

METHODOLOGY

Open Access



# The Gap Procedure: for the identification of phylogenetic clusters in HIV-1 sequence data

Irene Vrbik<sup>1\*</sup>, David A. Stephens<sup>1</sup>, Michel Roger<sup>2</sup> and Bluma G. Brenner<sup>3,4</sup>

## Abstract

**Background:** In the context of infectious disease, sequence clustering can be used to provide important insights into the dynamics of transmission. Cluster analysis is usually performed using a phylogenetic approach whereby clusters are assigned on the basis of sufficiently small genetic distances and high bootstrap support (or posterior probabilities). The computational burden involved in this phylogenetic threshold approach is a major drawback, especially when a large number of sequences are being considered. In addition, this method requires a skilled user to specify the appropriate threshold values which may vary widely depending on the application.

**Results:** This paper presents the *Gap Procedure*, a distance-based clustering algorithm for the classification of DNA sequences sampled from individuals infected with the human immunodeficiency virus type 1 (HIV-1). Our heuristic algorithm bypasses the need for phylogenetic reconstruction, thereby supporting the quick analysis of large genetic data sets. Moreover, this fully automated procedure relies on data-driven gaps in sorted pairwise distances to infer clusters, thus no user-specified threshold values are required. The clustering results obtained by the Gap Procedure on both real and simulated data, closely agree with those found using the threshold approach, while only requiring a fraction of the time to complete the analysis.

**Conclusions:** Apart from the dramatic gains in computational time, the Gap Procedure is highly effective in finding distinct groups of genetically similar sequences and obviates the need for subjective user-specified values. The clusters of genetically similar sequences returned by this procedure can be used to detect patterns in HIV-1 transmission and thereby aid in the prevention, treatment and containment of the disease.

**Keywords:** Clustering, Phylogenetics, HIV, Genetic distance estimation

## Background

In an age overwhelmed by a massive influx of data, the need for fast and effective clustering techniques has never been greater. This endeavour is particularly important in genetics where the sheer volume of data renders many popular clustering techniques prohibitive or ineffective. The present paper aims at developing new techniques for identifying clusters of genetically similar DNA sequences and pays particular attention to HIV-infected individuals from Quebec, Canada.

In a 2013 surveillance report released by the Public Health Agency of Canada (PHAC), it is estimated that a cumulative total of 78,511 cases of HIV have been

reported in Canada since 1985. Of the 2090 reported HIV cases in 2013, Quebec accounted for 21.7 %. This percentage is second only to the province of Ontario, which contributed 39.6 % of the total HIV cases in the PHAC report. Previous population-based studies involving the phylogenetic analysis of Quebec's primary HIV infection cohort have revealed clusters that correlate with distinct social networks and risk behaviours [1–3]. In other studies performed outside of Quebec, phylogenetic clusters have been used to provide crucial insights about the spread and transmission of the disease [4–9].

Although there are a number of programs available for clustering nucleotide sequences (e.g., BLASTClust [10], UPGMA and WPGMA [11], neighbor-joining (NJ) [12], and phyclust [13]), phylogenetic approaches have been

\*Correspondence: irene.vrbik@mcgill.ca

<sup>1</sup> Department of Mathematics & Statistics, McGill University, 805 Sherbrooke Street West, H3A 0B9 Montreal, Canada

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

ubiquitous in the literature involving HIV-1 transmission clusters. Broadly speaking, phylogenetics is the study of evolutionary relationships among organisms or taxon [14]. There are a number of programs for inferring phylogenies including, but not limited to, PAUP\* [15], BAMBE [16], BEAST [17], PHYLIP [18] RAxML [19] and MrBayes [20]. These relationships can be represented using a phylogenetic tree wherein branch lengths commonly reflect the estimated number of nucleotide substitutions between organisms. In the present study, the ‘tips’ (i.e., external nodes) of the tree represent sampled HIV-1 *pol* sequences and internal nodes can be viewed as the source of a chain of infections. We refer to all sequences rooted by a common interior node as *descendants* to the so-called *ancestor node*.

In the case of HIV, a *transmission cluster* describes a nonrandom aggregation of sequences from patients believed to share a recent common ancestor [21]. Graphically speaking, a transmission cluster corresponds to a particular branch (or monophyletic clade) in the phylogenetic tree. These transmission clusters are typically ascertained on the basis of high support—measured either by bootstrap percentages or Bayesian posterior probabilities—and sufficiently small genetic distances. One drawback of this procedure is that there is an onus on the user to determine the appropriate support/distance thresholds. In studies involving HIV, these have been reported to range anywhere from 70–99 % for bootstrap values, and 1–4.5 % for the genetic distance cutoff [21]. In addition to being data and user-specific, threshold values can also be affected by the statistical approach used to measure support. Namely, in a formal investigation conducted in [22], posterior probabilities were higher than their corresponding bootstrap values on average. Furthermore, reconstructing phylogenetic trees can be computationally intensive, especially when a large number of sequences are being considered. Despite these shortcomings, phylogenetic analysis has greatly improved our understanding of the epidemic, and remains at the forefront of cluster analysis on HIV sequences.

Herein, we present a new clustering algorithm, called the *Gap Procedure*, for identifying distinct clusters of genetically similar sequences in DNA data. This efficient and automated approach bypasses the need to estimate phylogenies and requires no user-specific threshold values. This distance-based clustering procedure relies on a dissimilarity matrix constructed using popular models for nucleotide substitution and returns a partition of the input data. Gaps in sorted pairwise distances are used to suggest groups of genetically related sequences. The frequency of these groupings are subsequently used to identify phylogenetic clusters. The efficacy and efficiency of the Gap Procedure is demonstrated on both simulated and HIV-1 *pol* sequence data.

## Methods

### Notation

Before discussing the details of our algorithm, we introduce the following notation. Let  $X = (X_1, \dots, X_N)$  be a collection of  $N$  aligned sequences of length  $L$  where  $X_i = (x_{i1}, \dots, x_{iL})$ . The  $j$ th position of the  $i$ th sequence,  $x_{ij}$ , is recorded as an alignment gap ‘-’ or one of the International Union of Pure and Applied Chemistry (IUPAC) codes: A, C, G, T, R, Y, M, K, S, W, H, B, V, D or N [23]. Let  $\mathbb{D}$  be a  $N \times N$  distance matrix whose  $ij$ th element is equal to the genetic distance between sequence  $X_i$  and  $X_j$  denoted by  $d(X_i, X_j)$ . In reference to sequence  $X_i$ , let  $d_i = (d(X_i, X_{[1]}), d(X_i, X_{[2]}), \dots, d(X_i, X_{[N-1]}))$  denote the sorted vector of pairwise distances between  $X_i$  and  $X_j$  (for all  $j \neq i$ ) such that  $d(X_i, X_{[1]}) \leq d(X_i, X_{[2]}) \leq \dots \leq d(X_i, X_{[N-1]})$ . We denote the difference between two adjacent elements in  $d_i$  by  $\delta_{ij} = d(X_i, X_{[j+1]}) - d(X_i, X_{[j]})$ . Finally, we denote a partition of the data by  $\mathcal{M} = \{\mathcal{X}_1, \dots, \mathcal{X}_G\}$  where  $\mathcal{X}_g$  is the set of sequences classified to the  $g$ th group.

### The algorithm

The Gap Procedure is a distance-based clustering algorithm which relies solely on a matrix of pairwise distances. There are a number of freely available packages for R [24] which can be used for evolutionary analysis. For instance, the ape package [25] contains the `dist.dna()` function which can compute a pairwise distance matrix for eleven substitution models; options include Jukes and Cantor 1969 (`jc69`) [26], Kimura 1980 (`K80`) [27] and Tamura and Nei 1993 (`TN93`) [28]. One potential drawback of this function is that it ignores sites with ambiguous nucleotides (i.e., the symbols R through N in the 15 letter IUCPAC nomenclature). Herein, distances are computed using adjusted versions of the `K80` distance formula that allow fractional values for counts on the number of transitional/transversional substitutions per site. These adjusted distances—which we will refer to as `aK80` distances—are described in Additional file 1.

For each individual sequence, the Gap Procedure defines a set of *nearest neighbours* that are subsequently used to determine a partition of the data. To be more specific, let  $c_i = \max\{\delta_{i1}, \delta_{i2}, \dots, \delta_{ik}\}$ , where  $k < N * 0.9$ . In other words,  $c_i$  corresponds to the largest ‘gap’ in the first 90 % of  $d_i$  values. As discussed in Additional file 2, this restriction was established to mitigate the effect of outlying observations. If we define  $k^*$  such that  $\delta_{ik^*} \geq \delta_{ik}$  for all  $k \neq k^*$  and  $d_i^* = d(X_i, X_{[k^*]})$ , then the nearest neighbour matrix  $\mathbb{N} = \{n_{ij}\}$  can be defined as an indicator matrix whose  $ij$ th element is given by

$$n_{ij} = \begin{cases} 1 & \text{if } d(X_i, X_j) \leq d_i^*, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

If  $n_{ij} = 1$ ,  $X_j$  is said to be a neighbour to  $X_i$ . Notice that  $\mathbb{N}$  is not necessarily symmetric. A graphical representation of this definition is provided in Fig. 1. In essence, the number of times observations share a neighbour contribute to their probability of being assigned to the same group. The precise details of this partitioning procedure are codified in Algorithm 1.

---

**Algorithm 1** The Gap Procedure

---

**Inputs:**  $N$  aligned sequences  $X = (X_1, \dots, X_N)$

**Output:** A partition of the data  $\mathcal{M} = \{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_G\}$

---

**Steps**

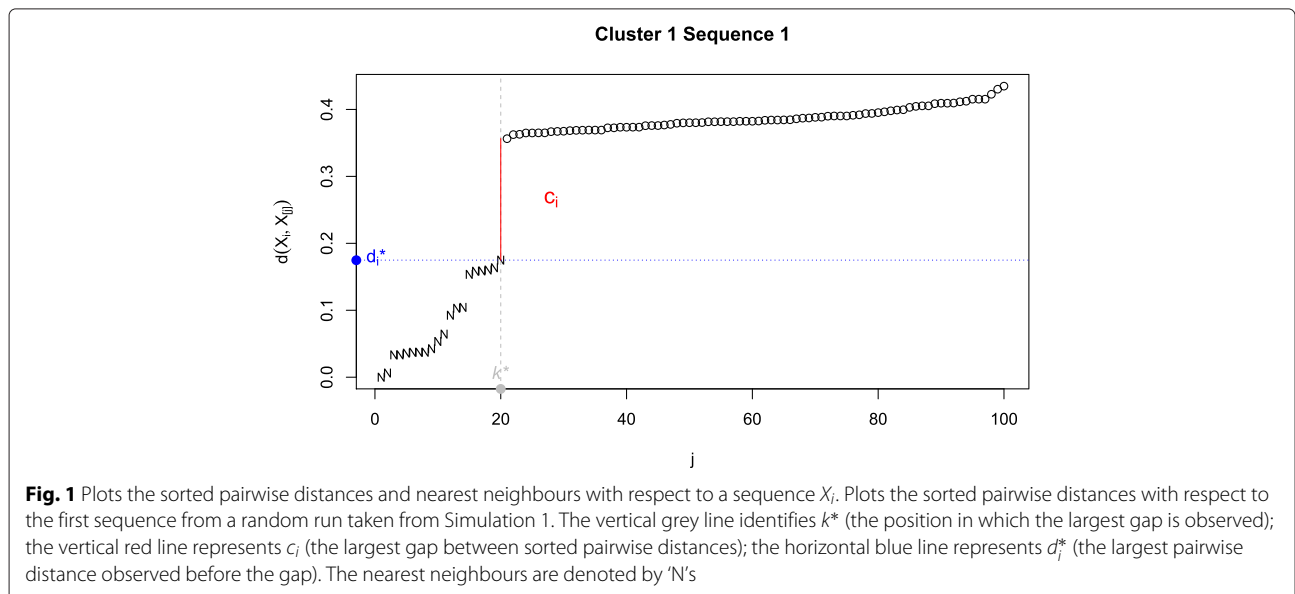
- 1: **procedure** GAP PROCEDURE( $X$ )
  - 2:     Compute the pairwise distance matrix  $\mathbb{D}_{N \times N}$
  - 3:     **for**  $i = 1, \dots, N$  **do**
  - 4:         Obtain the sorted pairwise vectors  $d_i$ .
  - 5:         Find the largest gap,  $c_i = \max\{\delta_{i1}, \delta_{i2}, \dots, \delta_{ik}\}$ , where  $k < N \times 0.9$
  - 6:         Find  $X_i$ 's nearest neighbours, i.e., compute  $\mathbb{N}$  as defined in (1).
  - 7:     **end for**
  - 8:     Construct a unique neighbours matrix  $\mathbb{U}_{M \times N}$  ( $M \leq N$ ) which comprise the unique rows of  $\mathbb{N}$ .
  - 9:     Multiply the rows of  $\mathbb{U}$  by the number of times it is duplicated in  $\mathbb{N}$ ; store it in  $\mathbb{P}_{M \times N} = \{p_{mi}\}$ .
  - 10:     Divide the columns of  $\mathbb{P}_{M \times N}$  by their corresponding sum; store it in  $\mathbb{Z}_{M \times N} = \{z_{mi}\}$ .
  - 11:     **if**  $z_{mi} > z_{m'i} \forall m' \neq m$  **then**
  - 12:         Assign  $X_i$  to cluster  $m$
  - 13: **end procedure**
- 

**Assessing clusters**

The efficacy of the Gap Procedure requires that sequences belonging to the same cluster are sufficiently similar and that the diversity between clusters is sufficiently large. In the literature, there are a variety of validation measures that can be used to test the compactness and separability of clusters, e.g., the Dunn index [29], the Calinski-Harabasz index [30], the C-index [31], the McClain-Rao index [32] and average Silhouettes [33, 34], to name a few. Herein, we assess the within-cluster distances with respect to the  $g$ th cluster, defined as  $S_w(g) = \{d(X_i, X_j) \mid i, j \in \mathcal{X}_g, i < j\}$  and the corresponding between-cluster set,  $S_b(g) = \{d(X_i, X_j) \mid i \in \mathcal{X}_g, j \notin \mathcal{X}_g, i < j\}$ . As a general guideline, we suggest that the Gap Procedure only be used when the 25 percentile of  $S_b(g)$  is larger than the 75 percentile of  $S_w(g)$  for all  $g = 1, \dots, G$ . Graphically speaking, the side-by-side boxplots of  $S_w(g)$  and  $S_b(g)$  should display little to no overlap for each group found using the Gap Procedure (see Additional file 3 for examples). Future work will aim at determining if numerical validation measures, such as the indices mentioned above, can be used in place of this visual diagnostic.

**Implementation and availability**

The Gap Procedure algorithm can be implemented using the GapProcedure package available on GitHub (<https://github.com/vrbiki/GapProcedure>). This R package includes functions for plotting the side-by-side boxplots mentioned in Section “Assessing clusters” as well as a vignette providing a step-by-step description of the algorithm and a quick demonstration. The GapProcedure package has been tested on Mac, Linux, and Windows.



## Results and discussion

In this section, we compare the results of the Gap Procedure with those obtained by the gold-standard phylogenetic threshold approach. In the threshold approach, clusters are ascertained on the basis of high clade support and low genetic distance. More precisely, given the topology of a phylogenetic tree, sequences are clustered together only if: (a) they belong to the same clade, (b) clade support (bootstrap values or posterior probabilities) exceeds  $T_c$ , and (c) the maximum within-cluster pairwise genetic distance is below  $T_d$ . As mentioned previously, the exact values of  $T_c$  and  $T_d$  vary between analyses. Clustering results are assessed using the Adjusted Rand Index (ARI) which measures the agreement between two partitions while accounting for chance [35]. In this particular analysis, a value of 1 corresponds to perfect agreement with the ‘true’ (i.e., simulated or expert-verified) clusters, whereas a value of 0 would be expected if clusters have been assigned at random.

Herein, phylogenetic trees were estimated using Randomized Axelerated Maximum Likelihood (RAxML) [19] and MrBayes [20]. RAxML is a program for inferring maximum likelihood trees with bootstrap support values whereas MrBayes performs a Bayesian analysis and produces summary trees with posterior probabilities. RAxML was implemented using the GTR +  $\Gamma$  model with 20 maximum likelihood searches and 100 bootstrap replicates (see RAxML manual for details). MrBayes was executed using the GTR + I +  $\Gamma$  substitution model and run until the average standard deviation of split frequencies—the statistic used by MrBayes to monitor convergence—dropped below a value of 0.01. For notational convenience, we refer to the clusters obtained using the threshold approach on the respective trees as ‘RAxML clusters’ and ‘MrBayes clusters.’ Although we used in-house code to implement the threshold approach, RAxML/MrBayes clusters could also be extracted using a program such as ClusterPicker [21]. All figures in this section were produced in R [24].

### Simulation studies

Using the `seqgen()` function available in the `phyclust` package [13], data were simulated by mutating DNA sequences along phylogenetic trees. The topology of the trees were generated at two stages *via* the `ms` program [13, 36]; for more details see Additional file 4. Sequences were mutated according to a General Time Reversible model which assumed rate heterogeneity and a proportion of invariable sites, i.e., the GTR + I +  $\Gamma$  model. For our simulation, sequences of length 800 were generated along trees made up of 4, 6, 20, or 50 transmission groups comprised of roughly 25 sequences per group<sup>1</sup>. For each  $G$ -group simulation, 100 random trees—thus 100 random data sets—were generated. Further simulations involving

tree topologies different than the ones considered herein are explored in Additional file 5.

When applied to the simulated data, the Gap Procedure achieved close to perfect classification (see Table 1). Accordingly, the average ARI values were close to 1 and the average number of clusters—where ‘clusters’ are defined to contain two or more members—roughly equalled the number of generated transmission clades ( $G$ ). As expected, the average number of singletons, i.e., unclustered sequences, was close to 0 for all simulations. Aside from achieving excellent clustering results, the average computation time was less than a second for Simulations 1, 2, and 3, and less than 7 seconds for Simulation 4. Note that the analysis was performed using an Intel Xeon E5-1650 (3.5 GHz) processor and includes the calculation of the pairwise distance matrix.

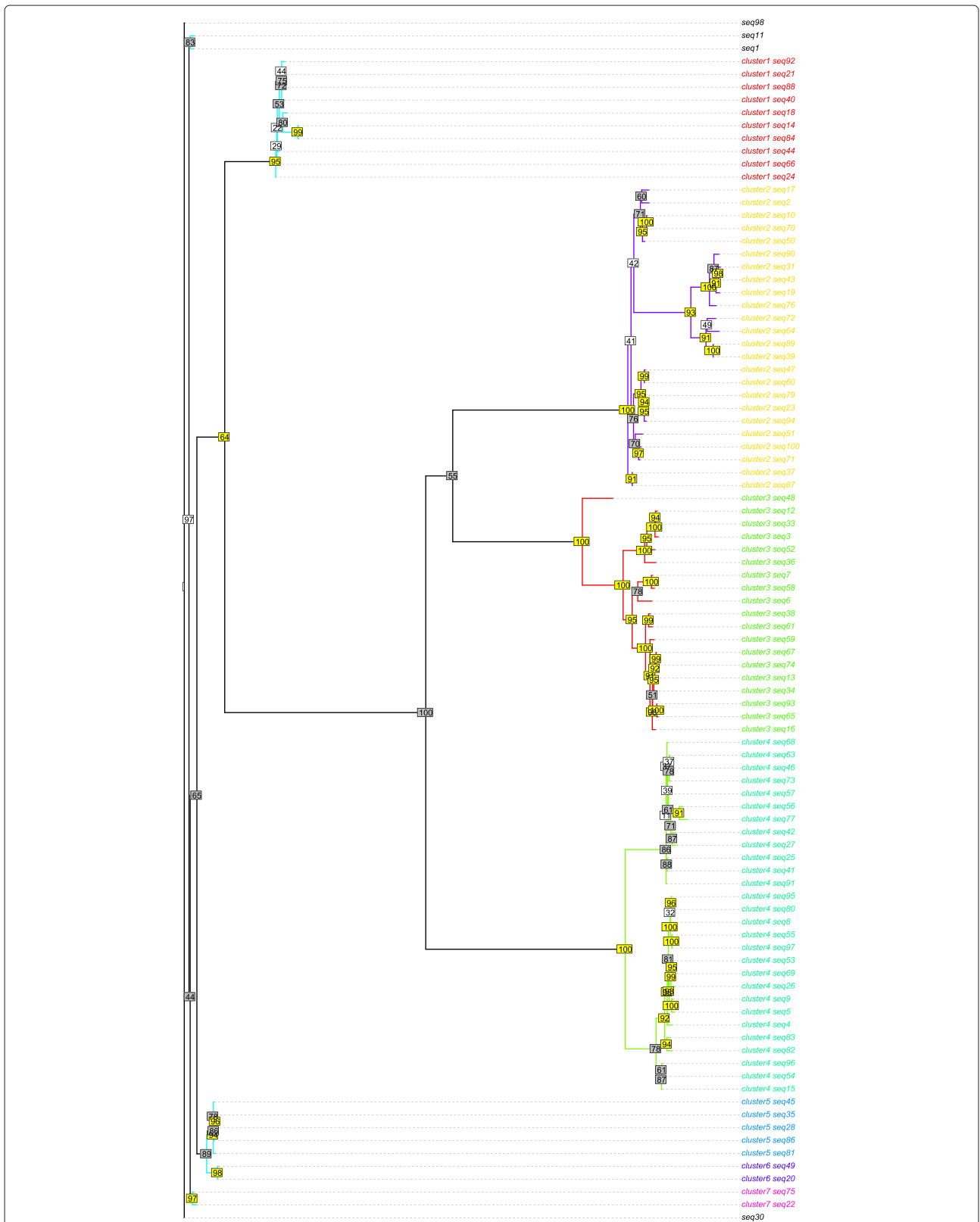
The results presented here correspond to RAxML and MrBayes clusters extracted using a clade support threshold ( $T_c$ ) of 90 % and a distance threshold ( $T_d$ ) of either 0.3 or 0.6 (the corresponding results for  $T_d = 0.4, 0.5$  are given in Additional file 6). Due to the computational complexity of these procedures, the results were based on a single run. The graphical representation of the RAxML and MrBayes clusters for Simulation 1 are shown in Figs. 2 and 3, respectively. High ( $\geq 90$ ), medium (50–90) and low ( $< 50$ ) clade credibility values are denoted by yellow, grey and white rectangles, respectively. The true (i.e., simulated) transmission clusters are designated using coloured edges whereas clusters found by the threshold procedure are provided using coloured tip labels (singletons are written in black). Note that the results for RAxML contain fewer observations, since the algorithm requires that duplicated sequences be removed before analysis.

Tables 2 and 3 summarize the results obtained using the threshold approach on the phylogenetic trees estimated using RAxML and MrBayes, respectively. As the results indicate, a distance threshold value of 0.3 was inadequate to capture the simulated clusters and resulted in poor ARI values. However, when  $T_d$  was raised to a value of 0.6, the

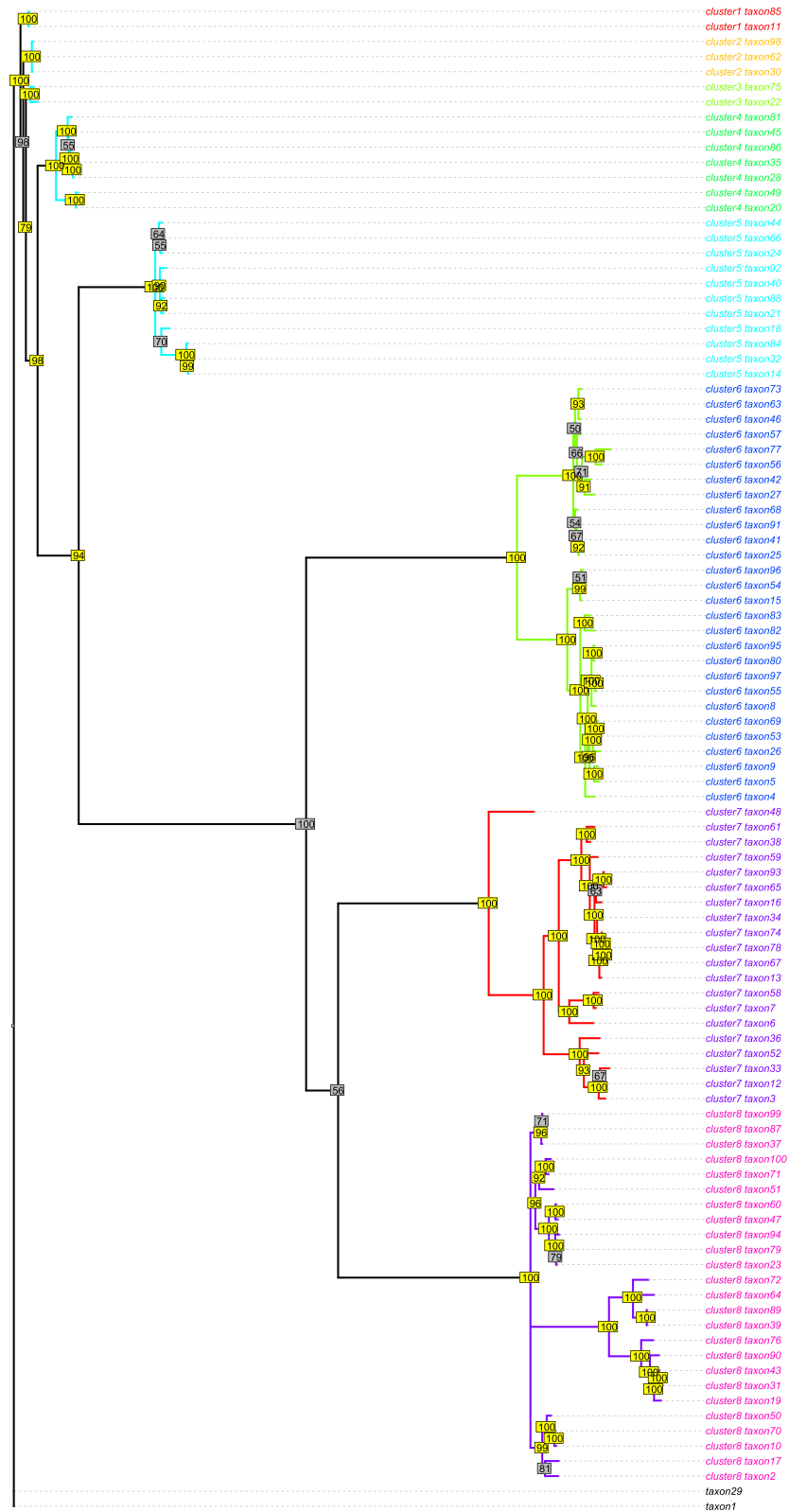
**Table 1** Clustering results for the Gap Procedure on simulated data

Sim	Data		Average			
	$N$	$G$	Time (in sec)	# clusters	# singletons	ARI
1	100	4	0.1108	4.25	0.04	0.9854
2	150	6	0.1370	6.39	0.04	0.9856
3	500	20	0.6073	22.49	0.13	0.9750
4	1250	50	6.6194	58.11	0.43	0.9694

The average clustering results (taken over 100 runs) obtained by the Gap Procedure when applied to the simulated data. The dissimilarity matrix was calculated using the `aR80` distance formula and sequences (of length 800) were mutated according to a GTR + I +  $\Gamma$  model



**Fig. 2** Clustering Results for RAxML with  $T_c = 90$ ,  $T_d = 0.6$ . The maximum likelihood phylogenetic tree ( $n = 94$ ) produced by RAxML for Simulation 1. High, ( $\geq 90$ ) medium (50–90) and low (< 50) bootstrap values are denoted by yellow, grey and white rectangles, respectively. Cluster indices are represented by coloured tip labels; singletons are denoted in black



**Fig. 3** Clustering Results for MrBayes with  $T_c = 90$ ,  $T_d = 0.6$ . The maximum likelihood phylogenetic tree ( $n = 100$ ) produced by RAxML for Simulation 1. High, ( $\geq 90$ ) medium (50–90) and low ( $< 50$ ) posterior probabilities are denoted by yellow, grey and white rectangles, respectively. Cluster indices are represented by coloured tip labels; singletons are denoted in black

**Table 2** Clustering results for RAxML on simulated data

	Sim	$T_c$	$T_d$	Time (in sec)	# clusters	# singletons	ARI
RAxML	1	90	0.3	2479.0	13	21	0.3662
	2	90	0.3	4654.0	13	10	0.7054
	3	90	0.3	41584.6	61	33	0.6206
	4	90	0.3	271593.7	167	70	0.4889
	1	90	0.6	2479.0	7	4	0.8757
	2	90	0.6	4654.0	9	5	0.8945
	3	90	0.6	41584.6	24	6	<b>0.9764</b>
	4	90	0.6	271593.7	54	2	<b>0.9922</b>

The clustering results (for a single run) obtained by RAxML when applied to the simulated data. The quoted run times represent the time it takes RAxML to produce a phylogenetic tree and obtain clade support values (conducted using 100 bootstrap replicates). RAxML clusters are obtained using a clade support threshold equal to  $T_c$  and distance thresholds of  $T_d$ . The ARI scores in bold indicate which runs performed better than the average score obtained using the Gap Procedure

ARI score improved dramatically and both methods outperformed the Gap Procedure in some cases. Although potential gain in ARI values can be achieved by adopting the threshold approach, we remark that the corresponding efficacy depends greatly on the user-specified  $T_d$  values. Furthermore, the programs used for estimating phylogenetic trees take substantially longer to run than the Gap Procedure. For instance, the trees produced for Simulation 4 took 3 days 3 hours 26 minutes for RAxML to complete and over 15 days for MrBayes. In stark contrast, the complete analysis for the Gap Procedure took an average of 6.6 seconds. Note that the exact gain in computational time will vary depending on the data.

### Quebec HIV-1 *pol* sequence data

Multiple studies have been conducted to improve our understanding of the HIV transmission dynamics in

**Table 3** Clustering results for MrBayes on simulated data

	Sim	$T_c$	$T_d$	Time (in sec)	# clusters	# singletons	ARI
MrBayes	1	90	0.3	3324.7	13	3	0.4642
	2	90	0.3	4243.6	19	7	0.5129
	3	90	0.3	144284.8	54	11	0.6565
	4	90	0.3	1328253.9	134	25	0.6269
	1	90	0.6	3324.7	8	2	0.8419
	2	90	0.6	4243.6	10	3	0.9011
	3	90	0.6	144284.8	24	6	<b>0.9768</b>
	4	90	0.6	1328253.9	52	3	<b>0.9927</b>

The clustering results (for a single run) obtained by MrBayes when applied to the simulated data. The quoted run times represent the time it takes MrBayes to estimate a phylogenetic tree with clade support (i.e., posterior probability) values. MrBayes clusters are obtained using a clade support threshold equal to  $T_c$  and distance thresholds of  $T_d$ . The ARI scores in bold indicate which runs performed better than the average score obtained using the Gap Procedure

Quebec [1–3]. Through the molecular surveillance of HIV-1 *pol* sequence data, researchers were able to link high rates of onward transmission to acute/early infection. Phylogenetic analysis was performed using maximum likelihood methods *via* BioEdit [37] and MEGA2 [38]. High bootstrap values (> 98 %) and sufficiently long branches on neighbour-joining (NJ) trees [12] were used to determine cluster membership. Manual inspection of polymorphisms and mutational motifs were used to validate clusters.

This section aims at rediscovering these transmission clusters using the automated Gap Procedure. The data, which was obtained from the provincial genotypic testing program (introduced in 1997) and the Quebec PHI cohort (established in 2001), has since been expanded to include 1517 sequences, each of length 810. Ethical approval for this cohort was granted by the Laboratoire de santé publique du Québec, and the Quebec Ministry of Health committee on confidentiality and access of information. Several subsets of this data (summarized in Table 4) were considered for the analysis. To test the efficacy of our approach, the resulting clusters were compared with those obtained by Brenner and colleagues. We henceforth refer to the latter as the ‘true’ or ‘gold-standard’ clusters.

Table 5 summarizes the clustering results obtained by the Gap Procedure when applied to the six subsets of the HIV-1 data summarized in Table 4. The number of correctly and incorrectly identified singletons are given under the heading “1 ✓” and “1 ✗”, respectively. The adjacent columns provide the total number of small clusters (2–4 members) and big clusters ( $\geq 5$  members); the total number of sequences belonging to the corresponding heading is reported in parentheses. Note that true singletons were removed before the ARI scores were calculated.

As indicated by the high ARI scores, there is a strong agreement between the true clusters and those found using our approach. In terms of cluster size, the Gap Procedure experienced some difficulty in distinguishing between small clusters and singletons. Consequently, when compared with the gold-standard, our approach found a greater number of small (2–4 member) groups. Despite this discrepancy, the Gap Procedure did well in identifying big ( $\geq 5$  member) clusters and obtained an ARI greater than 0.9 on all data sets considered.

In addition to its excellent clustering performance, this algorithm was extremely fast when compared with the competing approaches. For example, the Gap procedure took less than a second to run on the mibc data ( $N = 627, L = 810$ ). To produce the phylogenetic trees for the same data, RAxML and MrBayes took roughly 15 and 126 hours, respectively. In terms of clustering performance, the results of RAxML and MrBayes were highly variable (for a complete summary see Additional file 7). Using a range of threshold values, the ARIs produced by

**Table 4** A summary of the subset data taken from the HIV-1 sequence data

Name	Description	N	G <sup>b</sup>	Cluster size		
				1	2–4	≥ 5
all	Entire set	1517	169	533	108 (311)	61 (673)
men	Only males	1391	152	488	96 (276)	56 (627)
non.sing	Clustered sequences	984	169	0	108 (311)	61 (673)
nsm	Clustered males	903	152	0	96 (276)	56 (627)
big <sup>a</sup>	Sequences clustered to big <sup>a</sup> clusters	673	61	0	0 (0)	61 (673)
mibc	Males clustered to big <sup>a</sup> clusters	627	56	0	0 (0)	56 (627)

The total number of small and large-sized clusters are listed under the headings (2–4) and (≥ 5). The corresponding number of sequences belonging to each heading is given in parenthesis

<sup>a</sup>big<sup>a</sup> clusters are defined to have ≥ 5 members

<sup>b</sup>the number of clusters having ≥ 2 members

RAxML ranged anywhere from 0.0081 (with  $T_d = 0.01$  and  $T_c = 99$ ) to 0.8977 (with  $T_d = 0.07, 0.08, 0.09$  or  $0.1$  and  $T_c = 90$ ). For MrBayes clusters, the ARI ranged from as low as 0.0123 (with  $T_d = 0.01$  and  $T_c = 98$ ) to as high as 0.9969 (with  $T_d = 0.09$  and  $T_c = 99$ ).

**Conclusion**

A distance-based clustering algorithm for genetic HIV-1 sequence data has been presented. Unlike the competing threshold approach, the Gap Procedure is fully automated (i.e., it does not require any user-specific threshold values) and relies solely on pairwise distances. Results were obtained using the GapProcedure package wherein pairwise distances are calculated using adjusted K80 distance formula. Although this is the default setting of the algorithm, alternative dissimilarity matrices may be used in its place.

When compared with RAxML and MrBayes, our algorithm showed dramatic gains in computational time, owing greatly to the fact that it bypasses the construction of a phylogenetic tree. The resulting gains in efficiency supports the quick analysis of large genetic data sets. When applied to both simulated and HIV-1 *pol* sequence data, the Gap Procedure uncovered clusters that closely agreed with true or expert-verified clusters. These

encouraging results suggest that burdensome procedures involving the estimation of phylogenetic trees may not be required to infer distinct clusters of genetically similar DNA sequences.

**Endnote**

<sup>1</sup>Group membership is assigned according to a multinomial distribution.

**Additional files**

**Additional file 1: Describes the details for computing adjusted JC69 (aJC69) and adjusted K80 (aK80) distances. Figure S1.** Side-by-side boxplots comparing K80 and aK80 distances. (PDF 228 kb)

**Additional file 2: Details regarding the Gap Procedure’s scheme for combating the effects of outliers. Figure S1.** Plots the sorted pairwise distances for a 3-group simulation. **Figure S2.** Plots the sorted pairwise distances for a 3-group simulation with an outlier. (PDF 109 kb)

**Additional file 3: Side-by-side boxplots. Figure S1.** Boxplots for the within and between cluster pairwise distances for Simulation 1. **Figure S2.** Plots the sorted pairwise distances and nearest neighbours for Simulation 1. **Figure S3.** Boxplots for the within and between cluster pairwise distances for the mibc data. **Figure S4.** Plots the sorted pairwise distances and nearest neighbours for the mibc data. (PDF 195 kb)

**Additional file 4: Describes how to construct phylogenetic trees used for generating sequence data. Figure S1.** Plots the step-by-step construction a 6-group phylogenetic tree. (PDF 147 kb)

**Additional file 5: Performs simulation studies involving a variety of tree topologies. Figure S1.** The clustering results obtained by the Gap Procedure (ARI = 0.4489) when applied to a 4-group simulation with a star-like ancestor tree. **Figure S2.** The clustering results obtained by the Gap Procedure (ARI = 0.5994) when applied to a 4-group simulation with a random bifurcating ancestor tree. **Figure S3.** The clustering results obtained by the Gap Procedure (ARI = 0.4066) when applied to a 4-group simulation with a random bifurcating ancestor tree. **Figure S4.** The clustering results obtained by the Gap Procedure (ARI = 0.6152) when applied to a 4-group simulation with a random bifurcating ancestor tree. **Table S1.** The clustering results obtained by the Gap Procedure on Simulation 1 with varying ancestor/descendant tree heights. **Table S2.** The clustering results for the Gap Procedure on Simulation 1 while relaxing the star-phylogeny assumption. (PDF 182 kb)

**Additional file 6: Cluster results for the RAxML and MrBayes clusters on simulated data with  $T_d = 0.4, 0.5$ . Table S1.** Cluster results for the RAxML clusters on simulated data with  $T_d = 0.4, 0.5$ . **Table S2.** Cluster results for the MrBayes clusters on simulated data with  $T_d = 0.4, 0.5$ . (PDF 114 kb)

**Table 5** Clustering results for the Gap Procedure on HIV-1 data

Subset	Time (in sec)	1 ✓	1 ✗	2–4	≥ 5	ARI
all	10.56 sec	237	16	244 (619)	61 (645)	0.9170
men	8.261 sec	225	18	215 (536)	60 (612)	0.9097
non.sing	3.086 sec	–	12	125 (351)	57 (621)	0.9325
nsm	2.470 sec	–	11	109 (303)	54 (589)	0.9320
big	0.807 sec	–	3	5 (14)	61 (656)	0.9523
mibc	0.634 sec	–	3	5 (14)	56 (610)	0.9492

The ARI scores and running times of the Gap Procedure when performed on subsets of the HIV-1 data. The number of correctly (and incorrectly) identified singletons are listed under “1 ✓” (and “1 ✗”). The total number of for small and large-sized are listed under the headings (2–4) and (≥ 5). The corresponding number of sequences belonging to each class is given in parentheses



**Additional file 7: Cluster results for the RAxML and MrBayes clusters on the mibc data. Table S1.** The threshold values and corresponding clustering results for RAxML on the mibc data. **Table S2.** The threshold values and corresponding clustering results for MrBayes on the mibc data. (PDF 108 kb)

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

IV designed and implemented the algorithm, and drafted the manuscript. DS helped develop the concept and supervised the study at all stages. MR provide parts of the data, and BB organized the data and validated the phylogenetic clusters used for comparison in this study. All authors have read and approved the final manuscript.

### Acknowledgements

This work was supported by grants from the Canadian Institutes of Health Research (CIHR HHP-126781).

### Author details

<sup>1</sup>Department of Mathematics & Statistics, McGill University, 805 Sherbrooke Street West, H3A 0B9 Montreal, Canada. <sup>2</sup>Centre de recherche du CHUM, 900 rue Saint-Denis Tour Viger, H2X 0A9 Montreal, Canada. <sup>3</sup>Faculty of Surgery (Medicine), McGill University, 1010 Sherbrooke Street West, H3A 2R7 Montreal, Canada. <sup>4</sup>McGill University AIDS Centre, Jewish General Hospital, Côte-Sainte-Catherine, H3T1E2 Montreal, Canada.

Received: 14 May 2015 Accepted: 22 October 2015

Published online: 04 November 2015

### References

- Brenner BG, Roger M, Routy JP, Moisi D, Ntemgwana M, Matte C, et al. High rates of forward transmission events after acute/early HIV-1 infection. *J Infect Dis*. 2007;195(7):951–9.
- Brenner BG, Roger M, Moisi DD, Oliveira M, Hardy I, Turgel R, et al. Transmission networks of drug resistance acquired in primary/early stage HIV infection. *AIDS (London, England)*. 2008;22(18):2509.
- Brenner BG, Roger M, Stephens D, Moisi D, Hardy I, Weinberg J, et al. Transmission clustering drives the onward spread of the HIV epidemic among men who have sex with men in quebec. *J Infect Dis*. 2011;204(7):1115–9.
- Hué S, Pillay D, Clewley JP, Pybus OG. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc Natl Acad Sci U S A*. 2005;102(12):4425–429.
- Pao D, Fisher M, Hué S, Dean G, Murphy G, Cane PA, et al. Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. *Aids*. 2005;19(1):85–90.
- Ragonnet-Cronin M, Ofner-Agostini M, Merks H, Pilon R, Rekart M, Archibald CP, et al. Longitudinal phylogenetic surveillance identifies distinct patterns of cluster dynamics. *JAIDS J Acquir Immune Defic Syndr*. 2010;55(1):102–8.
- Chalmet K, Staelens D, Blot S, Dinakis S, Pelgrom J, Plum J, et al. Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. *BMC infect dis*. 2010;10(1):262.
- Hué S, Brown AE, Ragonnet-Cronin M, Lycett SJ, Dunn DT, Fearnhill E, et al. Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions. *AIDS*. 2014;28(13):1967–75.
- Lubelchek RJ, Hoehnen SC, Hotton AL, Kincaid SL, Barker DE, French AL. Transmission clustering among newly diagnosed HIV patients in Chicago, 2008 to 2011: Using phylogenetics to expand knowledge of regional HIV transmission patterns. *JAIDS J Acquir Immune Defic Syndr*. 2015;68(1):46–54.
- Dondoshansky I, Wolf Y. Blastclust . Bioinformatics Toolkit, Max-Planck Institute for Developmental Biology. 2008–2015. <http://toolkit.tuebingen.mpg.de/blastclust>.
- Sokal RR. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*. 1958;38:1409–38.
- Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4(4):406–25.
- Chen WC, Dorman K. phyclus: Phylogenetic Clustering (Phyloclustering). 2010. R package, <http://cran.r-project.org/package=phyclus>.
- Baxevanis AD, Ouellette BF. Bioinformatics: a Practical Guide to the Analysis of Genes and Proteins. New Jersey, USA: John Wiley & Sons; 2004. vol. 43.
- Swofford DL. PAUP\*. phylogenetic analysis using parsimony (and other methods). version 4. 2003. Sunderland Massachusetts: Sinauer Associates.
- Simon D, Larget B. Bayesian analysis in molecular biology and evolution (BAMBE), version 4.01a; 1999–2012. <http://en.bio-soft.net/tree/BAMBE.html>.
- Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evol Biol*. 2007;7(1):214.
- Plotree D, Plotgram D. PHYLIP-phylogeny inference package (version 3.2). *Cladistics*. 1989;5:163–6.
- Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma*. 2014;30(9):1312–3.
- Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 2012;61(3):539–42.
- Ragonnet-Cronin M, Hodcroft E, Hué S, Fearnhill E, Delpuch V, Brown AJ, et al. Automated analysis of phylogenetic clusters. *BMC bioinforma*. 2013;14(1):317.
- Alfaro ME, Zoller S, Lutzoni F. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol Biol Evol*. 2003;20(2):255–66.
- Cornish-Bowden A. IUPAC-IUB symbols for nucleotide nomenclature. *Nucleic Acids Res*. 1985;13:3021–30.
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. <https://www.R-project.org/>.
- Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. *Bioinforma*. 2004;20:289–90.
- Jukes TH, Cantor CR. Evolution of protein molecules. *Mamm Protein Metab*. 1969;11:21–132.
- Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J mol evol*. 1980;16(2):111–20.
- Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*. 1993;10(3):512–26.
- Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*. 1973;3(3):32–57.
- Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat theory Methods*. 1974;3(1):1–27.
- Hubert L, Schultz J. Quadratic assignment as a general data analysis strategy. *Br J Math Stat Psychol*. 1976;29(2):190–241.
- McClain JO, Rao VR. Clustis: A program to test for the quality of clustering of a set of objects. *J Mark Res*. 1975;12:456–60.
- Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53–65.
- Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis* vol. 344. New Jersey, USA: John Wiley & Sons; 2009.
- Hubert L, Arabie P. Comparing partitions. *J Classif*. 1985;2:193–218.
- Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinforma*. 2002;18(2):337–8.
- Hall TA. Bioedit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In: *Nucleic acids symposium series*; 1999. p. 95–98. Distributed by the author, website: [www.mbio.ncsu.edu/BioEdit/bioedit.html](http://www.mbio.ncsu.edu/BioEdit/bioedit.html).
- Kumar S, Tamura K, Jakobsen IB, Nei M. MEGA2: molecular evolutionary genetics analysis software. *Bioinforma*. 2001;17(12):1244–5.