


RESEARCH

Open Access

Using earth mover's distance for viral outbreak investigations



Andrew Melnyk^{1*} , Sergey Knyazev¹, Fredrik Vannberg², Leonid Bunimovich², Pavel Skums¹ and Alex Zelikovsky^{1,3}

From 15th International Symposium on Bioinformatics Research and Applications (ISBRA '19) Barcelona, Spain. 3-6 June 2019

Abstract

Background: RNA viruses mutate at extremely high rates, forming an intra-host viral population of closely related variants, which allows them to evade the host's immune system and makes them particularly dangerous. Viral outbreaks pose a significant threat for public health, and, in order to deal with it, it is critical to infer transmission clusters, i.e., decide whether two viral samples belong to the same outbreak. Next-generation sequencing (NGS) can significantly help in tackling outbreak-related problems. While NGS data is first obtained as short reads, existing methods rely on assembled sequences. This requires reconstruction of the entire viral population, which is complicated, error-prone and time-consuming.

Results: The experimental validation using sequencing data from HCV outbreaks shows that the proposed algorithm can successfully identify genetic relatedness between viral populations, infer transmission direction, transmission clusters and outbreak sources, as well as decide whether the source is present in the sequenced outbreak sample and identify it.

Conclusions: Introduced algorithm allows to cluster genetically related samples, infer transmission directions and predict sources of outbreaks. Validation on experimental data demonstrated that algorithm is able to reconstruct various transmission characteristics. Advantage of the method is the ability to bypass cumbersome read assembly, thus eliminating the chance to introduce new errors, and saving processing time by allowing to use raw NGS reads.

Keywords: Genetic relatedness, Transmission networks, Outbreaks investigations, K-mers, De Bruijn graph

Background

RNA viruses mutate at extremely high rates, forming an intra-host viral population of closely related variants (or quasi-species). Their high variability [1] allows them to evade the host's immune system and makes them particularly dangerous. Viral outbreaks pose a significant threat for public health, and, in order to deal with it, it is critical to infer transmission clusters, i.e., decide whether two viral samples belong to the same outbreak.

The progress of sequencing technologies made it possible to identify and sample intra-host viral populations

at great depth [2–7]. Consequently, contribution of sequencing technologies to molecular surveillance of viral outbreaks becomes more and more substantial. Genome sequencing of viral populations reveals similarities between samples, allows to measure viral genetic distance, and to facilitate outbreak identification and isolation. Computational methods can be used to infer transmission characteristics from sequencing data. MiSeq [8] is a popular NGS technology, that is used to sequence viral samples and detect rare viral mutations. Since MiSeq reads are short, their alignment and assembly for rapidly mutating RNA viruses is error-prone and complicated, which makes it appealing to develop an approach, that will allow to skip alignment and assembly steps.

* Correspondence: andrew.s.melnyk@gmail.com

¹Computer Science Department, Georgia State University, 25 Park Place NE, Atlanta, GA 30303, USA

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

In this paper, we apply an alignment- and assembly-free k -mer strategy to viral sequencing data. This strategy was initially introduced for analyzing NGS data in metagenomic studies, where reads come from multiple related and unrelated genomes (see [9]), as well as for RNA-seq quantification [10].

Indeed, it is relatively fast and easy to extract k -mers from reads, so that the complexity of viral distance measurement changes from read alignment and assembly to comparison of k -mer sets or distributions. Following [9], we build a De Bruijn graph for each sample, and then calculate Earth Mover’s Distance (EMD) between two k -mer distributions.

We applied the k -mer strategy to the following epidemiological tasks (T1-T5), where T1-T2 are applied to 2 hosts, and T3-T5 are applied to multiple hosts.

T1. Identification of relatedness:

Given: NGS reads from hosts A and B

Decide: Whether A and B are related (whether they belong to the same outbreak)

T2. Identification of transmission direction:

Given: NGS reads from hosts A and B

Decide: Whether host A infected B or B infected A

T3. Identification of transmission clusters:

Given: NGS reads from a set of hosts

Find: The transmission clusters corresponding to individual outbreaks

T4. Presence of outbreak source:

Given: NGS reads from a set of hosts

Decide: Whether outbreak source is present among sequenced hosts

T5. Identification of outbreak source:

Given: NGS reads from a set of hosts

Find: Outbreak source

Identifying whether 2 hosts belong to the same outbreak (T1) and transmission direction between them (T2) are tasks, that have to be solved in order to find

transmission chains. Another important task is to discover boundaries of an outbreak (T3). Once hosts, that belong to an outbreak are obtained, it is critical to design whether the source is among them (T4). Finally, identifying the main spreader of an outbreak (T5) is a crucial epidemiological task, by solving which outbreak spreading can be prevented.

We experimentally validated our approach on a dataset, that consists of a collection of HCV intra-host populations, sampled from 368 infected individuals [11].

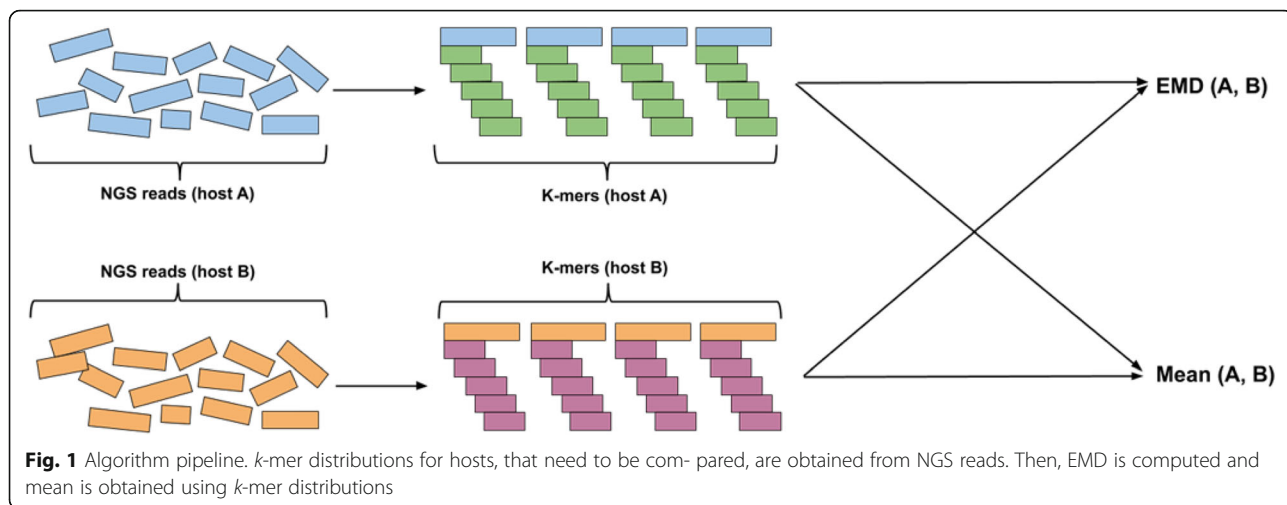
Outbreak collection contains:

- 175 HCV samples from 34 epidemiologically curated outbreaks, reported to Centers for Disease Control and Prevention in 2008–2013. Outbreaks contain from 2 to 33 samples. Epidemiological histories, including sources of infection, are known for 11 outbreaks.
- Collection of 193 epidemiologically unrelated HCV samples.

Obtained results are comparable with existing approaches [11, 12], but proposed algorithm is much faster, since it doesn’t rely on read assembly.

Methods

Our algorithms are based on finding the distance between populations using *Earth Movers’ Distance (EMD)* between distributions of k -mers in NGS data. The general pipeline of the algorithm (see Fig. 1) includes obtaining k -mer distributions from NGS reads for corresponding hosts and computing EMD between them. As a result, we obtain mean of hosts A and B $Mean(A, B)$ and $EMD(EMD(A, B))$ between them. We first describe how we find distances between k -mers and then describe how we find distance between samples.



Finding distances between k-mers in the De Bruijn graph

k-mer refers to a substring of length *k*. In our work, we use *De Bruijn graph* to calculate distance between k-mers. De Bruijn graph is the graph, that is constructed so that vertices represent every string over a finite alphabet of length *l*, and edges are added between vertices that have overlap of *l - 1*.

Once De Bruijn graph is constructed, distance between k-mers can be calculated as a length of shortest path between corresponding vertices using *breadth-first search* algorithm. In our algorithms, obtained graph is converted to undirected before shortest path computation.

Finding EMD between viral samples

Viral populations can be compared by comparing the corresponding *k-mer* distributions using EMD. First, *k-mer* distributions are obtained for each sample, so that they contain all *k-mers* and normalized frequencies.

EMD is a method, that allows to evaluate dissimilarity between two multi-dimensional distributions in some feature space where a distance measure between single features (*ground distance*) is given [13]. Distributions can be represented as *signatures* - sets of clusters, so that each cluster is represented by its mean and by the fraction of distribution that belongs to that cluster. Computation of EMD is based on solving the *transportation problem*, which can be formulated as following: for several suppliers, each with a given amount of goods, several consumers, each with limited capacity, and a cost of transporting a single unit of goods between each supplier-consumer pair, find a least-expensive flow of

goods from the suppliers to the consumers that satisfies the consumers' demand. EMD is calculated as the following: $EMD(P, Q) = \sum_{i=1}^m \sum_{j=1}^n f_{ij} d_{ij}$ where f_{ij} is the minimum-cost flow between supplier *i* and consumer *j*, and d_{ij} is the distance between *i* and *j*.

It should also be noted that EMD is usually normalized by the total flow, but we perform but we perform normalization of frequencies in *k-mer* distributions before EMD computation, which results in total flow always being equal to 1.

```

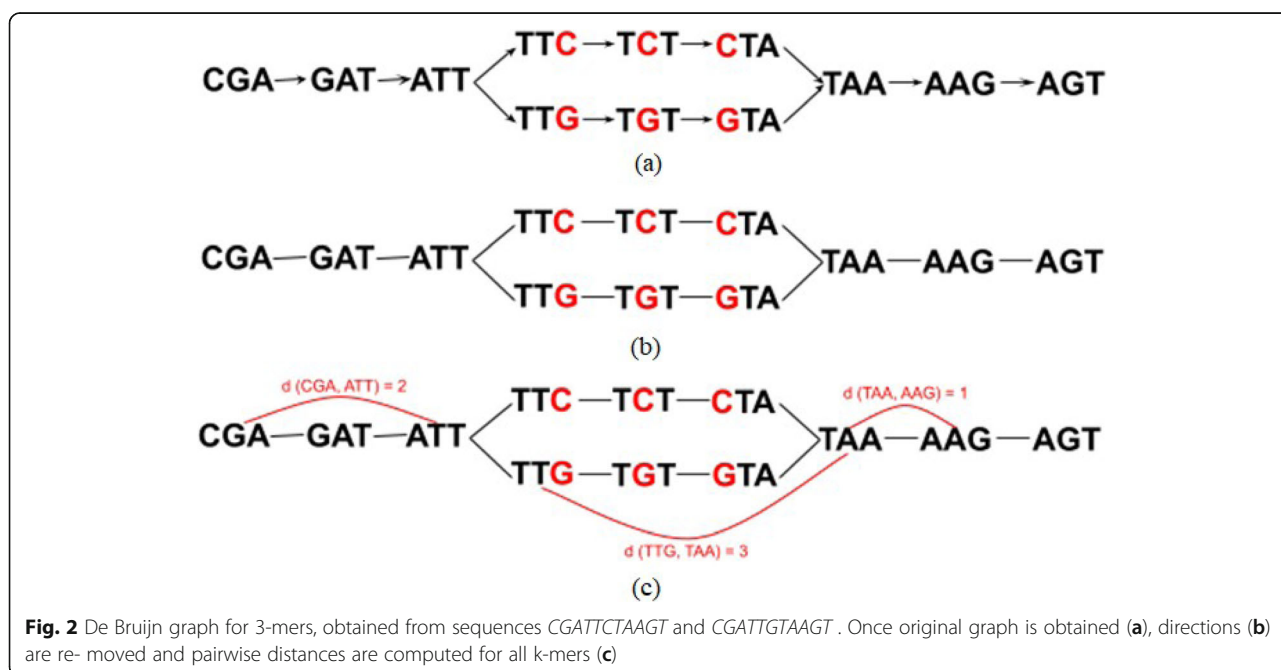
k-EMD distance computation
Input: Sets of sequencing reads from hosts A and B (SA and SB).
Output: k-EMD distance between A and B.
1 Produce k-mers from SA and SB:

    KMA ← k-mer distribution from SA
    KMB ← k-mer distribution from SB

2 Initialize distance matrix D(A, B): for any pair of k-mers x ∈ KMA and
  y ∈ KMB, find dist(A, B) in De Bruijn graph;
3 Compute EMD(KMA, KMB, D(A, B)).
    
```

Example of EMD computation

Constructing of the De Bruijn graph between two sequences *CGATTCTAAGT* and *CGATTGTAAGT* is shown on Fig. 2. Once original graph is obtained, directions are removed and pairwise distances are computed for all k-mers. Figure 3 describes an example of k-EMD distance computation. After k-mer distributions are generated for input sequences, EMD is computed as the work, where f_{ij} is the flow between histogram(k-mer distribution) elements *i* and *j* and d_{ij} is the corresponding distance between k-mers, which is obtained from De Bruijn graph (Fig. 2). This way, $EMD = 0.88$.



Mean k-mer distribution

Representing samples as k-mer distributions allows to estimate the center from a group of samples by introducing a mean host. We use the **maximum mean** k-mer distribution, which is obtained by finding the maximum observed frequency for each.

$$k\text{-mer } k_i f_i^{max} = \max_{1 \leq i \leq n} f_i \text{ and normalization } f'_i = \frac{f_i^{max}}{\sum_{1 \leq i \leq n} f_i^{max}}$$

Identification of relatedness

We train our algorithm on all given outbreaks and obtain minimal EMD between 2 unrelated hosts, which we use as a threshold t . To identify whether 2 hosts A and B are related, we compute EMD between them $EMD(A, B)$ and predict that they are related if $EMD(A, B) < t$, and unrelated otherwise.

Identification of transmission direction between hosts

To infer transmission direction between a pair of samples X and Y , we first compute a mean host $Mean(A, B)$.

Once $Mean(A, B)$ is obtained, we calculate EMD between mean host and hosts A and B $EMD(Mean(A, B), A)$ and $EMD(Mean(A, B), B)$. Host, that is closer to the maximum mean is assumed to be the transmission source, so that if $EMD(Mean(A, B), A) < EMD(Mean(A, B), B)$, we predict that the transmission happened from A to B (Fig. 4).

Identification of transmission clusters

To test hierarchical clustering, *single-linkage* algorithm was used. This method evaluates the similarity of two clusters based on their most similar members [14] and groups clusters in bottom-up order until certain termination condition is satisfied. In our algorithm, we use a distance criteria, so clusters are merged until distance between them exceeds a predefined distance threshold, which represents EMD between two closest unrelated samples in the dataset. This way, we obtain a partition, where some of the related hosts remain in different clusters. At this point, we proceed to the second stage of the algorithm, that allows to improve the clustering quality by merging the clusters, that contain related hosts by performing the following steps:

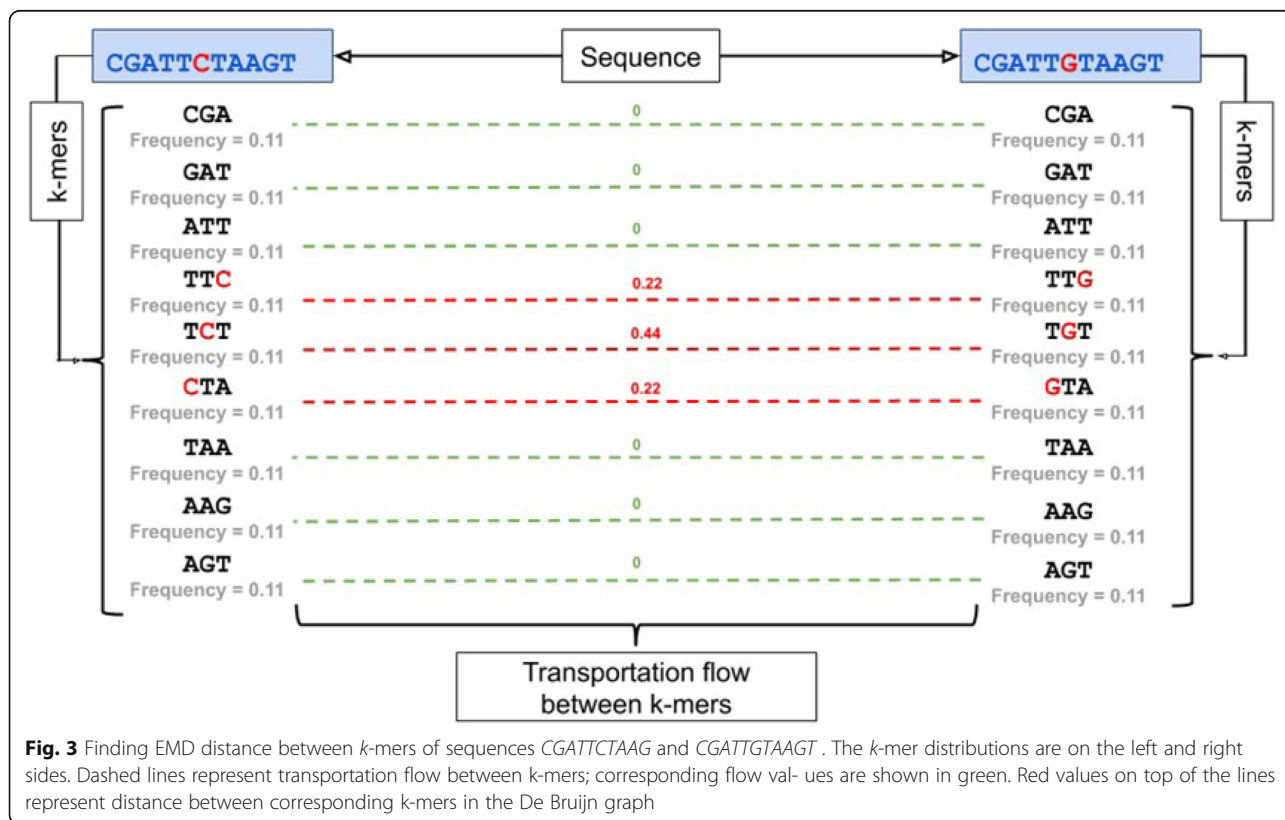
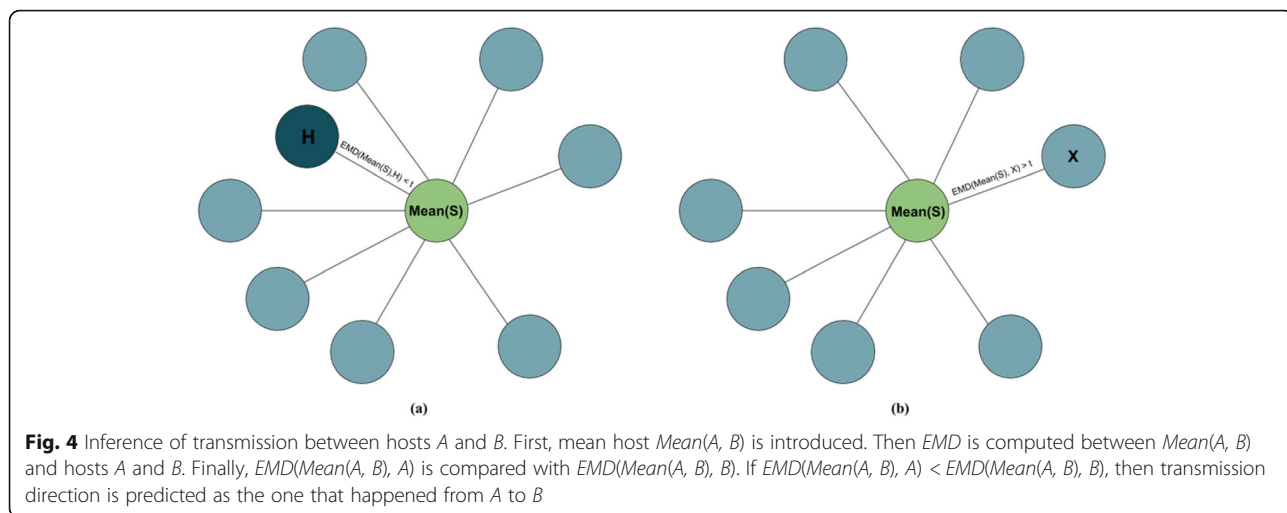


Fig. 3 Finding EMD distance between k-mers of sequences CGATTCTAAG and CGATTGTAAGT. The k-mer distributions are on the left and right sides. Dashed lines represent transportation flow between k-mers; corresponding flow values are shown in green. Red values on top of the lines represent distance between corresponding k-mers in the De Bruijn graph



1. For each cluster, obtained from hierarchical clustering, compute center as the mean of all hosts within the cluster;
2. For each center, obtained at the previous step:
 - Find distances to the furthest in-cluster host and closest host, that belongs to the different cluster;
 - If for cluster *A* there exists an ‘overlap’ (there is a host from cluster *B*, that is closer to the center than the furthest host, belonging to the same cluster (*A*)), merge *A* and *B*

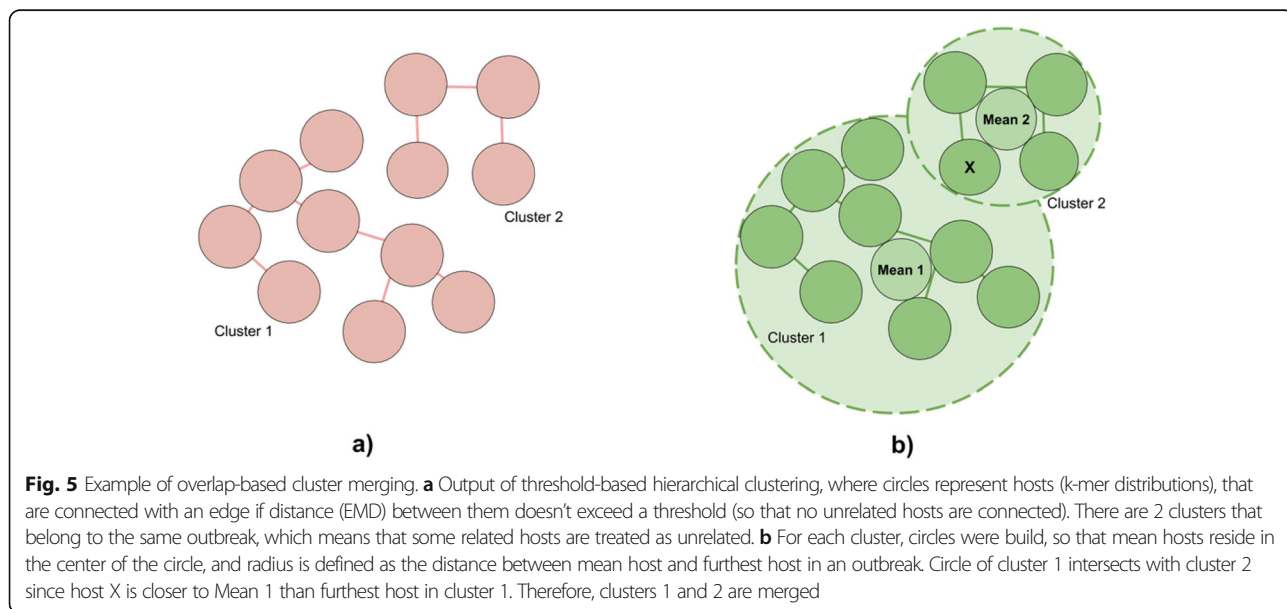
Example of the algorithm is demonstrated in Fig. 5. a) shows output of threshold-based hierarchical clustering, where circles represent hosts, that are connected with an edge if distance between them doesn’t exceed a

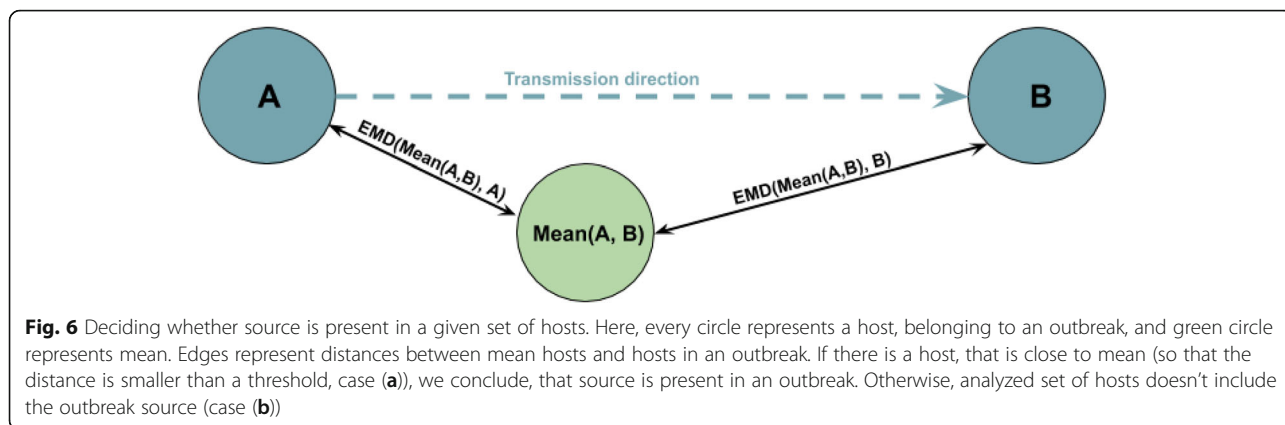
threshold. There are 2 clusters that belong to the same outbreak. b) shows how clusters are merged based on circle overlap. For each cluster, mean host of all hosts within the cluster is calculated (shown in the center). Circles with dashed borders have centers in respective mean hosts; their radiuses are calculated as distances between mean hosts and furthest in-cluster hosts. In the example, Mean 1 is closer to host *A* that to the furthest host from the same (left) cluster. This way, according to our algorithm, intersecting clusters collapse.

Deciding whether source is present in a set of hosts

To decide whether source is present in a set of sequenced hosts *S*, the following algorithm is applied:

1. Calculate mean $Mean(S)$ for all hosts within an outbreak;





- For every host H , calculate EMD between H and mean $Mean$ $EMD(Mean(S), H)$;
- If there exists a host, for which $EMD(Mean(S), H) < t$, source is present. To obtain threshold t , we train the algorithm on all outbreaks with known sources. For every such outbreak, we first calculate the mean host $Mean(S)$ and distances between mean and every host H in the outbreak $EMD(Mean(S), H)$, find the smallest distance and normalize it by the median distance from mean to host in an outbreak. After this, we repeat the procedure for the same outbreak, but discard the source. We define t as the minimal $EMD(Mean(S), H)$ for an outbreak without source, which maximizes accuracy, so that outbreaks, where source is present, have $EMD(Mean(S), H) < t$.

Source identification

To identify sources, we find a maximum mean host for an outbreak $Mean$ and calculate EMD between every host and $Mean$. Host with minimum $EMD(H, Mean)$ is assumed to be the source.

Runtime complexity

The algorithm uses Pele and Werman’s [15, 16] algorithm for fast EMD computation, which has a runtime complexity of $O(N^2 U \log N)$, where N is the number of nodes (k-mers), and U is an upper bound on the largest supply (flow) of any node (since frequencies are normalized, this is equal to 1). This way, k -mer EMD has a worst time complexity of $O(N^2 \log N)$.

Results

We validated our new algorithm on a publicly available dataset obtained from an epidemiological study of HCV outbreaks [11] Fig. 6.

Data sets

The data consists of 368 sequenced hosts where 175 of them belong to 34 annotated outbreaks. Among these annotated outbreaks, 11 have a known main spreader (Table 1). All outbreaks contain from 2 to 33 hosts. Every host is represented as an HCV intra-host population, obtained with end-point limiting-dilution (EPLD). All viral sequences represent a fragment of E1/E2 genomic region of length 264 bp. Data samples annotation consists of host and outbreak id along with abundance for every sequence. This way, we were able to interpret obtained experimental results.

We simulated MiSeq reads from known haplotypes by SimSeq [17] and created mixtures using abundances from original data.

Validation

Identification of relatedness

Viral populations from two samples are genetically related if they belong to the same outbreak and unrelated, otherwise. The genetic relatedness is validated on the union of both collections containing all outbreaks and unrelated samples. There are 67,528 host pairs (obtained from all 368 hosts). Among these pairs, 1007 represent related cases (so that both hosts in pair belong to the same annotated outbreak). We used EMD as predictor for relatedness. We measured the sensitivity of our method as following. First we determining the EMD value for all unrelated pairs, the minimum value we have chosen as a threshold which prohibits false-positive relatedness detection, the pairs which have EMD below the threshold are considered as related. Precision of our algorithm is 100%. We calculated the recall as a proportion of correctly predicted related pairs among all

Table 1 Outbreaks with known sources

Outbreak	AA	AC	AI	AJ	AQ	AW	BA	BB	BC	BJ	NH
# samples	3	4	15	3	9	19	6	7	2	4	33

Table 2 Validation results. k-EMD was tested on a dataset, that includes 34 out- breaks; MinDist, ReD and VOICE were validated earlier on a smaller dataset, that didn't include one of the outbreaks. For convenience, results for k-EMD contain 2 values - one for the smaller dataset, and one for the entire (34 outbreaks) set of hosts (values in parentheses)

Method	k-EMD	MinDist	MinDistB	ReD	VOICE-D	VOICE-S
Relatedness sensitivity, %	80.4 (90)	90	92.9	55.3	85.2	86.8
Clustering sensitivity, %	100 (100)	100	100	96.3	98.2	98.2
Direction accuracy, %	88.7 (90.4)	N/A	N/A	87.1	83.9	87.1
Source accuracy, %	80 (81.8)	50	40	90	80	90

known related pairs. Results are described in Table 2. Relatedness ROC is shown on Fig. 7.

Identification of transmission direction between hosts

Performance of algorithm when identifying transmission direction was calculated as a ratio of pairs of hosts with correctly predicted directions to all host pairs, where direction is known. Results are shown in Table 2.

Identification of transmission clusters

Precision for our algorithm is equal to 100%, since we don't merge hosts from different outbreaks. Similarities between true and estimated partitions were evaluated using an editing metric [18]. Given metric is defined as the minimum number of elementary operations, required to transform one partition into another, such as joining or partition of clusters [18]. Clustering recall was calculated similarly to [12], so that editing distance E was normalized by dividing it by the number of elementary operations N , required to transform trivial partition into singleton sets into true partition, which is equal to $n - k$, where n is the number of samples and k is the number of true clusters [12]:

$$Recall = \frac{E}{n - k} \times 100\%$$

Deciding whether outbreak source is present

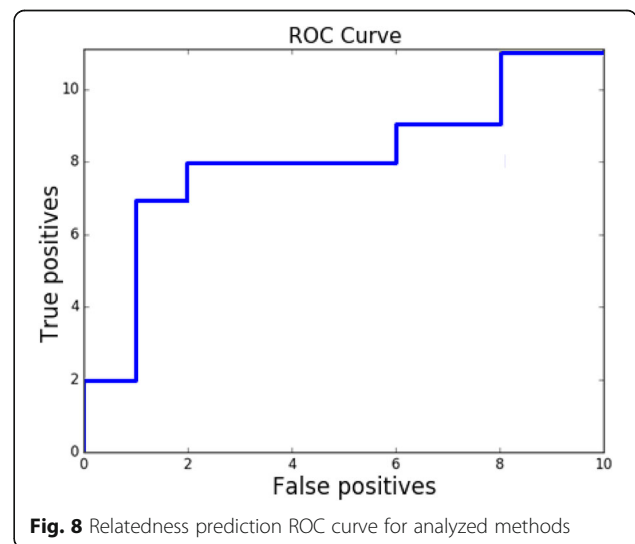
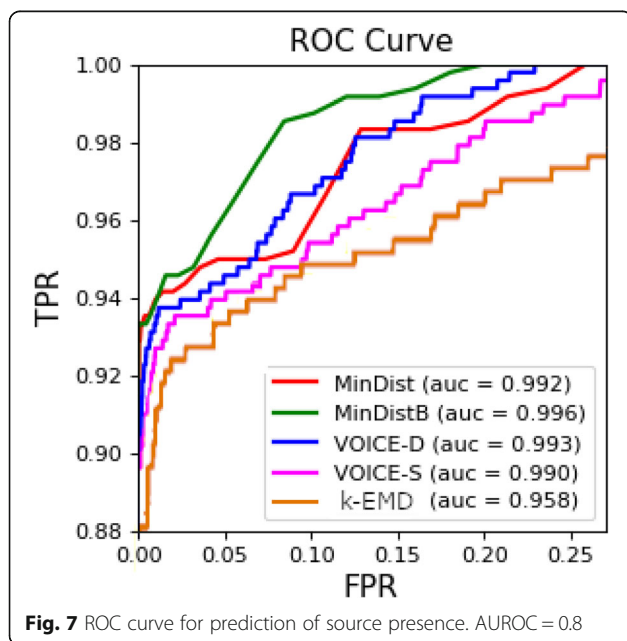
Source presence recall was calculated as the proportion of outbreaks with present source, that were correctly identified as such; precision - as the proportion of correctly identified outbreaks, where source is not present. Finally, specificity was calculated as the total number of outbreaks with present source, divided by the sum of total number of outbreaks with present source and the number of outbreaks, that were incorrectly identified to have a source present. For our algorithm, precision = 90%, specificity = 80%, and recall = 85%. ROC curve for source presence detection is shown on Fig. 8.

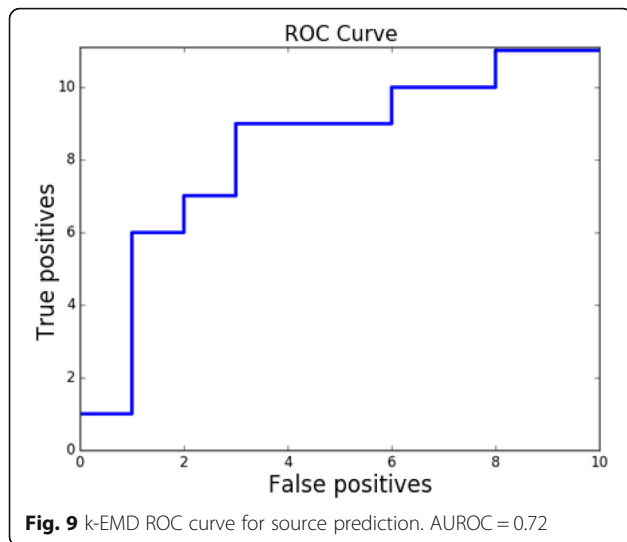
Identification of outbreak sources

Source identification accuracy is calculated as the percentage of outbreaks with correctly predicted sources for outbreaks with known sources. ROC curve for source presence detection is shown on Fig. 9.

Conclusions

Extracting haplotypes by EPLD is laborious and costly procedure and that prohibits previously developed





methods [12] from wide spread. On the other hand, viral samples can be easily sequenced by NGS, and that makes our novel method attractive. Furthermore, we can see that results in this article are comparable with those which were obtained using EPLD technology [12]. Moreover, our method allowed to decide whether the spreader get sequenced.

Application of molecular viral analysis to investigation of outbreaks and inference of transmission networks is a promising technique, that is available nowadays. However, it generates novel computational challenges. Given work introduced an algorithm for investigation of viral transmissions, that is based on analysis of the intra-host viral populations through k-mer decomposition. Proposed approach allows to cluster genetically related samples, infer transmission directions and predict sources of outbreaks. Validation on experimental data demonstrated that algorithm is able to reconstruct various transmission characteristics. It should be noted that even though there is still room for improvement when it comes to algorithm performance, advantage of the method is the ability to bypass cumbersome read assembly, thus eliminating the chance to introduce new errors, and saving processing time by allowing to use raw NGS reads.

Abbreviations

NGS: Next-generation sequencing; EMD: Earth mover's distance; HCV: Hepatitis C virus; RNA: Ribonucleic acid; EPLD: End-point limiting-dilution; ROC: Receiver operating characteristic

Acknowledgements

Authors would like to thank Georgia State University MBD fellowship for support.

About this supplement

This article has been published as part of BMC Genomics Volume 21 Supplement 5, 2020: Selected articles from the 15th International Symposium

on Bioinformatics Research and Applications (ISBRA-19): genomics. The full contents of the supplement are available online at <https://bmcbgenomics.biomedcentral.com/articles/supplements/volume-21-supplement-5>

Authors' contributions

AM designed, implemented and tested the algorithms; SK implemented and tested the algorithms; AZ, PS, FV and LB designed and implemented the algorithms, analyzed the results and supervised the research. All authors read and approved the final manuscript.

Funding

Publication costs are funded by Georgia State University Molecular Basis of Disease fellowship, NSF grants DBI-1564899 and CCF-1619110, and NIH grants 1R01EB025022-01 and 1R01EB025022-01.

Availability of data and materials

k-mer EMD is freely available at <https://github.com/amelnyk34/kemd>.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Computer Science Department, Georgia State University, 25 Park Place NE, Atlanta, GA 30303, USA. ²Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332, USA. ³I.M. Sechenov First Moscow State Medical University, Moscow 119991, Russia.

Received: 19 March 2020 Accepted: 12 June 2020

Published: 16 December 2020

References

- Drake JW, Holland JJ. Mutation rates among rna viruses. *Proc Natl Acad Sci*. 1999;96(24):13910-3.
- Eriksson N, Pachter L, Mitsuya Y, Rhee S-Y, Wang C, Gharizadeh B, Ronaghi M, Shafer RW, Beerwinkler N. Viral population estimation using pyrosequencing. *PLoS Comput Biol*. 2008;4(5):1000074.
- Archer J, Braverman MS, Taillon BE, Desany B, James I, Harrigan PR, Lewis M, Robertson DL. Detection of low-frequency pretherapy chemokine (cxcr4 motif) receptor 4-using hiv-1 with ultra-deep pyrosequencing. *AIDS (London, England)*. 2009;23(10):1209.
- Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, Bushman FD. Dna bar coding and pyrosequencing to identify rare hiv drug resistance mutations. *Nucleic Acids Res*. 2007;35(13):91.
- Wang W, Zhang X, Xu Y, Weinstock GM, Di Bisceglie AM, Fan X. High-resolution quantification of hepatitis c virus genome-wide mutation load and its correlation with the outcome of peginterferon-alpha2a and ribavirin combination therapy. *PLoS One*. 2014;9(6):100131.
- Skums P, Campo DS, Dimitrova Z, Vaughan G, Lau DT, Khudyakov Y. Numerical detection, measuring and analysis of differential interferon resistance for individual hcv intra-host variants and its influence on the therapy response. *In silico biology*. 2011;11(5):263-9.
- Campo DS, Skums P, Dimitrova Z, Vaughan G, Forbi JC, Teo C-G, Khudyakov Y, Lau DT. Drug resistance of a viral population and its individual intrahost variants during the first 48 hours of therapy. *Clin Pharmacol Therapeutics*. 2014;95(6):627-35.
- RK KW, Ravi MK. Miseq: A next generation sequencing platform for genomic analysis; 2018. p. 223-32.
- Mangul S, Koslicki D. Reference-free comparison of microbial communities via de bruijn graphs. In: proceedings of the 7th ACM international conference on bioinformatics, computational biology, and health informatics; 2016. p. 68-77. ACM.
- Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic rna-seq quantification. *Nat Biotechnol*. 2016;34(5):525.
- Campo DS, Xia G-L, Dimitrova Z, Lin Y, Forbi JC, Ganova-Raeva L, Punkova L, Ramachandran S, Thai H, Skums P, et al. Accurate genetic detection of

hepatitis c virus transmissions in outbreak settings. *J Infect Dis.* 2016;213(6):957–65.

12. Glebova O, Knyazev S, Melnyk A, Artyomenko A, Khudyakov Y, Zelikovsky A, Skums P. Inference of genetic relatedness between viral quasispecies from sequencing data. *BMC Genomics.* 2017;18:918. <https://doi.org/10.1186/s12864-017-4274-5>.
13. Rubner, Y., Tomasi, C., Guibas, L.J.: A metric for distributions with applications to image databases. 1998 IEEE International Conference on Computer Vision (1998).
14. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, pp. 382–385.
15. Pele O, Werman M. Fast and robust earth mover's distances. In: 2009 IEEE 12th international conference on computer vision; 2009. p. 460–7. IEEE.
16. Pele, O., Werman, M.: A linear time histogram metric for improved sift matching. In: *Computer Vision–ECCV 2008*, pp. 495–508. Springer, (2008)..
17. Benidt S, Nettleton D. Simseq: A nonparametric approach to simulation of rna-sequence datasets. *Bioinformatics.* 2015;31:9.
18. Deza MM, Deza E. *Encyclopedia of distances*: Springer; 2009.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

