Title

Characterization of Viral Populations by Using Circular Sequencing

Permalink
https://escholarship.org/uc/item/9fs9d6n5

Journal
Journal of Virology, 90(20)

ISSN
0022-538X

Authors
Whitfield, Zachary J
Andino, Raul

Publication Date
2016-10-15

DOI
10.1128/jvi.00804-14

Peer reviewed

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology

GEM

Characterization of Viral Populations by Using Circular Sequencing

Zachary J. Whitfield, Raul Andino

Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, California, USA

With the enormous sizes viral populations reach, many variants are at too low a frequency to be detected by conventional next-generation sequencing (NGS) methods. Circular sequencing (CirSeq) is a method by which the error rate of next-generation sequencing is decreased so that even low-frequency viral variants can be accurately detected. The ability to visualize almost the entire genetic makeup of a viral swarm has implications for epidemiology, viral evolution, and vaccine design. Here we discuss experimental planning, analysis, and recent insights using CirSeq.

Viruses replicate to enormous population sizes, resulting in many genomic variants existing at extremely low frequencies within the population. However, given the error rates associated with conventional next-generation sequencing (NGS), most of these rare variants go undetected. These variants initially arise due to the low replication fidelity inherent to most viruses, particularly RNA viruses, whose RNA-dependent RNA polymerases (RdRPs) display a high error rate of nucleotide incorporation. As a result of this constant error generation, RNA virus populations are believed to exist as a "quasispecies," a swarm of closely related viral genomes under continuous evolution and selection, resulting in the most fit population for the given environment (1). However, a given mutation may be more or less fit only within a specific context (such as being more resistant to a particular immunologic defense or being more efficient at a particular stage of the viral life cycle) (2). One illustration of this was a single point mutation facilitating Chikungunya virus' jump to the *Aedes albopictus* mosquito from *Aedes aegypti*, likely facilitating a recent outbreak on Reunion Island, France (3, 4). The dramatic effect a single variant can have on viral fitness illustrates the sensibility of the viral replication strategy: create a large population size with low-fidelity replication in order to continuously "sample" the so-called sequence space around the parental genome sequence. This is not to say that high mutation rates are always beneficial; viral mutation rates have been finely tuned by evolution. Perturbing this balance in either direction (higher or lower fidelity) typically attenuates a virus (5, 6). Nonetheless, with this constant generation of variants, a viral population can more rapidly adapt to changing conditions.

LIMITS OF CONVENTIONAL NEXT-GENERATION SEQUENCING

Given that adaptation depends of the mutation composition of the virus population, it is essential to understand the dynamics with which these mutants arise and proliferate. While conventional next-generation sequencing (NGS) can detect variants at a frequency of only about 1 in 1,000 (due to sequencing errors) (7), the mutation rates in RNA viruses can range from 10^{-4} to 10^{-6} per base. Consequently, many variants in a viral population that exist at low frequencies cannot be distinguished from noise using conventional NGS. The extraordinary amount of data generated by NGS enables even small error rates to cause significant numbers of sequencing errors. On top of this, errors in cDNA synthesis (reverse transcription and second-strand synthesis) as well as PCR amplification during library generation add to the potential for error. These errors look the same to a sequencer as true genetic variation does. As a result, conventional NGS can accurately detect only variants that have already risen to significant frequencies within the viral population. This gap between the NGS limit of detection and the mutation rate of RNA viruses means the mutational composition of the virus population remains unknown. A new sequencing protocol, called circular sequencing (CirSeq) (8, 9), was developed to increase the accuracy of conventional NGS and uncover the genetic structure of the viral quasispecies.

STUDYING VIRAL POPULATION DYNAMICS USING CIRCULAR SEQUENCING

CirSeq makes use of circularized RNA to help minimize downstream errors in sequencing. During the CirSeq workflow (9), viral genomic RNA is fragmented, and then molecules are self-ligated, circularizing the molecules. These circularized RNAs serve as a template for cDNA synthesis, and the resulting "rolling circle" replication product consists of head-to-tail repeats of the circularized RNA. Only mutations present in a majority of repeats on a given molecule are considered true variants in the RNA, while errors from PCR, reverse transcription, and base calling during library sequencing should not be found in all repeats. At the core of the error correction is the random nature of error in these processes. While PCR does tend to favor transition mutations, the location of any PCR-induced mutations is considered random (10–13). Illumina HiSeq technology can exhibit a bias in substitution rates after particular 3-mer motifs (14), though these motifs are generally given a low quality score and are not expected to occur in all cDNA repeats of a given molecule. CirSeq's error detection and correction approach lowers the theoretical limit of detection well below the typical mutation rate of RNA viruses. Consequently, all variants within a viral population can theoretically be identified using CirSeq.

In a viral population, four general categories of mutations can be identified: lethal, detrimental, neutral, and beneficial. Biologically speaking, a detrimental mutation has a negative impact on a virus' ability to replicate efficiently, while a beneficial mutation

Accepted manuscript posted online 20 July 2016

Citation Whitfield ZJ, Andino R. 2016. Characterization of viral populations by using circular sequencing. J Virol 90:8950–8953. doi:10.1128/JVI.00804-14.

Editor: S. Schultz-Cherry, St. Jude Children's Research Hospital

Address correspondence to Raul Andino, raul.andino@ucsf.edu.

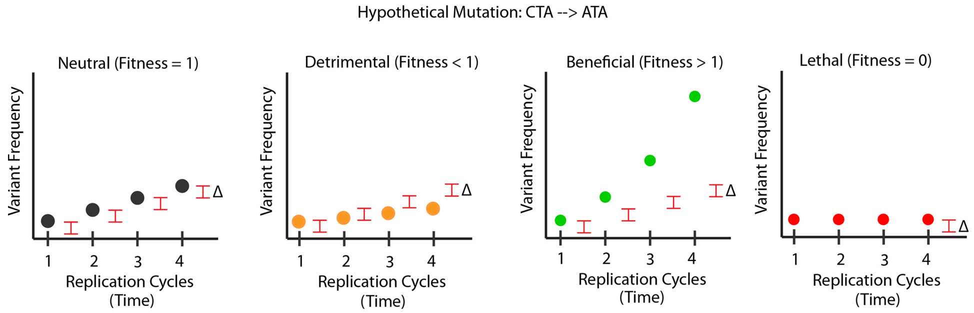

FIG 1 Illustration of variant fitness categories. Variants detected using CirSeq will be assigned a fitness based on the trajectory of their frequency during the course of the experiment. Here, a hypothetical C→A mutation can fall into one of four categories. The Δ symbol represents the inherent mutation rate for this type of variant (i.e., the overall mutation rate for C→A). Neutral mutations rise but only due to newly generated variants due to Δ. Detrimental variant frequencies may fall within the population, or could even rise, but at a level less than Δ. Beneficial mutations will increase in frequency within a population at a rate greater than Δ. Lethal mutations will be present in a population only at or below Δ.

improves some step of the viral life cycle so that it replicates more efficiently overall. A neutral mutation has little to no impact on viral replication. In the extreme case, a detrimental mutation can render a virus completely replication incompetent, in which case it is classified as a lethal mutation and should be quickly eliminated from the population. It is important to keep in mind that each type of variant can theoretically be the result of either a synonymous or nonsynonymous mutation (15–17).

These general mutation categories can also be described with respect to CirSeq analysis. Briefly speaking, neutral mutations rise in frequency only at the rate of that variant's mutation rate. A detrimental mutation is one whose frequency decreases within a population over time (or rises slower than that variant type's inherent mutation rate), while a beneficial mutation rises in frequency at a rate greater than that variant's inherent mutation rate. Lethal mutations are present in a population only at or below that variant type's mutation rate (these variants are continuously regenerated at that frequency over time) (Fig. 1).

The resolution of CirSeq permits the analysis of all four categories of mutants, even those at very low frequency. Basic CirSeq analysis (available at http://andino.ucsf.edu/toolsandprotocols) outputs the counts of each possible nucleotide at every genomic position in the reference. From this file, the detected frequency of every nucleotide variant at every position in the genome can be calculated (8, 9). Moving one step further, with the high coverage and decreased error rate of CirSeq, these frequencies make it possible to determine the overall mutation rates for every type of mutation (A→T, A→C, A→G, T→A, etc. . .). This is a powerful approach which has been used, for instance, to tease apart the contributions of a low-fidelity polymerase to each of these individual mutation rates (18). CirSeq (and deep sequencing in general) becomes more powerful when the dynamics of the population composition is examined over time. Applying CirSeq at specific instances of time (such as viral replication cycles and/or cell culture passages) allows for the assignment of fitness values to specific variants, including those at very low frequency (8). These fitness values are a quantitative approach to classifying variants as detrimental, lethal, neutral, or beneficial. Fitness can be calculated in a number of ways but generally rely on asking how the variant's frequency changes over time. Does its frequency rise, fall, or stay steady within the population? More specifically, we employ Bayesian computational methods, which take these properties into account and estimate the posterior probability distribution for the fitness of each allele in the virus population (8). Importantly, fitness assignments are independent of overall frequency within the population and are assigned based only on the change in frequency between time points. Therefore, very-low-frequency variants can still be assigned high or low fitness values.

With the ability to produce such a large volume of data, CirSeq lends deeper insight and analysis to a number of virus-specific applications. What are the dynamics of a viral population both when infecting a single individual and when spreading between individuals (or even species)? What evolutionary path does a virus follow when subverting the actions of a drug? Revisiting the concept of a viral quasispecies, how does the reservoir of low-frequency variants behave in the face of such challenges? Crucially, detection of low-frequency alleles allows for characterization of negative selection on a relatively short time scale. With the high sensitivity of CirSeq, the population dynamics of variants normally falling below the threshold of detection by NGS can still be analyzed. This analysis includes variants that exist but will never rise above conventional NGS limits of detection. By way of its high mutation rate, an RNA virus will explore its surrounding sequence space. However, as discussed above, some variants are lethal mutations, and so not all of a virus' surrounding sequence space can be sampled. The ability to identify variants that are not amenable to a given environment/experimental condition helps visualize the true sequence space a virus has available to it. This analysis could have implications in many aspects of viral evolution, including structure-function relationships, vaccine design, and epidemiology. With regard to variants that exhibit high fitness in the face of a particular challenge, how do they behave before that challenge is introduced? Do these variants tend to be neutral and then become beneficial after the introduction of some specific challenge? Or could a low-fitness mutant be maintained at a very low frequency only to "become" high fitness upon introduction of a challenge? Viral outbreaks can occur when a virus suddenly adapts to a new environment or becomes more virulent toward its current environment (or both). This adaptation starts with a minor variant establishing itself within a viral population and then taking over. CirSeq gives us a window with which to see how viral variants rise and fall in a population and the dynamics by which they do so.

To date, numerous studies have benefited from the power of CirSeq (8, 13, 18, 19). The initial studies demonstrated the ability to confidently detect ultralow-frequency variants within a ge-

TABLE 1 CirSeq checklist

Factors that must be considered in CirSeq library preparation
When planning and performing an experiment for CirSeq library preparation, it is important to ensure the following.
1. Large amounts (>1 μg) of viral RNA can be isolated for library preparation. This RNA should be relatively pure and free from contaminating cellular RNA.
● Contaminating RNA will reduce efficiency of sequencing runs, decreasing coverage and (potentially) significantly increasing cost of the experiment. A final sequencing depth of ~200,000 reads per position should be the goal.
● It is possible to take virus produced at each passage and amplify it at high MOI[a] (~10) in order to produce enough virus for RNA purification.
If the virus will be serially passaged in order to obtain fitness values, the following should be done or taken into consideration.
2. The population size used for each passage should be no less than 10^6 PFU (for typical RNA viruses).
● This helps to minimize the effects of genetic drift influencing frequency values of low-frequency variants.
3. The viral infections used for passaging should be performed at low MOIs (~0.1).
● This minimizes coinfection of cells, which could alter population dynamics of viral variants through complementation and competition.
4. The length of infection during passages should be taken into consideration.
● Ideally, it should last for one replication cycle, thus minimizing secondary infections during the process. This must be balanced with point 2 to ensure large enough population sizes for continued passaging.

[a] MOI, multiplicity of infection.

nome, at the organismal level (13) and within a poliovirus population (8). Furthermore, integrating serial passaging of an RNA virus with CirSeq allowed for a new level of insight into visualizing fitness distributions on a protein structure. This provides an important step in understanding the relationship between protein structure and function and identifying important, yet undiscovered, functional domains. As mentioned above, Korboukh et al. (18) utilized CirSeq to calculate individual mutation rates for every type of nucleotide substitution in two poliovirus populations, one with a wild-type RdRP and another with a low-fidelity RdRP (H273R). Fascinatingly, it was found that only a subset of mutation rates increased in the presence of H273R, suggesting that for some variants, the main source of mutation is polymerase independent (18). Mutations identified by CirSeq in cell culture passaging experiments have also shown a phenotype in animals. Xiao et al. used CirSeq to compare recombination-deficient and wild-type poliovirus (19). It was found that wild-type poliovirus accumulates a subset of beneficial mutations at a higher rate than its recombination-deficient counterparts. Together, these studies demonstrate a new lens through which viral population dynamics can be studied.

PROPER CirSeq EXPERIMENTAL DESIGN

In practical terms, CirSeq requires that a number of criteria be met to help ensure quality analysis further downstream (Table 1). To perform CirSeq efficiently, it is necessary to obtain at least 1 μg of pure viral RNA. Less RNA could lead to insufficient yields of the library (before amplification), risk reamplification of individual library molecules (9), and lead to less-accurate determination of low-frequency variants. This limitation may be relaxed if future improvements of the protocol allow for more-efficient processing and recovery of the viral RNA. Depending on the growth rate of the virus and purification techniques available (either at the virion or RNA level), attaining this much pure viral RNA can be a challenge. It is crucial to ensure beforehand that enough viral material can be obtained for the sequencing reaction. If serial passaging experiments are to be performed, it is also important to keep in mind the size of the viral population being passaged. Typically, at least 10^6 PFU should be passaged each time to help mitigate noise from genetic drift (8). Furthermore, the issue of coinfection during passages should be considered. When performing serial passages, significant coinfection may reduce the specific effect of individual variants. If two different viruses release their genome into the same cell, complementation and competition will occur, thereby disturbing the true fitness value of a given mutation. However, once RNA of sufficient quality and quantity is obtained, the CirSeq protocol uses all readily available commercial reagents.

Despite the incredible amount of data that can be obtained using CirSeq, there is, of course, room for improvement. Currently, CirSeq treats every variant detected in isolation. That is to say, viral haplotypes are not taken into consideration (due both to the short read length and the currently available analysis). Similarly, it will be important to incorporate epistatic interactions into the analysis of low-frequency variants. Understanding how two separate variants rely on one another and how they both establish themselves in the population will shed light on how the more-complex evolutionary landscapes of RNA viruses arise. Further optimization of the CirSeq workflow should also allow a relaxing of the minimum input of 1 μg viral RNA. Allowing for smaller amounts of input RNA will make CirSeq more amenable to situations where viral RNA is in short supply: tissue-derived samples, patient-derived samples, and poorly replicating/attenuated viruses. The experimental approaches discussed here, combined with these improvements will work to further inform attenuated vaccine design, as well as strategies viruses employ to cause rapid outbreaks such as those we have seen recently with Chikungunya, Ebola, and Zika viruses.

ACKNOWLEDGMENT

We thank Patrick Dolan for helpful comments on the manuscript.

FUNDING INFORMATION

This work, including the efforts of Raul Andino, was funded by HHS | NIH | National Institute of Allergy and Infectious Diseases (NIAID) (R01AI40085 and P01AI091575). This work was funded by University of California (CCACC). This work was funded by DOD | Defense Advanced Research Projects Agency (Prophecy program).

REFERENCES

1. **Andino R, Domingo E.** 2015. Viral quasispecies. Virology **479-480:**46–51. http://dx.doi.org/10.1016/j.virol.2015.03.022.
2. **Xue KS, Hooper KA, Ollodart AR, Dingens AS, Bloom JD.** 2016. Cooperation between distinct viral variants promotes growth of H3N2 influenza in cell culture. Elife **5:**e13974. http://dx.doi.org/10.7554/eLife.13974.
3. **Schuffenecker I, Iteman I, Michault A, Murri S, Frangeul L, Vaney MC, Lavenir R, Pardigon N, Reynes JM, Pettinelli F, Biscornet L, Diancourt L, Michel S, Duquerroy S, Guigon G, Frenkiel MP, Brehin AC, Cubito N, Despres P, Kunst F, Rey FA, Zeller H, Brisse S.** 2006. Genome microevolution of chikungunya viruses causing the Indian Ocean outbreak. PLoS Med **3:**e263. http://dx.doi.org/10.1371/journal.pmed.0030263.

4. **Tsetsarkin KA, Vanlandingham DL, McGee CE, Higgs S.** 2007. A single mutation in chikungunya virus affects vector specificity and epidemic potential. PLoS Pathog **3:**e201. http://dx.doi.org/10.1371/journal.ppat.0030201.
5. **Crotty S, Cameron CE, Andino R.** 2001. RNA virus error catastrophe: direct molecular test by using ribavirin. Proc Natl Acad Sci U S A **98:**6895–6900. http://dx.doi.org/10.1073/pnas.111085598.
6. **Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R.** 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature **439:**344–348. http://dx.doi.org/10.1038/nature04388.
7. **Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ.** 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol **30:**434–439. http://dx.doi.org/10.1038/nbt.2198.
8. **Acevedo A, Brodsky L, Andino R.** 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature **505:**686–690. http://dx.doi.org/10.1038/nature12861.
9. **Acevedo A, Andino R.** 2014. Library preparation for highly accurate population sequencing of RNA viruses. Nat Protoc **9:**1760–1769. http://dx.doi.org/10.1038/nprot.2014.118.
10. **McInerney P, Adams P, Hadi MZ.** 2014. Error rate comparison during polymerase chain reaction by DNA polymerase. Mol Biol Int **2014:**287430. http://dx.doi.org/10.1155/2014/287430.
11. **Vandenbroucke I, Van Marck H, Verhasselt P, Thys K, Mostmans W, Dumont S, Van Eygen V, Coen K, Tuefferd M, Aerssens J.** 2011. Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications. Biotechniques **51:**167–177. http://dx.doi.org/10.2144/000113733.
12. **Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B.** 2011. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A **108:**9530–9535. http://dx.doi.org/10.1073/pnas.1105422108.
13. **Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, Sawyer SL.** 2013. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proc Natl Acad Sci U S A **110:**19872–19877. http://dx.doi.org/10.1073/pnas.1319590110.
14. **Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C.** 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics **17:**125. http://dx.doi.org/10.1186/s12859-016-0976-y.
15. **Lauring AS, Acevedo A, Cooper SB, Andino R.** 2012. Codon usage determines the mutational robustness, evolutionary capacity, and virulence of an RNA virus. Cell Host Microbe **12:**623–632. http://dx.doi.org/10.1016/j.chom.2012.10.008.
16. **Burns CC, Campagnoli R, Shaw J, Vincent A, Jorba J, Kew O.** 2009. Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons. J Virol **83:**9957–9969. http://dx.doi.org/10.1128/JVI.00508-09.
17. **Burns CC, Shaw J, Campagnoli R, Jorba J, Vincent A, Quay J, Kew O.** 2006. Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region. J Virol **80:**3259–3272. http://dx.doi.org/10.1128/JVI.80.7.3259-3272.2006.
18. **Korboukh VK, Lee CA, Acevedo A, Vignuzzi M, Xiao Y, Arnold JJ, Hemperly S, Graci JD, August A, Andino R, Cameron CE.** 2014. RNA virus population diversity, an optimum for maximal fitness and virulence. J Biol Chem **289:**29531–29544. http://dx.doi.org/10.1074/jbc.M114.592303.
19. **Xiao Y, Rouzine IM, Bianco S, Acevedo A, Goldstein EF, Farkov M, Brodsky L, Andino R.** 2016. RNA recombination enhances adaptability and is required for virus spread and virulence. Cell Host Microbe **19:**493–503. http://dx.doi.org/10.1016/j.chom.2016.03.009.