Sequence Comparative Analysis Using Networks: Software for Evaluating De Novo Transcript Assembly from Next-Generation Sequencing

Ian Misner,[†,1] Cédric Bicep,[†,2] Philippe Lopez,[2] Sébastien Halary,[3] Eric Bapteste,[2] and Christopher E. Lane*,[1]

[1]Department of Biological Sciences, University of Rhode Island

[2]UMR CNRS 7138 Systématique, Adaptation, Evolution, Université Pierre et Marie Curie, Paris, France

[3]Département de Sciences Biologiques, Institut de recherche en biologie végétale, Université de Montréal, Montréal, Quebec, Canada

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: clane@mail.uri.edu.

Associate editor: Sudhir Kumar

Abstract

DNA sequencing technology is becoming more accessible to a variety of researchers as costs continue to decline. As researchers begin to sequence novel transcriptomes, most of these data sets lack a reference genome and will have to rely on de novo assemblers. Making comparisons across assemblies can be difficult: each program has its strengths and weaknesses, and no tool exists to comparatively evaluate these data sets. We developed software in R, called Sequence Comparative Analysis using Networks (SCAN), to perform statistical comparisons between distinct assemblies. SCAN uses a reference data set to identify the most accurate de novo assembly and the "good" transcripts in the user's data. We tested SCAN on three publicly available transcriptomes, each assembled using three assembly programs. Moreover, we sequenced the transcriptome of the oomycete *Achlya hypogyna* and compared de novo assemblies from Velvet, ABySS, and the CLC Genomics Workbench assembly algorithms. One thousand one hundred twenty-eight of the CLC transcripts were statistically similar to the reference, compared with 49 of the Velvet transcripts and 937 of the ABySS transcripts. SCAN's strength is providing statistical support for transcript assemblies in a biological context. However, SCAN is designed to compare distinct node sets in networks, therefore it can also easily be extended to perform statistical comparisons on any network graph regardless of what the nodes represent.

Key words: **transcriptome, de novo assembly, network, comparative genomics, oomycete, next-generation sequencing.**

Introduction

Advances in sequencing technologies have made genome-scale studies accessible to individual laboratories. Although most model systems rely on reference genomes for transcript assembly, nonmodel systems, or organisms without available reference genomes, must utilize de novo transcriptome assemblers (Zhao et al. 2011). The challenges of de novo transcriptome assembly are well documented (Normark et al. 1983; Cocquet et al. 2006; Martin et al. 2010; Grabherr et al. 2011; Martin and Wang 2011) and include varying transcript abundance, alternative splicing, and strand-specific expression of transcripts, which challenge the accuracy of assembly algorithms.

As transcriptome sequencing has become more common, algorithms used to assemble these data have become more numerous. Programs such as Velvet, ABySS, Oases, Trinity, and the CLC Genomics Workbench (CLC Bio Aarhus, Denmark) have been developed to specifically handle the difficulties associated with transcriptome data sets and have quickly become the "go to" programs for de novo transcript assembly (Zerbino and Birney 2008; Simpson et al. 2009; Grabherr et al. 2011; Schulz et al. 2012). Each has strengths, depending on the data and the needs of the user (Martin and Wang 2011).

The basic parameters that accompany the output of many assembly algorithms (e.g., transcript length distribution, median size [n50], and base quality) reflect the sequence composition and provide almost no indication as to whether the assembled transcripts represent plausible mRNA sequences that are sufficiently similar to an organism's genes. For this reason, time-consuming manual assessment via nucleotide or protein homology is required to determine the effectiveness of the assembly (Everett et al. 2011; Feldmeyer et al. 2011; Zheng et al. 2011). Additionally, if the study organism is distantly related to species for which comparative data are available, this approach may not provide conclusive results as to how well a transcriptome assembly recapitulates real mRNA. There does not currently exist a way to statistically compare the extent of transcript sequences highly similar to known biological sequences in different assemblies that result from iterations of a single algorithm or those derived from multiple assembly algorithms for the same data. With these limitations, we set out to assess transcriptome assembly in a statistically comparable way. We present the Sequence Comparative Analysis using Networks (SCAN) software, which utilizes distinct features of sequence similarity networks to make statistical comparisons among assemblies using one or several reference organisms.

Gene Similarity Networks

Gene similarity networks are graphs (fig. 1) illustrating sequence similarities in user-generated data sets (Holland et al. 2004; Huson and Bryant 2006; Bittner et al. 2010; Beauregard-Racine et al. 2011; Bhattacharya et al. 2013). Two nodes (sequences) are connected by an edge when there exists a relationship of similarity between the sequences (fig. 1*A*), as assessed by a Basic Local Alignment Search Tool (BLAST) search (Altschul et al. 1997). Importantly, two types of edges can be distinguished, depending on whether the similarity between two sequences occurs along more than 90% of each sequence (full homology) or over smaller portions (partial homology) (Alvarez-Ponce et al. 2013). On such a graph, nodes connected by edges fall into separate "connected components" (fig. 1*B*). Because they are mathematically based, measurable properties of connected components and indices of nodes in gene similarity networks (e.g., proportion of articulation points, proportion of nodes of degree one, and Jaccard Index [JI]; fig. 1*C*–*E*) can be used to compare the topological behaviors of transcripts and reference sequences in the network. Therefore, these networks can be used to compare a set of assembled transcripts to a data set of reference proteins.

SCAN identifies assembled transcripts that have no significant differences in statistical distribution of index values with respect to the reference data, according to five indices in the network: 1) proportion of reference and transcript data, 2) proportion of local and 3) global articulation points composed of reference and transcript nodes, 4) longest monochromatic chain, and 5) proportion of transcript/reference nodes with a degree of 1. More precisely, a global articulation point is a sequence whose removal in a connected component results in the disconnection of that component into smaller components. A monochromatic chain is a shortest path between two nodes that only comprises nodes with the same label (i.e., transcript or reference). SCAN calculates these statistics for each assembly condition. In addition, a sixth measure, the Jaccard Index (JI), applies to a pair of nodes, and measures the proportion of shared neighbors between these two nodes. In its present form, our gene similarity networks approach provides a quick method to compare large data sets using measureable and statistically informative connected component features. We tested SCAN on three public transcriptome data sets from the model organisms *Escherichia coli*, *Saccharomyces cerevisiae*, and *Plasmodium falciparum*, each assembled by three assembly programs (Blattner et al. 1997; Gardner et al. 2002; Otero et al. 2010). Moreover, we also investigated the discriminatory power of SCAN (i.e., its ability to detect sequences that behave like reference sequences in similarity networks) using four triplets of taxa. These triplets comprised two closely phylogenetically related fungal species and a more distant one, with comparable numbers of connected components in the network (supplementary materials, Supplementary Material online). Finally, we used SCAN to assess de novo transcriptome assemblies of the nonmodel oomycete *Achlya hypogyna* produced by ABySS, Velvet, and CLC Genomics Workbench (CLC) assembly algorithms. We used predicted proteins from the fully sequenced and annotated *Pythium ultimum* genome as the reference data set (Levesque et al. 2010) for SCAN, to assess the significant similarity of transcripts reconstructed from *A. hypogyna* mRNA data to known biological sequences.

It is not our goal to evaluate assembly algorithms. Instead, we provide analytical software for users to quickly identify the assemblies from their data that are more similar to known biological sequences. Such assemblies cluster with their homologs in a connected component. However, when assembly methods produce bad transcripts due to artifacts such as the production of chimeras or incomplete assembly, distinct patterns between reference genes and transcripts can appear (fig. 1*D*, *F*, *G*, and *H*) in sequence similarity networks. We show that all the assemblers produced both poor and well-assembled transcripts and that SCAN can efficiently compare multiple assemblies.

Results

Two versions of SCAN were developed: "SCAN" and "SCAN stringent" that can run either in single or multiprocessor nodes and are available from http://evol-net.fr (last accessed May 23, 2013). As documented later, the stringent version is preferred for biological applications, such as assembly assessment and phylogenomics.

Testing SCAN

SCAN uses the Kolmogorov–Smirnov (KS) test (*P*-value threshold of 0.05) to compare the distributions of index values for transcripts and reference nodes (except for the JI, see Materials and Methods). The null hypothesis is that test and reference transcripts are from the same population of biological sequences and should have comparable topological properties in sequence similarity network. Rejecting the null hypothesis indicates that assembly methods have potentially failed because properties of those test and reference transcripts have different distributions and are unlikely to be biologically similar. We have recently published a first test of this concept and most of these indices (Bhattacharya et al. 2013), but here we also verified that SCAN behaved according to our expectation on a fungal data set (see supplemental materials, Supplementary Material online).

To test SCAN's efficiency in terms of detection of "good" transcripts, we used three de novo assemblers (ABySS, CLC, and Oases) to assemble the transcriptome reads from three fully sequenced genomes (one bacteria, *E. coli*, and two eukaryotes: *Sac. cerevisiae* and *Pla. falciparum*). Good contigs were arbitrarily identified within each of these assemblies for each organism as the transcripts

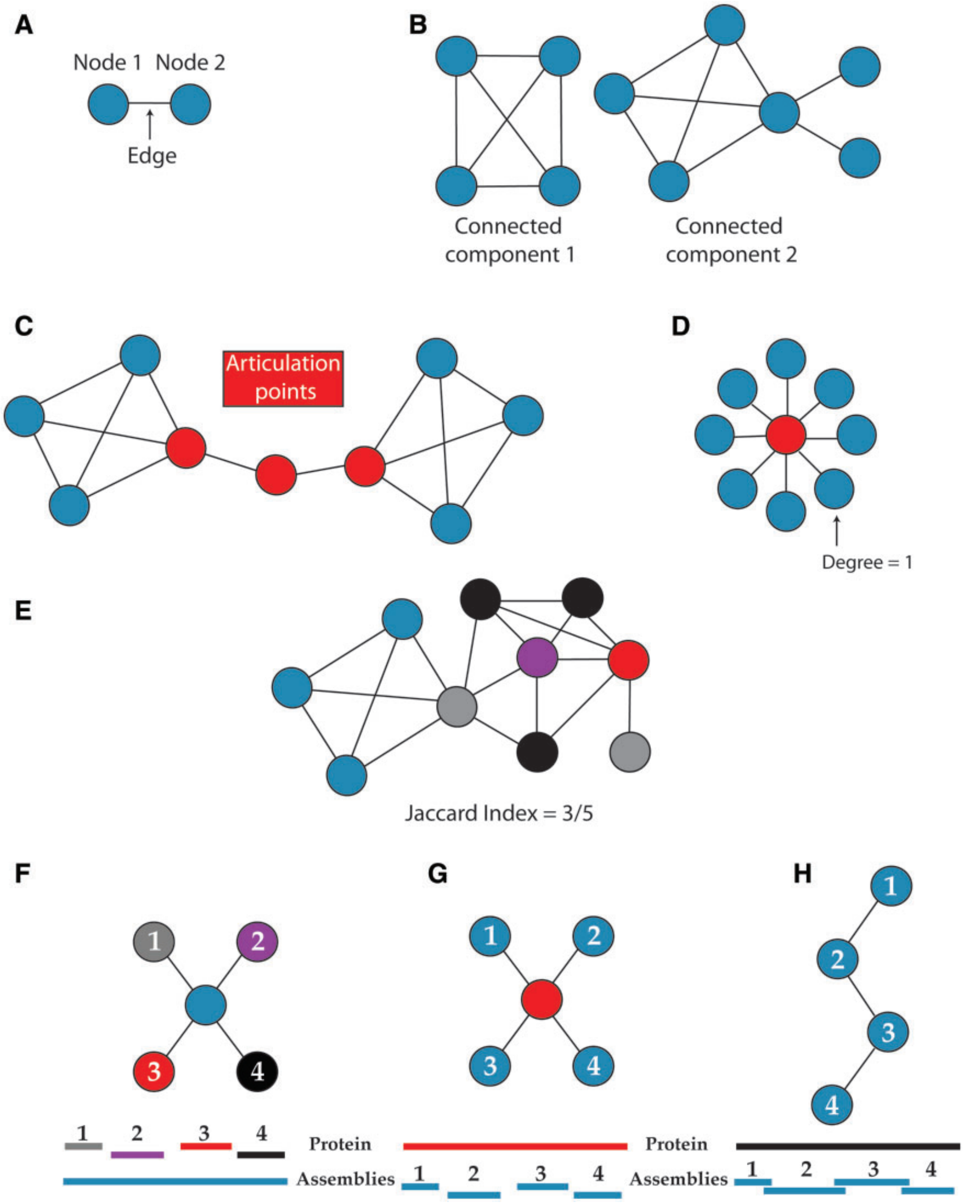

Fig. 1. Basic topologies found in sequence similarity networks. (*A*) Nodes represent sequences. Two nodes are connected by an edge when they show a similarity greater than a user-defined threshold. (*B*) Network graphs are made up of separate subnetworks, called connected components. (*C*) Any node that disconnects a connected component when removed is called a global articulation point (red). (*D*) The degree of a node is the number of its direct neighbor(s). Incomplete assembly may result in high proportion of transcripts with a degree of 1 (blue) compared with reference (red). (*E*) The JI is used to compare the neighborhood of two nodes. The red and purple nodes have a total of five neighbors (black/gray nodes) but only share the black nodes, so the JI for this pair is 3/5. If two nodes have strictly identical neighbors, the JI equals 1; if they do not share any neighbors, the JI equals 0. (*F*) When assembly methods produce chimeras (blue node) of unrelated proteins that connect the unrelated reference proteins (red, purple, black, and gray nodes), it produces a star-like pattern. (*G*) When assembly methods produce partial assemblies (blue nodes) that do not overlap with one another and are too short to connect to most of the full reference sequence (red node), a star-like pattern is also produced. (*H*) When assembly methods produce partial assemblies (blue nodes) that overlap with one another, it produces a chain-like pattern.

that match with more than 97% ID with their corresponding gene in the genome (supplementary table SI-2, Supplementary Material online). The number of transcripts that met this criterion varied between 14.90% and 43.90%. Thus, for each of these data sets, the number of good contigs was relatively low. This may be explained by a low coverage of the genome, by a poor choice of assembler criteria, or by a limited performance of these assemblers on these particular transcriptomes.

Nonetheless, we expected that SCAN should be able to distinguish good from bad transcripts within such assembled data sets, thereby proposing a "cleaned" subset of transcripts closely resembling their genes of origin. This cleaning step requires that SCAN solves a double challenge: rejecting as many of the poor transcripts as possible without losing too many of the good ones. To test how SCAN performed on these test data sets, we used two types of sequence similarity networks, including either 1) partial homology edges plus full homology edges or 2) full homology edges only (see Materials and Methods). The first type of network is very inclusive and can encompass even short assemblies (e.g., assemblies whose sequence of origin could not be entirely recovered). The second type of network is more stringent and should already eliminate short assemblies from the set of contigs in which good assemblies can be detected. Moreover, we tested two versions of SCAN: the default version and "SCAN stringent" on these networks. By default, SCAN operates with the set of indices described above to identify contigs with indistinguishable topological properties for most, or all these indices, with respect to their closest reference sequence in the network. SCAN stringent requires that the sequence of the contig is directly connected to (at least one of) its homolog(s) from the reference data set in the network.

This protocol allowed us to compare several conditions (table 1 and supplementary table SI-3, Supplementary Material online) to determine whether, and when, SCAN was most successful identifying genuinely good contigs, with the lowest rate of false-positive detection, and/or the lowest rate of miss. We observed that, irrespective of the transcriptome, a higher absolute number of good transcripts are recovered in networks with both partial and full homology edges. This is likely because the full homology condition eliminated many short, yet good, transcripts. Although using a network with both partial and full homology edges enhances the risk of false-positive detection, SCAN stringent massively reduces this risk (supplementary table SI-3, Supplementary Material online). For instance, the false-positive rate was 2.45% for ABySS, 3.84% for CLC, and 3.19% for Oases, in *Pla. falciparum* assemblies. False-positive rates were even lower in *E. coli* and *Sac. cerevisiae* (supplementary table SI-3, Supplementary Material online). Therefore, SCAN stringent is generally preferred to the default version for assembly evaluation and phylogenomic applications. However, one can also generally clean the assemblies using the default version of SCAN by raising the number of tests of indices that a sequence must pass to be considered as a good transcript. Although raising the number of tests of indices a transcript is required to pass (i.e., for which no statistical difference from a reference sequence is detected) tends to reduce the rate of false-positive detection, it also increases the rate of misses for both the stringent and the default version of SCAN (supplementary table SI-3, Supplementary Material online). Overall, it is clear that the stringent version of SCAN is not more conservative and performs better than the default version of SCAN to detect transcripts that are highly similar to reference sequences, irrespective of the type of networks (stringent or inclusive) on which it is used.

Application to Nonmodel Transcriptomes

SCAN stringent was used to assess transcriptome assemblies from the oomycete, *A. hypogyna*. The selected transcripts from each assembly returned significant values for five indices at different BLAST minimal % identity (ID) thresholds (ABySS = 50%, Velvet = 90%, CLC = 50%). SCAN indicated the CLC assembly resulted in the most transcripts (1,128) with statistically supported similarity to reference data, compared with the results of Velvet (49) and ABySS (937). Although the number of transcripts similar to the reference is low compared with the total number of transcripts in each assembly (CLC = 10.90%, Velvet = 0.12%, ABySS = 3.90%), the comparison to the number of analyzed transcripts (i.e., transcripts that made it into the network) is a more accurate reflection of SCAN's performance (CLC = 50.74%, Velvet = 36.01%, ABySS = 42.57%; table 2).

Reference and Network Selection

SCAN's calculations are based on distributions of transcript and reference nodes in a gene similarity network. As we have seen earlier, sequence similarity networks with different levels of stringency can be used, and SCAN will return larger absolute numbers of good contigs for more inclusive networks. Thus, unless one is exclusively interested in full sized good transcripts, the use of sequence similarity networks featuring both partial and full homology edges, coupled to the use of "SCAN stringent," will result in users identifying the highest number of good transcripts with limited false positives. Similarly, different references can be used to assess transcripts similarity to known sequences. The choice of a proper reference is not always trivial, therefore SCAN allows for the sequential use of multiple references. For each of these references, SCAN identifies transcripts that are highly similar to the reference genes. Therefore, different references can be jointly used to identify distinct sets of good transcripts (and collectively offer an even more complete detection of such transcripts). This "multi-reference" option is especially useful when a gene family has undergone an unusual evolutionary rate (or high rates of gene duplication or loss) in one reference but not in another. Furthermore, the use of multiple references may even allow one to choose with limited a priori information what reference is the best one overall for a transcriptome data set (e.g., the one with the highest number of genes similar to the transcripts).

To study the oomycetes transcripts, we compared them to five potential reference species before selecting *Pyt. ultimum* (see Materials and Methods). Regardless of the reference used, the CLC transcriptome assembly was the only assembly that consistently returned large numbers of quality transcripts (supplementary table SI-4, Supplementary Material online).

Proportion of Transcripts and References in Connected Components

SCAN calculations show that the CLC assembly had the most transcripts that resembled proteome data (good transcripts) (table 3; 2,132 transcripts, from the 1,716 connected components that pass the proportion test, P value > 0.05) for that index. This number was larger than ABySS (2,041) and Velvet (91) networks at the selected BLAST ID thresholds (table 3). Analyses of the ABySS and Velvet networks rejected the null hypothesis for the proportion index (table 3), indicating significant difference between test and reference data.

Table 1. Test of SCAN Efficiency for Different Transcriptomes, Assembly Programs, and a Majority of Network Indices.

Assembler	Number of Good Transcripts	Selected Network SCAN Stringent Mode	Selected Network SCAN Default Mode	Number of Good Transcripts SCAN Stringent Mode	Number of Good Transcripts SCAN Default Mode	Percentage Missed Good Transcripts (No. Transcripts) SCAN Stringent Mode	Percentage Missed Good Transcripts (No. Transcripts) SCAN Default Mode	Percentage False-Positive (No. Transcripts) SCAN Stringent Mode	Percentage False Positive (No. Transcripts) SCAN Default Mode
Plamodium falciparum, partial + full homology edges. Test passed for >3 indices									
ABySS	4,779	(>70% ID)	(>70% ID)	2,775	4,626	43.36 (2,072)	43.36 (2,072)	2.45 (68)	41.48 (1,919)
CLC	3,317	(>90% ID)	(>90% ID)	2,862	4,879	17.03 (565)	17.03 (565)	3.84 (110)	43.59 (2,127)
Oases	4,190	(>60% ID)	(>60% ID)	3,072	5,492	29.02 (1,216)	29.02 (1,216)	3.19 (98)	45.85 (2,518)
P. falciparum, full homology edges. Test passed for >3 indices									
ABySS	4,779	(>20% ID)	(>20% ID)	25	30	99.52 (4,756)	99.41 (4,751)	8.00 (2)	6.67 (2)
CLC	3,317	(>20% ID)	(>20% ID)	28	31	99.16 (3,289)	99.10 (3,287)	0.00 (0)	3.23 (1)
Oases	4,190	(>20% ID)	(>20% ID)	38	42	99.14 (4,154)	99.07 (4,151)	5.26 (2)	7.14 (3)
Saccharomyces cerevisiae, partial + full homology edges. Test passed for >3 indices									
ABySS	432	(>97% ID)	(>90% ID)	323	1,117	25.23 (109)	25.69 (111)	0.00 (0)	71.26 (796)
CLC	533	(>80% ID)	(>80% ID)	535	1,326	0.56 (3)	0.00 (0)	0.93 (5)	59.80 (793)
Oases	335	(>80% ID)	(>90% ID)	239	775	29.85 (100)	29.85 (100)	1.67 (4)	69.68 (540)
S. cerevisiae, full homology edges. Test passed for >3 indices									
ABySS	432	(>20% ID)	(>20% ID)	12	14	97.22 (420)	96.76 (418)	0.00 (0)	0.00 (0)
CLC	533	(>20% ID)	(>20% ID)	14	15	97.37 (519)	97.19 (518)	0.00 (0)	0.00 (0)
Oases	335	(>20% ID)	(>20% ID)	10	15	97.01 (325)	95.52 (320)	0.00 (0)	0.00 (0)
Escherichia coli, partial + full homology edges. Test passed for >3 indices									
ABySS	1,781	(>20% ID)	(>20% ID)	1,590	2,471	11.85 (211)	11.85 (211)	1.26 (20)	36.46 (901)
CLC	1,298	(>50% ID)	(>50% ID)	1,302	2,403	0.00 (0)	0.00 (0)	0.31 (4)	45.98 (1,105)
Oases	1,975	(>90% ID)	(>90% ID)	1,842	2,108	7.14 (141)	7.14 (141)	0.43 (8)	13.00 (274)
E. coli, full homology edges. Test passed for >3 indices									
ABySS	1,781	(>20% ID)	(>20% ID)	77	98	95.68 (1,704)	94.50 (1,683)	0.00 (0)	0.00 (0)
CLC	1,298	(>20% ID)	(>20% ID)	41	58	96.84 (1,257)	95.53 (1,240)	0.00 (0)	0.00 (0)
Oases	1,975	(>20% ID)	(>20% ID)	72	113	96.51 (1,906)	94.38 (1,864)	4.17 (3)	1.77 (2)

NOTE.—Three transcriptomes were assembled by distinct assemblers (column 1), producing different numbers of good transcripts (column 2). We used the stringent and default versions of SCAN to select the sequence similarity network (columns 3 and 4, respectively) with the highest absolute number of transcripts highly similar to the reference. For each version of SCAN, we report the number of hypothesized "good" transcripts (columns 5 and 6), the percentage of missed good transcripts (columns 6 and 7), and the percentage of false positives (columns 8 and 9).

Table 2. Assembly Results and Networks Features Analyzed by SCAN Stringent.

Assembler	ABySS	Velvet	CLC
Number of transcripts	23,996	41,420	10,349
n50 (bp)	531	673	900
k-mer's	55–64	29, 39, 49, 59, 69	NA
Number of transcripts/CC at 50%	2,201/1,566[a]	3,090/1,977	2,223/1,726[a]
Number of transcripts/CC at 60%	1,540/1,167	2,034/1,407	1,527/1,265
Number of transcripts/CC at 70%	844/674	1,101/782	812/692
Number of transcripts/CC at 80%	370/296	508/352	352/300
Number of transcripts/CC at 90%	112/86	136/92[a]	80/69
Number of transcripts/CC at 95%	39/31	38/26	30/24
Number of SCAN selected transcripts	937	49	1,128

Note.—CC, connected components.

[a]Selected as best assembly by SCAN.

Table 3. Number of Connected Components and Transcripts That Pass the Test for Various Assemblers and Indices.

Assembly Method	Centrality	Network Similarity	Number of CC	Quality Transcripts
AbySS	Proportion	50	1,552	2,041
AbySS	Articulation local	50	1,563	2,113
AbySS	Articulation global	50	1,565	2,151
AbySS	Degree one	50	1,565	2,151
AbySS	Monoch. chain	50	1,566	2,201
AbySS	Jaccard	50	798	983
Velvet	Proportion	90	92	136
Velvet	Articulation local	90	92	136
Velvet	Articulation global	90	92	136
Velvet	Degree one	90	92	136
Velvet	Monoch. chain	90	92	136
Velvet	Jaccard	90	36	52
CLC	Proportion	50	1,716	2,132
CLC	Articulation local	50	1,724	2,176
CLC	Articulation global	50	1,725	2,194
CLC	Degree one	50	1,726	2,223
CLC	Monoch. chain	50	1,726	2,223
CLC	Jaccard	50	974	1,167

Proportion of Transcripts and References Articulation Points in Connected Components

Articulation points are nodes whose removal disconnects a component, either locally (direct neighbors of the articulation point get disconnected) or globally (the removal of such nodes splits the initial component into several connected components). Chimeras could generate an excess of articulation points, for example, when two sequences that correspond to portions of different mRNAs have been mistakenly assembled into one single transcript (fig. 1*F*). Global articulation point produced 1,563 good transcripts for ABySS and 1,725 good transcripts for CLC and 92 for Velvet (table 3). SCAN indicated that CLC had the highest number of good transcripts for the local articulation point test (2,176) of all the networks examined (table 3).

Longest Monochromatic Chain

When assembly methods connect unrelated mRNA sequences, or fail to connect sequences from the same mRNA into a single transcript, an excess of chains of nodes corresponding to incorrect assemblies can form (fig. 2*C*). Length of these chains was estimated and compared with that of chains of sequences from the reference data in each connected component. In the SCAN selected networks, we were unable to identify any transcript assemblies that represented monochromatic chains as all the analyzed transcripts passed this index (table 3). The CLC assembly had the highest number of quality transcripts for that index with 2,223 transcripts.

Proportion Transcript and Reference of Degree One

Nodes with a degree of 1 likely correspond to poor assemblies, because failure to assemble short reads from the same gene or closely related paralogs will result in star-like connected components, with the incompletely assembled transcripts loosely connected (i.e., by a single edge) to the rest of the graph (fig. 1*G*). There were 50 transcripts with a degree of 1 in the ABySS assembly selected by SCAN (table 3). Velvet and CLC did not have any degree one transcripts in the selected assemblies as all transcripts passed this index (table 3).

Jaccard Index

If transcripts are successfully assembled, and if genes in the transcript data set have evolved in a similar fashion as the reference data (e.g., if the novel data did not undergo unusual amount of gene family expansion, elevated rates of evolution, or displayed other unusual genome features), we expect that transcripts would produce similar sequences to the reference. In this case, successfully assembled transcripts would be connected to the reference and have the same neighboring sequences in connected components (fig. 1*E*) of sequence similarity networks. The JI quantifies to what extent two nodes share the same neighbors by computing the ratio of common neighbors over the total number of neighbors. SCAN measured the JI for each transcript/reference pair in each input network.

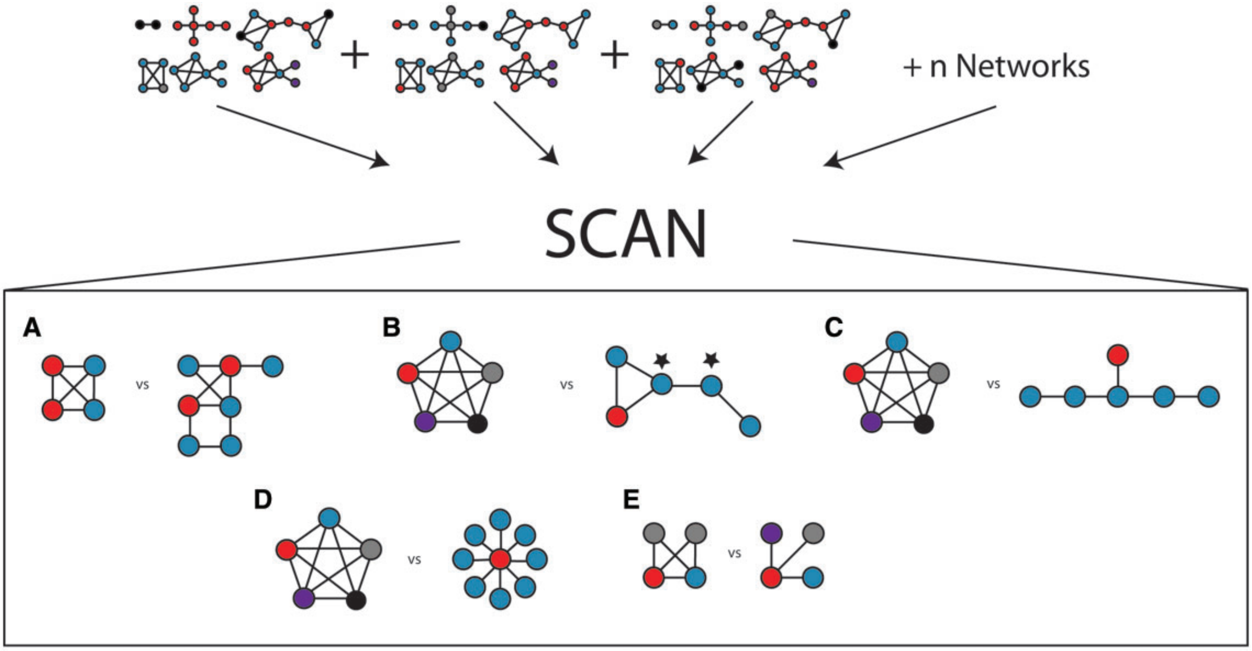

Fig. 2. SCAN flow chart diagramming the inputs, index tests, and how the appropriate test networks are identified. SCAN uses multiple network files as inputs where the user defines the two node types to be compared across, and within, the input networks. To identify similar network topologies, SCAN uses five indices to statistically quantify those topologies and the JI. (*A*) The proportion of test nodes (blue) and reference nodes (red) in each connected component is used as the first measure of similarity. Connected components with equal numbers of test and reference nodes (left) are indicative of similarity, whereas highly disproportionate ratios (right) indicate low similarity. (*B*) If test and reference nodes are similar, they will occur as articulation points (stars) in a near equal fashion (left). If they are different, this will not be the case (right). Articulation points can indicate improper assembly or chimerical sequence in gene networks. (*C*) Failed assemblies can produce chain-like patterns for test sequences as shown in right. Successful assemblies should produce no such chains, and if they do, test and reference chains should be of similar length (left). (*D*) Test nodes with a degree of 1 should occur at a similar rate as the reference (left). In case of problematic assembly, a higher proportion of test nodes with a degree of 1 can appear (right). (*E*) Using the JI, SCAN can identify when test and reference node pairs have similar neighbors (left) or not (right). Nodes from similar populations should share a common neighborhood.

The CLC assembly had the highest number of quality transcripts (1,167) sharing a similar neighborhood with the reference when compared with the assemblies from ABySS and Velvet (table 3).

Discussion

Correctly assembled transcriptomes yield a large number of transcripts that have comparable distributions of index values to sequences of related taxa in gene similarity networks and cluster with their homologs in a connected component. However, when assembly methods fail to recapitulate the actual transcriptome due to artifacts, such as the production of chimeras or to incomplete assembly, distinct patterns between reference genes and transcripts can appear (fig. 1*D*, *F*, *G*, and *H*).

In cases of incomplete assembly, we expect two extreme kinds of different behaviors from transcript sequences, with respect to reference proteins. Partial assemblies can loosely connect components, either producing chains of incorrect partial assemblies (fig. 1*H*) or a “cloud” of partial assemblies around the connected component. This occurs when sequences of partial assemblies do not overlap with one another and are too short to connect to most of the sequences in the component (fig. 1*G*). The result is connected components showing a high proportion of nodes whose removal increases the number of connected components, either at a local or global scale.

The second behavior results in sequences showing partial similarity with a reference node to which they are homologous. The resulting topology is a star-like pattern (fig. 1*D*) in which several partial transcripts (disconnected from one another) are directly connected to the same reference sequence. This situation would result in a connected component showing loosely connected nodes with an excess of nodes from the assembly having a degree of 1.

Indices Used in the Analysis

Using SCAN we were able to evaluate each connected component and node in the networks using six network indices. Our goal was to identify the largest pool of quality transcripts within the assemblies and networks examined. SCAN produces results on both of these levels. The initial output identifies which assembly has the highest number of quality transcripts and lists those transcripts. On a larger scale, SCAN can be used at the network level to identify networks that are significantly similar to the reference across all the connected components in that network. Individually, these indices have specific strengths in their ability to identify similarity, but it is the collective analysis of all six indices that gives SCAN its analytical power. “SCAN stringent” gets additional power from the direct connection between transcript and reference sequences in the network.

In each network of interest, SCAN first determines for each index, whether transcripts and reference have different topological properties in each connected component. This is calculated using a proportion test for the proportion of reference and transcript data, the proportion of local and global articulation points composed of reference and transcript nodes, the proportion of transcript/reference nodes with a degree of 1, and a KS test for the distribution of longest monochromatic chains of transcripts and reference sequences in all components. Connected components for which no difference is detected are assumed to contain good transcripts. Thus, SCAN can quantify the total number of good transcripts in a network and identify the best network as the one with the highest number of good contigs. However, in the rare event of a tie, SCAN uses the values calculated for each index in a different way. KS tests are used to compare the distributions of values for each index between reference and transcript nodes. SCAN considered the best network to be the one for which a greater number of KS tests are nonsignificant (e.g., showing no differences in the general distributions of reference and transcript sequences for more indices).

Quality Transcript Identification

For every transcript assembly examined, SCAN produces a list of good transcripts (table 2). The transcripts selected for this final list are those found at the intersection of connected components that could not reject the null hypothesis under the proportion test for an index and transcripts that passed five indices. This list of transcripts is highly conservative yet accurate, as multiple biological factors (e.g., gene family function, gene family expansion/contraction, lineage-specific changes, and genome duplication) will affect the topological properties of nodes in the network.

The ability to quickly indicate, with statistical support, transcripts that are biologically similar to the proteome of a related organism and useful for phylogenetic analyses in a comparative genomics framework is a valuable feature of SCAN. It is because of the family-level phylogenetic differences between *A. hypogyna* and *Pyt. ultimum* that SCAN only identifies 1,128 transcripts as statistically similar between the two organisms (P value ≥ 0.05 for > 5 tested indices), rather than an inability of CLC to assemble transcripts or SCAN to evaluate similarities between sequences. Because of the conservative nature of SCAN, the transcripts that pass all five tests should represent a reliable pool for phylogenomic applications.

An additional use for SCAN is parameter optimization using a single assembly program. Choosing parameters during an assembly is often as, or more important, than choosing an assembler program. When used to evaluate multiple networks produced under different assembly conditions, SCAN has the ability to evaluate program parameters and choose those producing contigs that best represent the reference data (data not shown).

We have provided novel software, SCAN, which can compare de novo assembled transcripts and reference sequences in similarity gene networks. SCAN's strength is providing statistical support for transcript assemblies in a biological context. This procedure requires limited computational infrastructure while providing robust analyses of thousands of genes in a short amount of time. SCAN's utility, however, could easily go far beyond evaluating transcriptome assemblies. As SCAN is designed to compare the topological properties of two node sets in networks, it can make statistical comparisons on any network graph regardless of what the nodes represent. To this end, future developments of the software will include a richer diversity of indices to broaden the comparative power of network-based analyses of large data sets of sequences.

Materials and Methods

Construction of Test Transcriptome Data Sets

Sequence data files for the evaluation tests of SCAN were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive for each organism used. Read sets for *E. coli* were ERR019652 and ERR019653, for *Sac. cerevisiae* SRR354092, SRR354093, and SRR354093, and for *Pla. falciparum* ERR185973 and ERR185974. All reads were trimmed on the CLC Genomics Workbench (v5.02) using default settings and duplicate reads were removed. The trimmed reads were used for all subsequent assemblies (table 4).

CLC assembly was completed with default settings for all data sets; parameters included a minimum contig length of 100 bp, 0.5 length fraction, and 0.8 similarity fraction with automatic word and bubble size. ABySS assemblies were run with default settings for each species with a coverage of five and erode of four, k-mer values for *E. coli* and *Sac. cerevisiae* were 19–31 and 21–31 for *Pla. falciparum* with a step interval of two. Transcripts from each species for each k-mer were combined into a single pool. Redundant transcripts and transcripts less than 100 bp were removed. Oases assemblies were completed using the python script included in the Oases package with a minimum transcript length of 100 bp and a coverage cutoff of 5. K-mer values for *E. coli* ranged between 19 and 31, 25–33 for *Sac. cerevisiae*, and 15–29 for *Pla. falciparum*.

RNA Extraction and Transcript Assembly

Cultures of *A. hypogyna* (ATCC 48635) were acclimatized at 25 °C for 2 weeks at 12:12 L:D before 16 Luria Broth

Table 4. Assembly Results for Each of the Test Species Used to Evaluate SCAN.

Assembler	Parameter	*Escherichia coli*	*Saccharomyces cerevisiae*	*Plasmodium falciparum*
CLC	n50 (bp)	248	196	195
	Number of transcripts	5,603	3,305	22,234
	Word size	21	20	23
	Bubble size	50	50	50
ABySS	n50 (bp)	349	208	276
	Number of transcripts	5,614	2,660	16,574
Oases	n50 (bp)	508	247	211
	Number of transcripts	4,495	1,678	21,248

50% concentration and 16 diH_2O 150 ml subcultures were created, each with three autoclaved hemp seeds. Cultures were then incubated at 4 °C, 15 °C, 25 °C, and 35 °C and harvested at 0.5, 1, 3, and 6 h in both light and dark conditions. Fresh material was ground under liquid nitrogen immediately after harvest, and ground material was placed directly in extraction buffer. RNA was extracted from *A. hypogyna* using the Qiagen RNeasy kit (Qiagen, CA) according to the manufacturer's protocol. RNA from all experiments was pooled, and quality was assessed using a Nanodrop 8000 (Thermo Scientific, CA).

Library construction and sequencing via the Illumina GAII were performed by Genome Quebec using one lane of single 108 bp reads. The data were trimmed on CLC to remove reads shorter than 70 bp and those reads whose cumulative bases with quality scores below 0.05. Default parameters for transcriptome assembly were used for CLC. The similarity threshold was set to 0.9, the insertion cost was set to 3, and the automatically generated k-mer was 26. ABySS (Simpson et al. 2009) was run with default settings using a set of k-mer values from 55 to 64 and then the assembled transcript sets were merged into a single pool, and redundant transcripts were removed. Velvet (Zerbino and Birney 2008; Zerbino et al. 2009; Zerbino 2010) transcript assemblies were done with default settings using a set of k-mer values including 29, 39, 49, 59, and 69, then transcript sets were merged as above.

Protein Translation

To identify the proper frame for each assembled transcript, DNA sequences from each assembly pool were compared with a local database containing the protein transcripts from *Drosophila melanogaster* (Adams et al. 2000), *Ectocarpus siliculosus* (Cock et al. 2010), *Fragilariopsis cylindrus* (Joint Genome Institute, USA), *Leishmania major* (Ivens et al. 2005), *Monosiga brevicollis* (King et al. 2008), *Mycosphaerella fijiensis* (Joint Genome Institute, USA), *Ostreococcus tauri* (Palenik et al. 2007), *Phaeodactylum tricornutum* (Bowler et al. 2008), *Phytophthora infestans* (Haas et al. 2009), *Phy. ramorum* (Tyler et al. 2006), *Phy. sojae* (Tyler et al. 2006), *Pla. falciparum* (Gardner et al. 2002), *Pyt. ultimum* (Levesque et al. 2010) *Saprolegnia parasitica* (Broad Institute, USA), and *Volvox carteri* (Prochnik et al. 2010) using BLASTX in conjunction with the OrfPredictor server (*e* value = 1e-5) (Min et al. 2005). All CLC transcripts had an identified protein translation, 17 ABySS and 84 Velvet transcripts did not show homology to sequences in the database and were excluded in subsequent analyses.

Reference Selection

The peronosporalean oomycete *Pyt. ultimum* was chosen by SCAN as the reference in the network for our *A. hypogyna* data. The genome sequence of the more closely related saprolegnian oomycete *Sap. parasitica* is available; however, this species has a highly divergent proteome as a result of its evolved specialization as a fish pathogen. *Pythium ultimum* lacks the highly duplicated genomic features of *Phytophthora* spp. and shares a necrotrophic lifestyle with *A. hypogyna* making it a better reference for our similarity network analysis approach (Tyler et al. 2006; Haas et al. 2009; Levesque et al. 2010). Despite *A. hypogyna* having a more distant evolutionary relationship to *Pyt. ultimum* than *Sap. parasitica*, *Pyt. ultimum* was preferred by SCAN to *Sap. parasitica* as a better reference.

Network Construction: Transcriptome Networks

We tested transcriptomes from three organisms, using the following reference data sets for each, downloaded from the NCBI. *Plasmodium falciparum* assemblies were analyzed in similarity networks with sequences from *P. cynomolgi*, *P. falciparum*, *P. knowlesi*, *P. vivax*, *Theileria parva*, and *T. annulata*. Assemblies from *E. coli str. K-12 substr. MG1655* were analyzed in similarity networks with sequences of *E. coli str. K-12 substr. MG1655*, *E. coli O157:H7 str. Sakai*, *E. coli SE11*, *E. coli O26:H11 str. 11368*, *Salmonella enterica*, and *Shigella dysenteriae*. Sequences in each of these data sets were BLASTed all against all (using the relevant BLASTP, BLASTN, BLASTX, and TBLASTN programs) using 1e-20 *e*-value cutoff and a maximum of 5,000 hits per query. The corresponding similarity network was then built by creating an edge between two sequences if the corresponding BLAST *e*-value was lower than 1e-20 and the identity percentage was greater than 20%. We used BLAST 2.2.21 to perform all possible pairwise comparisons. BLASTX (default parameters) was used to compare a transcript with a reference protein sequence, BLASTN (default parameters) was used to compare transcript sequences, TBLASTN (default parameters) was used to compare the reference protein sequences to the transcripts sequences, and BLASTP (default parameters) to compare reference sequences. Using these parameters, it is unlikely that closely related sequences would not be included in the network. All these steps are automated in the EGN software, freely available at http://www.evol-net.fr (last accessed May 23, 2013).

Network Construction: Oomycetes Network

Networks used in this study consisted of the translated transcripts from de novo assemblies of the *A. hypogyna* and the protein sequences from *Pyt. ultimum*, *Phy. infestans*, *Phy. ramorum*, *Thalassiosira pseudonana*, and *Sap. parasitica*. Networks were reconstructed as earlier, using protein sequences, as indicated in Armbrust et al. (2004), Tyler et al. (2006), Haas et al. (2009), and Levesque et al. (2010). In brief, all sequences were compared against one another using the BLAST algorithm (Altschul et al. 1997). Pairs of sequences were connected in a network if their BLAST *e*-value was less than 1e-20 and the two sequences presented >50%, >60%, >70%, >80%, >90%, or >95% similarity.

Network Analysis

We produced two types of networks: stringent networks with "full homology" edges (enforcing a >90% sequence length alignment threshold for pairs of sequences to be connected) and inclusive networks with both "full" and partial homology edges (when sequences presented a significant similarity for a shorter portion of their sequences). Using a custom python script, now implemented in SCAN, these gene similarity networks were filtered to contain only connected components with at least one reference node (i.e., *Pyt. ultimum* for the oomycetes data set) and one transcript node (e.g., *A. hypogyna* for this data set). In the *E. coli str. K-12 substr. MG1655*, *Sac. cerevisiae* uid 128, and *Pla. falciparum* analyses, every species in the network was used as a potential reference, and networks were filtered as above.

Filtered networks were analyzed using the custom R script, SCAN (or "SCAN stringent") (fig. 2), available at http://evol-net.fr (last accessed May 23, 2013). SCAN measures six distinct network features for each connected component: proportion of reference and transcript nodes, articulation point (local/global), longest monochromatic chain, proportion of nodes of degree one that are reference/transcript, and the JI. These analyses were performed on a computer with 2 quadcore Intel Xeon E5430 CPUs running at 2.66 GHz. The multithreaded version of SCAN processed the oomycete data sets described earlier in 15 min using eight cores.

For each connected component and for each index (except JI and longest monochromatic chain), we computed a proportion test (Newcombe 1998) between reference nodes and transcript nodes to determine whether their proportion in a connected component is significantly different (*P*-value threshold of 0.05). If the two proportions (i.e., of reference and transcript nodes) are not significantly different, we considered this connected component as "good" for this index.

For each index measure (except JI), the KS test (*P*-value threshold of 0.05) was used to compare the distributions of values of transcripts and reference nodes over all connected components. Under the KS test, the null hypothesis is that the two samples follow the same distribution. If so, the KS test *D* statistic equals 0, otherwise it is positive. A significantly large value of *D* allows for rejection of the null hypothesis, as measured by a *P* value estimating the probability for $D \geq 0$. SCAN utilizes the KS test to identify for each assembly, which network showed no significant differences in the distributions of index values for test and reference transcripts, for the highest number of indices.

The network we considered best was the one having the highest number of quality transcripts (according to proportion tests), and in the event of a tie, we selected the network having the highest number of distributions of indices for which no significant difference can be found between the reference and the transcript (according to the KS test). SCAN outputs a list of good quality transcripts from the selected network for each assembly method. To summarize, SCAN identifies transcripts that best resemble sequences from the biological reference. Each assembly is evaluated against the sequences of user-defined reference organisms in the network, using up to six network-based indices. Each connected component within the network is assessed based on: 1) the proportion of reference and transcript data, 2) the proportion of local and 3) global articulation points formed by reference and transcript data, 4) the longest monochromatic chain, 5) the proportion of transcript/reference data with a degree of 1, and 6) the JI of a pair of transcript and reference sequences. These indices are used to estimate topological features of the transcript sequences in comparison to reference data. Good transcripts are identified as those (directly connected to a reference sequence, in "SCAN stringent") that present topological properties, which cannot be distinguished from that of the reference sequences.

Proportion of Transcripts and References in Connected Components

The proportion of transcripts (nt) and reference (nr) sequences in each connected component was computed to test whether an assembly produced the same number of gene transcripts as exists in the reference connected component (fig. 2*A*). In cases of incomplete assembly, an excess of transcripts over reference proteins is expected in the connected component, because nt partial assemblies will connect to their $nr < nt$ homologs. Node proportion was calculated by dividing the total number of transcript (or reference) nodes in each connected component by the total number of nodes in that connected component.

Proportion of Transcript and Reference Articulation Points in Connected Components

For each connected component, we calculated the proportion of transcripts and reference nodes that were articulation points on two scales, local and global (fig. 2*B*). Local articulation points were tested by selecting all direct neighbor(s) for each transcript or reference node to create a new "local" connected component. If removing the transcript or reference node disconnected this "local" graph, this node was counted as a local articulation point. Global articulation points were tested by individually removing transcript or reference nodes in entire connected components. If removing an individual transcript or reference node disconnected the connected component, this node was counted as a global articulation point. The corresponding proportions were then calculated over the total number of nodes in the connected component. Both local and global articulation points are separate indices in the SCAN output.

Proportion Transcript and Reference of Degree One

The number of nodes of degree one for the transcript and the reference sequences for each connected component was calculated. We obtained their respective proportion by dividing these numbers by the total number of nodes in the connected component (fig. 2*D*). The two proportions were compared with a proportion test (*P*-value threshold > 0.05). If there is no significant difference between these proportions, this connected component is labeled as "good."

Longest Monochromatic Chain

The shortest path between any pair of nodes corresponds to the minimal number of edges required to connect these two nodes. If the shortest path between two nodes sharing a given label, that is, two nodes representing contigs from the same test condition, goes only across nodes with this given label, we call such a shortest path a monochromatic path (fig. 2C). For each connected component, SCAN computes the length of the longest monochromatic path connecting reference sequences and the length of the longest monochromatic path connecting transcripts generated in a given test condition. For components in which there is no monochromatic path, this value is 0. We did not compute any proportion test for this index because it is not a proportion. We considered all connected components in the network as "good," if the two distributions (transcript and reference) are not significantly different according to the KS test (*P*-value threshold of 0.05).

Jaccard Index

Unlike previous indices, the JI applies to a pair of nodes and not to an entire connected component. The JI was calculated for connected components in which a transcript node was directly connected to at least one reference node (fig. 2E). For each edge joining a reference and a transcript node, JI was calculated as the number of common neighbors divided by the total number of neighbors for these two nodes. Components for which at least one JI was greater than 0.9 were labeled as good.

Supplementary Material

Supplementary tables SI1–SI4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

I.M. carried out the assemblies, created networks, assisted with the conceptual aspects of the study, troubleshot SCAN, and analyzed the data. C.B. programmed SCAN, created networks, assisted with the statistical methods, and analyzed the data. P.L. helped conceive the study and assisted with network creation, statistical analysis, and analyzed the data. S.H. created the scripts to construct the networks. E.B. assisted in the conception of the study and it's coordination and analyzed the data. C.L. assisted in the conception of the study and its coordination. All authors wrote the manuscript. The authors thank two anonymous reviewers for insights that improved the final product. Sequencing of the *A. hypogyna* transcriptome was supported by the RI INBRE grant from NIH (no. 8P20GM103430-12). A fellowship for I.M. was provided by the USDA National Needs Graduate Program in Diseases of Marine Organisms (Grant no. 2008-38420-18737). S.H. was supported by a Genome Canada/Genome Quebec research grant (Genorem) to M. St-Arnaud and M. Hijri.

References

Adams MD, Celniker SE, Holt RA, et al. (196 co-authors). 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185–2195.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.

Alvarez-Ponce D, Lopez P, Bapteste E, McInerney J. 2013. Gene Similarity networks provide new tools for understanding eukaryote origins and evolution. *Proc Natl Acad Sci U S A.* 110:E1594–E603.

Armbrust EV, Berges JA, Bowler C, et al. (45 co-authors). 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306:79–86.

Beauregard-Racine J, Bicep C, Schliep K, Lopez P, Lapointe FJ, Bapteste E. 2011. Of woods and webs: possible alternatives to the tree of life for studying genomic fluidity in *E. coli*. *Biol Direct.* 6:39; discussion 39.

Bhattacharya D, Price DC, Bicep C, Bapteste E, Sarwade M, Rajah VD, Yoon HS. 2013. Identification of a marine cyanophage in a protist single-cell metagenome assembly. *J Phycol.* 49:207–212.

Bittner L, Halary S, Payri C, Cruaud C, de Reviers B, Lopez P, Bapteste E. 2010. Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. *Biol Direct.* 5:47.

Blattner FR, Plunkett G, Bloch CA, et al. (17 co-authors). 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453–1462.

Bowler C, Allen AE, Badger JH, et al. (77 co-authors). 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456:239–244.

Cock JM, Sterck L, Rouze P, et al. (77 co-authors). 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* 465:617–621.

Cocquet J, Chong A, Zhang G, Veitia RA. 2006. Reverse transcriptase template switching and false alternative transcripts. *Genomics* 88: 127–131.

Everett MV, Grau ED, Seeb JE. 2011. Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Mol Ecol Resour.* 11(1 Suppl):93–108.

Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M. 2011. Short read Illumina data for the *de novo* assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. *BMC Genomics* 12:317.

Gardner MJ, Hall N, Fung E, et al. (46 co-authors). 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419:498–511.

Grabherr MG, Haas BJ, Yassour M, et al. (21 co-authors). 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 29:644–652.

Haas BJ, Kamoun S, Zody MC, et al. (97 co-authors). 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461:393–398.

Holland BR, Huber KT, Moulton V, Lockhart PJ. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. *Mol Biol Evol.* 21:1459–1461.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 23:254–267.

Ivens AC, Peacock CS, Worthey EA, et al. (102 co-authors). 2005. The genome of the kinetoplastid parasite, *Leishmania major*. *Science* 309: 436–442.

King N, Westbrook MJ, Young SL, et al. (36 co-authors). 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451:783–788.

Levesque CA, Brouwer H, Cano L, et al. (49 co-authors). 2010. Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. *Genome Biol.* 11:R73.

Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. *Nat Rev Genet.* 12:671–682.

Martin J, Bruno VM, Fang Z, Meng X, Blow M, Zhang T, Sherlock G, Snyder M, Wang Z. 2010. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* 11:663.

Min XJ, Butler G, Storms R, Tsang A. 2005. OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res.* 33:W677–W680.

Newcombe RG. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med.* 17:857–872.

Normark S, Bergstrom S, Edlund T, Grundstrom T, Jaurin B, Lindberg FP, Olsson O. 1983. Overlapping genes. *Annu Rev Genet.* 17:499–525.

Otero JM, Vongsangnak W, Asadollahi MA, et al. (11 co-authors). 2010. Whole genome sequencing of *Saccharomyces cerevisiae*: from genotype to phenotype for improved metabolic engineering applications. *BMC Genomics* 11:723.

Palenik B, Grimwood J, Aerts A, et al. (42 co-authors). 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci U S A.* 104:7705–7710.

Prochnik SE, Umen J, Nedelcu AM, et al. (28 co-authors). 2010. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science* 329:223–226.

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28:1086–1092.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19: 1117–1123.

Tyler BM, Tripathy S, Zhang X, et al. (53 co-authors). 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313:1261–1266.

Zerbino DR. 2010. Using the Velvet *de novo* assembler for short-read sequencing technologies. *Curr Protoc Bioinformatics* Chapter 11:Unit 11.15.

Zerbino DR, Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18:821–829.

Zerbino DR, McEwen GK, Margulies EH, Birney E. 2009. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read *de novo* assembler. *PLoS One* 4:e8407.

Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P. 2011. Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* 12(14 Suppl):S2.

Zheng Y, Zhao L, Gao J, Fei Z. 2011. iAssembler: a package for *de novo* assembly of Roche-454/Sanger transcriptome sequences. *BMC Bioinformatics* 12:453.