Methods 91 (2015) 40–47

Contents lists available at ScienceDirect

Methods

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

CoVaMa: Co-Variation Mapper for disequilibrium analysis of mutant loci in viral populations using next-generation sequence data

Andrew Routh [a,b,c,*,1], Max W. Chang [d], Jason F. Okulicz [e,f], John E. Johnson [b], Bruce E. Torbett [a,*]

[a] Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA 92037, USA
[b] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA
[c] Department of Biochemistry and Molecular Biology, The University of Texas Medical Branch, Galveston, TX, USA
[d] Integrative Genomics and Bioinformatics Core, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[e] Infectious Disease Service, San Antonio Military Medical Center, Fort Sam Houston, TX 78234, USA
[f] Infectious Disease Clinical Research Program, Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA

ARTICLE INFO

Article history:
Received 10 August 2015
Received in revised form 18 September 2015
Accepted 21 September 2015
Available online 25 September 2015

Keywords:
Linkage disequilibrium
Flock House Virus
Human immunodeficiency virus protease
Covariation
RNAseq

ABSTRACT

Next-Generation Sequencing (NGS) has transformed our understanding of the dynamics and diversity of virus populations for human pathogens and model systems alike. Due to the sensitivity and depth of coverage in NGS, it is possible to measure the frequency of mutations that may be present even at vanishingly low frequencies within the viral population. Here, we describe a simple bioinformatic pipeline called CoVaMa (Co-Variation Mapper) scripted in Python that detects correlated patterns of mutations in a viral sample. Our algorithm takes NGS alignment data and populates large matrices of contingency tables that correspond to every possible pairwise interaction of nucleotides in the viral genome or amino acids in the chosen open reading frame. These tables are then analysed using classical linkage disequilibrium to detect and report evidence of epistasis. We test our analysis with simulated data and then apply the approach to find epistatically linked loci in Flock House Virus genomic RNA grown under controlled cell culture conditions. We also reanalyze NGS data from a large cohort of HIV infected patients and find correlated amino acid substitution events in the protease gene that have arisen in response to anti-viral therapy. This both confirms previous findings and suggests new pairs of interactions within HIV protease. The script is publically available at http://sourceforge.net/projects/covama.

1. Introduction

Viral populations are typically very diverse due to the high error-rate of their polymerases. As a result, just a single round of replication can generate many variant species of an original parent virus. The error rate of viral polymerases varies greatly among viral species, but is generally considered to be the highest for positive-strand RNA viruses [1]. This ability of viruses to sample a wide-range of mutational space is what gives rise to their ability to quickly evolve and adapt.

Next-Generation Sequencing (NGS) is well poised to characterize viral intra-host diversity and has been employed during the surveillance of viral epidemics, to monitor the development of drug resistance and importantly to further our fundamental understanding of viral evolution, their lifecycles and the functional components of their genomes. There are many bioinformatic software packages available that enable a user to detect and characterize mutations in viral genome from NGS data [2]. However, it is also important to characterize and understand how these mutations interact with one another and whether they arise independently or in a concerted manner due to a phenotypic co-dependence. There are many methods for haplotype phasing in the genome of higher-eukaryotes (reviewed in [3]). However, haplotype phasing in viruses is considerably more challenging due to the heterogeneous make-up of the viral populations, varying rates of homologous recombination [4], and the requirement for high sequence coverage to detect low-frequency events.

Abbreviations: FHV, Flock House Virus; NGS, Next-Generation Sequencing; LD, linkage disequilibrium; HIV, human immunodeficiency virus; SAM, Sequence Alignment Map; cis-RE, cis-acting response element; CoVaMa, Co-Variation Mapper; Nt, Nucleotide.

* Corresponding authors at: Department of Molecular and Experimental Medicine, The Scripps Research Institute, 10550 North Torrey Pines Rd, La Jolla, CA 92037, USA.

E-mail addresses: arouth@scripps.edu (A. Routh), betorbet@scripps.edu (B.E. Torbett).

[1] Present address: Department of Biochemistry and Molecular Biology, The University of Texas Medical Branch, Galveston, TX, USA.

http://dx.doi.org/10.1016/j.ymeth.2015.09.021
1046-2023/

To this end, a number of approaches have been proposed. These generally fall into two categories: direct and indirect approaches. Direct approaches extract information from individually sequenced fragments of DNA or sequence reads. VPhaser [5] and VPhaser 2.0 [6] employ a direct approach by measuring the frequency of nucleotide mutations, determining the probability of seeing two mutations together and comparing this prediction to the actual observed read data. This can sensitively infer intra-host diversity, but is limited by the length of the DNA fragments that are sequenced and the underlying error-rates of the sequencing platform. Indirect approaches employ mathematical or statistical models such as mutual information theory [7] and hidden-markov models [4] to infer the association of two mapped mutations from multiple individually sequenced fragments of DNA. However, this imposes certain assumptions about the expected and observed frequencies of distant mutations and the rates of recombination between them (either due to sequencing artifacts or the viral replication itself). Indirect approaches also include 'quasi-species reassembly' algorithms that aim to reconstruct the multiple variant viral genomes within a population, of which there are multiple strategies (reviewed in [8,9]).

Here, we describe a simple bioinformatic pipeline called CoVaMa (Co-Variation Mapper) that uses NGS data to search for evidence of covariation by measuring linkage disequilibrium (LD) within the mutational landscape of the virus sample. Our approach is 'direct' by virtue of counting the frequency of observed mapped nucleotides from the read alignment data and populating large matrices of contingency tables that correspond to every possible pairwise interaction of nucleotides in the viral genome. In this manner, no assumptions are made about the underlying rate or source of single nucleotide differences from the reference genome. Rather, significant epistatic pairs are reported by finding outlying LD values among the entire distribution of all other LD values measured. If provided with an open reading frame, CoVaMa is also able to detect epistatic amino acid pairs. Additionally, CoVaMa can merge the output from multiple read alignments in order to look for evidence of co-evolution of alleles within a population. CoVaMa can accept any read length as well as paired-end reads, although the computational demands of the pipeline scale accordingly with much longer reads.

In this manuscript, we use simulated NGS data to verify that expected pairs of mutations can be reliably and sensitively detected. We then demonstrate how co-variance can be found within the genome of Flock House Virus (FHV) when passaged in cell culture and analysed by RNAseq. Finally, we reanalyze RNAseq data from a large cohort of HIV infected patients who have failed to response to anti-viral therapy and find correlated mutations within the protease gene known to confer protease-inhibitor resistance as well as potentially novel associations. While we have applied this pipeline to explore RNA virus epistasis here, CoVaMa may also be applied in a broad range of settings including large DNA viruses, bacteria or even larger organisms but is limited by the computational resources available.

2. Materials and methods

2.1. Co-Variation Mapper – CoVaMa

A simple program scripted using Python was written to measure linkage disequilibrium in NGS datasets, called Co-Variation Mapper (CoVaMa). The scripts, associated manuals and test data are available at (http://sourceforge.net/projects/covama/). CoVaMa is cross-platform compatible requiring Python version 2.7 and Numpy (http://www.numpy.org/). CoVaMa is computationally intensive and run time can vary dramatically, ranging from a few minutes to a few hours depending upon the complexity, size and length of reads in a NGS dataset. However, we use an Apple Workstation with 16 Gb RAM and 8×2 core processors and have found this sufficient for the analyses described in this manuscript. The performance of CoVaMa can be improved considerably (up to $4\times$ improvement) by using the Python-interpreter and JIT compiler, PyPy (http://pypy.org/).

CoVaMa is split into three different scripts, separated to allow storing of data at intermediate stages of the analyses: (1) CoVaMa_Make_Matrices; (2) CoVaMa_Merge_Matrices; and (3) CoVaMa_Analyse_Matrices.

2.1.1. CoVaMa_Make_Matrices

Firstly, CoVaMa generates large matrices corresponding to every possible pair of nucleotides or amino acids in the reference genomic sequence. As illustrated in Fig. 1A, each point in these matrices is filled with a contingency table either 4×4 or 20×20 in size for nucleotide matrices and amino acid matrices respectively. The rows of the nucleotide contingency tables correspond to nts (A,T,G,C) at the 5′ nucleotide position of each nucleotide pair and likewise columns of the contingency table correspond to nts (A,T,G,C) at the 3′ nucleotide position. Similarly, each row and column of the 20×20 amino contingency tables corresponds to each of the 20 natural amino acids at the N and C termini respectively. The contingency tables in the matrices are generated using the Python defaultdict subclass in order to avoid building large numbers of unnecessary contingency tables that would otherwise occupy memory and reduce performance.

Next, CoVaMa extracts the relevant data contained within the Sequence Alignment/Map (SAM) files [10] and stores it in a temporary dictionary. Either single or paired-end alignment data can be used as input data. As it is likely that there are many duplicate reads with identical alignments due to PCR duplication or low-sample degeneracy, identical alignments are processed together as a group rather than individually to optimize performance and run time. An additional option (--Edge) provided in the command-line is to ignore a user-defined number of nucleotides at the 5′ and 3′ end of an aligned read. These portions of an aligned read are usually of the poorest quality and can often contain apparent variances from the reference genome due to the retention of short fragments of the Illumina adaptor sequences. Additionally, if the first or last nucleotides of a sequence read overlap the junction of a recombination event, they may erroneously map to the wild-type reference sequence but be counted as mismatched nucleotides. To mitigate this scenario, we recommend setting the --Edge option to the same value used for mismatch tolerance during the alignment of the sequence data.

Once the alignment data has been extracted, CoVaMa populates the contingency tables within the larger matrices. The alignment data provides the position in the reference genome that the reads map as well as any point mutations that may have occurred. Ambiguous ('N') nucleotides are ignored. Every nucleotide within an alignment is paired with every other nucleotide and the corresponding contingency table is populated according to which nucleotides are at found at each pair of loci. For example, for an alignment containing a 'T' at nt 20 and a 'C' at nt 40, CoVaMa will go to contingency table with the coordinates 20:40 in the nucleotide matrix and add 1 successful mapping to this table at coordinates corresponding to T:C, which is row 2, column 4 under the protocol used here. Pairs are assessed only once (i.e. nt 20 is paired with nt 40, but nt 40 is not paired with nt 20 as this would duplicate information) and a nucleotide cannot pair with itself. Therefore the number of nucleotide pairs in an alignment is given by: $n(n-1)/2$. Each read provided in the alignment data is also translated into an amino acid sequence in an open reading frame if provided by the user in the command line, up until a 'STOP'