Infection, Genetics and Evolution 32 (2015) 354–360

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Intrasegmental recombination does not contribute to the long-term evolution of group A rotavirus

Robert J. Woods *

University of Michigan Health System, Department of Internal Medicine, Division of Infectious Diseases, 3852 East Medical Center Drive, 3119 Taubman Center, Ann Arbor, MI 48109, USA

ARTICLE INFO

Article history:
Received 2 January 2015
Received in revised form 25 March 2015
Accepted 26 March 2015
Available online 3 April 2015

Keywords:
Intragenic recombination
3SEQ
Evolutionary constraint
Genome evolution
Viral evolution

ABSTRACT

Rotavirus is a genetically diverse pathogen with an eleven-segmented, double-stranded RNA genome. Intrasegmental recombination has been proposed as a potential mechanism to generate antigenic diversity and a possible route of escape from vaccine-imposed selective pressure. Here intrasegmental recombination was studied by performing a genome-wide scan across the eleven genome segments of 797 publically available rotavirus strains. Sixty-two sequences, or 0.7% of sequences analyzed, have evidence of intrasegmental homologous recombination. None of the specific recombination events is seen in more than one sequence. This uniqueness is consistent with either a spurious finding of recombination or the possibility that recombinant sequences arise naturally but are rapidly purged from the rotavirus population through selection. Arguments for the former explanation are presented. This analysis finds no evidence that intrasegmental recombination leads to ongoing transmission or plays a constructive role in rotavirus evolution. These results have practical implications for phylogenetic analyses and suggest a fundamental constraint that may have shaped rotavirus genome structure and evolution.

1. Introduction

RNA viruses are among the most evolutionarily adaptable organisms in the biological world, having an array of mechanisms for evolutionary change (Holland et al., 1982; Worobey and Holmes, 1999). Yet each of these mechanisms, mutation, migration, recombination, drift and selection, operate under constraints. Identifying the limitations imposed by these constraints offers a fundamental insight to the process of evolution, and to the underlying biology of the organism.

Group A rotavirus is a important cause of childhood mortality, leading to more than 400,000 childhood deaths globally each year (Tate et al., 2012). Two vaccines are available currently, and the WHO recommends vaccination in all countries (Zinser, 2009). As of August 2014, 69 countries have introduced rotavirus vaccines into their national vaccination programs. This global vaccination effort has the potential to impose intense selective pressure on the global rotavirus population. Describing the evolutionary mechanisms by which this pathogen may respond to this selective pressure can guide efforts to monitor for vaccine escape, and produce future generations of vaccine that are more robust to pathogen evolution.

Rotavirus has a double stranded RNA genome with eleven segments encoding six structural proteins (VP1–4, VP6 and VP7) and six nonstructural proteins (NSP1–6). NSP6 is as an out of frame protein entirely contained within the NSP5 coding sequence (Mattion et al., 1991). Rotavirus is known to evolve through accumulation of mutations, and through reassortment of whole genome segments (Estes and Kapikian, 2013). Rotavirus is a genetically diverse pathogen that is broadly present in wild and domesticated animals (Santos and Hoshino, 2005). Currently circulating human rotavirus strains represent multiple historical zoonotic introductions (Matthijnssens et al., 2008).

Little is known about whether genetic recombination within segments occurs. Several reports have identified sequences with intrasegmental recombination (Donker et al., 2011; Jere et al., 2011; Martínez-Laso et al., 2009; Parra, 2004; Phan et al., 2007a,b; Suzuki et al., 1998). Importantly, intrasegmental recombination has been reported to occur between the epitopes of the dominant antigenic protein, VP7 (Phan et al., 2007b). However the frequency of these events and there evolutionary implications are unclear. Specifically, it is unknown if these recombination events lead to sustained transmission, and whether all segments of rotavirus have similar rates of reassortment. Additionally, these studies failed to address alternative

* Address: 5510C MSRB I, SPC 5680, 1150 W. Medical Center Dr., Ann Arbor, MI 48109-5680, USA. Tel.: +1 734 763 2674; fax: +1 734 763 0544.
E-mail address: robertwo@umich.edu

http://dx.doi.org/10.1016/j.meegid.2015.03.035
1567-1348/

explanations for their finding such as the recombination occurring during PCR through template switching (Kanagawa, 2003).

The number of publically available genetic sequences for rotavirus is growing rapidly, opening the door to more sophisticated inference methods, including phylogenetic methods. Accounting for the role of recombination is critical to building appropriate phylogenetic models. This study utilized the growing number of publically available rotavirus sequences to perform a broad search for intrasegmental recombination. A large sample of whole genomes was identified and a systematic search for intrasegmental recombination was performed. The recombinant sequences identified were then individually analyzed for their contribution to ongoing evolution in rotavirus.

2. Material and methods

2.1. sequence download and processing

On November 14, 2014 a search of GenBank was performed for all sequences matching the search term "rotavirus" and having a length of 500–4000 nucleotides. Each sequence was assigned to rotavirus genome segment by performing a local blast against a consensus genome sequence with a liberal match score (Camacho et al., 2009). The strain name, year of collection, host species, and country of origin were extracted from the GenBank file using Biopython libraries in an automated script and entered into an SQLite database (Cock et al., 2009). This database was used to identify the subset of sequences belonging to strains with at least one sequence for each of the eleven segments of the genome.

2.2. Identification of recombinant sequences

Alignment for each segment from the whole genome set was performed using Muscle (Edgar, 2004), and visually inspected for appropriateness of alignment. Low quality sequences were removed (containing 10 or more unidentified 'N' nucleotides). Non-coding sequences were trimmed off, and sites with greater than 50% gaps were removed. The reading frame was preserved.

Recombinant sequences were identified with 3SEQ (Boni et al., 2007). 3SEQ analyzes three sequences at a time to calculate the non-parametric probability that one sequence is a recombinant of the other two. 3SEQ can be run on large dataset by comparing all possible triplets, and uses a statistical cutoff that corrects for multiple comparisons (Boni et al., 2007). Default setting of an adjusted p-value of 0.05 was used. The following straightforward procedure was used to identify a minimal set of sequences with no recombination: once sequences were identified as recombinant, they were removed from the alignment, and 3SEQ was re-run on the remaining sequences to determine if any additional recombinant sequences remained in the alignment. This process was repeated until no recombinant sequences could be identified. The approach was applied to each of the eleven genome segment alignments.

2.3. Analysis of recombinant sequences

In addition to identifying recombinant sequences, 3SEQ also identifies positions along the gene more closely related to each donor sequence. Visual inspection of the 3SEQ-identified cutoffs suggests they do not always accurately reflect the observed crossover points when there were multiple crossovers. Therefore, for each putative recombinant sequence, the location of crossover points was identified as follows:

1. The default crossover points were utilized from 3SEQ, which divided the entire gene sequences into two regions.
2. The sequences from the dataset that most closely match the putative recombinant in each of these two regions were identified. These matches are the sequences with the highest similarity, of those in our dataset, to the actual donor sequences for each recombinant. For brevity, these sequences will be referred to as 'donor' sequences.
3. Next, a sliding window of similarity was used to identify transition points along the recombinant sequence where similarity switched from one donor sequence to the other. These cutoffs, divided the sequence into regions likely derived from donor 1 and donor 2.

A default a sliding window of 100 nucleotides was used. This window size was manually adjusted in cases of high similarity or dissimilarity of the donor sequences, requiring a larger or smaller window respectively. Cutoffs determined by this sliding window are listed for each recombinant sequence (Supplementary data).

To determine if recombination contributes to the long-term evolution of rotavirus, I sought to identify whether any two sequences shared precisely the same recombination event, which would suggest the sequences were able to persist and be transmitted following recombination. To facilitate the ability to identify such an event, donor sequences and cutoffs identified above were used. The similarity of every sequence in the dataset in the region derived from donor 1 was plotted against the similarity in region derived from donor 2. Together these two regions cover the entire length of the alignment. Plotting the data in this way makes it readily visible when the sequences share the same recombination event, because they have high similarity in both region 1 and region 2. These plots are referred to as similarity scatter plots.

Genotype of the recombinant sequences, as well as of the donor from each recombinant sequences, was determined using the established nucleotide similarity cutoffs and a reference set kindly provided by Matthijnssen (Maes et al., 2009). Supplementary information was extracted from the GenBank files, or the primary paper when necessary, including the sequencing technology used to generate the sequences and whether the sample from which the sequence was obtained was known to be a mixed infections. Mixed infections are samples identified as having more than one sequence present for at least one segment. All sequences identified using the search algorithm described above were included in the analysis, however sequences that are both unverified and unpublished have been clearly labeled as such (Supplemental data), and should be treated with caution (Benson et al., 2012).

3. Results

3.1. Results of the GenBank search

The initial search returned in 37,400 sequences matching the search term "rotavirus" having a length between 500 and 4000 nucleotides. The majority of these sequences were from strains that did not have their entire genome sequenced. The number of strains with at least one sequence from every segment was 862. Sixty-five of these strains were removed due at least one segment having low quality as defined in Section 2.2. Thus, 797 strains were in the final dataset, with 9028 total sequences. Several strains had more than one sequence for some segments. The majority of strains, 96%, were from human infections, although 10 host species were represented (Table 1c). The sample has a global distribution, including 36 countries (Table 1b) and spans 33 years (Table 1a).

3.2. Identification of recombinant sequences

The algorithms described in Section 2.2 identified 71 potential recombinant sequences. Of these, five had short sequences of