HHS Public Access

Author manuscript

J Comput Biol. Author manuscript; available in PMC 2024 August 05.

Published in final edited form as:

J Comput Biol. 2023 April ; 30(4): 409–419. doi:10.1089/cmb.2022.0292.

Novel network method Major Minor Variation Clustering (MMVC) enables identification of poliovirus clusters with high-resolution linkages

Jiahui Tan[1], **Yutong Zhao**[1], **Cara C. Burns**[2], **Dechao Tian**[1], **Kun Zhao**[2]

[1]School of Public Health (Shenzhen), Sun Yat-sen University, Shenzhen, China

[2]Polio and Picornavirus Laboratory Branch, Division of Viral Diseases, Centers for Disease Control and Prevention, Atlanta, USA

Abstract

The Global Polio Eradication Initiative (GPEI) uses an outbreak response protocol that defines type 2 Sabin or Sabin-like virus as those with 0–5 nucleotides diverging from their parental strain in the complete VP1 genomic region. Sabin or Sabin-like viruses share highly similar genome sequences, regardless of their origin. Thus, it is challenging to distinguish viruses at a higher resolution to detect polio clusters or trace sources for local transmissions of viruses at an early stage. To identify type 2 Sabin or Sabin-like sources and improve our ability to map viral sources to campaigns during the polio endgame, we investigated the feasibility of a new method for genetic sequence analysis. We named the method Major Minor Variation Clustering (MMVC), which uses a network model to simultaneously incorporate sequence similarity in major and minor variants in addition to onset dates to detect fine-scale polio clusters. Each identified cluster represents a collection of sequences that are highly similar in both major and minor variants, enabling the discovery of new links between viruses. By applying the method to a published data set collected in Nigeria during 2009–2012, we found that the detected clusters identified using this method have several improvements over clusters derived from a phylogenetic tree approach. Extensive integrative data analysis reveals that sequences in the same cluster have higher genomic similarities and better agreement with onset dates. As a complement to current phylogenetic tree approaches, MMVC has the potential to improve epidemiological surveillance and investigation precision to guide polio eradication.

Keywords

Whole Genome Sequencing; Genomic epidemiology; Network modeling; Polio eradication

Address correspondence to Dechao Tian: tiandch@mail.sysu.edu.cn; Kun Zhao: vzt5@cdc.gov.

Authorship confirmation/contribution statement

K.Z., D.T., and C.C.B. conceived and designed the study. D.T. and K.Z. designed the method and analyses. J.T. and Y.Z. wrote the code, performed the experiments, and analyzed the data. K.Z., D.T., and C.C.B. interpreted the results. J.T. and Y.Z. wrote the original draft. K.Z., D.T, and C.C.B. reviewed and edited the manuscript. All authors have read and approved the manuscript.

Conflict of interest: none. The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention. This is an extended article from a conference presentation. No preprint was published.

Author disclosure statement

INTRODUCTION

Since the launch of the Global Polio Eradication Initiative (GPEI) in 1988, the number of wild poliovirus cases has declined by more than 99.99% (Jorba et al, 2019b). Thanks to the continued success of widely used attenuated oral polio vaccine (OPV), there were only five wild polio cases in Pakistan and Afghanistan combined in 2021 (WHO, 2022). Of the three serotypes of wild poliovirus, type 2 was certified as eradicated in 2015 and type 3 was certified as eradicated in 2018 (WHO, 2022). Despite the elimination of wild poliovirus in most parts of the world, polio persists in populations where high-quality vaccination programs may not reach every suspectable child due to economic, social, and/or political factors. Thus, poliovirus continues to threaten global health and hinders the fulfillment of the global polio eradication promise.

The Global Polio Laboratory Network monitors Sabin, Sabin-like poliovirus, and vaccine-derived poliovirus (VDPV) through acute flaccid paralysis (AFP) and environmental surveillance systems. In this study, we will be focusing on type 2 Sabin-like polioviruses because of the predominance of circulating VDPV2 over the past decade and the fact that Sabin-like polioviruses represent an early stage of VDPV2 evolution. Past dataset analyses present challenges in clustering due to low genetic diversity and limited availability of whole genome sequences. (Famulare et al, 2021; Valesano et al, 2021; Zhao et al, 2017)

Although the exact time periods of type 2 OPV (Sabin) vaccination campaigns are known, accurately mapping Sabin-like viral sequences obtained from an AFP surveillance system to the appropriate campaign time poses a major challenge during outbreak analysis, especially since type 2 Sabin-like viruses have a difference of less than 6 nucleotides (NTs) in the 903-NT VP1 genomic region. Analysis protocols utilizing the neighbor-joining (NJ) phylogenetic tree, maximum likelihood tree, and Bayesian tree, have proven to be as successful as a standard method for analyzing VDPV and wild polioviruses (WPVs) (Jorba J, 2018).

Phylogenetic tree approaches often use major variations obtained from genome consensus generated in the VP1 region (with a possible extension to the whole capsid) to characterize the VDPVs and WPVs during routine molecular surveillance. These approaches may overlook major variations (SNPs) in non-capsid regions and minor variations in the whole genome that may contain potent transmission information that could be used to establish possible linkages between a pair of viruses, particularly among highly similar sequences in Sabin-like viruses. In this study, we hypothesize that using major and minor variant frequencies in the whole genome with onset dates could improve the accuracy of mapping vaccination campaign dates. Additionally, to validate the initial hypothesis, we developed a new method named the Major Minor Variation Clustering (MMVC), which uses a well-established network framework for difficult datasets.

MATERIAL AND METHODS

2.1. Overview of MMVC

An overview of MMVC is illustrated in Figure 1. MMVC aims to identify clusters with genetic and epidemiological links to infer a possible campaign timing of Sabin-like poliovirus. From the sequencing data, MMVC extracts both major and minor SNPs for each individual samples (Fig. 1a). It then computes two Jaccard matrices whose elements are Jaccard indices between corresponding pairs of sequences based on their major and minor SNPs respectively, in which variations in VP1–4 have greater weighs than variations in other regions (e.g., 0.7 vs 0.3). Each of the computed Jaccard matrix is used to construct a k-nearest neighbor (KNN) graph (Fig. 1b).

A KNN graph is then used to produce an SNN network. An edge in a SNN network indicates high similarity in variation profiles of corresponding samples. Similarly, an SNN network for the samples using their onset dates is constructed. An edge between two samples indicates that a high proportion of samples having onset dates within the interval defined by the onset dates of the two samples. These three SNN networks (constructed using major SNPs, minor SNPs, and onset dates, respectively) form a multiplayer network, which is then merged into a weighted SNN graph (Fig. 1c). A modularity optimization method is used to divide the weighted SNN graph into multiple clusters (Fig. 1d) (Lu et al, 2021).

2.2. MMVC step-Measuring genetic similarity with the Jaccard similarity index

In the phylogenetic study of viral genomes, the Jaccard similarity index was widely used to quantify similarities of variations in the phylogenetic analysis of human, bacteria, and virus genomes (Comas et al, 2009; Ghosh et al, 2021; Wang et al, 2015; Yu et al, 2017a; Yu et al, 2017b). The Jaccard index, J(A, B), is defined as the intersection between two sequence sets S_1 and S_2, divided by the union of the two sequence sets (Seweryn et al, 2020).

$$JI(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1| + |S_2| - |S_1 \cap S_2|} \tag{1}$$

The polio capsid coded by the VP1–4 region protects viral nucleic acids and escorts them in and out of host cells. Variations in VP1–4 have been used to characterize polioviruses and quantify the relationship between viruses (Burns et al, 2013; Kirkegaard, 1990; Shaw et al, 2018; WHO, 2020). To compute the Jaccard similarity index between SNPs from two genome sequences, the Jaccard index assigns varying weights to SNPs located at various positions. Specifically, in Formula 2, C_S represents the capsid region, and NC_S represents non-capsid regions S, which are assigned weights at $w_1 = 70\%$ and $w_2 = 30\%$ respectively. The Jaccard similarity index of major SNPs for a pair of sequences is calculated below:

$$\mathrm{JI}_{\mathrm{major}}(S_1, S_2) = w_1 \times \mathrm{JI}_{\mathrm{major}}(C_{S_1}, C_{S_2}) + w_2 \times \mathrm{JI}_{\mathrm{major}}(NC_{S_1}, NC_{S_2}). \tag{2}$$

Similar to Formula (2), the Jaccard similarity index of minor SNPs, is computed as $JI_{minor}(S_1, S_2)$. The Jaccard index of the SNP sets of two viral isolates were used to determine their genetic closeness. A greater $JI(S_1, S_2)$ value implies that two sequences could share more SNPs.

2.3. MMVC step-network construction and clustering

MMVC takes three inputs to construct networks: major JI similarity matrix, minor JI similarity matrix and onset dates. For each input, a KNN network is constructed in which each node has exactly k interactions (neighbors). A KNN network is used to produce a SNN network two nodes are connected by an edge in the SNN network if the two nodes share a higher fraction of common neighbors (quantified by Jaccard similarity index) than a given cutoff c in the KNN network. The cutoff $c \in [0,1]$ is set for acceptable Jaccard similarity index when computing the neighborhood overlap for the SNN construction, and any edges with values less than or equal to c will be removed.

Next, three SNN networks (major, minor, and onset dates) are combined into one weighted SNN network. Specifically, the two SNN networks from JI matrices of major SNPs and minor SNPs are combined with weights equal to 0.7 and 0.3, respectively. After testing multiple weight assignments, the selected weight combination where major SNPs account for 70% and minor SNPs is 30% produced better results. The resulting network is then combined with the SNN network of onset dates with an equal weight (Fig. 1c).

Following the principles of intra-cluster density and inter-cluster sparsity, a Louvain community detection (Blondel et al, 2008) is calculated to identify a subset of nodes with a large number of internal edges and few external edges. The clustering of Sabin 2 sequences is achieved by dividing the graph into k subgraphs (communities).

2.4 Determining the number of clusters

We used a data-driven approach to determine the number of clusters. Specifically, vaccination campaigns are leveraged to choose the suitable number of clusters. Thirty isolate samples were collected in this study during a total of eight completed vaccination campaigns in Nigeria. Among the 30 samples, eight samples with an onset date earlier than the date of the earliest campaign activity (March 3, 2010) were considered as forming at least one cluster. Next, by varying the number of clusters from 6 to 9 only, subtle changes were noticed in detected clusters. Considering the information above, the number of clusters for the 30 samples collected here is chosen at 8.

To obtain 8 clusters by MMVC, we set $k = 5$ for the SNN algorithm and $c = 1/5$ for the acceptable Jaccard similarity index when computing the neighborhood overlap for the SNN construction (Zhu et al, 2020). The cluster resolution is equal to 1.73. As for NJ tree, the *hcut* function in R is used to divide the phylogenetic tree into 8 clusters.

Automatically detect the number of clusters in a network is a fundamental computational challenge in the cluster detection network science, which is beyond the scope of this paper. Thus, automatically detecting the number of clusters is not included as a feature of MMVC.

2.5. Data acquisition

The MMVC and NJ tree techniques are applied to a publicly available data set that contains stool samples collected from 30 individuals from routine AFP surveillance in Nigeria from 2009 to 2012. Eight possible supplementary immunization activities (SIAs) and corresponding dates during this time-period were collected in a previous study. The study included isolates categorized as type 2 “Sabin-like” utilizing real-time reverse transcription-PCR for intratypic differentiation and next generation sequencing at the US Centers for Disease Control and Prevention. The complete genome sequence of human poliovirus strain 2, Sabin 2, was downloaded from the NCBI (SRA BioSample SAMN04125839 to SAMN04126091 and NCBI accession number KJ170425 through KJ170677; AY184220 for the reference genome) (Famulare et al, 2016).

2.6. MMVC step-Poliovirus genome sequencing and assembly

The preliminary quality evaluation for each sample was generated using the FASTQC program. Geneious [“Geneious Prime 2021.2.2.(https://www.geneious.com)”] is used to pre-process the sample data. First, reads were trimmed by BBDuk [BBtools:BBMap - Bushnell B. - sourceforge.net/projects/bbmap/] with following settings: Set to Trim adapters to: trim from the right end, Kmer length is 27 and Maximum substitutions is 1; set to trim low quality, set Minimum Quality to 30. Reads with less than 20 bases are removed. Duplicated reads are removed by setting Kmer seed length to 27 and maximum substitutions to 1. Resulted reads are then mapped to the Sabin 2 reference sequence with the default parameter settings in Geneious.

Major and minor variations/SNPs are called using Geneious. A p-value represents the probability that a disagreement is a variant rather than a sequencing error, considering the given sum of qualities. The lower the p-value, the more likely the variation at the given position represents a real variant. A variant is called if its p-value is no more than 10^{-6}. After collecting genome-wide SNPs for each sample, a threshold of 0.5 is applied for [0, 1], and major variants are called if > 0.5 and minor variants if < 0.5. No 50–50 SNP situation has been observed.

RESULTS

3.1. Method comparison using major and minor SNP profiles

To compare the performance between NJ tree and MMVC, we visualized major and minor SNP profiles in each genome position among 30 type 2 Sabin-like poliovirus sequences. A heatmap confirmed that all sequences contain less than 6 NTs in VP1 region among major variations (Fig. 2a), consistent with prior knowledge (Famulare et al, 2016). Minor variations were spread over the whole genome (Fig. 2b).

Overall, the clustering results are largely consistent between MMVC and NJ tree. MMVC and NJ tree almost coincided on the two largest clusters (clusters #1, #2), and agreed on two singleton clusters (#6, #7) (Fig. 2c). MMVC clusters and NJ tree clusters have no remarkable differences in genetic distance between samples (Fig. S1).

We found that the MMVC cluster method had multiple improvements over NJ tree cluster analysis, as highlighted by the grey boxes in Fig. 2a. Each box highlights a scenario where genome sequences are assigned to the same cluster by NJ tree but share no major SNPs, indicating potentially weak NJ clusters. For example, genome sequences for samples with ID 25 and ID 10 (Table S1) shared no major SNPs but are clustered into cluster #1 by NJ tree (left grey box in Fig. 2a). Instead, the two samples are clustered into two different clusters (#1, #3) by MMVC (Fig. 2c).

Similar scenarios also exist for minor SNP profiles (Fig. 2b, 2d), which is expected as NJ tree does not use genetic distances in terms of minor SNPs whereas MMVC does. MMVC in these scenarios do not show a sacrifice in genetic distances among samples in the same clusters. In fact, samples in the same MMVC clusters have slightly shorter genetic distances than those in the same NJ tree clusters (Fig S1). Taken together, MMVC has multiple improvements over NJ tree in identifying genetically linked clusters.

3.2. Method comparison by onset dates and vaccination campaigns

The performance between NJ tree and MMVC methods is further compared with respect to onset dates and vaccination campaigns. Because we are focusing on Sabin-like viruses at an early stage in evolution and in places where routine immunization coverage was low, genome sequences in a cluster should not be from samples with a long span of onset dates. We found that the non-singleton clusters (those with at least two sequences) identified by MMVC has a narrower span of onset dates and higher proportion of sequences in a single cluster mapped to a prior campaign date than those by the NJ tree, which is expected as the MMVC method incorporates the onset date as input, while NJ tree method does not.

Among the non-singleton clusters #1-#5 identified by MMVC, 3 clusters (#2-#4) had samples with onset date spans that were within 8 days, 38 days, and 14 days of each other, respectively. In contrast, clusters #1, #2, and #3 identified by the NJ tree contained sequences that were collected within 12 months, 15 months, and 16 months of each other, respectively (Table S1). Within the one-month time window covered by MMVC clusters #2, #3, and #4, no samples were assigned to other clusters, suggesting consistency between genetic and temporal information within those three MMVC clusters. Closer examination showed that some MMVC clusters can be mapped to vaccination campaigns within two months (Fig. 3 and Table S1).

For this analysis, a genome sequence can be considered to have a potential to be mapped to a vaccination campaign if the campaign took place less than 90 days before the sample onset date, according to the molecular clock of poliovirus (Jorba et al, 2008). Genome sequences in MMVC cluster #4 were linked to two vaccination campaigns that took place on January 1 and February 1, 2011 while genome sequences in cluster #3 were mapped to two campaigns that took place on February 1 and March 1, 2011. In contrast, genome sequences in each of the non-singleton NJ clusters #1, #2, and #3 were mapped to more than three vaccination campaigns, as samples in each of those three clusters were collected spanning at least 12-month periods.

Cluster #1 and cluster #5 have samples that are within 12 months and 22 months of each other, respectively, suggested that onset date alone is not sufficient to link samples by MMVC and further improvement is needed. These results also highlight that when mapping clustered sequences to the campaign dates, onset date alone could be assigned with a greater weight when a cluster spans over 3 months. It is important to note that neither MMVC nor NJ tree use campaign dates as input data. These results demonstrates that MMVC was superior to NJ tree for mapping clustered sequences to vaccination campaigns.

3.3. Result evaluation by the number of shared major and minor SNPs

A superior method is expected to have more shared SNPs within a cluster and fewer shared SNPs between clusters. Shared mutational variants is indicative of possible transmission events and associated routes (Lickness et al, 2020; Worby et al, 2017). If two sequences share SNPs, the possibility that they were linked to the same transmission events and routes could be increased, which in turn enhances the possibility that they could be both traced to the same vaccination campaign. The difference in the number of shared SNPs between genome sequences was also used to compare the MMVC and NJ tree methods.

We found that genome sequences within MMVC clusters shared significantly (P=0.046) more major SNPs compared to genome sequences within NJ clusters (Fig. 4a). Particularly, each of 3 pairs (4.69%) of intra-cluster samples identified by MMVC shared 2 major SNPs. In contrast, no sequence pairs in the same NJ tree clusters shared 2 or more major SNPs. Regarding the number of shared minor SNPs between pairs of sequences from different clusters, MMVC clusters tended to have smaller (P=0.08) numbers of shared minor SNPs than NJ tree's inter-clusters (Fig. 4d).

MMVC and NJ clusters had no significant differences when looking at the number of shared minor SNPs within clusters and the numbers of shared major SNPs between clusters (P>0.1, Fig. 4b–c). Clustering type 2 Sabin-like polioviruses using SNP profiles is challenging due to low genetic diversity (Famulare et al, 2021; Jorba et al, 2019a). These results highlight that incorporating minor SNPs and onset dates as input data increases the number of shared major SNPs in MMVC clusters over NJ clusters which could in turn identify possible links between sequences which it may be overlooked by the NJ tree method. MMVC performed better than NJ tree methods due to the multilayer network design that incorporates 3 components in the algorithm.

DISCUSSION

New computational methods are needed to incorporate both genomic and epidemiologic profiles for differentiating highly similar sequences so that the goal of polio eradication can be met. To better identify linked clusters of type 2 Sabin-like polioviruses that may be mapped to known campaign dates, MMVC considers major SNPs, minor SNPs, and onset dates of samples. Utilizing a clustering framework for heterogeneous networks has multiple advantages over NJ tree method, such as higher proportion of similarities in major SNP profiles between samples in the same clusters (Fig. 2), higher number of shared major SNPs in the same clusters (Fig. 4), and better agreements between onset dates and vaccination campaigns (Fig. 3).

This study considers NJ tree method as the reference method. In addition, other tree-based methods (such as Maximum likelihood and Bayesian methods) were compared to MMVC. Varying the parameter settings using by phylogenetic tree methods, similar topologies were produced.

There were several fundamental differences between the proposed network approach and traditional tree-based methods. First, the network model allows for the existence of isolated nodes where no edge or branch was connected to nodes. In contrast, all end nodes must be connected by common ancestor nodes (internal nodes) and the root in the tree-based method. The presence allowance of isolated nodes, not required to be connected to others as in the NJ tree method, allows for a more straightforward identification of isolated individual sequences. This fundamental difference also leads to easier identification of clusters from a graph topology. Secondly, connections between nodes are allowed in a network model whereas a tree does not allow for such connections. Connections in a network are imperative to understanding the relationship between any pair of virus sequences. Thirdly, the accurate rooting of a phylogenetic tree is important to show the directionality of viral evolution. Assuming only a single root in a phylogenetic tree overlooks the existence of multiple viral sources (Sabin 2 root) during global polio vaccination campaigns. In other words, it is possible to observe and map multiple Sabin-like viruses to their Sabin-2 origins at various campaign times, even though the sequences may seem to be very similar due to low genetic diversity. MMVC relaxes the assumption of a single source in the tree and focuses on examining relationships between viral sequences.

Another major methodological contribution of our framework is that it utilizes a heterogeneous network model to integrate genomic and epidemiological data. The heterogeneous network model can be further modified to identify clusters by incorporating additional epidemiological data. For example, in additional to considering genomic and onset date information as we did in this analysis, other types of data that are collected by routine surveillance, such as genome sequences of environmental samples and relevant demographic data could be incorporated. These results also confirmed previous understanding that incorporating onset date can improve the clustering quality compared to using only genome sequence data. In addition, this framework can be easily adopted to support other pathogenic surveillance where the genomic data and epidemiological data is available.

This work has demonstrated that it is possible to improve existing tree-based method via MMVC with some limitations. First, the cluster number determination was based on the eight vaccination campaigns that were conducted during the study period. By mapping Sabin-like sequences to vaccination campaign dates, these campaigns were assumed to be the main (or probably sole) sources for the AFP cases. There are some singleton nodes in the network shown in Figure 2. MMVC cannot rule out the possibility of samples from other sources through routine immunization, travel and population mobility, and additional information may be needed for these isolated samples. Second, the improvement has been shown in 30 samples. The small sample size includes all available whole genome sequences of Nigerian isolates during this study period. During this period, WPV and VDPV sequencing had priority in routine surveillance. More sequencing in a well-designed and

well-representative sampling study will help further validate our observations in this study. Lastly, published data collected in the surveillance system is assumed to be error free. In the real-world, these data are still subject to collection errors, however, we assumed that errors in this study were negligible.

In summary, this study reveals that MMVC may outperform tree-based clustering method in multiple aspects. Firstly, MMVC integrates both major and minor variation frequencies and onset date information. The flexibility of this network model could enable MMVC to detect poliovirus clusters with stronger genetic and epidemiological linkages, when compared with tree-based methods such as the neighbor-joining tree (NJ). Secondly, the application of this method to 30 samples collected in Nigeria (2009–2012) reveals that clusters detected by MMVC produced more consistent results than the NJ tree method in mapping the clustered sequences to the eight documented vaccination campaign dates during that time. Thirdly, MMVC could identify refined clusters that shared more major variations. Therefore, MMVC has the potential to improve epidemiological interpretation/ analysis and investigation to guide polio eradication.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank William Weldon and Mark Sotir for their insightful comments and feedback.

Funding statement

D.T. was supported by the National Key Research and Development Program of China grant 2021YFC2300102 and GuangDong Basic and Applied Basic Research Foundation grant 2022A1515010043.

References

Blondel VD, Guillaume J-L, Lambiotte R, et al. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008;2008(10); doi: 10.1088/1742-5468/2008/10/p10008.

Burns CC, Shaw J, Jorba J, et al. Multiple independent emergences of type 2 vaccine-derived polioviruses during a large outbreak in northern Nigeria. J Virol 2013;87(9):4907–22; doi: 10.1128/JVI.02954-12. [PubMed: 23408630]

Comas I, Homolka S, Niemann S, et al. Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. PLoS One 2009;4(11):e7815; doi: 10.1371/journal.pone.0007815. [PubMed: 19915672]

Famulare M, Chang S, Iber J, et al. Sabin Vaccine Reversion in the Field: a Comprehensive Analysis of Sabin-Like Poliovirus Isolates in Nigeria. J Virol 2016;90(1):317–31; doi: 10.1128/JVI.01532-15. [PubMed: 26468545]

Famulare M, Wong W, Haque R, et al. Multiscale model for forecasting Sabin 2 vaccine virus household and community transmission. PLoS Comput Biol 2021;17(12):e1009690; doi: 10.1371/journal.pcbi.1009690. [PubMed: 34932560]

Ghosh N, Saha I, Sharma N, et al. Genome-wide analysis of 10664 SARS-CoV-2 genomes to identify virus strains in 73 countries based on single nucleotide polymorphism. Virus Res 2021;298:198401; doi: 10.1016/j.virusres.2021.198401. [PubMed: 33781798]

Jorba J, Campagnoli R, De L, et al. Calibration of multiple poliovirus molecular clocks covering an extended evolutionary range. J Virol 2008;82(9):4429–40; doi: 10.1128/JVI.02354-07. [PubMed: 18287242]

Jorba J, Diop OM, Iber J, et al. Update on Vaccine-Derived Poliovirus Outbreaks - Worldwide, January 2018-June 2019. MMWR Morb Mortal Wkly Rep 2019a;68(45):1024–1028; doi: 10.15585/mmwr.mm6845a4. [PubMed: 31725706]

Jorba J, Diop OM, Iber J, et al. Update on Vaccine-Derived Poliovirus Outbreaks — Worldwide, January 2018–June 2019. MMWR. Morbidity and Mortality Weekly Report 2019b;68(45):1024–1028; doi: 10.15585/mmwr.mm6845a4. [PubMed: 31725706]

Jorba J DO, Iber J, et al. Update on Vaccine-Derived Polioviruses - Worldwide, January 2017-June 2018. MMWR Morb Mortal Wkly Rep 2018;vol. 67,42 1189–1194; doi: 10.15585/mmwr.mm6742a5.

Kirkegaard K Mutations in VP1 of poliovirus specifically affect both encapsidation and release of viral RNA. J Virol 1990;64(1):195–206; doi: 10.1128/jvi.64.1.195-206.1990. [PubMed: 2152812]

Lickness JS, Gardner T, Diop OM, et al. Surveillance to Track Progress Toward Polio Eradication - Worldwide, 2018–2019. MMWR Morb Mortal Wkly Rep 2020;69(20):623–629; doi: 10.15585/mmwr.mm6920a3. [PubMed: 32437342]

Lu S, Conn DJ, Chen S, et al. MLG: multilayer graph clustering for multi-condition scRNA-seq data. Nucleic Acids Res 2021; doi: 10.1093/nar/gkab823.

Seweryn MT, Pietrzak M, Ma Q. 2020. Application of information theoretical approaches to assess diversity and similarity in single-cell transcriptomics. In Comput Struct Biotechnol J. 1830–1837.

Shaw J, Jorba J, Zhao K, et al. Dynamics of Evolution of Poliovirus Neutralizing Antigenic Sites and Other Capsid Functional Domains during a Large and Prolonged Outbreak. J Virol 2018;92(9); doi: 10.1128/JVI.01949-17.

Valesano AL, Taniuchi M, Fitzsimmons WJ, et al. The Early Evolution of Oral Poliovirus Vaccine Is Shaped by Strong Positive Selection and Tight Transmission Bottlenecks. Cell Host Microbe 2021;29(1):32–43 e4; doi: 10.1016/j.chom.2020.10.011. [PubMed: 33212020]

Wang C, Kao WH, Hsiao CK. Using Hamming Distance as Information for SNP-Sets Clustering and Testing in Disease Association Studies. PLoS One 2015;10(8):e0135918; doi: 10.1371/journal.pone.0135918. [PubMed: 26302001]

WHO. Standard operating procedures: responding to a poliovirus event or outbreak, version 3.1. World Health Organization. https://apps.who.int/iris/handle/10665/331895. License: CC BY-NC-SA 3.0 IGO 2020.

WHO. GPEI Polio Now; 2022. Available from: https://polioeradication.org/polio-today/polio-now/ [Last accessed].

Worby CJ, Lipsitch M, Hanage WP. Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. Am J Epidemiol 2017;186(10):1209–1216; doi: 10.1093/aje/kwx182. [PubMed: 29149252]

Yu C, Baune BT, Licinio J, et al. A novel strategy for clustering major depression individuals using whole-genome sequencing variant data. Scientific Reports 2017a;7(1):44389; doi: 10.1038/srep44389. [PubMed: 28287625]

Yu C, Baune BT, Licinio J, et al. A novel strategy for clustering major depression individuals using whole-genome sequencing variant data. Sci Rep 2017b;7:44389; doi: 10.1038/srep44389. [PubMed: 28287625]

Zhao K, Jorba J, Shaw J, et al. Are Circulating Type 2 Vaccine-derived Polioviruses (VDPVs) Genetically Distinguishable from Immunodeficiency-associated VDPVs? Comput Struct Biotechnol J 2017;15:456–462; doi: 10.1016/j.csbj.2017.09.004. [PubMed: 29276577]

Zhu X, Zhang J, Xu Y, et al. Single-Cell Clustering Based on Shared Nearest Neighbor and Graph Partitioning. Interdisciplinary Sciences: Computational Life Sciences 2020;12(2):117–130; doi: 10.1007/s12539-019-00357-4. [PubMed: 32086753]

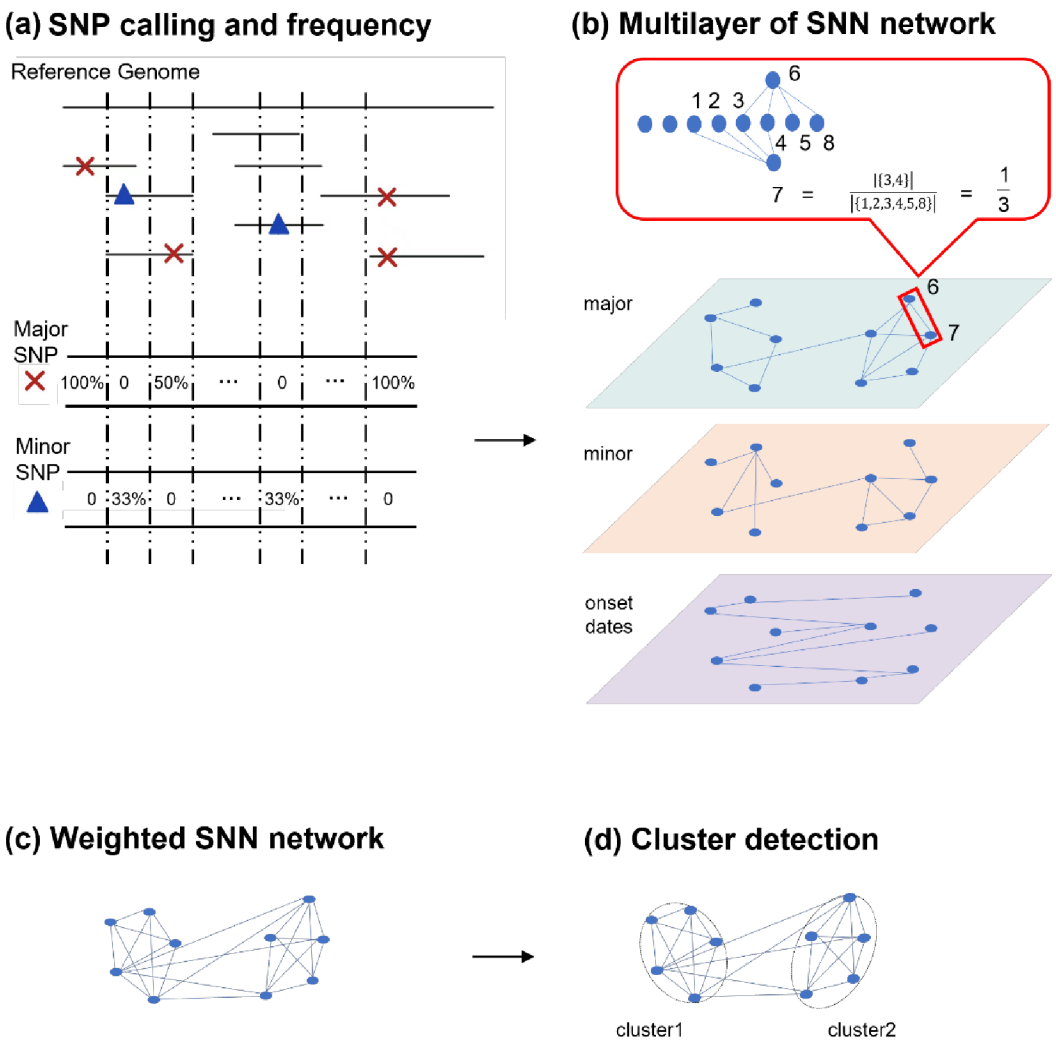

Figure 1.

Workflow of a novel network method, MMVC. (a) SNP Calling and frequency: Sample genomic sequences displayed as inputs for calling SNPs. Major (red cross) and minor SNP (blue triangle) frequencies that have been calculated. (b) Multilayer of SNN network: Sample similarities as computed by weighted Jaccard similarity indices. (c) Weighted SNN network: Network construction using onset date and sample similarities in SNP profiles. (d) Cluster detection: Network clustering by modularity optimization. For presentation purpose, the data shown here are over-simplified examples of the real data.

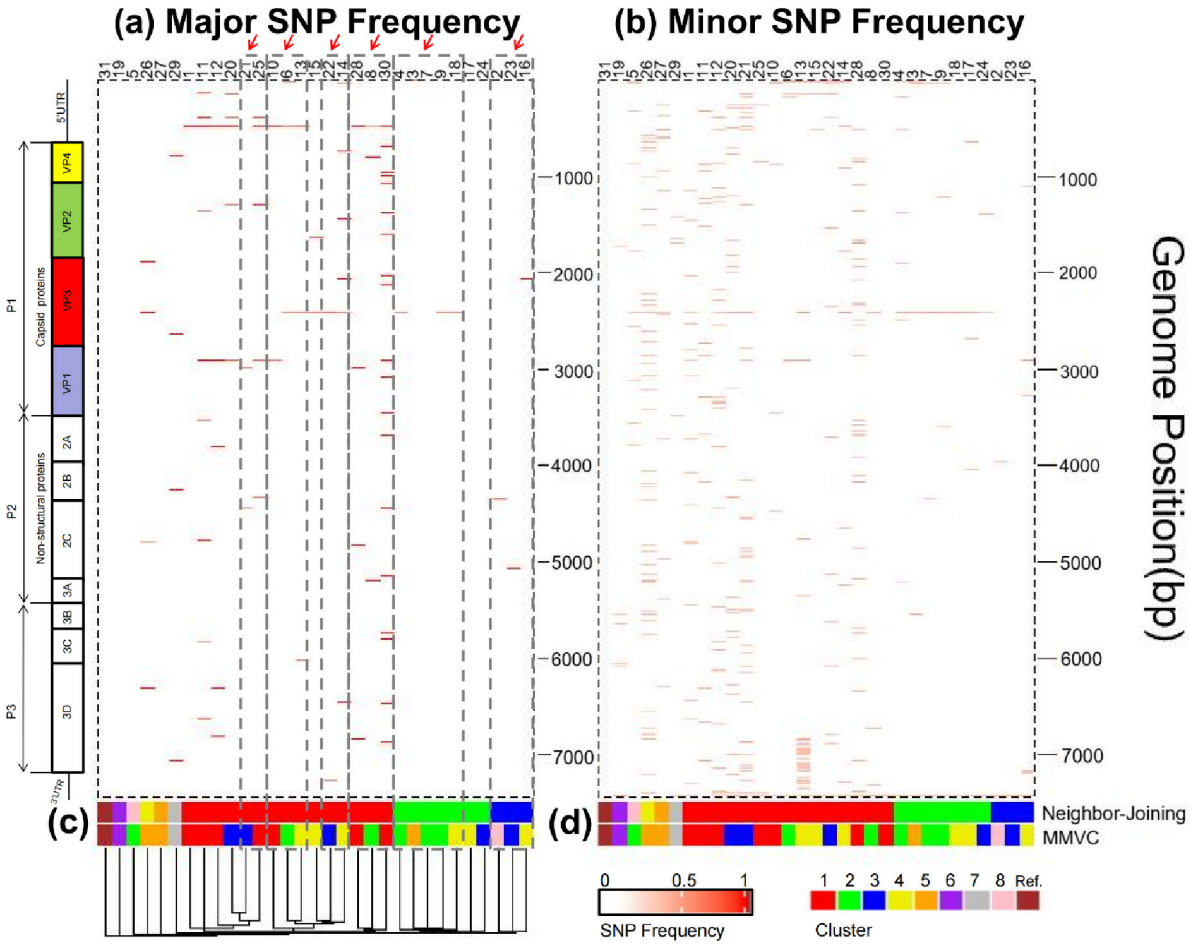

Figure 2.

Method comparison in terms of SNP profiles. (a) Major SNP profile frequency. (b) Minor SNP profile frequency. Grey dashed boxes which red arrows point to highlight questionable clusters by NJ tree algorithm: samples with different SNP profiles are clustered into the same clusters by NJ tree algorithm. Here, the x-axis represents samples and the y-axis represents genome coordinates. The red line represents a mutation in the sample at that position, and the color transparency represents the frequency of genetic mutation. (c) A neighbor-joining tree using major SNPs is annotated at the bottom of the Heatmap. Results from MMVC and neighbor-joining tree clustering are also annotated. (d) The same MMVC and neighbor-joining tree clustering results are annotated for minor SNP profiles. Different cells on the strip correspond to samples on the x-coordinate. The colors of the cells represent different clusters. Sample ID are generated by chronologically ordering the onset date.

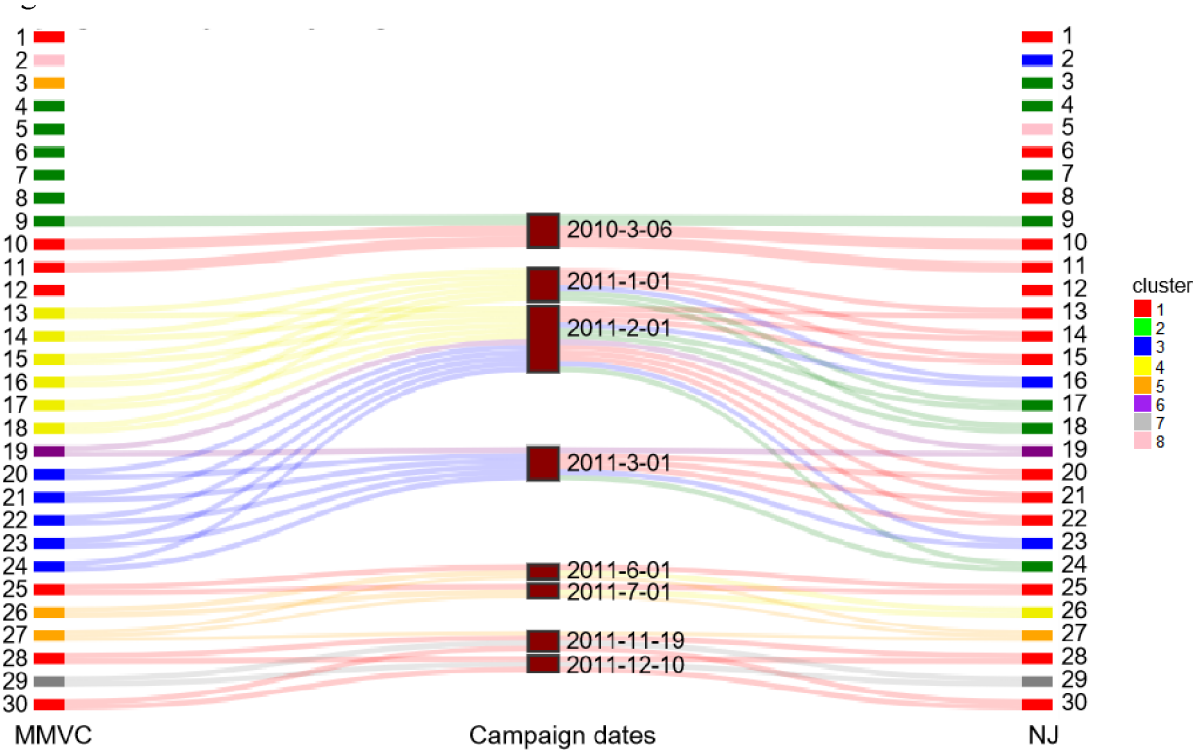

Figure 3.

Method comparison based on campaign activities. Nodes on left (MMVC) and right (NJ) represent the 30 samples, chronologically sorted by onset. The earliest to the most recent samples are shown top to bottom. The colors of the nodes represent their cluster assignments by NJ tree and MMVC, respectively (same as in Figure 2). The red nodes in the middle denote 8 vaccine campaigns. The edges between samples and campaigns are defined if a campaign is less than 3 months from onset date of a sample, indicating that samples could be possibly derived from the linked campaign in the 3 months prior to the onset date. Note that x-axis ticks are spaced to evenly allocate the 30 samples, instead of evenly spaced by onset date.

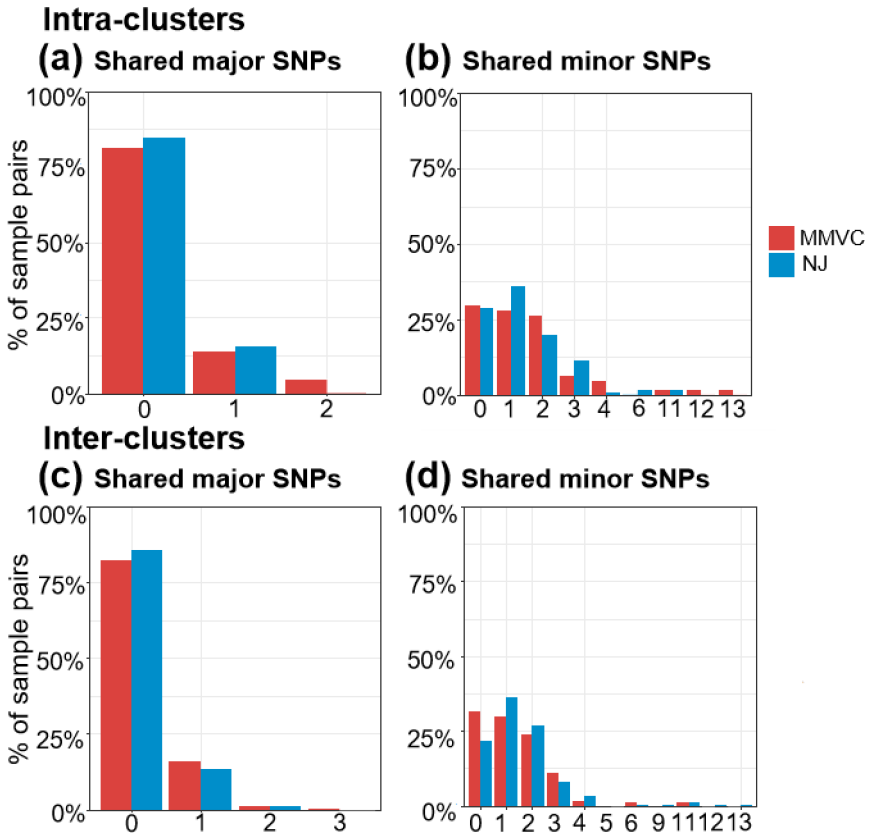

Figure 4.

Method comparison by the number of shared SNPs between samples. (a-b) shared major and minor SNPs between two samples in the same clusters; (c-d) shared major and minor SNPs between two samples from different clusters. *P* value for testing independence between two distributions are 0.046 (a), greater than 0.1 (b-c), and 0.085 (d).