The Author(s) *BMC Genomics* 2017, **18**(Suppl 4):372
DOI 10.1186/s12864-017-3732-4

BMC Genomics

RESEARCH Open Access

Efficient detection of viral transmissions with Next-Generation Sequencing data

Inna Rytsareva[1†], David S. Campo[1*†], Yueli Zheng[1], Seth Sims[1], Sharma V. Thankachan[2,3], Cansu Tetik[2], Jain Chirag[2], Sriram P. Chockalingam[4], Amanda Sue[1], Srinivas Aluru[2,4] and Yury Khudyakov[1]

From Fifth IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS 2015)
Miami, FL, USA. 15-17 October 2015

Abstract

Background: Hepatitis C is a major public health problem in the United States and worldwide. Outbreaks of hepatitis C virus (HCV) infections associated with unsafe injection practices, drug diversion, and other exposures to blood are difficult to detect and investigate. Molecular analysis has been frequently used in the study of HCV outbreaks and transmission chains; helping identify a cluster of sequences as linked by transmission if their genetic distances are below a previously defined threshold. However, HCV exists as a population of numerous variants in each infected individual and it has been observed that minority variants in the source are often the ones responsible for transmission, a situation that precludes the use of a single sequence per individual because many such transmissions would be missed.
The use of Next-Generation Sequencing immensely increases the sensitivity of transmission detection but brings a considerable computational challenge because all sequences need to be compared among all pairs of samples.

Methods: We developed a three-step strategy that filters pairs of samples according to different criteria: (i) a k-mer bloom filter, (ii) a Levenhstein filter and (iii) a filter of identical sequences. We applied these three filters on a set of samples that cover the spectrum of genetic relationships among HCV cases, from being part of the same transmission cluster, to belonging to different subtypes.

Results: Our three-step filtering strategy rapidly removes 85.1% of all the pairwise sample comparisons and 91.0% of all pairwise sequence comparisons, accurately establishing which pairs of HCV samples are below the relatedness threshold.

Conclusions: We present a fast and efficient three-step filtering strategy that removes most sequence comparisons and accurately establishes transmission links of any threshold-based method. This highly efficient workflow will allow a faster response and molecular detection capacity, improving the rate of detection of viral transmissions with molecular data.

Background

Hepatitis C virus (HCV) infects nearly 2.8% of the world's population and is a major cause of liver disease worldwide [1]. HCV infection is an important US public health problem, because it is the most common chronic blood-borne infection and the leading cause for liver transplantation [2]. Since 2007, HCV surpasses HIV as a cause of death in the US [3]. It is estimated that 2.7 million to 3.9 million people in the United States have chronic HCV infection and that more than 15,000 die each year from HCV-related disease, with mortality expected to rise in the coming years [4]. Approximately 80% of patients who become infected with HCV develop chronic Hepatitis and are at risk for advanced liver disease; 15–30% of these patients have progression to liver fibrosis and cirrhosis and up to 5% will die from liver failure due to cirrhosis or hepatocellular carcinoma [2].

* Correspondence: fyv6@cdc.gov
†Equal contributors
[1]Molecular Epidemiology and Bioinformatics, Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, GA, USA
Full list of author information is available at the end of the article

Outbreaks of hepatitis C virus (HCV) infections are associated with unsafe injection practices, drug diversion, and other exposures to blood and blood products. Given the long incubation period (up to 6 months) and that HCV infections can remain asymptomatic in >70% of infected persons for years, the detection and investigation of Hepatitis C outbreaks is a challenging task.

Molecular phylogenetic analyses of RNA viruses have been used frequently in the study of outbreaks and transmission chains [5–9], usually by analysing a single sequence per infected individual and comparing these sequences to ascertain if their genetic distances are below a previously defined threshold. However, HCV exists as a population of numerous variants in each infected individual and it has been observed that minority variants in the source are often the ones responsible for transmission, a situation that precludes the use of a single sequence per individual because many such transmissions would be missed [10]. Our laboratory has been using molecular analysis of Viral Hepatitis populations (rather than single sequence) for more than a decade [11–14] with a simple and accurate threshold-based approach for detecting HCV transmissions that streamlines molecular investigation of outbreaks, thus improving the public health capacity for rapid and effective control of hepatitis C [10].

Now with the advent of Next-Generation Sequencing (NGS) we expect an increase in the sensitivity of transmission detection due to the sampling of minority variants [10] but this advantage comes with a considerable computational challenge because all sequences need to be compared among all pairs of samples. For instance, for our relatively small dataset of 401 samples, a total of 80200 pairwise sample comparisons are performed, which account for 4.56×10^{10} pairwise sequence comparisons. Given that the molecular surveillance of HCV is just starting, this number will certainly grow in the near future and the detection of transmission will soon become impractical. We present an efficient three-step filtering strategy that removes 85.1% of all the pairwise sample comparisons and 91.0% of all pairwise sequence comparisons, accurately establishing which pairs of HCV samples are below the relatedness threshold.

Methods

Problem definition

Given $P = \{P_1, P_2, \ldots\}$, a set of samples where each P_i is associated with a set $S_i = \{S_i^1, S_i^2, \ldots\}$ of homologous sequences, enumerate all sample pairs (P_i, P_j) where any pairwise sequence comparisons (S_i^x, S_j^y) has an edit distance lower than the relatedness threshold, T (see Fig. 1). Given that every sample-pair needs to be considered, it yields an $O(n^2)$ algorithm, where n is the number of samples.

However, we have observed than less than 1% of all sample-pairs are linked by transmission in a typical study (see Fig. 2). Therefore, an exhaustive search over all pairs of sequences is very inefficient due to the fact that the great majority of sample pairs are above T. Briefly, it would be very advantageous to remove most of these pairs in order to reduce the number of computations needed to establish transmission on a set of samples.

Datasets

Sample description

We analyzed two set of files that cover the spectrum of genetic relationships among HCV cases. The "Unrelated dataset" is comprised of 401 HCV cases that are epidemiologically unrelated to each other and were obtained from national collections and other research projects [15, 16]. The "Related dataset" is comprised of 18 HCV

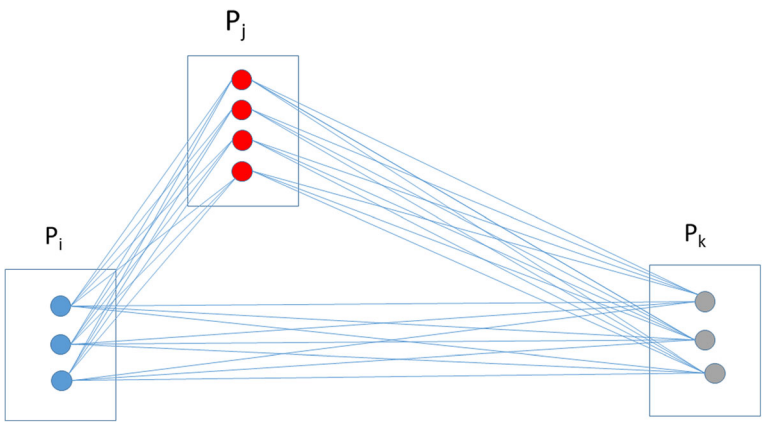

Fig. 1 Transmission detection overview. In this example, there are 3 samples: Pi contains 3 different sequences, Pj contains 4 and Pk contains 3. In addition, Pi and Pj are related, whereas Pk is unrelated to the other two. A total of 33 pairwise sequence comparisons must be performed to find the minimal distance between each pair of samples. The rationale of our approach is to quickly remove the sample-pair comparisons with zero probability of having a minimal distance lower than T

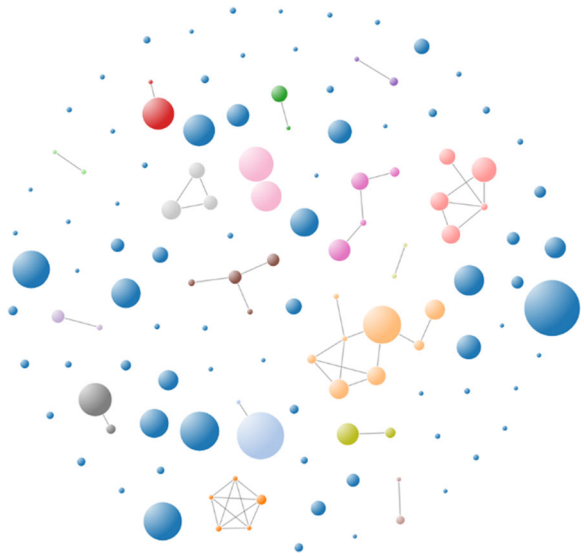

Fig. 2 Transmission network density This is an example of a real HCV transmission network obtained during an outbreak study. A link is drawn if the minimal edit distance between the two samples is smaller than T, whereas the size of the node is proportional to its genetic heterogeneity. In this particular example, only 0.8% of all sample-pairs are linked by transmission

cases from an outbreak where a surgical technician diverted drugs and infected patients at a health-care setting [17]. All samples in the related set are epidemiologically linked and their minimal edit distance is smaller than T (3.77%). The average number of different sequences per sample is 534.3

Experimental methods

For each sample, we used the sequences obtained and described in [10, 16]. Briefly, we amplified the E1/E2 junction of the HCV genome (306pb), which contains the Hyper Variable Region 1 region) using our nested PCR protocol as previously described [18]. PCR products were pooled and subjected to pyrosequencing using GS FLX Titanium Sequencing Kit (454 Life Sciences, Roche, Branford, CT). Low-quality reads were removed using the GS Run Processor v2.3 (Roche) and then processed by matching to the corresponding identifier. The NGS files were processed using the error correction algorithms KEC and ET [19].

Algorithms

We developed a three-step strategy that filters pairs of samples according to different criteria. Figure 3 shows an overview of the filtering strategy.

K-mer bloom filter

For a sequence pair (S_i^x, S_j^y) to be similar enough to link two samples, the following condition must be satisfied: the edit distance between S_i^x and S_j^y is < LT (Length x T). This means that when we align S_i^x and S_j^y, the lower bound of the maximal common substring is k = (L − LT)/(LT + 1), which for our particular T would be 26. We took advantage of this maximal common substring requirement and created for each sample a Bloom filter of all its 26-mers. A bloom filter is a space-efficient probabilistic data structure supporting dynamic set membership queries with false positives [20]. False positive matches are possible, but false negatives are not, thus a Bloom filter has a 100% recall rate [20]. For any pair of samples, we tested the intersection of k-mer sets: If it is empty, the sample pair is considered unrelated and removed from the sample-pair candidate list; if it is non-empty, the sample pair may be related and remains in the sample-pair candidate list.

Hamming radius filter

For each sample Pi in the database, we calculated and stored the following: (i) its Multiple Sequence Alignment (MSA); (ii) its Consensus, Ci, defined as the majority nucleotide state at each position in the alignment; and (iii) its Hamming radius, Ri, defined as the maximum edit distance found between the consensus and all other variants of the sample.

For any pair of samples we calculated a sample distance, Sd, defined as: Sd = dist(Ci, Cj) − (Ri + Rj). Each sample-pair is tested in this fashion and if Sd is higher than LT, it is removed from the sample-pair candidate list because these two samples cannot have any sequence-pair with a distance lower than T (see Fig. 4). If Sd is lower than the threshold, the sample pair may be related and remains in the sample-pair candidate list.

Identical sequences filter

We have previously estimated that 51.63% of sample-pairs coming from the same transmission cluster share at least one identical sequence [10]. Candidate pairs that share one or more sequences do not be need to be fully evaluated because their minimal distance is zero and therefore are ensured to be below the T. We take advantage of this fact to create a simple filter that quickly identifies those sample-pairs sharing identical sequences. In order to achieve this, we calculate for each sequence in a sample its hash "fingerprint" with a standard cryptographic function (MD5). The set of such strings is constructed for each sample file only once and then stored. When comparing sample-pairs, we check for intersection in their hash sets and if the size of the intersection is at least 1, then the sample-pair is considered related. If it is not, the sample pair remains in the sample-pair candidate list.

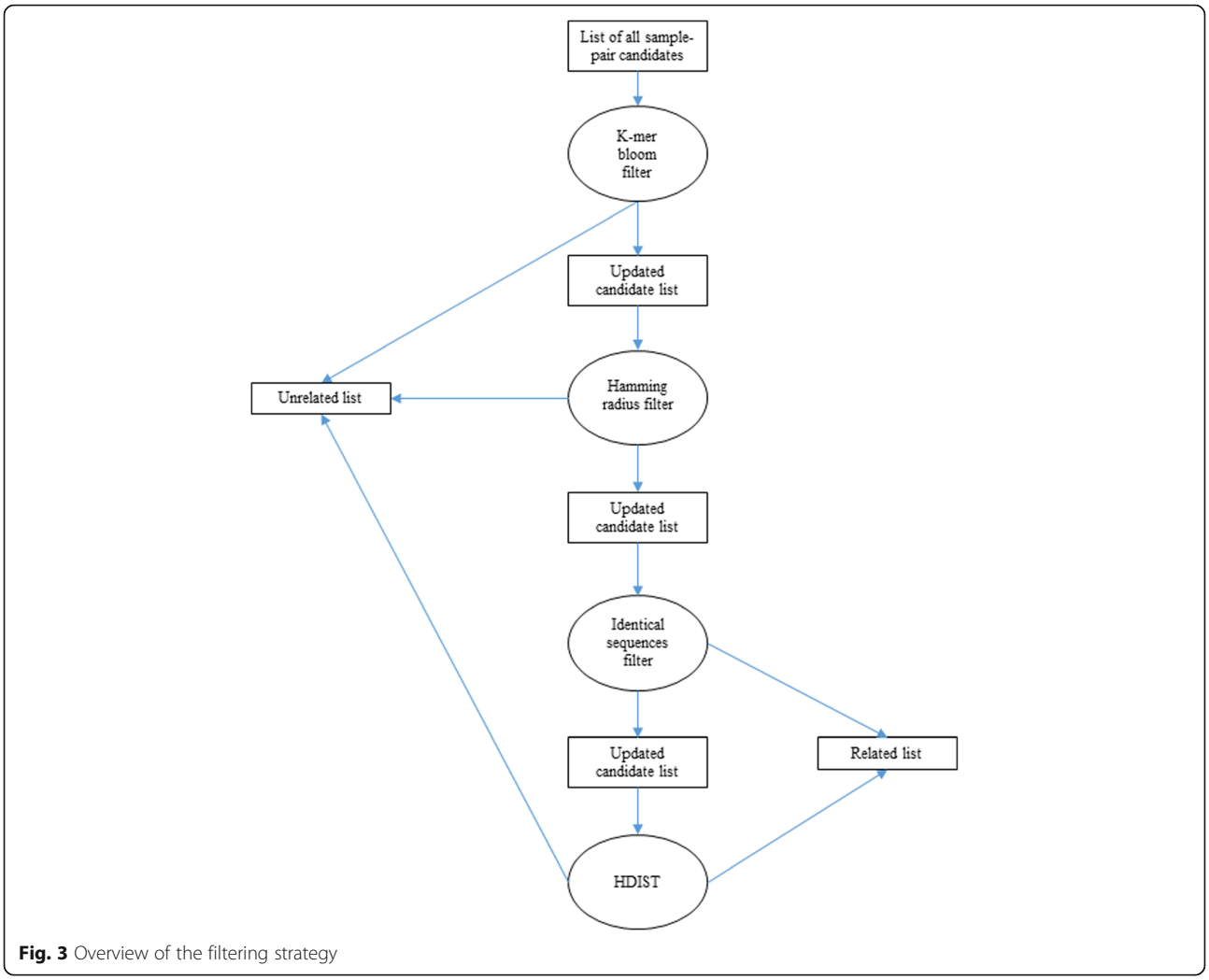

Fig. 3 Overview of the filtering strategy

Detection of transmission

For each sample-pair remaining in the candidate list, all its sequences are used to create a MSA, which is then used to calculate the edit distance between every pair of sequences. The two samples are considered related if the minimal edit distance between any of their sequences is smaller than T.

All edit distances were calculated with HDIST, a custom-made, highly optimized distance calculator that minimizes processor pipeline stalls and takes advantage of modern superscalar architecture. The procedure involves breaking a sequence pair into consecutive 3-mers, converting them into base 5 integers and using them as indices into a pre-calculated look-

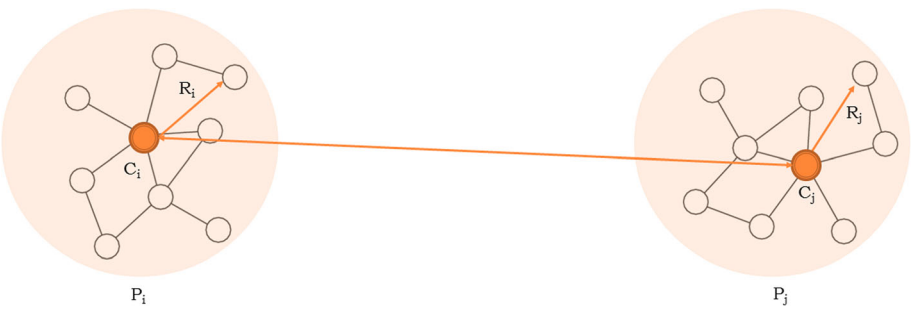

Fig. 4 Hamming radius filter. If Sd is higher than LT these two samples cannot have any sequence-pair with a distance lower than T

up table. The choice of 3-mers was made by testing different word sizes to maximize processor cache memory hits.

Results

Filtering strategy

We developed a three-step strategy that filters pairs of samples according to different criteria. The rationale of the approach is that the great majority of sample pairs are very different (unrelated) and it would be advantageous to remove these pairs in order to reduce the amount of computation needed to establish transmission on a set of samples. Every sample-pair is still considered, yielding an $O(n^2)$ algorithm, where n is the number of samples. However, the 3-step filtering strategy efficiently prunes most comparisons from the candidate list with much lower computational effort than the full distance calculation, even though both have the same order.

Filtering performance

For the Unrelated dataset, the whole algorithm can be performed under 5 min on a desktop computer, accurately removing 85.1% of all possible candidates and 91.0% of all pairwise sequence comparisons. The number of sample-pair candidates that are removed by each filter can be seen in Table 1. On this dataset, the best individual filter is the Hamming radius filter, which removes 84.7 of all sample-pairs. Only 302 candidates are removed by the k-mer bloom filter that are not removed by the Hamming radius filter, whereas 15404 candidates are removed by the Hamming radius that are not removed by the k-mer Bloom filter. With regard to the overlap, 52234 candidates are removed by both filters.

We studied the behavior of the bloom filter with different k-mer values. Figure 5 shows how the percentage of removed sample-pairs increases with the value of k. With our particular T value, the 26-mer bloom filter removes 65.5% of all sample-pairs are removed. As the k value increases, the percentage of removed sample-pairs increase very quickly. For instance, a common relatedness threshold used in HIV molecular epidemiology is 1%, which on this dataset yields a k-mer of 72 that filters 88.6% of all sample-pairs.

For the Related dataset, the whole algorithm can be performed under 10 s on a desktop computer, accurately identifying 51.6% of all possible candidates and removing them from the workflow (see Table 2). On this dataset, both the k-mer bloom and the Hamming radius filter do not remove any candidates, as is expected given that all of them are below T.

Table 1 Filtering results on the unrelated dataset

Filter	Individually	Serial workflow
k-mer bloom filter	52536 (65.5%)	52536 (65.5%)
Hamming radius filter	67940 (84.7%)	68242 (85.1%)
Identical sequences filter	0 (0.0%)	68242 (85.1%)

Number of candidate pairs removed by each filtering approach

Implementation

The k-mer bloom filter was implemented in JAVA, whereas the Hamming radius filter, the identical sequence filter and HDIST were implemented in Python and Cython. Although all the programs are available upon request, they are part of our recently developed web system for the advanced molecular detection of HCV transmissions, the Global Hepatitis Outbreak and Surveillance technology (GHOST, which will be described elsewhere). The web system includes the analytical methods described in this article, improving the accuracy and response time of transmission detection by integrating epidemiological evidence, NGS and data analysis. The tool is available to public health laboratories to identify outbreaks by simply uploading viral sequences.

Discussion

The utility of the "Identical sequences filter is only evident when there are samples coming from the same geographical region or from a suspected outbreak, as we have previously estimated that 51.63% of sample-pairs coming from the same transmission cluster share at least one identical sequence [10].

The Hamming radius filter seems to be outperforming the k-mer bloom filter on this dataset. However, the Hamming radius filter requires a pre-calculation step for each sample, which involves a MSA. This MSA can be performed efficiently with MAFFT but it has high memory requirements depending on the number of sequences. Therefore, the Hamming radius filter is contingent on the feasibility of the MSA, whereas the k-mer bloom filter is alignment free. This particular NGS dataset was obtained with 454 Life Sciences, where the average number of different sequences per sample is 534.3. Our initial tests on the Illumina MiSeq platform indicate that although the number of different sequences is around 15 times greater, the MSA step is still feasible.

The idea behind the Hamming radius to exclude sample-pairs could be generalized to exclude sequences within a patient that are too distant from the sequences of the other sample. We are currently using just the maximum distance to the consensus (radius), but all those distances could be used to filter a great amount of sequences that are very close to the consensus. A reduced number of sequences will decrease the number

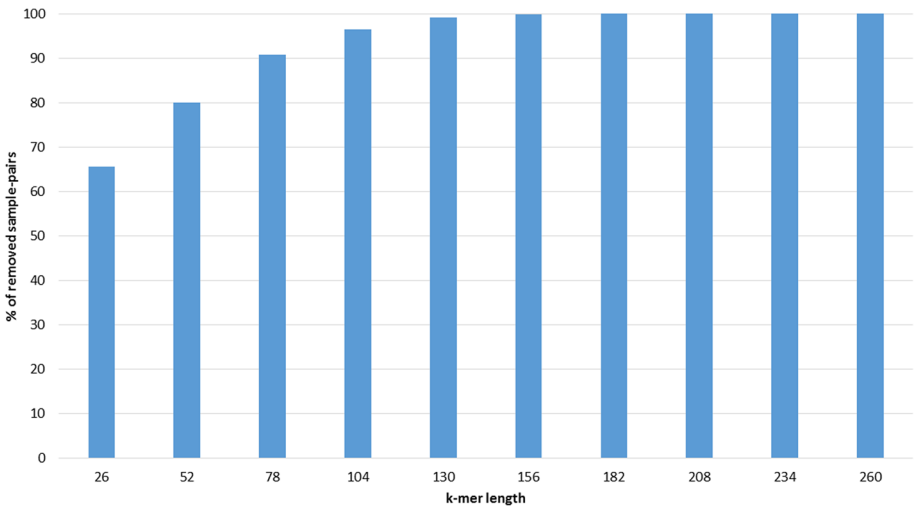

Fig. 5 Percentage of removed sample-pairs by the k-mer bloom filter

of pairwise comparisons that are calculated at the last HDIST step.

Until recently, molecular phylogenetic analyses of RNA viruses used a single viral sequence per patient to detect transmission. However, the advent of NGS data immensely increases the computational burden of this simple approach. Our proposed filtering strategy can be used for detecting transmissions of any heterogeneous virus where a threshold-based method has been validated.

The number of samples in our database is now in the order of 10^2, but it is constantly increasing with time as HCV molecular surveillance becomes more commonplace with the aid of cheaper and more effective NGS technologies. Just in the United States, it is estimated that 2.7 million to 3.9 million people have chronic HCV infection [4] and if we want to respond to a rapidly growing database of NGS data, there is a great need for our highly efficient workflow to accurately infer the network of HCV transmissions. The availability of this system for the detection of HCV transmissions will foster deeper involvement of public health researchers and practitioners in HCV outbreak investigation in the United States and worldwide. Improvement in molecular detection capacity also will increase the rate of detection of transmissions in the United States, thus providing opportunity for a rapid and effective response to the growing number of Hepatitis C outbreaks.

Table 2 Filtering results on the related dataset

Filter	Individually	Serial workflow
k-mer bloom filter	0 (0.0%)	0 (0.0%)
Hamming radius filter	0 (0.0%)	0 (0.0%)
Identical sequences filter	79 (51.6%)	79 (51.6%)

Number of candidate pairs removed by each filtering approach

Conclusions

We present a fast and efficient three-step filtering strategy that removes most sequence comparisons and accurately establishes transmission links of any threshold-based method. This highly efficient workflow will allow a faster response and molecular detection capacity, improving the rate of detection of viral transmissions with molecular data.

Acknowledgements

The authors are grateful to Scott Sammons (AMD, CDC) and the Scientific Computing team (SciComp, AMD, CDC) for their support and guidance on the development of computational tools accessible to the Public Health community.

Funding

All work and publication costs were funded by the Centers for Disease Control and Prevention, including the Advanced Molecular Detection program (AMD) and the Division of Viral Hepatitis (DVH).

Availability of data and material

Data can be made available on request.

Authors' contributions

DSC, IR and YK designed the study. IR and DSC designed and implemented the Hamming and the identical sequences filter. SVT, CT, JC, SPC and SA designed and evaluated the performance of the Maximal common substring approach. YZ and DSC designed and implemented the k-mer bloom filter. SS designed and implemented HDIST. IR, DSC and AS evaluated the performance of the pipeline on the experimental datasets. IR and DSC analysed all data. DSC and IR wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare they don't have any competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate
Not applicable.

About this supplement
This article has been published as part of BMC Genomics Volume 18 Supplement 4, 2017: Selected articles from the Fifth IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS 2015): Genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-18-supplement-4.

Publisher's Note

Author details
[1]Molecular Epidemiology and Bioinformatics, Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, GA, USA. [2]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [3]Department of Computer Science, University of Central Florida, Orlando, FL, USA. [4]Institute for Data Engineering and Science, Georgia Institute of Technology, Atlanta, GA, USA.

Published: 24 May 2017

References

1. Mohd Hanafiah K, Groeger J, Flaxman AD, Wiersma ST. Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. Hepatology. 2013;57(4):1333–42.
2. Alter M. Epidemiology of hepatitis C virus infection. World J Gastroenterol. 2007;13(17):2436–41.
3. Ly KN, Xing J, Klevens RM, Jiles RB, Ward JW, Holmberg SD. The increasing burden of mortality from viral hepatitis in the United States between 1999 and 2007. Ann Intern Med. 2012;156(4):271–8.
4. Ward JW. The hidden epidemic of hepatitis C virus infection in the United States: occult transmission and burden of disease. Topics Antiviral Med. 2013;21(1):15–9.
5. Spada E, Abbate I, Sicurezza E, Mariano A, Parla V, Rinnone S, Cuccia M, Capobianchi MR, Mele A. Molecular epidemiology of a hepatitis C virus outbreak in a hemodialysis unit in Italy. J Med Virol. 2008;80(2):261–7.
6. Bracho MA, Gosalbes MJ, Blasco D, Moya A, Gonzalez-Candelas F. Molecular epidemiology of a hepatitis C virus outbreak in a hemodialysis unit. J Clin Microbiol. 2005;43(6):2750–5.
7. Gonzalez-Candelas F, Bracho MA, Wrobel B, Moya A. Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source. BMC Biol. 2013;11:76.
8. Prosperi MC, De Luca A, Di Giambenedetto S, Bracciale L, Fabbiani M, Cauda R, Salemi M. The threshold bootstrap clustering: a new approach to find families or transmission clusters within molecular quasispecies. PLoS One. 2010;5(10):e13619.
9. Feray C, Bouscaillou J, Falissard B, Mohamed MK, Arafa N, Bakr I, El-Hoseiny M, Daly ME, El-Kafrawy S, Plancoulaine S, et al. A novel method to identify routes of hepatitis C virus transmission. PLoS One. 2014;9(1):e86098.
10. Campo D, Xia G, Dimitrova Z, Lin Y, Ganova-Raeva L, Punkova L, Ramachandran S, Thai H, Sims S, Rytsareva I, et al. Accurate genetic detection of hepatitis C virus transmissions in outbreak settings. J Infect Dis. 2015;213(6):957–65.
11. Nainan O, Alter M, Kruszon-Moran D, Gao F, Xia G, McQuillan G, Margolis H. Hepatitis C virus genotypes and viral concentrations in participants of a general population survey in the United States. Gastroenterology. 2006;131(2):478–84.
12. Thompson N, Novak R, White-Comstock M, Xia G, Ganova-Raeva L, Ramachandran S, Khudyakov Y, Bialek S, Williams I. Patient-to-patient hepatitis C virus transmissions associated with infection control breaches in a hemodialysis unit. J Nephrol Ther. 2012;S10:002.
13. Ganova-Raeva LM, Dimitrova ZE, Campo DS, Lin Y, Ramachandran S, Xia GL, Honisch C, Cantor CR, Khudyakov YE. Detection of hepatitis C virus transmission by use of DNA mass spectrometry. J Infect Dis. 2013;207(6):999–1006.
14. Ramachandran S, Purdy MA, Xia GL, Campo DS, Dimitrova ZE, Teshale EH, Teo CG, Khudyakov YE. Recent population expansions of hepatitis B virus in the United States. J Virol. 2014;88(24):13971–80.
15. Williams I. Epidemiology of hepatitis C in the United States. Am J Med. 1999;107(6B):2S–9S.
16. Campo DS, Dimitrova Z, Yamasaki L, Skums P, Lau DT, Vaughan G, Forbi JC, Teo CG, Khudyakov Y. Next-generation sequencing reveals large connected networks of intra-host HCV variants. BMC Genomics. 2014;15 (Suppl 5):S4. https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-S5-S4.
17. Warner AE, Schaefer MK, Patel PR, Drobeniuc J, Xia G, Lin Y, Khudyakov Y, Vonderwahl CW, Miller L, Thompson ND. Outbreak of hepatitis C virus infection associated with narcotics diversion by an hepatitis C virus-infected surgical technician. Am J Infect Control. 2015;43(1):53–8.
18. Ramachandran S, Xia GL, Ganova-Raeva LM, Nainan OV, Khudyakov Y. End-point limiting-dilution real-time PCR assay for evaluation of hepatitis C virus quasispecies in serum: performance under optimal and suboptimal conditions. J Virol Methods. 2008;V. 151(№ 2):217–24.
19. Skums P, Dimitrova Z, Campo DS, Vaughan G, Rossi L, Forbi JC, Yokosawa J, Zelikovsky A, Khudyakov Y. Efficient error correction for next-generation sequencing of viral amplicons. BMC Bioinformatics. 2012;13(Suppl 10):S6. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S10-S6.
20. Melsted P, Pritchard JK. Efficient counting of k-mers in DNA sequences using a bloom filter. BMC Bioinformatics. 2011;12:333.