Virus Research 239 (2017) 17–32

Contents lists available at ScienceDirect

Virus Research

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

Review

Recent advances in inferring viral diversity from high-throughput sequencing data

Susana Posada-Cespedes[a,b], David Seifert[a,b], Niko Beerenwinkel[a,b,*]

[a] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[b] SIB, Basel, Switzerland

ARTICLE INFO

Article history:
Received 24 June 2016
Received in revised form
23 September 2016
Accepted 24 September 2016
Available online 28 September 2016

Keywords:
Viral quasispecies
Genetic diversity
Haplotype reconstruction
Next-generation sequencing

ABSTRACT

Rapidly evolving RNA viruses prevail within a host as a collection of closely related variants, referred to as viral quasispecies. Advances in high-throughput sequencing (HTS) technologies have facilitated the assessment of the genetic diversity of such virus populations at an unprecedented level of detail. However, analysis of HTS data from virus populations is challenging due to short, error-prone reads. In order to account for uncertainties originating from these limitations, several computational and statistical methods have been developed for studying the genetic heterogeneity of virus population. Here, we review methods for the analysis of HTS reads, including approaches to local diversity estimation and global haplotype reconstruction. Challenges posed by aligning reads, as well as the impact of reference biases on diversity estimates are also discussed. In addition, we address some of the experimental approaches designed to improve the biological signal-to-noise ratio. In the future, computational methods for the analysis of heterogeneous virus populations are likely to continue being complemented by technological developments.

Contents

* Corresponding author
E-mail address: niko.beerenwinkel@bsse.ethz.ch (N. Beerenwinkel).

http://dx.doi.org/10.1016/j.virusres.2016.09.016

1. Introduction

The evolutionary dynamics of RNA viruses, such as the human immunodeficiency virus (HIV), the hepatitis C virus (HCV), or influenza virus, is characterized by high mutation rates, short generation times and large population sizes (Duffy et al., 2008). Under these conditions, a collection of non-identical but related genetic variants is able to co-exist within the host. This ensemble of variants has been referred to as a viral quasispecies (Domingo et al., 2005; Lauring and Andino, 2010). The term quasispecies was first used by Eigen and Schuster (1977), in the context of their work on molecular evolution (Eigen and Schuster, 1978, 1978). The quasispecies model was introduced by means of a theoretical framework using chemical kinetics to describe the mutation and selection processes governing the evolution of self-replicating macromolecules. In virology, the quasispecies model has been adopted to describe the evolutionary dynamics of RNA viruses at the population level (Nowak, 1992; Domingo and Holland, 1997).

Mutation and selection are one of the driving forces of evolution in RNA viruses. Largely due to the lack of proof-reading capability of the RNA polymerases (i.e., RNA-dependent RNA polymerase and RNA-dependent DNA polymerase or reverse transcriptase), RNA viruses exhibit high mutation rates (Duffy et al., 2008). For instance, the mutation rate of HIV-1 is on the order of 10^{-5} substitutions per position per generation (Duffy et al., 2008; Mansky and Temin, 1995). As a consequence of these high mutation rates, new viral strains are produced in every replication cycle by means of point mutations, insertions and deletions. Another common source of variability in RNA viruses is recombination. A recombination event can take place when at least two different viral strains infect the same cell, giving rise to a new strain which is a mosaic of its progenitors. On the other hand, selective pressures act upon the virus population as a whole, shaping the distribution of viral strains. For instance, in response to changing environments, the virus population quickly adapts by selecting preexisting strains with higher fitness (Bonhoeffer and Nowak, 1997). As a result, one or few viral strains dominate, surrounded by a large cloud of low-frequency variants.

The heterogeneous mixture of viral strains appears to confer numerous advantages to the virus population, including the ability to escape from the host's immune response (Nowak et al., 1991; Kuroda et al., 2010; Woo and Reifman, 2012; Borucki et al., 2013), and the development of resistance to vaccines (Gaschen et al., 2002) and antiviral drugs (Johnson et al., 2008). Furthermore, the existence of different viral strains has significant implications for viral pathogenesis, virulence, persistence and disease progression, and likely contributes to tissue tropism (Vignuzzi et al., 2006; Tsibris et al., 2009; Rozera et al., 2014). The robust adaptability featured by RNA viruses, which is related to their genetic heterogeneity is, thus, of clinical relevance. In fact, many of the infectious diseases which have jeopardized and still are a threat to public health are caused by RNA viruses, including HIV, HCV, Influenza virus, Ebola virus and Zika virus.

Before the establishment of HTS technologies, Sanger sequencing was the method of choice for analyzing virus samples. Even today, it remains the gold standard for many clinical applications. However, bulk sequencing only allows for determining the consensus sequence of the virus population. The consensus sequence is an aggregate of all variants within the population. Consequently, it is dominated by highly abundant strains and cannot be used to assess the linkage of mutations in individual variants (Wirden et al., 2005; Zagordi et al., 2010). Further experimental improvements, including isolation of individual viral strains through cloning (Domingo, 2015) or limiting dilutions (Palmer et al., 2005), allow to acquire a better, yet small, sample of the variants within the virus population. This is because these protocols are labor- and time-intensive and, thus, scalability remains a limiting factor.

The sensitivity and scalability issues are progressively being overcome by a set of newer technologies, which allow to produce massive volumes of genomic data in a relatively short time by parallelization of the sequencing reactions. These technologies are collectively referred to as high-throughput sequencing (HTS), massively parallel sequencing (MPS), next-generation sequencing (NGS) or ultra-deep sequencing (UDS). HTS technologies allow an in-depth characterization of the genetic diversity in heterogeneous virus populations by directly sequencing many of the viral strains. Furthermore, provided that the sequencing coverage is sufficiently high, it is possible to detect mutations present in less abundant strains, whereas consensus Sanger sequencing has a 20% detection threshold. However, low-frequency mutations are particularly relevant in the context of drug resistance, since they may facilitate viral adaptation leading to treatment failure (Metzner et al., 2009; Gianella and Richman, 2010; Avidor et al., 2013; Vandenhende et al., 2014). Therefore, studying the genetic diversity of the virus population as a whole is more informative than focusing solely on the dominant viral strains.

HTS technologies have the potential to provide a representative sample of the virus population. However, many HTS platforms generate large amounts of sequencing reads with short read lengths and relatively high error rates. These factors, in conjunction with errors associated with sample preparation (e.g., RNA extraction, reverse transcription and PCR amplification biases), pose computational and statistical challenges for inferring intra-host genetic diversity from HTS reads (Beerenwinkel et al., 2012; McElroy et al., 2014). For instance, many single-nucleotide variants (SNVs) are present at low frequencies and are therefore difficult to distinguish from technical errors. In addition, reconstructing the population structure from sequencing reads is challenging because the number of underlying viral strains is unknown, some of them exist at low relative abundances, and the diversity among strains can be low (i.e., some variants within the population exhibit a small genetic distance). From the technical perspective, reconstruction of full-length haplotypes is challenging because sequencing reads are typically shorter than the viral genome and do not cover the genome or the genetic region of interest uniformly. To this end, recent advances in single-molecule sequencing seem promising, as platforms commercialized by Pacific Biosciences and Oxford Nanopore offer very long reads (>10 kb). However, higher error-rates and lower throughput compared to predecessor HTS platforms still limit applicability of single-molecule sequencers.

Nevertheless, HTS technologies have already proven useful in different fields related to virology, including virus discovery (Cheval et al., 2011), characterization of virus biodiversity found in different environments (also known as virome profiling) (Hurwitz and Sullivan, 2013), estimation of fitness landscapes of viral populations (Seifert et al., 2015), characterization of intra-host virus diversity and population dynamics (Kuroda et al., 2010).

This review is structured as follows. First, we address experimental protocols which have been recently designed to overcome limitations associated with short and error-prone reads (Section 2). These sequencing protocols and accompanying data analysis pipelines have enabled correction of technical errors, as well as reconstruction of viral haplotypes. Next, acknowledging that alignment of sequencing reads is in most cases a prerequisite for subsequent analyses, strategies for read alignment are briefly discussed in Section 3, as well as remaining challenges. Lastly, we describe computational methods developed for studying the genetic diversity of virus populations from HTS reads (Section 4).