported by grants BFU 2011-23604, SAF 2014-52400-

R from Ministerio de Economía y Competitividad, S2013/ABI-2906 (PLATESA-CM) from Comunidad Autónoma de Madrid, and Fundación Ramón Areces. CIBERehd (Centro de Investigación en Red de Enfermedades Hepáticas y Digestivas) is funded by Instituto de Salud Carlos III. C.P. is supported by the Miguel Servet program of the Instituto de Salud Carlos III (CP14/00121).

Supplementary Material

Supplementary figures associated with this article can be found in the on line version at http://dx.doi.org/10.1016/j.virol.2016.03.017

REFERENCES

Abbate I., Cappiello G., Longo R. 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.

Andino R., Domingo E. (2015). Viral quasispecies. Virology 479-480, 46-51.

Boergeling Y., Rozhdestvensky T., Schmolke M. et al. (2015). Evidence for a novel mechanism of influenza virus-induced type I interferon expression by a defective RNA-encoded protein. PLoS Pathog 11, e1004924.

Bordería A.V., Rozen-Gagnon K., Vignuzzi M. (2016). Fidelity variants and RNA quasispecies. Curr Top Microbiol Immunol 392, 303-322.

Cabot B., Martell M., Esteban J.I. et al. (2000). Nucleotide and amino acid complexity of hepatitis C virus quasispecies in serum and liver. J Virol 74, 805-811.

Campagnola G., McDonald S., Beaucourt S., Vignuzzi M., Peersen O.B. (2015). Structure-function relationships underlying the replication fidelity of viral RNA-dependent RNA polymerases. J Virol 89, 275-286.

Chao A., Chiu C.H., Jost L. (2010). Phylogenetic diversity measures based on Hill numbers. Philos Trans R Soc Lond B Biol Sci 365, 3599-3609.

Chao A., Chiu C.H., Jost L. (2014a). Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through hill numbers. Annual Review of Ecology, Evolution, and Systematics 45, 297-324.

Chao A., Gotelli N.J., Hsieh T.C. et al. (2014b). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecological Monographs 81, 45-67.

Charif D., Lobry J.R. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U. (Ed.). Structural approaches to sequence evolution: molecules, networks, populations. New York: Springer Verlag.

Colwell R.K., Coddington J.A. (1994). Estimating terrestrial biodiversity through extrapolation. Philos Trans R Soc Lond B Biol Sci 345, 101-118.

Colwell R.K., Chao A., Gotelli N.J. 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., Gregori J., Esteban J.I. et al. (2014). Identification of host and viral factors involved in a dissimilar resolution of a hepatitis C virus infection. Liver Int 34, 896-906.

Dietz J., Schelhorn S.E., Fitting D. et al. (2013). Deep sequencing reveals mutagenic effects of ribavirin during monotherapy of hepatitis C virus genotype 1-infected patients. J Virol 87, 6172-6181.

Domingo E. (2016). Virus as populations. Amsterdam: Academic Press, Elsevier.

Domingo E., Perales C. (2012). From quasispecies theory to viral quasispecies: how complexity has permeated virology. Math Model Nat Phenom 7, 32-49.

Domingo E., Schuster P. (2016). What is a quasispecies? Historical origins and current

scope. Curr Top Microbiol Immunol 392, 1-22.

Domingo E., Sheldon J., Perales C. (2012). Viral quasispecies evolution. Microbiol Mol Biol Rev 76, 159-216.

Dray S., Dufour A.B. (2007). The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software 22, 1-20.

Eigen M., Biebricher C.K. (1988). Sequence space and quasispecies distribution. In: Domingo E., Ahlquist P., Holland J.J. (Eds.). RNA genetics. Boca Raton, FL: CRC Press, pp. 211-245.

Farci P. (2011). New insights into the HCV quasispecies and compartmentalization. Semin Liver Dis 31, 356-374.

Farci P., Shimoda A., Coiana A. et al. (2000). The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. Science 288, 339-344.

Fishman S.L., Branch A.D. (2009). The quasispecies nature and biological implications of the hepatitis C virus. Infect Genet Evol 9, 1158-1167.

Gotelli N.J., Chao A. (2013). Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In: Levin S.A. (ed.). Encyclopedia of biodiversity, second edition. Volume 5, pp. 195-211. Waltham, MA: Academic Press.

Grande-Pérez A., Sierra S., Castro M.G., Domingo E., Lowenstein P.R. (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., Salicrú M., Domingo E. et al. (2014). Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics 30, 1104-1111.

Hill M. (1973). Diversity and evenness: a unifying notation and its consequences. Ecology 54, 427-432.

Hortal J., Borges P.A., Gaspar C. (2006). Evaluating the performance of species richness estimators: sensitivity to sample grain size. J Anim Ecol 75, 274-287.

Larrat S., Kulkarni O., Claude J.B. et al. (2015). Ultradeep pyrosequencing of NS3 to predict response to triple therapy with protease inhibitors in previously treated chronic hepatitis C patients. J Clin Microbiol 53, 389-397.

Lauring A.S., Andino R. (2010). Quasispecies theory and the behavior of RNA viruses. PLoS Pathog 6, e1001005.

Li W.H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36, 96-99.

Magurran A.E. (2004). Measuring biological diversity. Oxford: Blackwell Science Ltd.

Magurran A.E., McGill B.J. (Eds.). (2010). Biological diversity. Frontiers in measurement and assessment. Oxford: Oxford University Press.

Nasu A., Marusawa H., Ueda Y. et al. (2011). Genetic heterogeneity of hepatitis C virus in association with antiviral therapy determined by ultradeep sequencing. PLoS One 6, e24907.

Nei M. (1987). Molecular evolutionary genetics. New York: Columbia University Press.

Nei M., Kumar S. (2000). Molecular evolution and phylogenetics. New York: Oxford University Press.

Nishijima N., Marusawa H., Ueda Y 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., 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.

Ojosnegros S., Agudo R., Sierra M. et al. (2008). Topology of evolving, mutagenized viral populations: quasispecies expansion, compression, and operation of negative selection. BMC Evol Biol 8, 207.

Ovreas L., Curtis T.P. (2010). Microbial diversity and ecology. In: Magurran A.E., McGill B.J. (Eds.). Biological diversity. Frontiers in measurement and assessment. Oxford: Oxford University Press.

Pages H., Aboyoun P., Gentleman R., DebRoy S. (2012). Biostrings: string objects representing biological sequences, and matching algorithms. R package version 2.24.1.

Paradis E., Claude J., Strimmer K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289-290.

Pawlotsky J.M., Germanidis G., Neumann A.U., Pellerin M., Frainais P.O., Dhumeaux D. (1998). Interferon resistance of hepatitis C virus genotype 1b: relationship to nonstructural 5A gene quasispecies mutations. J Virol 72, 2795-2805.

Perales C., Domingo E. (2016). Antiviral strategies based on lethal mutagenesis and error threshold. Curr Top Microbiol Immunol 392, 323-339.

Perales C., Henry M., Domingo E., Wain-Hobson S., Vartanian J.P.

(2011). Lethal mutagenesis of foot-and-mouth disease virus involves shifts in sequence space. J Virol 85, 12227-12240.

Pfeiffer J.K., Kirkegaard K. (2005). Increased fidelity reduces poliovirus fitness under selective pressure in mice. PLoS Pathog 1, 102-110.

Rao C.R. (1982). Diversity and dissimilarity coefficients: a unified approach. Theoretical Population Biology 21, 24-43.

Rao C.R. (2010). Quadratic entropy and analysis of diversity. Sankhya: The Indian Journal of Statistics 72-A, part 1, 70-80.

Schuster P. (2010). Genotypes and phenotypes in the evolution of molecules. In: Caetono-Anolles G. (Ed.). Evolutionary genomics and systems biology. Hoboken, NJ: Wiley-Blackwell, pp. 123-152.

Stavrou S., Crawford D., Blouch K., Browne E.P., Kohli R.M., Ross S.R. (2014). Different modes of retrovirus restriction by human APOBEC3A and APOBEC3G in vivo. PLoS Pathog 10, e1004145.

Sun Y., Jain D., Koziol-White C.J. et al. (2015). Immunostimulatory defective viral genomes from respiratory syncytial virus promote a strong innate antiviral response during infection in mice and humans. PLoS Pathog 11, e1005122.

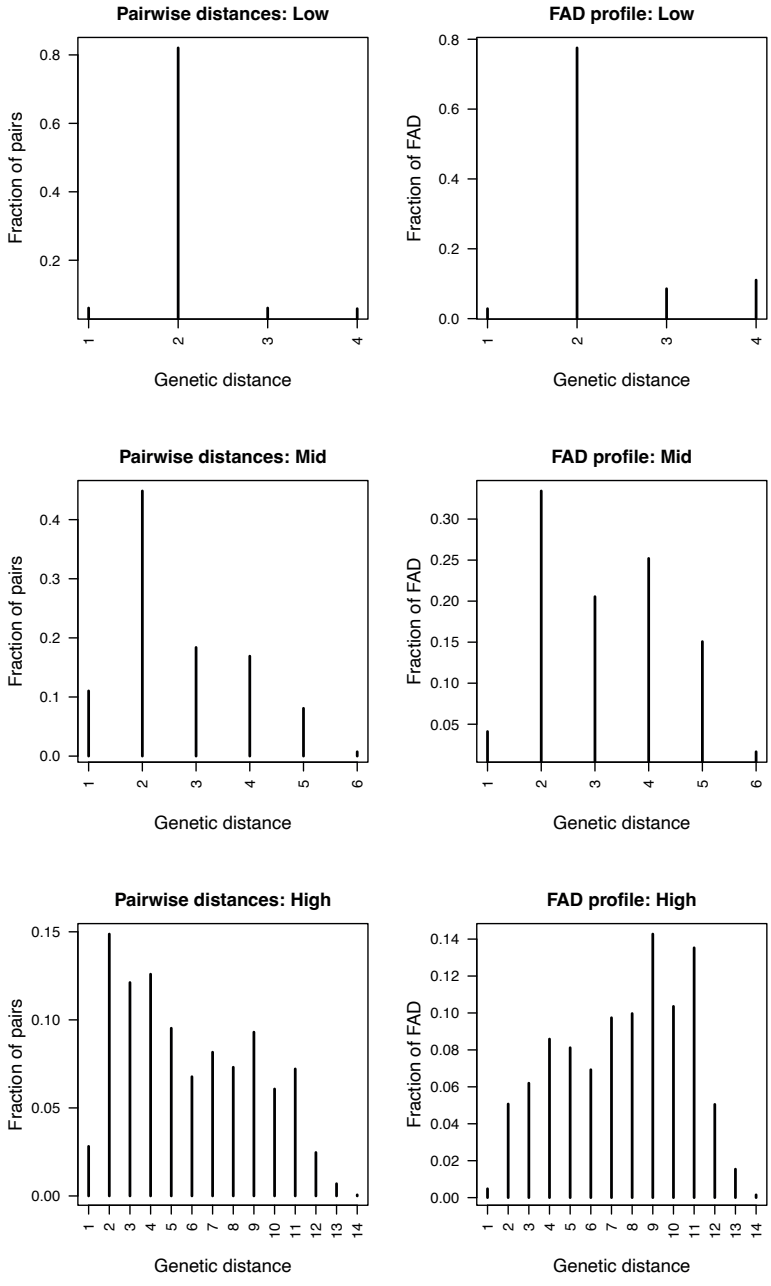
Vignuzzi M., Stone J.K., Arnold J.J., Cameron C.E., Andino R. (2006). Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439, 344-348.

Vignuzzi M., Wendt E., Andino R. (2008). Engineering attenuated virus vaccines by controlling replication fidelity. Nat Med 14, 154-161.

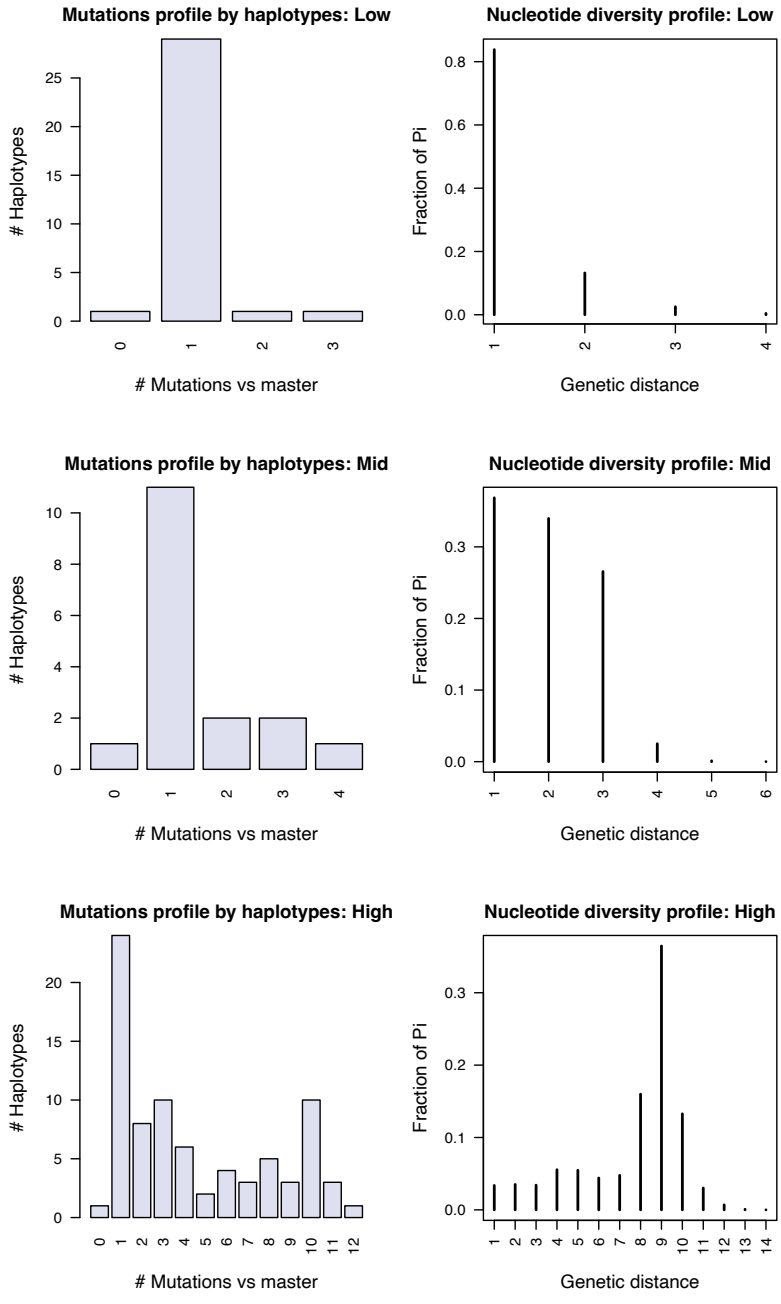
Walker B., Kinzig A., Langrid J. (1999). Plant attribute diversity, resilience, and ecosystem function: the nature and significance of dominant and minor species. Ecosystems 2, 95-113.

Wolinsky S.M., Korber B.T., Neumann A.U. et al. (1996). Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. Science 272, 537-542.

Supplementary Figure 1



Supplementary Figure 2



Section 3

Diversities, a tutorial

Abstract

The next document, a tutorial on quasispecies diversity, was first published as an appendix to Chapter 5, entitled "Hepatitis D virus (HDV) quasispecies study: experimental and bioinformatic analysis by next-generation sequencing methodology", in the book *Hepatitis D. Virology, management and methodology*, edited by Prof. Mario Rizzetto and Prof. Antonina Smedile [1]. This book is a reference text for the virology of hepatitis delta, a very small virus (1.7Kb in size) that requires simultaneous infection by hepatitis B virus to infect a host. Although HDV lacks its own polymerase, it is subject to considerable genetic diversity, among the highest rates of all known viruses. Chapter 5 also included some snippets of our code to process fastq files to obtain amplicon haplotypes and frequencies.

This tutorial focusses on quantification of the degree of diversity in the molecular composition of a quasispecies, as an extension of a previous review [2]. In some sense, the complexity of a quasispecies can be regarded as the biodiversity that exists in an ecological ensemble. We examine the diversity indices typically used in ecology to quantify this biodiversity [3], along with others classically used in population genetics [4]. The tutorial refers to computations on the data provided by multiple alignment of the haplotypes (in the form of amplicons) fully covering a genomic region of interest obtained by next generation sequencing (NGS).

Highlights

- Available diversity indices within each category (incidence, abundance, and function-related) are introduced and discussed.
- Calculations are shown step-by-step, using a simple dataset.
- Various profile plots are depicted as a means to visualize the composition of a quasispecies.
- Sample size and bias issues are discussed.
- Recommendations are given for choosing appropriate diversity indices to describe quasispecies in various scenarios.

REFERENCES

1. Gregori J, Quer J, Rodríguez-Frías F. Quasispecies complexity computations: a tutorial. In: Rizzetto M, Smedile A, eds. Hepatitis D. Virology, management and methodology. Rome: Il Pensiero Scientifico Editore, 2019.
2. Gregori J, Perales C, Rodríguez-Frías F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology 2016; 493:227-37. https://doi.org/10.1016/j.virol.2016.03.017
3. Magurran AE. Measuring biological diversity. Oxford: Blackwell Science, 2004.
4. Nei M. Molecular evolutionary genetics. 1st edition. New York: Columbia University Press, 1987; 276-9.

3. *Quasispecies complexity computations: a tutorial*

JOSEP GREGORI, JOSEP QUER, FRANCISCO RODRÍGUEZ-FRÍAS

Introduction

All viruses which pass through a RNA replication phase are found in what is known as a quasispecies. That is, a set of closely related genomes that may exhibit a huge number of variants but keeping a high degree of similarity among them. These variants are produced in the replication by the RNA-based RNA polymerases, which are error prone and lack the mechanism of error correction typical in most DNA polymerases.[1]

Taking HCV as an example, the replication error rate is estimated between $1{\cdot}10^{-4}$ and $1{\cdot}10^{-3}$ mutations per nucleotide per genomic replication (natural evolutionary rate of $1.5{\cdot}10^{-3}$ base substitutions/site/year). With a genome of 9600 bp, each time that a virion is replicated the number of errors introduced could be higher than 9 mutations. With viral loads of the order of 10^6 to 10^7 (6 to 7 logs), and a replication cycle of just a few hours, an estimate of 10^{12} virions are produced and eliminated daily. The variants are thus generated in the replication of the virions, but their viability and abundance in the quasispecies population is decided by their relative replicative capacity or fitness. The most fit variant will dominate the quasispecies, in the sense that it will become the most abundant variant in the population. Although a steady state could theoretically be reached, the pressure of the host immune system and/or therapeutic treatment causes variations in the fitness of all variants. New mutants may appear with enough capacity to become the dominants, and that could be outfitted in turn by variants to be produced. This is known as quasispecies dynamics. It is not evolution in the Darwinian sense, but a sort of genetic motion inside a confined space, the genetic space assigned to the corresponding subtype. Darwinian evolution could appear when a quasispecies by a highly improbable event is able to scape to this confinement and produce a new genotype or subtype. This tutorial focusses on the quantification of the degree of diversity of the mutant spectrum of a viral quasispecies (VQS), as an extension of a previous review.[2] In some sense the complexity of a quasispecies may be regarded as

the biodiversity which exists in an ecological ensemble, and we will explore diversity indices typically used in ecology to quantify this biodiversity,[3] along with others classically used in genetics.[4] The tutorial refers to computations on the data provided by the multiple alignment of the observed haplotypes (MAH) fully covering a genomic region of interest (amplicon), either by classical cloning followed by Sanger sequencing (SS) or by next generation sequencing (NGS). This analysis may be conducted by rows of the alignment (haplotypes or phenotype) or by columns (sites or genotype).

An example: quasispecies toy data

The data used in the tutorial to illustrate the computations, an amplicon of 50pb, provides a sufficiently simple case where computations may be followed almost by paper-and-pencil, and yet corresponds to a simplification of a deep sequenced amplicon of an HCV chronic patient, in the NS3 region.[5,6] Out of the originally sequenced 330pb amplicon most of the non-polymorphic sites have been removed and coverage has been rescaled to a total of 1000 reads. Figure A5.1 shows the alignment of the involved 10 haplotypes, and haplotypes ID, observed coverage and their frequencies in the viral quasispecies are given in Table A5.1. An even more simple case, taking just the first three haplotypes in this alignment, will be used to fully develop the proposed computations.

Quasispecies complexity by biodiversity indices

The diversity indices may be classified as incidence-based, abundance-based, and functional indices.[3] The functional indices may be subclassified as incidence-based and abundance-based. Incidence-based indices correspond to the number of observed entities, irrespective of their abundances, i.e. number of ob-

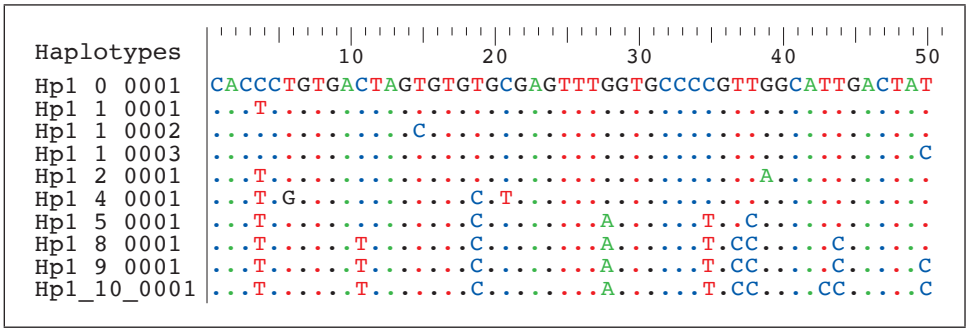


Figure A5.1
Multiple alignment of haplotypes (MAH) in toy quasispecies.

Table A5.1
Table of observed frequencies of each haplotype in the quasispecies. Hpl10 corresponds to the full alignment in Figure A5.1, while Hpl3 consists of the first three haplotypes with equivalent coverage. Under the mutations column, the number of differences between each haplotype and the dominant is given

ID	Mutations	Reads	Hpl10 frequency	Hpl3 frequency
Hpl_0_0001	-	464	0.464	0.821
Hpl_1_0001	1	62	0.062	0.110
Hpl_1_0002	1	39	0.039	0.069
Hpl_1_0003	1	27	0.027	-
Hpl_2_0001	2	37	0.037	-
Hpl_4_0001	4	16	0.016	-
Hpl_5_0001	5	33	0.033	-
Hpl_8_0001	8	54	0.054	-
Hpl_9_0001	9	248	0.248	-
Hpl_10_0001	10	20	0.020	-
Total		1000	1.000	1.000

served species in a given ensemble; each entity weighs the same in the computation. Abundance-based indices take into account the observed or estimated abundance of each entity; the contribution of each entity is weighted by their relative abundance in the population. The functional indices are computed on differences among traits of the observed entities. These may or may not be weighted by the relative abundance of each entity in the ensemble. In a VQS framework entities are the elements observed in the MAH.

Incidence-based or richness indices: how many are there?

Considering as entities the elements observed in the MAH (see Figure A5.1), the incidence indices are: the number of haplotypes (S), the number of polymorphic sites (P), and the number of mutations (M). An haplotype is each unique sequence found in the multiple alignment of sequences. A polymorphic site is a site where more than a single nucleotide may be observed in the alignment. A mutation is each unique nucleotide in the alignment which appears to be different with respect to the most abundant haplotype (the haplotype on top of the alignment, as in Figure A5.1).

Considering the data in Figure A5.1 we observe 10 haplotypes, that is ten different sequences; 14 polymorphic sites, that is 14 sites in the alignment where two or more nucleotides are observed and 14 mutations, that is 14 unique nucleotides observed in the alignment being different with respect to the dominant haplotype, on top.

P and M may be normalized to the amplicon length, 14/50=0.28, expressing then polymorphic sites and single mutations by site.

Note that, when sampling a population, both S, P and M are random variables rather than fixed values. And their distribution has the sample size as one of its parameters.

Abundance-based diversity indices: how frequent are they?

These indices take into account the observed entities and their relative abundance in the population. Each haplotype is represented by a number of genomes in the VQS, corresponding to its abundance. The two most common abundance-based indices used with VQS are the Shannon entropy and the Simpson index. The later may not be considered as a diversity index and a transformation known as the Gini-Simpson index might be preferred. When sampling a population, the sample size and the counts of each haplotype constitute the maximum likelihood estimates of the corresponding multinomial distribution. These parameter estimates are the basis of any abundance-based diversity index.

Shannon entropy: uncertainty in assignment

The Shannon entropy was developed in the frame of information theory and is a measure of uncertainty.[3 7] It represents the uncertainty in assigning a randomly sampled molecule (genome) to the corresponding haplotype in the VQS population.

In the VQS literature this index is rather popular, and it is commonly used in a normalized form, with rare exceptions. We recommend the unnormalized form given in equation (1), where S is the number of haplotypes in the VQS and p_i the relative abundance of each haplotype.

$$H_S(p) = -\sum_{i=1}^{S} p_i \ln(p_i) \qquad (1)$$

The normalized form, in the ecology sense, given by the equation (2) constitutes a measure of evenness (distribution uniformity) rather than diversity.[3]

$$H_{SH}(p) = -\sum_{i=1}^{S} p_i \ln(p_i) / \ln(S) \qquad (2)$$

Other used normalizations in the VQS literature may result in differential biases potentially bringing to wrong conclusions.[2 5]

The Shannon entropy obtained by the maximum likelihood estimates of haplotype proportions $\hat{p}_i=n_i/N$ is biased, the minimum bias estimator is expressed in equation (4), where N is the total number of reads, and $\hat{S}$ the estimated number of haplotypes in the quasispecies.[8]

A more substantial source of error occurs when the number of haplotypes in the population may not be properly estimated.

$$H_S(\hat{p}) = -\sum_{i=1}^{\hat{S}} \hat{p}_i \ln(\hat{p}_i) = -\sum_{i=1}^{\hat{S}} \left(\frac{n_i}{N}\right) \ln\left(\frac{n_i}{N}\right) \quad (3)$$

$$H_S(p) = H_S(\hat{p}) + \frac{\hat{S} - 1}{2N} + \ldots \quad (4)$$

This index varies from 0, when there is a single haplotype in the population, to $ln(S)$, when all haplotypes are equally abundant. Values above 4 are seldom observed in VQS with amplicon lengths below 500.

The most common criticisms to the use of this index are:

- units and meaning of difficult interpretation;
- an easy saturation that makes it less sensitive to abundance changes at high values;
- lack of sensitivity to the number of genetic differences among haplotypes in the VQS.

Applying equation 4 to the Hpl10 VQS data in Table A5.1 we obtain a value of 1.635. For the simpler Hpl3 VQS we may develop (4) as follows:

$$-(0.821\ ln(0.821) + 0.11\ ln(0.11) + 0.069\ ln(0.069)) + (3 - 1)/(2 \cdot 564) = 0.591$$

Gini-Simpson index: probability to be different

The Gini-Simpson index, equal to 1 minus the Simpson index, expresses the probability that two randomly sampled molecules from the viral population correspond to different haplotypes.[3] The nearly unbiased sample estimator is given by equation (5), where N is the number of reads.

$$H_{GS}(\hat{p}) = \left(\frac{N}{N - 1}\right) \left(1 - \sum_{i=1}^{\hat{S}} \hat{p}_i^2\right) \quad (5)$$

The interpretation of this index is more intuitive and clear than that of the Shannon entropy, nevertheless as a sum of squares it is scarcely sensitive to rare haplotypes, giving a higher weight to the common and abundant variants. As with the Shannon entropy this index saturates easily, and is non sensitive to the number of genetic differences among haplotypes in the VQS. H_{GS} has a range of variation from 0, when there is a single haplotype in the population, to asymptotically 1, when there are an infinity of equally abundant haplotypes.

Applying equation (5) to the Hpl10 VQS data in Table A5.1 we obtain a value of 0.7118. For the simpler Hpl3 VQS we may develop (5) as follows:

$$\left(\frac{564}{564-1}\right)(1 - (0.821^2 + 0.11^2 + 0.069^2)) = 0.310$$

Hill numbers: true diversity

Although both the Shannon entropy, the Gini-Simpson, and many of other published indices,[39] are considered measures of diversity, their units do not allow an easily interpretable and intuitive measure of real diversity. They are not linear with respect to the addition of new haplotypes, and they tend to saturate; the higher the number of haplotypes the less sensitive to frequency changes. That is, they show an asymptotic behaviour as the number of haplotypes increases. Mark Hill[10] developed a generalization of diversity measures in units of equally abundant species (6) that solved most of the observed inconsistencies, and that includes as particular cases a transformation of the Shannon entropy and the Gini-Simpson index. This generalization was reintroduced recently by Jost.[11]

$$ {}^{q}D(p) = \left(\sum_{i=1}^{S} p_i^q \right)^{1/(1-q)} \qquad (6) $$

A related form is the Rényi entropy,[9] which equals the logarithm of the Hill numbers (7).

$$ H_q(p) = \left(\frac{1}{1-q} \right) \ln \left(\sum_{i=1}^{S} p_i^q \right) \qquad (7) $$

As the order of the Hill number, q, increases, the measure of diversity becomes less sensitive to rare haplotypes. The Hill number of order q=0, ${}^{0}D$, is simply the number of haplotypes; ${}^{1}D$ is undefined but its limit as q tends to 1 is the exponential of the Shannon entropy; ${}^{2}D$ is the inverse of the Simpson index, that is the inverse of 1 minus the Gini-Simpson index. ${}^{\infty}D$ is the inverse of the relative abundance of the dominant haplotype, while ${}^{-\infty}D$ is the inverse of the relative abundance of the rarest haplotype. When q is 0, all haplotypes have the same weight and contribute equally to the measure; with increasing values of q the measure becomes progressively less sensitive to the rare haplotypes, and at infinity only the abundance of the dominant haplotype matters.

The Hill numbers are also called the true diversity of order q,[11] and measure the effective number of species; that is the number of equally abundant species D that are needed to obtain the same value of the generalized diversity of order q, ${}^{q}H$ (8), as shown in equation (9).

$$ {}^{q}H(p) = \sum_{i=1}^{S} p_i^q \qquad (8) $$

$$ {}^{q}H(p) = \sum_{i=1}^{S} p_i^q = \sum_{i=1}^{D} \left(\frac{1}{D} \right)^q = D^{1-q} \qquad (9) $$

The Hill numbers obey the replication principle by which if we have n equally diverse, equally large VQS with no haplotypes in common, the diversity of the pooled population must be n times the diversity of a single VQS.

The Hill numbers computed for the data of our example are collected in Table A5.2. For the Hpl10 VQS, given the frequency of the dominant haplotype, 0.464, the Hill numbers are limited at infinity by 1/0.464=2.16.

We may see that ${}^1D = exp(H_S)$ and that ${}^2D=1/(1 - H_{GS})$, ignoring the bias corrections in equations (4) and (5). For high sequencing depth these corrections have a very small impact. With VQS the most meaningful Hill numbers are those for *q=0, 1, 2*, and ∞.

Table A5.2
Hill numbers of different order q

	Order q					
VQS	**0**	**1**	**2**	**3**	**4**	**∞**
Hpl10	10	5.15	3.46	2.94	2.71	2.16
Hpl3	3	1.81	1.45	1.34	1.30	1.22

Functional diversity: how different are they?

Beyond abundance, one step further considers the differences among haplotypes in the VQS[4] by means of a matrix of genetic distances D. These distances could be as simple as the Hamming distance (see Table A5.3) divided by the amplicon length in number of nucleotides (50 in our example), or may consider just transitions or transversions, or synonymous changes by synonymous site, or the non-synonymous counterpart, or may be obtained through a model of nucleotide substitution (Jukes-Cantor, Kimura-80, etc).[12 13]

If we just consider distances between haplotypes, we obtain incidence-based functional diversity indices. If we consider all molecules sequenced and their abundances, that is the haplotype frequencies in the VQS population, we obtain abundance-based functional diversity indices.

Functional incidence-based diversity: haplotype (entity) level

Functional attribute diversity: cumulated difference

The *functional attribute diversity* (*FAD*) is an incidence-based functional diversity index equal to the sum of the elements in the matrix of genetic distances between haplotypes in the VQS $D_{i,j}$(10).[14]

$$\mathrm{FAD}(\mathrm{D}) = \sum_{i=1}^{S} \sum_{j=1}^{S} \mathrm{D}_{i,j} \qquad (10)$$

Table A5.3
Matrix of Hamming distances between pairs of haplotypes in the MAH

	Hpl_0_0001	Hpl_1_0001	Hpl_1_0002	Hpl_1_0003	Hpl_2_0001
Hpl_0_0001	0	1	1	1	2
Hpl_1_0001	1	0	2	2	1
Hpl_1_0002	1	2	0	2	3
Hpl_1_0003	1	2	2	0	3
Hpl_2_0001	2	1	3	3	0
Hpl_4_0001	4	3	5	5	4
Hpl_5_0001	5	4	6	6	5
Hpl_8_0001	8	7	9	9	8
Hpl_9_0001	9	8	10	8	9
Hpl_10_0001	10	9	11	9	10
	Hpl_4_0001	**Hpl_5_0001**	**Hpl_8_0001**	**Hpl_9_0001**	**Hpl_10_0001**
Hpl_0_0001	4	5	8	9	10
Hpl_1_0001	3	4	7	8	9
Hpl_1_0002	5	6	9	10	11
Hpl_1_0003	5	6	9	8	9
Hpl_2_0001	4	5	8	9	10
Hpl_4_0001	0	5	8	9	10
Hpl_5_0001	5	0	3	4	5
Hpl_8_0001	8	3	0	1	2
Hpl_9_0001	9	4	1	0	1
Hpl_10_0001	10	5	2	1	0

In our Hpl10 toy example the FAD is 9.88, that is the sum of elements in Table A5.3 divided by the amplicon length, 50. Table A5.4 shows the FAD values obtained from different genetic dissimilarities for both examples.

Developing for the Hpl3 VQS data, the sum of the top 3x3 elements in the matrix of Hamming distances (Table A5.3) amounts to 8 nucleotide differences, and dividing by the amplicon length, 50, we obtain 0.16.

Table A5.4
FAD values based on different types of genetic distance

	FAD			
VQS	**Raw**	**Transitions**	**Transversions**	**K-80**
Hpl10	9.88	9.52	0.36	11.71
Hpl3	0.16	0.16	0.00	0.165

Average mutation frequency by entity

The *mean mutation frequency by entity* (*Mfe*) takes one reference and measures the fraction of nucleotides in the MAH that are different with respect to this reference. See equation (11), where D_{1i} is the fraction of differences between the i-th haplotype and the reference sequence – taken as the first without loss of generality – and equivalent to the first column in Table A5.3 divided by the amplicon length, 50.

$$\mathrm{Mfe}\,(\mathrm{D}_1) = \frac{1}{\mathrm{S}} \sum_{\mathrm{i}=1}^{\mathrm{S}} \mathrm{D}_{1\mathrm{i}} \qquad (11)$$

The reference sequence may be either the dominant haplotype or the consensus sequence. With NGS the master sequence or dominant haplotype is usually preferred, as the consensus sequence might be a non existing entity in the quasispecies despite its population meaning.

In our Hpl10 example (Table A5.5) *Mfe* is 0.0820, and is obtained by the sum of the values in the first column of Table A5.3 divided by the amplicon length, 50, and divided by the number of haplotypes, 10.

Table A5.5
***Mfe* (entity level) values based on different types of genetic distance**

	Mutation frequency at the entity level			
VQS	**Raw**	**Transitions**	**Transversions**	**K-80**
Hpl10	0.0820	0.0800	0.0020	0.0972
Hpl3	0.0133	0.0133	-	0.0136

Developing equation 11 for the Hpl3 VQS:

$$\frac{1}{3}\ \frac{(0+1+1)}{50} = 0.0133$$

Nucleotide diversity by entity: average difference among haplotypes

It is a transformation of FAD, simply by dividing by the number of haplotype pairs, and it expresses the average difference among haplotypes in the MAH.

$$\pi_{\mathrm{e}}(\mathrm{D}) = \left(\frac{1}{\mathrm{S}\,(\mathrm{S}\,-\,1)} \right) \sum_{\mathrm{i}=1}^{\mathrm{S}} \sum_{\mathrm{j}=1}^{\mathrm{S}} \mathrm{D}_{\mathrm{i,j}} \qquad (12)$$

In our Hpl10 example (Table A5.6) π_e is 0.1098, and is obtained by the sum of the elements $D_{i,j}$ of the matrix of genetic distances, D (the values in Table A5.3 divided by 50), divided by the number of possible pairs of haplotypes, $10 \cdot 9$.

Table A5.6
π_e values based on different types of genetic distance

	π_e: nucleotide diversity at the entity level			
VQS	**Raw**	**Transitions**	**Transversions**	**K-80**
Hpl10	0.110	0.106	0.004	0.130
Hpl3	0.0267	0.0267	-	0.0275

Developing equation (12) for the Hpl3 VQS:

$$\frac{1}{3\cdot 2}\,\frac{(0+1+1+1+0+2+1+2+0)}{50} = 0.0267$$

Notice that the normalization in both *Mfe* and π_e, when dividing by the number of haplotypes or by the number of haplotype pairs, levels the value of these indices with growing number of haplotypes. They are not sensitive to the number of sequences in the MAH but to their average differences, and in this respect they reflect changes in the genetic structure of the quasispecies, not necessarily in the number of haplotypes or polymorphic sites.

Functional abundance-based diversity: virion (molecule) level

Average mutation frequency by molecule

The *proportion of different nucleotides* at the molecular level (*Mfm*), takes one reference and measures the fraction of nucleotides in the VQS population that are different with respect to this reference. See equation (13), where D_{1i} is the fraction of differences between the i-th haplotype and the master sequence, equivalent to the first column in Table A5.3 divided by the amplicon length, 50.

$$\mathrm{Mfm}(\hat{\mathrm{p}}, \mathrm{D}_1) = \sum_{\mathrm{i}=1}^{S} \hat{\mathrm{p}}_\mathrm{i}\ \mathrm{D}_{1\mathrm{i}} \qquad (13)$$

Note that by setting $\hat{p}_i$=1/*S*, ∀*i*: 1 . . . *S*, that is by assigning just one read to each haplotype, we obtain *Mfe* (11).

In our Hpl10 example (Table A5.7) *Mfm* is 0.0659, and is obtained by the cross product of the values in columns 2 and 3 in Table A5.1, and dividing by the number of sequenced nucleotides 50 · 1000.

Developing equation (13) for the Hpl3 VQS:

$$(0 \cdot 464 + 1 \cdot 62 + 1 \cdot 39)/(50 \cdot 564) = 0.00358$$

Table A5.7
Mfm values based on different types of genetic distance

	Mutation frequency at the molecular level			
VQS	**Raw**	**Transitions**	**Transversions**	**K-80**
Hpl10	0.0659	0.0656	0.0003	0.0801
Hpl3	0.00358	0.00358	-	0.00365

Nucleotide diversity: average difference among molecules

Taking into account the observed haplotype frequencies in the VQS and their differences, the *nucleotide diversity*, π_m, related to the Rao entropy in ecology,[15] corresponds to the mean genetic distance among molecules in the VQS.[4] The unbiased estimator of the neuclotide diversity in a sample is given by equation (14), where N is the sample size in number of reads or clones.

$$\pi_m(\hat{p}, D) = \left(\frac{N}{N-1}\right) \sum_{i=1}^{H} \sum_{j=1}^{H} \hat{p}_i \, D_{i,j} \, \hat{p}_j \qquad (14)$$

Note that by setting $\hat{p}_i = 1/S$, $\forall i$: 1 . . . S and with $N=S$, that is by assigning just one read to each haplotype, we obtain π_e (12).

In our Hpl10 example (Table A5.8) π_m is 0.08635, and is obtained in matrix form by the product $p^T D p$, where p^T is the transpose of the vector of frequencies (fourth column in Table A5.1) and D the matrix of dissimilarities (the values in Table A5.3 divided by 50); the bias correction factor is 1000/999.

Table A5.8
π_m values based on different types of genetic distance

	π_m: nucleotide diversity at the entity level			
VQS	**Raw**	**Transitions**	**Transversions**	**K-80**
Hpl10	0.0864	0.0857	0.000630	0.105
Hpl3	0.00651	0.00651	-	0.00664

Developing equation (14) for the Hpl3 VQS:

$$(564/(564-1))(464 \cdot 0 \cdot 464 + 464 \cdot 1 \cdot 62 + 464 \cdot 1 \cdot 39 + 62 \cdot 1 \cdot 464 + 62 \cdot 0 \cdot 62 + 62 \cdot 2 \cdot 39 + 39 \cdot 1 \cdot 464 + 39 \cdot 2 \cdot 62 + 39 \cdot 0 \cdot 39)/(50 \cdot 564 \cdot 564) = 0.00651$$

See Figure A5.2 and Table A5.9 with the full hierarchy of diversity indices discussed in this tutorial.

Table A5.9
Summary of diversity values of the toy data

						Functional			
	Incidence (entity)			**Abundance (molecular)**		**Incidence (entity)**		**Abundance (molecular)**	
VQS	*S*	*P*	*M*	H_S	H_{GS}	*Mfe*	π_e	*Mfm*	π_m
Hpl10	10	14	14	1.64	0.712	0.0820	0.110	0.0659	0.0864
Hpl3	3	2	2	0.591	0.309	0.0133	0.0267	0.00358	0.00651

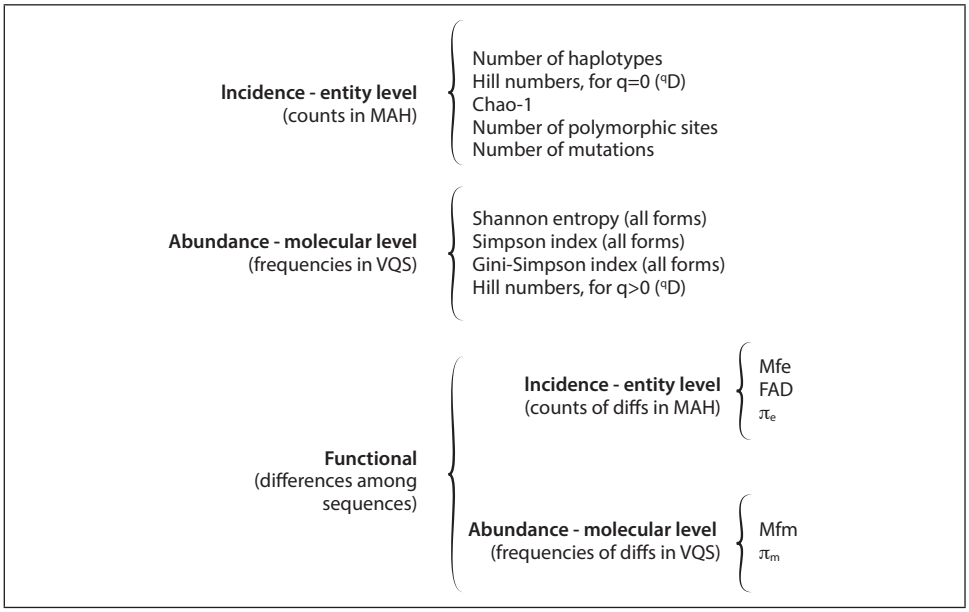


Figure A5.2
Classification of VQS diversity indices.

Quasispecies profiles: dissecting diversity

Quasispecies complexity may be visualized through different plots. These plots are highly informative of the VQS haplotypes distribution structure, providing richer information than any single abundance-based or functional diversity index, and a graphical means to compare samples. Most of these plots show the parts which are summarized in a diversity index.

Hill numbers profile

The *Hill numbers profile*, in Figure A5.3, plots ${}^{q}D$ at increasing values of q. As q increases ${}^{q}D$ diminishes, approaching asymptotically to the inverse of the frequency of the dominant haplotype as q tends to infinity. The highest drop is usually observed from q=0 to q=1, and from q=1 to q=2. It is steepest when there is a high number of rare haplotypes. And remains flat when all haplotypes are equally abundant.

When plotting Hill profiles of a set of samples in a single plot, the use of the Rényi entropy (7) – the natural logarithm of Hill numbers – could better visualize the crossing of curves at q values below 2.

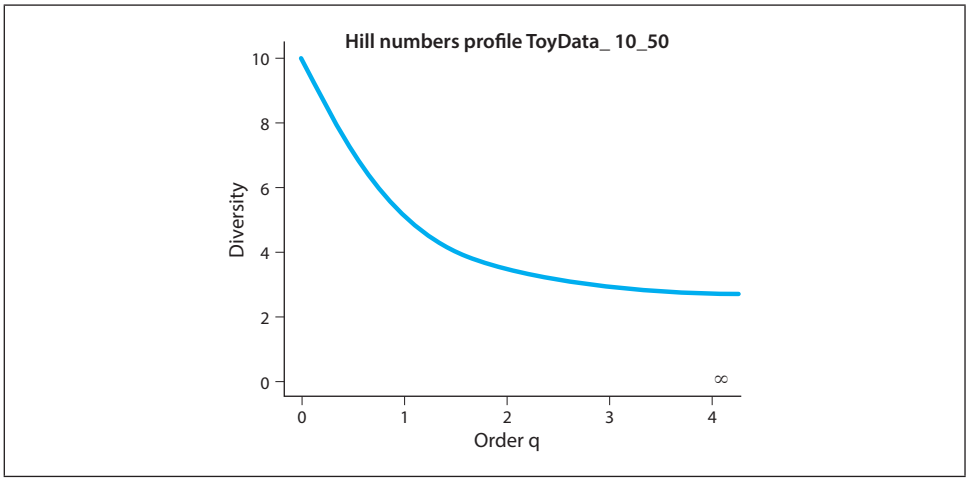


Figure A5.3
Hill numbers profile.

Montserrat plots

The terms in the sum of *Mfm* (13) may be visualized in the *Montserrat plot* where each bar in the plot gives the frequency of an haplotype in the VQS. The haplotypes are ordered, first by number of mutations respect to the dominant, and then by decreasing order of abundance within the number of mutations (Figure A5.4). *Mfm* results of the sum of the products of the number of mutations of each haplotype with respect to the dominant (abscissa) by the frequency of the corresponding haplotype (ordinate). Note that a given value of *Mfm* may be obtained from infinity different VQS – that is from infinite combinations of terms adding to the same value; the Montserrat plot distinguishes easily among these VQS and shows the structure of the quasispecies in terms of the abundance and genetic richness of its components.

Mutations plots

While the Montserrat plot shows the VQS structure by phenotype, the *mutations plot* (Figure A5.5) shows the VQS structure by genotype. On a rug representing the sequence of the dominant haplotype the different bars represent the mutations and their frequency. A distinctive colour is used for each nucleotide. Few abundant mutations may radically influence the values of π_m and *Mfm*. A lower genomic barrier to

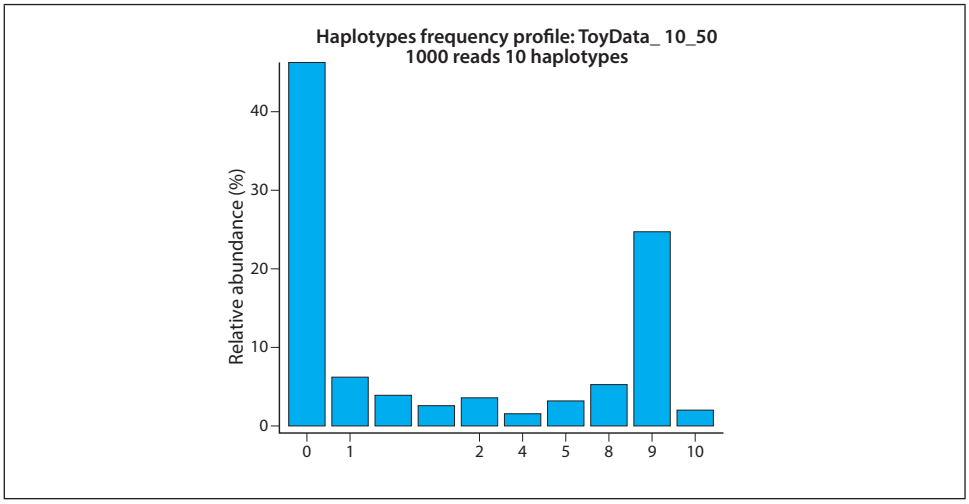


Figure A5.4
Montserrat plot of the Hpl10 QVS.

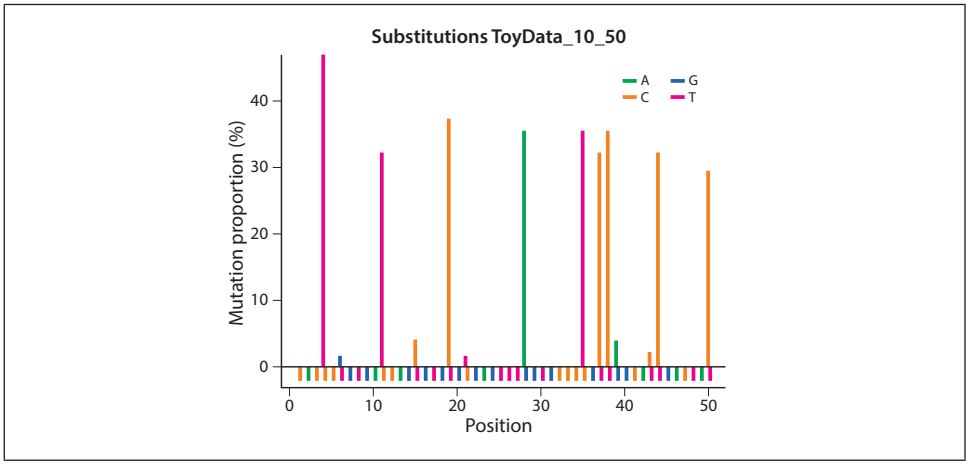


Figure A5.5
Mutations frequency by site.

resistant antiviral variants could be favoured instead by a high number of mutations at moderate or low level. Note that both cases may produce equivalent values of *Mfm* or π_m. The mutations plot shows the contribution of each mutation to *Mfm* and π_m.

Polymorphic sites profile

The *polymorphic sites profile* (Figure A5.6) depicts the number and importance of all polymorphic sites in the MAH. A progressively higher abundance threshold is applied and the polymorphic sites with mutation frequencies below the threshold are excluded. The number of polymorphic sites passing the filter is plotted vs the abundance threshold. As is the case with the Montserrat plot and the mutations plot, this graphic helps to visualize the contribution of each polymorphic site to *Mfm* and π_m.

FAD profiles

The *FAD profiles* contribute to visualize the structure of the matrix of pairwise genetic distances between haplotypes in the VQS. The first *FAD profile* (Figure A5.7) shows the fraction of pairs of haplotypes at each observed genetic distance.

The second profile (Figure A5.8) shows the aggregate contribution of all pairs of haplotypes at each observed genetic distance to the *FAD* value. That is, the sum of all elements in the genetic distance matrix whose value corresponds to each observed genetic distance divided by *FAD*.

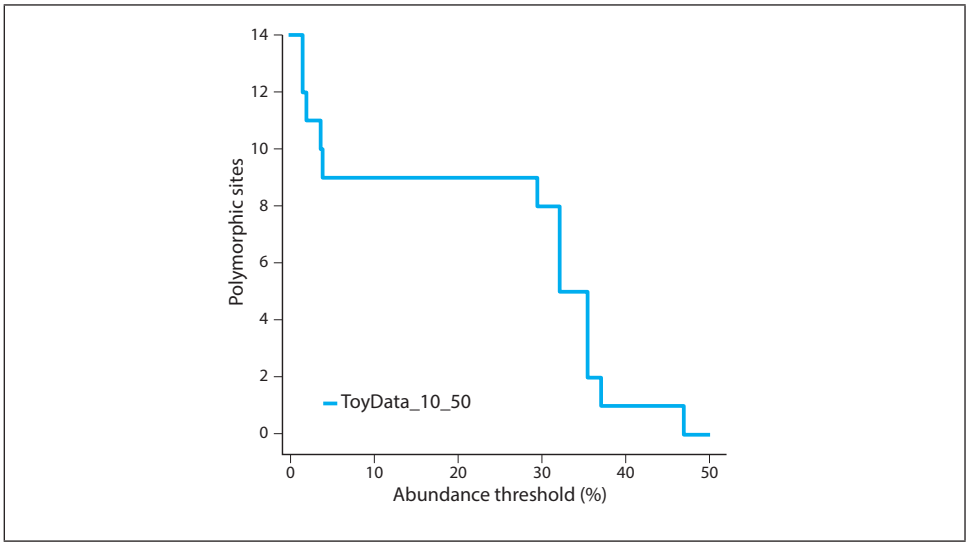


Figure A5.6
Profile showing the number of polymorphic sites as we increase the abundance threshold.

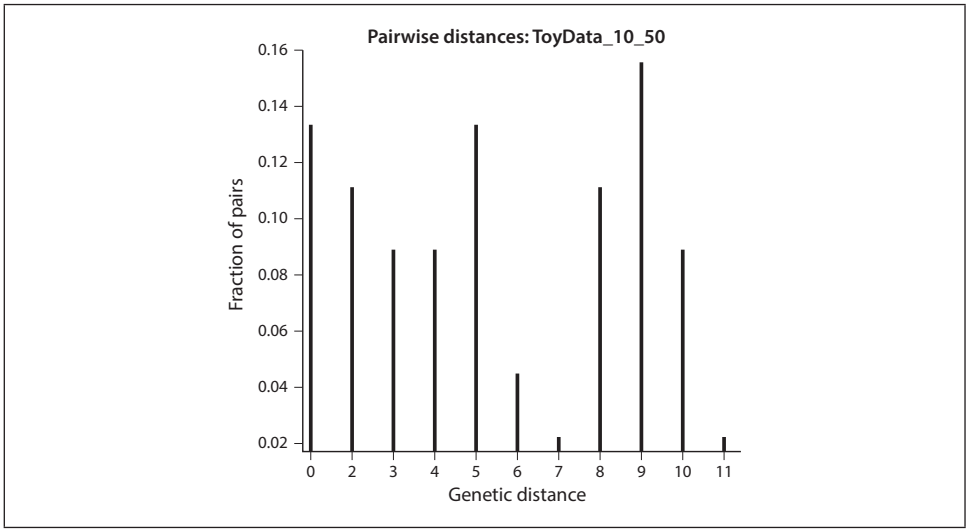


Figure A5.7
Fraction of pairs of haplotypes at each observed genetic distance.

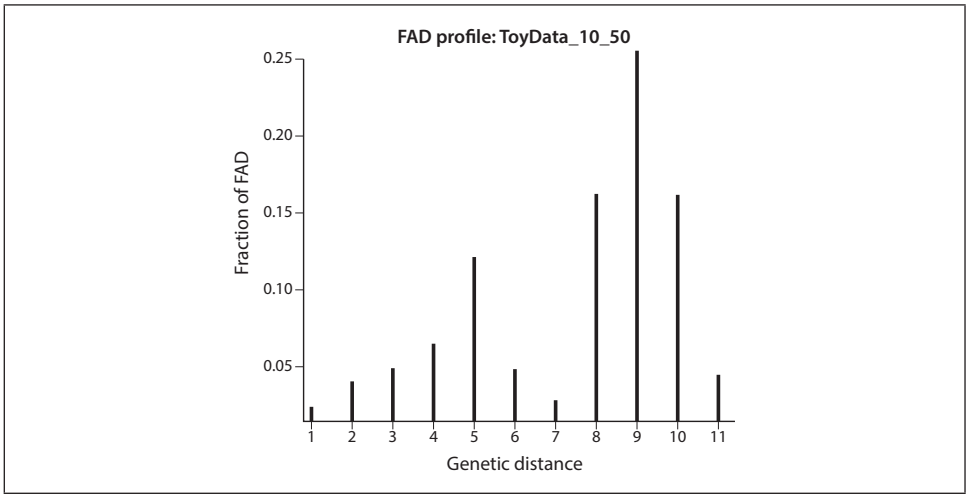


Figure A5.8
Contribution to the FAD of pairs of haplotypes at each observed genetic distance.

Nucleotide diversity profile

This profile (Figure A5.9) shows the aggregate contribution of all pairs of haplotypes at each observed genetic distance to π_m. These contributions are weighted by the frequency of the corresponding haplotypes as in equation (14).

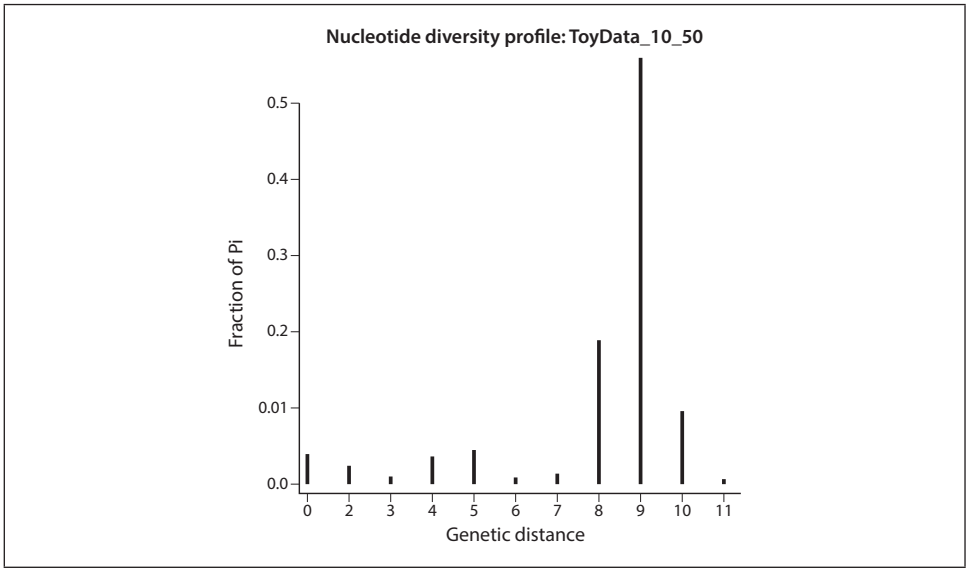


Figure A5.9
Contribution of all pairs of haplotypes at each observed genetic distance to the nucleotide diversity π_m.

Sampling quasispecies and bias

All what has been said so far considers that we have a perfect picture of the quasispecies composition, which is just an idealization. Sampling a few thousands of molecules out of a population of trillions will provide in the best of the cases a poor representation of the complexity of a quasispecies. Despite the big improvement provided by the current NGS technologies our sampling capabilities are still very modest. On the other hand a non negligible fraction of the NGS data will be discarded to limit errors and artifacts. Part of the rejected sequences are true quasispecies components which cannot be distinguished from the noise. The selection of an appropriate noise level is critical in the characterization of VQS complexity. A too low level will inflate diversity, while a too high level will too much smooth the estimates.

There are three major challenges to face. These are of clinical, technical and statistical nature. The first would be to clearly set the frequency above which a variant becomes clinically relevant, if given the dynamics of a quasispecies such threshold might exist. The second would be to increase the length and accuracy of amplicon sequencing. And this involves very high coverages, ensure that there is no primers bias, and keep the error rate as low as possible. The third would be to minimize the differential bias in the computation of VQS complexity of samples to be compared.

Given the affordable sample size, compared to the huge population we wish to study, the diversity indices presented above will give values downward biased respect

to the whole population. And this bias is a function of the sample size. With current sequencing methods, no absolute value of VQS diversity may be computed, just values tied to the sample size. This dependency is highest for all incidence-based indices, the higher the sample size, the higher the chances to observe a higher number of haplotypes, polymorphic sites, or mutations. It is strong for the Shannon entropy. And it is modest for the Gini-Simpson index, and the mutation frequency and the nucleotide diversity at the molecular level.[8 3 5 16]

The main problem arises when comparing two or more samples, even when assembling a table of VQS complexity values of a set of samples. When the sizes of these samples are not balanced enough these values should be corrected to make them comparable. That is to assure that all of them have the same level of bias. According to our experience with NGS, down-sampling all samples to the size of the smallest in the set, followed by haplotype fringe trimming at a given frequency threshold[5] balances the biases in all the samples, even with incidence-based diversity indices.

Balancing biases

The NGS data available for VQS complexity computations has passed through a series of data treatment steps involving usually an abundance filter; by which all haplotypes with abundances below a threshold considered the noise level are discarded. Given the properties of the multinomial distribution, the effect of such filter on the number of final haplotypes is sample-size dependent.

Table A5.10 and Figure A5.10 show the results of a simulation where samples of 2000 reads and 5000 reads have been repeatedly sampled form the same empirical HCV haplotype distribution obtained at a high depth (80,000 reads) from a high complexity amplicon in the NS3 region of a chronically infected patient.[5 6]

Table A5.10
Prominent statistics of the simulations shown in the boxplots of Figure A5.10. Computations after abundance filtering

Hill number	**^0D**		**^1D**		**^2D**	
Sample size	**2000**	**5000**	**2000**	**5000**	**2000**	**5000**
Mean	51	39.2	12.4	10.4	5.35	4.92
Min	37	30	9.56	8.97	4.47	4.31
5%	44	35	11	9.63	4.93	4.66
50%	51	39	12.3	10.4	5.34	4.92
95%	58	44	13.9	11.3	5.81	5.19
Max	65	51	15.3	12.2	6.35	5.58

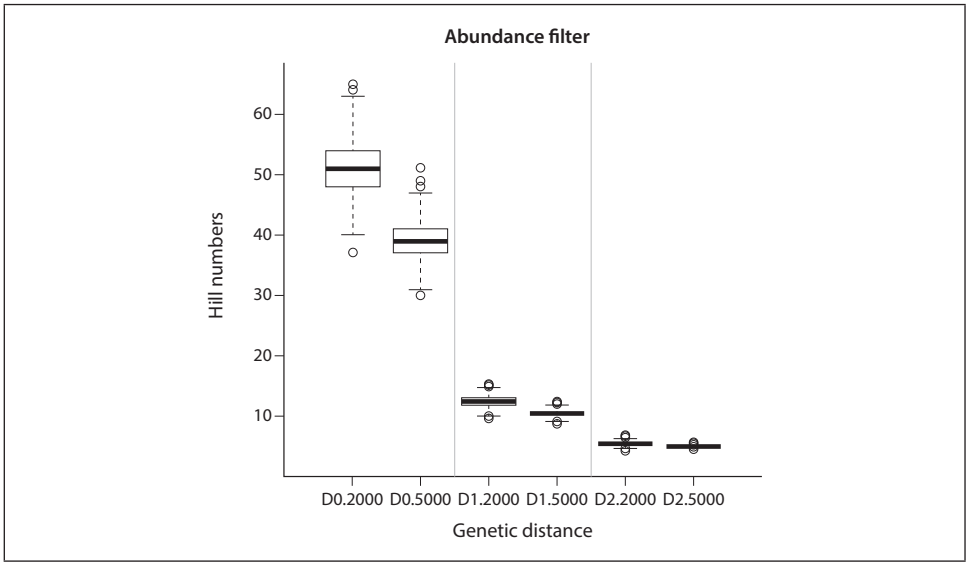


Figure A5.10
Boxplots with the distribution of Hill numbers obtained in 2000 simulations of pairs of samples of sizes 2000 and 5000 reads taken form an empirical distribution of an HCV amplicon sequenced at a very high depth (80,000 reads). Hill numbers have been computed after the abundance filter at noise level (0.2%). Abscissa labels distinguish between order of Hill numbers – D0, D1 and D2 – and between sample sizes – 2000 and 5000 reads.

The noise level is taken at 0.2% and all haplotypes with abundances below this threshold are excluded after sampling. The boxplots show the distribution of Hill numbers of order 0, 1 and 2 obtained in 2000 simulations, where a significant bias towards the small samples can be easily appreciated. The number of haplotypes, ${}^0D$, passing the filter is consistently higher for the samples of 2000 reads than for the samples of 5000 reads. The differential bias, as expected, diminishes as the order of the Hill number increases.

An algorithm which minimizes this differential bias is down-sampling followed by fringe trimming (DSFT) the bigger sample, and fringe trimming the smaller sample. This operation is done with the quality filtered NGS data but previous to abundance filtering. Down-sampling is done by simply rescaling the observed haplotype frequencies to the size of the smaller sample and rounding to the nearest integer. Fringe trimming is done by excluding all haplotypes with frequencies below a confidence level (0.95) given that are represented in the population at an abundance below or equal to the noise level. The threshold frequency is then the 0.95 quantile of a binomial distribution, B(n, p), with n equal to the size of the smaller sample, and p equal to the noise level. Table A5.11 and Figure A5.11 show the result of this algorithm on the simulations in Figure A5.10. At the cost of loosing information, both in the big and the small samples, the differential bias is minimized and the values of diversity become comparable.

Table A5.11
Prominent statistics of the simulations shown in the boxplots of Figure A5.11. Computations after DSFT

Hill number	**^0D**		**^1D**		**^2D**	
Sample size	**2000**	**5000**	**2000**	**5000**	**2000**	**5000**
Mean	22	21.4	7.97	7.8	4.38	4.32
Min	17	17	6.53	6.88	3.71	3.84
5%	19	20	7.18	7.3	4.04	4.11
50%	22	21	7.95	7.79	4.38	4.31
95%	25	23	8.79	8.31	4.73	4.54
Max	29	26	9.67	8.99	5.09	4.83

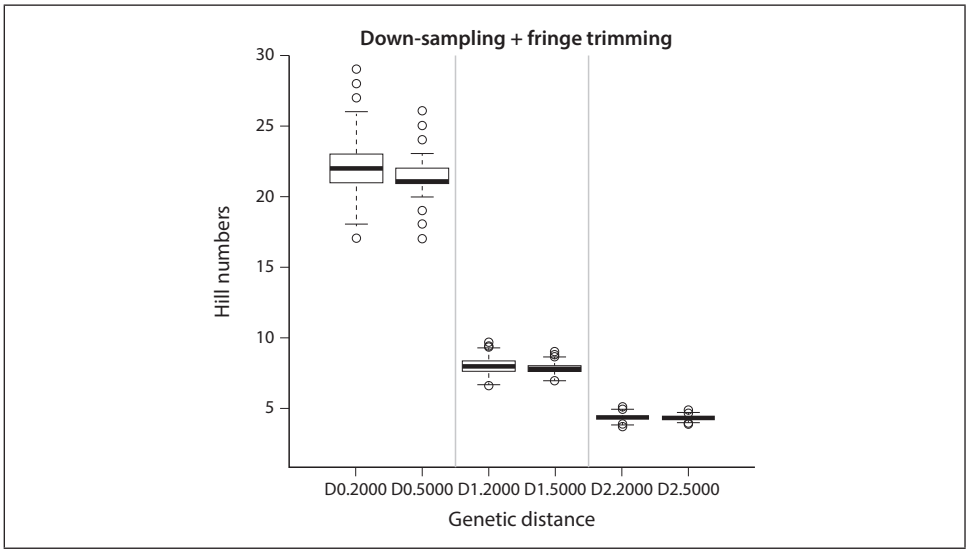


Figure A5.11
Boxplots with the distribution of Hill numbers obtained in 2000 simulations as in Figure A5.10. Hill numbers have been computed after DSFT at noise level (0.2%).

When computing VQS complexity measures, given the amount of true data excluded in the filters, it is recommended to take an abundance threshold below the real noise level. Provided that sequencing depth is high enough, half of it might be appropriate.

Which index to use?

Even though the tutorial has restricted the range of biodiversity indices available in the literature[3 9] to just very few, there is still the obvious question of which of them could be more appropriate or should be preferred with VQS. Each index shows just one aspect of VQS complexity, a variable which is necessarily multimensional. Each index could be seen as a summarization over a subspace of the VQS complexity. The main recommendations in this respect are as follows:

- Take a set of diversity indices including incidence, abundance and functional indices, to obtain a multidimensional representation of VQS complexity.
- Use Hill numbers of order 1 and 2 because they can be more informative that the corresponding H_S and H_{GS}, and are less affected by saturation.
- When comparing samples, all incidence-based indices including the Hill numbers of order below 2, all forms of Shannon entropy, and the functional incidence-based indices (*FAD*, *Mfe*, and π_e) should be appropriately corrected, as they are sensitive to sample size differences. On the other hand, the Gini-Simpson index, the Hill numbers of order above 2, and the functional abundance-based indices (*Mfm* and π_m) are less sensitive to rare haplotypes and more robust against sample size differences.
- Incidence-based indices are best indicated in a mutagenesis scenario; for example to characterize VQS subject to lethal mutagenesis. Abundance-based indices are strongly correlated with current haplotype fitness, and could be best indicated in any evolutionary scenario where fitness is a relevant parameter.
- Compare profile plots as well as the selected indices. A given index may show the same value even with divergent profiles.

As most indices are strongly correlated, a multidimensional approach[17] could be to take the first or first two principal components, in a principal components analysis (PCA) of the scaled matrix of selected diversity indices, to summarize a complexity measure. This contributes to reducing noise and to characterize VQS complexity in a data set by just one or two values. On the other hand a multidimensional scaling (MDS) of D could complement a phylogenetic analysis of a single sample. In a multisample scenario, the matrix of genetic distances among samples[4] could also be submitted to MDS to obtain a reduced dimension representation of all samples in the dataset.

Concluding remarks

The availability of several diversity indices is important to define viral quasispecies at the molecular level. There are at least four main reasons to quantify the complexity of viral quasispecies:[2]

1. complexity is one of the parameters that predict adaptability of viral quasispecies to complex environments;
2. mutant spectrum complexity is one of the factors identified as predictors of viral disease progression and response to treatments;
3. a reduction of mutant spectrum complexity in sequential viral samples alerts of important evolutionary events, particularly the occurrence of a sweeping selection episode or a population bottleneck;
4. an effective antiviral mutagen in a lethal mutagenesis design should produce an increase of mutant spectrum complexity, at least in a transient fashion.

The evaluation of virus population complexity for biological inferences has as one of its major complications that the diversity profile of a viral quasispecies is not a constant parameter. The generation of a new haplotype is subjected to the uncertainties of mutant generation, and relative abundance of any new haplotype is influenced by past and present fitness levels of the relevant genomes, in interaction with other members of the mutant swarm. Consecutive expansions and contractions of diversity may be observed. A contraction may occur when a new haplotype with much higher fitness than those currently dominant emerges, eventually resulting in a substantial increase in viral load despite a transient reduction of complexity. A change in relative fitness among haplotypes may be due to a change in the environment, to the production of a new superior mutant, or both. Eventually a nearly stationary state with very high diversity may be reached, provided the environment does not change and the haplotype repertoire approaches optimal fitness for the environmental requirements.

The indices and methods discussed in the previous sections should allow the characterization of a quasispecies in terms of its complexity. The different proposed plots could also help in the interpretation of complexity profiles and facilitate the comparison of sequential samples. We hope that the systematization introduced (Figure A5.2) might help to better understand the type of information provided by each index, and guide the problem-specific selection of the set of most adequate indices in each case.

REFERENCES

1. Domingo E, Parrish CR, Holland JJ, eds. (2008). Origin and evolution of viruses. 2nd ed. London: Academic Press, Elsevier.
2. Gregori J, Perales C, Rodríguez-Frías F et al. (2016). Viral quasispecies complexity measures. Virology 493:227-37.
3. Magurran AE (2004). Measuring biological diversity. Oxford: Blackwell Science.
4. Nei M (1987). Molecular evolutionary genetics. NY: Columbia University Press.
5. Gregori J, Salicrú M, Domingo E (2014). Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics 30:1104-11.
6. Cubero M, Gregori J, Esteban JI et al. (2014). Identifcation of host and viral factors involved in a dissimilar resolution of hepatitis C virus infection. Liver Int 34:896-906.
7. Shannon CE (1948). A mathematical theory of communication. Bell System Technical Journal 27:379-423; 623-56.
8. Hutcheson K (1970). A test comparing diversities based on the Shannon formula. J Theor Biol 29:151-4.

9. Magurran AE, McGill BJ, eds. (2011). Biological diversity, frontiers in measurement and assessment. Oxford: Oxford University Press.

10. Hill M (1973). Diversity and evenness: a unifying notation and its consequences. Ecology 54:427-32.

11. Jost L (2006). Entropy and diversity. OIKOS 113(2):363-75.

12. Yang Z (2008). Computational molecular evolution. Oxford: Oxford Series in Ecology and Evolution.

13. Felsenstein J (2004). Inferring phylogenies. Suderlan (MA): Sinauer Associates Inc.

14. Walker B, Kinzig A, Langridge J (1999). Plant attribute diversity, resilience, and ecosystem function: the nature and significance of dominant and minor species. Ecosystems 2:95-113.

15. Rao CR (1982). Diversity and dissimilarity coefficients: a unified approach. Theor Pop Biol 21:24-43.

16. Chao A, Gotelli NJ, Hsieh T et al. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecological Monographs 81:45-67.

17. Rencher AC (2002). Methods of multivariate analysis. 2nd ed. New York: John Wiley & Sons.

Section 4

Quantifying mutagenesis: rare haplotype load

Abstract

RNA viruses replicate with a template-copying fidelity that lies close to their extinction threshold. Increases in the mutation rate due to the effect of nucleotide analogue treatments can drive viruses to extinction. This transition is the basis of an antiviral strategy known as lethal mutagenesis. In this study, we introduce a new diversity index, the rare haplotype load (RHL), to describe NS5B (polymerase) mutant spectra of hepatitis C virus populations passaged in the absence or presence of the mutagenic agents favipiravir or ribavirin. We found that the RHL increase is more prominent in mutant spectra in which expansion was due to the action of nucleotide analogues than in those where expansion resulted from multiple passages in the absence of these mutagens. Statistical tests for paired mutagenized versus non-mutagenized samples with 14 diversity indices showed that RHL consistently provides the highest standardized effect of the difference caused by mutagenic treatment with ribavirin or favipiravir. The results indicate that enrichment of the viral quasispecies with very low frequency minority genomes can serve as a robust marker for lethal mutagenesis. The diagnostic value of RHL determination from deep sequencing data is relevant for experimental studies on enhanced viral mutagenesis and for pharmacological evaluations of inhibitors suspected to have mutagenic activity.

Highlights

- RHL is a reliable index to diagnose expansion of a mutant spectrum associated with mutagenic treatment.
- RHL offers the means to determine whether base or nucleoside analogues showing antiviral activity affect viral RNA replication by direct inhibition of polymerase function or by enhanced mutagenesis, thus contributing to clarify uncertainties about the action of a drug.
- Regarding the strength of correlations between the diversity indices and mutagenesis, the RHL is followed by incidence-based indices, such as the number of haplotypes, number of mutations, number of polymorphic sites, and the functional attribute diversity (FAD).
- Although mutation frequency and nucleotide diversity are widely used to describe mutant spectra, these indices exhibit a poor correlation with mutagenesis treatment.
- RHL was statistically more robust than other correlated diversity indices, was unbiased, and was scarcely sensitive to sample size within the limits of the study.

Note

The study describing this experiment on HCV mutagenesis showed that the classical function-related indices, mutation frequency and nucleotide diversity, were unable to explain the diversity introduced by mutagenic treatment. The incidence indices were the most sensitive to treatment effects, but they are highly sensitive to sample size. Our attention then turned to a different system of characterization, the aggregation of molecules with sequences present at low abundance levels. This method is further elaborated in the next article (Section 5), in which a quasispecies fitness partition into four fractions (QFF) is proposed. The QFF is contemplated as a summary of haplotype distribution in the quasispecies, with biological meaning. Finally, the last article in the series, Section 6, follows the evolution of an in-host quasispecies by measuring the dissimilarity (distance) between haplotype distributions in serial samples.

4. *Rare haplotype load as marker for lethal mutagenesis*

JOSEP GREGORI, MARÍA EUGENIA SORIA, ISABEL GALLEGO, MERCEDES GUERRERO-MURILLO, JUAN I. ESTEBAN, JOSEP QUER, CELIA PERALES, ESTEBAN DOMINGO

ABSTRACT

RNA viruses replicate with a template-copying fidelity, which lies close to an extinction threshold. Increases of mutation rate by nucleotide analogues can drive viruses towards extinction. This transition is the basis of an antiviral strategy termed lethal mutagenesis. We have introduced a new diversity index, the rare haplotype load (RHL), to describe NS5B (polymerase) mutant spectra of hepatitis C virus (HCV) populations passaged in absence or presence of the mutagenic agents favipiravir or ribavirin. The increase in RHL is more prominent in mutant spectra whose expansions were due to nucleotide analogues than to multiple passages in absence of mutagens. Statistical tests for paired mutagenized versus non-mutagenized samples with 14 diversity indices show that RHL provides consistently the highest standardized effect of mutagenic treatment difference for ribavirin and favipiravir. The results indicate that the enrichment of viral quasispecies in very low frequency minority genomes can serve as a robust marker for lethal mutagenesis. The diagnostic value of RHL from deep sequencing data is relevant to experimental studies on enhanced mutagenesis of viruses, and to pharmacological evaluations of inhibitors suspected to have a mutagenic activity.

Introduction

The mutant spectra of RNA viruses are a reflection of their evolutionary history, as well as important determinants of virus adaptability. Concerning control of viral diseases, mutant spectrum dynamics is an obstacle for the efficacy of therapeutic interventions due to selection of treatment-escape viral mutants. The antiviral agents to combat RNA viruses include those directed to specific viral targets [direct-acting antiviral agents (DAAs)], and those that inhibit cellular functions needed for the completion of the virus life cycle. The viral RNA-dependent RNA polymerase (RdRp) is the target of several effective antiviral agents. Some of them, notably base or nucleoside analogues, are intracellularly converted into their active nucleotide counterparts. The

discovery that ribavirin (1-β-D-ribofuranosyl-1-*H*-1,2,4-triazole-3-carboxamide) is mutagenic for poliovirus [1] introduced a new perspective in the antiviral mechanism of some nucleotide analogues. Three alternative – not mutually exclusive – mechanisms of anti-RdRp activity by nucleotide analogues have been described: RNA chain termination, inhibition of RNA synthesis without chain termination, and inhibition associated with viral genome mutagenesis.

Nucleotide analogue-induced mutagenesis is equivalent to a decrease of copying fidelity by the viral RdRp. Quasispecies theory predicts a maximum amount of genetic information that can be transmitted for a given average copying fidelity. This concept is mathematically formulated in the form of an error threshold relationship. An increase in mutation rate drives the population across the error threshold, into error catastrophe, equated with loss of inheritable information [2, 3]. The error threshold applies to finite populations in variable fitness landscapes, and its position in a fidelity scale depends also on the degree of adaptation of the mutant ensemble to the environment [3]. The error threshold concept has found experimental support in studies on the negative effects of chemical mutagenesis on the survival of RNA viruses ([4–8], among other studies). The convergence of theoretical and experimental results opened the way to lethal mutagenesis as an antiviral strategy [9].

The licensed antiviral nucleoside analogues favipiravir (T-705; 6-fluoro-3-hydroxy-2-pirazinecarboxamide) and ribavirin are mutagenic for several RNA viruses. Ribavirin has been used as antiviral agent for decades [10, 11], and only recently shown to be mutagenic for several RNA viruses [1, 12, 13]. Favipiravir has been licensed as an anti-influenza agent in Japan having potent antiviral activity against different influenza virus strains (types A, B and C) including those resistant to neuraminidase and M2 inhibitors [14]. Favipiravir has also been effective to inhibits the replication of other RNA viruses in vitro and in animal models, including flavi-, noro-, alpha-, bunya-, arena-, filovirus and other RNA virus for which no antiviral therapy is currently available [reviewed in (15, 16)]. Favipiravir is converted intracellularly into the ribofuranosyl 5'-triphosphate metabolite (favipiravir-RTP) and is this form it can be recognized as a pseudopurine by the RdRp [14, 17]. Its selective inhibition of RdRp implicates a wider anti-viral spectrum with a limited cell damage compared with other mutagens such as ribavirin.

Incorporation of favipiravir-RTP in the nascent viral RNA could result in lethal mutagenesis, as has been proposed for influenza virus [18], norovirus [7], hepatitis C virus (HCV) [19], foot-and-mouth disease virus [20], West Nile virus [21], Dengue virus [22] and Ebola virus [8], coxsackievirus B3 [23]. It is not clear whether favipiravir acts as RNA chain terminator, inhibitor, mutagen or by a combination of these mechanisms; its dominant mode of action may depend on the virus-host system and concentration of the active form. In order to cause lethal mutagenesis, favipiravir-RTP needs to be incorporated into the RNA without causing immediate chain termination. It is possible that both lethal mutagenesis and chain termination occur depending on the available concentration of favipiravir-RTP. It has been hypothesized that incorporation of low

levels of favipiravir-RTP could result in full-length extension of the viral RNA, leading to lethal mutagenesis and lower infectivity [24].

The standard way to distinguish an RNA virus mutagen from a non-mutagenic inhibitor, is that a mutagenic inhibitor promotes an increase of mutant spectrum diversity, and a decrease of the virus specific infectivity (defined as the ratio of the number of infectious units to the amount of viral RNA in the virus population) (reviewed in [25, 26]). The application of deep sequencing to the analysis of viral populations has introduced several new diversity indices that allow a more detailed description of mutant spectra [27–29]. Some diversity indices were adopted from ecology, and are classified in three groups: incidence (based on the count of entities in a multiple alignment of haplotypes), abundance (that considers both, counts of entities and their frequency), and functional (based on differences among the observed haplotypes) [27]. The value of alternative diversity indices to diagnose the mechanism underlying the expansion of mutant spectra is an unsolved issue.

We have adapted a HCV serial passage design to study the genetic and phenotypic diversification of HCV in Huh-7.5 reporter cells in absence of cellular evolution [30–32]. The parental (plasmid-derived) HCV population was passaged in absence or presence of ribavirin or favipiravir. Populations whose mutant spectrum was expanded in absence of drugs were also subjected to mutagenesis. The design produced several HCV populations for comparative mutant spectrum analyses. NS5B amplicons were analyzed to quantify mutant spectrum complexity. We describe a new diversity index, the rare haplotype load (RHL), and show that its variation outstands among that of other diversity indices to characterize mutant spectra in their transition into error catastrophe. RHL may help in the understanding of quasispecies dynamics, and in the clarification of the mechanisms of action of antiviral agents.

Results

The rare haplotype load of hepatitis C virus populations

HCV RNA expressed from plasmid Jc1FLAG2(p7-nsGluc2A) (genotype 2a) [33] was transfected into Huh-7 Lunet cells and amplified in Huh-7.5 cells to produce the initial virus population HCV p0 [30]. HCV p0 was subjected to 200 serial passages in Huh-7.5 reporter cells in the absence of any drug. The populations at passage 100 (HCV p100) and at passage 200 (HCV p200) displayed increased replication in Huh-7.5 reported cells [31, 32]. HCV p0, HCV p100 and HCV p200 were further passaged either in the absence of any drug or in the presence of favipiravir or ribavirin (Fig 1A). Infectious progeny levels were those expected from previous quantifications of inhibition of HCV p0 by favipiravir [19] and ribavirin [34]; the sustained HCV p100 and HCV p200 production in the presence of the drugs is expected from the fitness-associated HCV resistance to antiviral agents [31, 32, 35] (Fig 1B).

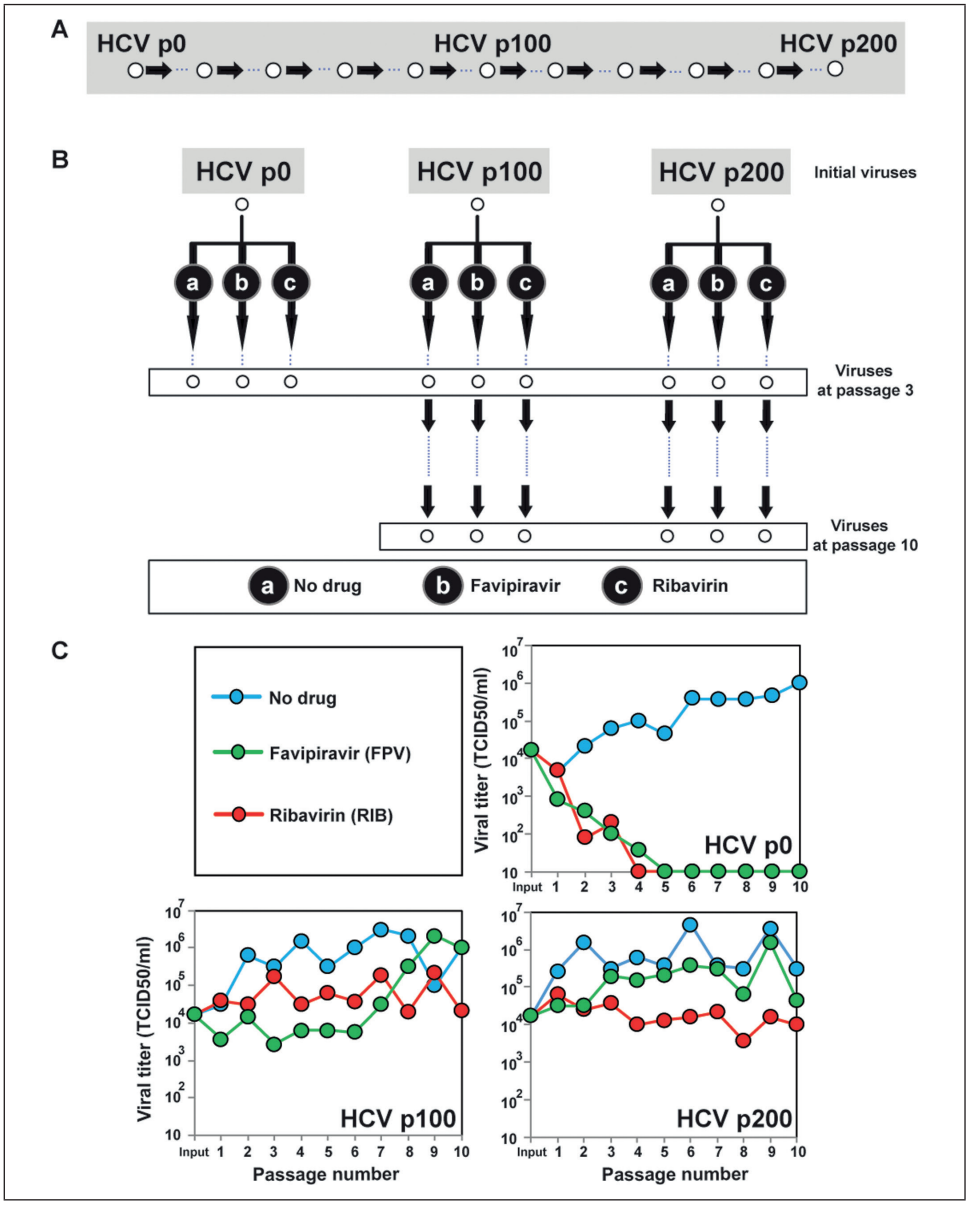


Fig 1. Experimental design and infectious HCV progeny production in absence or presence of favipiravir or ribavirin.
(**A**) Passage of HCV p0 in Huh-7.5 reporter cells to derive high fitness HCV p100 and HCV p200 populations. (**B**) Serial passages of HCV p0, HCV p100 and HCV p200 in absence of drugs (No drug) or the presence of 400 μM favipiravir, or 100 μM ribavirin. (**C**) Infectious progeny production during 10 serial passages in absence or presence of the drugs. Details on the infections are given in "Materials and methods".
https://doi.org/10.1371/journal.pone.0204877.g001

Intracellular viral RNA was sequenced in MiSeq™ with 2x300 mode with v3 chemistry, and fastq files were analyzed as previously described [27, 29] to obtain forward and reverse consensus haplotypes with abundances not below 0.1%, median coverage 147,000 reads, interquartile range (IQR) 75570–226100. The fasta files obtained for each sample were further subjected to DSFT for diversity indices computation. The resulting median coverage was of 139200 with IQR 71,480–210,600 reads.

We introduce the rare haplotype load (RHL) as a new diversity index which may be considered intermediate between incidence and abundance indices. In the context of this work we define as rare those haplotypes with abundances below a given threshold (1%), and as load the fraction of molecules in the quasispecies belonging to these haplotypes. Translating this concept to next generation sequencing is not an easy task as we wish to take into account full reads in the range of abundances below the instrument noise level. Our approach consisted in taking all reads corresponding to haplotypes common to the forward and reverse strand with no previous abundance filtering and computing the RHL as the fraction of reads belonging to haplotypes with abundances below 1%. We suppose that technical noise affects equally all samples in the experiment and that the distinctive effect would be caused by the treatment. This index was not submitted to any sample size correction.

Comparison of RHL with other diversity indices

A full set of diversity indices was computed for each sample: Hpl, number of haplotypes; PolySites, number of polymorphic sites; nMuts, number of mutations; Shannon, Shannon entropy; GiniS, Gini-Simpson index; q1D, Hill number of order 1; q2D, Hill number of order 2; qInfD, Hill number of order infinity; FAD, functional attribute diversity; Mf.e, mutation frequency by entity; Pi.e, nucleotide diversity by entity; Pi, nucleotide diversity, and Mf, mutation frequency. The correlation among all indices (including RHL) resulting from all samples in the experimental design, shows a structure with three groups (Fig 2): G1: RHL, Hpl, FAD, nMuts and PolySites; G2: Shannon, GiniS, q1D, q2D and qInfD; and G3: Mf.e, Pi.e, Mf and Pi. The three indices more correlated to RHL are Hpl 0.895, FAD 0.854 and Shannon 0.826. RHL falls within the group of incidence indices, but it results from the aggregation of abundances. Its high correlation with Shannon and q1D denotes properties of abundance-based indices, while its high correlation with FAD confers to RHL properties of functional incidence. These three properties were expected from the definition of RHL, and the correlations provide an empirical prove of the computations adequacy.

Association tests

Table 1 summarizes the results of the association tests of each diversity index considered, including RHL, to mutagenicity, for each drug (favipiravir and ribavirin) sorted by decreasing order of standardized effect. No distinction has been made of am-

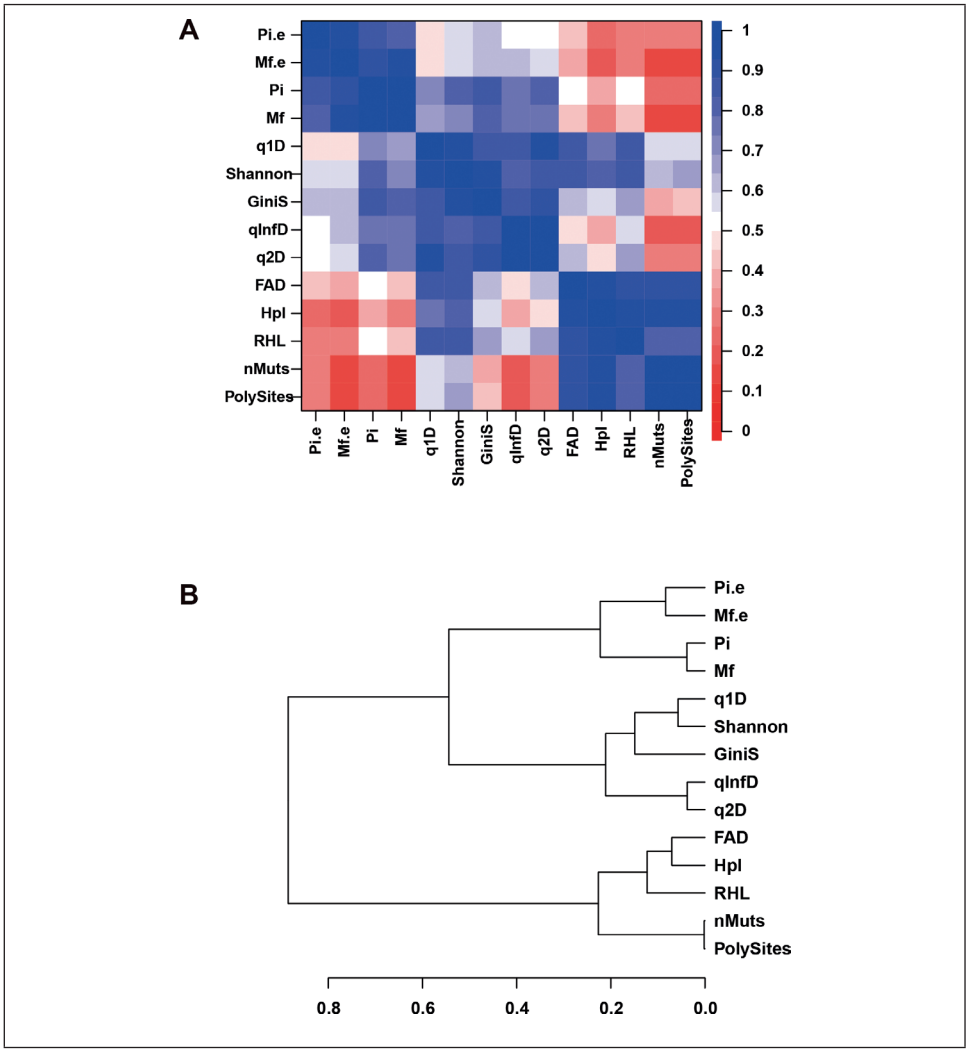


Fig 2. Correlation among diversity.
(**A**) Plot illustrating the correlation between diversity indices in this study. The correspondences between colors and correlation values is shown on the right bar. (**B**) Hierarchical clustering of diversity indices computed with the square root of one minus the correlation matrix, as measure of dissimilarity.
https://doi.org/10.1371/journal.pone.0204877.g002

plicon or treatment length. Hence the results represent averaging over the full NS5B region sequenced, and over the two treatment lengths. The RHL is the index with the highest standardized effect among all in both mutagenic treatments, with adjusted p-values of the order of 10^{-4}. Top indices are also FAD, Hpl, Shannon, nMuts, and PolySites. No significant association is found in neither treatment, for Mf and Pi, both at

Table 1. Wilcoxon signed rank test of paired samples, treatment vs control.
Rows sorted in descending order of standardized effect. Estimate: median of treatment difference. SD: standard deviation. StdEffect: standardized effect (median divided by sd). p.value: Wilcoxon test p-value. adj.pv: multiple test adjusted p-value by the Bonferroni method.
https://doi.org/10.1371/journal.pone.0204877.t001

	Ribavirin						Favipiravir				
	Estimate	SD	StdEffect	p.value	adj.pv		Estimate	SD	StdEffect	p.value	adj.pv
RHL	1.49E+01	5.09E+00	2.920	3.05E-05	4.27E-04	**RHL**	1.81E+01	3.86E+00	4.680	3.05E-05	4.27E-04
nMuts	2.05E+01	1.00E+01	2.040	3.58E-04	5.01E-03	**Shannon**	8.76E-01	4.22E-01	2.080	3.05E-05	4.27E-04
Hpl	2.25E+01	1.11E+01	2.020	3.61E-04	5.06E-03	**FAD**	1.74E+01	9.33E+00	1.860	6.10E-05	8.54E-04
PolySites	2.00E+01	9.92E+00	2.010	3.58E-04	5.01E-03	**Hpl**	2.75E+01	1.54E+01	1.790	5.45E-04	7.62E-03
FAD	1.21E+01	6.55E+00	1.850	3.05E-05	4.27E-04	**GiniS**	2.15E-01	1.39E-01	1.550	2.14E-04	2.99E-03
Shannon	6.99E-01	3.89E-01	1.800	3.05E-05	4.27E-04	**q1D**	5.18E+00	3.33E+00	1.550	3.05E-05	4.27E-04
GiniS	1.62E-01	1.27E-01	1.280	3.05E-04	4.27E-03	**nMuts**	2.60E+01	1.77E+01	1.470	1.59E-03	2.22E-02
q1D	4.04E+00	3.95E+00	1.020	3.05E-05	4.27E-04	**PolySites**	2.40E+01	1.71E+01	1.400	1.18E-03	1.65E-02
qInfD	7.74E-01	7.64E-01	1.010	5.80E-04	8.12E-03	**q2D**	1.56E+00	1.74E+00	0.901	3.05E-04	4.27E-03
q2D	1.58E+00	2.06E+00	0.766	3.05E-04	4.27E-03	**qInfD**	6.71E-01	9.05E-01	0.741	1.31E-03	1.84E-02
Mf.e	6.52E-04	1.55E-03	0.420	5.35E-02	7.49E-01	**Pi**	1.22E-03	2.04E-03	0.596	1.77E-02	2.47E-01
Pi	5.44E-04	1.34E-03	0.407	4.73E-02	6.62E-01	**Pi.e**	5.73E-04	1.40E-03	0.410	8.44E-02	1.00E+00
Mf	5.38E-04	1.37E-03	0.393	2.77E-02	3.88E-01	**Mf**	7.04E-04	2.32E-03	0.304	3.19E-02	4.46E-01
Pi.e	3.32E-04	1.17E-03	0.284	1.51E-01	1.00E+00	**Mf.e**	3.85E-04	1.72E-03	0.223	2.11E-01	1.00E+00

the abundance and at the entity level. RHL is still the top indicator when distinguishing among amplicons or treatment length, followed by FAD and Hpl.

Logistic regression

As a further step towards characterizing RHL as mutagenicity marker, samples were relabeled as under mutagenicity (Mut) if treated with favipiravir or ribavirin, or not mutagenized (control) for passages in absence of drug. Each group includes the variability due to the factors amplicon, drug, and treatment length. Then a univariate logistic regression was fit with each of the diversity indices, including RHL (Figs 3 and S1). The regression to RHL resulted in the best value of the Aikake Information Content (AIC), the highest area under the ROC curve (AUC), and the lowest leave-one out cross-validation (LOOCV) error rate to classification (Table 2). According to the LOOCV error rate the predictive capacity of these indices follows the order RHL, Hpl, PolySites, nMuts and FAD (S2 Fig).

When this analysis was performed separately for each treatment length, for 10 passes both RHL, Shannon and q1D resulted in the lowest values of AIC and LOOCV, for 3 passes RHL, FAD and Hpl resulted in the lowest values of AIC and LOOCV (S3

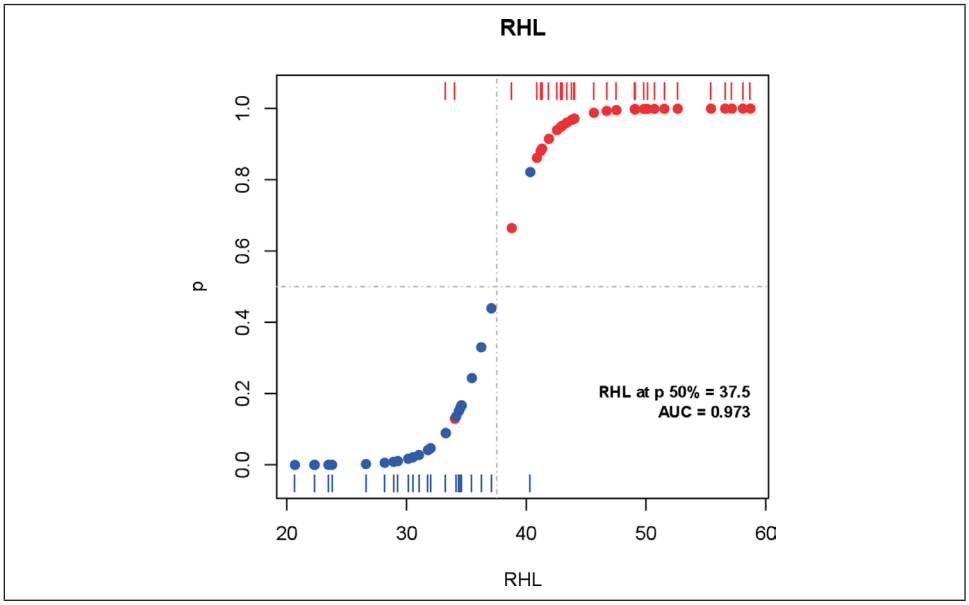


Fig 3. Logistic regression plot of mutagenicity over RHL.
Red bars at the top depict the values of RHL for HCV samples subject to mutagenesis. Blue bars at the bottom depict RHL values for control samples. Corresponding blue and red dots on the fitted logistic curve show the predicted probability of mutagenesis for each sample, with a predicted 50% probability at a RHL value of 37.5%. Area under the ROC curve 0.973. Only two samples under mutagenic drug treatment and one control sample were mistakenly classified.
https://doi.org/10.1371/journal.pone.0204877.g003

and S4 Figs). No multivariate logistic regression model resulted better that RHL as a single predictor in the whole dataset.

These results prove that within the experimental design of this study RHL is the most sensitive diversity index for predicting mutagenic effects, with independence of factors such as amplicon, base line passage, drug, and treatment length.

In-silico study to test the robustness and unbiasedness of RHL

To study the robustness of RHL and its possible dependence on sample size, an in-silico study on the full set of fasta files with no exclusion was performed. Robustness was evaluated by comparing RHL with the number of haplotypes, Hpl. According to the main statistics of the distribution of median values obtained after 2000 simulation cycles on each fasta file, RHL is on average five times less variable than Hpl, in terms of the coefficient of variation (CV) and of interquartile dispersion (QD) (Table 3). The main statistics of the relative error of the median value of RHL in the sample replicates indicate a median error of 0.04% with a maximum of 1.99%. This confirms that the observed value of RHL in a sample is unbiased and not influenced by the sample size. No sample

Table 2. Results of the logistic regressions.
Dev: Residual deviance. AIC: Aikake Information Content. Sensit: Sensitivity. Specif: Specificity. AUC: area under the ROC curve. Err: Classification error rate. TenFoldCV: Ten fold cross-validation error rate. LOOCV: Leave-one-out cross-validation error rate.
https://doi.org/10.1371/journal.pone.0204877.t002

Index	Dev	AIC	Sensit	Specif	AUC	Err	TenFoldCV	LOOCV
RHL	19.6	23.6	0.933	0.958	0.973	0.056	0.057	0.055
Hpl	24.2	28.2	0.867	0.917	0.968	0.111	0.123	0.111
PolySites	44.2	48.2	0.800	0.875	0.899	0.167	0.176	0.165
nMuts	45.3	49.3	0.800	0.875	0.897	0.167	0.176	0.172
FAD	36.5	40.5	0.867	0.792	0.928	0.167	0.182	0.181
Shannon	49.7	53.7	0.833	0.750	0.863	0.204	0.208	0.204
q1D	50.4	54.4	0.800	0.792	0.863	0.204	0.215	0.221
q2D	62.5	66.5	0.633	0.833	0.738	0.278	0.290	0.278
qInfD	62.8	66.8	0.633	0.833	0.724	0.278	0.298	0.312
GiniS	62.4	66.4	0.767	0.542	0.738	0.333	0.345	0.367
Pi.e	74.1	78.1	1.000	0.000	0.557	0.444	0.485	0.465
Pi	72.3	76.3	0.700	0.292	0.597	0.482	0.539	0.516
Mf	71.4	75.4	0.633	0.417	0.589	0.463	0.520	0.529
Mf.e	73.4	77.4	0.933	0.208	0.562	0.389	0.494	0.571

Table 3. Results of the in-silico study to characterize the biasedness and robustness of RHL with respect to Hpl.
The table provides a summary of distributional values for the % error with respect to the true value of RHL (top row); the coefficient of variation (CV) observed for RHL and Hpl, rows second and third; the ratio of both CVs in row fourth; the interquartile deviation (QD) observed for RHL and Hpl, rows fifth and sixth; and the ratio of both QDs in the last row.
https://doi.org/10.1371/journal.pone.0204877.t003

	Min.	Q1	Median	Mean	Q3	Max.
RHL.err %	0.00	0.02	0.04	0.25	0.32	1.99
RHL.cv	0.0048	0.0077	0.0098	0.0109	0.0134	0.0211
Hpl.cv	0.0279	0.0399	0.0443	0.0505	0.0548	0.1086
CV.ratio	2.1	3.2	4.7	5.3	6.9	13.1
RHL.qd	0.0057	0.0097	0.0120	0.0147	0.0184	0.0378
Hpl.qd	0.0000	0.0488	0.0606	0.0637	0.0741	0.1538
QD.ratio	0.0	2.9	5.1	5.2	7.0	13.7

size correction was needed for RHL, contrary to most diversity indices [27, 29]. Thus, RHL is far more stable and robust than Hpl and other diversity indices to characterize a virus transition into error catastrophe with deep sequencing data.

Discussion

Comparison of diversity indices for complexity evaluation of mutant spectra of HCV populations has unveiled that RHL is a reliable index to diagnose mutant spectrum expansions associated with a mutagenic treatment. Previous studies with HCV quantified average 5.1-fold (range 3.8–6.6) increases of mutation frequency following 200 serial large population passages in Huh-7.5 cells, and 3.5-fold (range 1.6–5.6) increases as a result of up to 5 serial passages in the presence of favipiravir or ribavirin [19, 32, 34]. Despite comparable or even larger mutation frequency increases associated with multiple passages compared with mutagenic treatments, RHL stood as a reliable, robust and unbiased marker for lethal mutagenesis. RHL is less influenced by standard serial passages.

A salient informative role of RHL can be interpreted in the light of current evidence of the molecular mechanisms that underlie the transition of RNA viruses towards an extinction threshold. One event is suppression of viable genome replication by defective genomes that are produced as a result of mutagenesis [36–38]. This is an extension to mutagenized populations of the capacity of mutant spectra to suppress replication of high fitness cognate populations [39]. Mutant-dependent interference was formulated as the lethal defection model of virus extinction [40] that has as one of its consequences that during mutagenesis infectivity is lost earlier than the capacity of viral RNA to replicate, thus leading to decrease of specific infectivity [19, 20, 34, 41, 42]. Examination of individual biological clones of viruses that remain viable amidst a mutagenic treatment evidenced 200-fold reduction in infectivity, with 8-fold increase in mutation frequency [43]. Therefore, the high RHL value in mutagenized populations is likely to reflect a fundamental property of mutant spectra subjected to continuous mutagenesis in which many low fitness genomes are generated. Such genomes, because of the continuous input of new mutations, do not have the opportunity of fitness recovery thus replenishing a low fitness sub-swarm captured by the RHL value. An increase in the proportion of minority (low frequency) mutations has been observed in lethal mutagenesis experiments both in cell culture [19, 20] and *in vivo* [8]. In mutant spectrum expansions that occur under basal mutation rate, RHL is expected to be less abundant because no enhanced mutagenesis jeopardizes opportunities for fitness gain, a tendency documented for RNA viruses when allowed unrestricted replication in a constant environment [31, 32, 44, 45].

Regarding diversity index adequacy to characterize lethal mutagenesis, RHL is followed by the highly correlated incidence-based indices RHL, Hpl, nMuts, PolySites and FAD, the latter probably because its entity level quality prevails under the con-

ditions of our study. In contrast, Mf and Pi, despite being widely used in the description of mutant spectra, exhibit poor correlation with mutagenesis treatment. We also examined by logistic regression the capacity of each index to discriminate between a history of mutagenesis and non-mutagenesis accompanying a mutant spectrum expansion. Sorting of indices by LOOCV error rate placed RHL on top, followed by HpI, PolySites, nMuts, FAD and Shannon. No discriminating capacity is observed for Mf, Pi, Mf.e and Pi.e, in agreement with the poor results of these indices in the association tests. The performed logistic regression has aimed at a more complex scenario, recognizing a mutagenic state independently of population history, where the signal could be blurred and affected by different phases of quasispecies dynamics, either of expansion or contraction of its mutant spectrum.

It could be anticipated that multivariate models such as logistic PCLR and PLSLR might describe the mutagenic effects more accurately than individual indices by adding the contributed predictive capacity of different indices despite its high correlation, in the sense that they could have a higher incidence with samples under mutagenic effect. But no logistic multivariate model beats RHL as a single predictor.

The information we seek to be captured with RHL lies below technical noise. Our approach has consisted in supposing that technical noise affects equally all samples in the experiment and that the distinctive effect would be caused by mutagenesis; that level will include both authentic rare haplotypes and those that are introduced by technical noise.

Deep sequencing has become an important tool to analyze viral populations subjected to mutagenic treatments [8, 22, 46]. The ranking of diversity indices to best characterize mutant spectra subjected to lethal mutagenesis is relevant to a growing body of fundamental and applied studies in virology and microbial genetics in general. Multiple high and low fidelity RNA virus mutants have been characterized [47], and how such mutants modify diversity indices is an open question that may shed light on the biological consequences of altered polymerase fidelity. Also, a large RHL questions the meaning of lethality of mutations in viral and microbial populations [41, 48–51]. Specifically, it is not clear whether the genomes that contribute to the RHL are slow replicators that can participate in evolutionary events, or are dead-end products transiently kept in viral populations by complementation [41, 52]. Ranking of diversity indices may provide also relevant information on the adaptative dynamics under enhanced mutagenesis [53], or the action of mutagenic agents on plant viruses [54]. From the perspective of pharmacology, RHL offers the means to distinguish whether base or nucleoside analogues that display antiviral activities affect viral RNA replication by direct inhibition of polymerase function or by enhanced mutagenesis, thus contributing to clarify uncertainties of drug action. The mechanism of activity of nucleotide analogues has consequences for the types of drug combinations that used together or sequentially can exert a more suppressive antiviral effect [55–58]. Studies with additional viruses, fidelity mutants, and nucleotide analogues are needed to provide a clearer picture of the relevance of different diversity indices to characterize

mutant spectra with alternative evolutionary histories. Cellular heterogeneity in important clinical disorders such as cancer parallels the population dynamics and the collective behavior of RNA viruses. Evaluation of a possible applicability of RHL determination as a marker in the mutagenic spectra generated during different cellular tumogeneric processes could also be considered. The present study, however, points towards RHL as a valuable marker for lethal mutagenesis of virus, and emphasizes that the choice of diversity indices to describe mutant spectra is not trivial.

Materials and methods

Cells, viruses, infections, and drugs

Huh-7.5 and Huh-7.5 reporter human hepatoma cell lines were grown in Dulbecco's modified Eagle's medium, and controlled as previously described [30, 32, 59, 60]. HCV p0 is the parental viral population obtained by electroporation into Huh-7.5-Lunet cells of a transcript of plasmid Jc1FLAG2(p7-nsGluc2A) (a chimera of J6 and JFH-1, genotype 2a) [33], and amplified in Huh-7.5 cells [30]. HCV p100 and HCV p200 resulted from population HCV p0 passaged 100 and 200 times, respectively, in Huh-7.5 reporter cells, as described [32]. Fitness of HCV p100 and HCV p200 relative to HCV p0 was measured in different growth-competition experiments between virus pairs. In initial determinations at a total MOI of 0.03 $TCID_{50}$/cell, HCV p100 fitness was 2.2±0.4 that of HCV p0 [31]. Subsequent determinations gave 1.28±0.34 at a MOI of 0.03 $TCID_{50}$/cell, and 1.10±0.02 $TCID_{50}$/cell at a MOI of 1 $TCID_{50}$/cell; the corresponding values for HCV p200 relative to HCV p0 were 1.33±0.46 and 1.17±0.02, respectively [32]. Infectious HCV was titrated as previously described [32]: serially diluted samples were applied to Huh-7.5 cells in 96-well plates (6,400 cells/well seeded 16 h earlier), and three days post-infection, cells were washed with PBS, fixed with ice-cold methanol, and stained to detect anti-NS5A monoclonal antibody 9E10 [61]. Titrations were performed in triplicate, and titers expressed as $TCID_{50}$/ml. Favipiravir (T-705) (Atomax Chemicals Co. Ltd) and ribavirin (Sigma) were prepared and used as previously described [19, 34, 35]. Their concentrations were chosen to produce comparable inhibition of HCV p0 progeny production.

RNA extraction, cDNA amplification and deep sequencing

Total intracellular viral RNA was extracted from infected cells using the Qiagen RNeasy kit (Qiagen, Valencia, CA, USA), according to the manufacturer's instructions. RT-PCR was carried out using AccuScript (Agilent Technologies), with specific oligonucleotide primers (S1 Table) The amplicons covered the following genomic regions: A1, spanning genomic residues 7626 to 7962; A2, residues 7941 to 8257; and A3, residues 8229 to 8653. Negative controls without template RNA were included in parallel to ascertain the absence of cross-contamination by template nucleic acids. PCR prod-

ucts were purified (QIAquick Gel Extraction kit), quantified (Pico Green™ assay), and analyzed for quality (Bioanalyzer) prior to MiSeq™ Illumina® sequencing.

Experimental design

The experiment is described schematically in Fig 1A and 1B, as described above. Four factors have been considered. (i) Amplicon, with three levels A1, A2 and A3. Different regions in the ORF are submitted to different functional restrictions. (ii) Base-line passage, with analyses at passes 0, 100 and 200, where starts quasispecies evolution in absence or presence of treatment. (iii) Treatment, with three levels: no drug, favipiravir and ribavirin. (iv) Treatment passages, with analyses at passes 3 and 10. (Fig 1).

Bioinformatics and statistics

All computations were done in the R environment and language (Team R 2017).

Fastq data treatment

The fastq files obtained from the MiSeq™ were subjected to the following treatment. A haplotype-centric data analysis pipeline was developed on targeted samples by amplicons following described procedures [62, 63] adapted to the MiSeq™ Illumina® platform in a paired-end 2x300 mode. It involved the following steps:

- Quality control of fastq files by inspecting profiles of per-site quality, read length and general instrument parameters of quality.
- In paired-end experiments overlap paired reads by FLASh [64] imposing a minimum of 20 bp overlapped with a maximum of 10% mismatches.
- Quality profiles of FLASh reads.
- Demultiplex reads by identifying oligonucleotide sequences within windows of expected positions in the sequenced reads.
  1. By MID (10 bp oligonucleotide) distinguishing samples from different patients/origins, only one mismatch is allowed.
  2. By specific primer (20-30 bp oligonucleotides) distinguishing different regions in the genome, and the two strands, up to three mismatches are allowed.
  3. Trim MIDs and primers.
  4. As a result, obtain a fasta file by each combination of MID, primer and strand in the run, where the reads were collapsed to haplotypes with corresponding observed frequencies.
- Align haplotypes in each fasta file to the wild type reference sequence or the master sequence in the file (most abundant haplotype) and quality filter.
  1. Discard haplotypes not covering the full amplicon.
  2. Discard haplotypes with more than two indeterminations, three gaps or more than 30% differences with respect to the reference.

3. Repair accepted indeterminations and gaps as per the reference sequence.

- Intersect haplotypes in both strands with a minimum abundance of 0.1%.
  1. Select haplotypes in both strands with abundances not below 0.1%.
  2. Discard haplotypes unique to one strand.
  3. Take coverage of haplotypes passing the filter as the sum of reads in both strands.
- The final haplotypes are called consensus haplotypes, and are the basis of the downstream analysis, except for the rare haplotypes load. Final yield 15–25% with respect to raw reads.

The pipeline consists in a set of R [65] scripts using objects and functions in packages Biostrings [66], ShortRead [67], and ape [68].

Rare haplotypes load (RHL)

It is computed as the fraction of reads in the sample belonging to haplotypes common to the forward and reverse strands with abundance below a given threshold. In the present work 1%, 0.1% and 0.01% have been studied as thresholds, finally taking 1% as the most informative and reliable.

Down sampling and fringe trimming (DSFT)

To compensate for possible biases in diversity indices due to differences in sample size [27, 29] we used down-sampling followed by fringe trimming, which consists in the following steps: (i) start with fasta files collecting the set of consensus haplotypes with abundance not below 0.1%; (ii) compute the total number of reads in each fasta file in the analysis, and take the minimum as the reference size; (iii) re-size the read number of each haplotype in each fasta file to the reference size; (iv) filter out all haplotypes below 0.2% with 95% confidence. These are the haplotypes and frequencies used in the computation of all diversity indices, except for RHL.

Association tests

Association tests of all indices with mutagenicity were computed by the non-parametric Wilcoxon signed rank test for paired samples, comparing the diversity values of the mutagenic treatment samples versus the paired control samples, and correcting the p-values for multi-test by the Bonferroni method; we did not distinguish among amplicons. Function *wilcox.test* with alternate as greater and in paired mode, and *p.adjust* in package 'stats' were used in the computations. Rather than declaring association at any p-value threshold, the models were ranked according to descending order of the standardized effect. The most associated index is considered to be that with the highest effect and still with a low p-value.

Logistic regression

Logistic regression was used to fit a predictive model of mutagenic activity regardless of the other factors in the design. Computations were performed by function *glm* in package 'stats'. To assess the predictive error rate of a fitted model, a sample was considered as under mutagenic effects if the fitted probability was above 0.5. The parameters used in the assessment were: residual deviance, Aikake Information Content (AIC), sensitivity, specificity, area under the ROC curve, and classification error rate. To minimize bias due to overfitting, the cross-validation error rate under 10-fold cross-validation (TenFoldCV) and under leave-one-out cross validation (LOOCV) was used. The models were sorted by increasing order of LOOCV. Function *cv.glm* in package 'boot' was used for the computation of CV error rates [69].

In-silico study

An in-silico analysis was performed to assess the robustness of RHL when compared with Hpl. To this aim, a fasta file of a sample with all haplotypes common to the forward and reverse strands with no previous abundance filter was used to sample and compute its RHL. Then the following steps were repeated 2000 times: (i) take a sample of 40,000 reads from the population; (ii) compute the RHL of this sample, (iii) filter out all haplotypes with abundance below 0.1% and not common to both strands; (iv) DSFT to 20,000 reads with haplotypes not below 0.2% with 95% confidence; (v) count Hpl. In each cycle the number of final reads, haplotypes, and RHL was computed and registered. With the set of 2000 values computed for each fasta file, the mean, median, coefficient of variation (CV), and interquartile dispersion coefficient (QD) were determined. A robust index is that showing the lowest CV and QD.

Free available R package.

An R package collecting all important functions, the package manual, and tutorial vignettes are freely available from GitHub at
https://github.com/VHIRHepatiques/QSutils

Supporting information

S1 Table. Oligonucleotides used to amplify and sequence HCV p0, HCV p100 and HCV p200 virus subjected to serial passages in the absence or presence of 400 μM favipiravir and 100 μM ribavirin.
https://doi.org/10.1371/journal.pone.0204877.s001 (DOCX)

S1 Fig. Logistic regression plots of mutagenicity over each diversity index considered.
A plot for each diversity index. As in Fig 3, bars at the top and bottom depict the values of each diversity index for HCV samples subject to mutagenesis or control. Dots on the logistic curve represent the predicted probability of mutagenesis.
https://doi.org/10.1371/journal.pone.0204877.s002 (PDF)

S2 Fig. AUC and LOOCV error rates (see p. 106).
(*Top*) Barplot with AUC values for each diversity index considered. (*Bottom*) Barplot with LOOCV error rate values for the logistic regression to each single diversity index. All samples included.
https://doi.org/10.1371/journal.pone.0204877.s003 (PDF)

S3 Fig. AUC and LOOCV error rates.
(*Top*) Barplot with AUC values for each diversity index considered. (*Bottom*) Barplot with LOOCV error rate values for the logistic regression to each single diversity index. Samples with a control/treatment of three passes only.
https://doi.org/10.1371/journal.pone.0204877.s004 (PDF)

S4 Fig. AUC and LOOCV error rates.
(*Top*) Barplot with AUC values for each diversity index considered. (*Bottom*) Barplot with LOOCV error rate values for the logistic regression to each single diversity index. Samples with a control/treatment of ten passes only.
https://doi.org/10.1371/journal.pone.0204877.s005 (PDF)

Acknowledgments

We thank Dr. Charles M. Rice for the supply of plasmid Jc1FLAG2(p7-nsGluc2A) and helpful advice for the implement of HCV replicon in cell culture.

Author contributions

- **Conceptualization**: Josep Gregori, Esteban Domingo.
- **Data curation**: Josep Gregori, Mercedes Guerrero-Murillo, Esteban Domingo.
- **Formal analysis**: Josep Gregori, Celia Perales.
- **Investigation**: María Eugenia Soria, Isabel Gallego, Josep Quer, Celia Perales.
- **Methodology**: Josep Gregori, María Eugenia Soria, Isabel Gallego, Celia Perales.
- **Project administration**: Josep Quer.
- **Resources**: Juan Ignacio Esteban, Josep Quer.
- **Software**: Josep Gregori, Mercedes Guerrero-Murillo.
- **Supervision:** Juan Ignacio Esteban, Josep Quer, Celia Perales, Esteban Domingo.
- **Validation**: Juan Ignacio Esteban, Josep Quer, Celia Perales, Esteban Domingo.
- **Visualization**: Celia Perales.
- **Writing ± original draft**: Josep Gregori, Celia Perales, Esteban Domingo.
- **Writing ± review & editing**: Isabel Gallego, Juan Ignacio Esteban, Josep Quer, Esteban Domingo.

REFERENCES

1. Crotty S, Maag D, Arnold JJ, Zhong W, Lau JY, Hong Z, et al. The broad-spectrum antiviral ribonucleoside ribavirin is an RNA virus mutagen. Nat Med 2000; 6(12):1375-9. https://doi.org/10.1038/82191 PMID: 11100123
2. Eigen M, Schuster P. The hypercycle. A principle of natural self-organization. Berlin: Springer, 1979.
3. Schuster P. Quasispecies on Fitness Landscapes. Curr Top Microbiol Immunol 2016; 392:61-120. https://doi.org/10.1007/82_2015_469 PMID: 26597856
4. Holland JJ, Domingo E, de la Torre JC, Steinhauer DA. Mutation frequencies at defined single codon sites in vesicular stomatitis virus and poliovirus can be increased only slightly by chemical mutagenesis. J Virol 1990; 64:3960-2. PMID: 1695258
5. Domingo E, Schuster P. Quasispecies: from theory to experimental systems. Switzerland: Springer, 2016.
6. Mullins JI, Heath L, Hughes JP, Kicha J, Styrchak S, Wong KG, et al. Mutation of HIV-1 genomes in a clinical population treated with the mutagenic nucleoside KP1461. PLoS One 2011; 6(1):e15135. https://doi.org/10.1371/journal.pone.0015135 PMID: 21264288
7. Arias A, Thorne L, Goodfellow I. Favipiravir elicits antiviral mutagenesis during virus replication in vivo. eLife 2014; 3:e03679. https://doi.org/10.7554/eLife.03679 PMID: 25333492
8. Guedj J, Piorkowski G, Jacquot F, Madelain V, Nguyen THT, Rodallec A, et al. Antiviral efficacy of favipiravir against Ebola virus: a translational study in cynomolgus macaques. PLoS Med 2018; 15(3):e1002535. https://doi.org/10.1371/journal.pmed.1002535 PMID: 29584730
9. Loeb LA, Essigmann JM, Kazazi F, Zhang J, Rose KD, Mullins JI. Lethal mutagenesis of HIV with mutagenic nucleoside analogs. Proc Natl Acad Sci USA 1999; 96(4):1492-7. PMID: 9990051
10. Sidwell OW, Simon LN, Witkowski JT, Robins RK. Antiviral activity of virazole: review and structure activity relationships. Prog Chemotherapy 1974; 2:889-903.
11. Smith RA, Kirkpatrick W. Ribavirin: a broad spectrum antiviral agent. New York: Academic Press Inc., 1980.
12. Graci JD, Cameron CE. Mechanisms of action of ribavirin against distinct viruses. Rev Med Virol 2006; 16(1):37-48. https://doi.org/10.1002/rmv.483 PMID: 16287208
13. Beaucourt S, Vignuzzi M. Ribavirin: a drug active against many viruses with multiple effects on virus replication and propagation. Molecular basis of ribavirin resistance. Curr Opin Virol 2014; 8:10-5. https://doi.org/10.1016/j.coviro.2014.04.011 PMID: 24846716
14. Furuta Y, Takahashi K, Kuno-Maekawa M, Sangawa H, Uehara S, Kozaki K, et al. Mechanism of action of T-705 against influenza virus. Antimicrob Agents Chemother 2005; 49(3):981-6. https://doi.org/10.1128/AAC.49.3.981-986.2005 PMID: 15728892
15. Furuta Y, Komeno T, Nakamura T. Favipiravir (T-705), a broad spectrum inhibitor of viral RNA polymerase. Proc Jpn Acad Ser B Phys Biol Sci 2017; 93(7):449-63. https://doi.org/10.2183/pjab.93.027 PMID: 28769016
16. Delang L, Abdelnabi R, Neyts J. Favipiravir as a potential countermeasure against neglected and emerging RNA viruses. Antiviral Res 2018; 153:85-94. https://doi.org/10.1016/j.antiviral.2018.03.003 PMID: 29524445
17. Delang L, Segura Guerrero N, Tas A, Querat G, Pastorino B, Froeyen M, et al. Mutations in the chikungunya virus non-structural proteins cause resistance to favipiravir (T-705), a broad-spectrum antiviral. J Antimicrob Chemother 2014; 69(10):2770-84. https://doi.org/10.1093/jac/dku209 PMID: 24951535
18. Baranovich T, Wong SS, Armstrong J, Marjuki H, Webby RJ, Webster RG, et al. T-705 (Favipiravir) induces lethal mutagenesis in influenza A H1N1 viruses in vitro. J Virol 2013; 87(7):3741-51. https://doi.org/10.1128/JVI.02346-12 PMID: 23325689
19. de Avila AI, Gallego I, Soria ME, Gregori J, Quer J, Esteban JI, et al. Lethal mutagenesis of hepatitis C virus induced by favipiravir. PLoS One 2016; 11(10):e0164691. https://doi.org/10.1371/journal.pone.0164691 PMID: 27755573
20. de Avila AI, Moreno E, Perales C, Domingo E. Favipiravir can evoke lethal mutagenesis and extinction of foot-and-mouth disease virus. Virus Res 2017; 233:105-12. https://doi.org/10.1016/j.virusres.2017.03.014 PMID: 28322918
21. Escribano-Romero E, Jimenez de Oya N, Domingo E, Saiz JC. Extinction of West Nile Virus

by favipiravir through lethal mutagenesis. Antimicrob Agents Chemother 2017; 61(11):e01400-17.

22. Qiu L, Patterson SE, Bonnac LF, Geraghty RJ. Nucleobases and corresponding nucleosides display potent antiviral activities against dengue virus possibly through viral lethal mutagenesis. PLoS Negl Trop Dis 2018; 12(4):e0006421. https://doi.org/10.1371/journal.pntd.0006421 PMID: 29672522

23. Abdelnabi R, Morais ATS, Leyssen P, Imbert I, Beaucourt S, Blanc H, et al. Understanding the mechanism of the broad-spectrum antiviral activity of favipiravir (T-705): key role of the F1 motif of the viral polymerase. J Virol 2017; 91(12):e00487-17.

24. Jin Z, Smith LK, Rajwanshi VK, Kim B, Deval J. The ambiguous base-pairing and high substrate efficiency of T-705 (Favipiravir) Ribofuranosyl 5'-triphosphate towards influenza A virus polymerase. PLoS One 2013; 8(7):e68347. https://doi.org/10.1371/journal.pone.0068347 PMID: 23874596

25. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiol Mol Biol Rev 2012; 76(2):159-216. https://doi.org/10.1128/MMBR.05023-11 PMID: 22688811

26. Perales C, Domingo E. Antiviral strategies based on lethal mutagenesis and error threshold. Curr Top Microbiol Immunol 2016; 392:323-39. https://doi.org/10.1007/82_2015_459 PMID: 26294225

27. Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology 2016; 493:227-37. https://doi.org/10.1016/j.virol.2016.03.017 PMID: 27060566

28. Seifert D, Beerenwinkel N. Estimating fitness of viral quasispecies from next-generation sequencing data. Curr Top Microbiol Immunol 2016; 392:181-200. https://doi.org/10.1007/82_2015_462 PMID: 26318139

29. Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodriguez-Frias F, et al. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics 2014; 30(8):1104-11. https://doi.org/10.1093/bioinformatics/btt768 PMID: 24389655

30. Perales C, Beach NM, Gallego I, Soria ME, Quer J, Esteban JI, et al. Response of hepatitis C virus to long-term passage in the presence of alpha interferon: multiple mutations and a common phenotype. J Virol 2013; 87(13):7593-607. https://doi.org/10.1128/JVI.02824-12 PMID: 23637397

31. Sheldon J, Beach NM, Moreno E, Gallego I, Pineiro D, Martinez-Salas E, et al. Increased replicative fitness can lead to decreased drug sensitivity of hepatitis C virus. J Virol 2014; 88(20):12098-111. https://doi.org/10.1128/JVI.01860-14 PMID: 25122776

32. Moreno E, Gallego I, Gregori J, Lucia-Sanz A, Soria ME, Castro V, et al. Internal disequilibria and phenotypic diversification during replication of hepatitis C virus in a noncoevolving cellular environment. J Virol 2017; 91(10):e02505-16. https://doi.org/10.1128/JVI.02505-16 PMID: 28275194

33. Marukian S, Jones CT, Andrus L, Evans MJ, Ritola KD, Charles ED, et al. Cell culture-produced hepatitis C virus does not infect peripheral blood mononuclear cells. Hepatology 2008; 48(6):1843-50. https://doi.org/10.1002/hep.22550 PMID: 19003912

34. Ortega-Prieto AM, Sheldon J, Grande-Perez A, Tejero H, Gregori J, Quer J, et al. Extinction of hepatitis C virus by ribavirin in hepatoma cells involves lethal mutagenesis. PLoS One 2013; 8(8):e71039. https://doi.org/10.1371/journal.pone.0071039 PMID: 23976977

35. Gallego I, Sheldon J, Moreno E, Gregori J, Quer J, Esteban JI, et al. barrier-independent, fitness-associated differences in sofosbuvir efficacy against hepatitis C virus. Antimicrob Agents Chemother 2016; 60(6):3786-93. https://doi.org/10.1128/AAC.00581-16 PMID: 27067341

36. González-López C, Arias A, Pariente N, Gómez-Mariano G, Domingo E. Preextinction viral RNA can interfere with infectivity. J Virol 2004; 78(7):3319-24. https://doi.org/10.1128/JVI.78.7.3319-3324.2004 PMID: 15016853

37. Crowder S, Kirkegaard K. Trans-dominant inhibition of RNA viral replication can slow growth of drug-resistant viruses. Nat Genet 2005; 37(7):701-9. https://doi.org/10.1038/ng1583 PMID: 15965477

38. Perales C, Mateo R, Mateu MG, Domingo E. Insights into RNA virus mutant spectrum and lethal mutagenesis events: replicative interference and complementation by multiple point mutants. J Mol Biol 2007; 369(4):985-1000. https://doi.org/10.1016/j.jmb.2007.03.074 PMID: 17481660

39. de la Torre JC, Holland JJ. RNA virus quasispecies populations can suppress vastly superior mutant progeny. J Virol 1990; 64(12):6278-81. PMID: 2173792

40. Grande-Pérez A, Lázaro E, Lowenstein P, Domingo E, Manrubia SC. Suppression of viral infectivity through lethal defection. Proc Natl Acad Sci USA 2005; 102(12):4448-52. https://doi.org/10.1073/pnas.0408871102 PMID: 15767582

41. Domingo E, Perales C. Quasispecies and virus. Eur Biophys J 2018; 47(4):443-57.

42. Crotty S, Cameron CE, Andino R. RNA virus error catastrophe: direct molecular test by using ribavirin. Proc Natl Acad Sci USA 2001; 98(12):6895-900. https://doi.org/10.1073/pnas.111085598 PMID: 11371613

43. Arias A, Isabel de Avila A, Sanz-Ramos M, Agudo R, Escarmis C, Domingo E. Molecular dissection of a viral quasispecies under mutagenic treatment: positive correlation between fitness loss and mutational load. J Gen Virol 2013; 94(Pt 4):817-30. https://doi.org/10.1099/vir.0.049171-0 PMID: 23239576

44. Grande-Pérez A, Gómez-Mariano G, Lowenstein PR, Domingo E. Mutagenesis-induced, large fitness variations with an invariant arenavirus consensus genomic nucleotide sequence. J Virol 2005; 79(16):10451-9. https://doi.org/10.1128/JVI.79.16.10451-10459.2005 PMID: 16051837

45. Novella IS, Duarte EA, Elena SF, Moya A, Domingo E, Holland JJ. Exponential increases of RNA virus fitness during large population transmissions. Proc Natl Acad Sci USA 1995; 92(13):5841-4. PMID: 7597039

46. Rawson JM, Landman SR, Reilly CS, Bonnac L, Patterson SE, Mansky LM. Lack of mutational hot spots during decitabine-mediated HIV-1 mutagenesis. Antimicrob Agents Chemother 2015; 59(11):6834-43. https://doi.org/10.1128/AAC.01644-15 PMID: 26282416

47. Borderia AV, Rozen-Gagnon K, Vignuzzi M. Fidelity variants and RNA quasispecies. Curr Top Microbiol Immunol 2016; 392:303-22. https://doi.org/10.1007/82_2015_483 PMID: 26499340

48. Jee J, Rasouly A, Shamovsky I, Akivis Y, Steinman SR, Mishra B, et al. Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. Nature 2016; 534(7609):693-6. https://doi.org/10.1038/nature18313 PMID: 27338792

49. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci USA 2011; 108(23):9530-5. https://doi.org/10.1073/pnas.1105422108 PMID: 21586637

50. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc 2014; 9(11):2586-606. https://doi.org/10.1038/nprot.2014.170 PMID: 25299156

51. Gregory MT, Bertout JA, Ericson NG, Taylor SD, Mukherjee R, Robins HS, et al. Targeted single molecule mutation detection with massively parallel sequencing. Nucleic Acids Res 2016; 44(3):e22. https://doi.org/10.1093/nar/gkv915 PMID: 26384417

52. Agol VI, Gmyl AP. Emergency services of viral RNAs: repair and remodeling. Microbiol Mol Biol Rev 2018; 82(2):e00067-17.

53. Arribas M, Cabanillas L, Kubota K, Lazaro E. Impact of increased mutagenesis on adaptation to high temperature in bacteriophage Qbeta. Virology 2016; 497:163-70. https://doi.org/10.1016/j.virol.2016.07.007 PMID: 27471955

54. Diaz-Martinez L, Brichette-Mieg I, Pineno-Ramos A, Dominguez-Huerta G, Grande-Perez A. Lethal mutagenesis of an RNA plant virus via lethal defection. Sci Rep 2018; 8(1):1444. https://doi.org/10.1038/s41598-018-19829-6 PMID: 29362502

55. Perales C, Agudo R, Tejero H, Manrubia SC, Domingo E. Potential benefits of sequential inhibitor-mutagen treatments of RNA virus infections. PLoS Pathog 2009; 5(11):e1000658. https://doi.org/10.1371/journal.ppat.1000658 PMID: 19911056

56. Moreno H, Grande-Pérez A, Domingo E, Martin V. Arenaviruses and lethal mutagenesis. Prospects for new ribavirin-based interventions. Viruses 2012; 4(11):2786-805. https://doi.org/10.3390/v4112786 PMID: 23202505

57. Iranzo J, Perales C, Domingo E, Manrubia SC. Tempo and mode of inhibitor-mutagen antiviral therapies: a multidisciplinary approach. Proc Natl Acad Sci USA 2011; 108(38):16008-13. https://doi.org/10.1073/pnas.1110489108 PMID: 21911373

58. Perales C, Iranzo J, Manrubia SC, Domingo E. The impact of quasispecies dynamics on the use of therapeutics. Trends in microbiology 2012; 20(12):595-603. https://doi.org/10.1016/j.tim.2012.08.010 PMID: 22989762

59. Blight KJ, McKeating JA, Rice CM. Highly permissive cell lines for subgenomic and genomic hepatitis C virus RNA replication. J Virol 2002; 76(24):13001-14.

https://doi.org/10.1128/JVI.76.24.13001-13014.2002 PMID: 12438626

60. Jones CT, Catanese MT, Law LM, Khetani SR, Syder AJ, Ploss A, et al. Real-time imaging of hepatitis C virus infection using a fluorescent cell-based reporter system. Nat Biotechnol 2010; 28(2):167-71. https://doi.org/10.1038/nbt.1604 PMID: 20118917

61. Lindenbach BD, Evans MJ, Syder AJ, Wolk B, Tellinghuisen TL, Liu CC, et al. Complete replication of hepatitis C virus in cell culture. Science 2005; 309(5734):623-6. https://doi.org/10.1126/science.1114016 PMID: 15947137

62. Gregori J, Esteban JI, Cubero M, Garcia-Cehic D, Perales C, Casillas R, et al. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One 2013; 8(12):e83361. https://doi.org/10.1371/journal.pone.0083361 PMID: 24391758

63. Ramírez C, Gregori J, Buti M, Tabernero D, Camos S, Casillas R, et al. A comparative study of ultradeep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. Antiviral Res 2013; 98(2):273-83. https://doi.org/10.1016/j.antiviral.2013.03.007 PMID: 23523552

64. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 2011; 27(21):2957-63. https://doi.org/10.1093/bioinformatics/btr507 PMID: 21903629

65. Team R. R: a language and environment for statistical computing. Vienna, Austria, 2017. Available from: https://www.r-project.org/

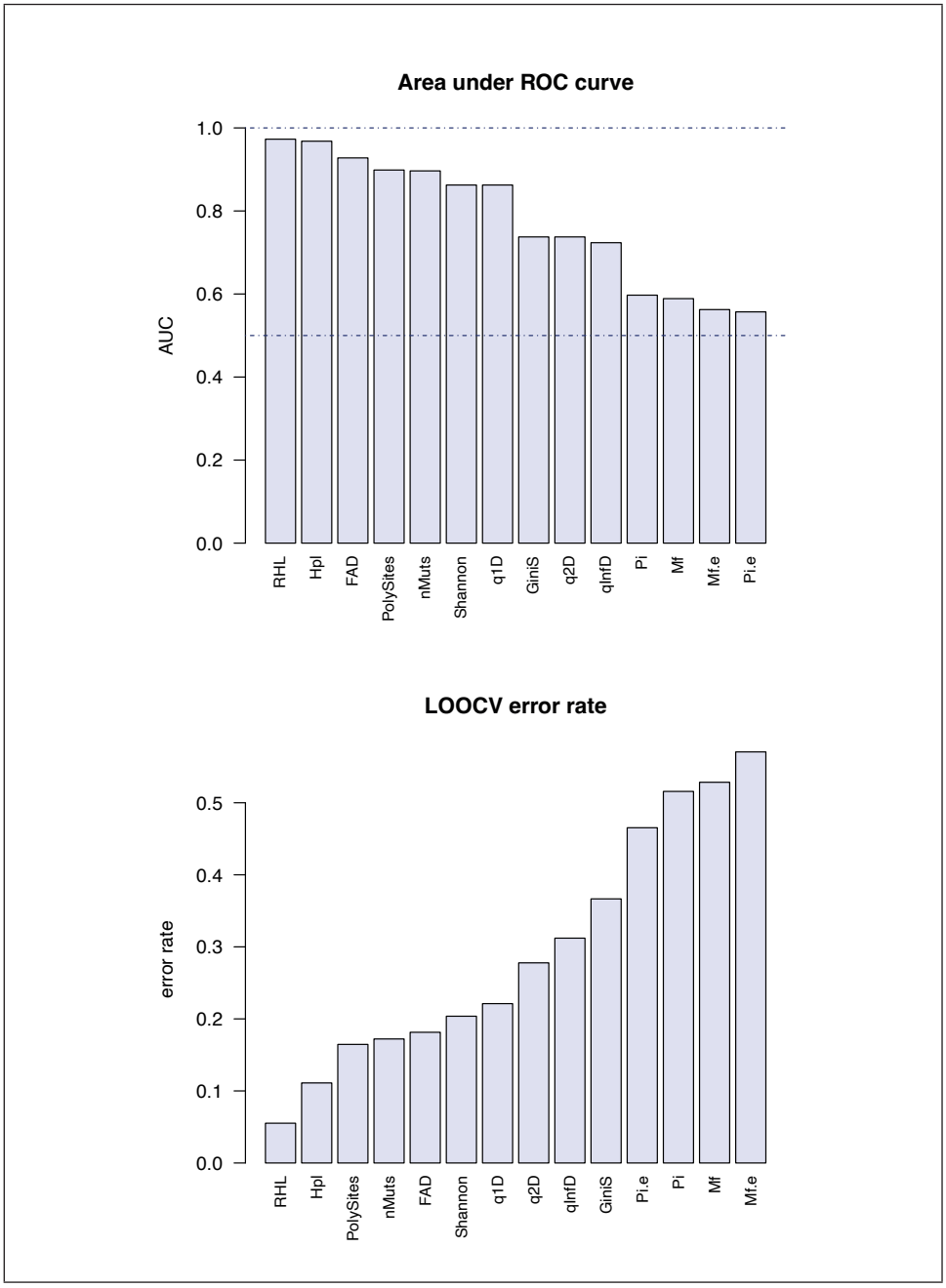
66. Pages H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: string objects representing biological sequences, and matching algorithms. R package version 2.38.4.

67. Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics 2009; 25(19):2607-8. https://doi.org/10.1093/bioinformatics/btp450 PMID: 19654119

68. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 2004; 20(2):289-90. PMID: 14734327

69. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York: Springer, 2009.

Supplementary Figure 2 (S2)



SECTION 5

Quasispecies fitness partition

Abstract

The changes occurring in viral quasispecies populations during infection have been monitored using diversity indices, nucleotide diversity, and several other indices to summarize the quasispecies structure in a single value. In this study, we present a method to partition quasispecies haplotypes into four fractions according to their fitness: the master haplotype, rare haplotypes at two levels (those below 0.1%, and those between 0.1% and 1%), and a fourth fraction that we term "emerging haplotypes", present at frequencies above 1%, but lower than that of the master haplotype. We propose that by determining the changes occurring in the volume of the four quasispecies fitness fractions (QFF), together with those of the Hill number profile (HNP), we will be able to visualize and analyze the molecular changes in the composition of a quasispecies over time. To develop this concept, we used three data sets: a technical clone of the complete SARS-CoV-2 spike gene, a subset of data previously used in a study of rare haplotypes, and data from a clinical follow-up study of a patient chronically infected with HEV and treated with ribavirin. The viral response to ribavirin mutagenic treatment was a selection of a rich set of synonymous haplotypes. The mutation spectrum was very complex at the nucleotide level, but at the protein (phenotypic/functional) level the pattern differed, showing a highly prevalent master phenotype. We discuss the putative implications of this observation in relation to mutagenic antiviral treatment.

Highlights

- A method is proposed to partition viral quasispecies into four fractions according to fitness.
- The fractions were combined with the Hill number profile to characterize the viral population.
- This combined concept is presented using a SARS-CoV-2 clone, and controlled HCV data.
- The method is especially sensitive to two types of quasispecies evolution: mutagenesis and emergence of treatment-resistant haplotypes, as well as a combination of these.
- Clinical application of this method is shown using samples from an HEV-infected patient receiving ribavirin.
- The information provided can be used to determine the effects of treatment at the molecular level.
- Early discontinuation of ribavirin treatment expanded nucleotide diversity.
- High nucleotide diversity was compatible with maintained functionality.
- High quasispecies diversity may reduce the antiviral effectiveness of ribavirin.

5.
Quasispecies fitness partition to characterize the molecular status of a viral population. Negative effect of early ribavirin discontinuation in a chronically infected HEV patient

JOSEP GREGORI , SERGI COLOMER-CASTELL, CAROLINA CAMPOS, MARTA IBAÑEZ-LLIGOÑA, DAMIR GARCIA-CEHIC, ARIADNA RANDO-SEGURA, CAROLINE MELANIE ADOMBIE, ROSA PINTÓ, SUSANNA GUIX, ALBERT BOSCH, ESTEBAN DOMINGO, ISABEL GALLEGO, CELIA PERALES, MARIA FRANCESCA CORTESE, DAVID TABERNERO, MARIA BUTI, MAR RIVEIRO-BARCIELA, JUAN I. ESTEBAN, FRANCISCO RODRÍGUEZ-FRÍAS, JOSEP QUER

ABSTRACT

The changes occurring in viral quasispecies populations during infection have been monitored using diversity indices, nucleotide diversity, and several other indices to summarize the quasispecies structure in a single value. In this study, we present a method to partition quasispecies haplotypes into four fractions according to their fitness: the master haplotype, rare haplotypes at two levels (those present at <0.1%, and those at 0.1-1%), and a fourth fraction that we term *emerging haplotypes*, present at frequencies >1%, but less than that of the master haplotype. We propose that by determining the changes occurring in the volume of the four quasispecies fitness fractions together with those of the Hill number profile we will be able to visualize and analyze the molecular changes in the composition of a quasispecies with time. To develop this concept, we used three data sets: a technical clone of the complete SARS-CoV-2 spike gene, a subset of data previously used in a study of rare haplotypes, and data from a clinical follow-up study of a patient chronically infected with HEV and treated with ribavirin. The viral response to ribavirin mutagenic treatment was a selection of a rich set of synonymous haplotypes. The mutation spectrum was very complex at the nucleotide level, but at the protein (phenotypic/functional) level the pattern differed, showing a highly prevalent master phenotype. We discuss the putative implications of this observation in relation to mutagenic antiviral treatment.

Keywords

Quasispecies, deep-sequencing, variability, rare haplotypes, fitness, mutagens.

1. Introduction

Viral quasispecies [1] are intrinsically dynamic entities, with new genomes (haplotypes) being created during each replication cycle, mainly produced by the activity of an error-prone viral polymerase. The fate of each genome depends on its fitness [2]; that is, its capacity for replication in competition with other genomes in the quasispecies within the current environment. At a given time, the approximate fitness of a viral genome in a specific sample can be inferred from its frequency in the quasispecies [3, 4]. That is, by the relative number of molecules belonging to the genome at that time point expressed as a fraction of the total. These frequencies can be considered a summary of the current molecular state of the quasispecies. The information gained by monitoring the molecular status of a viral quasispecies in an infection is particularly useful for following the viral response to a monoclonal antibody [5, 6] or mutagenic agent [7, 8]. Changes in the haplotype frequencies indicate the type of effects produced.

To characterize a viral quasispecies, we had to find an optimal balance between an analysis with high final coverage of shorter genomic fragments or with lower final coverage of larger fragments. The emergence of single-molecule real-time (SMRT™) sequencing has opened the possibility to sequence large fragments, including entire viral genomes, in a single read as has been reported in influenza virus [9], hepatitis C virus [10], and human immunodeficiency virus [11]. However, SMRT™ technology typically has lower coverage (maximum 4M reads per run) and higher error rates than Illumina techniques [12, 13]. SMRT™ is proven to have great value for assembling genomes [14] but requires further development to be useful for quasispecies analysis. For the present study, we needed very extensive final coverage ($\geq 10^5$ reads) to limit potential bias caused by a small sample size [3]. Hence, we used the MiSeq™ Illumina® instrument, which enables analysis of amplicons in a size range of 300-500 base-pairs (bp) at very high final depth and with acceptable error levels. The methods we propose are independent of the sequencing platform used. The only requirement is to have a set of high-quality haplotypes with their frequencies in the quasispecies. For this purpose, the sequencing data obtained here were treated to preserve the integrity of an amplicon's full-length reads; reads were either accepted or refused attending to their quality; however, they were never trimmed, except for the primers. Thus, the term *haplotype* used here refers to amplicon haplotypes, not to viral haplotypes, and the analyses are done amplicon-to-amplicon or on a single amplicon.

In a previous publication [7], we introduced the term *rare haplotype load* (RHL) in the context of a controlled experiment with HCV-infected hepatocytes treated with two mutagenic agents, the nucleoside analogues ribavirin (1-β-D-ribofuranosyl-1-*H*-1,2,4-triazole-3-carboxamide) and favipiravir (T-705; 6-fluoro-3-hydroxy-2-pirazinecarboxamide), and one inhibitor, sofosbuvir (isopropyl (2S)-2-[[[(2R,3R,4R,5R)-5-(2,4-dioxopyrimidin-1-yl)-4-fluoro-3-hydroxy-4-methyl-tetrahydrofuran-2-yl] methoxy-phenoxy-phosphoryl]amino]propanoate) [7]. The RHL refers to the fraction

of genomic molecules in the quasispecies representing low- to very low-fitness haplotypes, and it was used as a biomarker of mutagenesis. In this study, we extend the concept of quasispecies partitioning according to their fitness (frequency within the population) into four fractions or haplotype categories: the master or dominant haplotype (present at the highest frequency within the quasispecies); the RHL divided into two levels, haplotypes present at less than 0.1% and haplotypes present from 0.1% to 1%; and a fourth category including emerging haplotypes, defined as those present at frequencies below the master value and above 1%. Emerging haplotypes are considered to have the potential to proliferate, attain higher frequencies, and possibly overtake the current master.

Here, we propose to represent quasispecies evolution as the changes observed in the volume (fraction of molecules) of the four quasispecies fitness fractions (QFF) combined with the Hill number profile (HNP) [15], which quantifies the effective number of haplotypes. This combination provides a means to visualize and analyze molecular changes in the composition of a quasispecies over time. Three data sets were used for this purpose: (1) a technical clone of the complete SARS-CoV-2 spike gene [16], sequenced using 12 overlapping amplicons; (2) a subset of data previously used in a study of the RHL as a marker of lethal mutagenesis [7]; and (3) data from a clinical follow-up study of a patient chronically infected with hepatitis E virus (HEV) and treated with ribavirin (RBV), reported here for the first time.

2. Results

The technical clone is the simplest example (Figure 1). Despite its name, the commercial product carries multiple errors (although at low levels) due to the chemical yield at each synthesis step, which by definition never reaches 100% [17]. The errors observed include mainly deletions, which in the present case were bioinformatically corrected, but there were also some point mutations, possibly carry-overs from previous synthesis steps, which were left as they were. Some of these errors might also have been caused during the PCR amplification or sequencing steps. In Figure 1A, the QFF plot shows the <0.1% fraction which has a median of 0.11 (interquartile range [IQR] 0.02), the 0.1% to 1% fraction with a median of 0.011 (IQR 0.0058), no emerging haplotypes, and the master haplotype with a median of 0.876 (IQR 0.023). The HNP (Figure 1B) curve shows a steep decrease from $q = 0$ to $q = 1$. The curves remain asymptotically flat from $q = 1.5$ onwards, and there is only a small difference from $q = 3$ to $q = \infty$.

The controlled experiment with cultured HCV-infected human hepatoma cells treated with mutagens or inhibitors yielded richer plots (Figure 2A), which are characteristic of the effects of treatment. The median QFF and IQR values of the three amplicons for each fraction and under each condition are given in Table 1. Inhibitor treatment resulted in lower fractions of emerging haplotypes and RHL haplotypes with respect to the control, as well as higher master frequencies. In contrast, mutagen

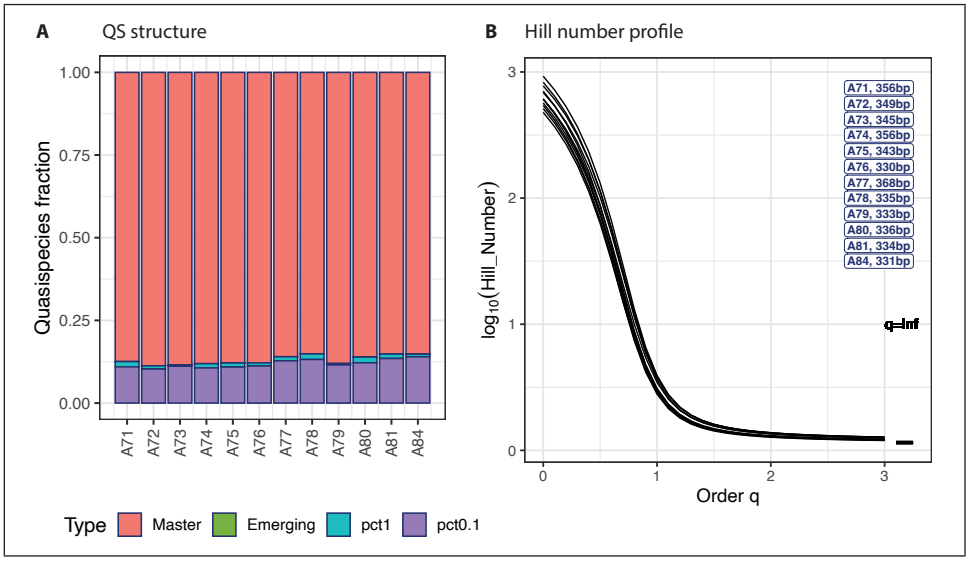


Figure 1. Quasispecies fitness fractions (**A**) and Hill number profile (HNP) plots (**B**) for a SARS-CoV-2 technical clone. The labels on the right of the HNP are sorted top-down in decreasing order of Hill number at q = 2; the ARTIC amplicon nomenclature is used (A71 to A84). Pct1 low fitness (0.1 < Freq ≤ 1%), pct0.1 very low fitness (Freq ≤ 0.1%). Each bar or curve corresponds to an amplicon of the S gene. Rarefied values are represented. Coverage range: 51,102-299,349 reads; median 76,954 reads.

exposure yielded higher RHL fractions and a depressed master relative to the control. In Figure 2B, the HNP is shown with labels sorted top-down in decreasing order of Hill number at q = 2. As the lower QFFs increase in volume and the master volume decreases, the curves remain at higher levels, and show a larger difference between q = 3 and q = ∞. Some curves cross over each other. This generally happens when there is one quasispecies with a large number of haplotypes having limited diversity in the frequencies, and a second with a smaller number of haplotypes but with higher diversity in the frequencies.

The changes occurring in the viral quasispecies following sequential RBV administration in the clinical case of HEV infection are shown in Figure 3A. In the first sample (baseline), analyzed on day 5 after the diagnosis (23 May 2018), the viral quasispecies already had a high burden of rare haplotypes at both <0.1% and 0.1-1%, with the master haplotype at <50%. The first mutagenic treatment (RBV 600 mg) increased the RHL while reducing the master haplotype. Treatment was stopped on day 158 following the diagnosis, and 28 days later (20 November 2018), the same master haplotype predominated in the quasispecies (61.7%), but at a low viral load (3 logs).

Two months later (245 days since diagnosis, 18 January 2019), in the absence of treatment, the quasispecies had diversified to a higher level than was seen at baseline, with the 0.1% RHL reaching 50%. Nine months later (502 days since diagnosis, 2 Octo-

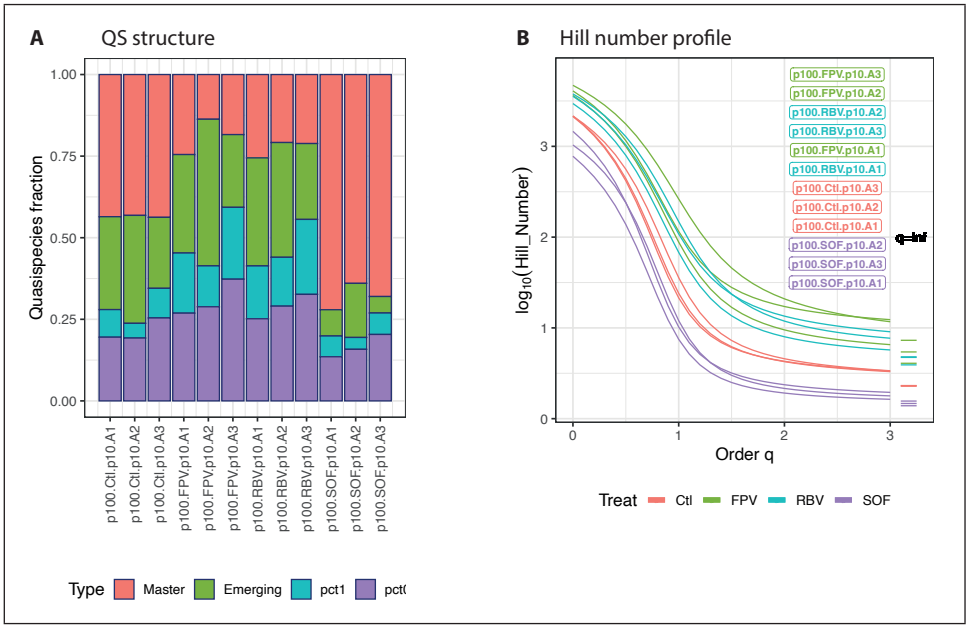


Figure 2. QFF (**A**) and HNP (**B**) plots for the HCV dataset. The labels on the right of the HNP are sorted top-down in decreasing order of Hill number at q = 2. Ctl control; FPV favipiravir; RBV ribavirin; SOF sofosbuvir. A1 to A3 refer to the amplicons analyzed. Pct1 low fitness (0.1 < Freq ≤ 1%), pct0.1 very low fitness (≤0.1%). Rarefied values are represented. Coverage range 48,057-335,535 reads; median 206,354 reads.

Table 1. Median (interquartile range) values of each fraction by treatment condition over the three amplicons.

	Master	**Emerging**	**RHL_1_0.1**	**RHL_0.1**
Control	0.436 (0.0030)	0.283 (0.0566)	0.085 (0.0232)	0.197 (0.0304)
FPV	0.184 (0.0541)	0.298 (0.111)	0.186 (0.0453)	0.290 (0.0519)
RBV	0.211 (0.0235)	0.331 (0.0586)	0.160 (0.0404)	0.293 (0.0366)
SOF	0.680 (0.0403)	0.081 (0.0577)	0.064 (0.0142)	0.158 (0.0348)

ber 2019), the viral load had increased to more than six logarithms, and a larger RBV dose (800 mg) was prescribed. One month later (530 days since diagnosis, 30 October 2019), the quasispecies showed a structure very similar to that of the previous analysis in QFF terms, but the viral load had decreased by 2 logs. Suddenly, one month later (558 days since diagnosis, 27 November 2019) while the patient was still under treatment, the master sequence recovered to >50%, and remained at the same level

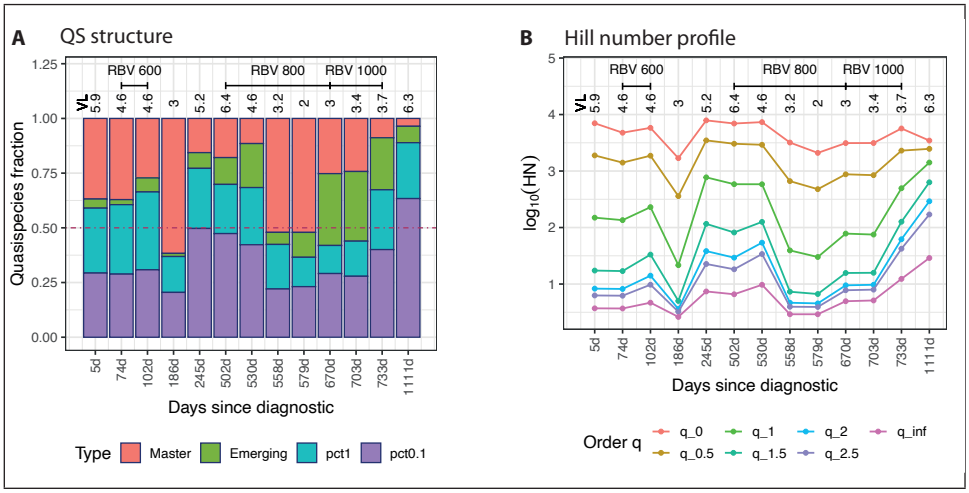


Figure 3. QFF (**A**) and HNP (**B**) plots for the HEV patient follow-up. The HNP is plotted here as cross-sections of the profile at given q values. Each line corresponds to a q value; the lines show how this value changes over time. Rarefied values are represented. On the x-axis, days since the diagnosis for each sample. Coverage range: 53, 307-503,770 reads; median 328,271 reads.

for an additional 20 days (day 579 since diagnosis, 18 December 2019), whereas the viral load decreased to 2 logs. Three months later (670 days since diagnosis, 18 March 2020), the volume of emerging haplotypes had increased while the master haplotype showed a >50% decline with respect to the previous time point, with slightly higher viral loads. This structure was maintained for another month, at similar viral loads. Finally, one month later (733 days since diagnosis, 20 May 2020), when the master haplotype had further declined to <10%, treatment had to be stopped. One year later (1111 days since diagnosis, 2 June 2021), in the absence of treatment, the master haplotype was present at 3.6%, the emerging volume was 7.9%, and >88.5% of the quasispecies was comprised of haplotypes with frequencies <1%. The HNP (Figure 3B) shows how the Hill numbers, at selected q values, changed over time.

The UPGMA tree of the master sequences of all samples (Figure 4A) shows that the same master haplotype was maintained from the time of the diagnosis up to 530 days later (30 October 2019). From then on, the master differed at each time point except the last one (1111 days since diagnosis, 2 June 2021) when the master was the same as the sample at 670 days (18 March 2020). The UPGMA tree (Figure 4B) based on the net genetic population distances, D_A, computed from the top 50 haplotypes in each sample, displays a similar structure. The samples at 558 and 579 days since diagnosis (27 November 2019 and 18 December 2019) show a divergence in the structure and in the master sequence. These correspond to the lowest viral loads in the follow-up, with the master sequence predominating in the quasispecies.

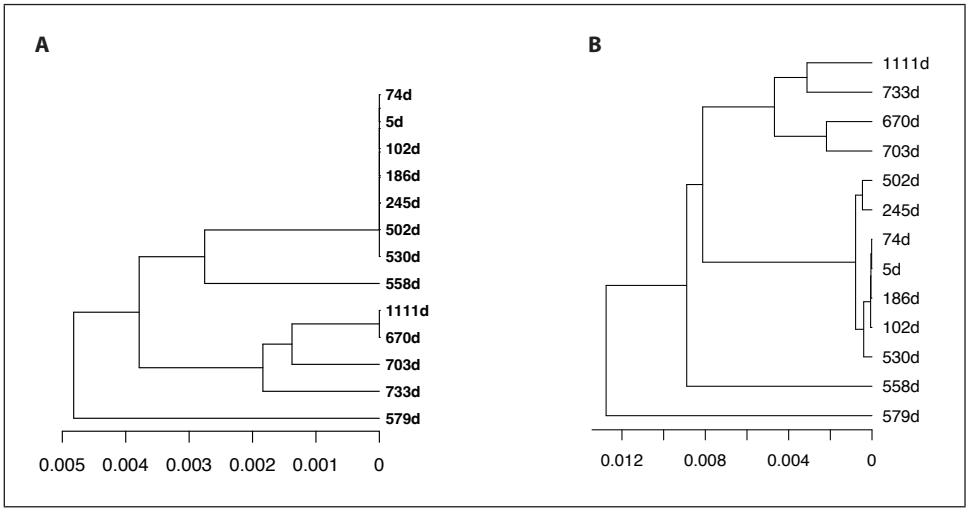


Figure 4. UPGMA tree of the master haplotypes based on raw nucleotide distances (**A**), and quasispecies tree based on the D_A population distances (**B**) taking the top 50 haplotypes in each sample. Each sample is labeled as days since the diagnosis.

To understand why the profile of the last sample (1111 days since diagnosis, 2 June 2021), analyzed after one year with no treatment, showed a highly mutagenized quasispecies, the nucleotide haplotypes were translated to amino acids and recollapsed to obtain amino acid haplotypes (phenotypes) with their corresponding frequencies (Figure 5A, B). The last sample (1111 days since diagnosis, 2 June 2021) shows a master phenotype accounting for >80.9% of the molecules. The master phenotype clearly predominated in the quasispecies along treatment, except in two samples (at end of treatment with RBV 800 mg and during treatment with RBV 1000 mg), where there was a change in the predominant amino acid sequence (Figure S1a). At the other time points, the same master sequence predominated in the quasispecies. These findings indicate that although a quasispecies can have a highly mutated spectrum at the nucleotide level, it may remain almost unchanged at the amino acid level, suggesting that functionality is at least transiently maintained (Figures S1a, b and S2).

The number of synonymous haplotypes corresponding to the master phenotype steeply increased after the treatment discontinuations (Figure 6). That is, the various haplotypes identical to the master phenotype generated during treatment were able to easily increase in frequency when it was stopped, as they all had highest functional fitness.

A UPGMA tree of master phenotypes based on their Grantham amino acid distances, and a quasispecies tree based on D_A population distances were constructed using the top 20 most frequent phenotypes in each sample (Supplementary Figures S1a and b). Apparently, viral treatment led to production of a rich set of haplotypes

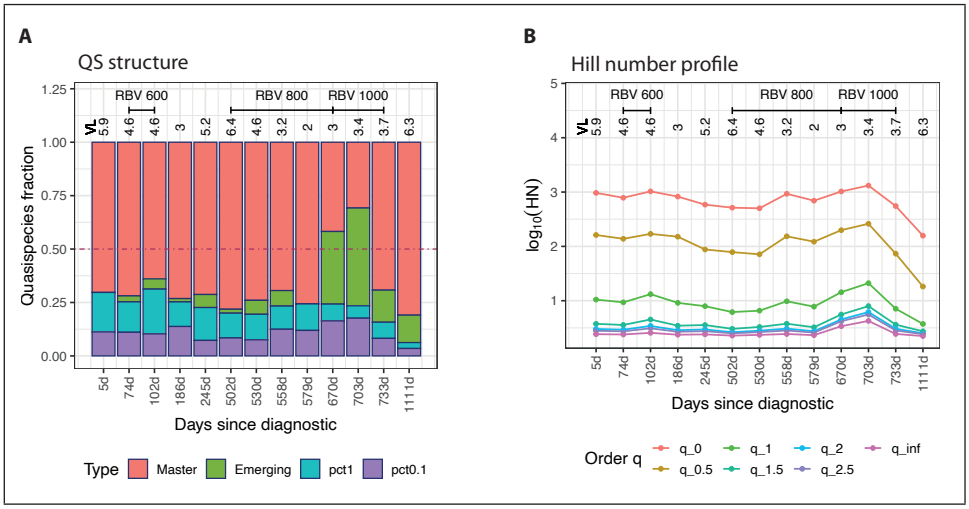


Figure 5. QFF (**A**) and HNP (**B**) plots for the quasispecies as amino acid haplotypes (phenotypes) for the HEV follow-up. Viral loads (VL) are expressed as logarithms. RBV, ribavirin. Pct1 low fitness (0.1 < Freq ≤ 1%), pct0.1 very low fitness (≤0.1%). Rarefied values are represented.

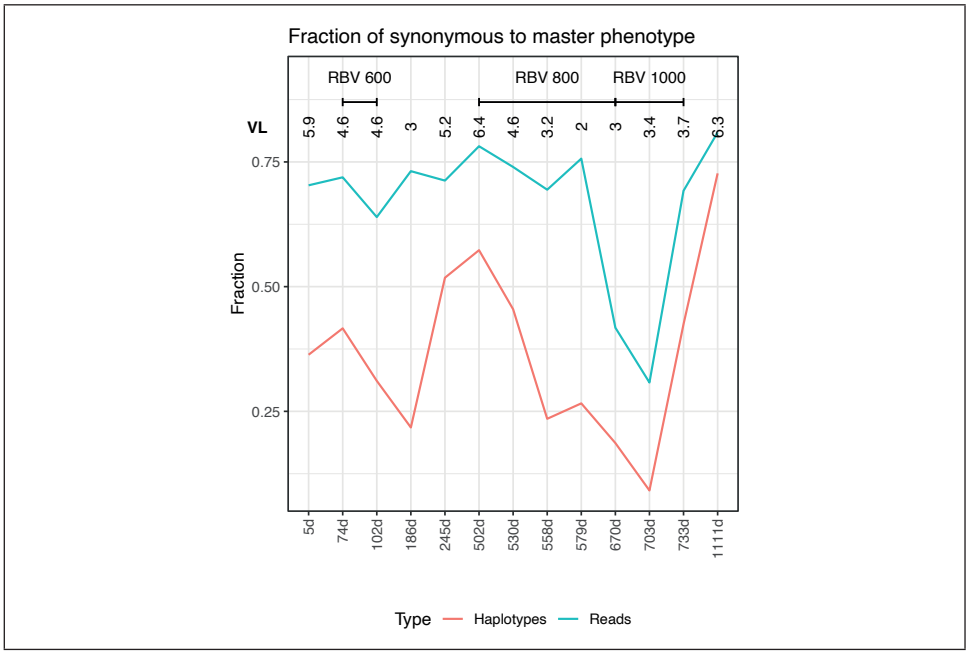


Figure 6. Fraction of haplotypes synonymous to the master phenotype (orange), and fraction of reads for these haplotypes (turquoise). On the x-axis, days since the diagnosis for each sample. Viral loads (VL) are expressed as logarithms. RBV, ribavirin.

synonymous to the master phenotype, all with equal or comparable fitness, which proliferated when treatment was discontinued. This was evidenced by the decline in the master haplotype volume and the increase in emerging haplotypes synonymous to the master. As a whole, the resulting quasispecies might be more resistant to further mutagenic treatment and better fit to its current environment. A large number of highly fit haplotypes could correspond to a large number of molecular pathways to escape a treatment.

A final illustration of this conclusion is provided in Supplementary Figures S3 and S4, where Montserrat plots depict the distribution of the 1000 most abundant haplotypes and all phenotypes in the last sample (1111 days since diagnosis, 2 June 2020). In Montserrat plots, haplotypes are sorted by number of substitutions with respect to the master first, and by decreasing order of abundance second [18]. Each successive peak represents additional differences with respect to the master. In correspondence, Supplementary Figure S5 shows the mean number of differences with respect to the master haplotype/phenotype per read, at both the nucleotide level (substitution load) and amino acid level (mutation load). Notably, the ratio of the nucleotide substitution load to the amino acid mutation load was 13.31 in the last sample (one year without treatment), only 1.24 in the baseline sample, shortly after diagnosis, and 6.57 when the last treatment was discontinued. In addition, the mutation load (amino acid level) in the last sample was the lowest value in the series, despite the highly diverse quasispecies.

3. Discussion

The presence of a broad repertoire of genomes in the mutant spectra of RNA viruses represents a challenge for annotating and describing the genomes and the neighbor relationships among them. Several procedures have been developed to rank subgroups of related sequences within mutant spectra [19, 20, 21, 22, 23]. Alternative approaches have monitored mutant spectrum composition and diversification through quantification of diversity indices [4, 24] and haplotype mapping using two-dimensional neural networks [25, 26].

In our previous studies we introduced a number of new diversity indices, which, when adequately combined, provided information on both the abundance and divergence of haplotypes within a viral population [4]. In the present report, we have gone one step further, creating a procedure to divide a quasispecies population into four fractions using a fitness partition (QFF). The advantages of this approach include better statistical properties than most diversity indices and inclusion of the biological features of the population. The QFF procedure goes beyond a previous proposal restricted to the two fractions at lowest frequency in the mutant ensemble [7]. In addition, we used the Hill number profile [15] to provide a complementary view of the quasispecies structure. The profile is an enriched summary that assigns increasing

weights to the haplotype frequencies in the quasispecies as the order q increases. In addition, the units provided (number of equally fit haplotypes) contribute to the interpretation of the results.

Rarefaction by repeated resampling provided an efficient down-sampling of mutant distributions in the quasispecies samples to the minimum coverage, thereby allowing comparisons regardless of the sample size. Hill numbers of orders below two are highly dependent on sample size and must be corrected. However, it is advisable to correct all diversity values. QFF is compatible with other ranking procedures used to study quasispecies, as the proposed fractioning can be applied to genome subsets obtained by other means. In particular, it can be a useful complement to self-organized fitness maps based on artificial neural networks [26]. It may also provide support to precisely define the SARS-CoV-2 mutant spectra which, according to recent results, are populated by a large proportion of low-frequency haplotypes [27, 28].

The rationale behind the QFF definition and its main contribution to quasispecies analysis resides in the biological meaning of each fraction. Observation of a significant fraction of molecules corresponding to very low fitness haplotypes is only consistent with the presence of a mechanism that can generate mutants at a high rate, but cannot increase their frequency relative to competing genomes. Furthermore, concerning the response of HEV to RBV, the course of any treatment giving rise to resistant variants is only consistent with a decline in volume of the master haplotype together with a parallel increase in emerging haplotypes.

We used both the QFF and HNP to visualize and analyze two simple case models: the first, a SARS-CoV-2 technical clone, and the second, populations from a controlled experiment in which a clonal HCV population was serially passaged in cultured human hepatoma Huh-7.5 cells. These two datasets were used to demonstrate the capability of the proposed QFF/HNP combined method. Finally, the procedure was applied to a complex clinical case: follow-up of an HEV-infected patient treated with various RBV doses with two discontinuations, to illustrate its performance for clinical purposes. As this is a single example, the discussion provided for this case should be considered explanatory of how the information obtained with the QFF procedure can be of clinical value.

Lethal mutagenesis is a useful antiviral approach that consists of driving viral genomes to extinction – pushing the virus to cross the error catastrophe threshold by increasing the viral mutation rate above the maximum level compatible with infectivity – without mutagenizing the host cells [29, 30, 31]. A recent example of successful application of lethal mutagenesis is the use of molnupiravir against COVID-19 [32]. Several other cases have been reviewed [8]. Currently, no specific drugs have been approved for HEV infection; RBV is the main option as an off-label drug. RBV is a broad-spectrum mutagenic agent [33, 34, 35] that can increase mutation rates and result in extinction of the virus by lethal mutagenesis [36]. However, during RBV treatment, a reduction in the effective antiviral dose by low adherence or early treatment interruption can allow residual viral replication and production of rescue var-

iants with decreased RBV sensitivity or altered replication fitness [37, 38, 39]. This could lead to selection of mutations resistant to RBV and the appearance of resistant variants.

In our study, deep-sequencing of samples from an HEV-infected patient receiving RBV treatment showed that the mutagenic agent led to a reduction in the most highly represented sequence (master) at the nucleotide level, together with a significant increase in the number of rare haplotypes, findings in agreement with a mutagenic effect of RBV on this virus. When the first round of RBV treatment (600 mg) was stopped, HEV relapsed, showing an increase in viral load and re-acquisition of the master sequence that had predominated before treatment was started. In the second and third rounds of RBV treatment, with increases in the drug concentration to 800 mg and 1000 mg and evidence of good adherence to therapy, the viral load remained unchanged and even increased, while the master sequence decreased once again. Notably, after stopping 1000 mg RBV treatment, the frequency of the master genome declined even further despite a three-log viral load increase.

However, when we examined the quasispecies at the protein (phenotypic/functional) level, the pattern drastically changed, showing a highly predominant master phenotype, despite the complexity of the mutation spectrum observed at the nucleotide level. On stopping treatment, the synonymous variants generated had the same replication capacity as the original master genome and were able to proliferate, leading to a high diversity of genomes that could all express the same phenotype. The reason why the dynamics of the mutant spectrum involved haplotypes with silent mutations is unknown. Furthermore, we cannot exclude that some of the variants produced might have had lower sensitivity to RBV. Both these reasons, a large reservoir of functional genomes and decreased sensitivity to treatment, could explain the viral load increase in the presence of a high dose of mutagen. Failures in HEV RBV treatment have been described, but detection of RBV-resistant mutations requires sequencing the full HEV genome [38, 39, 40]. In our case, we assumed that the effects of RBV on the sequenced amplicon would be similar to the impact on the remainder of the genome, as the virus cannot drive the mutagenic effects. Future approaches using the new SMRT circular consensus sequencing technology may provide further support of this amplicon to genome relationship.

The mutagenic effect of RBV on HEV could lead to viral extinction, but it also involves a risk of accumulating advantageous mutations and selecting fitness-enhancing ones [36]. In our study, while the patient was receiving antiviral treatment, a number of synonymous haplotypes were produced, which as a whole seemed to be stronger against further treatment and improved accommodation of the quasispecies to its current environment. The findings from our patient suggest that mutagenic antiviral therapy should ideally be combined with other antivirals. When this is not possible, as in HEV infection, treatment should be maintained, even after serum RNA tests negative, to avoid relapses which could lead to selection of fitness-enhancing mutations and treatment failure.

The QFF approach, although simple and straightforward, has some technical limitations, mainly due to the current state of high-throughput sequencing technology. The sequence length and error level are two aspects that limit each other. We can sequence amplicons up to slightly more than 500 bp in length with an acceptable error level using paired-end technology, but we cannot sequence full-length viral genomes of a few thousand base pairs at high depth (10^4-10^5) with low error levels and high coverage. In this study, we used amplicons larger than 300 bp and assumed that either the amplicon underwent effects similar to those that would occur in the whole genome, or that the amplicon was the target to study. Another limitation relates to a factor observed in all -*omics*, where the experimental design is of the utmost importance to avoid bias. Even in balanced designs, batch effects should be taken into account. In our case, we were not interested in a detailed account of point mutations and indels; rather, we aimed to provide a picture of the macroscopic structure of a quasispecies. This information is of value, as high-resolution deep sequencing unveils myriads of low frequency mutations that should correspond to an extensive repertoire of minority genomes [28]. In providing this general view, we can accept a certain presence of artifactual haplotypes, provided that all samples in the experiment show the same noise level, hence the need for an accurate experimental design. Filtering above the error level to avoid all artifacts would involve a considerable loss of information and jeopardize the type of analysis we are proposing. The impact of filtering all haplotypes below 0.1% or 1% on the total reads number, that is the information loss, can be seen in Figure 1, Figure 2 and Figure 3. Nevertheless, low-level filters of very few reads per haplotype (e.g., 1-10 reads at 1×10^5 coverage) will have an impact on the number of haplotypes – that is, on the Hill numbers of low order (<1.5) – but will have a minimum effect on the number of reads. It could be helpful to perform a sensitivity analysis of the various diversity indices used with respect to this threshold. These limitations and warnings are equally applicable to any quasispecies study, whatever the indices or variables used, and are not exclusive to the methods proposed here.

We propose a simple method for monitoring the changes occurring in a quasispecies at the molecular level, involving fitness fractions and the Hill number profile. This combined method, which provides an easily interpretable visualization of quasispecies evolution in viral terms, was applied to two simple cases as a demonstration, and to samples from an RBV-treated HEV patient to illustrate its clinical value. The method is based on bioinformatic treatment of sequencing data to obtain a set of high-quality amplicon haplotypes with their corresponding frequencies to represent the quasispecies structure. Use of next-generation sequencing technology in combination with a good experimental design provides exceptional opportunities to study complex quasispecies and follow their evolution at the molecular level, in both research laboratories and clinical settings.

4. Materials and methods

4.1. Samples

Two datasets were used to develop quasispecies molecular characterization with QFFs and the HNP:

- A technical clone of the SARS-CoV-2 *S* gene (Twist Synthetic SARS-CoV-2 RNA Control 2 MN908947.3, TWIST Biosciences, South San Francisco, CA, USA) sequenced in 12 amplicons [17]. Commercial Twist Synthetic SARS-CoV-2 RNA controls consist of six non-overlapping 5-Kb fragments generated from in vitro transcription of gene fragments. The synthetic controls were diluted at 1:10 to a concentration of 1×10^5 copies per microliter, PCR-amplified following the Sub-ARTIC v3 protocol [41] using a set of 28 primers (A71 to A84) covering the full *S* gene, and sequenced on a MiSeq™ Illumina® system [42]. The haplotypes and corresponding frequencies in this analysis included all haplotypes common to both DNA strands after a previous filter at 2 reads. That is, at a minimum of 2 + 2 reads.
- Three HCV amplicons from samples taken from a controlled experiment, in which HCV-infected human hepatoma cells were observed in the presence or absence of RBV, favipiravir, or sofosbuvir [43, 44]. Briefly, HCV p0 was the parental viral population obtained by electroporation of a transcript of plasmid Jc1FLAG2(p7-nsGluc2A) (a chimera of J6 and JFH-1, genotype 2a) [45] into Huh-7.5-Lunet cells and amplification in Huh-7.5 cells [46]. HCV p100 resulted from passaging the HCV p0 population 100 times in Huh-7.5 reporter cells [47]. HCV p100 was subsequently passaged 10 additional times in the presence of favipiravir (T-705) (Atomax Chemicals Co., Ltd., Shenzhen, China), RBV (Sigma, Kawasaki, Japan), or sofosbuvir. Drug concentrations were adjusted to produce comparable inhibition of HCV p0 progeny production. The amplicons sequenced covered the following HCV genomic regions: A1, spanning genomic residues 7626 to 7962; A2, residues 7941 to 8257; and A3, residues 8229 to 8653. The haplotypes and corresponding frequencies in this analysis included all haplotypes common to both strands, with no previous abundance filter; that is, a minimum of 1 + 1 reads.

In addition, we describe the quasispecies findings from the clinical follow-up case of a 27-year-old patient who acquired chronic HEV infection after undergoing two kidney transplantations. The patient received three different RBV regimens (Table 2). First, 600 mg per day for 3 months, which led to a significant reduction in viral load without achieving undetectable HEV RNA. Treatment was stopped. The patient relapsed, and a second treatment with RBV 800 mg daily was prescribed, with a new reduction in HEV levels. At month 5, RBV dosage was increased to 1000 mg daily for two additional months. Treatment was discontinued because of a lack of antiviral response, and viral load jumped three logs at 10 days after stopping treatment. A single amplicon covering genomic positions 6323 to 6734 on the ORF2 region was se-

Table 2. Follow-up data, with dates and intervals, viral loads expressed as logarithms, and clinical observations. EOT, end of treatment.

Date (Y-M-D)	Interval (Days)	Days since diagnosis	Sample ID	LogVL	Observations
2018-05-18	0	0		5.91	Diagnosis
2018-05-23	5	5	S01	5.87	
2018-07-31	69	74	S03	4.60	Ribavirin 600 mg
2018-08-28	28	102	S04	4.60	
2018-10-23	56	158		1.54	EOT
2018-11-20	28	186	S06	3.04	Relapse
2019-01-18	59	245	S08	5.18	
2019-10-02	257	502	S10	6.43	Ribavirin 800 mg
2019-10-30	28	530	S12	4.62	
2019-11-27	28	558	S14	3.18	
2019-12-18	21	579	S16	2.04	
2020-03-18	91	670	S17	3.04	Ribavirin 1000 mg
2020-04-20	33	703	S18	3.40	
2020-05-20	30	733	S20	3.68	EOT
2020-06-17	28	761		4.45	Relapse
2021-06-02	350	1111	S24	6.28	

quenced. The haplotypes and corresponding frequencies in this analysis included all haplotypes common to both DNA strands after a previous filter at 2 reads. That is, a minimum of 2 + 2 reads.

4.2. Processing the sequencing data

The aim of the sequencing data treatment was to discard error-bearing reads while preserving full-length read integrity, so that haplotypes that completely cover the amplicon with their respective frequencies were incorporated. The steps in this process are the following:

- obtain Fastq files with Illumina® 2 × 300 bp paired-end reads;
- recover full amplicon reads with FLASH [48] (min. 20 bp overlap, max. 10% mismatches). The 300 bp reads, when overlapped, result in reads covering complete ~400-500 bp amplicons;

- remove full reads with 5% or more bases below a Phred score of Q30;
- demultiplex and trim primers (max three differences accepted);
- collapse reads (molecules) to haplotypes (amplicon-genomes) and their frequencies; the frequencies were calculated per haplotype of each amplicon;
- in certain cases, remove all haplotypes below a fixed frequency threshold;
- remove all haplotypes that are not common to both DNA strands.

The final obtained haplotypes and their frequencies were the basis for all further calculations.

The SARS-CoV-2 dataset consists of 12 amplicons (min 330 bp, max 368 bp, median 340 bp). The HCV dataset consists of three amplicons (312, 318, and 423 bp). Finally, the HEV study is based on single 363 bp amplicons (Supplementary Table S1). The amplicon sizes provided are the final result after primer trimming.

4.3. Quasispecies fitness partitions

At a given time, a quasispecies is usually comprised of a highly predominant haplotype, a few low- to medium-frequency genomes, various rare haplotypes with very low fitness but still able to replicate to some level, and some defective genomes unable to replicate. This composition can be modeled using the set of frequencies of all haplotypes as parameters of a multinomial distribution (Equation (1)),

$$\Pi = \{p_1, p_2, \ldots, p_n\} with \sum_{i}^{n} p_i = 1 \tag{1}$$

where p_1, p_2, ..., p_n represent the various haplotypes, arranged in order of decreasing frequency. The parameters, p_i, are sorted in decreasing order without a loss of generality. In this way, the quasispecies can be partitioned into fractions limited by frequency thresholds of interest, as in Equation (2), where a partition into four fractions is illustrated,

$$\begin{gathered} \Pi_1 = \{p_1, p_2, \ldots, p_k\}, \forall p_i : p_i \geq p_k \\ \Pi_2 = \{p_{k+1}, p_{k+2}, \ldots, p_l\}, \forall p_i : p_l \leq p_i < p_k \\ \Pi_3 = \{p_{l+1}, p_{l+2}, \ldots, p_m\}, \forall p_i : p_m \leq p_i < p_l \\ \Pi_4 = \{p_{m+1}, p_{m+2}, \ldots, p_n\}, \forall p_i : p_n \leq p_i < p_m \\ p'_1 = \sum_{i}^{k} p_i; p'_2 = \sum_{k+1}^{l} p_i; p'_3 = \sum_{l+1}^{m} p_i; p'_4 = \sum_{m+1}^{n} p_i; with \sum_{i=1}^{4} p'_i = 1 \end{gathered} \tag{2}$$

and where, p'_1, p'_2, p'_3, and p'_4 represent the four fractions.

In the typical quasispecies structure mentioned above, the four fractions can be defined as follows:

- **Master**: the fraction of molecules belonging to the most frequent haplotype; that is, the one present at the highest percentage ($p'_1 = p_1$);
- **Emerging**: the fraction of molecules present at a frequency >1% and less than the master percentage, belonging to haplotypes that are able to compete with the predominant one and possibly replace it (p'_2);
- **Low fitness**: the fraction of molecules present at frequencies from 1% to 0.1%, belonging to haplotypes that have a low probability of progressing to higher frequencies (p'_3);
- **Very low fitness**: the fraction of molecules present at frequencies <0.1% belonging to haplotypes with very low fitness and to defective genomes. The likely fate of these molecules individually is degradation, but the fraction is continuously fed with new very low fitness genomes produced by replication errors or by host editing activities (p'_4).

The evolutionary trends occurring in a viral quasispecies can be characterized by determining the changes that take place in the molecular volume of these fractions as a function of time.

The coefficient of variation (CV) of a proportion, p, for a given sample size, N (i.e., the standard deviation expressed in expected value units), is given in Equation (3). When the proportion is very small, it can be approximated by calculating the square root of the inverse of the product of the sample size multiplied by the proportion, as in Equation (4),

$$CV\left[\frac{x}{N}\right] = CV(p,N) = \frac{sd(p,N)}{p} = \sqrt{\frac{(1-p)}{Np}} \tag{3}$$

$$CV(p,N) \approx \sqrt{\frac{1}{Np}} \tag{4}$$

where x is the observed count, N is the sample size, and p is the observed proportion. In the experiments described in this study, the coverage (sample size, N) ranged from 10^4 to 10^6, with the average larger than 1×10^5, and our aim was to observe haplotypes present in very low proportions (p); that is, <0.1% (<1.e-3). Individually, haplotypes considered to have very low fitness will show a high CV, which means that some of them can be easily overlooked in a single sample. Nevertheless, when grouped together, as is seen in the above partition (the p' in Equation (2)), they amount to a much higher proportion than when counted individually and become more statistically stable. Thus, the fraction of molecules in the quasispecies belonging to haplotypes having very low fitness, p'_4, becomes more stable to sampling and less dependent on the sample size [7] than any single haplotype at these frequencies.

4.4. Hill numbers

In addition to the partition described above, we can determine the diversity profile of a quasispecies with the use of Hill numbers [4, 15, 49]. Based on the expression of the generalized diversity index, of order q, ${}^qH(p)$ given in Equation (5),

$$ {}^qH(p) = \sum_{i=1}^{H} p_i^q \tag{5} $$

the Hill number of order q, ${}^qD$, of a quasispecies corresponds to the number of equally fit haplotypes comprising a quasispecies with the same general diversity, ${}^qH$, as the original quasispecies, as is shown in Equation (6).

$$ \sum_{i=1}^{H} p_i^q = \sum^{D} \left(\frac{1}{D}\right)^q = D\left(\frac{1}{D}\right)^q = D^{1-q} \tag{6} $$

This results in Equation (7),

$$ {}^qD(p) = \left(\sum_{i=1}^{H} p_i^q\right)^{1/(1-q)} \tag{7} $$

where ${}^qD(p)$ is the Hill number of order q, calculated from the haplotype frequencies observed in the quasispecies, p_i. The diversity indices, ${}^qD(p)$, obey the replication principle [4] and are expressed in intuitive units (number of equally fit haplotypes). In ecology, the replication principle states that if we have N equally large, equally diverse groups (quasispecies), and no species (genomes) in common, the diversity of the pooled groups must be N times the diversity of a single group.

At increasing values of q starting from 0, we obtain diversity values that are also a transformation of other classical diversity indices:

- at $q = 0$, the Hill number is the number of haplotypes;
- at $q = 1$, it corresponds to the exponential of Shannon entropy;
- at $q = 2$, it is the inverse of the Simpson index; and
- at $q = \infty$, it is the inverse of the predominant haplotype.

The Hill number profile of a quasispecies is the curve we obtain from $q = 0$ to 3, plus the value at infinity. The curve becomes asymptotic beyond order 2 and reaches its minimum at infinity. As a result of the quasispecies values spanning a large range – more than three orders of magnitude – the Hill number profile is best represented in log_{10} units.

4.5. Abundance filter effect on haplotype distribution

The goal of step 6 in the sequencing data treatment described above is usually to limit technical errors (PCR + sequencing) to a level suitable for the purposes of the study, while maintaining the integrity of full amplicon reads. The required frequency threshold can be established through the use of clones that have been processed in parallel with the clinical samples and sequenced in the same run [50].

The immediate effect of this filter is removal of all haplotypes with abundances below the threshold, which in probabilistic terms, corresponds to truncating the distribution. Let p_k be this threshold, so that the haplotypes removed are those at frequencies p_{k+1}, p_{k+2}, ..., p_n. The resulting truncated quasispecies will show haplotype frequencies resulting from normalization of the remaining haplotype frequencies, (p_1, p_2, ..., p_k), as described in Equation (8),

$$\Pi' = (p_1, p_2, \ldots, p_k) / \sum_{i=1}^{k} p_i \tag{8}$$

where p_i are the frequencies calculated from read counts before the filter. In this manner, the original frequencies are simply re-scaled. The truncation represents a loss of information, in the sense that below the error level, whatever it may be, there are authentic haplotypes that would equally be rejected.

4.6. Sample size dependence

Diversity indices are dependent to a varying extent on the sample size [4], and this dependence has to be taken into account when comparing values from different samples. In the present study, we used rarefaction to correct differences due to sample size. From the set of samples to be compared, the minimum coverage was taken as the sample size reference. Each sample then underwent 1000 resampling cycles (i.e., sampling with repositioning), taking the frequencies of all haplotypes in the sample as the probabilities, and the minimum coverage in the set of samples to be compared as the sample size. In each cycle, each diversity index was calculated from the resulting sampling. At the end of the 1000 cycles, averages and standard deviations were computed for each diversity index in the study. In this way all diversity index values were referenced to the same sample size (coverage).

In previous research, we suggested a faster alternative to this process, calling it down-sampling with fringe trimming [3, 4]. Although it provided good correction of sample size bias, it resulted in a loss of information that can be critical in some situations.

4.7. Distance between quasispecies, quasispecies dendrograms, and multidimensional scaling plots (MDS)

A quasispecies can be seen as a genetic population, and the methods used to study diversification in a genetic population can also be used to determine the distance or dissimilarity between different quasispecies. There are several useful methods to quantify distance. When examining the Hamming distance, or any genetic distance, between pairs of haplotypes in two populations (quasispecies), the nucleotide divergence (D_A) formula by Matoshi Nei [51] measures the net genetic distance between them, correcting the full genetic distance by subtracting their mean intra-population genetic diversity. This same method can be applied to calculate the inter-quasispecies phenotype distance, in this case substituting the matrices of genetic distances between pairs of haplotypes for the matrices of distances between amino acid haplotypes. The distance between proteins can be computed by the method of Grishin [52] to obtain dissimilarities between proteins from substitution matrices, such as PAM or BLOSUM. Alternatively, they can be computed directly from matrices of distances between pairs of amino acids, using methods such as that of Fitch [53], or Grantham [54]. When the quasispecies are closely related, as in the HEV follow-up study here, an alternative type of distance or dissimilarity of interest can be obtained directly from the haplotype frequencies of the two quasispecies, regardless of the genetic distance between them, by applying the method of Yue and Clayton [55]. This method can also be applied to the quasispecies fitness fractions introduced above.

The distances or dissimilarities obtained can be used to plot quasispecies dendrograms or trees, or multidimensional scaling maps, to help visualize how the quasispecies in a study are related.

4.8. Software and statistics

All computations were done in R (v4.0.3) [56] with in-house scripts, using the Biostrings [57], ShortRead [58], and QSutils [59] packages from Bioconductor [60], as well as ape [61], tidyverse [62], and ggplot2 [63].

Supplementary Materials

The following supporting information can be downloaded at:
https://www.mdpi.com/article/10.3390/ijms232314654/s1

Author contributions

JG: Conceptualization, methodology, software, formal data analysis, visualization, writing original draft; SC-C: Methodology, investigation, visualization, writing original draft; CC: Investigation, visualization; MI-L: Software, data curation;

DG-C: Methodology, investigation, project administration; AR-S: Samples, investigation, and resources; CMA: Investigation, visualization; RP: Conceptualization, resources, writing review and editing; SG: Methodology, writing – review; AB: Supervision, funding acquisition; ED: Supervision, writing – review and editing; IG: Resources, data curation; CP: Resources, formal analysis; MFC: Methodology, visualization; DT: Investigation, writing editing; MB: Conceptualization, resources, writing – review and editing; MR-B: Resources, data analysis, validation; JIE: Conceptualization, resources, validation, writing review and editing; FR-F: Conceptualization, funding acquisition, data analysis, writing review and editing; JQ: Conceptualization, methodology, validation, supervision, funding acquisition, and writing original draft.

All authors have read and agreed to the published version of the manuscript.

Funding

This study was partially supported by Plan Estratègic de Recerca i Innovació en Salut (PERIS) – Direcció General de Recerca i Innovació en Salut (DGRIS), Catalan Health Ministry, Generalitat de Catalunya; Centro para el Desarrollo Tecnológico Industrial (CDTI) from the Spanish Ministry of Economy and Business, grant number IDI-20200297; grant PI19/00301 from Instituto de Salud Carlos III cofinanced by the European Regional Development Fund (ERDF), and Gilead's biomedical research project GLD21/00006. Work at CBMSO in Madrid was supported by grant PID2020-113888RB-100 from Ministerio de Ciencia e Innovación and S2018/BAA-4370 (PLATESA2 from Comunidad de Madrid/FEDER).

Institutional review board statement

The study was approved by Vall d'Hebron University Hospital Ethics Committee, with reference number PR(AG)259-2020.

Informed consent statement

Informed consent was obtained from all subjects involved in the study.

Data availability statement

The genomic nucleotide sequences included in this study have been submitted in the GENBank repository database as BioProject ID PRJNA876218.

Acknowledgments

We thank Celine Cavallo for English language support.

Conflicts of interest

The authors declare no conflict of interest.

REFERENCES

1. Domingo E., Escarmis C., Lazaro E., Manrubia S.C. Quasispecies dynamics and RNA virus extinction. Virus Res 2005; 107: 129-139.
2. Domingo, E. Virus as populations. Composition, complexity, dynamics and biological implications. In: Virus as populations, 1st ed. London: Academic Press, 2016; 1-412.
3. Gregori J., Salicrú M., Domingo E. et al. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics 2014; 30: 1104-1111.
4. Gregori J., Perales C., Rodriguez-Frias F., Esteban J.I., Quer J., Domingo E. Viral quasispecies complexity measures. Virology 2016; 493: 227-237.
5. Vellas C., Del Bello A., Debard A. et al. Influence of treatment with neutralizing monoclonal antibodies on the SARS-CoV-2 nasopharyngeal load and quasispecies. Clin Microbiol Infect 2022; 28: 139.e5-139.e8.
6. Perales C., Martín V., Ruiz-Jarabo C.M., Domingo E. Monitoring sequence space as a test for the target of selection in viruses. J Mol Biol 2005; 345: 451-459.
7. Gregori J., Soria M.E., Gallego I. et al. Rare haplotype load as marker for lethal mutagenesis. PLoS One 2018; 13: e0204877.
8. Perales C., Gallego I., de Ávila A.I. et al. The increasing impact of lethal mutagenesis of viruses. Future Med Chem 2019; 11: 1645-1657.
9. Nakano K., Shiroma A., Shimoji M. et al. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Hum Cell 2017; 30: 149-161.
10. Bull R.A., Eltahla A.A., Rodrigo C. et al. A method for near full-length amplification and sequencing for six hepatitis C virus genotypes. BMC Genom 2016; 17: 247.
11. Dilernia D.A., Chien J.T., Monaco D.C. et al. Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing. Nucleic Acids Res 2015; 43: e129.
12. Stoler N., Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. NAR Genom Bioinforma 2021; 3: lqab019.
13. Amarasinghe S.L., Su S., Dong X., Zappia L., Ritchie M.E., Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol 2020; 21: 30.
14. Wenger A.M., Peluso P., Rowell W.J. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 2019; 37: 1155-1162.
15. Hill M.O. Diversity and evenness: a unifying notation and its consequences. Ecology 1973; 54: 427-432.
16. Carcereny A., Martínez-Velázquez A., Bosch A. et al. Monitoring emergence of the SARS-CoV-2 B.1.1.7 variant through the Spanish National SARS-CoV-2 Wastewater Surveillance System (VATar COVID-19). Environ Sci Technol 2021; 55: 11756-11766.
17. Twist Synthetic SARS-CoV-2 RNA Control 2 MN908947.3. Available online: https://www.twistbioscience.com/es/resources/product-sheet/twist-synthetic-sars-cov-2-rna-controls (accessed on 12 November 2022).
18. Cubero M., Gregori J., Esteban J.I. et al. Identification of host and viral factors involved in a dissimilar resolution of a hepatitis C virus infection. Liver Int 2014; 34: 896-906.
19. Baccam P., Thompson R.J., Fedrigo O., Carpenter S., Cornette J.L. PAQ: Partition Analysis of Quasispecies. Bioinformatics 2001; 17: 16-22.
20. Töpfer A., Marschall T., Bull R.A., Luciani F., Schönhuth A., Beerenwinkel N. Viral quasispecies assembly via maximal clique enumeration. PLoS Comput Biol 2014; 10: e1003515.
21. Skums P., Zelikovsky A., Singh R. et al. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics 2018; 34: 163-170.
22. Ahn S., Ke Z., Vikalo H. Viral quasispecies reconstruction via tensor factorization with successive read removal. Bioinformatics 2018; 34: i23-i31.
23. Henningsson R., Moratorio G., Bordería A.V., Vignuzzi M., Fontes M. DISSEQT-DIStribution-based modeling of SEQuence space Time dynamics. Virus Evol 2019; 5: vez028.
24. Beerenwinkel N., Zagordi O. Ultra-deep sequencing for the analysis of viral populations. Curr Opin Virol 2011; 1: 413-418.
25. Lorenzo-Redondo R., Delgado S., Morán F., Lopez-Galindez C. Realistic three dimensional fitness landscapes generated by self organizing maps for the analysis of

experimental HIV-1 evolution. PLoS One 2014; 9: e88579.

26. Delgado S., Perales C., García-Crespo C. et al. A two-level, intramutant spectrum haplotype profile of hepatitis C virus revealed by self-organized maps. Microbiol Spectr 2021; 9: e0145921.

27. Gregori J., Cortese M.F., Piñana M. et al. Host-dependent editing of SARS-CoV-2 in COVID-19 patients. Emerg Microbes Infect 2021; 10:1777-1789.

28. Martínez-González B., Soria M.E., Vázquez-Sirvent L. et al. SARS-CoV-2 mutant spectra at different depth levels reveal an overwhelming abundance of low frequency mutations. Pathogens 2022; 11: 662.

29. De Avila A.I., Gallego I., Soria M.E. et al. Lethal mutagenesis of hepatitis C virus induced by favipiravir. PLoS One 2016; 11: e0164691.

30. Perales C., Agudo R., Tejero H., Manrubia S.C., Domingo E. Potential benefits of sequential inhibitor-mutagen treatments of RNA virus infections. PLoS Pathog 2009; 5: e1000658.

31. Githaka J.M. Molnupiravir does not induce mutagenesis in host lung cells during SARS-CoV-2 treatment. Bioinform Biol Insights 2022; 16: 11779322221085076.

32. Gordon C.J., Tchesnokov E.P., Schinazi R.F., Götte M. Molnupiravir promotes SARS-CoV-2 mutagenesis via the RNA template. J Biol Chem 2021; 297: 100770.

33. Cameron C.E., Castro C. The mechanism of action of ribavirin: lethal mutagenesis of RNA virus genomes mediated by the viral RNA-dependent RNA polymerase. Curr Opin Infect Dis 2001; 14: 757-764.

34. Dietz J., Schelhorn S.-E., Fitting D. et al. Deep sequencing reveals mutagenic effects of ribavirin during monotherapy of hepatitis C virus genotype 1-infected patients. J Virol 2013; 87: 6172-6181.

35. Cuevas J.M., González-Candelas F., Moya A., Sanjuán R. Effect of ribavirin on the mutation rate and spectrum of hepatitis C virus in vivo. J Virol 2009; 83: 5760-5764.

36. Todt D., Meister T.L., Steinmann E. Hepatitis E virus treatment and ribavirin therapy: viral mechanisms of nonresponse. Curr Opin Virol 2018; 32: 80-87.

37. Todt D., Gisa A., Radonic A. et al. In vivo evidence for ribavirin-induced mutagenesis of the hepatitis E virus genome. Gut 2016; 65: 1733-1743.

38. Debing Y., Gisa A., Dallmeier K. et al. A mutation in the hepatitis E virus RNA polymerase promotes its replication and associates with ribavirin treatment failure in organ transplant recipients. Gastroenterology 2014; 147: 1006-1008.

39. Lhomme S., Kamar N., Nicot F. et al. Mutation in the hepatitis E virus polymerase and outcome of ribavirin therapy. Antimicrob Agents Chemother 2015; 60: 1608-1614.

40. Debing Y., Ramière C., Dallmeier K. et al. Hepatitis E virus mutations associated with ribavirin treatment failure result in altered viral fitness and ribavirin sensitivity. J Hepatol 2016; 65: 499-508.

41. ARTIC Network. Available online: https://artic.network/ncov-2019 (accessed on 12 November 2022).

42. Andrés C., Garcia-Cehic D., Gregori, J. et al. Naturally occurring SARS-CoV-2 gene deletions close to the spike S1/S2 cleavage site in the viral quasispecies of COVID19 patients. Emerg Microbes Infect 2020; 9: 1900-1911.

43. Gallego I., Gregori J., Soria M.E. et al. Resistance of high fitness hepatitis C virus to lethal mutagenesis. Virology 2018; 523: 100-109.

44. Gallego I., Sheldon J., Moreno E. et al. Barrier-independent, fitness-associated differences in sofosbuvir efficacy against hepatitis c virus. Antimicrob Agents Chemother 2016; 60: 3786-3793.

45. Marukian S., Jones C.T., Andrus L. et al. Cell culture-produced hepatitis C virus does not infect peripheral blood mononuclear cells. Hepatology 2008; 48: 1843-1850.

46. Perales C., Beach N.M., Gallego I. et al. Response of hepatitis C virus to long-term passage in the presence of alpha interferon: multiple mutations and a common phenotype. J Virol 2013; 87: 7593-7607.

47. Moreno E., Gallego I., Gregori J. et al. Internal disequilibria and phenotypic diversification during replication of hepatitis C virus in a noncoevolving cellular environment. J Virol 2017; 91: e02505-16.

48. Magoc T., Salzberg S.L. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 2011; 27: 2957-2963.

49. Chao A., Gotelli N.J., Hsieh T.C. et al. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol Monogr 2014; 84: 45-67.

50. Gregori J., Esteban J.I., Cubero M. et al. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One 2013; 8: e0083361.

51. Nei M. Molecular evolutionary genetics. New York, NY: Columbia University Press, 1987.

52. Grishin V.N., Grishin N. V. Euclidian space and grouping of biological objects. Bioinformatics 2002; 18: 1523-1534.

53. Fitch W.M. An improved method of testing for evolutionary homology. J Mol Biol 1966; 16: 9-16.

54. Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974; 185: 862-864.

55. Yue J.C., Clayton M.K. A similarity measure based on species proportions. Commun Stat-Theory Methods 2005; 34: 2123-2131.

56. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2019.

57. Pages H., Aboyoun P., Gentleman R., DebRoy S. Biostrings: string objects representing biological sequences, and matching algorithms. R package 2.38.4; 2012. Available online: https://bioc.ism.ac.jp/packages/3.2/bioc/html/Biostrings.html (accessed on 12 November 2022).

58. Morgan M., Anders S., Lawrence M., Aboyoun P., Pages H., Gentleman R. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics 2009; 25: 2607-2608.

59. Guerrero-Murillo M., Gregori J. QSutils: quasispecies diversity. R Package Version 1.0.0; 2018. Available online: https://bioconductor.org/packages/release/bioc/html/QSutils.html_2018 (accessed on 22 November 2022).

60. Gentleman R.C., Carey V.J., Bates D.M. et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004; 5: R80.

61. Paradis E., Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 2019; 35: 526-528.

62. Wickham H. Welcome to master the tidyverse. J Open Source Softw 2019; 4.

63. Valero-Mora P.M. ggplot2: elegant graphics for data analysis. J Stat Soft Book Rev 2010; 35: 1-3.

Supplementary Materials

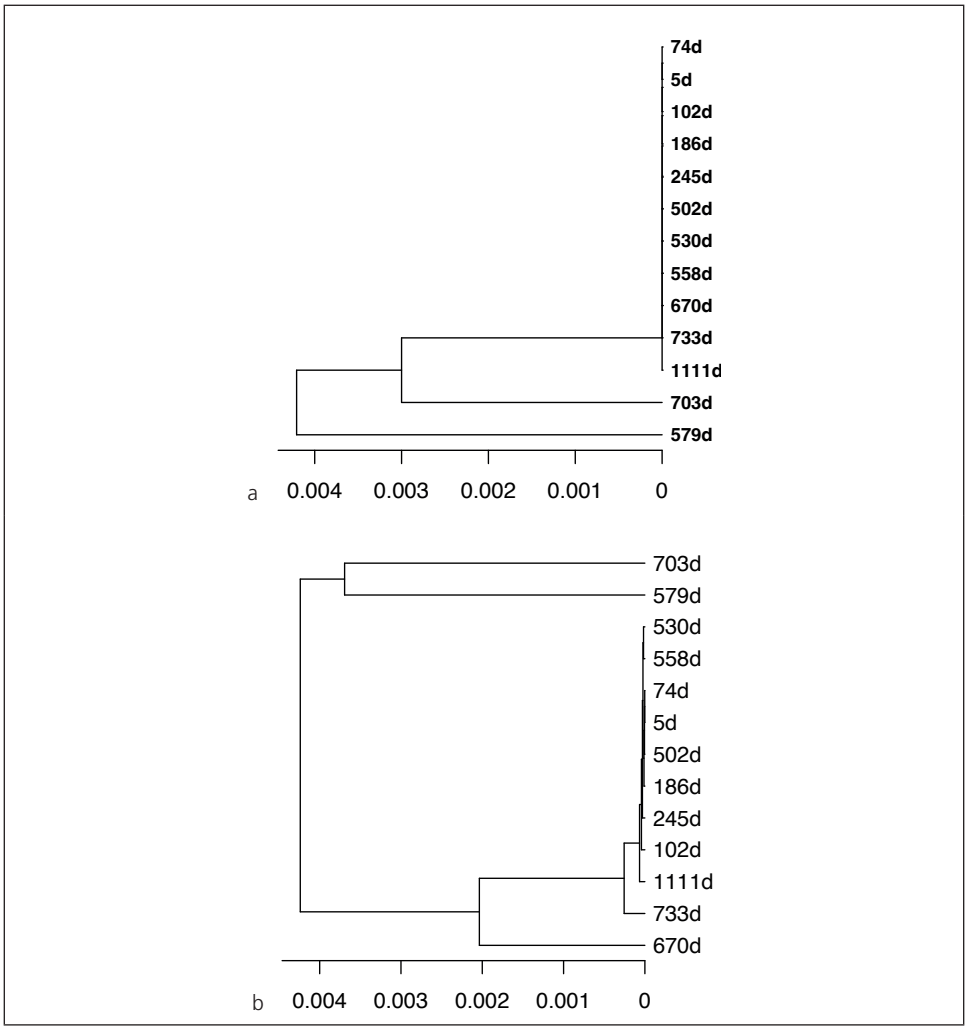


Figure S1. UPGMA tree of the master phenotypes based on Grantham amino acid distances (a), and quasispecies tree based on the D_A population distances (b), taking the top 20 phenotypes in each sample.

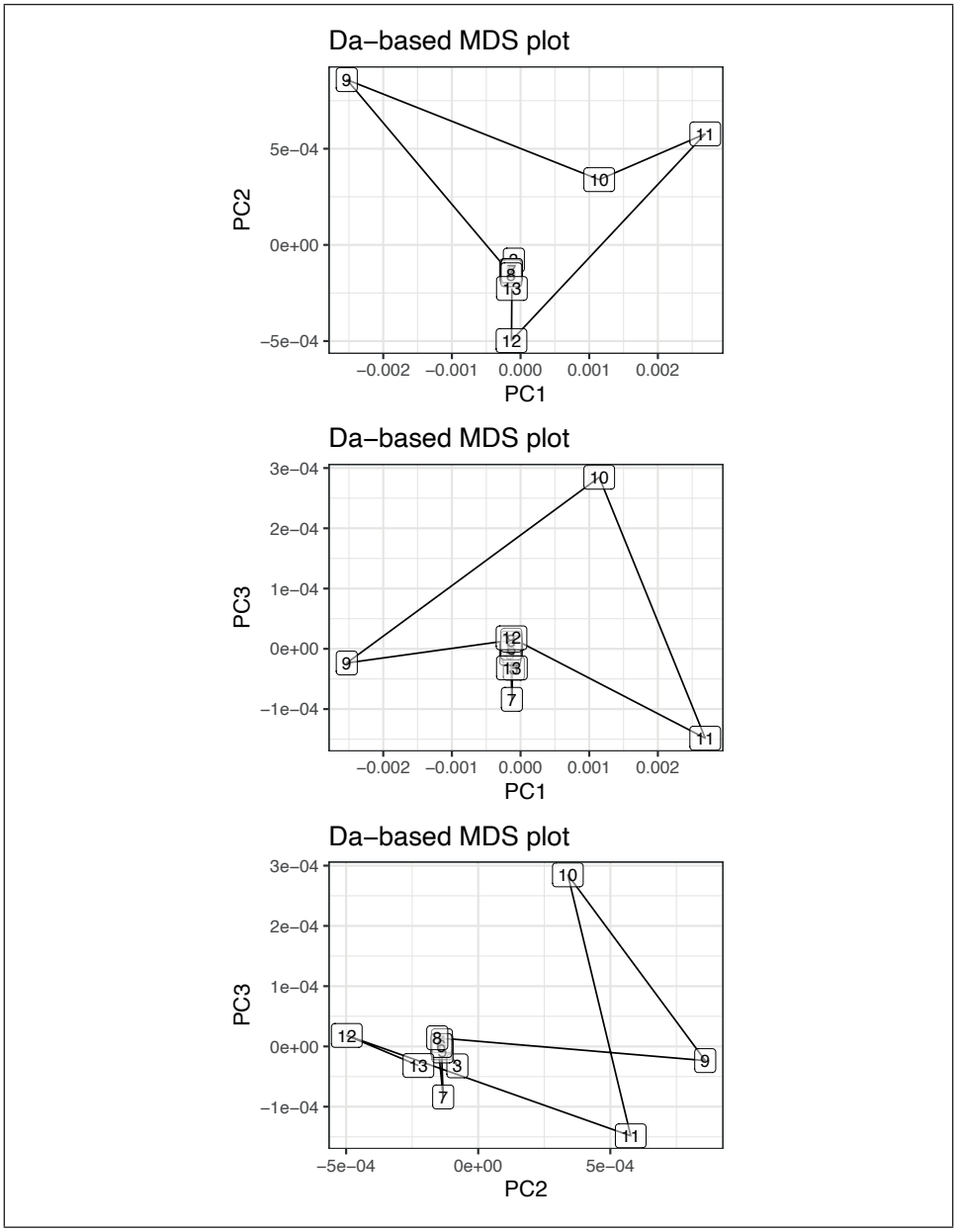


Figure S2. Multidimensional scaling plot of phenotype distances between quasispecies, taking the top 20 phenotypes in each sample.

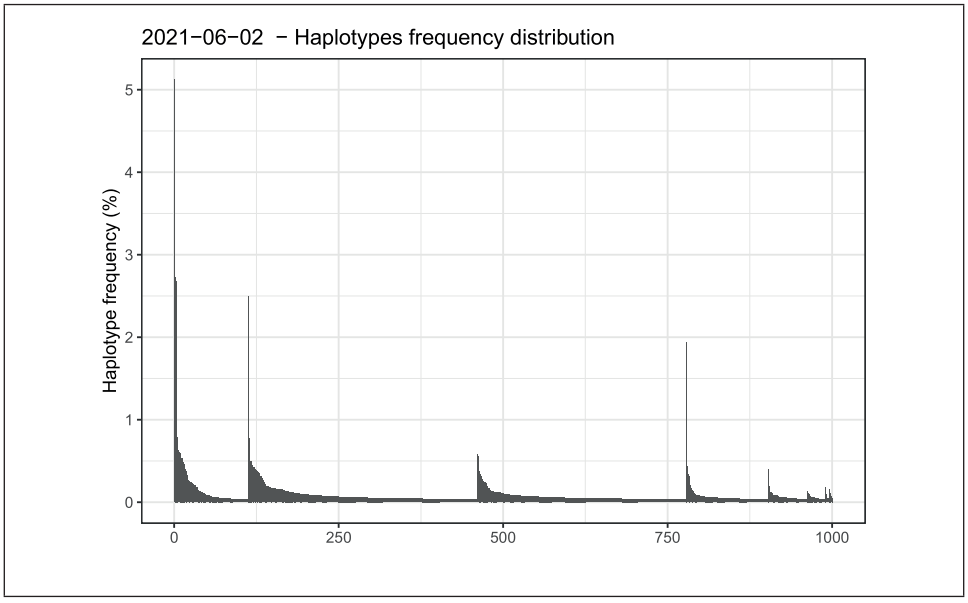


Figure S3. Montserrat plot with the distribution of the 1000 most abundant haplotypes in the last sample (1111 days since diagnosis, 2 June 2020).

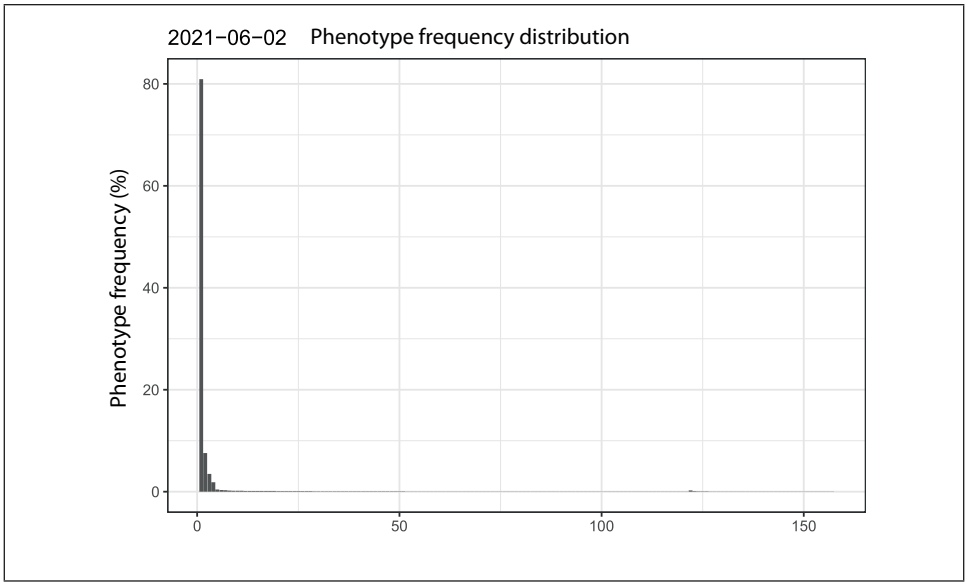


Figure S4. Montserrat plot with the distribution of the 1000 most abundant phenotypes in the last sample (1111 days since diagnosis, 2 June 2020).

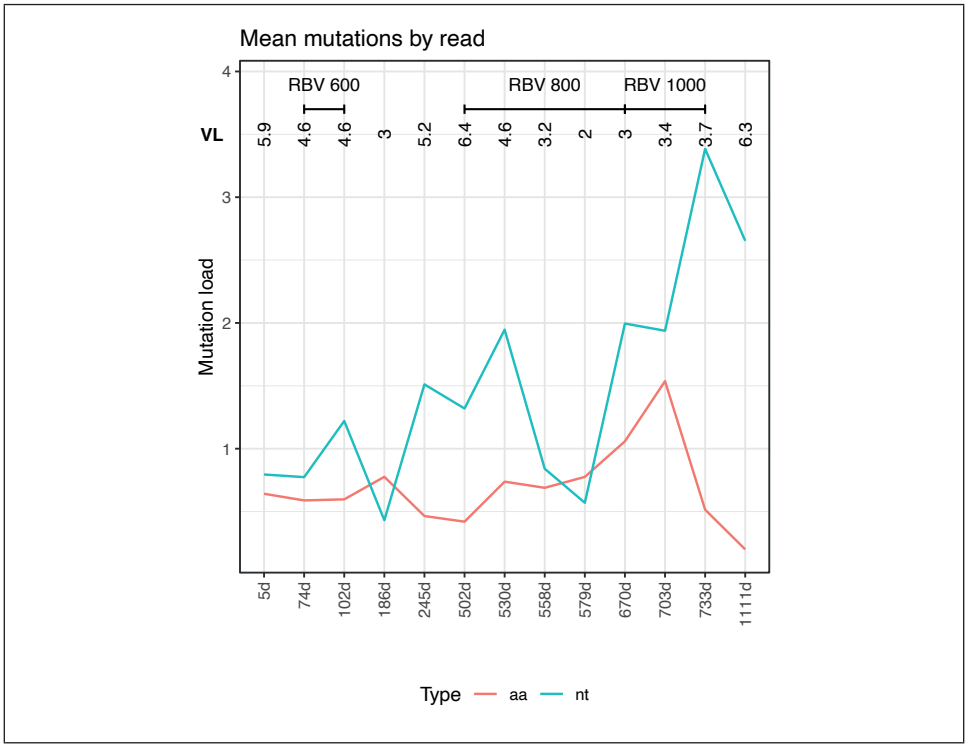


Figure S5. Mean number of substitutions per read with respect to the master haplotype, at the nucleotide level (turquoise), and mean number of mutations per read with respect to the master phenotype at the amino acid level (orange).

Section 6

Similarity between haplotype distributions

Abstract

The molecules in a viral quasispecies can be regarded as individuals from competing species in a single ecosystem, where the species are the various haplotypes and the ecosystem is the host. The status of a quasispecies is represented by the frequencies of the existing haplotypes, a multinomial distribution where each category corresponds to a different haplotype. The same approach can be considered for phenotype distribution, where applicable. The changes or evolution a quasispecies undergoes in a host can be monitored by determining the relative frequencies of haplotypes in samples obtained at established intervals (sequential sampling), and computing the similarity or distance between pairs of sequential samples. Three indices of similarity and their corresponding distances are examined in this study. The index of common molecules (C_m), the index of distribution overlap (O_v) and the Yue-Clayton index (Y_C). The mutual correlations between these indices, and their correlation with the genetic distance between populations (D_A), are reported based on extensive simulated data. Some examples are tabulated and plotted to help understand the results. The proposed distances can then be used to obtain plots such as quasispecies dendrograms or multidimensional scaling plots to help visualize the changes that have occurred. These methods are illustrated using simulated data of quasispecies evolution, and clinical data from a chronically infected HEV patient treated with three different mutagen regimens and followed for three years.

Highlights

- C_m expresses how two quasispecies are related, in the sense of having common haplotypes, even when the haplotype frequencies differ. When all haplotypes in two quasispecies are identical, the index yields a value of 1, even though the relative proportions may differ considerably.
- O_v expresses to what degree two distributions are similar, both in haplotypes and frequencies. O_v may yield low values even when all haplotypes of two quasispecies are identical.
- Y_C results in high values when the fraction of common haplotypes is high and their proportions are similar. Weakly sensitive to low frequency haplotypes, its value is driven by the most frequent haplotypes in both quasispecies.
- The distances C_m and D_A show the weakest correlations.
- The pairs O_v and Y_C, O_v and D_A, and Y_C and D_A show the strongest correlations.
- Despite the correlations, we recommend the use of all these distances, C_m, O_v, Y_C, and D_A. They are sensitive to different aspects of quasispecies composition similarity, and their relationship is indicative of the type of changes arisen between the two samples.

6.
Quantifying in-host quasispecies evolution

JOSEP GREGORI, MARTA IBAÑEZ-LLIGOÑA, JOSEP QUER

ABSTRACT

What takes decades, centuries or millennia to happen with a natural ecosystem, it takes only days, weeks or months with a replicating viral quasispecies in a host, especially when under treatment. Some methods to quantify the evolution of a quasispecies are introduced and discussed, along with simple simulated examples to help in the interpretation and understanding of the results. The proposed methods treat the molecules in a quasispecies as individuals of competing species in an ecosystem, where the haplotypes are the competing species, and the ecosystem is the quasispecies in a host, and the evolution of the system is quantified by monitoring changes in haplotype frequencies. The correlation between the proposed indices is also discussed, and the R code used to generate the simulations, the data and the plots is provided. The virtues of the proposed indices are finally shown on a clinical case.

Keywords

Quasispecies evolution, distributions similarity, quasispecies fitness partition, viral treatment, mutagenesis.

1. Introduction

All viruses that pass through an RNA replication phase are found in what is known as a quasispecies. That is, a set of closely related genomes that may exhibit a huge number of variants but keeping a high degree of similarity among them in a host. These variants are produced during the replication by the RNA-dependent RNA polymerases, which are error prone and lack the mechanism of error correction typical in most DNA polymerases [1].

Quasispecies are dynamical entities subject to evolution, generating new variants at each replication cycle, while losing the less fit and those unable to replicate. A qua-

sispecies at a given time point may be described in molecular terms by the existing different genomes (haplotypes) and their frequencies (the number or fraction of molecules with the same sequence), the haplotype distribution. That is, a multinomial distribution where each category corresponds to a different haplotype. The evolution of this dynamic entity may then be represented by the changes observed in this distribution, as new categories appear and others disappear, and as their frequencies vary.

The extent of changes of a quasispecies in a host, between two time points, may be quantified by the genetic distance between the two viral populations [2], by the changes in quasispecies diversity indices [3], and by the distance or dissimilarity between the two haplotype distributions [4]. In this report, we discuss three selected indices used to compute the similarity between two haplotype distributions and their implications. With quasispecies simulated data, we show their particularities and correlations, and use plots to help in the interpretation of results. Finally, a clinical HEV dataset, from a recent publication, is used to illustrate the practical use of these indices. They are particularly useful in the clinical follow up of a patient, where the compared quasispecies are highly related, and where the genetic distance between them may not suffice to describe the observed changes.

In the context of NGS, we denote each distinct genome as an *haplotype*, and each molecule sampled as a *read*. We shall be using this terminology throughout the paper.

2. Results

2.1. Simulated pairs of quasispecies

To quantify the extent of changes in a quasispecies, we compare the quasispecies composition at two time points. The pairs of quasispecies used to illustrate the results and discussion are obtained by a simple simulation with a limited number of haplotypes whose frequencies vary randomly within given constraints, and where a random number of these haplotypes are common to both quasispecies. The simulation aims to obtain closely related quasispecies as we could find, a few weeks or months apart, in a host. We simulate 10,000 pairs of related quasispecies, computing their similarity, and genetic distance. The simulated pairs are illustrated in the form of a table and a figure, confronting the haplotype distributions in both quasispecies, as in Table 1 and Figure 1.

The index of Commons, C_m (Equation (1)), is strongly indicative of quasispecies relatedness. When the two quasispecies have all their haplotypes identical, this index results in a value of 1, even when the proportions are highly dissimilar. On the other hand, the Overlap index, O_v (Equation (2)), may result in low values even when all haplotypes of both quasispecies are identical. Finally, the Yue–Clayton index, Y_C (Equation (3)), results in high values when the fraction of common haplotypes is high, and their proportions are similar. The overlap between distributions is better illustrated with a plot like Supplementary Figure S1.

Table 1. Two closely related quasispecies. *Hpl* haplotype ID, *nA* reads in quasispecies A, *nB* reads in quasispecies B, *pA* and *pB* corresponding frequencies (%).

Hpl	*nA*	*nB*	*pA*	*pB*
A	2000	1400	50.06	38.36
B	600	800	15.02	21.92
C	400	310	10.01	8.49
D	100	0	2.5	0
E	0	70	0	1.92
F	0	30	0	0.82
G	15	0	0.38	0
H	600	750	15.02	20.55
I	200	220	5.01	6.03
J	50	0	1.25	0
K	0	70	0	1.92
L	30	0	0.75	0

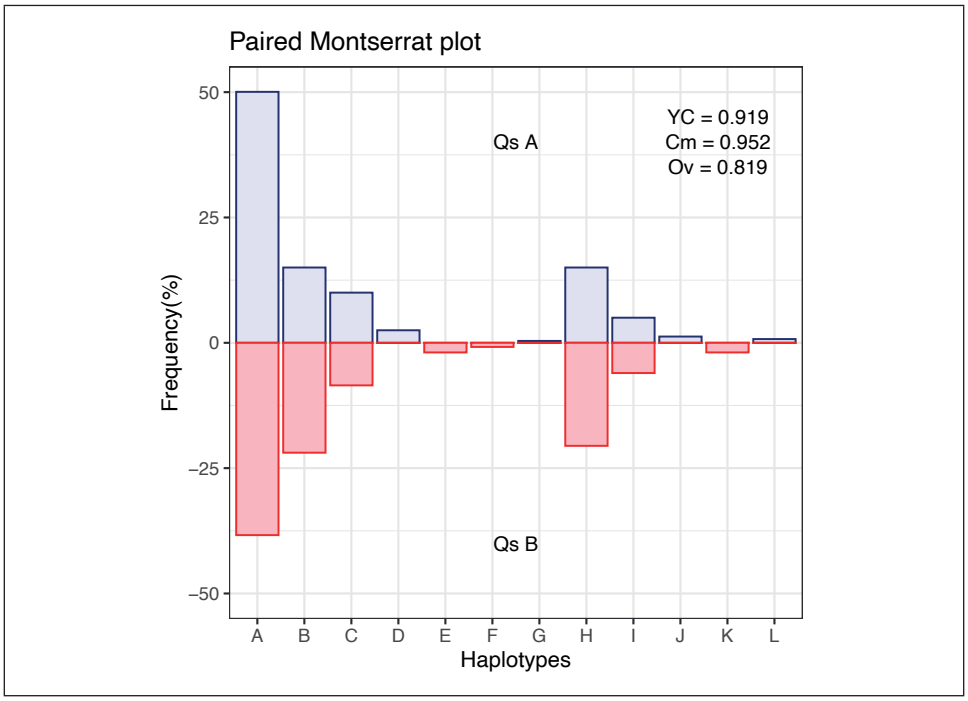


Figure 1. Montserrat plot with paired haplotype distribution.

A summary of the values of similarity indices obtained from the simulated quasispecies pairs is given in Table 2, along with the number of pairs resulting in a similarity value over 0.5, 0.75 and 0.9. The histograms for the three indices are given in Supplementary Figure S2. On the other hand, Table 3 and Figure 2 show the distribution of the three indices for the 2698 simulated pairs resulting in C_m values over 0.75, that is, highly related. The histograms for the corresponding nucleotide diversities and genetic distances (Equations (4)-(7)) are given in Supplementary Figure S3.

Table 2. Summary of similarity values between pairs of quasispecies.

	O_v	C_m	Y_c
Min.	0.00070	0.01075	0.00000
1stQ	0.04237	0.46139	0.00994
Median	0.10081	0.61189	0.03628
Mean	0.14544	0.60412	0.09800
3rdQ	0.20982	0.76187	0.12192
Max.	0.84055	1.00000	0.97111
Over 0.50	245	6944	336
Over 0.75	10	2698	62
Over 0.90	0	692	12

Table 3. Summary of similarity values with a C_m over 0.75.

	O_v	C_m	Y_c
Min.	0.0034	0.7500	0.0005
1stQ	0.0886	0.7923	0.0336
Median	0.1891	0.8395	0.1049
Mean	0.2289	0.8501	0.1863
3rdQ	0.3395	0.9022	0.2737
Max.	0.8405	1.0000	0.9711

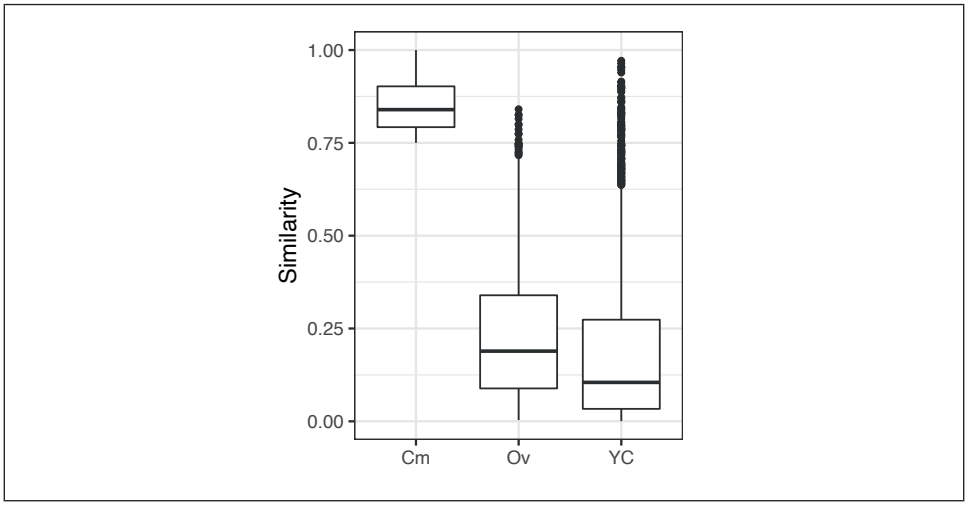


Figure 2. Boxplots with the distributions of the three indices, for all simulated pairs which result in values of C_m over 0.75.

2.1.1. Correlations

The eventual redundancy in the information provided by these indices, and by the genetic distance, may be assessed by inspecting the correlation coefficient between them, taking the 10,000 simulated values, as in Table 4.

Table 4. Correlation between similarity indices and with genetic distance.

	C_m	O_v	Y_C	D_A
C_m	1.0000	0.4616	0.4256	–0.3442
O_v	0.4616	1.0000	0.9372	–0.7961
Y_C	0.4256	0.9372	1.0000	–0.8011
D_A	–0.3442	–0.7961	–0.8011	1.0000

These correlations may be further illustrated by the joint density plots in Figures 3-6, from which result the following observations:

- C_m and O_v: at high O_v values only high values of C_m may occur, on the other hand at very low O_v values almost all C_m values are possible. This is consistent with the definition of both indices.
- Y_C and O_v: are highly correlated and seemingly do not convey significant additional information.

- C_m and D_A: at low C_m values, D_A takes high values, but at high C_m values, D_A spans the highest range of values; the lower D_A values correspond to high values of C_m.
- O_v and D_A: at high O_v values, D_A takes the lower values, but at low O_v values, D_A spans a high range of values.

The information provided by C_m, O_v, and D_A complement each other and offer different faces of the same comparison. C_m expresses how related the two quasispecies are, in the sense of having common haplotypes, even if the frequencies are different. O_v expresses how similar both distributions are, both in haplotypes and frequencies. Additionally, D_A provides the net genetic distance between the two quasispecies, taken as populations of viruses.

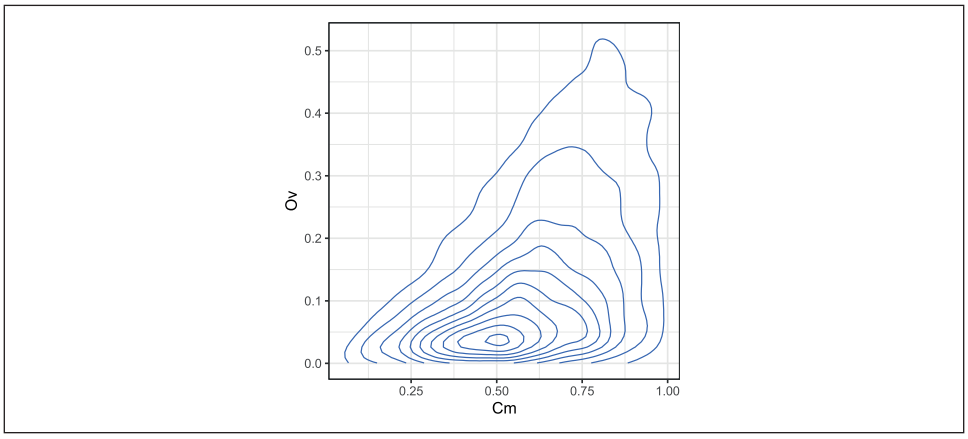


Figure 3. Joint density of C_m and O_v.

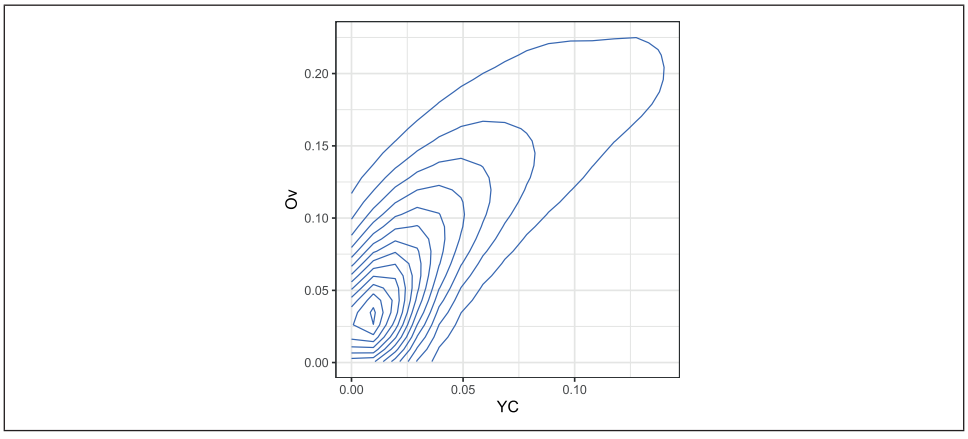


Figure 4. Joint density of Y_C and O_v.

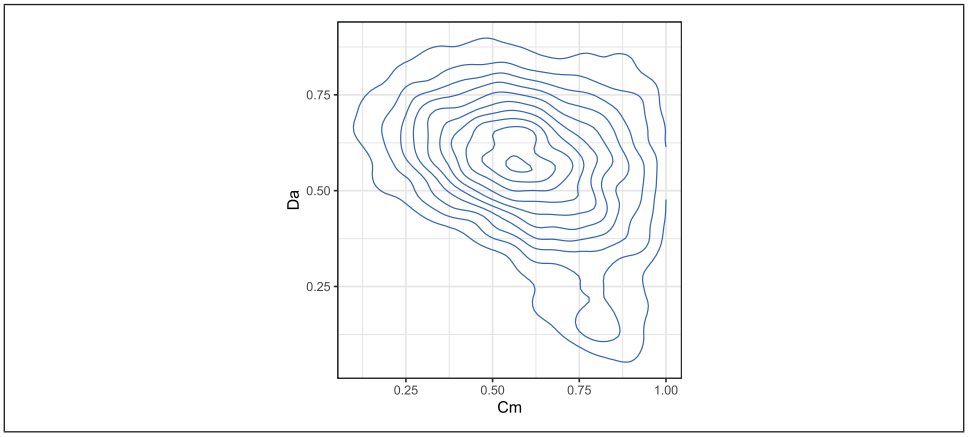


Figure 5. Joint density of C_m and D_A.

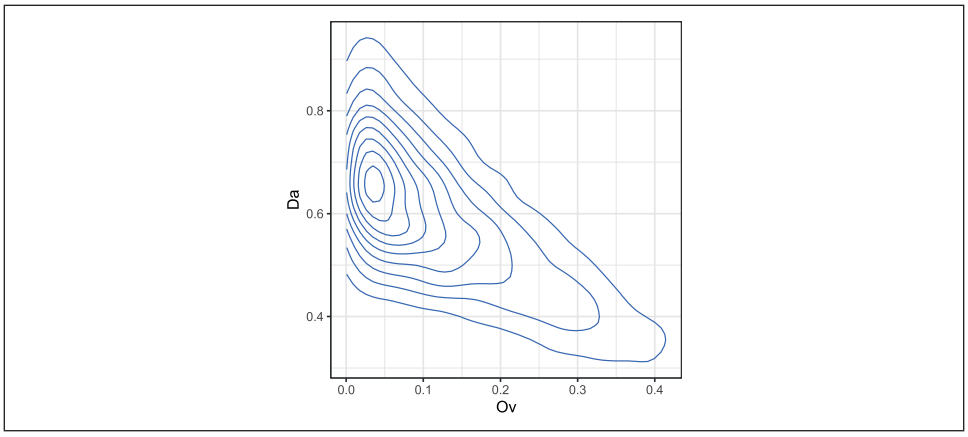


Figure 6. Joint density of O_v and D_A.

2.1.2. Illustration of selected pairs

From the set of simulated pairs, a few are selected attending to the values of the three indices, to help in the understanding and interpretation of these indices, and are plotted in Supplementary Figures S6-S14. Table 5 shows a summary of these examples.

To improve the visualization of haplotypes unique to either quasispecies, the proportions of both quasispecies are sorted according to the order in decreasing value of the proportions of the quasispecies A. The haplotypes unique to quasispecies B will be placed on the right of the plot, or at the bottom of the table.

Table 5. Summary of selected examples. *Idx* index of the simulated quasispecies pair, *nHpl* number of total haplotypes in the pair, nC_m number of haplotypes in common, C_m, O_v, and Y_C similarity indices, D_A genetic distance.

Idx	*nHpl*	nC_m	C_m	O_v	Y_C	D_A	Suppl. Figure
4213	16	8	0.9420	0.7477	0.8900	0.0180	S6
7426	15	9	0.9899	0.7232	0.8310	0.0711	S7
774	17	7	0.8611	0.0137	0.0071	0.7719	S8
5463	18	6	0.8521	0.0093	0.0039	0.6835	S9
3053	17	7	0.5741	0.2250	0.1789	0.4961	S10
5955	17	7	0.6312	0.1149	0.0417	0.5403	S11
1159	18	6	0.7983	0.5691	0.6454	0.1483	S12
2528	18	6	0.7946	0.6503	0.8052	0.0743	S13
345	16	8	0.7308	0.4823	0.6668	0.1199	S14

These selected pairs are shown on the joint density plots of C_m and D_A in Supplementary Figure S4, and C_m and O_v in Supplementary Figure S5, in order to illustrate their position with respect to the bulk of the simulation.

Supplementary Figures S6 and S7 show two typical examples with high values in the three indices.

Supplementary Figures S8 and S9 show two examples with high C_m but very low O_v and Y_C values. This situation arises when there is a number of common haplotypes with mid to high proportions in one quasispecies and very low in the second.

Supplementary Figures S10 and S11 show intermediate cases with C_m below 0.70, and feeble values of O_v and Y_C.

Finally, Supplementary Figures S12-S14 show cases with higher values of O_v and Y_C.

2.2. Simulated treatment

As described under Methods, the evolution of a quasispecies with a shrinking dominant haplotype, an emerging haplotype, and a set of minority haplotypes subject to quasispecies dynamics was simulated. The result is represented in Figure 7, where the evolution in the frequencies of each of the 40 haplotypes is shown at each of the simulated evolution steps, with the corresponding dominant haplotype labelled with a + sign.

The fitness partition (QFP) analysis applied to the simulated samples in the follow-up example (Figure 8) show the four fitness fractions (QFF) in the form of a shrinking dominant haplotype in parallel with an increasing volume of molecules belonging to emer-

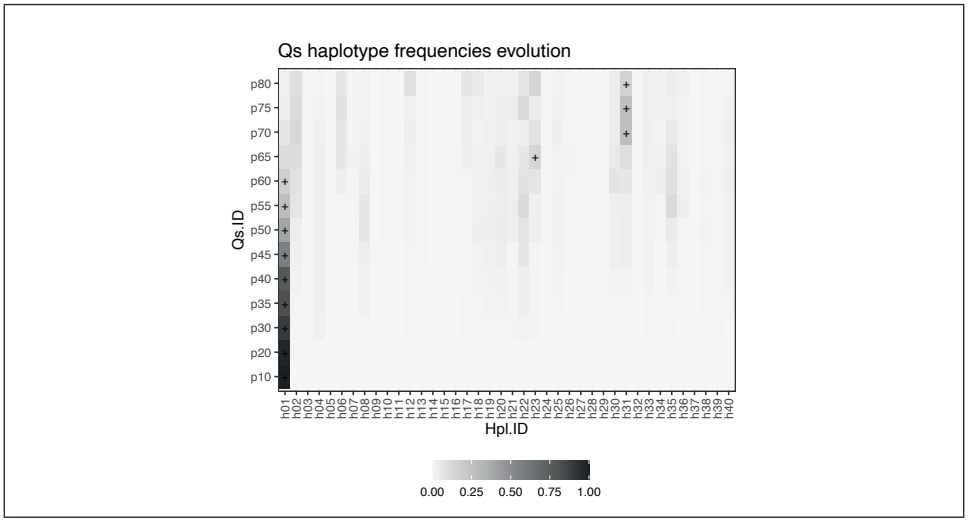


Figure 7. Haplotype distributions in the simulated follow-up example. The dominant haplotype at each step is labeled with a + sign.

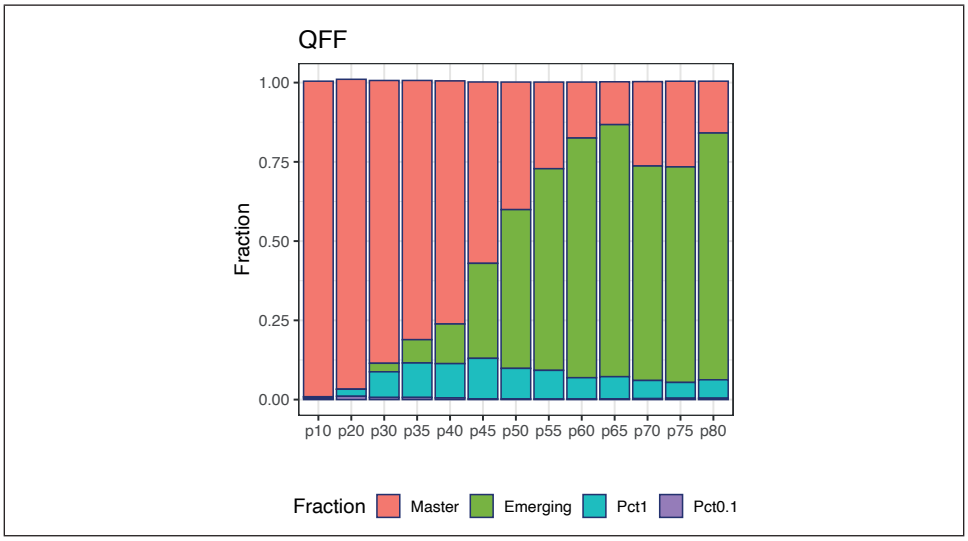


Figure 8. Quasispecies fitness partition of the simulated follow-up.

gent haplotypes as a side effect of a treatment, generating resistant variants. This figure constitutes a summary of the full quasispecies distributions illustrated in Figure 7.

The similarity indices discussed above (C_m Equation (1), O_v Equation (2), and Y_C Equation (3)), take values from 0 to 1, and may be easily transformed into distances

by the rule *Distance* = 1 – *Similarity*. Figure 9 illustrates the matrix of Yue-Clayton distances for the simulated treatment.

These distances may then be used to construct quasispecies dendrograms, or transformed by multidimensional scaling (MDS) to plot maps showing the relationships between the quasispecies. D_A genetic distances (Equation (7)) may be used in the same way to get dendrograms or MDS maps, as shown in Supplementary Figures S15-S18.

Note that by the very definition of the simulation used in the follow-up example, all quasispecies pairs show a C_m similarity index of 1, as all samples in the series share the same 40 haplotypes, although at varying frequencies. In real cases both the C_m, the O_v or the Y_C, and the D_A will be informative about the quasispecies evolution, showing different aspects about the changes produced. Additionally, the QFF contributes an interesting summary of quasispecies evolution.

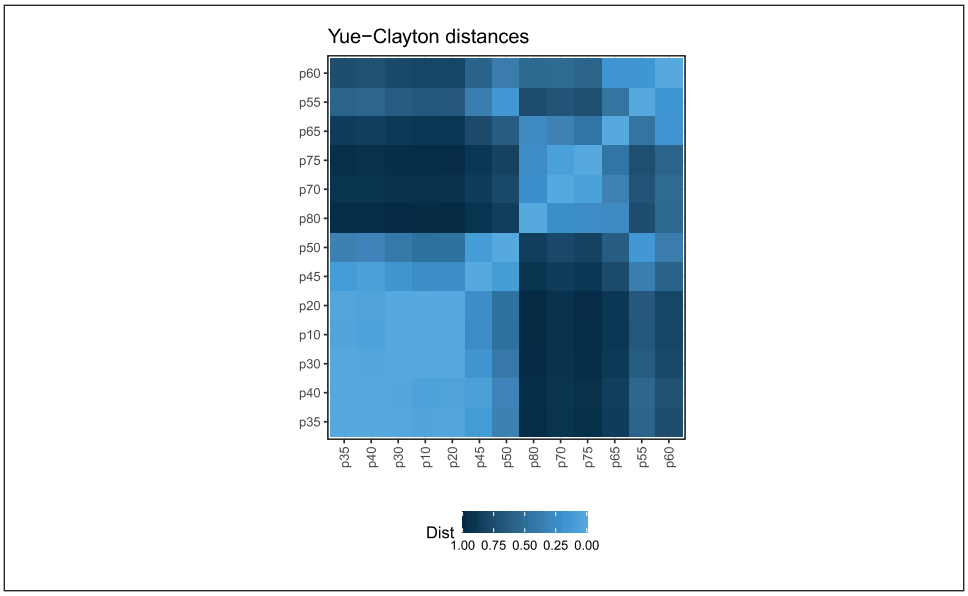


Figure 9. Matrix of Yue-Clayton distances between quasispecies haplotype distributions.

2.3. A clinical case

This is the clinical follow-up of a patient chronically infected by HEV who underwent an off-label treatment with ribavirin for three years [5]. The treatment involved three regimens (600, 800, and 1000 mg/day) with discontinuations caused by adverse effects, followed by relapses.

This dataset is of particular interest here, because it involves the follow-up of a patient infected by a zoonotic virus, HEV, treated with a mutagenic agent, with the fol-

low-up spanning over three years of treatment. In this case, the naturally high genetic diversity of HEV quasispecies is enhanced by the treatment with a mutagen.

The behavior of the three indices, in this case, is illustrated in Figure 10, where the similarities between each pair of sequential samples is shown, comparing sequential haplotype distributions on the left, and corresponding phenotype distributions on the right. The impact of the mutagenic treatment is evidenced by the sequential decrease in C_m, whose behavior is smoother than that of O_v or Y_C. The continued decrease in C_m value indicates that the proportion of molecules with sequences corresponding to haplotypes common to the two compared quasispecies is shrinking, consistent with the expected results of a mutagenic treatment, which generates new variants at an enhanced rate. The new variants will increase in abundance or fade according to their replicative fitness. The drop in C_m is especially marked when each treatment is initiated, especially those at 800 and 1000 mg/day, but these are followed by a small correction upwards. On the other hand, despite the radical changes observed in the haplotype composition, the analysis by phenotype composition shows that the functionality was maintained over a significant period of time, thanks to the generation of a rich set synonymous variants, and until the 800 mg/day regimen took effect. The changes observed in phenotype composition near the end of treatment, together with the observed increase in viral load may indicate that, either some resistance to the treatment was generated, or that the rich set of synonymous haplotypes generated and selected during the treatments contributed to generate a more resilient quasispecies [6]. The

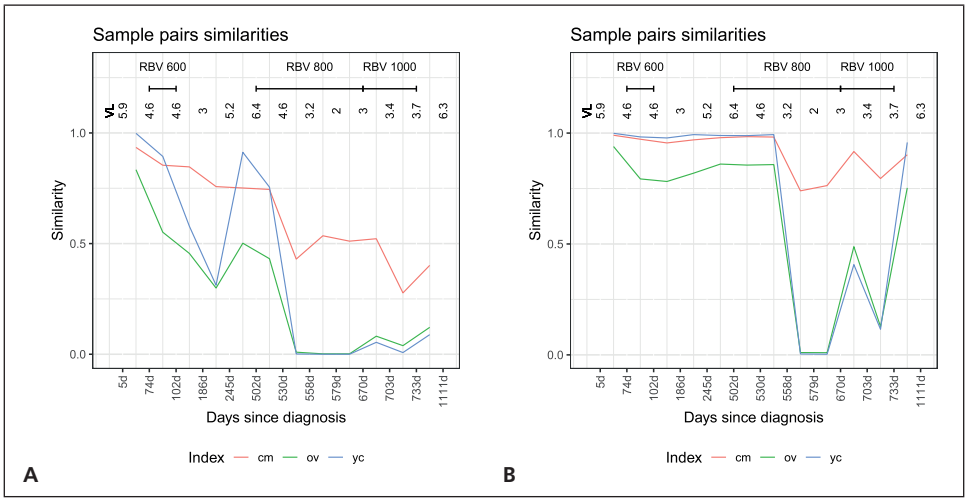


Figure 10. Distribution similarities in the composition of pairs of sequential samples in the HEV clinical case. (**A**) Haplotypes. (**B**) Phenotypes. Each point is the result of the comparison of two sequential samples, and is depicted in between the two compared samples. The segments above the figures show the time spanned by each treatment. Each sample is labeled as days since diagnosis. (VL viral load in logarithms, RBV ribavirin, cm C_m, ov O_v, yc Y_C).

similarity in the phenotype distribution between the end-of-treatment sample and that taken one year after is very high. This figure also shows that the indices O_v and Y_C are highly correlated, as previously shown with the simulated data. The C_m and O_v similarities in this dataset are plotted in Figure 11 over the 2D-density of the simulated data to show the correspondence between this clinical case and the simulated data.

The QFF profile of haplotypes and phenotypes of this case was presented and analyzed in the previous publication [5], and provides an interesting complementary and consistent view of this quasispecies evolution.

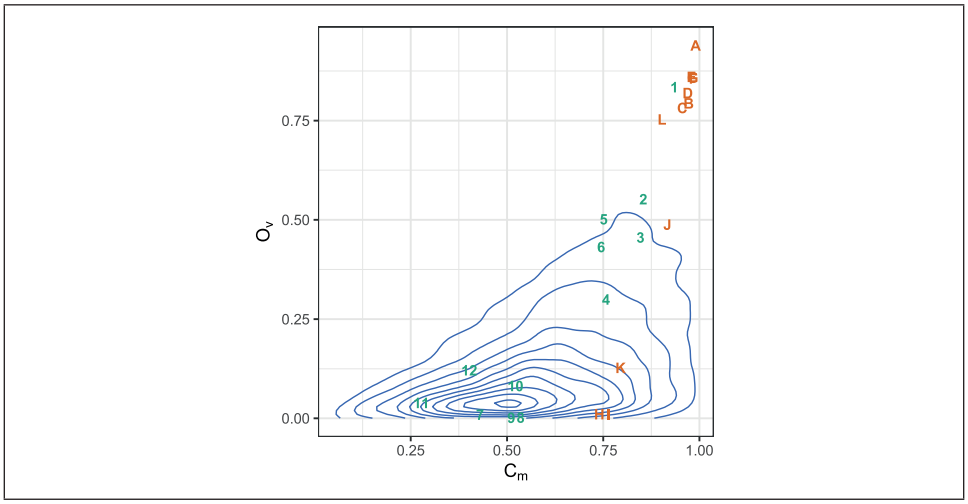


Figure 11. Observed similarities in the HEV clinical case plotted over the 2D-density of the simulated data. Similarities in haplotype distribution between pairs of sequential samples are labeled with numbers in increasing order, similarities in phenotype distribution are labeled with alphabetically ordered capital letters. The points on the top-right show high similarity by both indices.

3. Discussion

The proposed methods are intended to be used in the analysis of changes occurred in a in-host quasispecies along time, as a consequence of the host immune system or of an external action, like a treatment. The quasispecies are treated as entities (closed ecosystems or genetic populations), where the respective distribution of molecules is compared, in contrast with the more widespread comparison of summary values such as diversity indices (i.e., Shannon entropy), or of genetic diversification (i.e., nucleotide diversity) [3].

In a recent paper [5], we introduced the Quasispecies Fitness Partition (QFP) in four fractions (QFF), also described under Methods, and we recommended its use together with the Hill Numbers Profile (HNP) to visualize the evolution of a quasis-

pecies. Those methods were used in a deep exploration of a clinical case of an HEV infection treated with ribavirin. As part of the discussion, we proposed the use of distances between haplotype distributions as an alternative or complement to the use of genetic distances between quasispecies. This paper comes to explore three selected indices of similarity between haplotype distributions, from which the corresponding distances may be obtained.

Here, we have used simulated data aimed only at producing closely related quasispecies, similar to what could be observed in the follow-up of a single patient, with enough simplicity to be tabulated and plotted. However, to put in clinical context the methods here described, we have added the data of a clinical follow-up of an HEV chronically infected patient treated with a mutagen, spanning three years of observation, and different treatment regimens. Since HEV is an RNA virus having very high mutation rates, on the range of 10^{-3} to 10^{-4} substitutions/base/replication cycle [7], similar to other highly clinically relevant viruses such as HCV or HIV, the tools presented can be extrapolated to the vast majority if not all RNA viral infections.

The simulation of a substantial number of paired quasispecies allowed us to illustrate particular cases of interest, contributing to the interpretation of results, and also to estimate the correlations between the three indices (C_m, O_v and Y_C), and with the quasispecies genetic distance, D_A. The correlation values show the pairs O_v and Y_C, O_v and D_A, and Y_C and D_A as highly correlated, with C_m the most independent of the others. Despite this high correlation we recommend the use of three distances, C_m, O_v or Y_C, and D_A. Nevertheless, for distant quasispecies the four distances will contribute valuable information.

The use of these distances is shown with the simulated data of a quasispecies treatment (Figure 7), the changes experienced by the quasispecies with samples taken at given evolutionary steps are summarized in the QFF plot, Figure 8. The relationship between the quasispecies is shown in the form of a matrix of Y_C distances, Figure 9, from which we obtain a dendrogram by hierarchical clustering with the average method, Supplementary Figure S15, or a MDS map, Supplementary Figure S17. Using D_A distances we may obtain an alternative dendrogram, Supplementary Figure S16, or an alternative MDS plot, Supplementary Figure S18.

A key point with all these methods is the availability of quasispecies haplotypes with corresponding frequencies. The classical and more widespread NGS data analysis procedures for viruses, like Galaxy [8], i.e., limit sequencing errors by trimming the reads at their ends, where the quality is poorer, by a number of nucleotides, attending to instrument quality scores, using different algorithms. As a result of this trimming the coverages are uneven, even within the same amplicon, which prevents the direct obtention of amplicon haplotypes. In [5, 9], for instance, we describe the method used by our group to obtain high quality amplicon haplotypes in sequencing viral quasispecies samples. It is simply based on respecting the integrity of full reads, with no trimming, except for the primers. The quality filters are executed on full reads. This requires high sequencing quality and very high coverage to get a comprehensive pic-

ture of an infection that may involve viral loads higher than 10^6 copies/mL of blood. Currently we are only able to obtain high quality amplicon haplotypes of a size slightly over 500 bp, with coverages of the order of 10^5 reads per amplicon, sequencing with Illumina instruments. Despite this limitation, quasispecies genomes may be studied amplicon by amplicon. On the other hand, when the monitored treatment is by a direct acting agent that targets a specific region of the genome a single amplicon may suffice [9]. There are a number of inferential methods for reconstructing full viral haplotypes from short reads, but they have limitations, require of special computational resources for high coverages, and perform poorly with samples of high genetic diversity, according to a recent review evaluation of them [10].

The clinical case presented has given the opportunity to show a practical application of the proposed methods. This dataset with thousands of haplotypes in each sample, and coverages in the range of 5×10^4 to 5×10^5 reads, shows a correlation between the three indices consistent with what has been observed with the more modest simulated pairs of quasispecies entailing very few haplotypes; nevertheless, a critical aspect in the simulations was to ensure a close relationship between pairs of quasispecies, as it is the case in the follow-up of a patient, the main objective of this work.

The advantage of the described methods is that they provide rich summaries and visual tools to monitor the changes occurring in a viral quasispecies at the molecular level, with time. This facilitates the interpretation of the biological changes in the quasispecies, and also provides a means to diagnose possible outcomes of a treatment when monitoring a patient, as seen with the discussed HEV clinical case.

In the case of mutagenic treatments, we recommend this method, combined with the method of quasispecies fitness fractions (QFF), and the Hill numbers profile (HNP) [5]. When the quasispecies evolution rate is low compared to mutagenic scenarios, the QFF may result as insufficient to evidence changes in the quasispecies, and the proposed indices could be more sensitive to changes.

4. Materials and methods

4.1. Data

4.1.1. Simulation of paired quasispecies

To quantify the extent of changes (evolution) of a quasispecies, we compare the quasispecies composition at two time points. The paired quasispecies needed to illustrate the results and discussion are obtained by simulation as described in the following method:

1. **Distribution pattern:** 20,000 random occurrences of a geometric distribution, with parameter $p = 0.2$, are generated, simulating 20,000 reads of over 35 haplotypes. The frequencies of this distribution are used as pattern distribution on which to apply random selection criteria of frequencies.

2. **Select frequencies for quasispecies A:** from the above pattern distribution, 12 frequencies are randomly selected to represent the composition of quasispecies A.
3. **Select frequencies for quasispecies B:** from a new pattern distribution generated with the same parameters as above, randomly select 12 frequencies to represent the composition of quasispecies B.
4. **Confront both simulated quasispecies:** the two quasispecies are composed together of 20 haplotypes, some common to both quasispecies, some unique to either one. Assign randomly the 12 frequencies of quasispecies A among the 20, and do the same with the 12 frequencies of quasispecies B. Remove from the 20 any haplotype not populated (0 reads in both quasispecies).

A single cycle of this simulation results in the distributions of two paired quasispecies, which are given as shown in Table 1, and may be represented, confronting both distributions, as in Figure 1. The chosen numbers of reads and haplotypes in the simulation are arbitrary, a simplification of real life cases, but complex enough to compose a quasispecies.

The simulated pairs of quasispecies are related because of the result of a random selection of 12 haplotypes each from a common source of 20. On the other hand, the random selection of frequencies results in varying proportions for each haplotype and varying coverages (total number of reads) for each quasispecies. In this way, in each pair, we consider quasispecies B as the result of an evolution from quasispecies A.

The R code is provided in the Supplementary Materials.

4.1.2. Simulation of a viral treatment follow-up

The previous simulation aimed to generate pairs of quasispecies, more or less distant, as a result of certain evolution from the first to the second, and it was intended to help in the understanding and interpretation of the similarity indices and the correlations between them.

A second simulated dataset aims to generate a sequence of quasispecies that could be the result of an external treatment which generates resistant variants as a side effect. The quasispecies will consist of 40 haplotypes of three types:

1. The dominant haplotype, initially at a frequency of 99.9% evolving at a pace of a constant uniformly distributed between 0.85 and 1.05, at each evolution step.
2. A minoritary haplotype initially at (0.1/39)%, and evolving at a pace of a constant uniformly distributed between 0.95 and 1.25, at each evolution step.
3. The remaining 38 haplotypes, initially at (0.1/39)%, and evolving at a pace of a constant uniformly distribution between 0.8 and 2.5. Only a random number of these, between 2 and 10, are submitted to evolution at each step. The remaining are left as they were.

In this way, samples are sequentially generated at evolution steps 10, 20, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, and 80. The resulting haplotype distributions are plotted in Figure 7. The R code is provided in the Supplementary Materials.

4.1.3. A clinical HEV case

This dataset is taken from a recent publication [5], which shows the negative effects of early treatment discontinuation by a mutagenic agent of an HEV chronically infected patient. This dataset is used to show an example of application of the proposed method to a practical case. Briefly, this is the clinical follow-up case of a 27-year-old patient who acquired chronic HEV infection after undergoing two kidney transplantations. The patient received three different RBV regimens (600 mg/day, 800 mg/day, and 1000 mg/day) with discontinuations caused by adverse effects, followed by relapses.

A single amplicon covering genomic positions 6323 to 6734 on the HEV ORF2 region was sequenced, for each of 13 sequential samples taken from May 2018 to June 2021. The coverage range of the final dataset is 53, 307-503,770 reads, with a median of 328,271 reads per sample/amplicon, covering the full amplicon, and enabling the obtainment of amplicon haplotypes and corresponding frequencies. The number of haplotypes per sample are in the range 1688-7881, with a median number of 5602.

4.2. Methods

4.2.1. Similarity between distributions

The similarity between two distributions may be quantified by a rich set of different indices [4]. In this report, we use three of them:

1. Commons: as the fraction of reads belonging to haplotypes populated in both quasispecies.

$$C_m = \frac{1}{2}\sum_i \{(p_i + q_i)\, I(p_i > 0 \wedge q_i > 0)\} \tag{1}$$

2. Overlap: as the sum of the minimum proportion of common haplotypes.

$$O_v = \sum_i min(p_i, q_i) \tag{2}$$

3. Yue–Clayton: this index takes fuller account of all proportion information, considering the proportions of both common and unique haplotypes. [11]

$$Y_C = \frac{\sum_i p_i\, q_i}{\sum_i p_i^2 + \sum_i q_i^2 - \sum_i p_i\, q_i} \tag{3}$$

The three indices vary from 0 (no similitude) to 1 (equal quasispecies). The disimilarity, or distance, between two distributions may be computed as 1 minus the similarity index.

4.2.2. Genetic distance between quasispecies

The nucleotide distance between two quasispecies [2], *X* and *Y*, may be estimated by:

$$D_{XY} = \sum_{i \in X} \sum_{j \in Y} p_i \, d_{ij} \, q_j \tag{4}$$

where p_i and q_j are the proportion of the *i*-th haplotype in quasispecies *X*, and that of the *j*-th haplotype in quasispecies *Y*, and *dij* is the genetic distance between both haplotypes. The sum extends over all haplotypes in both quasispecies. This distance is interpreted as the average number of nucleotide substitutions between the reads from quasispecies *X* and quasispecies *Y*.

Taking into account the nucleotide diversity of each quasispecies [2], that is the average number of nucleotide substitutions for a random pair of reads in the quasispecies, D_X and D_Y, which may be estimated by:

$$D_X = \frac{N_X}{N_X - 1} \sum_{i \in X} \sum_{j \in X} p_i \, d_{ij} \, p_j \tag{5}$$

$$D_Y = \frac{N_Y}{N_Y - 1} \sum_{i \in Y} \sum_{j \in Y} q_i \, d_{ij} \, q_j \tag{6}$$

where N_X and N_Y are the number of reads in each quasispecies, then the net nucleotide substitutions between the two quasispecies [2] is estimated by:

$$D_A = D_{XY} - (D_X + D_Y)/2 \tag{7}$$

D_A will be taken as the net genetic distance between two quasispecies.

The quasispecies pairs are simulated in a way that all haplotypes are considered to have a single substitution with respect to the master haplotype in the first quasispecies. In this way, the matrix of distances between all pairs of haplotypes in both quasispecies has the form:

$$D : \left\{ \begin{array}{r} d_{ij} = 0, \forall \, i = j \\ d_{ij} = 1, \forall \, i = 1 \, and \, j > 1 \\ d_{ij} = 1, \forall \, j = 1 \, and \, i > 1 \\ d_{ij} = 2 \, otherwise \end{array} \right\} \tag{8}$$

4.2.3. Quasispecies Fitness Partition (QFP)

A quasispecies, at a given time, understood as a viral population, is usually comprised of a predominant haplotype, a few low- to medium-frequency genomes, various rare haplotypes with very low fitness but still able to replicate at some level, and some defective genomes unable to replicate. This composition can be modeled using the set of frequencies of all haplotypes in the quasispecies as parameters of a multinomial distribution, $\prod = \{p_i, p_2, ..., p_n\}$ with $\sum_{i=1}^{n} p_i = 1$. Where p_i is the frequency in the quasispecies of the i-th haplotype. The parameters, p_i, are sorted in decreasing order without a loss of generality.

In this way, the quasispecies can be partitioned into fractions limited by frequency thresholds of interest [5], as is in Equation (9), where a partition into four fractions (QFF) is illustrated, and where, p'_1, p'_2, p'_3 and p'_4 represent the four fractions.

$$\begin{aligned} \Pi_1 &= \{p_1, p_2, ..., p_k\}, \quad \forall p_i : p_i \geq p_k \\ \Pi_2 &= \{p_{k+1}, p_{k+2}, ..., p_l\}, \quad \forall p_i : p_l \leq p_i < p_k \\ \Pi_3 &= \{p_{l+1}, p_{l+2}, ..., p_m\}, \quad \forall p_i : p_m \leq p_i < p_l \\ \Pi_4 &= \{p_{m+1}, p_{m+2}, ..., p_n\}, \quad \forall p_i : p_n \leq p_i < p_m \end{aligned} \tag{9}$$

$$p'_1 = \sum_{i=1}^{k} p_i; \quad p'_2 = \sum_{i=k+1}^{l} p_i; \quad p'_3 = \sum_{i=l+1}^{m} p_i; \quad p'_4 = \sum_{i=m+1}^{n} p_i \tag{10}$$

$$\Pi' = \{p'_1, p'_2, p'_3, p'_4\} \tag{11}$$

In the typical quasispecies structure mentioned above, the four fractions can be defined as follows:

1. **Master**: the fraction of molecules belonging to the most frequent haplotype; that is, the one present at the highest percentage ($p'_1 = p_1$).
2. **Emerging**: the fraction of molecules presents at a frequency greater than 1% and smaller than the master percentage, belonging to haplotypes that are potentially able to compete with the predominant one and possibly replace it (p'_2).
3. **Low fitness**: the fraction of molecules presents at frequencies from 0.1% to 1%, belonging to haplotypes that have a low probability of progressing to higher frequencies (p'_3).
4. **Very low fitness**: the fraction of molecules presents at frequencies below 0.1%, belonging to haplotypes with very low fitness and to defective genomes. The likely fate of these molecules individually is degradation, but the fraction is continuously fed with new very low fitness genomes produced by replication errors or by host editing activities (p'_4).

This partition represents a summarization of the full haplotype distribution, where changes in each fraction have a straightforward biological meaning, and allow for the

interpretation of the effects caused by the current environment, or by the administration of an external agent.

4.3. Software and statistics

All computations were done in R (v4.0.3) [12], using packages ape [13], tidyverse [14], and ggplot2 [15]. The full code of the simulations and computations is provided in the Supplementary Materials. The session info follows:

```
sessionInfo()
R version 4.0.3 (2020-10-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19043)

Matrix products: default

Random number generation:
RNG:     Mersenne-Twister
Normal:  Inversion
Sample:  Rounding

locale:
[1] LC_COLLATE=Catalan_Spain.1252  LC_CTYPE=Catalan_Spain.1252
[3] LC_MONETARY=Catalan_Spain.1252 LC_NUMERIC=C
[5] LC_TIME=Catalan_Spain.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] forcats_0.5.1   stringr_1.4.0   dplyr_1.0.7     purrr_0.3.4
[5] readr_2.0.0     tidyr_1.1.3     tibble_3.1.3    ggplot2_3.3.5
[9] tidyverse_1.3.1

loaded via a namespace (and not attached):
[1] Rcpp_1.0.7       cellranger_1.1.0 pillar_1.6.2     compiler_4.0.2
[5] dbplyr_2.1.1     tools_4.0.2      digest_0.6.27    jsonlite_1.7.2
[9] lubridate_1.7.10 lifecycle_1.0.0  gtable_0.3.0     pkgconfig_2.0.3
[13] rlang_0.4.11     reprex_2.0.1     cli_3.0.1        rstudioapi_0.13
[17] DBI_1.1.1        haven_2.4.3      xml2_1.3.2       withr_2.4.2
[21] httr_1.4.2       fs_1.5.0         generics_0.1.0   vctrs_0.3.8
[25] hms_1.1.0        grid_4.0.2       tidyselect_1.1.1 glue_1.4.2
[29] R6_2.5.0         fansi_0.5.0      readxl_1.3.1     farver_2.1.0
[33] tzdb_0.1.2       modelr_0.1.8     magrittr_2.0.1   backports_1.2.1
[37] scales_1.1.1     ellipsis_0.3.2   rvest_1.0.1      assertthat_0.2.1
[41] colorspace_2.0-2 labeling_0.4.2   utf8_1.2.2       stringi_1.7.3
[45] munsell_0.5.0    broom_0.7.9      crayon_1.4.1
```

Supplementary Materials

The following supporting information can be downloaded from https://www.mdpi.com/article/10.3390/ijms24021301/s1

Author contributions

Conceptualization: J.G. and J.Q.; Methodology: J.G. and M.I.-L.; Software: J.G. and M.I.-L.; Formal analysis: J.G.; Investigation: M.I.-L.; Resources: J.Q.; Writing, original draft preparation: J.G.; Writing, review and editing: J.G. and J.Q.; Visualization: J.G. and M.I.-L.; Validation: J.G.; Supervision: J.G. and J.Q.; Funding acquisition: J.Q.

All authors have read and agreed to the published version of the manuscript.

Funding

This study was partially supported by Pla Estratègic de Recerca i Innovació en Salut (PERIS) – Direcció General de Recerca i Innovació en Salut (DGRIS), Catalan Health Ministry, Generalitat de Catalunya; the Spanish Network for the Research in Infectious Diseases (REIPI RD16/0016/0003) from the European Regional Development Fund (ERDF); Centro para el Desarrollo Tecnológico Industrial (CDTI) from the Spanish Ministry of Economy and Business, grant number IDI-20200297; grant PI19/00301 and PI22/00258 from Instituto de Salud Carlos III cofinanced by the European Regional Development Fund (ERDF), and Gilead's biomedical research project GLD21/00006.

Data availability statement

The R code used to generate the simulated data, and used in the computations is provided in the Supplementary Materials.

Conflicts of interest

The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

HNP: Hill Numbers Profile
MDS: Multidimensional Scaling
NGS: Next Generation Sequencing
QFF: Quasispecies Fitness Fractions
QFP: Quasispecies Fitness Partition
RHL: Rare Haplotype Load

REFERENCES

1. Domingo E., Parrish C.R., Holland J.J. (Eds.). Origin and evolution of viruses, 2nd ed. London, UK: Academic Press, Elsevier, 2008.
2. Nei M. Molecular evolutionary genetics, 1st ed. New York, NY: Columbia University Press, 1987; 276-279.
3. Gregori J., Perales C., Rodriguez-Frias F., Esteban J.I., Quer J., Domingo E. Viral quasispecies complexity measures. Virology 2016; 493: 227-237.
4. Cha S.-H. Comprehensive survey on distance/similarity measures between probability density functions. Int J Math Model Methods Appl Sci 2007; 1: 300-307.
5. Gregori J., Colomer-Castell S., Campos C. et al. Quasispecies fitness partition to characterize the molecular status of a viral population. Negative effect of early ribavirin discontinuation in a chronically infected HEV patient. Int J Mol Sci 2022; 23: 14654.
6. Gallego I., Gregori J., Soria M.E. et al. Resistance of high fitness hepatitis C virus to lethal mutagenesis. Virology 2018; 523: 100-109.
7. Domingo E. Virus as populations: composition, complexity, dynamics, and biological. In: Virus as populations, 1st ed. London, UK: Academic Press, Elsevier, 2016; 1-412.
8. Galaxy Community Hub. Available online: https://galaxyproject.org/ (accessed on 7 January 2023).
9. Soria M.E., Gregori J., Chen Q. et al. Pipeline for specific subtype amplification and drug resistance detection in hepatitis C virus. BMC Infect Dis 2018; 18: 446.
10. Eliseev A., Gibson K.M., Avdeyev P. et al. Evaluation of haplotype callers for next-generation sequencing of viruses. Infect Genet Evol 2020; 82: 104277.
11. Yue J.C., Clayton M.K. A similarity measure based on species proportions. Commun Stat Theory Methods 2005; 34: 2123-2131.
12. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2020. Available online: https://www.R-project.org/ (accessed on 7 January 2023).
13. Paradis E., Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 2019; 854: 526-528.
14. Wickham H. Welcome to Master the Tidyverse. J Open Source Softw 2019; 4: 1686.
15. Wickham H. ggplot2: elegant graphics for data analysis. New York, NY: Springer, 2016. Available online: https://ggplot2.tidyverse.org (accessed on 7 January 2023).

Supplementary Materials

1. Distribution overlap

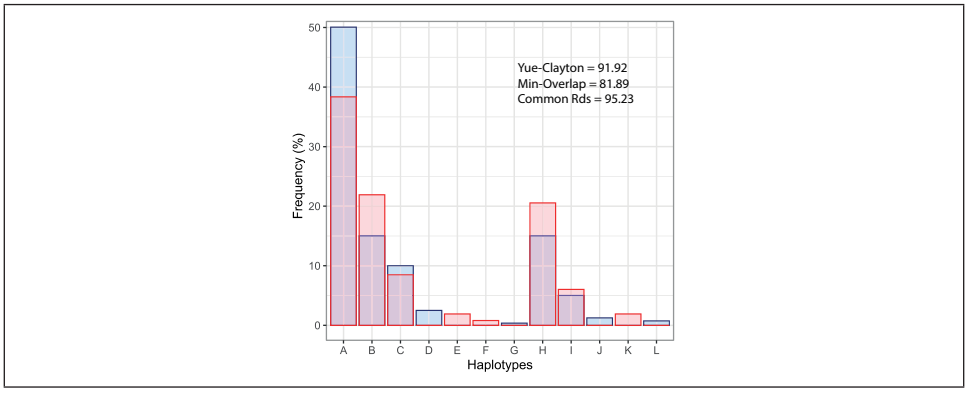


Figure S1. Montserrat plot illustrating distribution overlap.

2. Histograms of simulated values

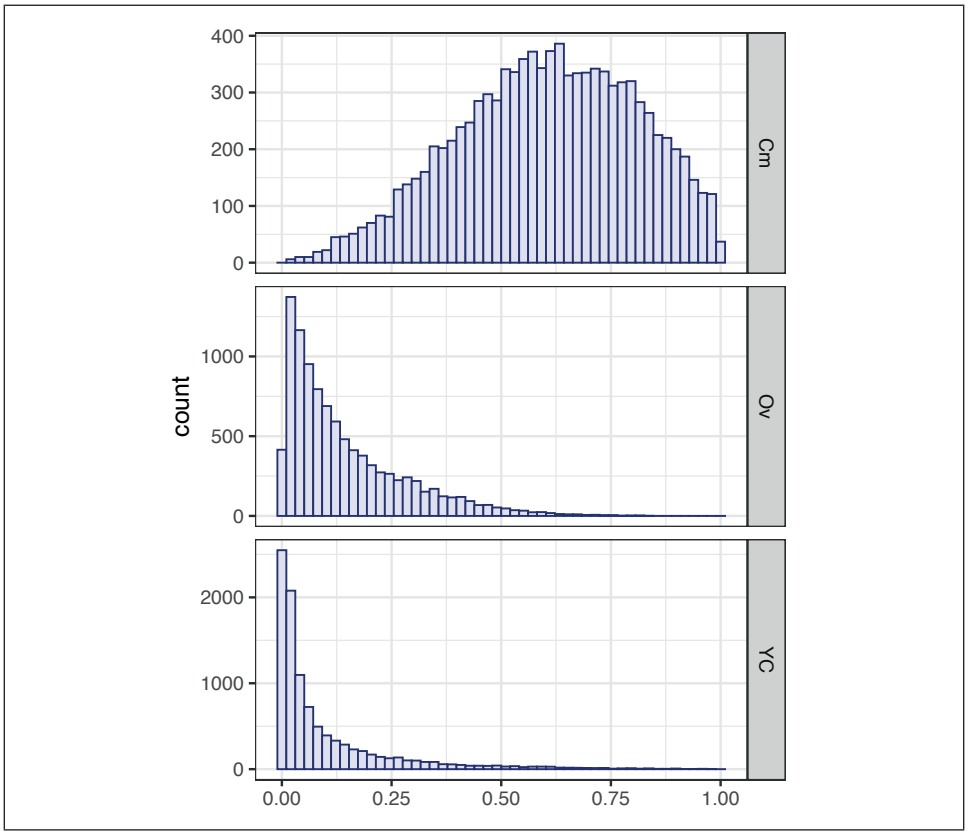


Figure S2. Histograms of values of similarity for the simulated pairs of quasispecies.

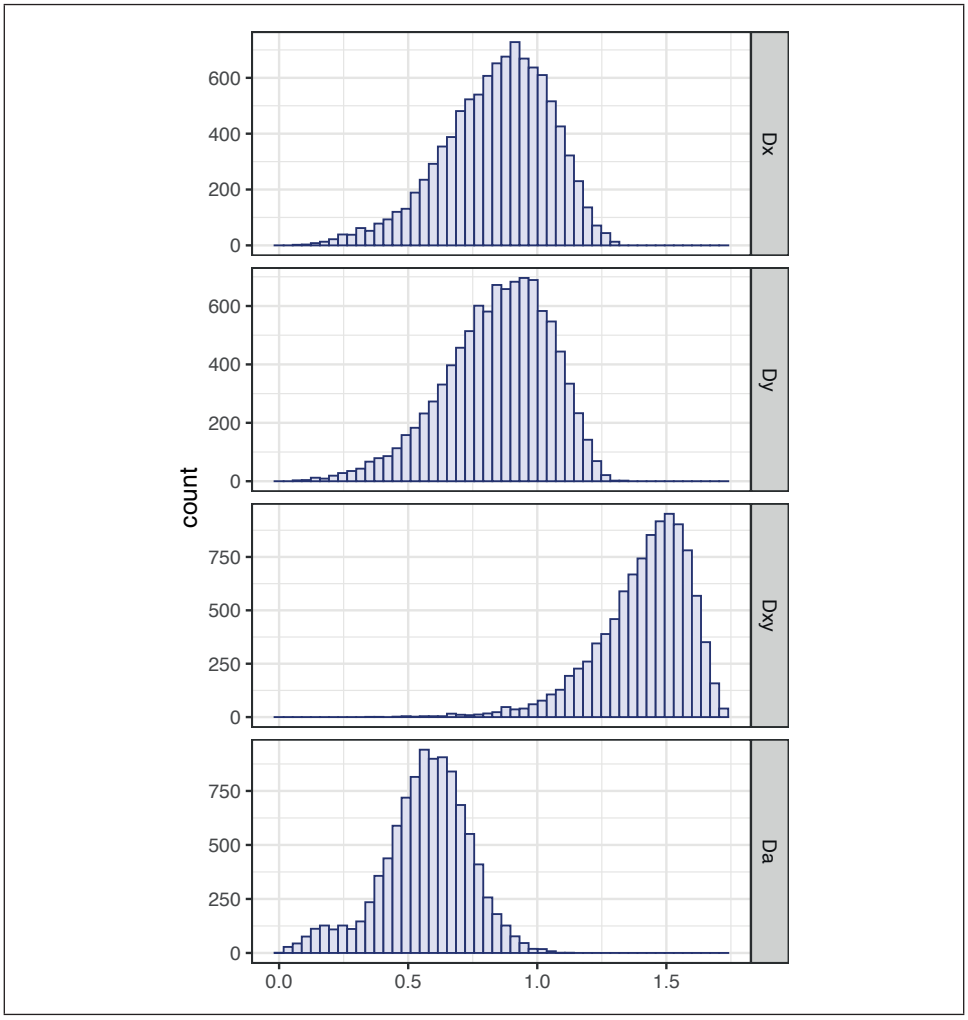


Figure S3. Histograms of values of nucleotide diversity and genetic distance for the simulated pairs of quasispecies.

3. Tables and figures of selected quasispecies pairs

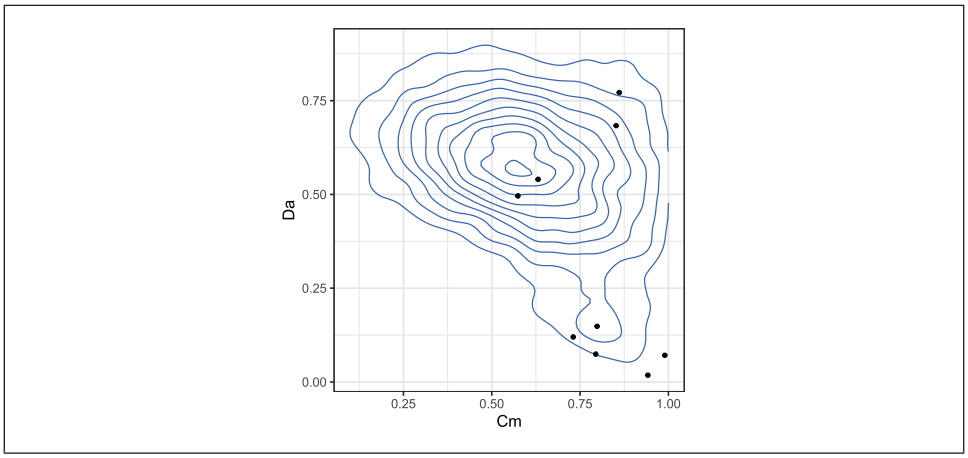


Figure S4. Selected pairs plotted on the C_m and D_A density plot.

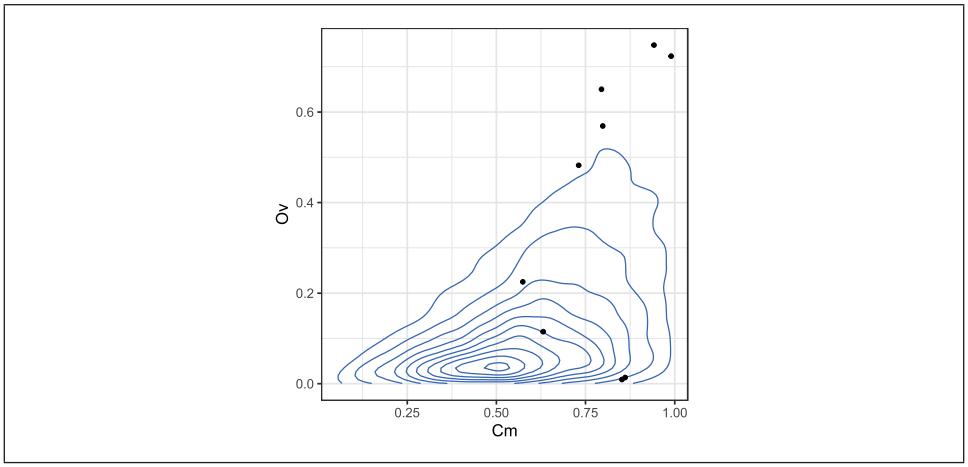


Figure S5. Selected pairs plotted on the C_m and O_v density plot.

Hpl	*nA*	*nB*	*pA*	*pB*	*Commons*	*Overlap*
A	5230	3808	52.30	38.08	45.19	38.08
B	3453	3059	34.53	30.59	32.56	30.59
C	528	981	5.28	9.81	7.55	5.28
D	355	15	3.55	0.15	1.85	0.15
E	233	0	2.33	0.00	0.00	0.00
F	85	0	0.85	0.00	0.00	0.00
G	56	90	0.56	0.90	0.73	0.56
H	39	0	0.39	0.00	0.00	0.00
I	8	4	0.08	0.04	0.06	0.04
J	7	402	0.07	4.02	2.05	0.07
K	4	0	0.04	0.00	0.00	0.00
L	1	842	0.01	8.42	4.22	0.01
M	0	614	0.00	6.14	0.00	0.00
N	0	21	0.00	0.21	0.00	0.00
O	0	163	0	1.63	0	0
P	0	2	0	0.02	0	0

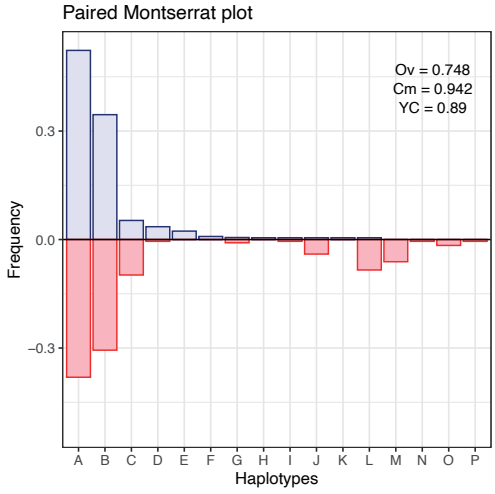


Figure S6. Simulated pair of index 4213.

Hpl	*nA*	*nB*	*pA*	*pB*	*Commons*	*Overlap*
A	3666	3142	36.66	31.42	34.04	31.42
B	2794	3899	27.94	38.99	33.47	27.94
C	1081	2541	10.81	25.41	18.11	10.81
D	909	2	9.09	0.02	4.56	0.02
E	766	79	7.66	0.79	4.23	0.79
F	474	115	4.74	1.15	2.95	1.15
G	265	2	2.65	0.02	1.34	0.02
H	22	5	0.22	0.05	0.14	0.05
I	13	25	0.13	0.25	0.19	0.13
J	6	0	0.06	0.00	0.00	0.00
K	3	0	0.03	0.00	0.00	0.00
L	1	0	0.01	0.00	0.00	0.00
M	0	145	0.00	1.45	0.00	0.00
N	0	14	0.00	0.14	0.00	0.00
O	0	33	0	0.33	0	0

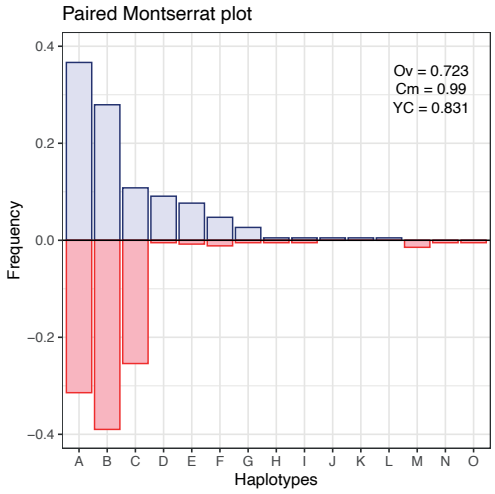


Figure S7. Simulated pair of index 7426.

Hpl	*nA*	*nB*	*pA*	*pB*	*Commons*	*Overlap*
A	4453	15	44.53	0.15	22.34	0.15
B	2834	4	28.34	0.04	14.19	0.04
C	1149	4	11.49	0.04	5.77	0.04
D	849	0	8.49	0.00	0.00	0.00
E	371	0	3.71	0.00	0.00	0.00
F	157	0	1.57	0.00	0.00	0.00
G	65	2377	0.65	23.77	12.21	0.65
H	47	6144	0.47	61.44	30.96	0.47
I	38	0	0.34	0.00	0.00	0.00
J	34	0	0.34	0.00	0.00	0.00
K	1	121	0.01	0.07	0.04	0.01
L	1	7	0.01	0.07	0.04	0.01
M	0	92	0.00	0.92	0.00	0.00
N	0	309	0.00	3.09	0.00	0.00
O	0	18	0	0.18	0	0
P	0	898	0	8.98	0	0
Q	0	11	0	0.11	0	0

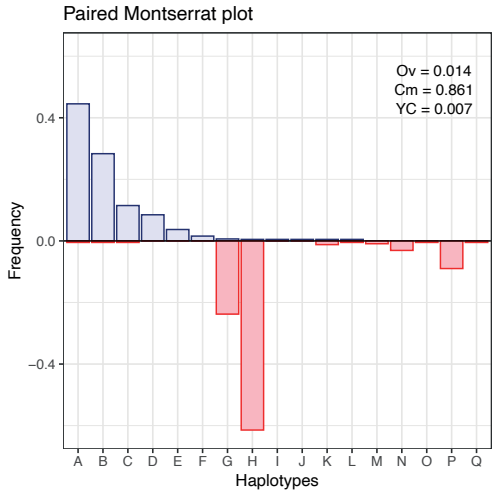


Figure S8. Simulated pair of index 774.

Hpl	*nA*	*nB*	*pA*	*pB*	*Commons*	*Overlap*
A	4360	1	43.60	0.01	21.81	0.01
B	3299	4	32.99	0.04	16.52	0.04
C	1473	0	14.73	0.00	0.00	0.00
D	604	0	6.04	0.00	0.00	0.00
E	89	0	0.89	0.00	0.00	0.00
F	72	0	0.72	0.00	0.00	0.00
G	48	2070	0.48	20.70	10.59	0.48
H	30	3076	0.30	30.76	15.53	0.30
I	13	0	0.13	0.00	0.00	0.00
J	9	4142	0.09	41.42	20.76	0.09
K	2	0	0.02	0.00	0.00	0.00
L	1	1	0.01	0.01	0.01	0.01
M	0	411	0.00	4.11	0.00	0.00
N	0	1	0.00	0.01	0.00	0.00
O	0	6	0	0.06	0	0
P	0	15	0	0.15	0	0
Q	0	264	0	2.64	0	0
R	0	9	0	0.09	0	0

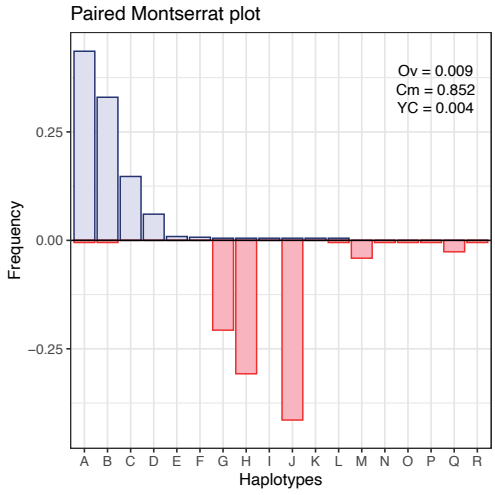


Figure S9. Simulated pair of index 5463.

Hpl	*nA*	*nB*	*pA*	*pB*	*Commons*	*Overlap*
A	1140	0	34.19	0.00	0.00	0.00
B	701	132	21.02	3.95	12.49	3.95
C	548	1424	16.43	42.73	29.58	16.43
D	361	50	10.83	1.51	6.17	1.51
E	291	0	8.73	0.00	0.00	0.00
F	145	0	4.36	0.00	0.00	0.00
G	58	2	1.75	0.05	0.90	0.05
H	30	12	0.90	0.35	0.63	0.35
I	29	6	0.87	0.18	0.53	0.18
J	22	0	0.67	0.00	0.00	0.00
K	8	0	0.23	0.00	0.00	0.00
L	1	474	0.03	14.21	7.12	0.03
M	0	77	0.00	2.31	0.00	0.00
N	0	1124	0.00	33.72	0.00	0.00
O	0	2	0	0.05	0	0
P	0	30	0	0.89	0	0
Q	0	1	0	0.03	0	0

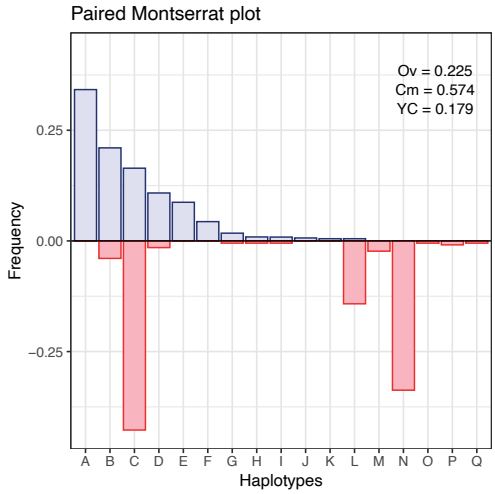


Figure S10. Simulated pair of index 3053.

Hpl	*nA*	*nB*	*pA*	*pB*	*Commons*	*Overlap*
A	1167	1	35.02	0.04	17.53	0.04
B	926	0	27.79	0.00	0.00	0.00
C	573	338	17.20	10.13	13.67	10.13
D	528	0	15.85	0.00	0.00	0.0
E	88	0	2.63	0.00	0.00	0.00
F	29	243	0.87	7.30	4.09	0.87
G	9	185	0.28	5.54	2.91	0.28
H	6	0	0.17	0.00	0.00	0.00
I	2	1239	0.07	37.16	18.62	0.07
J	2	414	0.07	12.41	6.24	0.07
K	1	0	0.03	0.00	0.00	0.00
L	1	4	0.03	0.12	0.08	0.03
M	0	111	0.00	3.34	0.00	0.00
N	0	9	0.00	0.28	0.00	0.00
O	0	637	0	19.1	0	0
P	0	141	0	4.24	0	0
Q	0	11	0	0.34	0	0

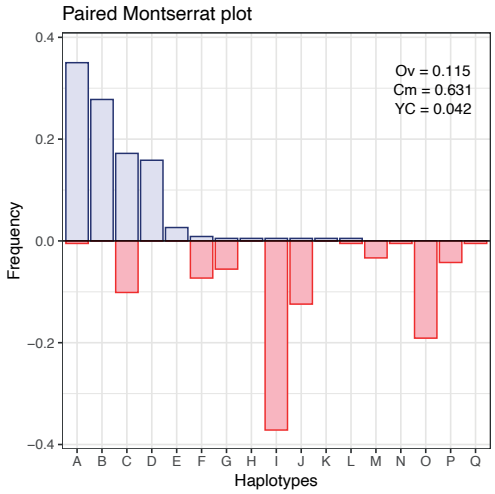


Figure S11. Simulated pair of index 5955.

Hpl	*nA*	*nB*	*pA*	*pB*	*Commons*	*Overlap*
A	2376	3108	47.53	62.16	54.85	47.53
B	1568	0	31.37	0.00	0.00	0.00
C	478	2	9.56	0.04	4.80	0.04
D	250	354	5.01	7.07	6.04	5.01
E	208	1184	4.16	23.69	13.93	4.16
F	43	0	0.86	0.00	0.00	0.00
G	34	0	0.68	0.00	0.00	0.00
H	26	0	0.51	0.00	0.00	0.00
I	8	5	0.15	0.10	0.13	0.10
J	4	6	0.07	0.12	0.10	0.07
K	4	0	0.07	0.00	0.00	0.00
L	1	0	0.03	0.00	0.00	0.00
M	0	2	0.00	0.04	0.00	0.00
N	0	28	0.00	0.55	0.00	0.00
O	0	216	0.00	4.33	0.00	0.00
P	0	1	0.00	0.02	0.00	0.00
Q	0	68	0.00	1.36	0.00	0.00
R	0	26	0.00	0.53	0.00	0.00

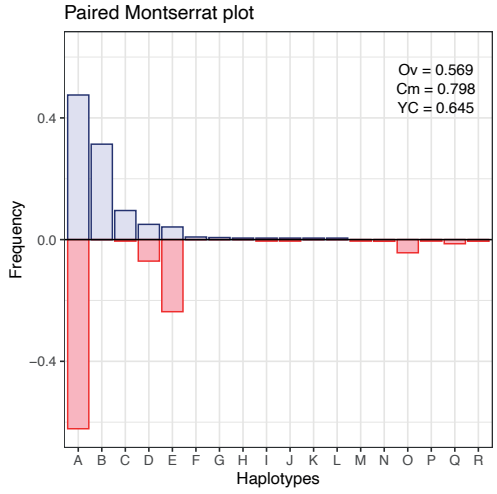


Figure S12. Simulated pair of index 1159.

Hpl	*nA*	*nB*	*pA*	*pB*	*Commons*	*Overlap*
A	5732	4936	57.32	49.36	53.34	49.36
B	1559	3222	15.59	32.22	23.91	15.59
C	1198	0	11.98	0.00	0.00	0.00
D	487	0	4.87	0.00	0.00	0.00
E	416	4	4.16	0.04	2.10	0.04
F	320	0	3.20	0.00	0.00	0.00
G	204	0	2.04	0.00	0.00	0.00
H	66	0	0.66	0.00	0.00	0.00
I	7	0	0.07	0.00	0.00	0.00
J	7	1	0.07	0.01	0.04	0.01
K	1	9	0.01	0.09	0.05	0.01
L	1	2	0.01	0.02	0.02	0.01
M	0	21	0.00	0.21	0.00	0.00
N	0	2	0.00	0.02	0.00	0.00
O	0	134	0.00	1.34	0.00	0.00
P	0	10	0.00	0.10	0.00	0.00
Q	0	1628	0.00	16.28	0.00	0.00
R	0	30	0.00	0.30	0.00	0.00

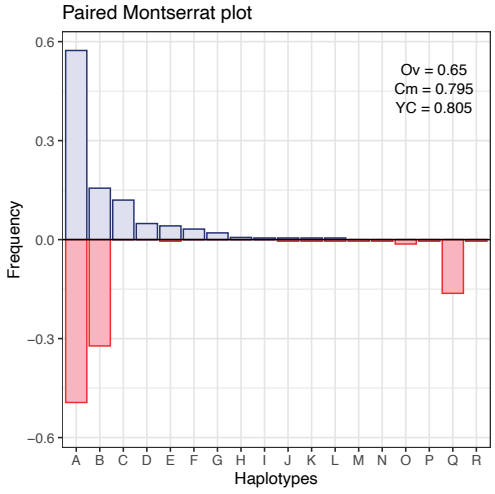


Figure S13. Simulated pair of index 2528.

Hpl	*nA*	*nB*	*pA*	*pB*	*Commons*	*Overlap*
A	3290	2354	65.79	47.07	56.43	47.07
B	662	0	13.25	0.00	0.00	0.00
C	558	1	11.17	0.03	5.60	0.03
D	283	20	5.66	0.41	3.04	0.41
E	72	0	1.43	0.00	0.00	0.00
F	64	17	1.28	0.34	0.81	0.34
G	40	1	0.80	0.03	0.42	0.03
H	12	0	0.24	0.00	0.00	0.00
I	10	114	0.20	2.29	1.25	0.20
J	6	59	0.13	1.18	0.66	0.13
K	1	0	0.03	0.00	0.00	0.00
L	1	488	0.02	9.75	4.89	0.02
M	0	734	0.00	14.69	0.00	0.00
N	0	33	0.00	0.66	0.00	0.00
O	0	3	0.00	0.06	0.00	0.00
P	0	1174	0.00	23.49	0.00	0.00

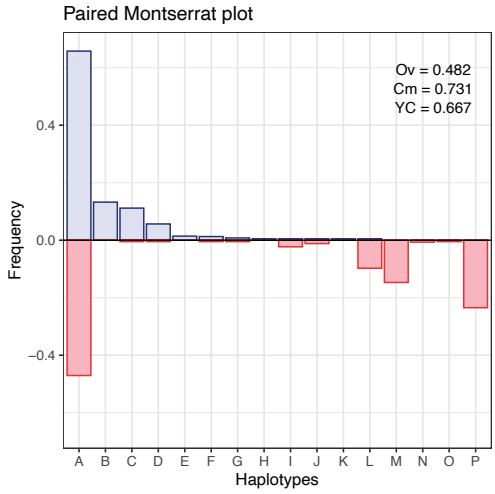


Figure S14. Simulated pair of index 345.

4. Simulated evolution

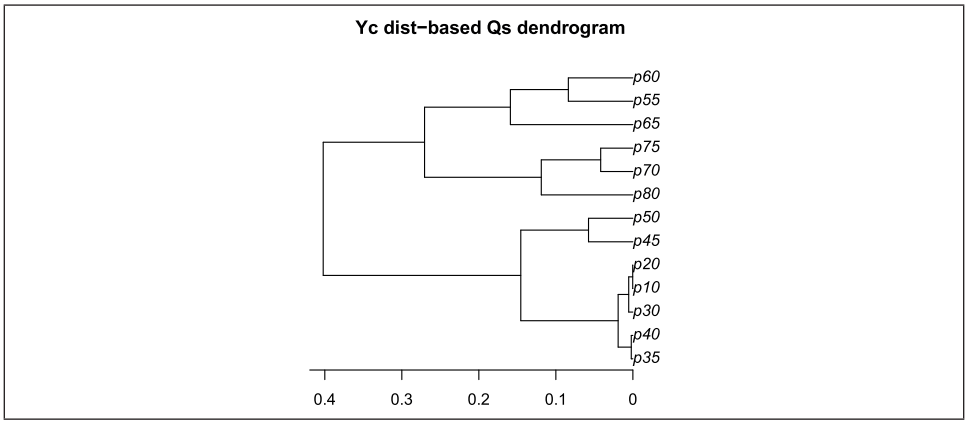


Figure S15. Quasispecies dendrogram based on Yue-Clayton distribution distances.

Figure S16. Quasispecies dendrogram based on D_A genetic distances.

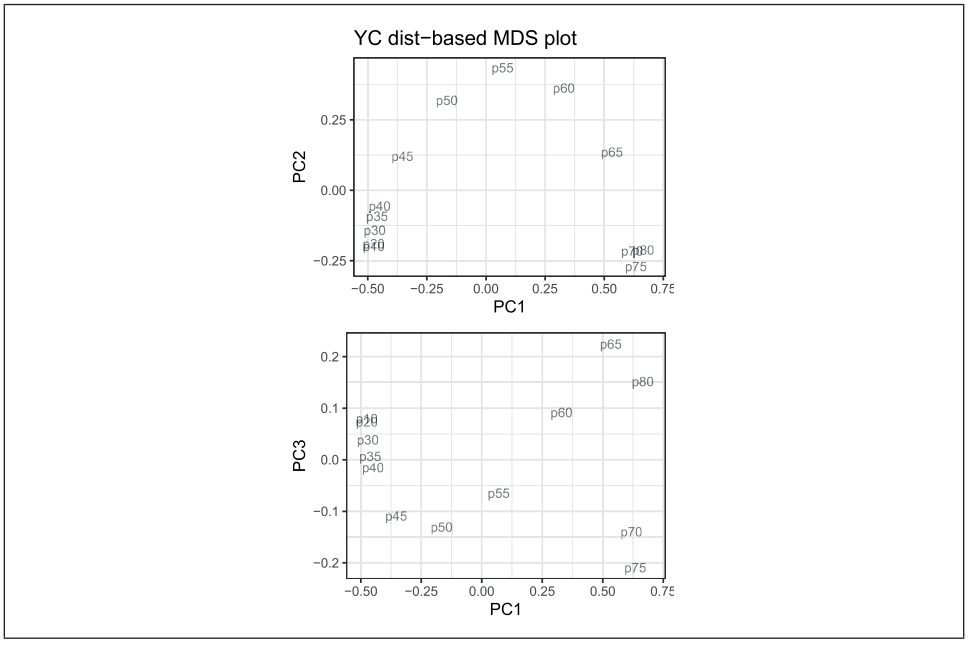


Figure S17. MDS plot based on Yue-Clayton distribution distances.

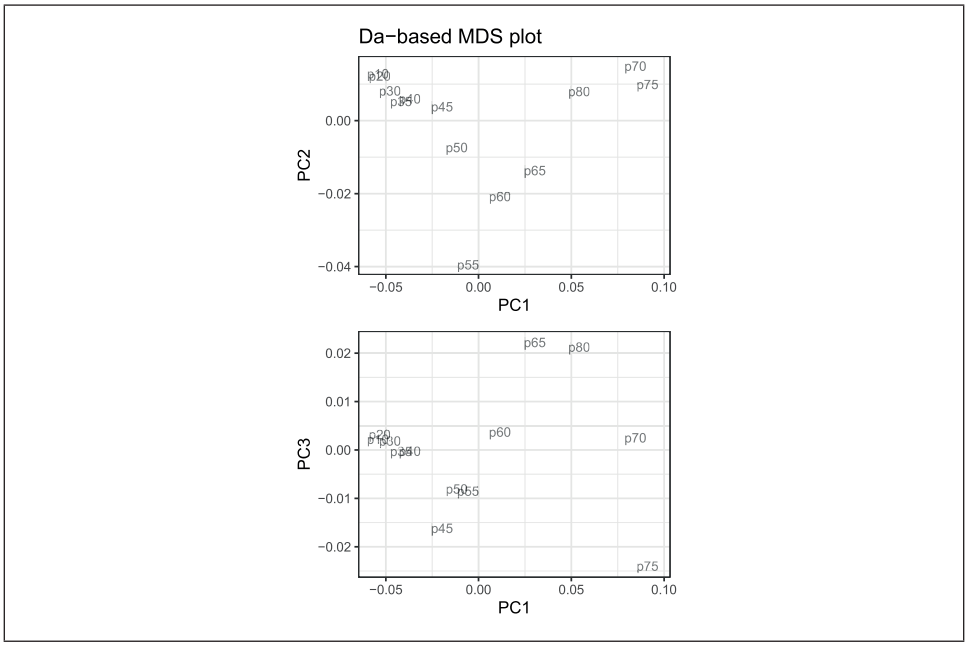


Figure S18. MDS plot based on D_A genetic distances.

Concluding remarks

The journey initiated by exploring Shannon entropy and mutation frequency [1], as basic indices of viral quasispecies diversity with NGS data, led us to explore and propose other indices classically used in the field of biodiversity. Hierarchical classification of these indices contributed to a better understanding of the type of information each one provides [2]. Next, study of mutagenesis cases, in a controlled laboratory setting in liver cell lines and in clinical samples, led to formulation of new indices based on haplotype fitness fractions: first, the rare haplotype load (RHL) [3] at various levels, and then haplotype partition into four fractions, the quasispecies fitness fractions (QFF) [4]. Finally, as measures complementary to the previous methods, indices of similarity between haplotype distributions were explored to follow in-host quasispecies changes [5]. Along the journey, a number of graphical tools were proposed to visualize the composition and evolution of viral quasispecies.

Take-home messages

Diversity as a dimensional reduction

- The information obtained by NGS of a quasispecies sample consists in a multiple alignment of accepted haplotypes and the corresponding vector of frequencies (read counts).
- Computation of a diversity index corresponds to a projection of part of this multidimensional information into a single value.
- In quasispecies terms, this is a tremendous reduction of information.
- Because of this extreme reduction, several viral populations could yield the same diversity index value even though they differ.
- All diversity indices that result from a sum of terms can give the same results for rather different situations. The contribution to mutation frequency of a single haplotype with an abundance of 10% and a single difference with respect to the master,

is the same as the contribution of 100 haplotypes at the same distance from the master at an abundance of 0.1%.

- The same applies to nucleotide diversity. Although mutation frequency and nucleotide diversity provide functional diversity information, these indices are unable to distinguish between the two quite different genetic situations mentioned above.
- To a different extent, the same applies to abundance-based diversity indices, such as Shannon entropy or the Simpson index.
- These considerations justify the need for a multidimensional approach to characterize quasispecies diversity; that is, the use of several diversity indices contributing complementary information.

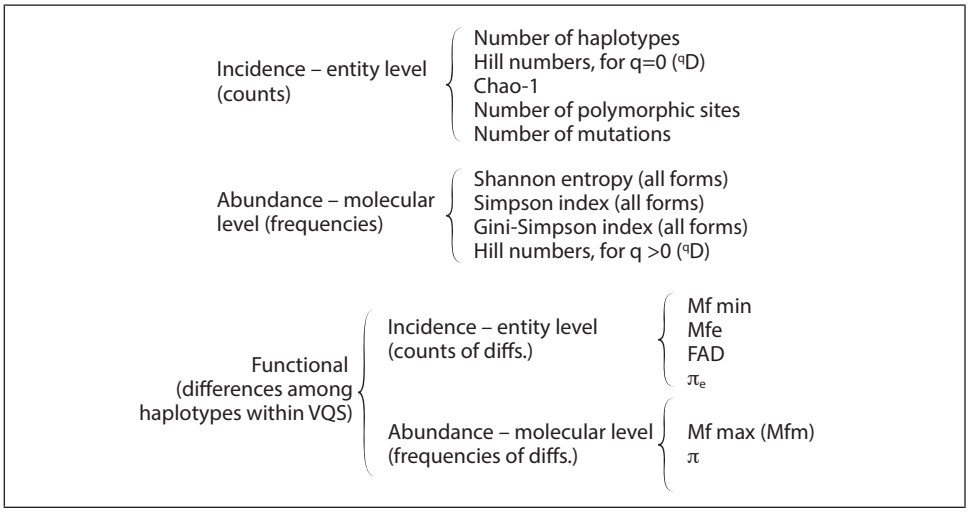


Figure 1. Reproduction of a figure in [2] with the classification of a few diversity indices. See Section 2. Mf: mutation frequency. π: nucleotide diversity. FAD: functional attribute diversity.

Type of information

- Classification of diversity indices into three groups – incidence, abundance and function-related – provides insights into the type of information each delivers and helps in the interpretation of the results.
- Each group in the classification answers a different question in population terms:
  – **Incidence**. How many (different) are there?
  – **Abundance**. How many are there of each of them?
  – **Function-related**. How different from each other are they?
- Nevertheless, for most indices, there is a second level of information beyond this distinction. For example, number of haplotypes, number of polymorphic sites, and

number of mutations are incidence indices that also provide implicit functional information.
- All diversity indices of incidence are highly informative about genetic diversity, even though they ignore the extent of the genetic difference between haplotypes. They are, however, very sensitive to sample size.

Diversity profiles

- Computation of any diversity profile corresponds to a projection of part of the information gathered onto a curve over a two dimensional space.
- In this respect, diversity profiles are richer than any single diversity index.
- Diversity profiles are visual tools.
- The QFF and the Hill numbers profile (HNP) are both recommended profiles.

Quasispecies distances

- Two quasispecies are distant to the extent that they are different.
- Several types of distances are useful for computing quasispecies distances. The differences can be genetic, phenotypic, or distributional.
- As occurs with diversity indices, computation of a distance corresponds to a significant reduction in information, with respect to two compositional data sets.
- The net nucleotide distance between quasispecies (D_A) is closely related to the computation of nucleotide diversity, and may experience the same limitations as those seen for mutation frequency and nucleotide diversity.
- The distance between haplotype distributions in two quasispecies provides a complementary view.
- When haplotypes are translated to phenotypes, computation of distances between phenotypes provides another rich complementary view.
- Various visual tools can be applied to any matrix of distances between a set of related quasispecies: dendrograms obtained from hierarchical clustering, and two- or three dimensional maps by multidimensional scaling.

Sample size issues

- Because of the difficulty of obtaining identical sample sizes for all samples in an NGS study, methods are needed to correct the diversity values analyzed to an equivalent sample size.
- Resampling to the minimum acceptable size is a recommended rarefaction method. Despite that rarefaction is defined as repeated resampling without replacement, the big numbers implied with the recommended coverages make practically no difference between the two methods [6], being faster with replacement.
- Although sample size corrections are possible, the amount of information con-

veyed by a study will be limited by the size of the smallest sample. The minimum acceptable sample size should be established as a part of the study's experimental design.

- Some indices are more sensitive than others to sample size; this sensitivity is linked to the order of the index. The order of an index refers to type of terms implied. Those taking squares of proportions or products of two proportions are of order 2. At increasing powers, the influence of terms with low frequencies quickly fades, making the corresponding indices less sensitive to sample size. With Hill numbers, the order ranges from 0 to infinity, number of haplotypes is of order 0, exponential of Shannon entropy of order 1, inverse of Simpson index of order 2, and inverse of the master frequency of order infinity. The following list provides a hierarchy of sample size sensitivity from highest to lowest.
  - Number of haplotypes, number of polymorphic sites, and incidence-based indices in general.
  - Abundance based indices of order <2, such as Shannon entropy.
  - Abundance based indices of order 2, such as the Simpson index, or the nucleotide diversity.
  - Fitness fractions resulting in significant amounts, such as the RHL, or the master frequency.
  - Abundance-based indices of higher order, such as the inverse of the master frequency.

Technical and experimental noise

- Reverse transcription (RT), polymerase chain reaction (PCR), and sequencing (NGS) errors are unavoidable, but observation of a quasispecies structure requires a deep view.
- Despite the use of various error filters in NGS data treatment, both true and artefactual haplotypes coexist at low abundance levels.
- The RHL ([3], Section 4), and QFF ([4], Section 5) studies show the amount of information loss implied when filtering at a minimum abundance level to avoid errors.
- The best strategy to avoid biased results is a perfectly balanced experimental design, with high and even coverages.
- Hence, it is important to use a good experimental design and avoid unnecessary abundance filters to obtain a comprehensive picture of quasispecies composition.

Quasispecies as dynamic systems

A single quasispecies sample provides no information about what the quasispecies was or what it will be.

REFERENCES

1. Gregori J, Salicrú M, Domingo E, et al. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics 2014; 30(8):1104-11. https://doi.org/10.1093/bioinformatics/btt768
2. Gregori J, Perales C, Rodríguez-Frías F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology 2016; 493:227-37. https://doi.org/10.1016/j.virol.2016.03.017
3. Gregori J, Soria ME, Gallego I, et al. Rare haplotype load as marker for lethal mutagenesis. PLoS One 2018; 13(10):e0204877. https://doi.org/10.1371/journal.pone.0204877
4. Gregori J, Colomer-Castell S, Campos C, et al. Quasispecies fitness partition to characterize the molecular status of a viral population. Negative effect of early ribavirin discontinuation in a chronically infected HEV patient. Int J Mol Sci 2022; 23:14654. https://doi.org/10.3390/ijms232314654
5. Gregori J, Ibañez-Lligoña M, Quer J. Quantifying in-host quasispecies evolution. Int J Mol Sci 2023; 24(2):1301. https://doi.org/10.3390/ijms24021301
6. Cameron ES, Schmidt PJ, Tremblay BJ, Emelko MB, Müller KM. Enhancing diversity analysis by repeatedly rarefying next generation sequencing data describing microbial communities. Sci Rep 2021; 11(1):22302. https://doi.org/10.1038/s41598-021-01636-1

Applicability constraints

Through the articles discussed here, we provide the means to perform statistically sound tests and comparisons to determine quasispecies diversity based on NGS data. Nevertheless, beyond any p-value or effect size, researchers should be conscious of the constraints imposed by sampling from a dynamic system. For example, in case-control studies, a significant number of patients should be included to take into account the added source of variability.

Despite the advances in characterizing quasispecies, questions as simple as classifying an infection as acute or chronic by analysis of a single sample remain elusive. Because of the dynamic nature of quasispecies, making predictions based on a single sample taken at a given time could be compared to deciding who is guilty of a crime after seeing a single random frame of a film.

The evolution rate of a quasispecies is a priori unknown, and depends on several factors; among them, the viral load, but also the evolutionary pressure imposed by external factors such as treatments and the patient's immune system. Any clinical prediction about quasispecies evolution should be based on data provided by a set of samples encompassing a significant time period, as with the HEV patient reported in reference [1]. By the same rule, comparing single samples from two unrelated quasispecies can also be misleading: one could be passing through a phase of diversity contraction or expansion, whereas the other could be in the opposite phase. These considerations advise caution and much care before reaching any conclusions based on quasispecies composition data.

A final note of warning concerns the risk of amplification bias. As has been discussed throughout the book, our developments are based on quasispecies haplotypes and frequencies, and because of current NGS technical limitations, are focused on data based on amplicons. This requires the use of specific amplicon primer pairs for PCR amplification, which implies a risk of amplification bias due to dissimilar efficiencies of the primer pair on the various haplotypes comprising the quasispecies. Putative amplification bias can be experimentally evaluated by determining the mean

efficiency of the primer pair with respect to a quasispecies by computation of the slope of the quasispecies standard dilution curve ([2], [3]).

REFERENCES

1. Gregori J, Colomer-Castell S, Campos C, et al. Quasispecies fitness partition to characterize the molecular status of a viral population. Negative effect of early ribavirin discontinuation in a chronically infected HEV patient. Int J Mol Sci 2022; 23:14654. https://doi.org/10.3390/ijms232314654
2. Booth CS, Pienaar E, Termaat JR, Whitney SE, Louw TM, Viljoen HJ. Efficiency of the polymerase chain reaction. Chem Eng Sci 2010; 65(17):4996-5006. https://doi.org/10.1016/j.ces.2010.05.046
3. Svec D, Tichopad A, Novosadova V, Pfaffl MW, Kubista M. How good is a PCR efficiency estimate: recommendations for precise and robust qPCR efficiency assessments. Biomol Detect Quantif 2015; 3:9-16. https://doi.org/10.1016/j.bdq.2015.01.005

A collection of selected publications by the research group at VHIR in Vall d'Hebron Barcelona Hospital Campus, this book deals with the diversity, complexity, and evolution of viral quasispecies. It reports the tools devised to monitor and quantify the changes in quasispecies composition and describes the developments attained in laboratory to distill complexity into something simple but still informative. The articles here contained correspond to research initiated in 2011 and cover the progress in quasispecies characterization up to 2023. The challenge faced by the authors was to characterize viral quasispecies in terms of their diversity. The progress in their work includes the use of visual tools to represent viral diversity in simple terms, while retaining, whenever possible, high biological meaning.
The book starts with an introduction and a historical note that narrates the transition from molecular cloning to NGS in viral quasispecies studies, including development of the software used by the research group to obtain amplicon haplotypes with their frequencies from NGS data. The related articles are listed and briefly described, and a section of the book is devoted to each of them. The volume ends with general closing remarks and a note about the meaning and implications of acquiring samples from a dynamic system.

"… based on the pioneering contributions of the authors, using deep sequencing to unveil the composition of pathogenic RNA viruses (…) and to interpret treatment responses and failures in terms of quasispecies dynamics (…) the book is both informative and tutorial. The authors take advantage of having shared expertise in bioinformatics and clinical medicine for many years. The collection of articles (...) will guide the reader into understanding the mathematic formulations conductive to diversity index calculation, and how the results find an application to the clinical setting."

Esteban Domingo

€ 45,00 (i.i.)