



QS-Net: Reconstructing Phylogenetic Networks Based on Quartet and Sextet

Ming Tan^{1†}, Haixia Long^{2†}, Bo Liao^{1,2*}, Zhi Cao^{1*}, Dawei Yuan³, Geng Tian³, Jujuan Zhuang⁴ and Jialiang Yang^{2,5*}

¹ College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, ² School of Information Science and Technology, Hainan Normal University, Haikou, China, ³ Geneis (Beijing) Co. Ltd., Beijing, China, ⁴ Department of Mathematics, Dalian Maritime University, Dalian, China, ⁵ Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, United States

OPEN ACCESS

Edited by:

Marco Antonioti,
University of Milano-Bicocca, Italy

Reviewed by:

Gianluca Della Vedova,
University of Milano-Bicocca, Italy
Mohammed El-Kebir,
University of Illinois at Urbana-
Champaign, United States

*Correspondence:

Bo Liao
dragonbw@163.com
Zhi Cao
66384436@qq.com
Jialiang Yang
jialiang.yang@mssm.edu

[†]These authors have contributed
equally to this work.

Specialty section:

This article was submitted to
Bioinformatics and
Computational Biology,
a section of the journal
Frontiers in Genetics

Received: 24 September 2018

Accepted: 11 June 2019

Published: 24 July 2019

Citation:

Tan M, Long H, Liao B, Cao Z,
Yuan D, Tian G, Zhuang J and Yang J
(2019) QS-Net: Reconstructing
Phylogenetic Networks Based
on Quartet and Sextet.
Front. Genet. 10:607.
doi: 10.3389/fgene.2019.00607

Phylogenetic networks are used to estimate evolutionary relationships among biological entities or taxa involving reticulate events such as horizontal gene transfer, hybridization, recombination, and reassortment. In the past decade, many phylogenetic tree and network reconstruction methods have been proposed. Despite that they are highly accurate in reconstructing simple to moderate complex reticulate events, the performance decreases when several reticulate events are present simultaneously. In this paper, we proposed QS-Net, a phylogenetic network reconstruction method taking advantage of information on the relationship among six taxa. To evaluate the performance of QS-Net, we conducted experiments on three artificial sequence data simulated from an evolutionary tree, an evolutionary network involving three reticulate events, and a complex evolutionary network involving five reticulate events. Comparison with popular phylogenetic methods including Neighbor-Joining, Split-Decomposition, Neighbor-Net, and Quartet-Net suggests that QS-Net is comparable with other methods in reconstructing tree-like evolutionary histories, while it outperforms them in reconstructing reticulate events. In addition, we also applied QS-Net in real data including a bacterial taxonomy data consisting of 36 bacterial species and the whole genome sequences of 22 H7N9 influenza A viruses. The results indicate that QS-Net is capable of inferring commonly believed bacterial taxonomy and influenza evolution as well as identifying novel reticulate events. The software QS-Net is publically available at <https://github.com/Tmyiri/QS-Net>.

Keywords: phylogenetic network, reticulate evolution, sextet, bacterial taxonomy, influenza reassortment

INTRODUCTION

Phylogenetic tree is usually utilized to show the evolutionary history of a set of biological entities or taxa. However, the tree-like topology cannot represent reticulate evolutionary events, such as horizontal gene transfer (HGT), hybridization, recombination, or reassortment, which have been shown to be critical in genotypic diversity, related phenotypes, estimations of evolutionary history, and virus emergence and immune evasion (Fenderson and Bruce, 2008; Vijaykrishna et al., 2015; Bastide et al., 2018). For example, HGT, also known as lateral gene transfer (LGT), promotes the diversification of microorganisms on the evolutionary time scale. This mechanism

can change the types and characteristics of bacteria and plays a major role in the genetic diversity of bacteria (Ochman et al., 2000). In the long run, it may be the dominant force affecting genes in most prokaryotes. Recombination is a major source of genotypic diversity and a core force for the formation of genome and related phenotypes (Leducq et al., 2017). Reassortment is responsible for most antigenic shifts of influenza virus (Nelson et al., 2008). Hybridization has been shown to be the main evolutionary mechanism for plants and some animals (Rieseberg et al., 2000; Yu et al., 2011).

A phylogenetic network can serve as an alternative to phylogenetic tree. When the evolutionary history of a sequence set contains reticulate events (Huson et al., 2010), generally speaking, phylogenetic networks can be divided into explicit and implicit networks. The implicit phylogenetic networks, such as split network, are often adopted to illustrate incompatible data and capture conflicting signals in a data set. With the increasing sequencing data, phylogenetic networks have become more and more important in molecular evolution.

Over the past decades, many methods have been proposed for reconstructing phylogenetic trees or networks. The most common type of method reconstructs a network directly from the original character data, usually through a parsimony or maximum-likelihood criterion. Methods in this category include Spectronet (Huber et al., 2002), maximum pseudo-likelihood estimation (Yu and Nakhleh, 2015), HGT maximum parsimony (Park et al., 2010), PhyloNetwork (Solís-Lemus et al., 2017), inferring phylogenetic networks using PhyloNet (Wen et al., 2018), and SNaQ (Claudia and Cécile, 2016). However, these methods are inefficient computationally and tend to overestimate the actual number of reticulate events in the evolutionary history (Huelsenbeck, 1995; Park et al., 2010). The second widely used method is the distance-based method, which first builds a genetic distance matrix for a taxa set and then reconstructs the phylogenetic network from the distance matrix. Methods in this category include Neighbor-Net (Bryant and Moulton, 2004), Split-Decomposition (Bandelt and Dress, 1992), FastME (Lefort et al., 2015), ASTRID (Vachaspati and Warnow, 2015), tree-average distances method (Willson, 2013), and large-scale Neighbor-Joining with NINJA (Wheeler, 2009). The distance-based methods are very fast compared with character-based methods, but they have a disadvantage in terms of reconstruction accuracy. The third kind of methods reconstructs phylogenetic networks from weighted triplets and quartets because they can retain more information than distances. Methods in this category include local maximum likelihood using triplets (Ranwez and Gascuel, 2002), Quartet-Net (Yang et al., 2013), tree with strong combinatorial evidence (Berry and Gascuel, 2000), QNet (Grünwald et al., 2007), SuperQ (Grünwald et al., 2013), DistiQue (Sayyari and Mirarab, 2016), level 1 network from a dense quartet (Keijsper and Pendavingh, 2014), and weighted QMC (Avni et al., 2015). In addition, there are other methods using statistical models such as stochastic local search method (Tria et al., 2010), clusters (Van Iersel et al., 2010), Bayesian inference (Zhang et al., 2017), statistical model (Pickrell and Pritchard, 2012), and Monte Carlo method (Eslahchi et al., 2010).

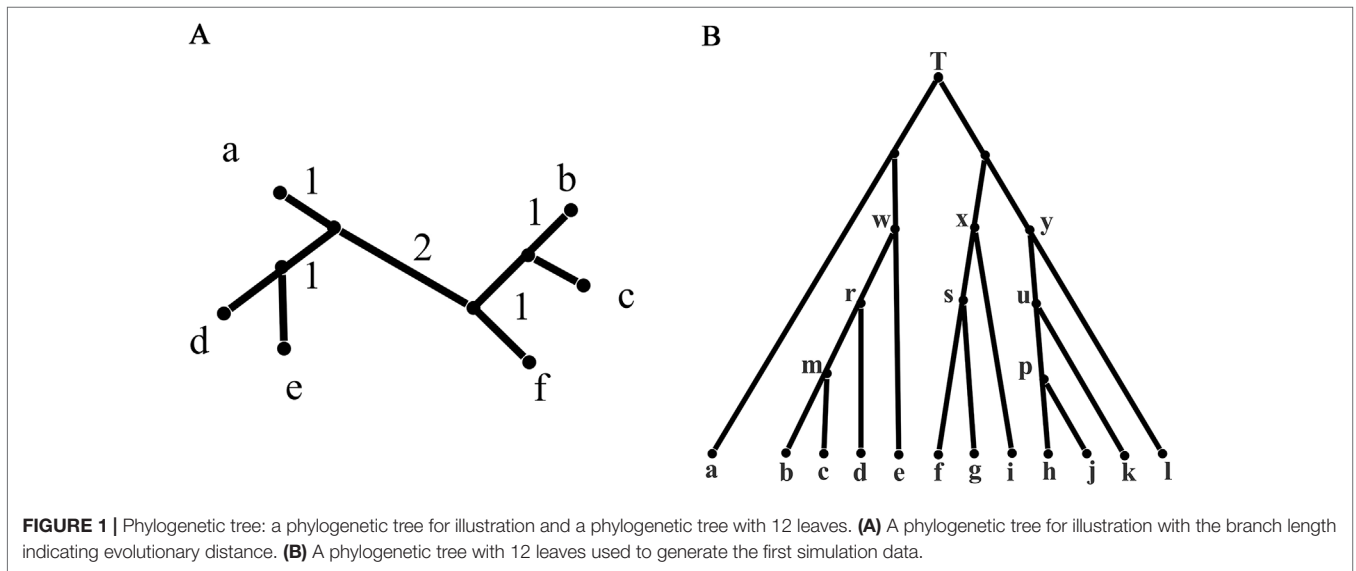
Quartet-Net (Yang et al., 2013) is a method for reconstructing phylogenetic networks from a set of weighted triplets and quartets, which uses parsimony information sites to calculate triplet and quartet weights directly from multiple sequence alignment (MSA). Based on the calculated triplet and quartet weights, Quartet-Net then performs a split expanding process to obtain all full splits and their weights, which will transform to an evolutionary tree or network. The method is a generalization of Split-Decomposition (Bandelt and Dress, 1992). In this paper, we further generalize Quartet-Net and propose a novel method called QS-Net to reconstruct evolutionary networks based on weighted quartets and sextets. The analysis of artificial and real data sets shows that this method can reconstruct a more accurate phylogeny when the sequence data are generated from complicated evolutionary scenarios involving many reticulate events and identifies novel reticulate evolution and reassortment events.

MATERIALS AND METHODS

Background: Split and Split Weight

For a taxa set $S = \{S_1, S_2, \dots, S_n\}$ of size n , a split consisting of two disjoint non-empty subsets of S is denoted by $A | B$ that is, A and B . If A and B contain all the taxa in S , then $A | B$ is called a full split; otherwise, it is called a partial split. In a phylogenetic tree, each edge is a full split that divides the tree into two parts, while in a phylogenetic network, a group of parallel edges with equal length represents a full split. If $|A| = 1$ or $|B| = 1$, the split $A|B$ is called a trivial split. For example, the phylogenetic tree in **Figure 1A** contains five trivial full splits, such as $a|bcdef$, and three non-trivial full splits $d|abcf$, $bc|adef$, and $ade|bcf$. In general, a split $A|B$ with $|A| = m$ and $|B| = n$ is called an $m|n$ split. In addition, $W(A|B)$ represents the evolutionary distance between taxa groups A and B . If A or B contains more than two taxa, then $W(A|B)$ calculates the distance between the common ancestor of A and B . For example, $W(a|de) = 2$, $W(d|ae) = 1$ in **Figure 1A**, $W(a|d)$ represents the evolutionary distance between taxa a and d , and therefore, through **Figure 1A** and these definitions, we can get this equation $W(a|d) = W(a|de) + W(ae|d)$.

For an MSA, a simple parsimony-based method is used to estimate the weights of quartets and sextets. For example, if the character in a site is the same for taxa a , b , and c and for taxa d , e , and f , but different for a and d , then the site is defined to support the split $abc | def$. For any sextet $abc | def$, its weight $W(abc|def)$ is defined to be the proportion of total number of sites supporting it in the MSA. The weight of a quartet say $ab|cd$ is calculated in a similar way. After all the quartet and sextet weights are obtained, an ever-expanding process is performed based on these weights to all full splits and their weights. As shown in previous literatures (Bandelt and Dress, 1992; Yang et al., 2013), reconstructing a phylogenetic tree or network is equivalent to calculating all the full splits and their weights. Thus, we have obtained the reconstructed tree or network by this process, which could be shown by a software SplitsTree4 (Huson and Bryant, 2006).



Ever-Expanding Process Based on Quartet and Sextet Weights

As represented by equation $W(a|d) = W(a|de) + W(ae|d)$, there is such an equation $W(abc|def) = W(abc|defg) + W(abcg|def)$, which can be seen as adding a new taxon g to either side of a split $abc|def$. If $W(abc|def) = 0$, then $W(abc|defg) = 0$ and $W(abcg|def) = 0$. If taxa group $A_1 \subseteq A$ and $B_1 \subseteq B$, or $A_1 \subseteq B$ and $B_1 \subseteq A$, we call the split $A|B$ displays $A_1|B_1$. It is proven in Bandelt and Dress (1992) that $W(A|B) \leq W(A_1|B_1)$. Therefore, a split with zero weight cannot be further expanded to larger splits with positive weights.

For a taxa set S with size n , there are $10 \binom{n}{6}$ sextets. We first calculate the weights of all quartets and sextets from the MSA, and then we expand them to get all full split weights using an ever-expanding process. Suppose there is a septet of $abc|defg$ type, we have $W(abc|defg) = W(abc|def) - W(abcg|def)$, and there is a similar equation for $W(abcg|def)$, so the weight of $W(abc|defg)$ can be obtained by similar continuous calculations, as follows.

$$\left\{ \begin{array}{l} W(abc|defg) = W(abc|def) - W(abcg|def) \\ W(abcg|def) = W(abg|def) - W(abg|cdef) \\ W(abg|cdef) = W(abg|cde) - W(abfg|cde) \\ W(abfg|cde) = W(afg|cde) - W(afg|bcde) \\ W(afg|bcde) = W(afg|bcd) - W(aefg|bcd) \\ W(aefg|bcd) = W(efg|bcd) - W(efg|abcd) \\ W(efg|abcd) = W(efg|abc) - W(defg|abc) \end{array} \right.$$

Combining the above equations, we have

$$\begin{aligned} W(abc|defg) &= \frac{1}{2} \{W(abc|def) - W(abg|def) + \\ &W(abg|cde) - W(afg|cde) + W(afg|bcd) - \\ &W(efg|bcd) + W(efg|abc)\} \end{aligned} \quad (1)$$

For $|B| \geq 4$, taking minimum over all possible cases, we have

$$\begin{aligned} W(abc|B) &= \max \left\{ \frac{1}{2} \min_{defg \in B} \{W(abc|def) - W(abg|def) + \right. \\ &W(abg|cde) - W(afg|cde) + \\ &W(afg|bcd) - W(efg|bcd) + W(efg|abc)\}, 0 \} \end{aligned} \quad (2)$$

When $|A|=4$ and $|B|=4$, the weight of the 4|4 split

$$\begin{aligned} W(A|B) &= \min \left\{ \min_{a \in A} \{W(A - a|B) - W(A - a|B + a)\}, \right. \\ &\left. \min_{a \in B} \{W(A|B - a) - W(A + a|B - a)\} \right\} \end{aligned} \quad (3)$$

where $A - A'$ for two sets A and A' denotes set difference (subtraction).

For example $A = \{a, b, c, d\}$, $B = \{e, f, g, h\}$, there are eight equations for $W(abcd|efgh)$,

$$W(abcd|efgh) = \begin{cases} W(abc|efgh) - W(abc|defgh) \\ W(abd|efgh) - W(abd|cefgh) \\ W(acd|efgh) - W(acd|befgh) \\ W(bcd|efgh) - W(bcd|aefgh) \\ W(abc|efg) - W(abc|dh|efg) \\ W(abc|efh) - W(abc|dg|efh) \\ W(abc|efgh) - W(abc|df|efgh) \\ W(abc|efgh) - W(abc|fgh) \end{cases}$$

For any split $A|B$ with $|A| \geq 4$ and $|B| \geq 4$, we traverse the elements in A and B and take out four taxa for each calculation. Suppose $a, b, c, d \in A$ and $e, f, g, h \in B$, and we have

$$W(A|B) = \min_{abcd \in A, efgh \in B} \{W(abcd|efgh)\} \quad (4)$$

For any $2|n$ split of $ab|B$ type with $c, d, e \in B$, we calculate their weight by formula (5) referred in Quartet-Net (Yang et al., 2013),

$$W(ab|B) = \max \left\{ \frac{1}{2} \min_{cde \in B} \{W(ab|cd) - W(ae|cd) + W(ae|bc) - W(bc|de) + W(ab|de)\}, 0 \right\} \quad (5)$$

Finally, for any trivial split of $a|S - a$ type with $b, c \in S - a$ in a taxa set S , we calculate the weight as follows (see also Yang et al., 2013):

$$W(a|S - a) = \min_{bc \in S - a} \left\{ W(a|bc) - \sum_{a \in A; bc \in B} A|B \right\} \quad (6)$$

Formulas (1) – (6) are used to calculate all full splits by decomposing sextet weights iteratively.

The QS-Net Method

QS-Net takes an MSA as input. Suppose that there are n taxa in the taxa set S , which are arranged in the order of 1, 2, 3, ..., n . In the initialization step, all triplet, quartet, and sextet weights are calculated directly from the MSAs. We calculate the weights of full splits in the following ways.

- Full split of type $A|S - A$ with $|A| \geq 3$ and $|S - A| \geq 3$: for the first six taxa—1, 2, 3, 4, 5, and 6—there are 10 sextets. We store these sextets together with their weights in a set X_1 . QS-Net then iteratively adds $i=7, 8, \dots, n$ to the left and right parts of the splits stored in X_1 and use equations (2)–(4) to calculate the weights of newly generated splits. Noticing that the only splits that cannot be generated in this way are of type $i|j|k|S_1 - \{i, j, k\}$ with $j = i - 1, i - 2, \dots, 2$ and $k = j - 1, j - 2, \dots, 1$, we calculate their weights using equation (2) and add them to X_1 . At the end of each iteration, the splits with a weight of zero are removed because they cannot be further expanded to have a positive weight. After the last iteration, all full splits of type $|A| \geq 3$ and $|S - A| \geq 3$ have been calculated.
- $2|n - 2$ full splits: These splits can be calculated using equation (5). In practice, we use Quartet-Net to calculate their split weights.
- Trivial $(1|n - 1)$ full splits: These splits can be calculated by equation (6).

By the above procedures, we calculate the weights of all full splits. Similar to Yang et al. (2013), it is usually advisable to filter the non-trivial full splits with very low split weights, which tend to be false positives. In practice, we remove splits with weight less than $c\%$ of the average weight, where c is a user-defined threshold setting to be 1 in this study. The output file containing all non-zero full splits and their weights is stored in NEXU file format, which can be visualized using SplitsTree4 (Huson and Bryant, 2006). The time complexity of QS-Net is $O(n^{10})$.

RESULTS AND DISCUSSIONS

To demonstrate QS-Net, we analyzed three artificial data sets and two real data sets. The artificial data sets were generated

from a simple tree phylogeny, a phylogenetic scenario with three reticulate events, and a more complicated phylogenetic scenario with five reticulate events. The purpose is to show that the QS-Net method can accurately reconstruct all kinds of evolutionary histories from simple to complicated ones. The real data include a bacterial taxonomy data consisting of 36 bacterial species and the whole genome sequences of 22 H7N9 influenza A viruses downloaded from NCBI influenza database.

The software Dawg (Cartwright, 2005) with model GTR + Gamma + I was used to generate three artificial data sets. The substitution rate is 0.01; the sequence length of the tree is 10,000 bp; the sequence length of the network containing three evolutionary events is 80,000 bp, while the sequence length of the network containing five evolutionary events is 320,000 bp because they are a concatenation of eight and 32 feasible trees. To avoid randomness, we performed 100 Dawg runs on each of the three artificial data sets and applied the 100 MSAs of each data set to QS-Net together with other four popular methods: Quartet-Net (Yang et al., 2013), Neighbor-Net (Bryant and Moulton, 2004), Split-Decomposition (Bandelt and Dress, 1992), and Neighbor-Joining (Saitou and Nei, 1987).

Analysis on the Tree Data

The tree data were generated from Figure 1B with 12 leaves. For brevity, we only listed reconstructed taxa set in the left or right block containing fewer number of taxa (Supplementary Material: Table S1). For example, split $bd|acefghijkl$ was listed as bd . We then normalized each split by the weight of a split successfully constructed by all methods. All trivial full splits were not listed because they can be successfully reconstructed by all five methods. As shown in Table 1, all five methods can successfully reconstruct all full splits in the 100 runs of the tree data; the accuracy is equal to the experimental bootstrap value divided by the real bootstrap value. The true-positive split result represents all splits in the real phylogenetic history of the simulated data sets. We listed the number of true-positive splits obtained by the five methods on all simulated data sets in Table 2. If a method can reconstruct the true-positive split once in 100 runs, we determined that the true-positive split can be obtained by this method. In addition to true-positive results, other split results reconstructed by the method are false-positive splits, which typically have very few weight values. Except for Neighbor-Joining, the other four methods reconstructed some false-positive splits (here we only list false-positive splits with a bootstrap value ≥ 10). For example, Quartet-Net and QS-Net reconstruct two additional split al and ae with bootstrap values of 10 and 26, respectively (see Table 3). This is because QS-Net and Quartet-Net methods use the same calculation formula for split of $2|n$ type. Neighbor-Net identifies 35 false-positive splits with bootstrap value ranging from 10 to 40. These false-positive splits may be caused by some random mutations in the tree data set.

TABLE 1 | Comparison of accuracy (the total bootstrap value obtained from the experimental results is divided by the bootstrap BV value) between QS-Net and four other methods.

Data set	QS-Net	Quartet-Net	Neighbor-Net	Split-Decomposition	Neighbor-Joining
Tree	100%	100%	100%	100%	100%
Network (3)	100%	100%	70.16%	67.24%	36%
Network (5)	100%	94.74%	58.89%	46.76%	23.68%

Network (3) is the phylogenetic network with three reticulate events, while Network (5) is the phylogenetic network with five reticulate events.

TABLE 2 | The number of true-positive results can be obtained by five methods.

Data set	True	QS-Net	Quartet-Net	Neighbor-Net	Split-Decomposition	Neighbor-Joining
Tree	9	9	9	9	9	9
Network (3)	25	25	25	21	23	9
Network (5)	38	38	36	30	22	11

The "True" column represents the real number of true-positive splits of the simulated data.

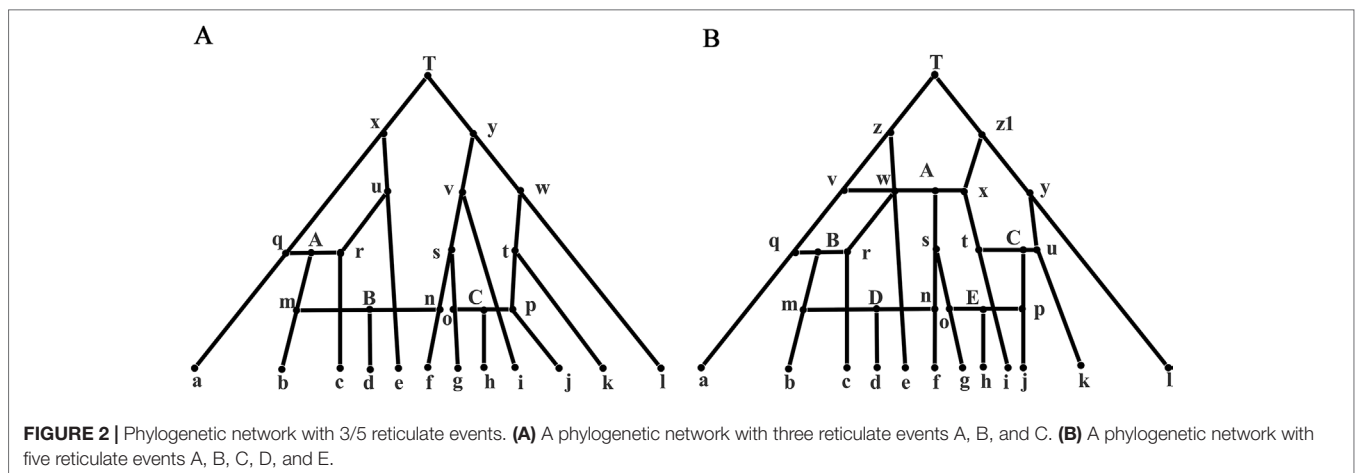
TABLE 3 | The number of false-positive results obtained by five methods.

Data set	QS-Net	Quartet-Net	Neighbor-Net	Split-Decomposition	Neighbor-Joining
Tree	2	2	35	4	0
Network (3)	4	4	16	1	0
Network (5)	4	4	4	1	0

Analysis on the Network Data with Three Reticulate Events

The network data were generated from **Figure 2A** containing three reticulate events A, B, and C, which can be decomposed into eight feasible underlying trees. A feasible tree can be obtained by cutting off one branch respectively at A, B, and C. For example, we can get an underlying tree by cutting off the three edges qA, mB, and oC in the three reticulate events. The sequence data of a taxon m were generated by concatenating partial sequence data from q and partial sequence data from r. All true splits and splits reconstructed by the five methods are listed in **Supplementary Material: Table S2**. The weight of the true split is the sum of the

split weights in eight feasible trees. Similarly, we normalized each split with the weight of split ab and multiplied it by 4. As can be seen from the **Table 1**, QS-Net and Quartet-Net accurately reconstructed all true splits in all 100 runs, while Neighbor-Net, Split-Decomposition, and Neighbor-Joining failed to reconstruct a large number of true splits. For example, Neighbor-Net failed to reconstruct split gh, fgi, and fgh in more than 90 runs, and Split-Decomposition was unable to reconstruct split bce and bcde in all 100 runs (**Supplementary Material: Table S2**). Neighbor-Joining obtained even worse result with 16 true splits missing, which is reasonable because Neighbor-Joining only reconstructs trees and retains the strongest compatible splits.



Analysis on the Network Data with Five Reticulate Events

Supplementary Material Table S3 lists all true splits and splits reconstructed from the five methods from the network data. The data set was generated from Figure 2B with a complicated phylogenetic scenario containing five reticulate events. Similarly, the weight of the true split is the sum of the weights of the splits in 32 feasible trees. We normalized each split with the weight of split ce. As can be seen from the Table 1, only QS-Net method obtains 100% accuracy in all 100 runs, while the other four methods fail to reconstruct some splits in most runs. For example, Quartet-Net failed in reconstructing split fgi and afg in all 100 runs. In addition to the two splits, Neighbor-Net also cannot reconstruct split hj, bcd, and bcde in more than 90 runs (Supplementary Material: Table S3), which happens because Neighbor-Net reduces splits to make the split system planar. Split-Decomposition and Neighbor-Joining still performed poorly. In addition, all methods except for Neighbor-Joining reconstructed some false-positive splits.

Analysis on the Bacterial Data

The bacterial data set was used in Takahashi and Kryukov (2009) for the analysis of phylogenetic relationships among bacterial species. This data set consists of 36 bacterial genomes containing concatenated sequence of seven genes (16S rRNA, 23S rRNA, gyrB, pyrH, recA, rpoA, and rpoD). The 36 species were divided into three different groups based on different GC content (32–38%, 50–53%, and 64–69%), containing 14, 11, and 11 species, respectively. We took the GC-rich data consisting of 11 bacterial species and a data of 25 species containing both GC-poor and GC-rich bacteria. The MSAs of both data were generated by ClustalW (Larkin et al., 2007) and further fed into to QS-Net, Quartet-Net (Yang et al., 2013), Neighbor-Net (Bryant and Moulton, 2004), Split-Decomposition (Bandelt and Dress, 1992), and Neighbor-Joining (Saitou and Nei, 1987). We ran the program on an MSI laptop with 2.8-GHz processor and 8-GB memory. A comparison of runtime between QS-Net and Quartet-Net on all data sets is shown in Table 4; the time statistics for three artificial data sets are the average of all 100 runtimes. The Neighbor-Joining method has the least runtime, and all other three methods can produce results in less than 2 s on all data sets. The reconstructed results were then viewed by SplitsTree4 (Huson and Bryant, 2006). Only three split networks reconstructed by QS-Net and Quartet-Net method on bacterial data set are shown in Figures 3 and 4.

Figure 3 shows the phylogenetic network of 11 GC-rich bacterial sequence data set by using QS-Net, which is basically

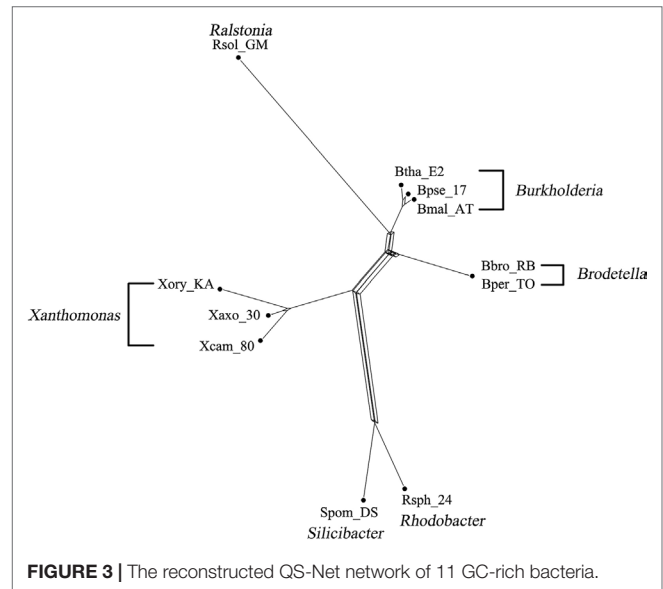


FIGURE 3 | The reconstructed QS-Net network of 11 GC-rich bacteria.

consistent with the experimental results in Takahashi and Kryukov (2009). The reconstructed networks of 25 GC-poor or GC-rich (32–38% and 64–69%) sequence data set reconstructed by QS-Net and Quartet-Net are shown in Figures 4A, B, respectively. As can be seen from the figures, the differences between QS-Net and Quartet-Net are quite obvious. There are two distinct parallelograms that represent the reticulate evolution event in the reconstructed network in Figure 4A but not in Figure 4B, which might be neglected by Quartet-Net due to its inability to identify complicated reticulate events. The numbers of full splits reconstructed by the five methods on bacterial data set and the influenza data set are also listed in Table 5. QS-Net constructs a moderate total number of splits among all comparison methods, probably because the full resolution of taxa is not achieved. In the GC-rich data set, Neighbor-Net constructs three more splits than does QS-Net, while in the GC-poor and GC-rich data set, Neighbor-Net constructs 29 more splits than does QS-Net. In addition, by comparing Figures 3 and 4A, it can be found that GC content may have an important influence on the evolutionary history of bacteria.

Analysis on the Influenza Data

The data set consisted of the full genome sequence of 22 H7N9 influenza A viruses aligned by ClustalW (Larkin et al., 2007). These viruses have major relations with the H7N9 virus (Gao et al., 2013) that appeared in China in 2013, which caused human mortality. We estimated the phylogenetic relationships

TABLE 4 | A comparison of runtime between QS-Net and Quartet-Net on all data sets.

Method	Tree	Network (3)	Network (5)	GC rich	GC poor and rich	Influenza
QS-Net	1.25 s	6.02 s	24.39 s	0.92 s	9.49 min	3.22 min
Quartet-Net	0.20 s	1.05 s	4.05 s	0.19 s	10.17 s	4.54 s

of these 22 influenza A viruses using Quartet-Net and QS-Net. The results are shown in **Figures 5A, B**, respectively. **Table 5** lists the numbers of full splits reconstructed by the five methods on bacterial data set and the influenza data set. General split networks do not actually represent explicit evolutionary

events, which makes the interpretation and comparison of reconstruction methods on real data set difficult. So we list the number of splits built by various methods. As can be seen in **Table 4**, QS-Net reconstructs 47 full splits, while Quartet-Net reconstructs 45 full splits.

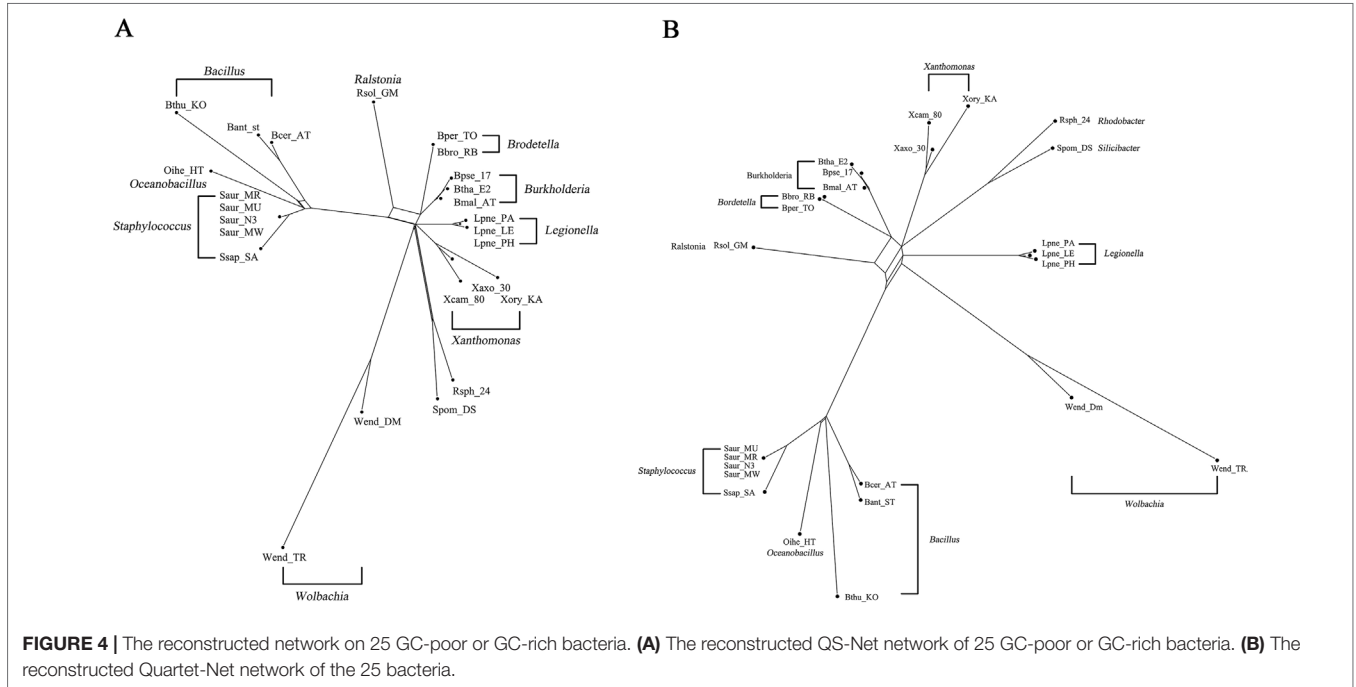
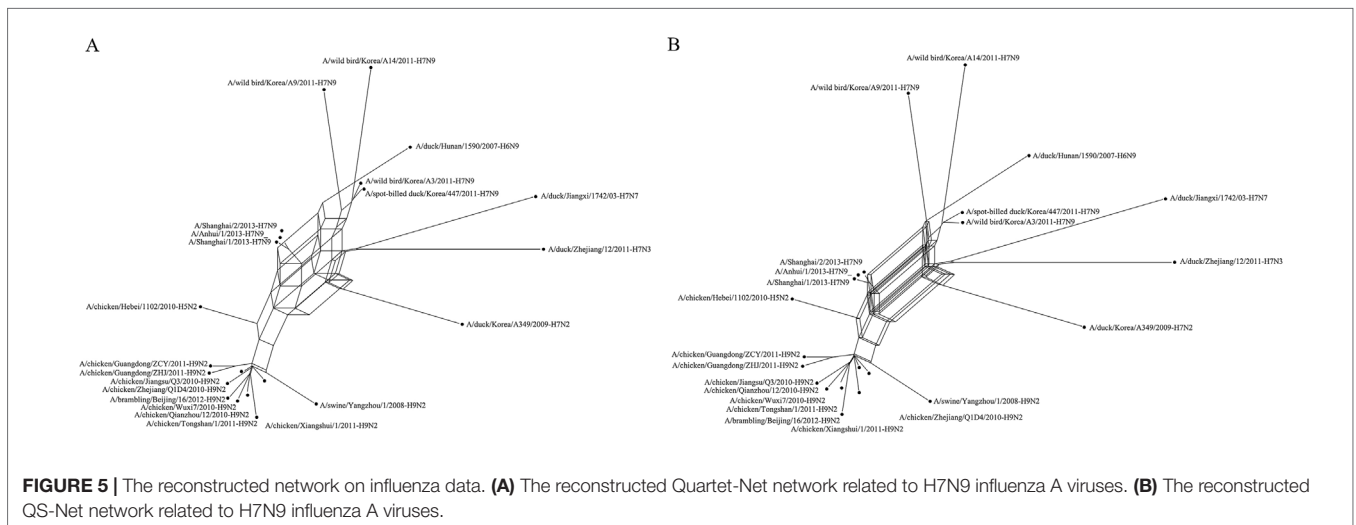


TABLE 5 | The number of full splits reconstructed by five methods on bacterial data set and the influenza data set.

Data set	QS-Net	Quartet-Net	Neighbor-Net	Split-Decomposition	Neighbor-Joining
GC rich	26	22	29	23	19
GC poor and rich	48	45	77	48	47
Influenza	47	45	68	36	41



The three viruses that caused human death (A/Shanghai/1/2013, A/Shanghai/2/2013, and A/Anhui/1/2013) were combined. The phylogenetic network indicates that these H7N9 viruses may be derived from the reassortment from influenza subtypes, including avian-origin H7N9 viruses, H9N2 viruses, and H7N3 viruses. In **Figure 5B** (constructed by QS-Net), the internal region surrounded by H7N9, H7N7, and H7N3 is more complex than **Figure 5A** (constructed by Quartet-Net), which indicates that the true evolutionary history of H7N9 influenza A viruses is very complex. Of course, the real evolutionary history is unknown, but at least the results constructed by QS-Net are consistent with a few previous findings.

CONCLUSIONS

QS-Net is a method generalizing Quartet-Net. Both simulation studies and real data analyses show that QS-Net has the potential to reconstruct more accurate phylogenetic relationships than its competitors like Quartet-Net and Neighbor-Net. However, the method runs slower than other algorithms, and the major computational difficulty lies in the calculation of 3|4 splits. Nevertheless, the difficulty will be partially resolved with the development of high-speed computers and parallel algorithms. Thus, we believe QS-Net will be useful in identifying more complex reticulate events that will be ignored by other network reconstruction algorithms.

REFERENCES

- Avni, E., Cohen, R., and Snir, S. (2015). Weighted quartets phylogenetics. *Syst. Biol.* 64 (2), 233–242. doi: 10.1093/sysbio/syu087
- Bandelt, H. J., and Dress, A. W. M. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.* 1 (3), 242–252. doi: 10.1016/1055-7903(92)90021-8
- Bastide, P., Solis-Lemus, C., Ané, C., William Sparks, K., and Kriebel, R. (2018). Phylogenetic comparative methods on phylogenetic networks with reticulations. *Syst. Biol.* 67 (5), 800–820. doi: 10.1093/sysbio/syy033
- Berry, V., and Gascuel, O. (2000). Inferring evolutionary trees with strong combinatorial evidence. *Theor. Comput. Sci.* 240 (2), 271–298. doi: 10.1016/S0304-3975(99)00235-2
- Bryant, D., and Moulton, V. (2004). Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* 21 (2), 255–265. doi: 10.1093/molbev/msh018
- Cartwright, R. A. (2005). DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics* 21 Suppl 3 (Suppl_3), iii31. doi: 10.1093/bioinformatics/bti1200
- Claudia, S. L., and Cécile, A. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* 12 (3), e1005896. doi: 10.1371/journal.pgen.1005896
- Eslahchi, C., Habibi, M., Hassanzadeh, R., and Mottaghi, E. (2010). MC-Net: a method for the construction of phylogenetic networks based on the Monte-Carlo method. *BMC Evol. Biol.* 10 (1), 254. doi: 10.1186/1471-2148-10-254
- Fenderson and Bruce, (2008). Molecular biology of the cell, 5th Edition. *Shock* 30 (1), 100. doi: 10.1097/01.shk.0000286288.33338.f6
- Gao, R., Cao, B., Hu, Y., Feng, Z., Wang, D., Hu, W., et al. (2013). Human infection with a novel avian-origin influenza A (H7N9) virus. *N. Engl. J. Med.* 368 (20), 1888. doi: 10.1056/NEJMoa1304459
- Grünewald, S., Forslund, K., Dress, A., and Moulton, V. (2007). QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets. *Mol. Biol. Evol.* 24 (2), 532–538. doi: 10.1093/molbev/msl180

AUTHOR CONTRIBUTIONS

JY and BL conceived the concept of the work and designed the experiments. MT, HL, ZC, and JZ performed literature search. MT, HL, DY, and GT collected and analyzed the data. MT and JY wrote the paper. All authors have approved the final manuscript.

FUNDING

This work was supported by Hainan Provincial Innovation research team (No. 2019CXTD405), National Natural Science Foundation of China (No. 61762034), Hainan Provincial Natural Science Foundation of China (No.618MS057, No.617122) , Hainan Provincial major scientific and technological plans (No. ZDKJ2017012), Natural Science Foundation of Hunan, China (Nos. 2018JJ2461 and 2018JJ3568), New Century Excellent Talents in university (No. NCET-10-0365), National Nature Science Foundation of China (Nos 11171369, 61272395, 61370171, 61300128, 61472127, 61572178, 61672214, and 61702054).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.00607/full#supplementary-material>

- Grunewald, S., Spillner, A., Bastkowski, S., Bogershausen, A., and Moulton, V. (2013). SuperQ: computing supernetworks from quartets. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10 (1), 151–160. doi: 10.1109/TCBB.2013.8
- Huber, K. T., Langton, M., Penny, D., Moulton, V., and Hendy, M. (2002). Spectronet: a package for computing spectra and median networks. *Appl. Bioinformatics* 1 (3), 159–161. doi: 10.1385/ABAB:97:1:01
- Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Syst. Biol.* 44 (1), 17–48. doi: 10.1093/sysbio/44.1.17
- Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23 (2), 254–267. doi: 10.1093/molbev/msj030
- Huson, D. H., Rupp, R., and Scornavacca, C. (2010). *Phylogenetic networks: concepts, algorithms and applications*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511974076
- Keijsper, J. C. M., and Pendavingh, R. A. (2014). Reconstructing a phylogenetic level-1 network from quartets. *Bull. Math. Biol.* 76 (10), 2517–2541. doi: 10.1007/s11538-014-0022-z
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23 (21), 2947–2948. doi: 10.1093/bioinformatics/btm404
- Leducq, J.-B., Henault, M., Charron, G., Nielly-Thibault, L., Terrat, Y., Fiumera, H. L., et al. (2017). Mitochondrial recombination and introgression during speciation by hybridization. *Mol. Biol. Evol.* 34 (8), 1947–1959. doi: 10.1093/molbev/msx139
- Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* 32 (10), 2798–2800. doi: 10.1093/molbev/msv150
- Nelson, M. I., Viboud, C., Simonsen, L., Bennett, R. T., Griesemer, S. B., George, K. S., et al. (2008). Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. *PLoS Pathog.* 4 (2), e1000012. doi: 10.1371/journal.ppat.1000012
- Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405 (6784), 299–304. doi: 10.1038/35012500

- Park, H. J., Jin, G., and Nakhleh, L. (2010). Bootstrap-based support of HGT inferred by maximum parsimony. *BMC Evol. Biol.* 10 (1), 131. doi: 10.1186/1471-2148-10-131
- Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8 (11), e1002967. doi: 10.1371/journal.pgen.1002967
- Ranwez, V., and Gascuel, O. (2002). Improvement of distance-based phylogenetic methods by a local maximum likelihood approach using triplets. *Mol. Biol. Evol.* 19 (11), 1952–1963. doi: 10.1093/oxfordjournals.molbev.a004019
- Rieseberg, L. H., Baird, S. J. E., and Gardner, K. A. (2000). Hybridization, introgression, and linkage evolution. *Plant Mol. Biol.* 42 (1), 205–224. doi: 10.1023/A:1006340407546
- Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4 (4), 406. doi: 10.1093/oxfordjournals.molbev.a040454
- Sayyari, E., and Mirarab, S. (2016). Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction. *BMC Genomics* 17 (Suppl 10), 101–113. doi: 10.1186/s12864-016-3098-z
- Solis-Lemus, C., Bastide, P., and Ané, C. (2017). PhyloNetworks: a package for phylogenetic networks. *Mol. Biol. Evol.* 34 (12), 3292–3298. doi: 10.1093/molbev/msx235
- Takahashi, M., and Kryukov, K. N. (2009). Estimation of bacterial species phylogeny through oligonucleotide frequency distances. *Genomics* 93 (6), 525–533. doi: 10.1016/j.ygeno.2009.01.009
- Tria, F., Caglioti, E., Loreto, V., and Pagnani, A. (2010). A stochastic local search algorithm for distance-based phylogeny reconstruction. *Mol. Biol. Evol.* 27 (11), 2587. doi: 10.1093/molbev/msq154
- Vachaspati, P., and Warnow, T. (2015). ASTRID: accurate species trees from internode distances. *BMC Genomics* 16 (10), S3. doi: 10.1186/1471-2164-16-S10-S3
- Van Iersel, L., Kelk, S., Rupp, R., and Huson, D. (2010). Phylogenetic networks do not need to be complex: using fewer reticulations to represent conflicting clusters. *Bioinformatics* 26 (12), i124–i131. doi: 10.1093/bioinformatics/btq202
- Vijaykrishna, D., Mukerji, R., and Smith, G. J. D. (2015). RNA virus reassortment: an evolutionary mechanism for host jumps and immune evasion. *PLoS Pathog.* 11 (7), e1004902. doi: 10.1371/journal.ppat.1004902
- Wen, D., Zhu, J., Yu, Y., and Nakhleh, L. (2018). Inferring phylogenetic networks using PhyloNet. *Syst. Biol.* 67 (4), 735–740. doi: 10.1093/sysbio/syy015
- Wheeler, T. J. (2009). *Large-scale Neighbor-Joining with NINJA*. Berlin Heidelberg: Springer. doi: 10.1007/978-3-642-04241-6_31
- Willson, S. J. (2013). Reconstruction of certain phylogenetic networks from their tree-average distances. *Bull. Math. Biol.* 75 (10), 1840–1878. doi: 10.1007/s11538-013-9872-z
- Yang, J., Grünewald, S., and Wan, X. F. (2013). Quartet-Net: a quartet-based method to reconstruct phylogenetic networks. *Mol. Biol. Evol.* 30 (5), 1206–1217. doi: 10.1093/molbev/mst040
- Yu, Y., and Nakhleh, L. (2015). A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* 16 (10), S10. doi: 10.1186/1471-2164-16-S10-S10
- Yu, Y., Nakhleh, L., Than, C., and Degnan, J. H. (2011). Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Syst. Biol.* 60 (2), 138–149. doi: 10.1093/sysbio/syq084
- Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. (2017). Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.* 35 (2), 504–517. doi: 10.1101/124982

Conflict of Interest Statement: Authors DY and GT were employed by company Geneis (Beijing) Co. Ltd. The remaining authors declare no competing interests.

Copyright © 2019 Tan, Long, Liao, Cao, Yuan, Tian, Zhuang and Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.