A simple method to test the reproducibility of the phylogenetic reconstructions: the molecular systematics of cyanobacteria as a case study

Andrea Paparini[1*], Elvina Lee[1], Andrew Bath[2], Cameron Gordon[2] & Una M. Ryan[1]

[1]Vector– and Water–Borne Pathogen Research Group, School of Veterinary & Life Sciences, Murdoch University, WA, Australia; [*]Corresponding author e–mail: a.paparini@murdoch.edu.au, tel.: +61 8 9360 7649
[2]Water Quality Branch, Water Corporation, 629 Newcastle Street, Leederville, Western Australia 6007

Abstract: Molecular systematics uses currently available data to produce the best approximation to the true (un–observable) phylogeny of a taxon. Molecular phylogeny complements morphological identification and classification of organisms, in order to infer their evolutionary relationships. In the current era dominated by cultivation–independent surveys, testing the potential technical and analytical pitfalls and limitations of environmental DNA surveys appears crucial. Sequence–based phylogenetic reconstructions rely on three main steps: alignment, alignment curation and tree building. Several independent options and settings can be adopted at each step, but it is well known that their choice (or combination) can significantly affect the topology of the phylogenetic tree obtained and skew the reliability of the resultant systematics. For the present study, five alignment algorithms, two curation options and three tree–building methods were used to infer the phylogeny of three orders of cyanobacteria, based on four validated markers widely used for this phylum: 16S rRNA, 16S–23S ITS, *cpc*BA–IGS and *rpo*C1. Compared to the alignment algorithm or the curation stringency used, the tree–building method was found to have the greatest effect on the resultant tree topology. This result was consistent for all loci, including the genetically–constrained (protein–coding) locus *rpo*C1. The reproducibility of the tree topology was clearly visualized and measured for each locus. This paper presents pitfalls in cyanobacteria systematics and implements a simple and rapid method, applicable to any locus and organism, to identify aberrant results and assess the reproducibility of phylogenetic reconstructions.

Key words: 16S rRNA, 16S–23S ITS, cyanobacteria, molecular phylogeny, phycocyanin operon, *rpo*C1

Introduction

Cyanobacteria are ancient photosynthetic bacteria with a cosmopolitan distribution, important ecological roles and global socio–economic relevance (Whitton & Potts 2000; Sciuto & Moro 2015). Members of this phylum have traditionally been identified and classified based on morphological characteristics, which determined their taxonomic distinction based on phenotypic properties (Castenholz 2001; Palinska & Surosz 2014) .

Pleomorphism, uncultivability, cryptic diversity, convergent evolution and other factors have greatly limited the reliability of this approach resulting in discrepancies, misnomers, confusing nomenclature and a puzzling systematics (Lyra et al. 2001; Komárek 2006; 2010; Palinska & Surosz 2014). To alleviate these limitations, DNA– or protein–based methods complementary to morphological analyses have been developed for identification, typing, traceability and classification (Willame et al. 2006; Valerio et al. 2009). The small ribosomal subunit RNA (16S rRNA) gene and its internal transcribed spacer (ITS) region, the protein–coding gamma subunit of the DNA–dependent RNA polymerase (*rpo*C1), the phycocyanin operon, consisting of the two *cpc*B–*cpc*A genes and their variable intergenic region (*cpc*BA–IGS) are commonly used phylogenetic markers (Neilan 1995; Fergusson & Saint 2000; Castenholz 2001; Coenye & Vandamme 2003; Komárek 2006; Lee et al. 2014). The 16S rRNA to 23S rRNA ribosomal DNA internal transcribed spacer region (16S–23S ITS) has also been suggested to be useful, for lower level discrimination of cyanobacterial *taxa* (Otsuka et al. 1999; Premanandh et al. 2006).

Once the sequences are obtained, the evolutionary relationships can be inferred by means of phylogenetic reconstructions, consisting of three main steps: alignment, alignment curation and tree building

(Holder & Lewis 2003; Harrison & Langdale 2006; Yang & Rannala 2012; De Bruyn et al. 2014).

During the first step, gaps are added to a matrix of data so that the nucleotides in one column are related to each other by descent from a common ancestral residue (Holder & Lewis 2003). ClustalW (Thompson et al. 1994) is probably the most widely used classical progressive alignment algorithm (Landan & Graur 2009); MAFFT (Katoh et al. 2002) and MUSCLE (Edgar 2004) are faster, progressive aligners including iteration and refinement, while PRANK (Löytynoja 2014) and PAGAN (Löytynoja et al. 2012) also distinguish between insertions, deletions and vary the costs between opening and extending gaps based on phylogenetic information.

Highly variable regions may be characterized by high rates of insertion or deletion of bases (INDELs) in the aligned sequences, which the algorithms resolve by introducing gaps of various lengths and frequencies. Alignment errors/artefacts and/or discrepancies are known to accumulate at these regions (Misof et al. 2014) and have been shown to significantly change the outcome of the phylogenetic reconstruction (Morrison & Ellis 1997; Ogden & Rosenberg 2006; Landan & Graur 2009; Liu et al. 2009). Consequently, these error prone sites may be removed either manually or automatically by programs like Gblocks (Castresana 2000; Talavera & Castresana 2007), REAP (Hartmann & Vision 2008) or NOISY (Dress et al. 2008).

During the final step, a phylogenetic tree, (a graph simulating the ancestor–descendant relationships between organisms or gene sequences), is constructed using a variety of methods based on either distance or characters (Yang & Rannala 2012). Computationally–fast distance–based methods such as neighbour–joining (NJ), calculate the pairwise distances between sequences and group the most similar sequences accordingly (Van de Peer 2009). As the molecular clock hypothesis doesn't always hold, the simplicity of this method underestimates the complexity of the phylogenetic–inference problem and approaches like maximum likelihood (ML) or Bayesian inference (BI), that take into account rate variation across lineages, were introduced to obtain better estimates of divergence times (Holder & Lewis 2003).

Phylogenetic reconstructions can be used to identify unknown isolates, track infections or contaminations, discover novel taxa or support classifications, formulated on the basis of characters of a different nature (e.g., molecular– and morphological–systematics). It should be noted however, that the obtained trees (consisting of relationships and divergence times) are not directly observed, but are instead statistically–inferred from available data. This implies that the robustness of the reconstruction can vary and that, starting from a given dataset, multiple (sometimes equally plausible) scenarios can be obtained. While tree scores can be used to identify the most probable tree (e.g., the plausibility of the mutations that a particular tree would require to explain the data), the congruency between the inferred (unobservable) molecular phylogeny of a given taxon and the generated systematics cannot be easily tested. For this reason, noting only tree scores and branch statistical supports gives only partial and/or skewed indications of the "true" evolutionary relationships between the various lineages. In this context, accepting a tree without assessing the variation of all the topologies obtained experimentally can be inadequate (Morrison & Ellis 1997).

The goal of the present study was to investigate the combined effect of alignment algorithm, alignment curation option and tree–building method, on the reproducibility of phylogenetic reconstructions of cyanobacterial *taxa* with the most commonly used and validated phylogenetic markers: 16S rRNA, *rpoC*1, *cpc*BA–IGS and 16S–23S ITS.

Materials and Methods

Isolation, cultivation and sequencing of cyanobacterial isolates. Forty–nine cyanobacteria isolates from a variety of freshwater habitats (n = 20) surveyed in Western Australia were isolated, cultured and DNA was extracted as described previously (Lee et al. 2014). Partial fragments of the 16S rRNA hypervariable region (467 bp), *rpo*C1 (612 bp), 16S–23S interspacer region (variable length) and the phycocyanin intergenic spacer region (*cpc*BA–IGS) (approximately 585 bp) were amplified as previously described (Palenik & Haselkorn 1992; Robertson et al. 2001; Janse et al. 2003; McGregor & Rasmussen 2008). Amplicons corresponding to the expected length were excised from the electrophoresis gel and sequenced on an Applied Biosystem 3730 DNA Analyzer (Applied Biosystems, USA).

Multiple sequence alignment and curation. Figure 1 gives an overview of the study workflow. The input set consisted of globally–trimmed sequences (16S rRNA: n = 112; 16S–23S ITS: n = 87; *cpc*BA–IGS: n = 95; *rpoC*1: n = 94), generated in our laboratory or retrieved from GenBank by BLAST–searches. Global trimming of the input set was performed manually in MEGA5 (Tamura et al. 2011) to eliminate terminal gaps, by constructing temporary ClustalW alignments (Larkin et al. 2007) before global trimming. The temporary alignments were then dissolved and five new alignments were generated under the default settings of each program: ClustalW v.1.82, MAFFT v.6.712, MUSCLE v.3.7, PAGAN v.0.44 and PRANK v.100223 (Katoh et al. 2002; Edgar 2004; Larkin et al. 2007; Löytynoja et al. 2012; Löytynoja 2014). Alignments were curated (i.e., degapped), remotely (Dereeper et al. 2008), by Gblocks v.0.91b (Talavera & Castresana 2007) on default settings. The p–distance scores for all (uncurated and curated) alignments were then computed by the PAST software v.3.08 (Hammer et al. 2001).

Tree–building and multivariate analyses. The appropriate models of nucleotide substitution for each alignment was determined using JModelTest2 (Darriba et al. 2012) on the CIPRES Science Gateway v.3.3 (Miller et al. 2010). Phylogenetic trees were constructed using distance (neigh-

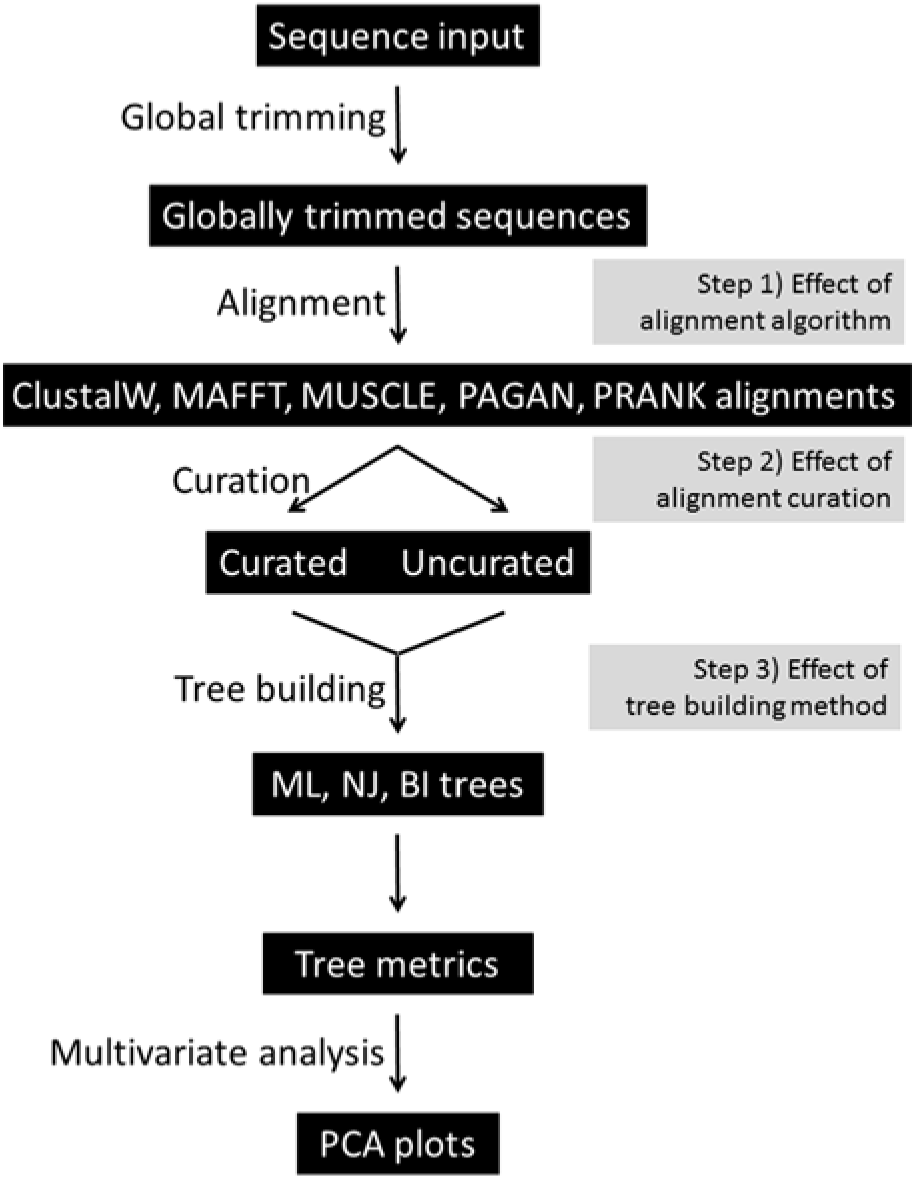
Sequence input
Global trimming
Globally trimmed sequences
Alignment
Step 1) Effect of alignment algorithm
ClustalW, MAFFT, MUSCLE, PAGAN, PRANK alignments
Curation
Step 2) Effect of alignment curation
Curated
Uncurated
Tree building
Step 3) Effect of tree building method
ML, NJ, BI trees
Tree metrics
Multivariate analysis
PCA plots

Fig. 1. Workflow: Overview of the workflow performed. Cyanobacterial sequences from four loci (16S rRNA, *rpo*C1, 16S–23S–ITS or *cp-c*BA–IGS) were aligned using five alignment algorithms (ClustalW, MAFFT, MUSCLE, PAGAN, PRANK), prior to optional alignment curation (i.e., de–gapping). Trees were built using maximum likelihood (ML), neighbour joining (NJ) and Bayesian inference (BI). Multivariate analysis was performed based on a number of metrics computed from each phylogenetic tree. The output of the multivariate analysis consisted of principal component analysis (PCA) plots highlighting the topological differences between trees, generated using alternative methods adopted during steps 1 to 3.

bour–joining–NJ), maximum likelihood (ML), and Bayesian inference (BI), using MEGA5, RAxML v.8.0 or SiMBa v.1.0 (Stamatakis et al. 2008; Tamura et al. 2011; Ronquist et al. 2012; Mishra & Thines 2014). Optimal nucleotide substitution models and gamma– and invariable–rates were chosen for each reconstruction, based on the JModelTest2 results (Darriba et al. 2012). For the NJ and ML reconstructions, tree reliability was evaluated with bootstrap analysis of 500 replicates, while default settings were used for BI (No. generations: 1,000,000; burnin fraction: 0.25; No. of runs: 2; No. of chains: 4; outgroup set to *G. violaceous*), with substitution model and gamma– and invariable–rates based on JModelTest2 results (Darriba et al. 2012). The trees generated were visualised using FigTree v.1.4.2 (Rambaut 2014). From each of the trees generated, a set of five tree metrics were calculated using TreeStat v.1.2 (Rambaut 2008). These figures were then imported into PAST v.3.08 (Hammer et al. 2001) to compute multivariate analyses (principal component analyses–PCA) and generate PCA plots.

The tree metrics considered describe general features of the tree, its shape and topology. These were: 1) tree length (i.e., sum of branch lengths), 2) tree height (i.e., height of the root of the tree), 3) treeness (i.e., proportion of total length of tree taken up by internal branches; interpreted as a signal/signal+noise measure) (cf. (Phillips et al. 2001)), 4) N_bar (i.e., mean number of nodes above an external node) (Kirkpatrick & Slatkin 1993), and 5) cherry count (i.e., number of internal nodes that have only tips as children (McKenzie & Steel 2000). Although these measures are somehow inter–correlated, they capture different aspects of the tree shape and may be used for comparisons (Agapow & Purvis 2002).

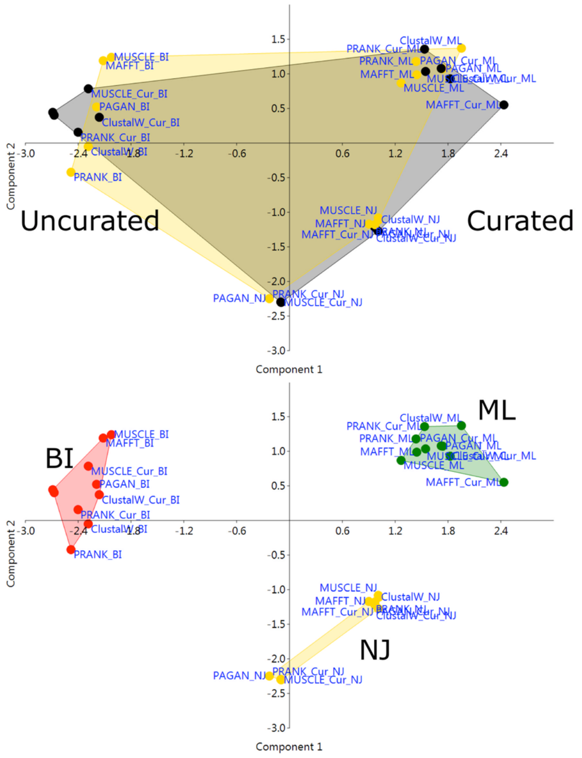

Fig. 2. 16S rRNA: Principal component analysis (PCA) plot highlighting the topological differences between trees generated using curated or uncurated alignments (panel A) and different tree–building methods (panel B). Sequence input consisted of cyanobacterial sequences from the 16S rRNA locus. Trees were built using maximum likelihood (ML), neighbour joining (NJ) and Bayesian inference (BI).

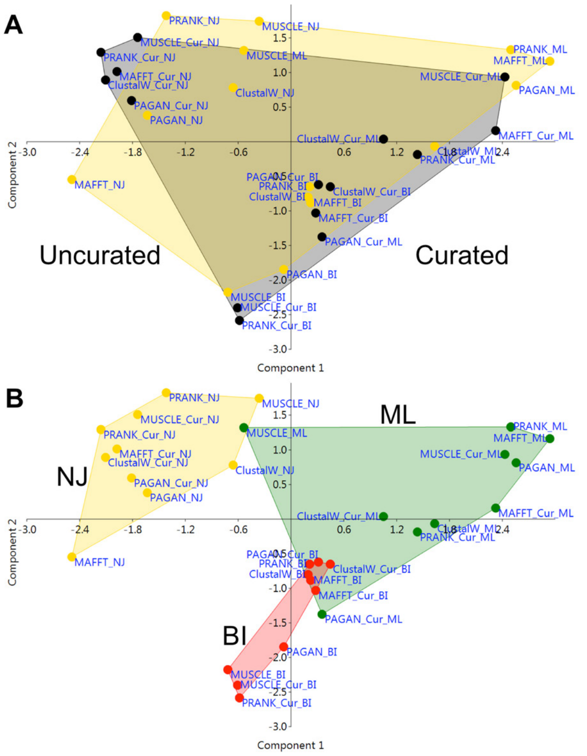

Fig .3. *rpo*C1: Principal component analysis (PCA) plot highlighting the topological differences between trees generated using curated or uncurated alignments (panel A) and different tree–building methods (panel B). Sequence input consisted of cyanobacterial sequences from the *rpo*C1 locus. Trees were built using maximum likelihood (ML), neighbour joining (NJ) and Bayesian inference (BI).

Results

Comparison of alignment algorithms and effect of curation

Depending on the alignment algorithm used, the final alignment length varied: at all four loci, PAGAN always produced the longest alignments, while ClustalW alignments were consistently the shortest. The effect of alignment algorithm and curation were more evident at the loci containing variable regions (16S–23S ITS and *cpc*BA–IGS) than at the conserved or protein coding loci (16S rRNA and *rpo*C1).

For the uncurated alignments at the 16S rRNA and *rpo*C1 loci, there were negligible differences, between the five algorithms, in: alignment length and fraction of conserved–, variable– and parsimony–informative–sites (data not shown). These alignments remained very similar, also after curation (there was < 20% change in alignment length, compared to pre–curation). This was reflected in the negligible differences between the five algorithms for the p–distance values, measured both pre– and post–curation (data not shown).

Conversely, large differences were seen when comparing the output of the various alignment algorithms at the hypervariable, highly–gapped 16S–23S ITS and *cpc*BA–IGS loci, where the longest alignments were three– and two–times longer than the shortest alignments, respectively. Curation also had a significant effect (especially for 16S–23S ITS) in reducing the alignment length by at least 65%. These large fluctuations in alignment length, fraction of conserved sites etc., were reflected in the computed p–distance values. At the16S–23S ITS (in particular) and *cpc*BA–IGS loci, the various alignment algorithms had different p–distance values both pre– and post–curation (data not shown).

Comparison of phylogenetic reconstructions

For each of the four loci, a total of 30 trees were generated, starting from 5 curated and 5 uncurated alignments (n = 120 trees in total).

Effect of alignment algorithm on tree topology

Multivariate analysis based on the five tree metrics used allowed for the comparison of the shape and topology of trees produced by different alignments. For 16S rRNA, there were virtually no differences, as shown by the areas connecting the trees produced by the same algorithm which were clearly overlapping in the PCA plot (data not shown). Only two trees (PRANK/uncu-

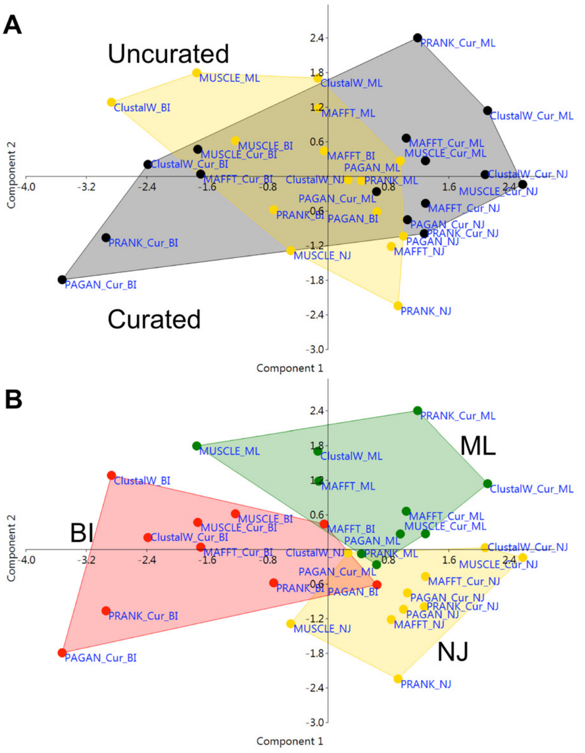

Fig. 4. *cpc*BA–IGS: Principal component analysis (PCA) plot highlighting the topological differences between trees generated using curated or uncurated alignments (panel A) and different tree–building methods (panel B). Sequence input consisted of cyanobacterial sequences from the *cpc*BA–IGS locus. Trees were built using maximum likelihood (ML), neighbour joining (NJ) and Bayesian inference (BI).

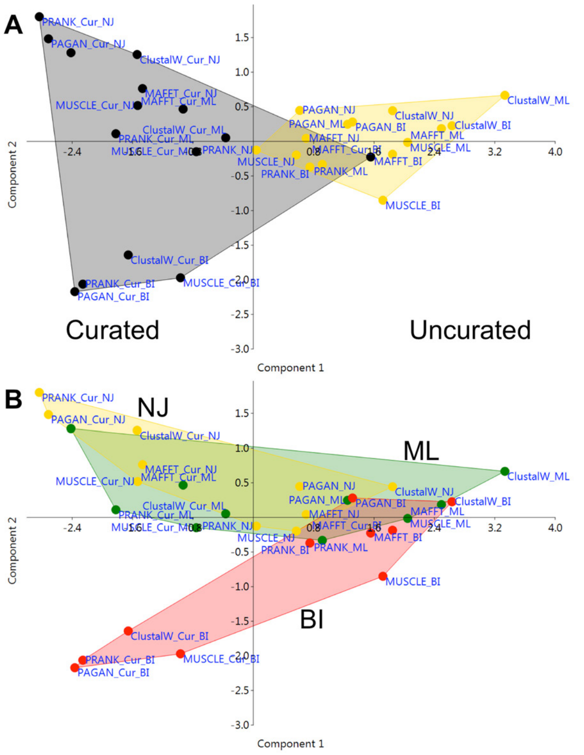

Fig. 5. 16S–23S–ITS: Principal component analysis (PCA) plot highlighting the topological differences between trees generated using curated or uncurated alignments (panel A) and different tree–building methods (panel B). Sequence input consisted of cyanobacterial sequences from the 16S–23S–ITS locus. Trees were built using maximum likelihood (ML), neighbour joining (NJ) and Bayesian inference (BI).

rated/BI and MAFFT/curated/ML) appeared slightly spatially segregated from the others on the PCA plot, based on their metrics.

Also for *rpo*C1, the differences were very small, with possibly up to seven trees showing a minor spatial segregation on the PCA plot (MAFFT/uncurated/NJ, PRANK/curated/NJ; PRANK/uncurated/NJ, MUSCLE/uncurated/NJ, MAFFT/uncurated/ML, MAFFT/curated/ML, MUSCLE/curated/BI) (data not shown).

On the other hand, for the 16S–23S ITS and *cpcBA*–IGS loci, some differences were noticeable and outlying tree topologies were relatively more evident (these trees appeared spatially segregated on the PCA plot, based on the metrics considered) (data not shown). For 16S–23S ITS, trees with possibly peculiar topologies were: MUSCLE/curated/BI, MUSCLE/uncurated/BI and ClustalW/uncurated/ML. For *cpc*BA–IGS, there was a core of trees with comparable topologies, surrounded by a large number of trees (≈10) appearing segregated from the others, based on their metrics (PAGAN/curated/BI, PRANK/curated/BI, ClustalW/curated/BI, ClustalW/uncurated/BI, MUSCLE/uncurated/ML, ClustalW/uncurated/ML, PRANK/curated/ML, ClustalW/curated/ML, MUSCLE/curated/NJ, PRANK/uncurated/NJ) (data not shown).

Effect of alignment curation on tree topology

Alignment curation affected tree topology in a clearly locus–specific manner. Differences between topologies of the trees produced from curated or uncurated alignments ranged from insignificant to extreme, increasing from 16S rRNA (curated alignment = uncurated alignment), to *rpo*C1, *cpc*BA–IGS and 16S–23S ITS (curated alignment ≠ uncurated alignment) (Figs. 2–5, Panel A). For the highly–gapped 16S–23S ITS hypervariable locus, the effects of curation were very obvious, with mainly two trees only (PRANK/uncurated/NJ and MAFFT/curated/BI) clearly intersecting the opposite cluster in the PCA plot (thus showing a topology more similar to the opposite cluster) (Fig. 5a).

Effect of tree–building method on tree topology

The tree–building method had a strong effect on tree topology, based on the metrics considered. Although locus specific, the extent of this effect was obvious for all loci considered. The 16S–23S ITS was the only marker where similar tree shapes were obtained from different tree–building methods (Fig. 5b). At this locus, NJ and ML trees largely coincided, whilst most trees obtained by BI were different. Conversely, at all other loci, the three tree–building methods produced considerably different topologies, with divergences between the methods progressively increasing from *cpc*BA–IGS, to *rpo*C1 and 16S rRNA. Interestingly, for the 16S rRNA locus, there was also high consistency among trees produced by the same method. For instance, all NJ trees (or ML or BI) showed virtually identical topology, irrespective of alignment algorithm or curation option (Fig. 2b). This shows that, for 16S rRNA, the choice of tree–building method is the most important factor to consider when planning the phylogenetic reconstruction method. Tree–building method, particularly for this marker, is by far more critical than the choice of alignment algorithm or curation option (based on the dataset and metrics analysed in the present study).

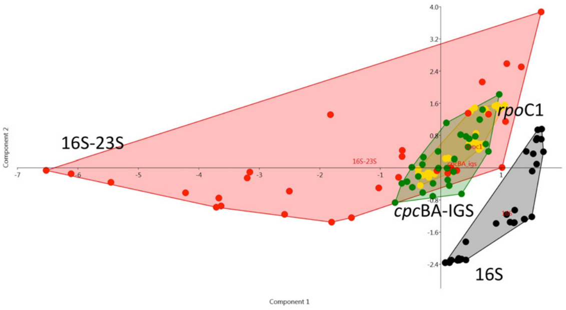

Fig. 6. Loci comparison: Principal component analysis (PCA) plot highlighting the reproducibility of trees generated by various combinations of phylogenetic reconstruction options, starting from cyanobacterial sequences from four loci: 16S rRNA, *rpo*C1, 16S–23S–ITS or *cpc*BA–IGS. Multivariate analysis was performed based on a number of metrics computed from each tree. The figure shows the principal component analysis (PCA) plot, output of the multivariate analysis.

Discussion

Initially conceived to infer evolutionary relationships based on traditional (morphological and physiological) characters, phylogenetic reconstructions have become extremely popular with the advent of culture–independent typing techniques and DNA sequencing technologies. The molecular systematics of a group is proposed through phylogenetic trees (graphical representations of the relationships among taxa) substantiated by the currently–available data, with the aim of inferring the "true" (un–observable) phylogeny of the taxon. In this context, the limitations shown by the more traditional approaches used for identification (e.g., morphological, biochemical etc.), have contributed to, and justified, the explosion of molecular classification of cyanobacteria.

In most of the 120 trees produced during the study, the Nostocales generally formed a monophyletic shallow–branching cluster, irrespective of the locus. The Chroococcales and Oscillatoriales were more deep–branching but their evolutionary relationships varied, depending on the reconstruction workflow and locus (the number of taxa per locus differed). As only three orders were included in the analysis, the degrees of resolution (soft polytomies) and of paraphyly of the orders yielded the major difference between the trees obtained from the same locus. Reconstructing the molecular phylogeny of the phylum cyanobacteria, however, lies beyond the objectives of the present paper. Instead, our study mainly aimed at quantifying and visualizing the uncertainties, associated with alternative reconstruction methods and at assessing the consequences of alternative analytical choices. The output of this study can still be used to suggest guidelines for investigating the molecular phylogeny of cyanobacteria.

The combination of conserved and hypervariable regions within the 16S rRNA, have made this marker the gold standard for the systematics of bacteria and archaea since the molecule was first sequenced in the late 70's (Brosius et al. 1978). This locus is important also for the phylum cyanobacteria (Seo & Yokota 2003; Lee et al. 2014; Rehakova et al. 2014) and the greater availability of 16 rRNA–based studies should be exploited to identify a "prototype tree", representing the currently accepted systematics (Tomitani et al. 2006; Howard–Azzeh et al. 2014; Rehakova et al. 2014). This step is essential to pinpoint the spatial position in the PCA plot of the "16S rRNA prototype tree", in comparison to all other 16S rRNA trees with alternative topologies under test. Moreover this tree can also be used for comparisons across loci.

During the present study, coherence with the topology of trees published in previous seminal papers (Tomitani et al. 2006; Howard–Azzeh et al. 2014; Rehakova et al. 2014) was found in several (similar) trees, produced by the ML tree–building method. This suggests that, with our dataset, this method was adequate to obtain reliable 16S rRNA cyanobacterial systematics.

When a benchmark "prototype tree" cannot be chosen a priori, or when more trees are equally plausible, the method described in the present paper study can only assess the reproducibility of the reconstructions, obtained from a given dataset using alternative options. However, optional display of metrics–associated vectors in the PCA plots (these are called biplots) can be very useful to identify possible relationships between the metrics' value and the spatial pattern in the PCA plot. For instance, one ML 16S rRNA tree, generated based on a curated MUSCLE alignment, was characterized by numerous monophyletic clades and fully resolved taxa and this is a combination of options that appeared successful in our hands, for 16S rRNA. Paraphyletic clades of Oscillatoriales appeared basal to the tree, while the Chroococcales (e.g., *Microcystis* spp., *Synechococcus* sp.) fell into two strongly supported clades (100% bootstrap support). The Nostocales formed a large monophyletic group with strong bootstrap support (96%) (data not shown). Interestingly, the topology of this very robust tree (MUSCLE/curated/ML) produced a treeness value higher than all other trees. During our study, high values of treeness (and possibly N_bar and cherry counts) were generally associated with fewer unresolved clades and (soft) polytomies.

As Fig. 6 shows, the metrics associated with the 16S–23S ITS trees varied significantly and had higher average values, than the other loci. This translated into some 16S–23S ITS trees being characterized by fewer polytomies and more bifurcations and sister taxa. Conceivably, this level of resolution is more useful and appropriate for reconstructions focusing on shallower phylogenies (e.g., species or sub–species level), rather than at phylum level. It is tempting to speculate that, once the systematics of a particular *taxon* is well–known, one metric (or a few) could be used as a proxy of "quality" of newly generated trees. Future implementations of the approach presented in this paper should also include metrics associated with the statistical support of branches.

Genetic constraints associated with protein coding loci are expected to produce fewer alignments gaps, and to increase the reproducibility of the results in response to the alternative workflows adopted. The present investigation clearly confirms that, because compared to 16S–23S ITS, trees at the *rpo*C1 locus (and to a lesser extent *cpc*BA–IGS) clustered tightly showing relatively little influence by the reconstruction method. In contrast, the set of topologies for the first locus were highly divergent (Fig. 6). These findings imply that for *rpo*C1 and *cpc*BA–IGS the combination of analytical options made for the reconstruction should be less critical than for 16S–23S ITS.

As multiple plausible trees can be obtained, the comparison of topologies should be treated with caution as none of the tree metrics used in the present paper has been tested as a universal proxy of accuracy. When other tree–building methods are used (e.g., minimum evolution) trees showing minimum tree length (shortest tree length) may be selected as "true" trees (Van de Peer 2009). However, previous studies have also shown that, although curation results in shorter trees, this does not necessarily yield improved accuracy for other tree–building methods (Liu et al. 2009). One finding from the present analysis somewhat contrary to previous reports (Morrison & Ellis 1997; Lindgren & Daly 2007; Liu et al. 2009; Löytynoja 2012), is that alignment parameters did not appear to affect phylogenetic reconstructions and evolutionary inferences. This difference may be due to the present study directly comparing the topology of the trees generated with various alignment algorithms; conversely, previous studies directly scored the alignments generated, in comparison to a "prototype alignment".

Earlier reports (Ogden & Rosenberg 2007; Liu et al. 2009; Sedaghatinia et al. 2009; Löytynoja et al. 2012; Varon & Wheeler 2012), employed either simulated data or simulated differences from a known data set, such that a "true–alignment" and a "true–tree" were known. These provided benchmark tools to score any experimenatlly–obtained output against. However, in the present study, the correct alignment was unknown and therefore we were unable to benchmark the accuracy of the alignment algorithm, its curation and the tree–building method.

For the first time, a simple and universal method, applicable to any locus or organism, shows the different effects, produced by alternative alignment algorithms, gap–treatment options and tree–building methods, on the tree topology. While choosing one tree against another can be difficult, especially when a "reference tree" is unavailable (like in the case of markers less validated than 16S rRNA), this paper highlights the consequences of choosing one workflow over another. Although molecular methods and the resultant classifications (molecular systematics) should be still considered invaluable tools, pitfalls and limitations of these approaches must be kept in mind.

Acknowledgments
The authors wish to thank Prof Michael J Wise (University of Western Australia) for constructive discussions and criticism. This project was funded under the Australian Research Council Linkage Project scheme (LP100100160, UR), with financial– and in kind–support from the Water Corporation of Western Australia, and Murdoch University.

References

Agapow, P.–M. & Purvis, A. (2002): Power of eight tree shape statistics to detect nonrandom diversification: A comparison by simulation of two models of cladogenesis. – Syst. Biol. 51**:** 866–872.

Brosius, J.; Palmer, M.L.; Kennedy, P.J. & Noller, H.F. (1978): Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli. – Proc. Natl. Acad. Sci. U. S. A. 75**:** 4801–4805.

Castenholz, R.W. (2001): Phylum BX. Cyanobacteria. – In: Boone, D.R.; Castenholz, R.W. & Garrity, G.M. (eds.): Bergey's Manual of Systematic Bacteriology (2nd Edition). – pp. 473–599, Springer–Verlag, New York.

Castresana, J. (2000): Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. – Mol. Biol. Evol. 17: 540–552.

Coenye, T. & Vandamme, P. (2003): Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. – FEMS Microbiol. Lett. 228**:** 45–49.

Darriba, D.; Taboada, G.L.; Doallo, R. & Posada, D. (2012): jModelTest 2: more models, new heuristics and parallel computing. – Nat. Methods 9: 772–772.

De Bruyn, A.; Martin, D.P. & Lefeuvre, P. (2014): Phylogenetic Reconstruction Methods: An Overview. – In: Besse, P. (ed.): Molecular Plant Taxonomy: Methods and Protocols. – pp. 257–277, Humana Press, Totowa.

Dereeper, A.; Guignon, V.; Blanc, G.; Audic, S.; Buffet, S.; Chevenet, F.; Dufayard, J.F.; Guindon, S.; Lefort, V.; Lescot, M.; Claverie, J.M. & Gascuel, O. (2008): Phylogeny.fr: robust phylogenetic analysis for the non–specialist. – Nucleic Acids Res. 36: W465–469.

Dress, A.W.M.; Flamm, C.; Fritzsch, G.; Gruenewald, S.; Kruspe, M.; Prohaska, S.J. & Stadler, P.F. (2008): Noisy: Identification of problematic columns in multiple sequence alignments. – Algorithms Mol. Biol. 3.

Edgar, R.C. (2004): MUSCLE: a multiple sequence alignment method with reduced time and space complexity. – BMC Bioinformatics 5: 113.

Fergusson, K.M. & Saint, C.P. (2000): Molecular phylogeny of Anabaena circinalis and its identification in environmental samples by PCR. – Appl. Environ. Microbiol. 66: 4145–4148.

Hammer, Ø.; Harper, D.A.T. & Ryan, P.D. (2001): PAST: Paleontological Statistics Software Package for Education and Data Analysis. – Palaeontol. Electronica 4: 9.

Harrison, C.J. & Langdale, J.A. (2006): A step by step guide to phylogeny reconstruction. – Plant J. 45: 561–572.

Hartmann, S. & Vision, T. (2008): Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? – BMC Evol. Biol. 8: 95.

Holder, M. & Lewis, P.O. (2003): Phylogeny estimation: traditional and Bayesian approaches. – Nat. Rev. Genet. 4: 275–284.

Howard–Azzeh, M.; Shamseer, L.; Schellhorn, H.E. & Gupta, R.S. (2014): Phylogenetic analysis and molecular signatures defining a monophyletic clade of heterocystous cyanobacteria and identifying its closest relatives. – Photosyn. Res. 122: 171–185.

Janse, I.; Meima, M.; Kardinaal, W.E.A. & Zwart, G. (2003): High–resolution differentiation of cyanobacteria by using rRNA–internal transcribed spacer denaturing gradient gel electrophoresis. – Appl. Environ. Microbiol. 69: 6634–6643.

Katoh, K.; Misawa, K.; Kuma, K. & Miyata, T. (2002): MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. – Nucleic Acids Res. 30: 3059–3066.

Kirkpatrick, M. & Slatkin, M. (1993): Searching for evolutionary patterns in the shape of a phylogenetic tree. – Evolution 47: 1171–1181.

Komárek, J. (2006): Cyanobacterial taxonomy: Current problems and prospects for the integration of traditional and molecular approaches. – Algae 21: 349–375.

Komárek, J. (2010): Recent changes (2008) in cyanobacteria taxonomy based on a combination of molecular background with phenotype and ecological consequences (genus and species concept). – Hydrobiologia 639: 245–259.

Landan, G. & Graur, D. (2009): Characterization of pairwise and multiple sequence alignment errors. – Gene 441: 141–147.

Larkin, M.; Blackshields, G.; Brown, N.; Chenna, R.; McGettigan, P.; McWilliam, H.; Valentin, F.; Wallace, I.; Wilm, A.; Lopez, R.; Thompson, J.; Gibson, T. & Higgins, D. (2007): Clustal W and clustal X version 2.0. – Bioinformatics 23: 2947–2948.

Lee, E.; Ryan, U.M.; Monis, P.; McGregor, G.B.; Bath, A.; Gordon, C. & Paparini, A. (2014): Polyphasic identification of cyanobacterial isolates from Australia. – Water Res. 59: 248–261.

Lindgren, A.R. & Daly, M. (2007): The impact of length–variable data and alignment criterion on the phylogeny of Decapodiformes (Mollusca: Cephalopoda). – Cladistics 23: 464–476.

Liu, K.; Nelesen, S.; Raghavan, S.; Linder, C.R. & Warnow, T. (2009): Barking up the wrong treelength: The impact of gap penalty on alignment and tree accuracy. – IEEE–ACM Transactions on Computational Biology and Bioinformatics 6: 7–21.

Löytynoja, A. (2012): Alignment methods: Strategies, challenges, benchmarking, and comparative overview. – In: Anisimova, M. (ed.): Evolutionary Genomics. – pp. 203–235, Humana Press, New York.

Löytynoja, A. (2014): Phylogeny–aware alignment with PRANK. – Methods Mol. Biol. 1079: 155–170.

Löytynoja, A.; Vilella, A.J. & Goldman, N. (2012): Accurate extension of multiple sequence alignments using a phylogeny–aware graph algorithm. – Bioinformatics 28: 1684–1691.

Lyra, C.; Suomalainen, S.; Gugger, M.; Vezie, C.; Sundman, P.; Paulin, L. & Sivonen, K. (2001): Molecular characterization of planktic cyanobacteria of Anabaena, Aphanizomenon, Microcystis and Planktothrix genera. – Int. J. Syst. Evol. Microbiol. 51: 513–526.

McGregor, G.B. & Rasmussen, J.P. (2008): Cyanobacterial composition of microbial mats from an Australian thermal spring: a polyphasic evaluation. – FEMS Microbiol. Ecol. 63: 23–35.

McKenzie, A. & Steel, M. (2000): Distributions of cherries for two models of trees. – Math. Biosci. 164: 81–92.

Miller, M.A.; Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Gateway Computing Environments Workshop (GCE), 14 Nov 2010. 1–8.

Mishra, B. & Thines, M. (2014): siMBa—a simple graphical user interface for the Bayesian phylogenetic inference program MrBayes. – Mycol. Prog. 13: 1255–1258.

Misof, B.; Meusemann, K.; von Reumont, B.M.; Kuck, P.; Prohaska, S.J. & Stadler, P.F. (2014): A priori assessment of data quality in molecular phylogenetics. – Algorithms Mol. Biol. 9.

Morrison, D.A. & Ellis, J.T. (1997): Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of Apicomplexa. – Mol. Biol. Evol. 14: 428–441.

Neilan, B.A. (1995): Identification and phylogenetic analysis of toxigenic cyanobacteria by multiplex randomly amplified polymorphic DNA PCR. – Appl. Environ. Microbiol. 61: 2286–2291.

Ogden, T.H. & Rosenberg, M.S. (2006): Multiple sequence alignment accuracy and phylogenetic inference. – Syst. Biol. 55: 314–328.

Ogden, T.H. & Rosenberg, M.S. (2007): Alignment and topological accuracy of the direct optimization approach via POY and traditional phylogenetics via ClustalW plus PAUP*. – Syst. Biol. 56: 182–193.

Otsuka, S.; Suda, S.; Li, R.; Watanabe, M.; Oyaizu, H.; Matsumoto, S. & Watanabe, M.M. (1999): Phylogenetic relationships between toxic and non–toxic strains of the genus Microcystis based on 16S to 23S internal transcribed spacer sequence. – FEMS Microbiol. Lett. 172: 15–21.

Palenik, B. & Haselkorn, R. (1992): Multiple evolutionary origins of prochlorophytes, the chlorophyllb–containing prokaryotes. – Nature 355**:** 265–267.

Palinska, K.A. & Surosz, W. (2014): Taxonomy of cyanobacteria: a contribution to consensus approach. – Hydrobiologia 740**:** 1–11.

Phillips, M.J.; Lin, Y.H.; Harrison, G.L. & Penny, D. (2001): Mitochondrial genomes of a bandicoot and a brushtail possum confirm the monophyly of australidelphian marsupials. – Proc. R. Soc. London Ser. B: Biol. Sc. 268**:** 1533–1538.

Premanandh, J.; Priya, B.; Teneva, I.; Dzhambazov, B.; Prabaharan, D. & Uma, L. (2006): Molecular characterization of marine cyanobacteria from the Indian subcontinent deduced from sequence analysis of the phycocyanin operon (*cpc*B–IGS–*cpc*A) and 16S–23S ITS region. – J. Microbiol. 44**:** 607–616.

Rambaut, A. (2008): TreeStat [online]. Available from http://tree.bio.ed.ac.uk/software/treestat/.

Rambaut, A. (2014): FigTree [online]. Available from http://tree.bio.ed.ac.uk/software/figtree/.

Rehakova, K.; Johansen, J.R.; Bowen, M.B.; Martin, M.P. & Sheil, C.A. (2014): Variation in secondary structure of the 16S rRNA molecule in cyanobacteria with implications for phylogenetic analysis. – Fottea 14**:** 161–178.

Robertson, B.R.; Tezuka, N. & Watanabe, M.M. (2001): Phylogenetic analyses of *Synechococcus* strains (cyanobacteria) using sequences of 16S rDNA and part of the phycocyanin operon reveal multiple evolutionary lines and reflect phycobilin content. – Int. J. Syst. Evol. Microbiol. 51**:** 861–871.

Ronquist, F.; Teslenko, M.; van der Mark, P.; Ayres, D.L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M.A. & Huelsenbeck, J.P. (2012): MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. – Syst. Biol. 61**:** 539–542.

Sciuto, K. & Moro, I. (2015): Cyanobacteria: the bright and dark sides of a charming group. – Biodivers. Conserv. 24**:** 711–738.

Sedaghatinia, A.; Atan, R.B.; Arifin, K. & Murad, M. (2009): Comparison and evaluation of multiple sequence alignment tools in bioinformatics. – International Journal of Computer Science and Network Security 9**:** 51–56.

Seo, P.S. & Yokota, A. (2003): The phylogenetic relationships of cyanobacteria inferred from 16S rRNA, gyrB, rpoC1 and rpoD1 gene sequences. – J. Gen. Appl. Microbiol. 49**:** 191–203.

Stamatakis, A.; Hoover, P. & Rougemont, J. (2008): A rapid bootstrap algorithm for the RAxML web servers. – Syst. Biol. 57**:** 758–771.

Talavera, G. & Castresana, J. (2007): Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. – Syst. Biol. 56**:** 564 – 577.

Tamura, K.; Peterson, D.; Peterson, N.; Stecher, G.; Nei, M. & Kumar, S. (2011): MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. – Mol. Biol. Evol. 28**:** 2731–2739.

Thompson, J.D.; Higgins, D.G. & Gibson, T.J. (1994): CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position–specific gap penalties and weight matrix choice. – Nucleic Acids Res. 22**:** 4673–4680.

Tomitani, A.; Knoll, A.H.; Cavanaugh, C.M. & Ohno, T. (2006): The evolutionary diversification of cyanobacteria: Molecular–phylogenetic and paleontological perspectives. – Proc. Natl. Acad. Sci. U. S. A. 103**:** 5442–5447.

Valerio, E.; Chambel, L.; Paulino, S.; Faria, N.; Pereira, P. & Tenreiro, R. (2009): Molecular identification, typing and traceability of cyanobacteria from freshwater reservoirs. – Microbiology 155**:** 642–656.

Van de Peer, Y. (2009): Phylogenetic inference based on distance methods. – In: Lemey, P.; Salemi, M. & Vandamme, A.M. (eds.): Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing (2nd Edition). – pp. 142–180, Cambridge University Press, New York.

Varon, A. & Wheeler, W.C. (2012): The tree alignment problem. – BMC Bioinformatics 13.

Whitton, B.A. & Potts, M. (2000): Introduction to the Cyanobacteria. – In: Whitton, B.A. & Potts, M. (eds.): The ecology of cyanobacteria. Their diversity in time and space. – pp. 1–11, Springer–Netherlands, Rotterdam.

Willame, R.; Boutte, C.; Grubisic, S.; Wilmotte, A.; Komarek, J. & Hoffmann, L. (2006): Morphological and molecular characterization of planktonic cyanobacteria from Belgium and Luxembourg. – J. Phycol. 42**:** 1312–1332.

Yang, Z. & Rannala, B. (2012): Molecular phylogenetics: principles and practice. – Nat. Rev. Genet. 13**:** 303–314.

Received October 16, 2015
Accepted December 22, 2015