Algorithm for History Reconstruction of Viral Recombination Events: Preliminary Results

Conference Paper · July 2018

DOI: 10.1109/IWOBI.2018.8464134

3 authors, including:

Gabriel Gonzalez
University College Dublin
72 PUBLICATIONS **875** CITATIONS
SEE PROFILE

Esteban Meneses
Costa Rican Institute of Technology (ITCR)
16 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Algorithm for History Reconstruction of Viral Recombination Events: Preliminary Results

Carlos Gómez
School of Computing
Costa Rica Institute of Technology
Cartago, Costa Rica
cestebangv1@estudiantec.cr

Gabriel González
Research Center for Zoonosis Control
Hokkaido University, Japan
gagonzalez@czc.hokudai.ac.jp

Esteban Meneses
School of Computing
Costa Rica Institute of Technology
Advanced Computing Laboratory
Costa Rica National High Technology Center
emeneses@cenat.ac.cr

***Abstract*—An accurate understanding of the evolutionary history of virus species could help in the development of prevention measures or new treatments. Most current tools offer phylogenetic analysis focused on mutations as the main evolutionary mechanism and ignore inconsistencies caused by recombination events. This work intends to describe an algorithm for history reconstruction of recombination events detected from a set of viral genomic sequences. Preliminary results are provided from the analysis of some sequences of Human Adenovirus D with the proposed algorithm.**

I. Introduction

The reconstruction of the recombination events history is an important step in the correct characterization of the horizontal transmission of genetic material that leads to the origin of new viral strains or viral species. It helps to expose the functional and metabolic pathways used by the virus to survive and thrive in the infected host. This knowledge could lead to the development of new treatments, prevention measures and clinical applications.

Recombination is a relevant mechanism in the evolutionary process of most viruses. It can impact positively in the generation of new opportunities that help viruses to survive in new environments or hosts by gaining advantage over specific pressures or conditions. Therefore, many professionals such as clinicians, molecular and evolutionary biologist, and epidemiologist are interested in the analysis and characterization of viral recombination to offer a better idea of how to compensate and control the benefits granted by this evolutionary method.

This work proposes a method able to answer the question of how to build the most likely history of a viral species. It provides information about the recombination events order of occurrence, so that we have a better idea where in that history those events took place.

Despite the existence of several up to date algorithms to detect signals of recombination events between multiple sequences, there are few approaches to analyze different genetic sequences and deduce the most likely chronological order of those recombination events in the origin of a novel strain.

Several approaches reconstruct a phylogenetic tree or network that best explains the history of a specific species. However, some of them completely ignore recombination events considering all inconsistencies as mutations [1], and some others generate a network with all recombination events but do not detect inconsistencies and do not explain the order in which those events occurred.

In Section II a background is described in order to revisit several concepts involved in the proposed algorithm to order recombination events into the most likely chronological history. Section III contains the proposed method description which has been implemented to collect the results provided in Section IV. Finally in Section V some conclusions and future work are presented.

II. Background

A. Virus classification

Viruses enclose a huge diversity of genetic compositions [2]. They can be grouped depending on whether they have RNA or DNA genomes; although some of them may have DNA during one period of their life span and RNA in another time.

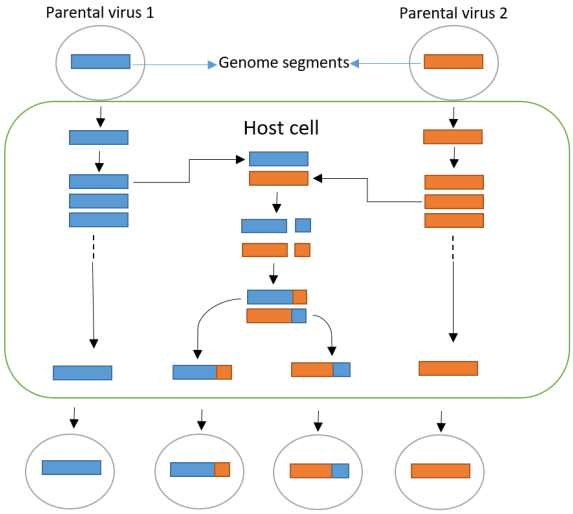
From an architectural point of view, viruses can be classified as single stranded or double stranded. There are some cases in which the viral genome is double stranded in some regions but single stranded in the other ones. Also, the genome can come in single segment or multiple segments and circular or linear segments.

Another option to classify RNA and single stranded DNA genomes is to take in account the polarity of their messenger RNA. If the translatable information is in the sense strand, we call it plus strand (+); but if it is in the antisense strand, we denote it as minus strand (-). In some cases the viral genomes could be ambisense (+/-) [3].

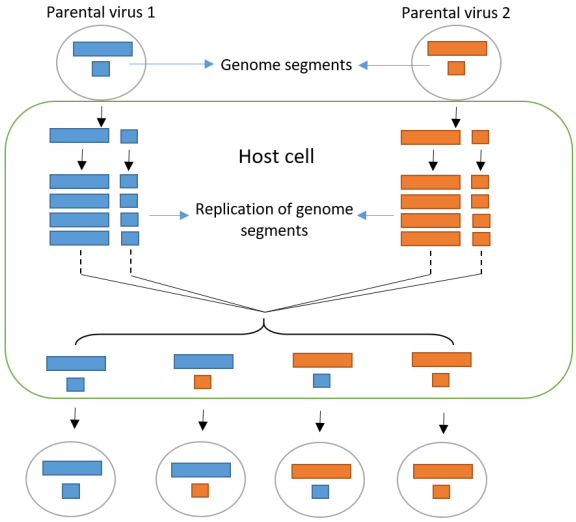
B. Human Adenovirus

Human adenoviruses (HAdVs) are part of the Adenoviridae family, their DNA genome is linear double stranded with lengths between 30 and 37 kilobase pairs. The genetic material is contained in a non enveloped icosahedral nucleocapsid.

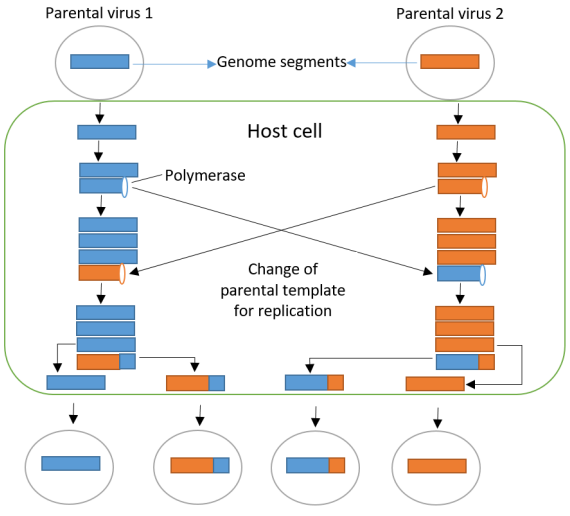
HAdVs enclose thirteen genes able to code around forty distinct proteins. They are divided into seven species identified by a letter from A to G, and each of them consist of different number of types (4, 11, 5, 43, 1, 2, and 1, respectively) [4].



(a)



(b)



(c)

Fig. 1. Viral recombination mechanisms: (a) break-rejoin of incompletely linked genes, (b) independent assortment, (c) copy-choice of incompletely linked genes

HAdV-D is the species with more types and affects only to humans. The great variety of HAdV-D types is associated to frequent recombination events that occurs among virus of this species types. Recombination between related segments in distant types genomes has been identified as a crucial evolutionary way in which so many types of HAdV-D have been produced. The quick generation of HAdV-D types has induced many widespread epidemics linked to different diseases globally.

Several diseases such as respiratory, gastrointestinal, and ocular infections are mainly caused by HAdVs. They are usually mortal in children and people with immunological deficiencies.

C. Recombination mechanisms

Viral recombination takes place when the genome of at least two different viral types interact in the same host during the replication process causing the generation of new genomic sequences with genes of both parental viruses. Usually there will be a parent that contribute with the major part of the genome, while the other parent just supply a minor portion of its genome. The viruses involved in the coinfection could have been originated from different times and places.

Recombination usually happens between viruses of the same type. Figure 1a illustrates the break-rejoin recombination mechanism. Re-assortment or independent assortment is another mechanism that implies the exchange of segments during the replication process in viruses that contain segmented genomes, the genome segments are separated and assorted arbitrary as described in Figure 1b.

Other RNA recombination mechanism is copy-choice. Under this model, the generation of a recombinant RNA sequence with mixed genomes is produced when the RNA polymerase switches from a donor RNA parent to an acceptor RNA parent during synthesis without detaching the new RNA chain. The RNA polymerase resolves viral replication, specifically in most RNA viruses it is referred as RNA-dependent RNA polymerase but in retroviruses it is reverse transcriptase [5]. Figure 1c shows the process steps.

The majority of phylodynamic studies neglect the effects of recombination or require the use of genomic regions with low recombination frequency. It usually happens because the implementation of general recombination models involve several computational and theoretical issues in phylodynamics.

D. Recombination events detection

RDP4 is a recombination detection program [6]. It offers several methods to identify, describe and visualize recombinant events from a set of virus genome sequence alignments [7].

This tool provides information about the sequences involved in a recombination event including the closest sequences to the parental ones that coinfected the same host to give raise to the event. Also, the likely recombinant segment position is estimated from the detected recombination breakpoints.

In addition to the original RDP method [8], other methods such as BootScan [9], MaxChi [10], Chimaera [11], 3Seq [12],

Geneconv [13], Lard [14], and Siscan [15] are used to detect recombination events. From the recombination signals obtained by the previous procedures, RDP4 estimates breakpoint positions using a hidden Markov model and then determines the recombinant sequence through algorithms such as EEEP [16], PhylPro [17], and VisRD [18].

Other method for recombination events detection measures distances between unrooted topologies related to the number of recombinations. Then from a prior distribution of distances and a Bayesian hierarchical model phylogenetic inconsistencies due to recombinations are detected [19].

E. Molecular clock

Molecular clocks are estimations about the date when two or more genetic sequences have diverged. It is computed through a genetic distance that provides a measure of how many changes have occurred between genetic sequences from their divergence, and a calibration rate that gives information about the predicted number of genetic changes per unit of time [20].

The approach or model to estimate both the genetic distance and the calibration rate must be selected carefully in order to obtain accurate molecular clock values. For example, counting the number of nucleotides that are different between two DNA sequences is a simple way to get their genetic distance; however, a nucleotide in the same position might have changed more than one time. Therefore, simple approaches could introduce miscalculations that will move away the estimation of the real value.

The calibration rate can be assumed constant when comparing sequences of the same species. In that case, the genetic distance provides enough information to determine if a sequence diverged more recently than another one from a common ancestor.

F. Related Work

In order to describe the ancestry of a set of recombined sequences, it is possible to build a complex graph that includes horizontal evolution such as coalescence and recombination events. Some tools have been developed for the inference of ancestral recombination graphs (ARGs) [21].

For example, Bacter [22] [23] allows the inference of ARGs from bacterial genomes including recombinant edges describing the conversion events and sites affected. Also, it offers the inference of tract lengths, recombination rates, population dynamics, and substitution rate. The inferences resulting from this tool depend heavily on the data set size, the sampling procedure, and the recombination and mutation rates of the population under study. Therefore, in case of using Bacter for viral genomes it would be necessary to analyze all those variables first.

A significant disadvantage of methods for ARG inference such as Bacter which uses a Markov Chain Monte Carlo algorithm is that they require high computational power and a parametric model that accurately describes the genome in study. Consequently, analyses often take long time and are highly approximate.

III. Method description

The proposed algorithm requires as inputs a file with the viral genomic aligned sequences in FASTA format and another file with the recombination events information obtained from applications such as RDP4 Beta 4.95 in CSV format. The information required for each event is a unique ID used for general reference, the recombinant sequence, the major parent, the minor parent, and the position where the recombinant segment begins and ends.

Genetic sequences were obtained from GenBank which is the genetic sequence database of the National Institute of Health. It is an annotated compilation of all DNA sequences openly available. The DNA DataBank of Japan, the European Nucleotide Archive, and GenBank at National Center for Biotechnology Information constitute the International Nucleotide Sequence Database Collaboration from which GenBank is part. Those three organizations transfer data everyday. When a nucleotide or protein sequence is submitted to GenBank, it is assigned an accession number which is a unique identifier for the sequence.

The sequences alignment was done by using the Iterative Refinement Method algorithm of MAFFT which is a multiple sequence alignment program [24]. According to the authors the CPU time is notoriously reduced in comparison with other existing approaches. MAFFT uses the fast Fourier transform in order to identify similar regions quickly by converting aminoacid sequences into a volume and polarity values sequence. Then, a scoring system helps to reduce CPU time and increase the accuracy of alignments. A progressive method and an iterative refinement method integrate two different heuristics implemented by the MAFFT.

Once the input files are read and parsed, several genetic distances need to be computed. Given a recombination event with the major and minor parent of the recombinant sequence, four distances can be defined as follows:

- D1: distance from major parent to recombinant sequence excluding all recombinant segments
- D2: distance from minor parent to recombinant sequence excluding all recombinant segments
- D3: distance between minor parent and recombinant segments
- D4: distance between major parent and recombinant segments

After evolutionary distances for all recombination events are estimated, each event is represented as a small directed graph as shown in Figure 2a, where each event node is linked to its associated genetic sequence nodes by directed edges weighted with the corresponding distance values. A directed graph (DG) is built by merging common sequence nodes among all events. Some events could have the same input sequence reported as minor parent, major parent, or recombinant. It makes several events sharing the same node and very likely will produce internal cycles where recombinant nodes, associated to one event, can be a parent node for other events, and so on until

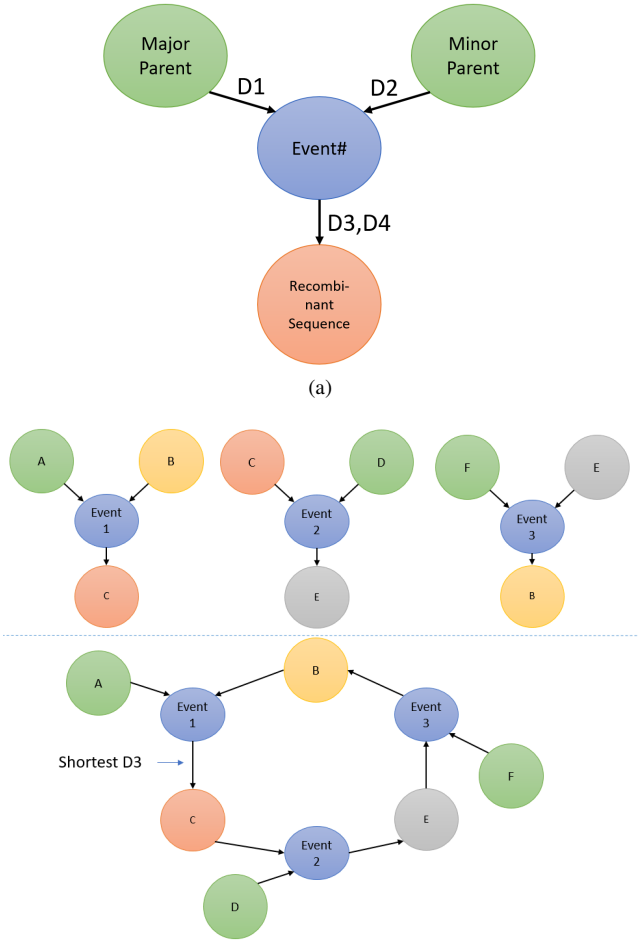


Fig. 2. Directed graph representations for (a) a single recombination event, (b) three different events

reach an event where its recombinant node is a parent of the initial event, as shown in Figure 2b.

A. Breaking cycles

The initial histories are represented as direct acyclic graphs (DAGs), so if there are cycles they need to be broken by removing specific edges that link event nodes to recombinant nodes. Only one edge per cycle needs to be broken, so it is selected by the shortest D3 distance, by this way the event node with the shortest distance in the cycle becomes a leaf indicating that it's the most recent event in that cycle. For example, in the history shown in Figure 2b it could be assumed that $Event\ 1$ has the least $D3$ distance, so the cycle could be broken by removing the edge that goes from $Event\ 1$ node to sequence C node. By this way the DAG presented in Figure 3 is obtained and considered as a possible history.

B. Choosing best history

For each generated history a metric is computed, it is an average of fractions that indicates how many event edges in all paths from roots to leaves satisfy that previous edge has a larger D3 distance than the current edge. The metric values go

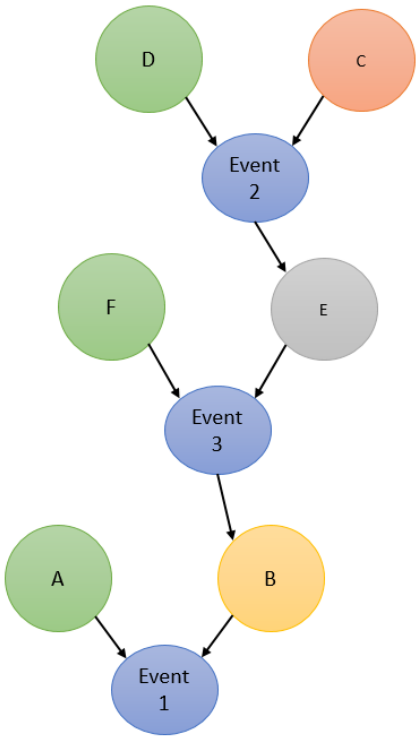


Fig. 3. Directed acyclic graph for three different recombination events

from zero to one, and the closer to one it is better. Closer to one means most of the graph event nodes are properly ordered, while closer to zero means less event nodes are properly ordered.

A mathematical expression for the consistency metric can be found in Equation 1, where M is the whole history metric, L is the number of paths going from roots to leaves, and m_i is the partial metric for each path i. The m_i value is computed according to Equation 2, where P_i is the total number of event nodes E_j in the path i, and the function λ is expressed by Equation 3 which returns 0 if distance D3 of an existing previous event E_{j-1} in path i is less than the distance D3 of current event E_j, and returns 1 otherwise.

$$M = \frac{\sum_{i=1}^{L} m_i}{L} \tag{1}$$

$$m_i = \frac{\sum_{j=1}^{P_i} \lambda(E_j)}{P_i} \tag{2}$$

$$\lambda(E_j) = \begin{cases} 0, \ if\ \exists\ D3(E_{j-1}) \mid D3(E_{j-1}) < D3(E_j) \\ 1, \ otherwise \end{cases} \tag{3}$$

D3 has been chosen as the main distance because when a recombination event occurs a segment of minor parent is copied or transferred to the recombinant sequence. From that moment that segment can suffer mutations that are directly reflected in D3 and through the mutation rate we could know when the

event took place. D2 and D4 have the inconvenience that they tell about the parent sequences divergence but not precisely about the event. D1 can tell about the event occurrence time from the major parent perspective but it is more complex to get reliable values because it is necessary to make sure other recombinant segments are not included into the estimation. Of course D1, D2, and D4 could be useful to double check the ordering makes sense but using D3 is enough.

C. Ordering events

Once the best history is chosen based on the consistency metric, a topological order of events starting from root nodes to leaf nodes is obtained through a breadth-first search (BFS) process. With this order it's possible to classify the events by depth level and know which ones necessarily had to happen first and which ones afterwards. Events that belong to the same level could be treated as concurrent but an order between them could be set by using the distance value between minor parent and recombinant segments. The less the distance the most recent the event, so this provides information about which one might happen before the others.

D. Visualizing history

In order to visualize the history, it can be drawn as a DAG; however, it is inconvenient as nodes might not be located for a clear visualization. Our implemented option is a diagram with genomic sequences as vertical bars and events distributed by levels from top to bottom as horizontal arrows that go from minor parents to major parents and then to recombinant sequence, meaning that the minor parent segment was recombined into the major parent to produce the recombinant sequence. Several events could be grouped into the same level represented by an horizontal rectangle. It means that they could be treated as concurrent events, although a slight order can be provided based on the D3 distance.

For instance, it could be assumed that the DAG in Figure 3 is the best history, so it can be represented as the diagram shown in Figure 4, where the six sequences involved correspond to the blue vertical bars and the events are drawn with two arrows that go in orange from minor parent to major parent and then in green from the major parent to the recombinant sequence. Also, the order is represented in three levels depicted as horizontal rectangles and in general the order from top to bottom goes from the oldest event to the most recent event.

IV. Results

The method described in the previous section and summarized in the flow diagram presented in Figure 5 was implemented with the support of the following libraries: Biopython 1.71 and NetworkX 2.1. An example of a recombination history reconstruction is presented for eight selected DNA sequences of Human Adenovirus D shown in Table I. Also, the adenovirus type for each sequence is provided.

The recombination events detected by RDP4 from the eight sequences is presented in Table II. All sequences are involved in many events and play different roles depending on the time

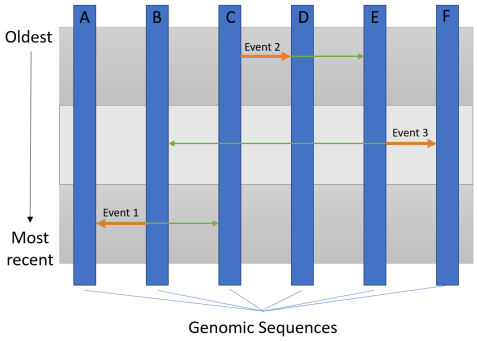


Fig. 4. Diagram of a final history showing three recombination events ordered by the proposed method

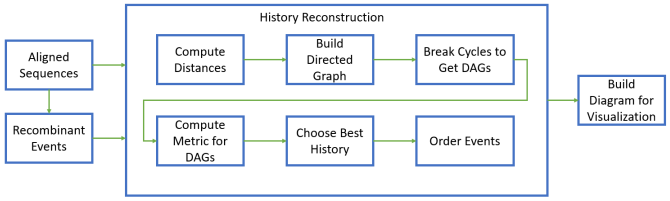


Fig. 5. Flow diagram for the recombination events history reconstruction implementation

TABLE I
Sequences of Human Adenovirus D

Sequence ID	GenBank Accession Number	Adenovirus Type
A	$AJ854486.1$	9
B	$HM770721.2$	56
C	$KF268201.1$	P33H56F56
D	$AP012285.1$	65
E	$JQ326209.1$	19
F	$AB765926.1$	81
G	$FJ404771.1$	22
H	$JN226761.1$	42

in which they occurred. Through the begin and end position of the recombinant segment, its length can be computed. The shortest recombinant segment length is 115 nucleotides, the longest is 16838, and the average length is 3665 nucleotides.

With genomic sequences and RDP4 recombinant events data as inputs for the method proposed, the history can be represented as the DAG shown in Figure 6. The lack of a layout to draw the graph nodes in a hierarchical order makes it hard to appreciate an events history, so it was required some extra processing in order to provide an ordered list of events and draw separately a history diagram for better visualization. From the showed DAG it is possible to observe some details such as some leaf event nodes in the graph periphery, and a couple of root nodes (sequences B and E). Currently the implementation just ensures it is a DAG and provides information about the events order, graph metrics were not exposed for a deeper analysis.

For better understanding of the events order it was created

TABLE II
RECOMBINATION EVENTS DETECTED BY RDP4 FROM SEQUENCES OF HUMAN ADENOVIRUS D

ID	Recombinant	Major	Minor	Begin	End
1	D	F	B	27120	33218
2	F	H	G	14699	31430
3	C	B	A	1454	18292
4	H	C	E	27764	29326
5	G	C	E	27768	29280
6	A	H	D	31811	32878
7	D	E	F	34295	35426
8	A	B	E	18279	21167
9	G	H	A	4006	12952
10	E	D	A	1310	14594
11	G	H	B	18274	26722
12	G	H	A	33540	35841
13	D	F	B	18529	19312
14	F	D	B	1472	3859
15	E	H	C	18702	22170
16	D	B	G	720	1145
17	H	E	D	29482	29810
18	C	B	A	23075	27119
19	G	F	C	30540	30897
20	D	E	G	29834	32224
21	D	E	H	27242	27487
22	D	E	H	33427	33722
23	G	B	D	13089	14076
24	C	E	A	14830	16588
25	C	B	A	19421	20385
26	G	H	B	14806	17583
27	F	C	H	1008	1123
28	D	A	F	2206	2715

the history diagram in Figure 7 which has a metric of 0.8067 with L=25. In order to interpret the diagram correctly, each vertical blue bar represents a genomic sequence, each couple of horizontal arrows are events that are ordered from the oldest at the top to the most recent at the bottom. The first orange arrow goes from minor parent to major parent, and the second green arrow goes from major parent to the recombinant sequence. Also, each event is enumerated according to the ID in Table II.

The events are grouped in five strict order levels depicted as gray rectangles. Events inside the same level group can be considered concurrent, but events located in a top level necessarily happened first than events located in a bottom level. According to distance D3 it is possible to give events inside same level a slight order as shown in the diagram.

V. CONCLUSION

We describe a method to build a history of recombination events from a set of viral sequences, based mainly on the genetic distance between parental and recombinant segments involved on each event. Although the method does not infer or provide information about specific dates for when the events

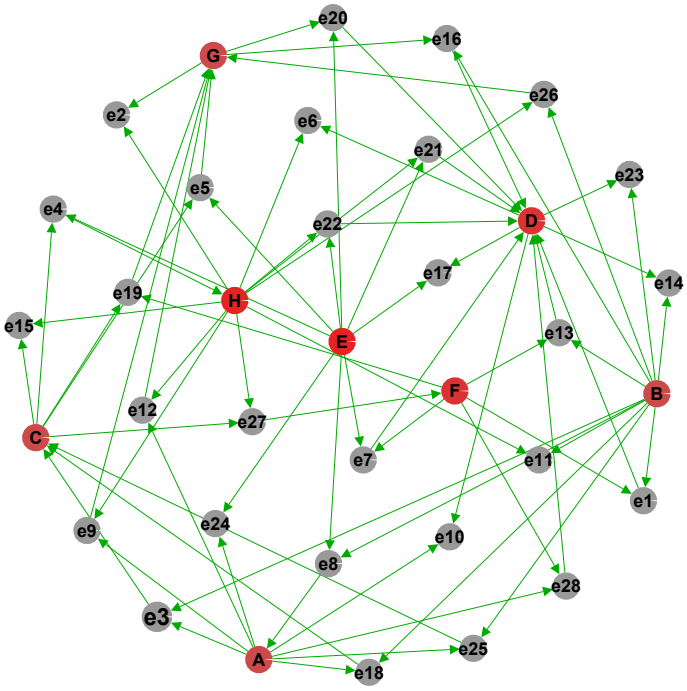


Fig. 6. Directed acyclic graph representing a recombination events history

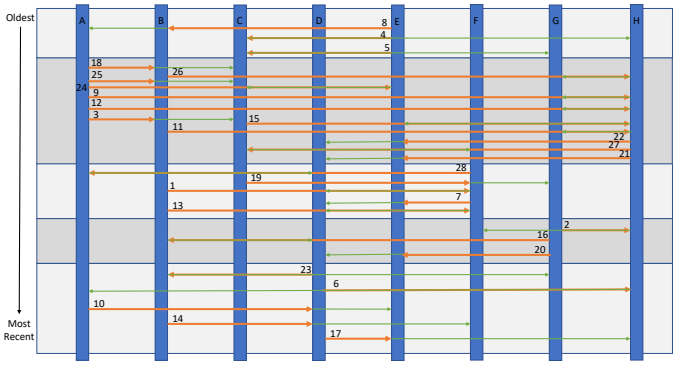


Fig. 7. Possible history of the recombination events of eight HAdVs

took place, it offers a relative order that allows to interpret whether a recombination event occurred first or later than another one. We also propose a metric to establish whether a recombination events history better explains the occurrence order from the given data, but it's still required an evaluation method that enables comparison with other recombination history representations.

APPENDIX
IMPLEMENTED ALGORITHM TO COMPUTE THE HISTORY CONSISTENCY METRIC

For better understanding of how the metric described in section III.B is computed check the pseudo-code shown in Figure 8. In this context a simple path in a recombination events DAG will never contain repeating nodes and it has the restriction that it just can go from a root to a leaf node. Once all possible simple paths has been collected, they are ordered

$history \leftarrow inputHistory$
$roots \leftarrow inputHistoryRoots$
$leaves \leftarrow inputHistoryLeaves$
$\{The\ algorithm\ inputs\ are\ a\ history\ represented\ as\ a\ DAG,\ a\ list\ of\ the\ DAG\ root\ nodes,\ and\ a\ list\ of\ the\ DAG\ leaf\ nodes\}$
$\{Get\ the\ first\ N\ longest\ simple\ paths\ from\ roots\ to\ leaves\ that\ contain\ all\ nodes\ in\ history\}$
$paths \leftarrow simplePaths$
for all $path \in paths$ **do**
  $lastDistance \leftarrow 0$
  $totalComparisons \leftarrow 0$
  $positiveComparisons \leftarrow 0$
  for all $event \in path$ **do**
    $currentDistace \leftarrow event\ distance\ D3$
    if $lastDistance \neq 0$ **then**
      $totalComparisons \leftarrow totalComparisons + 1$
      if $lastDistance > currentDistance$ **then**
        $positiveComparisons \leftarrow positiveComparisons + 1$
      end if
      $lastDistance \leftarrow currentDistance$
    else
      $lastDistance \leftarrow currentDistance$
    end if
  end for
  if $totalComparisons \neq 0$ **then**
    $partialMetric \leftarrow positiveComparisons/totalComparisons$
  else
    $partialMetric \leftarrow 1\ \{The\ path\ is\ just\ one\ event\}$
  end if
  $metric \leftarrow metric + partialMetric$
end for
$metric \leftarrow metric/pathsLength$
return $metric$

Fig. 8. Algorithm to compute the consistency metric for a given recombination event history

according to their length (number of nodes) and the first N longest paths are selected for the rest of the algorithm. The number N is variable but the minimum value depends on the number of paths necessary to cover all existing nodes, that is to say each node must appear at least once on the selected paths. The maximum value would be all possible simple paths on the DAG but depending on the graph its necessary to limit the value because of the computational resources availability.

REFERENCES

[1] M. H. Schierup and J. Hein, "Consequences of recombination on traditional phylogenetic analysis." *Genetics*, vol. 156, no. 2, pp. 879–891, Oct 2000, 11014833[pmid]. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461297/

[2] M. Pérez-Losada, M. Arenas, J. C. Galán, F. Palero, and F. González-Candelas, "Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences," *Infection, Genetics and Evolution*, vol. 30, pp. 296 – 307, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S156713481400478X

[3] D. Baltimore, "Expression of animal virus genomes." *Bacteriol Rev*, vol. 35, no. 3, pp. 235–241, Sep 1971, 4329869[pmid]. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC378387/

[4] G. Gonzalez, K. O. Koyanagi, K. Aoki, and H. Watanabe, "Interregional coevolution analysis revealing functional and structural interrelatedness between different genomic regions in human mastadenovirus d," *Journal of Virology*, vol. 89, no. 12, pp. 6209–6217, 2015. [Online]. Available: http://jvi.asm.org/content/89/12/6209.abstract

[5] E. Simon-Loriere and E. C. Holmes, "Why do rna viruses recombine?" *Nat Rev Micro*, vol. 9, no. 8, pp. 617–626, Aug 2011. [Online]. Available: http://dx.doi.org/10.1038/nrmicro2614

[6] D. P. Martin, B. Murrell, M. Golden, A. Khoosal, and B. Muhire, "Rdp4: Detection and analysis of recombination patterns in virus genomes," *Virus Evol*, vol. 1, no. 1, p. vev003, May 2015, vev003[PII]. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5014473/

[7] E. U. Martynova, C. Schal, and D. V. Mukha, "Effects of recombination on densovirus phylogeny," *Archives of Virology*, vol. 161, no. 1, pp. 63–75, Jan 2016. [Online]. Available: https://doi.org/10.1007/s00705-015-2642-5

[8] D. Martin and E. Rybicki, "Rdp: detection of recombination amongst aligned sequences," *Bioinformatics*, vol. 16, no. 6, pp. 562–563, 2000. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/16.6.562

[9] M. O. Salminen, J. K. Carr, D. S. Burke, and F. E. McCutchan, "Identification of breakpoints in intergenotypic recombinants of hiv type 1 by bootscanning," *AIDS Research and Human Retroviruses*, vol. 11, no. 11, pp. 1423–1425, 1995, pMID: 8573403. [Online]. Available: https://doi.org/10.1089/aid.1995.11.1423

[10] J. M. Smith, "Analyzing the mosaic structure of genes," *Journal of Molecular Evolution*, vol. 34, no. 2, pp. 126–129, 1992. [Online]. Available: https://doi.org/10.1007/BF00182389

[11] D. Posada and K. A. Crandall, "Evaluation of methods for detecting recombination from dna sequences: Computer simulations," *Proceedings of the National Academy of Sciences*, vol. 98, no. 24, pp. 13 757–13 762, 2001. [Online]. Available: http://www.pnas.org/content/98/24/13757

[12] M. F. Boni, D. Posada, and M. W. Feldman, "An exact nonparametric method for inferring mosaic structure in sequence triplets," *Genetics*, vol. 176, no. 2, pp. 1035–1047, 2007. [Online]. Available: http://www.genetics.org/content/176/2/1035

[13] M. Padidam, S. Sawyer, and C. M. Fauquet, "Possible emergence of new geminiviruses by frequent recombination," *Virology*, vol. 265, no. 2, pp. 218 – 225, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0042682299900569

[14] E. C. Holmes, M. Worobey, and A. Rambaut, "Phylogenetic evidence for recombination in dengue virus." *Molecular Biology and Evolution*, vol. 16, no. 3, pp. 405–409, 1999. [Online]. Available: http://dx.doi.org/10.1093/oxfordjournals.molbev.a026121

[15] M. J. Gibbs, J. S. Armstrong, and A. J. Gibbs, "Sister-scanning: a monte carlo procedure for assessing signals in recombinant sequences," *Bioinformatics*, vol. 16, no. 7, pp. 573–582, 2000. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/16.7.573

[16] R. G. Beiko and N. Hamilton, "Phylogenetic identification of lateral genetic transfer events," *BMC Evolutionary Biology*, vol. 6, no. 1, p. 15, 2006. [Online]. Available: https://doi.org/10.1186/1471-2148-6-15

[17] G. F. Weiller, "Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences." *Molecular Biology and Evolution*, vol. 15, no. 3, pp. 326–335, 1998. [Online]. Available: http://dx.doi.org/10.1093/oxfordjournals.molbev.a025929

[18] P. Lemey, M. Lott, D. P. Martin, and V. Moulton, "Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning," *BMC Bioinformatics*, vol. 10, no. 1, p. 126, 2009. [Online]. Available: https://doi.org/10.1186/1471-2105-10-126

[19] L. de Oliveira Martins, E. Leal, and H. Kishino, "Phylogenetic detection of recombination with a bayesian prior on the distance between trees," *PLOS ONE*, vol. 3, no. 7, pp. 1–13, 07 2008. [Online]. Available: https://doi.org/10.1371/journal.pone.0002651

[20] L. Bromham and D. Penny, "The modern molecular clock," *Nat Rev Genet*, vol. 4, no. 3, pp. 216–224, Mar 2003. [Online]. Available: http://dx.doi.org/10.1038/nrg1020

[21] P. G. Cámara, A. J. Levine, and R. Rabadán, "Inference of ancestral recombination graphs through topological data analysis," *PLoS Comput*

Biol, vol. 12, no. 8, p. e1005071, Aug 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4988722/

[22] T. G. Vaughan, D. Welch, A. J. Drummond, P. J. Biggs, T. George, and N. P. French, "Inferring ancestral recombination graphs from bacterial genomic data," *Genetics*, vol. 205, no. 2, pp. 857–870, 2017. [Online]. Available: http://www.genetics.org/content/205/2/857

[23] X. Didelot, D. Lawson, A. Darling, and D. Falush, "Inference of homologous recombination in bacteria using whole-genome sequences," *Genetics*, vol. 186, no. 4, pp. 1435–1449, 2010. [Online]. Available: http://www.genetics.org/content/186/4/1435

[24] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata, "Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform," *Nucleic Acids Res*, vol. 30, no. 14, pp. 3059–3066, Jul 2002, gkf436[PII]. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC135756/