Gilbert *et al. BMC Genomics* 2011, **12**:456
http://www.biomedcentral.com/1471-2164/12/456

RESEARCH ARTICLE Open Access

Array Comparative Genomic Hybridizations: Assessing the ability to recapture evolutionary relationships using an *in silico* approach

Luz B Gilbert[1*], Lee Chae[2], Takao Kasuga[3] and John W Taylor[4]

Abstract

Background: Comparative Genomic Hybridization (CGH) with DNA microarrays has many biological applications including surveys of copy number changes in tumorogenesis, species detection and identification, and functional genomics studies among related organisms. Array CGH has also been used to infer phylogenetic relatedness among species or strains. Although the use of the entire genome can be seen as a considerable advantage for use in phylogenetic analysis, few such studies have questioned the reliability of array CGH to correctly determine evolutionary relationships. A potential flaw in this application lies in the fact that all comparisons are made to a single reference species. This situation differs from traditional DNA sequence, distance-based phylogenetic analyses where all possible pairwise comparisons are made for the isolates in question. By simulating array data based on the *Neurospora crassa* genome, we address this potential flaw and other questions regarding array CGH phylogeny.

Results: Our simulation data indicates that having a single reference can, in some cases, be a serious limitation when using this technique. Additionally, the tree building process with a single reference is sensitive to many factors including tree topology, choice of tree reconstruction method, and the distance metric used.

Conclusions: Without prior knowledge of the topology and placement of the reference taxon in the topology, the outcome is likely to be wrong and the error undetected. Given these limitations, using CGH to reveal phylogeny based on sequence divergence does not offer a robust alternative to traditional phylogenetic analysis.

Background

The field of comparative genomics, particularly in microbes, has benefited greatly by the proliferation of whole genome sequences. The advantages of having sequences from multiple, related organisms range from improving annotation to characterizing the genetic basis of major phenotypic differences between strains or species [1-4]. While there are organisms for which multiple sequences from several different strains or species are available http://www.genomesonline.org/, in many cases resource limitations restrict the number of sequencing projects for members of the same genus. An appealing alternative to characterize sequence polymorphisms among related organisms is array Comparative Genome Hybridization (array CGH). This technique is attractive because microarrays made from genome sequence or even random DNA fragments from just one individual can be used to study the phylogenetic relationships among many closely related species.

Array CGH (aCGH) for two color array platforms uses DNA samples from a reference individual and a test individual, each labelled with a different fluorescent dye, and competitively hybridizes them to an array composed of immobilized DNA fragments from the reference individual [5-8]. This technique has primarily been used to characterize gene copy number changes and deletion events and has been applied extensively in the study of human tumorogenesis and bacterial pathogens [9-14].

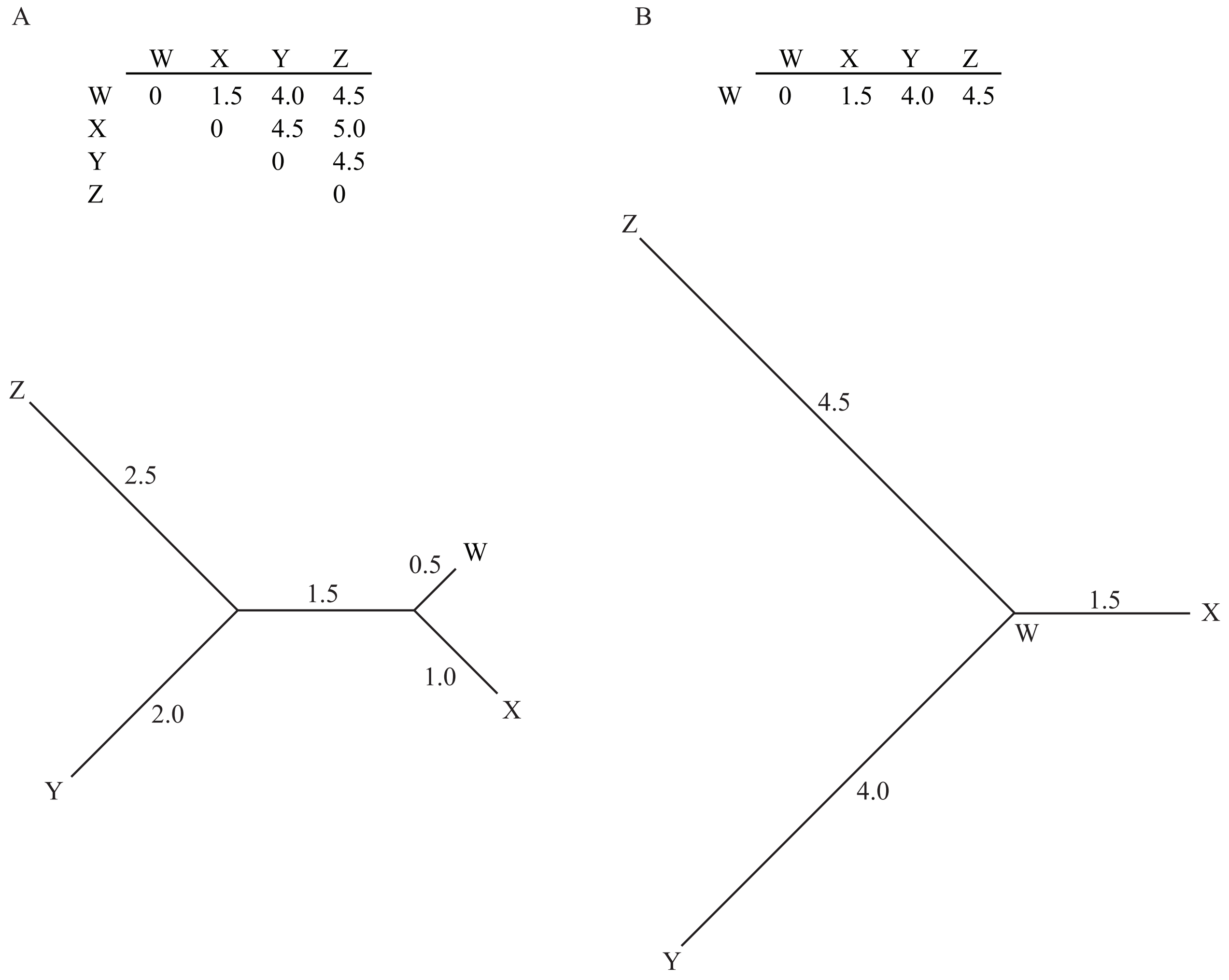
The relative ease with which aCGH provides large amounts of discriminating information between individuals makes it a very attractive technique to determine relatedness. Comparisons between aCGH derived trees and trees based on DNA sequence of one or a few loci

* Correspondence: gilbert@lrsv.ups-tlse.fr
[1] Laboratoire de Recherche en Sciences Végétales, UMR CNRS-Université Paul Sabatier 5546, Chemin de Borde Rouge - Auzeville 31326, Castanet Tolosan, France
Full list of author information is available at the end of the article

have supported this assumption [15-18]. However, the commonly used ribosomal DNA sequences do not necessarily provide enough resolution at the species and subspecies level to accurately resolve a tree [19,20]. Indeed, it has been shown that a detailed multi-locus phylogenetic analysis is often necessary to accurately resolve a species tree [21-27].

Practitioners of aCGH have inferred evolutionary relationships using distance methods implemented by programs such as Cluster, Genespring, and Acuity [28-30]. Several studies have used aCGH to compare bacterial pathogens and infer relatedness of strains from traditional cluster analysis [31-37]. A recent study utilized both human and bovine clustered aCGH data to make inferences on evolutionary relatedness [38]. Other studies have applied distance and parsimony techniques to build phylogenies from their array data [17,39-44]. Another compared MLST to aCGH trees derived by Bayesian inference [37].

A complicating factor not addressed in these studies is that typically a single individual is represented on an array. Having a single reference species or strain is at odds with traditional sequence based phylogenetic distance methods, where all pairwise comparisons among the taxa are used in constructing a species tree. This difference is illustrated in Figure 1. Having a single reference (Figure 1B) provides direct comparisons between all taxa and the reference taxon W, but provides no direct information about the relationships among the non-reference taxa X, Y, and Z. The assumption made for array CGH phylogeny is that the massively parallel nature of the technique provides sufficient evolutionary signal to infer the true relationships between all the taxa, in a manner similar to clustering analysis of gene



	W	X	Y	Z
W	0	1.5	4.0	4.5
X		0	4.5	5.0
Y			0	4.5
Z				0

	W	X	Y	Z
W	0	1.5	4.0	4.5

Figure 1 The single reference design. 1A shows an example of a pairwise distance matrix for the tree below the matrix. In 1B the distances of other taxa to W, typical of a single reference design. Below it is the star topology constructed from these distances when no other information is available to differentiate distances among X, Y, and Z.

expression data [28]. A further complication is that genes absent in the reference strain cannot be included in the analyses.

In this study we ask, what effect does having a single reference taxon have on recovering phylogenetic relationships? What effects do varying amounts of sequence divergence among taxa have on recovering the "true" tree? What impact do different tree topologies have? Does the position of the reference taxon in that topology have any importance or do all taxa serve equally well as the reference? Are distance and parsimony tree construction methods equivalent when applied to this type of data?

In order to address these questions, we chose to simulate CGH array data to avoid the added complexity of experimental noise associated with microarray technology. We chose the *Neurospora crassa* genome as a starting point for our simulated data, and evolved the sequence based on one of three simple topologies using empirically determined rates of nucleotide substitution. We modelled a very conservative system- allowing only nucleotide substitution, not gene gain or loss as that would complicate the tree-building process to generate optimal data with a high but realistic amount of phylogenetic signal. This *in silico* approach allowed us to vary all of the parameters necessary to test these questions, e.g., tree topology, position of reference taxon, and mean substitution rate and its variation, which would be impossible with any single set of empirical data.

Methods

CGH Data Simulation

We chose 70 mer sequences to represent the length of probes in a standard long oligo array platform because they have been a popular choice for groups investing in arrays for their chosen organism [7,45,46]. From the genome sequence of the filamentous fungus *Neurospora crassa*, 70 mer oligonucleotides were designed to provide hybridization probes for each of the 10,200 ORFs in the release 3 annotation, with an additional 300 probes devoted to non-coding regions [47-50]. This set of probe sequences was designated as the ancestral sequence and, along with a phylogenetic tree specifying the relationships and genetic distances among the taxa, was input into the sequence evolution program ROSE v1.3 to evolve sequences for related taxa [51].

Because not all genes in a genome evolve at the same rate, empirical data were analyzed to determine an appropriate distribution of sequence divergence to evolve our sequences. Based on the whole genome comparisons for different yeast species [4,52], a normal curve with a range of average polymorphisms was approximated to model sequence heterogeneity among related taxa across all coding regions. Three values of polymorphism appropriate for the detection limits of a 70 mer array, were chosen to model the distribution, 5 ± 1.44%, 7.5 ± 2.2%, and 10 ± 2.9%. The standard deviation of this distribution was empirically derived from the publicly available comparative genome analysis of different yeast species [1].

To simulate variation in evolutionary rate among genes, for each 70 mer, the branch lengths of the input tree were multiplied by a scaling factor, randomly drawn from the distribution of evolutionary rates. By varying the mean and standard deviation of the distribution of evolutionary rates, we varied the amount of evolutionary distance between the taxa of our chosen tree topology. These parameters were used to evolve sequences for each of 10,500 genes for each taxon using the Jukes-Cantor model of evolutionary change, allowing nucleotide substitution but not indels. This simple model of evolutionary change models a scenario where species are close relatives and the problem of multiple substitutions at single nucleotide positions is minimal, i.e. where the signal to noise ratio is highest. The genetic distances between all pairs of taxa were calculated for each locus using dnadist from the PHYLIP package v3.6 [53]. These distances are used to approximate DNA-DNA hybridization levels.

To simulate the use of a microarray based on a single reference taxon, one species from each tree was chosen as the reference and only those pairwise comparisons involving the reference individual were saved, corresponding to the genetic distance from the reference species to all other taxa. These distances were combined into one supermatrix with 10,500 genes for *n* taxa. This process is illustrated in Figure 2. This supermatrix represents the experimental design of a typical single-reference CGH experiment and was the basis for phylogenetic tree construction using PAUP version 4.0b10 for unix [54]. Additionally, as a control, sequence alignments for all taxa were concatenated and used to calculate, using PAUP, both distance based and parsimony trees.

To evaluate the effect of using much longer oligomers to construct an array, as with cDNA arrays, a limited number of simulations were conducted for probes of 500 bases. The results using the longer arrays were not substantially different from those of the 70 base trials (see supplemental data).

Tree Construction

To construct a distance tree, a Pearson correlation-based distance matrix and a Euclidean distance matrix were derived from the original supermatrix using the statistical package R (http://www.R-project.org, R Development Core Team, 2006). Neighbor-joining was used to make phylogenetic trees based on these matrices, employing the minimum evolution criteria in PAUP

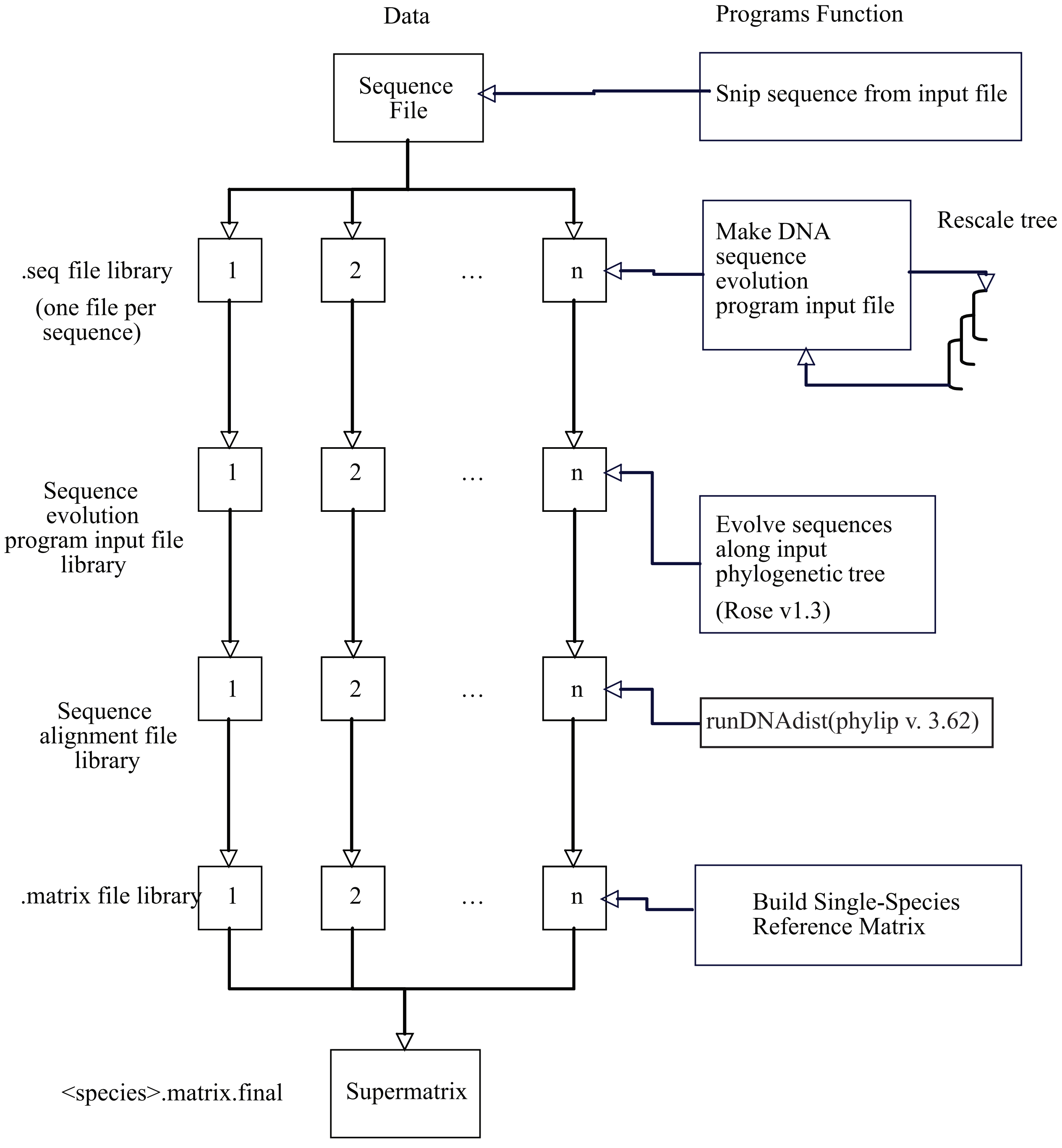

Data
Programs Function
Sequence File
Snip sequence from input file
.seq file library
(one file per sequence)
1
2
…
n
Make DNA sequence evolution program input file
Rescale tree
Sequence evolution program input file library
Evolve sequences along input phylogenetic tree
(Rose v1.3)
Sequence alignment file library
runDNAdist(phylip v. 3.62)
.matrix file library
Build Single-Species Reference Matrix
<species>.matrix.final
Supermatrix

Figure 2 Program Flow. The diagram indicates the simulation program flow used to construct the supermatrix, our simulated CGH data matrix. Variables that can be modified include the number of sequences input, the topology of the tree used, and the amount of divergence among taxa in the simulated alignment, and the choice of evolutionary model to evolve and calculate genetic distance among taxa (see methods). Neighbor-Joining (NJ) and Parsimony Majority-Rule (PMR) trees were constructed from the supermatrix.

[54]. To convert simulated DNA hybridization levels to characters with discrete states (0,1), simulated hybridization levels above the third quartile, representing the most divergent genes, were converted to 0 and those below to 1. This simple discretization method was amongst the most powerful when considering various gene categorization methods in our accompanying experimental analysis and was the simplest to implement *in silico*. Maximum parsimony was then used to infer phylogenetic tree topologies [54].

Tree to tree distance metric quantification

To assess the outcome of a simulation, we compared the phylogenetic tree used to produce the data sets to the phylogenetic tree recovered from the CGH simulation. Initially, we measured the differences between the input tree and that produced by the simulation with two tree-to-tree distance metrics - symmetric distance and agreement subtree - implemented in PAUP [54]. Symmetric distance (SymD) determines the number of branches that must be rearranged or collapsed to make two topologies identical. If the topologies are identical and no branches need to be moved, the step size is 0. One step indicates a single collapsed branch and a rearrangement between two taxa is scored as two steps [55]. The agreement subtree metric, D1, counts the number of taxa that must be pruned from the trees make their topologies identical [56].

Cophenetic correlation distributions

Neither the symmetric nor the agreement subtree metric compares branch lengths between the input and output trees. To investigate differences in branch length, we calculated the cophenetic correlation coefficient (CCC, from the cophenetic.phylo and cor functions implemented in the R stats and ape packages [57,58]), which measures how well a tree fits the data used to create it on a scale of 0 to 1, with 1 indicating a complete fit.

In this work, we used the CCC to compare the tree topology used to initiate the simulated data set (input tree) with the distance matrices resulting from the various CGH analyses. The R functions extrapolate distances from the branch lengths from the starting topology to compare to the distance matrix estimated by the simulation. The converse comparisons, where tree topologies resulting from simulations are compared to a distance matrix calculated from the starting topology, were very similar to the first comparisons and are not presented here.

Results

The efficacy of CGH as a phylogenetic method as judged by two metrics (SymD or D1) varied considerably with topology (balanced, pectinate or empirical *Neurospora*), tree building algorithm (NJ with Pearson or Euclidian distances, or parsimony), and location of the reference taxon relative to the other taxa (basal or derived positions). The numerical scores for both tree-building methods are provided as Additional Files 1 and 2: Tables S1 and S2.

The parameter with the least effect was substitution rate (5%, 7.5% or 10%). The simplest topology, the balanced tree, was recovered most successfully. This special case is not likely to be found in nature, and it should be noted that permuting the branches individually may lead to different results. For the other two topologies, the mixture of close and more distantly related taxa presented a complication for the different tree-building algorithms. For both the pectinate topology and the *Neurospora* phylogeny, the relationship of the reference taxa to its sister taxon was not consistently well resolved. For the pectinate tree, the output tree differed from the input tree from between 0 to 8 steps (SymD) or 0 to 2 for distance D1, with scores growing progressively worse based on the choice of reference taxon. For the *Neurospora* topology, results also vary for the different reference positions and for the various tree-building algorithms, with step differences between input and output trees of 0 to 10, (SymD) or 0 to 4 (D1).

Generally, the parsimony trees were more sensitive to the changes in average nucleotide substitutions than those made by the Neighbor-Joining method. When sequence divergence was high (10%), parsimony phylogenies were poorly resolved. At 10% divergence, we checked the effect of changing the threshold for considering a hybridization difference to be diverged, that is, from 1 (present) to 0 (absent). Altering the threshold from 75% (where the most diverged 25% of the genes were given a score of 0) to 20%, 50%, 80% or 90%, however, failed to significantly improve the performance of parsimony analysis (see Additional File 2: Table S2).

To report the details of our simulation we have organized the results to feature the input tree topology and tree-building algorithm. The four major variables tested in this simulation: tree topology, placement of reference taxon, tree building algorithm, and substitution rate, are discussed separately as much as possible given their inherent interrelationships.

Balanced topology (Figure 3)

Figure 3A shows an eight taxa tree in a balanced topology. The balanced topology is the simplest in terms of the relationships among the taxa because no matter which taxon is chosen as the reference, it will have a close sister taxon, a pair of taxa at a middle distance and two, most distant, taxon pairs. The three distances, sister, middle and distant, will not change as the reference taxon changes. Therefore, we expected to find no effect of choosing different reference taxa and, as can be seen in Figure 3B-D, the symmetric distances comparing input and output trees were identical when taxon A or taxon H was chosen as the reference taxon. This simple topology is recovered with moderate success with both distance-based NJ and parsimony tree construction but there are differences in the efficacy of the different tree construction algorithms and simulated substitution rates.

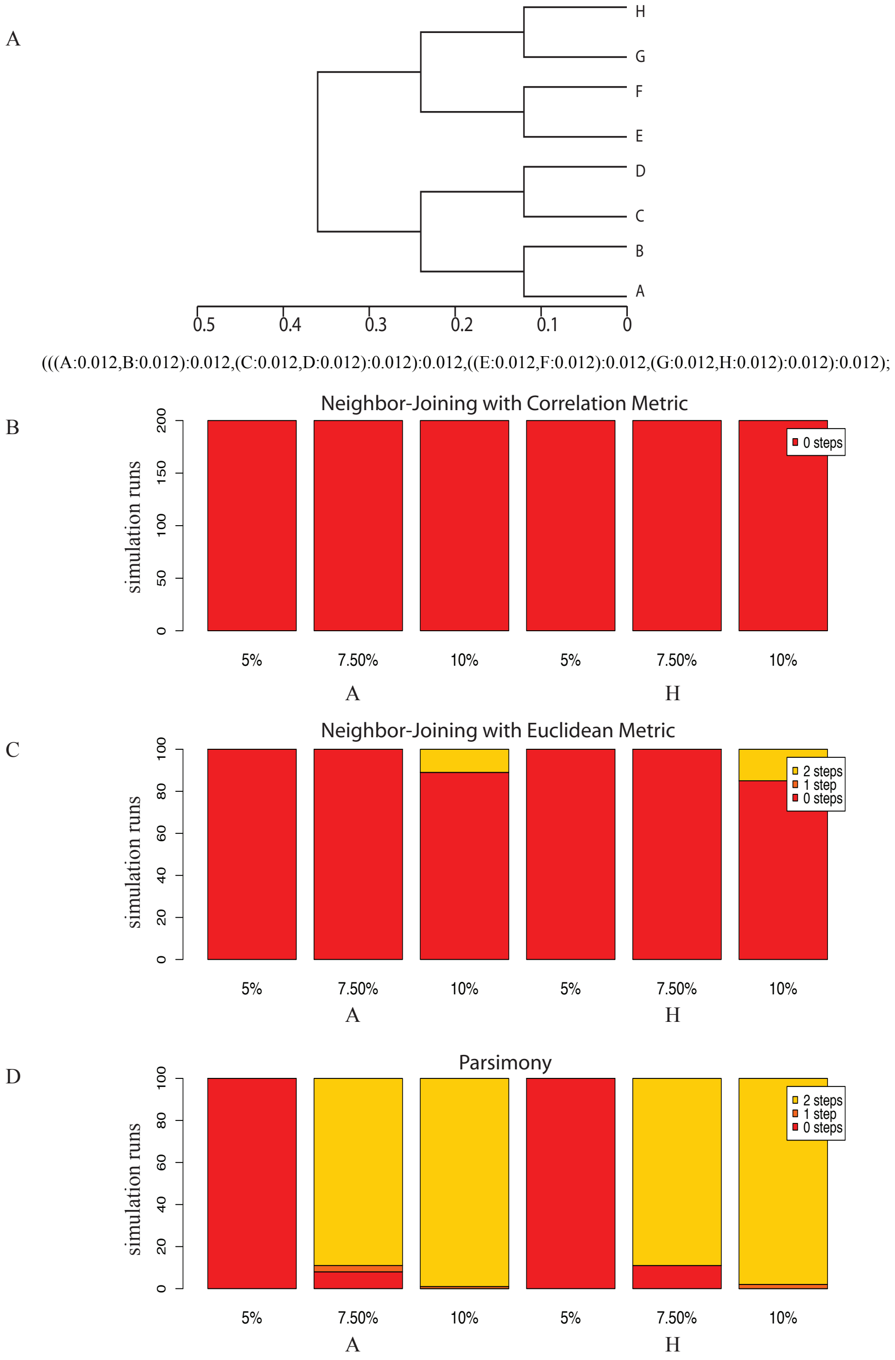


Figure 3 Balanced Topology Results. Figure 3A is a cladogram of an eight taxa balanced tree topology. In our simulation a sequence alignment was created using this input tree as a guide, where the branch lengths govern the number and pattern of substitutions. A number randomly drawn from a Gaussian distribution, empirically determined from published yeast genomic data, was used to multiply these branch lengths to produce a set of genes evolved for a range of substitution rates. The average of each range of substitution rates is indicated under each column. Figure 3B shows the symmetric distances for 200 simulation runs of the correlation-based NJ analysis evolved at the average sequence divergence indicated. Columns 1-3 use taxon A as the reference, Columns 4-6 use taxon H. Zero steps indicate perfect agreement between the input topology and the trees output by the simulation (red portions). One step away indicates a single collapsed branch (orange portions). A mispairing of two taxa is given a score of two (yellow portions). Scores in subsequent figures are cumulative and color-coded as indicated in the chart legends. Figure 3C shows the stacked histogram for the symmetric distances from the input topology for 100 replicate NJ trees calculated with the Euclidean distance metric. The order for reference taxa is the same as above. The stacked histogram in Figure 3D is of the symmetric distances from the input tree for 100 replicate 50% Majority-Rule Parsimony (PMR) trees. An example of this topology, two steps away from the reference, is given in Figure 4B.

NJ with the Pearson and Euclidean distance metric

When NJ was used with the Pearson correlation-based distance metric to make phylogenetic trees, comparison of input and output tree topologies by either symmetric distance or distance D1 showed no difference between the two topologies, regardless of nucleotide substitution rate (Figure 3B). When NJ was used with the Euclidean distance metric, there were discrepancies between the input and output topologies, but only at the highest rate of nucleotide substitution (Figure 3C).

Parsimony Analysis

With parsimony analysis (Figure 3D), substitution rate had a more dramatic effect. At 5% nucleotide substitution, there was no difference between the input and output topologies in all simulations, but, at substitution rates of 7.5% and 10% the input tree was recovered in no more than 8 of 100 simulations. At 7.5% and 10% sequence divergence, the pair of sister taxa closest to the reference taxon did not group together in more than 95% of simulations. Instead, these sister taxa collapsed to the base of the group as a polytomy or they resolved into separate lineages more than 95% of the time, as shown in the example in Figure 4A and 4B.

Pectinate topology (Figure 5)

This 12-taxon tree (Figure 5A) shows a modified pectinate form in which at most four taxa have the same relationships to the other members of the tree, i.e., S1 to S4. Each of these four taxa has a close sister taxon and then shows increasingly longer distances to four other taxon pairs. This situation is in marked contrast to the balanced tree in Figure 3A, where all members had the same relationships to the other members. To assess the effect of the position of the reference taxon, we designated three different taxa as the reference, S1, S6, and S12 (Figures 5B, C, and 5D).

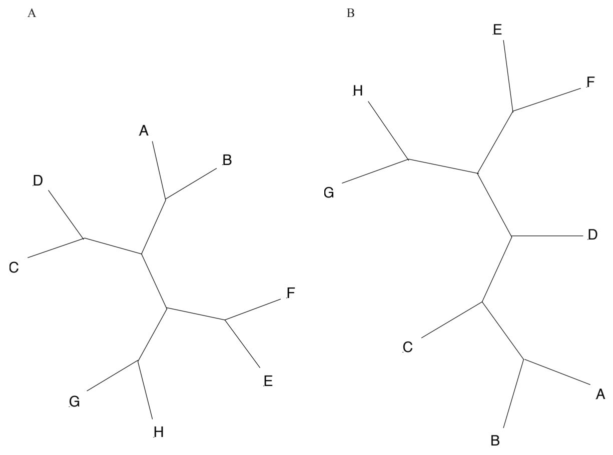


Figure 4 The expected and recovered balanced topologies. Figure 4A illustrates the eight taxa balanced topology as an unrooted cladogram. Figure 4B is the Euclidean NJ tree, when taxon A is the reference, as an unrooted cladogram and is two steps away according to the symmetric distance from the desired topology in Figure 4A. It was evolved at 7.5% average nucleotide substitutions.

NJ with the Pearson and Euclidean distance metric

The simulations show that the pectinate topology is not recovered as frequently as the balanced topology. NJ using the Pearson correlation-based distance recovered the input tree most frequently, with almost 2/3 of the trees 0 steps away. Choice of the reference taxon had a strong effect with the pectinate tree. For example, with NJ using Pearson's correlation distances, the input tree was recovered at least 95% of the time when the most basal (S12) or the most derived taxon (S1) were used as reference, but never when an interior taxon (S6) was designated as reference (Figure 5B).

With NJ using Euclidean distances, the input topology was recovered in 100% of simulations when the most derived taxon was the reference taxon, but almost never when either an internal (S6) or basal (S12) taxon was designated as reference (Figure 5C). With NJ using Euclidean distances and with parsimony, input trees were recovered 100% of the time when the most derived taxa were the reference, but never when either internal or basal taxa were designated the reference taxon. In fact, use of the basal taxa as reference made the output trees at least 6 steps different from the input trees with both NJ using Euclidean distance or parsimony.

Parsimony Analysis

With parsimony analysis, the input topology was recovered in all simulations when the most derived taxon was the reference, but never when an internal or basal taxon was so designated (Figure 5D). The parsimony analysis (Figure 5D) behaved very similarly to the Euclidean-based NJ analyses (Figure 5C). However, parsimony analysis was marginally worse than NJ because taxa S1 and S2 occasionally collapsed into a polytomy with the S3-S4 pair, Which increased the number of steps between input and output trees (SymD), as did higher substitution rates. Example trees are given in Figure 6.

Neurospora topology (Figure 7)

The *Neurospora* phylogeny is based on observed distances among species in this genus and is used here as an example of a natural topology to begin the simulation process [24]. This 11 taxon phylogeny combines elements of the balanced and pectinate topologies (Figure 7A in that it contains a mixture of close and distantly related taxa, some of which are connected with small internal branches. These features present a challenge for array CGH tree-building algorithms.

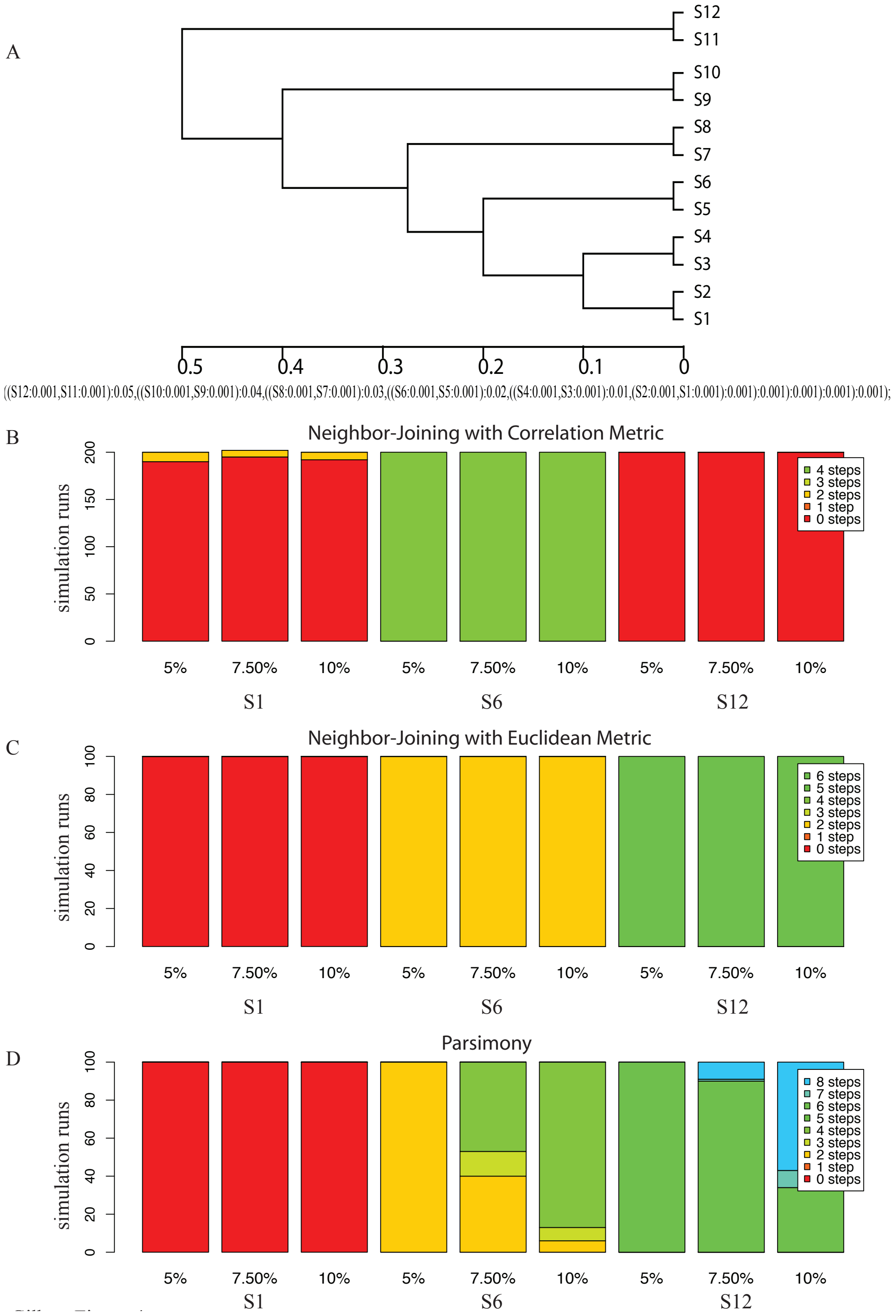


Figure 5 Pectinate Topology Results. Figure 5A shows the twelve taxa pectinate topology input into the simulation pipeline. Figure 5B shows the stacked histograms for the symmetric distances for 200 replicate correlation-based NJ. Columns 1-3 show S1 as the reference taxa, 4-6 show S6, and 7-9 represent taxon S12. Figure 5C give the stacked histogram for the symmetric distance for the Euclidean distance based NJ trees, 100 replicates. The order for reference taxa is the same as above. Figure 5D gives the corresponding stacked histogram for 100 replicate PMR trees. Two examples of this topology, six and eight steps away from the reference, are given in Figure 6B and 6C.

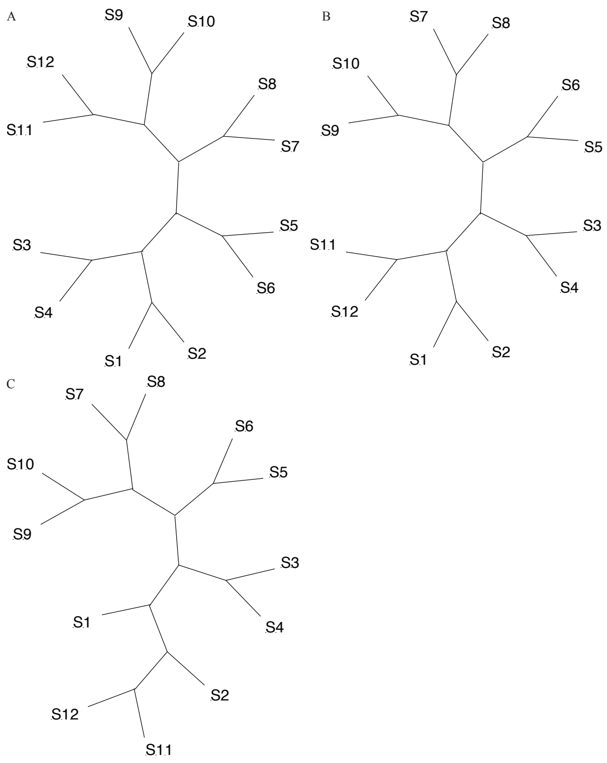


Figure 6 The expected and recovered pectinate topologies. Figure 6A shows the twelve taxon pectinate topology drawn as an unrooted cladogram. 6B show an example where S12 acts as the reference for the NJ Euclidean method, it is 6 steps away according to the symmetric distance from the reference tree in A. C shows another example tree where S12 is the reference but for the parsimony algorithm, and is eight steps away. Both 6B and 6C were evolved at 7.5% average nucleotide substitutions.

As with the other topologies, we evaluated the effects of choosing different reference taxa, *N. discreta, N. crassa* and *N. tetrasperma*, in addition to the effects of different methods of phylogenetic analysis and different rates of nucleotide substitution. In summary, the phylogenetic method had the strongest effect on recovery of the input topology. NJ with Pearson's correlation (Figure 7B) performed far better than either the NJ with Euclidean distance or parsimony methods. Neighbor-Joining with Euclidean distances (Figure 7C) never recovered the input phylogeny and with parsimony (Figure 7D), the input tree was almost never recovered.

NJ with the Pearson and Euclidean distance metric

With NJ using Pearson's correlation (Figure 7B), designating the basal taxon (*N. discreta*) as reference resulted in recovery of the input tree 100% of the time. The same result was seen when *N. tetrasperma* was designated reference, but only 64% of simulations did when *N. crassa* A was designated reference. Instead, output trees were occasionally two to four steps different from the input tree, resulting from a failure to resolve short internal branches separating branches among taxa within clade 2 and by placing the long branch of *N. discreta* incorrectly. When *N. tetrasperma* or *N. discreta* was the reference, these short connective branches were resolved almost 100% of the time. When the nucleotide substitution rate was increased from 5% to either 7.5% or 10%, a greater percentage of simulations recovered the input topology, suggesting that more polymorphisms help differentiate those short branches for this distance metric.

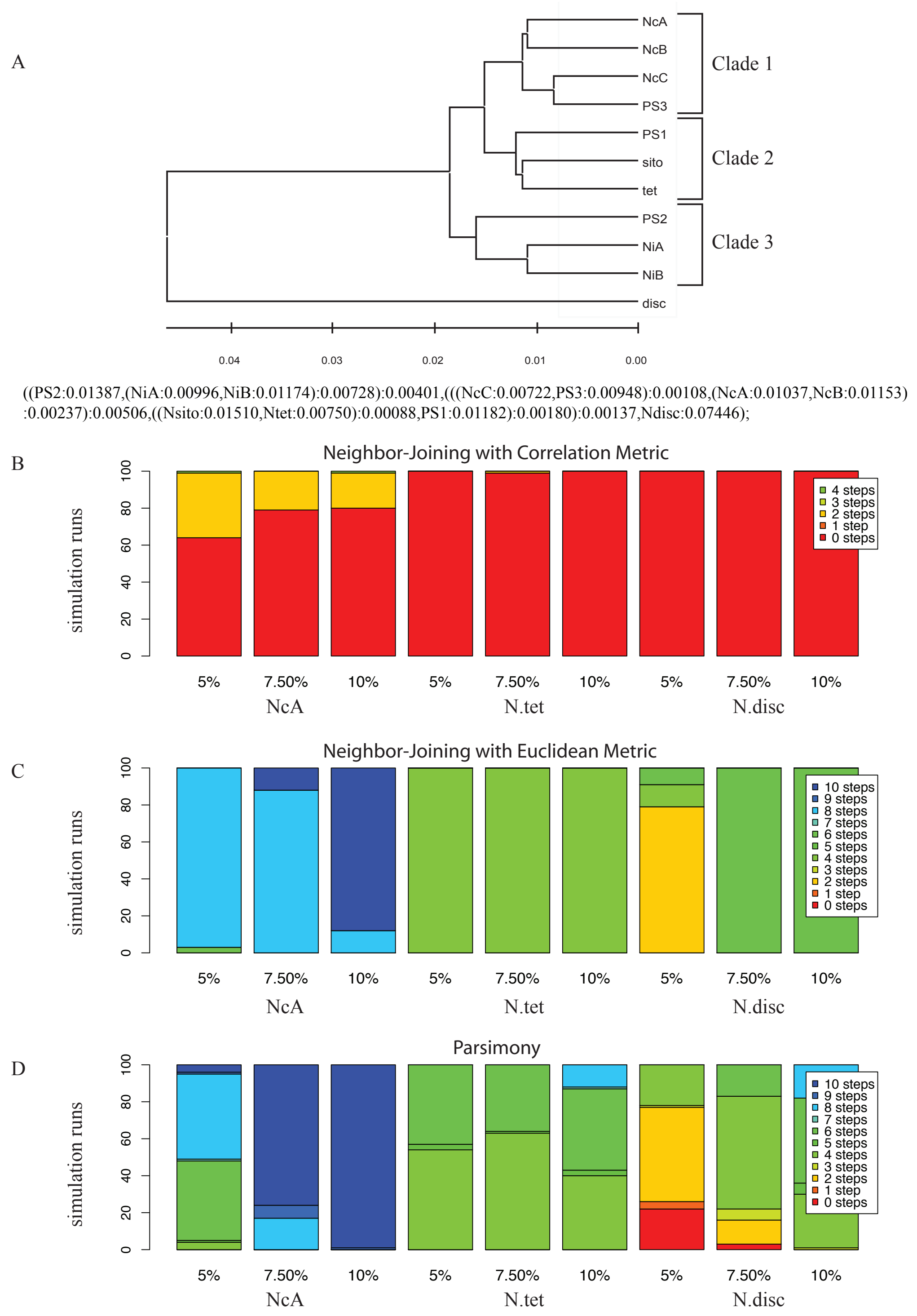
With NJ using Euclidean distance, the input topology was never recovered and the majority of output trees were 4 or more steps different from the input topology (Figure 7C). In these trees, many sister taxa relationships devolved into polytomies when small internal supporting branches were left unresolved. When *N. discreta* was designated as the reference taxon, a higher proportion of CGH trees were more similar to the input tree - especially when the nucleotide substitution rate was lowest. *N. crassa* A performed most poorly as the reference with most output trees at least 6 and as many as 10 steps distant from the input tree (Figure 8A and 8B). When *N. tetrasperma* was the reference, the short internal branches remained unresolved, and when either *N. crassa* A or *N. tetrasperma* were the reference taxon, there was a tendency to erroneously group *N. discreta* with clade 3 (Figure 8C). At low mutation rates, *N. discreta* and *N. crassa* A performed better as the reference, although the discrepancies between the input and output trees worsened as the substitution rates increased. (Figure 8D). Close examination of these trees suggests that a saturation of polymorphisms made shorter branches difficult to resolve.

Parsimony Analysis

With parsimony, again the input tree was recovered only when the most basal taxon, *N. discreta*, was reference and then only 22% of the time at 5% nucleotide substitution and only 3% of the time at 7.5% nucleotide substitution. When *N. crassa* A or *N. tetrasperma* were designated as reference taxa, parsimony analysis was unable to correctly place the most divergent taxon, *N. discreta*. For these trees, *N. discreta* was incorrectly placed in a polytomy within clade 3 instead of outside of it. With increasing sequence differences, *N. discreta* was placed within clade 2 or between clades 1 and 2 (Additional File 3: Figure S1).

Probe Length

Increasing the length of the probe sequence in the microarray from 70 nt to 500 nt might be expected to improve the recovery of input topologies by providing more polymorphisms for phylogenetic analysis. To test this hypothesis, 25 replicate runs at 10% average



((PS2:0.01387,(NiA:0.00996,NiB:0.01174):0.00728):0.00401,(((NcC:0.00722,PS3:0.00948):0.00108,(NcA:0.01037,NcB:0.01153):0.00237):0.00506,((Nsito:0.01510,Ntet:0.00750):0.00088,PS1:0.01182):0.00180):0.00137,Ndisc:0.07446);

Figure 7 Neurospora Topology Results. Figure 7A gives the topology of the natural phylogeny for the eight conidiating species of *Neurospora*, modified from Dettman 2003a. Figure 7B gives the stacked histogram shows the proportion of NJ trees in 200 replicates that are 0 or more steps away from the input topology. The first three columns show the symmetric distance using *N. crassa* clade A (NcA) as the reference taxa for the average sequence divergence indicated. Columns 4-6 show the results using *N. tetrasperma* (tet) as the reference taxa. Columns 7-9 use *N. discreta* (disc) as the reference. Figure 7C gives the stacked histogram of the symmetric distance for 100 NJ trees built with the Euclidean distance metric. No trees were less than 2 or more steps away. Figure 7D gives the stacked histogram showing the proportion of PMR trees in 100 replicates that are 0 or more steps away from the input topology. Three examples of this topology, between four to eight steps away from the reference, are given in Figure 8B, 8C, and 8D.

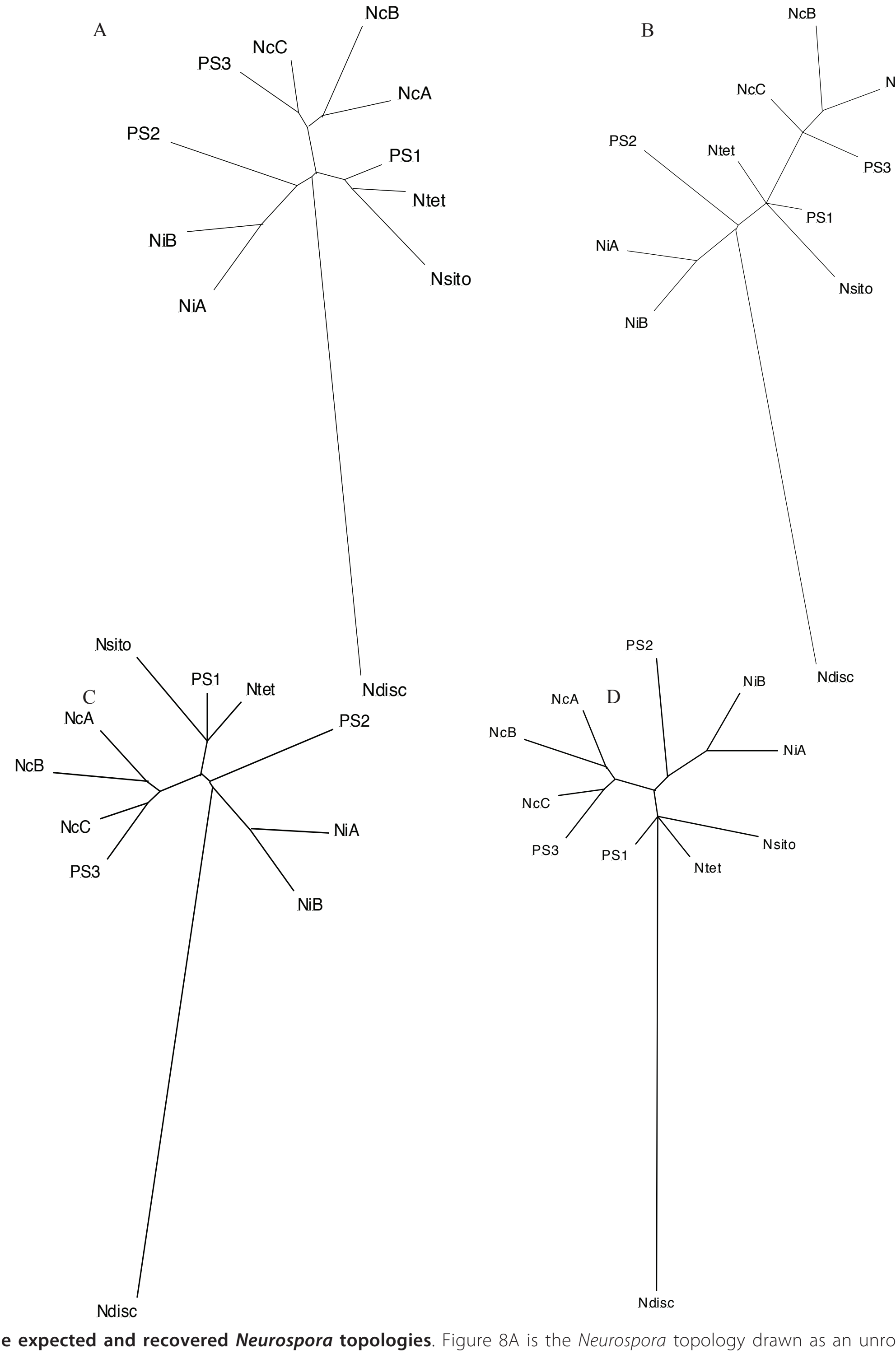


Figure 8 The expected and recovered *Neurospora* topologies. Figure 8A is the *Neurospora* topology drawn as an unrooted phylogram. Figure 8B shows an example were *N. crassa A* is the reference for the Euclidean NJ that is 8 steps away from A according to the symmetric distance. Figure 8C shows an example with *N. tetrasperma* as the reference for the Euclidean NJ that is 4 steps away from the reference according to the symmetric distance. Figure 8D shows an example with *N. discreta* as the reference that is 6 steps away according to the symmetric distance by the Euclidean NJ method. B, C, and D were evolved at 7.5% average nucleotide substitutions.

sequence divergence were completed using 500 bases as the length of the probe for the Neurospora topology. There was no significant improvement in recovery of the input topology regardless of which taxon was designated as the reference taxon with the longer alignments (Additional File 2: Table S2).

Cophenetic Correlations

To evaluate differences in the branch lengths, irrespective of topology, of the input and output trees obtained by NJ using Pearson's correlation or Euclidean distances, we determined the cophenetic correlation coefficient (CCC) for the *Neurospora* distance trees (Figures 9 and 10). When assayed by topology, NJ with Pearson's correlation was more successful than the Euclidean in finding the topologically correct tree. However, when assayed by the CCC, the Euclidean method was better at estimating the branch lengths. This result was true not only for the *Neurospora* tree, but also for the pectinate tree. For the balanced tree, the CCC was slightly better for the NJ using Pearson's correlation than with Euclidean distance. The CCC also varied moderately with the choice of reference taxon and slightly with the percent substitution rate. Distributions for the other topologies are included in the Additional Files 4, 5, 6, and 7: Figures S2, S3, S4, and S5.

Discussion

Array Comparative Genomic Hybridization (array CGH) is a technique in which the genomic DNAs of a reference individual and a close relative are hybridized to a microarray made of probes designed to match the genome of the reference individual. To assess the appropriateness of using microarray data for phylogenetic analysis, we asked if having a single reference species, to which all other taxa must be compared, limits the ability to accurately recover evolutionary relationships. In our simulations, to make phylogenies from CGH data, we used both distance and parsimony phylogenetic methods. We assessed the effects of tree topology, the position of the reference taxon in the topology, the rate of nucleotide substitution, and the length of microarray probes. We assessed the difference between the input and output phylogenies using two tree-to-tree distance metrics and also assessed the difference in rates of evolution for two topologies as measured from branch length differences.

Our results show that under specific conditions, using CGH to quantify inter-genomic sequence variation can yield data that support the input topology. However, it is just as common to get a tree with little resemblance to the true tree. The user would be unaware of this, as studies often have no prior knowledge of two key parameters, the topology and the position of the reference taxon. Unfortunately, these are the goals that phylogenetic inference is designed to determine. CGH, however, may have a role to play in assessing sequence divergence, by identifying genes that are evolving rapidly, but that are still present across taxa, that would be good candidates for multi-locus sequence analysis.

Summary of Simulation Results

Tree topology and the position of the reference taxon had the greatest effects on successful recovery of the input tree by CGH. The balanced topology was recovered more successfully than the pectinate or *Neurospora* topology, probably because all taxa occupy the same position relative to their neighbors in the balanced topology. For the pectinate topology, CGH was most successful with reference taxa in the most derived position. This result is not unexpected, as the single reference design of most CGH arrays constricts the dataset in such a way that discards information that would resolve irregular distances. In effect it is similar to constricting multi-dimensional scaling data to a single plane- that in this case is dictated by the choice of reference taxon. For the *Neurospora* topology, CGH success was poorest with the reference in a derived position, but for this topology the analytical method had a stronger influence than the position of the reference taxon.

In general, the NJ distance method was more successful at recovering the input topology than was parsimony, and distance data matrices made using Pearson's correlation coefficient performed better than those made by Euclidean distance. For the correlation, in the tree-building process it was found to be more expedient to exclude the reference as the correlation method was not able to cope with the reference taxon's lack of variability in the evolved data. However, branch lengths were better recovered with NJ using a Euclidean distance matrix than with the Pearson's coefficient. While the Pearson's correlation is more robust to missing data and is less sensitive to small variations than Euclidean distance, the Euclidean potentially preserves more information about the genetic distance [59].

For parsimony analysis all "species" hybridization values were binned using a single cutoff, a simple approach we found adequate for distribution of sequence polymorphisms simulated here. Modifying the threshold for assigning shared character states (1,1) from the 75th percentile to the 80th percentile or higher occasionally improved the results, but only for some locations of the reference taxon and topologies.

Varying the rate of nucleotide substitution had little effect on recovery of the input topology by CGH, although at high substitution rates parsimony analysis performed better when the threshold was high for

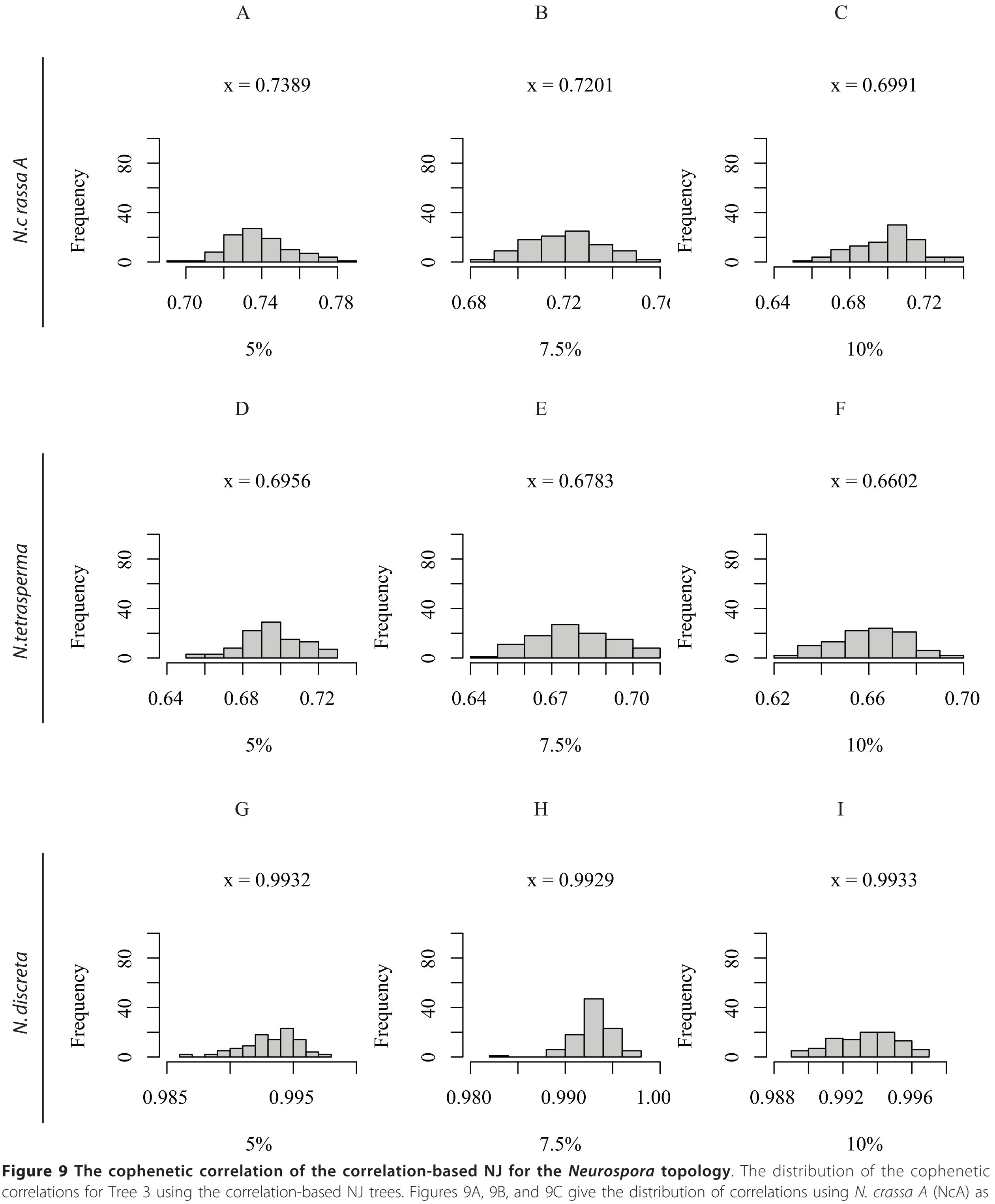


Figure 9 The cophenetic correlation of the correlation-based NJ for the *Neurospora* topology. The distribution of the cophenetic correlations for Tree 3 using the correlation-based NJ trees. Figures 9A, 9B, and 9C give the distribution of correlations using *N. crassa A* (NcA) as the reference for the average sequence divergence shown. Figures 9D, 9E, and 9F use *N. tetrasperma* (tet) as the reference taxa. Figures 9G, 9H, and 9I represent reference taxon *N. discreta* (disc). The average correlation for each set replicates is shown with each distribution.

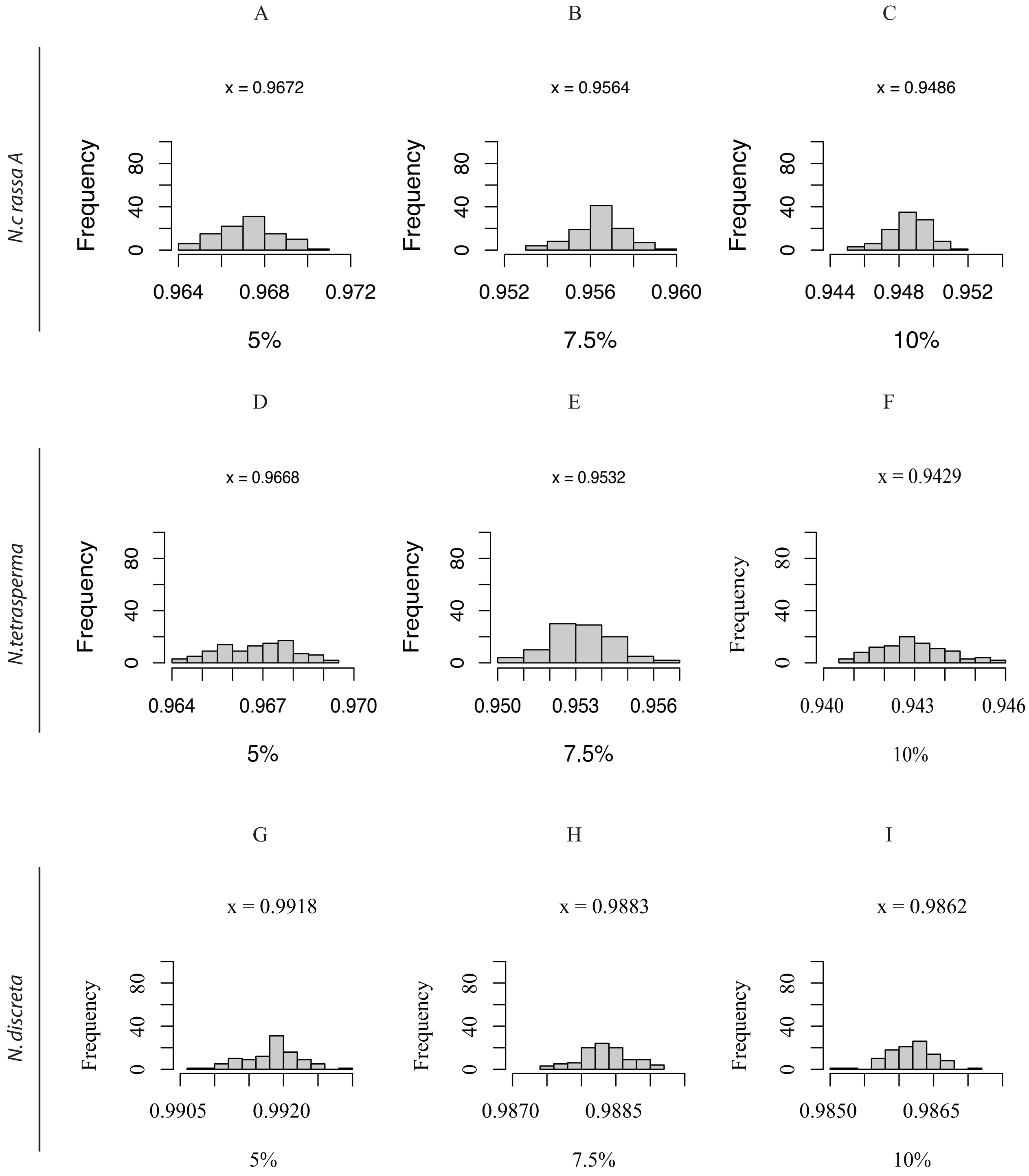


Figure 10 The cophenetic correlation of the Euclidean-based NJ for the *Neurospora* topology. The distribution of the cophenetic correlations for Tree 3 for 100 replicate Euclidean distance based NJ Trees. Histograms 10A, 10B, and 10C are *N. crassa A* (NcA) reference based. Histograms 10D, 10E, 10F are *N. tetrasperma* (tet) reference based. Histograms 10G, 10H, and 10I represent reference taxon *N. discreta* (disc). Note that the correlations are high, because of the correspondence in branch lengths even though the topologies aren't similar.

scoring a hybridization spot as absent. Increasing the length of the probe from 70 bases to 500 bases had no substantial effect on recovery of the input topology.

Of particular importance to a correct CGH phylogeny is the choice of taxa included in the analysis. Taxa with reasonable distance to each other, avoiding mixtures of close and more distantly related taxa, are more easily captured. It is also best to have a reference taxon in the most basal position of the tree, or second-best the most derived position. *A priori*, these two parameters are never known except in test cases, making CGH problematic for practical use in phylogenetics.

Comparison to other studies using CGH to infer phylogenetic relationships

Initial studies of CGH for phylogeny include research with *Salmonella typhimurium* and *Bordetella pertussis* microarrays. For the *Bordetella* study a parsimony CGH tree has at least one major difference when compared to the available MLEE tree [39], the latter being an unreliable method as it only determines phenotypes and is not very reproducible. NJ and Parsimony CGH trees for *Salmonella* are also at least one step away from several sequence trees used for comparison [60].

Other researchers utilizing *Bacillus, Streptococcus*, and *Ralstonia* microarrays report congruence between CGH-derived topologies constructed using hierarchical clustering, NJ, or Parsimony methods of bacterial clusters or species groups and DNA sequence trees based on several loci or MLEE [41,61,62]. All of these studies highlight the similarities of their trees to the previously published work but say little about the differences of their CGH-based trees to those derived via more traditional methods though the *Streptococcus* study found genetic relatedness only within clonal complexes, not between them. Though one or two branch rearrangements may not be substantial difference to the researcher, it demonstrates the limitations of this phylogenetic technique by failing to resolve a perfect tree.

It is possible that a CGH phylogeny from a multi-species array would better reflect evolutionary distances among taxa. While answering this question is outside the scope of this study at least one other, Wan et al [44], has addressed it. Unlike our study, which focused on a single species array, Wan et. al analyze the effects of the composition of a multi-species array and propose a bias-correction algorithm for uneven species content. Using both experimental and simulated data for a mixed *Enteroccocus* species array they find a CGH tree one step away from a 5 gene MLST tree of the same species. In their *in silico* study they see longer branch lengths from reference to other taxa in reconstructed trees, as we do, but stress that the clustering of groups should be unaffected in most cases.

In a more recent study aCGH data from bovine and human data as well as simulated data were examined by the developers of a wavelet based de-noising algorhythm aimed at quantifying structural variation at the genome level [38]. Though the authors restricted their study of evolutionary relatedness to clusters of intra-species data they expressed an interest in future phylogenetic analysis utilizing their algorithm as a starting point. The large scale structural variations simulated by the authors were quite different from the closely related relationships we chose to model in this work. This different approach to the same question did not address the factors we found most pertinent - the topology of underlying tree and choice of reference taxon.

Conclusions

Our results show that CGH cannot be counted on to reveal genomic sequence divergence reliably enough to recover a known phylogeny. Consequently, relying on this technique to recover an unknown phylogeny is problematic, particularly when there is no traditional phylogeny to compare to or when an existing phylogeny is based only a few genes or obsolete techniques like MLEE.

The results from our *in silico* study presented here are further substantiated by a thorough analysis of experimental data in our accompanying manuscript and by the various pro-CGH phylogeny papers enumerated above which detail minor, but significant, differences between their trees and previously published works.

Given the drawbacks of CGH for phylogeny, resulting in high uncertainty in the correct topology, it would be perhaps be more prudent to utilize the cross-taxa CGH data to identify a set of genes suitable for MLST, those still present but with a high to moderate amount of variation, then pursue a traditional phylogenetic analysis.

Additional material

Additional file 1: Additional Table S1. Tree to Tree distances for all Neighbor-Joining Results.

Additional file 2: Additional Table S2. Tree to Tree distances for all Parsimony Results.

Additional file 3: Additional Figure S1. Example topologies of Neurospora Parsimony simulation results in pdf format.

Additional file 4: Additional Figure S2 Correlation-Based CCC distributions. Balanced topology Cophenetic Correlations Coefficient Distributions for the correlation-based Neighbor-Joining Analysis in pdf format.

Additional file 5: Additional Figure S3 Euclidean-Based CCC distributions. Balanced topology Cophenetic Correlations Coefficient Distributions for the Euclidean-based Neighbor-Joining Analysis in pdf format.

Additional file 6: Additional Figure S4 Correlation-Based CCC distributions. Pectinate topology Cophenetic Correlations Coefficient Distributions for the correlation-based Neighbor-Joining Analysis in pdf format.

Additional file 7: Additional Figure S5 Euclidean-Based CCC distributions. Pectinate topology Cophenetic Correlations Coefficient Distributions for the Euclidean-based Neighbor-Joining Analysis in pdf format.

Acknowledgements

Special thanks to Michael B. Eisen for conceiving the study. Thanks to Charles Yong for his work scripting essential parts of the simulation pipeline. Thanks to Sandrine Dudoit and Jeffrey P. Townsend for statistical advice. Thanks to Tom Sharpton and Jason Staich for scripting advice. Special thanks to Elizabeth Turner and Tracy K. Powell for editorial assistance. Funding for LBG provided by the Ford Pre-Doctoral and Dissertation Year Fellowships. Funding for TK was provided by NIH Program Project Grant GM068087 to N. Louise Glass, PMB, UC Berkeley.

Author details

[1]Laboratoire de Recherche en Sciences Végétales, UMR CNRS-Université Paul Sabatier 5546, Chemin de Borde Rouge - Auzeville 31326, Castanet Tolosan, France. [2]Department of Plant Biology, Carnegie Institution for Science, Stanford, California 94305, USA. [3]USDA ARS, Plant Pathology Department, UC Davis, Davis, CA, 95616, USA. [4]Department of Plant and Microbial Biology, University of California Berkeley, Berkeley CA 94720, USA.

Authors' contributions

LBG participated in the simulation design and coordination and drafted the manuscript. LC, TK, and JWT participated in the design of the study. LC and LBG scripted code for the completion of the simulation. LBG completed the statistical analysis. All authors read and approved the final manuscript.

Received: 7 April 2010 Accepted: 21 September 2011
Published: 21 September 2011

References

1. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423(6937)**:241.
2. Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, Baquero F, Berche P, Bloecker H, Brandt P, Chakraborty T, *et al*: **Comparative Genomics of Listeria Species.** *Science %R 101126/science1063447* 2001, **294(5543)**:849-852.
3. Galagan JE, Calvo SE, Cuomo C, Ma L-J, Wortman JR, Batzoglou S, Lee S-I, Basturkmen M, Spevak CC, Clutterbuck J, *et al*: **Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae.** *Nature* 2005, **438(7071)**:1105-1115.
4. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, de Montigny J, Marck C, Neuveglise C, Talla E, *et al*: **Genome evolution in yeasts.** *Nature* 2004, **430(6995)**:35.
5. Hardiman G: **Microarray platforms - comparisons and contrasts.** *Pharmacogenomics* 2004, **5(5)**:487-502.
6. Hazen SP, Kay SA: **Gene arrays are not just for measuring gene expression.** *Trends in Plant Science* 2003, **8(9)**:413.
7. Fan J-B, Chee MS, Gunderson KL: **Highly parallel genomic assays.** *Nat Rev Genet* 2006, **7(8)**:632-644.
8. Dorrell N, Hinchliffe SJ, Wren BW: **Comparative phylogenomics of pathogenic bacteria by microarray analysis.** *Current Opinion In Microbiology* 2005, **8(5)**:620-626.
9. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays.** *Nat Genet* 1999, **23**:41-46.
10. Oostlander AE, Meijer GA, Ylstra B: **Microarray-based comparative genomic hybridization and its applications in human genetics.** *Clinical Genetics* 2004, **66**:488-495.
11. Pinkel D, Albertson DG: **Array comparative genomic hybridization and its applications in cancer.** *Nat Genet* 2005, **37**:S11-S17.
12. Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, Patil N, Wolff RK, Chee MS, Reid BJ, Lockhart DJ: **Genome-wide Detection of Allelic Imbalance Using Human SNPs and High-density DNA Arrays.** *Genome Res* 2000, **10(8)**:1126-1137.
13. Barrett MT: **Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA.** *Proc Natl Acad Sci USA* 2004, **101**:17765-17770.
14. Huyghe A, Francois P, Schrenzel J: **Characterization of microbial pathogens by DNA microarrays.** *Infection Genetics and Evolution* 2009, **9(5)**:987-995.
15. Takahito Watanabe YMSOHI: **A new approach to species determination for yeast strains: DNA microarray-based comparative genomic hybridization using a yeast DNA microarray with 6000 genes.** *Yeast* 2004, **21(4)**:351-365.
16. Murray AE, Lies D, Li G, Nealson K, Zhou J, Tiedje JM: **DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes.** *Proc Natl Acad Sci USA* 2001, **98**:9853-9858.
17. Porwollik S, Boyd EF, Choy C, Cheng P, Florea L, Proctor E, McClelland M: **Characterization of Salmonella enterica subspecies I genovars by use of microarrays.** *Journal Of Bacteriology* 2004, **186(17)**:5883-5898.
18. Watanabe T, Murata Y, Oka S, Iwahashi H: **A new approach to species determination for yeast strains: DNA microarray-based comparative genomic hybridization using a yeast DNA microarray with 6000 genes.** *Yeast* 2004, **21(4)**:351-365.
19. Fox GE, Wisotzkey JD, Jurtshuk P Jr: **How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity.** *Int J Syst Bacteriol* 1992, **42(1)**:166-170.
20. Keswani J, Orkand S, Premachandran U, Mandelco L, Franklin MJ, Whitman WB: **Phylogeny and taxonomy of mesophilic Methanococcus spp. and comparison of rRNA, DNA hybridization, and phenotypic methods.** *Int J Syst Bacteriol* 1996, **46(3)**:727-735.
21. Dettman JR, Jacobson DJ, Taylor JW: **A multilocus genealogical approach to phylogenetic species recognition in the model eukaryote Neurospora.** *Evolution* 2003, **57(12)**:2703-2720.
22. Hillis DM: **Inferring complex phytogenies.** *Nature* 1996, **383(6596)**:130-131.
23. Rokas A, Williams BL, King N, Carroll SB: **Genome-scale approaches to resolving incongruence in molecular phylogenies.** *Nature* 2003, **425(6960)**:798-804.
24. Dettman JR, Harbinski FM, Taylor JW: **Ascospore Morphology Is a Poor Predictor of the Phylogenetic Relationships of Neurospora and Gelasinospora.** *Fungal Genetics and Biology* 2001, **34(1)**:49.
25. Taylor JW, Jacobson DJ, Kroken S, Geiser DM, Hibbett DS, Fisher MC: **Phylogenetic Species Recognition and Species Concepts in Fungi.** *Fungal Genetics and Biology* 2000, **31(1)**:21-32.
26. Avise JC, Wollenberg K: **Phylogenetics and the origin of species.** *Proceedings of the National Academy of Sciences* 1997, **94(15)**:7748-7755.
27. Avise JC, Ball RM: **Principles of genealogical concordance in species concepts and biological taxonomy.** In *Oxford surveys in evolutionary biology.* Edited by: Antonovics DFaJ. Oxford, U. K.: Oxford Univ. Press; 1990:45-67.
28. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95(25)**:14863-14868.
29. Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.** *Statistical Applications in Genetics and Molecular Biology* 2004, **3(1)**:Article 3.
30. Grewal AaC A: **Tools for Analyzing Microarray Expression Data.** *Journal of Lab Automation* 2000, **5(5)**:62-64.
31. Rasmussen TB, Danielsen M, Valina O, Garrigues C, Johansen E, Pedersen MB: **Streptococcus thermophilus Core Genome: Comparative Genome Hybridization Study of 47 Strains.** *Appl Environ Microbiol* 2008, **74(15)**:4703-4710.
32. Israel DA, Salama N, Krishna U, Rieger UM, Atherton JC, Falkow S, Peek RM: **Helicobacter pylori genetic diversity within the gastric niche of a single human host.** *Proc Natl Acad Sci USA* 2001, **98**:14625-14630.
33. Chan K, Baker S, Kim CC, Detweiler CS, Dougan G, Falkow S: **Genomic Comparison of Salmonella enterica Serovars and Salmonella bongori by Use of an S. enterica Serovar Typhimurium DNA Microarray.** *J Bacteriol* 2003, **185(2)**:553-563.
34. Dorrell N, Mangan JA, Laing KG, Hinds J, Linton D, Al-Ghusein H, Barrell BG, Parkhill J, Stoker NG, Karlyshev AV: **Whole genome comparison of**

Campylobacter jejuni human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res* 2001, **11**:1706-1715.

35. Wolfgang MC, Kulasekara BR, Liang X, Boyd D, Wu K, Yang Q, Miyada CG, Lory S: **Conservation of genome content and virulence determinants among clinical and environmental isolates of Pseudomonas aeruginosa.** *PNAS %R 101073/pnas0832438100* 2003, **100(14)**:8484-8489.
36. Igboin CO, Griffen AL, Leys EJ: **Porphyromonas gingivalis Strain Diversity.** *Journal of Clinical Microbiology* 2009, **47(10)**:3073-3081.
37. Solheim M, Aakra A, Snipen L, Brede D, Nes I: **Comparative genomics of Enterococcus faecalis from healthy Norwegian infants.** *BMC Genomics* 2009, **10(1)**:194.
38. Mitra A, Liu G, Song J: **A Genome-Wide Analysis of Array-Based Comparative Genomic Hybridization (CGH) Data to Detect Intra-Species Variations and Evolutionary Relationships.** *PLoS ONE* 2009, **4(11)**:e7978.
39. Cummings CA, Brinig MM, Lepp PW, van de Pas S, Relman DA: **Bordetella Species Are Distinguished by Patterns of Substantial Gene Loss and Host Adaptation.** *J Bacteriol* 2004, **186(5)**:1484-1492.
40. Winzeler EA, Castillo-Davis CI, Oshiro G, Liang D, Richards DR, Zhou Y, Hartl DL: **Genetic Diversity in Yeast Assessed With Whole-Genome Oligonucleotide Arrays.** *Genetics* 2003, **163(1)**:79-89.
41. Read TD, Peterson SN, Tourasse N, Baillie LW, Paulsen IT, Nelson KE, Tettelin H, Fouts DE, Eisen JA, Gill SR, *et al*: **The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria.** *Nature* 2003, **423(6935)**:81.
42. Hinchliffe SJ, Isherwood KE, Stabler RA, Prentice MB, Rakin A, Nichols RA, Oyston PCF, Hinds J, Titball RW, Wren BW: **Application of DNA Microarrays to Study the Evolutionary Genomics of Yersinia pestis and Yersinia pseudotuberculosis.** *Genome Res* 2003, **13(9)**:2018-2029.
43. Edwards-Ingram LC, Gent ME, Hoyle DC, Hayes A, Stateva LI, Oliver SG: **Comparative Genomic Hybridization Provides New Insights Into the Molecular Taxonomy of the Saccharomyces Sensu Stricto Complex.** *Genome Res* 2004, **14(6)**:1043-1051.
44. Wan Y, Broschat SL, Call DR: **Validation of mixed-genome microarrays as a method for genetic discrimination.** *Applied and Environmental Microbiology* 2007, **73(5)**:1425-1432.
45. Carvalho B, Ouwerkerk E, Meijer GA, Ylstra B: **High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides.** *J Clin Pathol* 2004, **57(6)**:644-646.
46. Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, Meijer GA: **BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH).** *Nucleic Acids Research* 2006, **34(2)**:445-450.
47. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma L-J, Smirnov S, Purcell S, *et al*: **The genome sequence of the filamentous fungus Neurospora crassa.** *Nature* 2003, **422(6934)**:859.
48. Kasuga T, Townsend JP, Tian C, Gilbert LB, Mannhaupt G, Taylor JW, Glass NL: **Long-oligomer microarray profiling in Neurospora crassa reveals the transcriptional program underlying biochemical and physiological events of conidial germination.** *Nucl Acids Res %R 101093/nar/gki953* 2005, **33(20)**:6469-6485.
49. Mannhaupt G, Montrone C, Haase D, Mewes HW, Aign V, Hoheisel JD, Fartmann B, Nyakatura G, Kempken F, Maier J, *et al*: **What's in the genome of a filamentous fungus? Analysis of the Neurospora genome sequence.** *Nucl Acids Res %R 101093/nar/gkg293* 2003, **31(7)**:1944-1954.
50. Tian C, Kasuga T, Sachs MS, Glass NL: **Transcriptional Profiling of Cross Pathway Control in Neurospora crassa and Comparative Analysis of the Gcn4 and CPC1 Regulons.** *Eukaryotic Cell* 2007, **6(6)**:1018-1029.
51. Stoye J, Evers D, Meyer F: **Rose: generating sequence families.** *Bioinformatics* 1998, **14(2)**:157-163.
52. Kellis M, Birren BW, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae.** *Nature* 2004, **428(6983)**:617.
53. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author.** *Department of Genome Sciences, University of Washington, Seattle* 2005.
54. Swofford DL: **PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).** *Portable version 4.0b10 for Unix* Sinauer Associates, Sunderland, Massachusetts; 2004.
55. Penny D, Hendy MD: **The Use of Tree Comparison Metrics.** *Systematic Zoology* 1985, **34(1)**:75-82.
56. Goddard W, Kubicka E, Kubicki G, McMorris FR: **The agreement metric for labeled binary trees.** *Mathematical Biosciences* 1994, **123(2)**:215.
57. R_Development_Core_Team: R: **A language and environment for statistical computing.** Vienna, Austria.: R Foundation for Statistical Computing; 2006.
58. Paradis E, Strimmer K, Claude J, Jobb G, Opgen-Rhein R, Dutheil J, Noel Y, Bolker B: **ape: Analyses of Phylogenetics and Evolution.** 2006, R package version 1.8-2. edn.
59. Moore DS, McCabe GP: **Introduction to the Practice of Statistics.** New York: W.H. Freeman and Company;, Fourth 2003.
60. Porwollik S, Wong RM-Y, McClelland M: **Evolutionary genomics of Salmonella: Gene acquisitions revealed by microarray analysis.** *PNAS* 2002, **99(13)**:8956-8961.
61. Guidot A, Prior P, Schoenfeld J, Carrere S, Genin S, Boucher C: **Genomic structure and phylogeny of the plant pathogen Ralstonia solanacearum inferred from gene distribution analysis.** *Journal of Bacteriology* 2007, **189(2)**:377-387.
62. Dagerhamn J, Blomberg C, Browall S, Sjostrom K, Morfeldt E, Henriques-Normark B: **Determination of Accessory Gene Patterns Predicts the Same Relatedness among Strains of Streptococcus pneumoniae as Sequencing of Housekeeping Genes Does and Represents a Novel Approach in Molecular Epidemiology.** *J Clin Microbiol* 2008, **46(3)**:863-868.

doi:10.1186/1471-2164-12-456
Cite this article as: Gilbert *et al.*: Array Comparative Genomic Hybridizations: Assessing the ability to recapture evolutionary relationships using an *in silico* approach. *BMC Genomics* 2011 **12**:456.