ABSTRACT

Title of Dissertation:	INFERRING DINOFLAGELLATE GENOME STRUCTURE, FUNCTION, AND EVOLUTION FROM SHORT-READ HIGH-THROUGHPUT RNA-SEQ
	Theodore Robert Gibbons, Doctor of Philosophy, 2015
Dissertation directed by:	Professor Charles F. Delwiche, Cell Biology and Molecular Genetics

Dinoflagellates are a diverse and ancient lineage of globally abundant algae that have adapted to fill a diverse array of important ecological roles. Despite their importance, dinoflagellate genomes remain relatively poorly understood because of their enormous size. It is suspected that dinoflagellate genomes have expanded through rampant gene duplication, possibly using a lineage-specific mechanism that involves reinsertion of mature transcripts back into the genome, and that may rely on spliced leader trans-splicing for reactivation and processing of recycled transcripts. Draft genomes have recently been published for two extremely small endosymbiotic species. These genomes confirm expansion of nearly 10k gene families, relative to other eukaryotes. In the more complete genome, evidence for transcript recycling based on relict spliced leader sequences was found in over 5,500 genes. Genomic efforts in larger dinoflagellates have focused instead on transcriptome sequencing, but transcriptomes assembled from short-read HTS data contain very little evidence for rampant gene duplication, or for trans-splicing. I have shown that apparent disagreement with hypotheses related to ubiquitous trans-splicing and widespread gene duplication are the result of technological limitations. By leveraging the statistical power of high-throughput sequencing, I found that spliced

leader suffixes as short as six nucleotides are sufficient for positive identification. I also found that isoform sequences from families of conserved paralogs are systematically collapsed during assembly, but that many of these consensus sequences can be identified using a custom SNP-calling procedure that can be combined with traditional clustering based on pairwise sequence alignment to obtain a more complete picture of gene duplication in dinoflagellates. Efficient, automated homology detection based on pairwise sequence alignment is an equally challenging problem for which there is much room for improvement. I explored alternative metrics for scoring alignments between sequences using a popular procedure based on BLAST and Markov clustering, and showed that simplified metrics perform as well or better than more popular alternatives. I also found that Markov clustering of protein sequences suffers from a serious false positive problem when compared against manual curation, suggesting that it is more appropriate for pre-clustering of very large data sets than as a complete clustering solution.

INFERRING DINOFLAGELLATE GENOME STRUCTURE, FUNCTION, AND EVOLUTION FROM SHORT-READ HIGH-THROUGHPUT RNA-SEQ

by

Theodore Robert Gibbons

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory committee:
Professor Charles F. Delwiche, Chair
Professor Najib M. El-Sayed
Professor Carl L. Kingsford
Professor Thomas D. Kocher
Professor Stephen M. Mount

Preface

The completion of my PhD is the culmination of a lifetime of support from too many people to possibly name. I was very fortunate to have known and been supported by both of my parents and all four of my grandparents growing up, all of whom encouraged me to pursue higher education and fostered my passion for biology by allowing me to keep, and often even helping me catch and raise, many exotic pets throughout my formative years. I was also very fortunate to have an uncle with more experience caring for exotic pets, and to have been able to volunteer at the Tulsa Zoo as a teenager.

My early exposure to computers was somewhat less fortunate, as the first two in my house came preloaded with Windows 95 and Millennium Edition, respectively. It wasn't until halfway through undergrad that I discovered computers could be used to do things *faster* than a person could accomplish with pen and paper. My now extensive training in Computer Science cannot be attributed to any formal education, but instead stems from a series of friendships with computer scientists, most of whom I met through the Oklahoma chapter of Triangle Fraternity, the UMD Center for Bioinformatics and Computational Biology, and my student internship at Computomics and the University of Tübingen.

I would also like to give special thanks to two of my high school teachers, Rebecca "Doc" Simcoe and John Waldron, for helping me through a particularly difficult time in my life; my friends at the Pirate House who tolerated me coding through most social gatherings, and always saw me safely home from DC; all of the members of my committee for guiding me through different stages of graduate school; and Drs. Tsvetan Bachvaroff and Endymion Cooper, who acted as unofficial committee members in this final year.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SL.....spliced leader

DinoSL.....Dinoflagellate Spliced Leader

BLASTp.....basic local alignment search tool for protein sequences

MCL.....Markov clustering

CAMERA .....Community Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis

MMETSP .....Marine Microbial Eukaryotic Transcriptome Sequencing Project

EST .....expressed sequence tag

HTS.....high-throughput sequencing

RNA-Seq .....RNA sequencing

bp .....base pair

Kbp .....kilo base pair

Mbp.....mega base pair

Gbp .....giga base pair

pg .....picogram

CCMP .....Culture Collection of Marine Phytoplankton

HPI.....hours post infection

Po.glac .....*Polarella glacialis*

Ka.vene.....*Karlodinium veneficum*

Am.cart.....*Amphidinium carterae*

Gy.inst.....*Gyrodinium instriatum*

AmxAsSp.....*Amoebophrya* sp. ex *Akashiwo sanguinea* (dinospores)

AmxAs48.....*Amoebophrya* sp. ex *Akashiwo sanguinea* (48 HPI)

AmxKven .....*Amoebophrya* sp. ex *K. veneficum*

Pr.hoff.....*Prorocentrum hoffmannianum*

Pr.mica.....*Prorocentrum micans*

Pr.mini .....*Prorocentrum minimum*

Pr.3122.....*Prorocentrum* sp. CCMP3122

contig .....contiguously assembled sequence

CEGMA.....conserved eukaryotic genes mapping approach

CEG .....conserved eukaryotic gene (CEGMA cluster)

KOG.....eukaryotic orthologous groups

ORF ..... open reading frame

CDS ..... coding sequence

HMM ..... hidden Markov model

Pfam..... protein family

GO ..... gene ontology

PE..... paired end (sequencing)

QC..... quality control

RACE..... rapid amplification of cDNA ends

nt ..... nucleotide

poly(A)..... poly-adenylated

UTR ..... untranslated region

ppt ..... parts per thousand

COG..... clusters of orthologous groups

ECK ..... expanded CEGMA KOGs

SBS ..... self bit score

BSR..... bit score ratio

AL ..... anchored alignment length

BAL ..... bit score over anchored alignment length

E-value..... expectation value

NLE ..... negative common (base ten) log of the expectation value

1 Introduction

Dinoflagellates are globally abundant algae, estimated to represent about half of all identifiable phytoplankton within the sunlit ocean (de Vargas et al., 2015). They are important primary producers of oxygen, omega-3 fatty acids (Henderson & Mackinley, 1991), and of general biomass. Some are famous for their beautiful bioluminescence (Wilson & Hastings, 1998). Others are infamous for deadly coastal blooms, staining the waters an ominous deep red, suffocating and starving both themselves and other local marine life, then releasing a variety of neurotoxins (Band-Schmidt et al., 2010; Ignatiades & Gotsis-Skretas, 2010; Reguera et al., 2014; Watkins, Reich, Fleming, & Hammond, 2008).

Dinoflagellates are monophyletic group, believed to have diverged from other alveolates nearly a billion years ago, and from model plants and opisthokonts over 1.5 billion years ago (Bhattacharya, Yoon, Hedges, & Hackett, 2009). In this time, they have adapted to a diversity of ecosystems and lifestyles. The most abundant species are phototrophic, although they do not share a single plastid origin (Delwiche, 1999), and not all photosynthetic species are free-living plankton. *Symbiodinium* are endosymbionts that sustain coral and provide their color (Roth, 2014). When this relationship fails, the corals appear "bleached" and can starve, leading to the destruction of entire reef habitats. Many dinoflagellates are active hunters that only photosynthesize to supplement their diets between meals (P. J. Hansen, 2011), while others are exclusively heterotrophic (Schnepf & Elbrächter, 1992). Among the non-photosynthetic species, *Amoebophrya* have received a lot of attention because they parasitize other dinoflagellates, including many of the bloom-forming species (Park, Yih, & Coats, 2004). *Amoebophrya* are also

interesting as representatives of the Syndiniales, a class of dinoflagellates best known for parasitizing a variety of hosts, and that is sister to the core Dinophyceae, which contains the majority of characterized species (Bachvaroff et al., 2014).

Amoebophrya and Suessiales (composed of *Symbiodinium* and *Polarella*) are also important because they have some of the smallest genomes among dinoflagellates (LaJeunesse, Lambert, Andersen, Coffroth, & Galbraith, 2005). Small is a relative term, however, as other dinoflagellates have some of the largest genomes on earth. The largest dinoflagellate genome measured to date is that of *Prorocentrum micans*, with a haploid size estimated to be over 110 Gbp (LaJeunesse et al., 2005; Veldhuis, Cucci, & Sieracki, 1997). The tight correlation between cell size and genome size reported in the same study suggests that larger species, such as *Akashiwo sanguinea* (Menden-Deuer & Lessard, 2000), may have genomes approaching a terabase, but these are too large to be confirmed using flow cytometry.

The extreme size of dinoflagellate genomes has made them particularly challenging to study, and what little is known suggests that they differ from all model organisms in very fundamental ways (Senjie Lin, 2011; Wisecaver & Hackett, 2011). They are one of a small assortment of disparate eukaryotes to trans-splice short mini exons called *spliced leaders* onto the 5' ends of mature mRNA transcripts (Lidie & van Dolah, 2007; H. Zhang & Lin, 2009; H. Zhang et al., 2007). They package their genomes into an atypical crystalline structure that remains permanently condensed, even through cellular division (Bouligand & Norris, 2001; Gautier, Michel-Salamin, Tosi-Couture, McDowall, & Dubochet, 1986; Gavrila, 1977). They also possess many highly duplicated genes that are often arranged into tandem arrays with fairly short intergenic sequences (Bachvaroff &

Place, 2008; Beauchemin et al., 2012; Bertomeu & Morse, 2004; S. Lin et al., 2015; Liu & Hastings, 2006; Mendez, Delwiche, Apt, & Lippmeier, 2015; Sano & Kato, 2009; Shi, Zhang, & Lin, 2013; H. Zhang, Hou, & Lin, 2006; H. Zhang & Lin, 2003). It is therefore possible that large dinoflagellate genomes are unusually gene rich for their size, and that many of the genes have been highly duplicated near each other within the genome, creating long stretches of genomic sequences that are rife with gene-length repeats, greatly complicating assembly of genome sequences from short-reads (Kingsford, Schatz, & Pop, 2010; Wetzel, Kingsford, & Pop, 2011). This means that sequencing a large dinoflagellate genome will require not simply large amounts of data, but large amounts of long-read data with low error rates.

Long-read high-throughput sequencing (HTS) technologies are only now becoming affordable for bacteria and smaller eukaryotes. The dinoflagellate genomics research community has therefore focused on identifying tractable targets for short-read HTS. A draft assembly of the 1.5-Gbp *Symbiodinium minutum* genome was published in 2013, yet more than half of it remains unassembled due to the presence of many highly repetitive sequences (Shoguchi et al., 2013). A more recent effort with *Symbiodinium kawagutii* was more successful, reconstructing nearly 80% of the 1.2-Gbp genome (S. Lin et al., 2015). Both studies provided valuable information about gene family expansion, and revealed a strong bias for genes to be oriented along the same strand, as they are in *Trypanosomes*. The *S. kawagutii* assembly also contains relict spliced leaders embedded immediately upstream from 15% of the genes, suggesting that they were duplicated through reinsertion of mature transcripts back into the genome (Slamovits & Keeling, 2008).

How these numbers compare with the genomes of larger dinoflagellates remains unknown. Sequencing efforts within the free-living Dinophyceae have instead focused on transcriptomes (Beauchemin et al., 2012; Cooper, Bentlage, Gibbons, Bachvaroff, & Delwiche, 2014; Keeling et al., 2014; Meyer et al., 2015; Ryan, Pepper, & Campbell, 2014; H. Zhang, Zhuang, Gill, & Lin, 2013; S. Zhang et al., 2014; Y. Zhang, Zhang, Lin, & Wang, 2014). Transcriptome sequencing offers a convenient way to obtain what are arguably the most important parts of a genome, yet it is not the same as exome sequencing. Inferring genome structure from transcriptome data can be extremely challenging, and the potential for misassembly of short-read data introduces further complications. Despite this, the volume of data generated using short-read HTS has provided information about dinoflagellate genomes on an unprecedented scale.

My work has focused on characterizing gene duplication in dinoflagellates, using only short-read high-throughput RNA sequencing data. This has required careful application of a wide variety of sophisticated methods, and the development of many custom tools and pipelines. I have identified and characterized several significant flaws within popular methods for transcriptome sequencing and analysis. Where possible, I have leveraged the statistical power of deep sequencing to reach beyond these limitations, and confirmed broad support for several important open hypotheses on a genomic scale across a diverse set of dinoflagellates, most of which have very large genomes that will likely remain unsequenceable for years to come.

2 Unassembled paralogs are a significant source of paralogy underestimation in dinoflagellates

2.1 Background

Dinoflagellates are a major component of global aquatic microbial diversity that have adapted to a wide variety of lifestyles and ecological roles (de Vargas et al., 2015; Hackett, Anderson, Erdner, & Bhattacharya, 2004; Park et al., 2004). They are significant global producers of oxygen, biomass, polyunsaturated fatty acids, and an assortment of neurotoxins (Harrington, Beach, Dunham, & Holz, 1970; Kalaitzis, Chau, Kohli, Murray, & Neilan, 2010). Despite decades of research, their genomes remain poorly characterized because of their enormous size and unusual structure (Hackett et al., 2004; LaJeunesse et al., 2005; Senjie Lin, 2011; Veldhuis et al., 1997; Wisecaver & Hackett, 2011). One particularly interesting feature of dinoflagellate genomes is the presence of many highly duplicated genes.

Extreme gene duplication within dinoflagellates was first discovered while studying diel regulation of bioluminescence genes in *Lingulodinium polyedrum* (Lee, Mittag, Sczekan, Morse, & Hastings, 1993). Several studies have since revealed the presence of other highly duplicated genes within a variety of dinoflagellates (Bachvaroff & Place, 2008; Beauchemin et al., 2012; Bertomeu & Morse, 2004; S. Lin et al., 2015; Liu & Hastings, 2006; Mendez et al., 2015; Sano & Kato, 2009; Shi et al., 2013; H. Zhang et al., 2006; H. Zhang & Lin, 2003). The mechanism of duplication is unknown, but targeted genomic sequencing has revealed that many of these genes are conspicuously arranged into tandem arrays, separated by short intergenic spacers. This suggests that dinoflagellate genomes have expanded through rampant duplication of individual genes, rather than

entire chromosomes, and that they may have unusually high gene densities for their size (Beauchemin et al., 2012; Hou & Lin, 2009).

Other duplicated dinoflagellate genes, such as form II Rubisco, are arranged into long transcriptional units (TU) containing multiple coding units (CUs) (Rowan, Whitney, Fowler, & Yellowlees, 1996; H. Zhang & Lin, 2003), which are translated together, then cleaved into active proteins. Subsequent discovery of the dinoflagellate trans-splicing spliced leader (DinoSL) fueled speculation that other genes might be expressed into polycistronic pre-mRNA that is cleaved into mature during trans-splicing (H. Zhang et al., 2007), as has been observed in trypanosomes (Preußer, Jaé, & Bindereif, 2012), although this remains unconfirmed (Beauchemin et al., 2012). Learning more about these genes and their impact on dinoflagellate genomic structure, function, and evolution, will require being able to reliably identify other highly duplicated genes for further targeted studies.

In other contexts, highly duplicated genes are a nuisance to be avoided. Many powerful computational and molecular techniques, such as phylogenetic inference or the creation of gene knockout mutants, rely on the use of single copy genes. When sequences with unidentified paralogs are mistakenly interpreted as being single-copy, it can cause experiments to fail, or worse, can lead to spurious results from which researchers might unknowingly draw incorrect conclusions. Access to complete genome sequences could eliminate these problems, but the free-living Dinophyceae have some of the largest genomes ever measured, making them impractical to sequence using even the latest high-throughput sequencing (HTS) technologies.

The haploid genome of *Prorocentrum micans* has been estimated to be over 110 Gbp (LaJeunesse et al., 2005; Veldhuis et al., 1997), with estimates varying widely due to methodological challenges. The genomes of even larger species, such as *Prorocentrum hoffmannianum* (Herrera-Sepúlveda et al., 2015), *Gyrodinium instriatum* (Freudenthal & Lee, 1963), and *Akashiwo sanguinea* (Menden-Deuer & Lessard, 2000), are so large that they cannot even be measured using standard flow cytometric methods. For comparison, the largest genome sequenced to date is that of the 20 Gbp genomes of the loblolly pine (Neale et al., 2014).

Not all dinoflagellates have such enormous genomes (LaJeunesse et al., 2005). *Amoebophrya* and *Symbiodinium*, best known as intracellular parasites of other dinoflagellates and intracellular symbionts of coral, respectively, have some of the smallest cell and estimated genome sizes among the dinoflagellates (LaJeunesse et al., 2005). Draft assemblies of the relatively small 1.5 and 1.2 Gbp genomes of *Symbiodinium minutum* and *Symbiodinium kawagutii* genomes were recently published, providing a wealth of insights about dinoflagellate genomic structure (S. Lin et al., 2015; Shoguchi et al., 2013). By analyzing genes that were found to be homologous with annotated sequences from other eukaryotes, they identified nearly 10k gene families that had expanded within *Symbiodinium*, of which 2.5k and 1.25k were unique to *S. minutum* or *S. kawagutii*, respectively.

How these numbers compare to the repeat-rich unassembled half of the *S. minutum* genome, or the genomes of the larger dinoflagellates, remains unknown. In these larger dinoflagellates, recent *omics efforts have instead focused on the more tractable challenge of transcriptome sequencing using short-read HTS (Cooper et al., 2014;

Keeling et al., 2014; Meyer et al., 2015; Ryan et al., 2014; H. Zhang et al., 2013; S. Zhang et al., 2014; Y. Zhang et al., 2014). Curiously, these studies have not reported any evidence for extreme gene duplication, other than to say that some duplicated genes appear to be highly conserved (Beauchemin et al., 2012).

I show here that families of conserved paralogs are misrepresented in *de novo* assemblies from short-read HTS as a result of being collapsed down into small numbers of consensus sequences, and that this phenomenon is widespread within a diverse set of free-living dinoflagellates. The assembled sequences representing families of unassembled paralogs can be identified through the use of a simplified SNP-calling procedure. Combining this procedure with traditional clustering methods based on pairwise sequence alignment provides a much more complete picture of gene duplication, without requiring a reference genome or targeted sequencing.

2.2 Results and Conclusions

2.2.1 RNA Sequencing and Assembly

Messenger RNA from eleven species of dinoflagellates were sequenced for this study, generating over half a billion usable paired-end (PE) Illumina RNA-Seq reads containing over 100 billion nucleotides of usable transcriptome sequence data (Table S1). All libraries were sequenced in multiplexed runs on Illumina HiSeq instruments, except for *P. glacialis*, which was sequenced in its own lane on an Illumina GAIIx shortly before the release of the HiSeq instruments. The GAIIx reads do not appear to be of lower quality than the HiSeq reads (Figure 1), although they were shorter (78 bp vs. 101 bp), and fewer of them were obtained compared to most of the multiplexed HiSeq libraries. On the opposite extreme, six *G. instriatum* libraries were extracted from subcultures that

had been exposed to a variety of growth conditions, then multiplexed together into a full HiSeq lane for that organism, generating 150 million usable PE reads and representing more than a quarter of the total data used in this study.

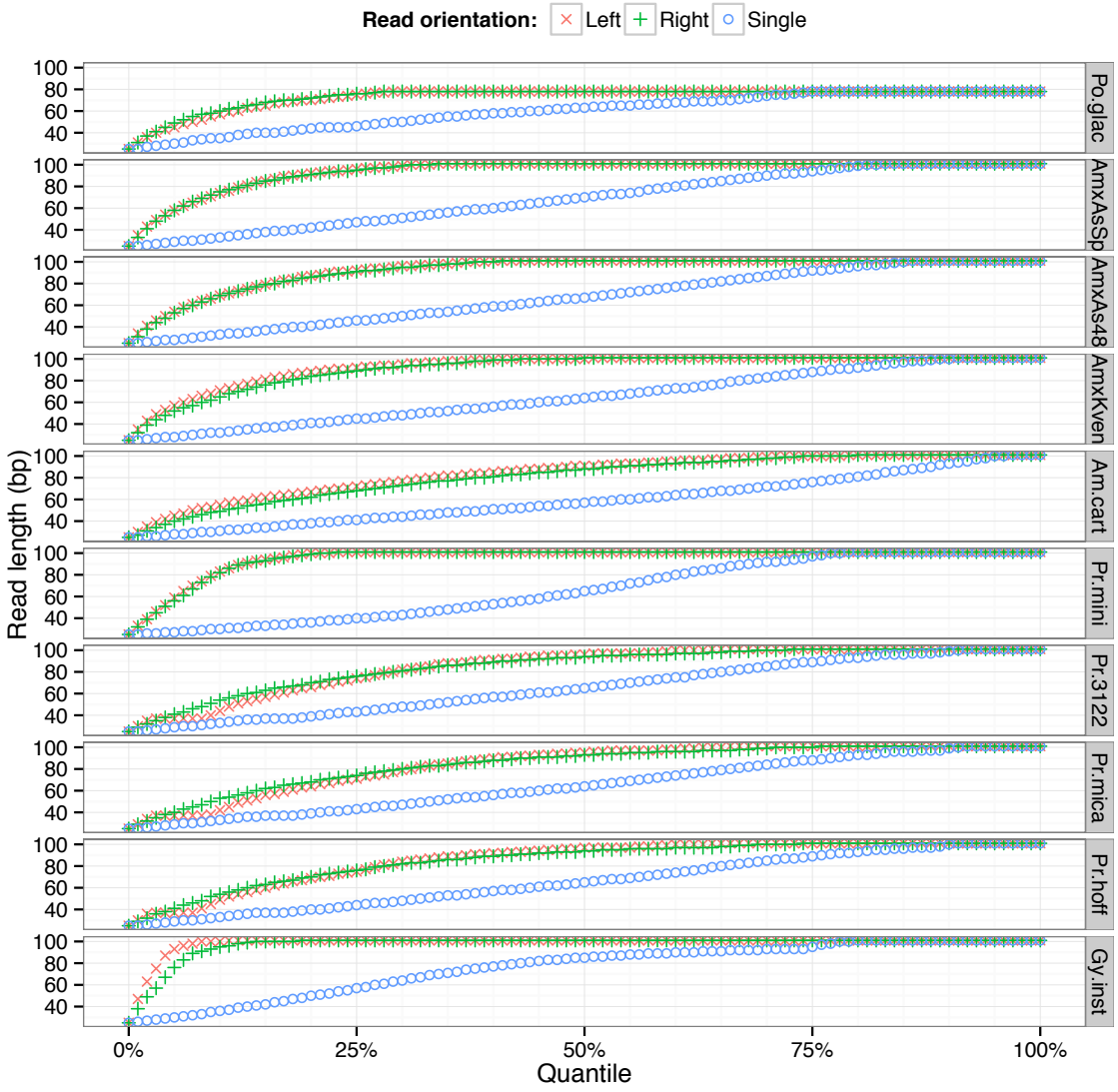


Figure 1 - Read lengths by quantile for each assembly

Lengths of trimmed reads by length-ordered quantile and read orientation, faceted by the assembly to which they contributed (see Methods for assembly abbreviations). Colored shapes indicate orientation of the reads. Paired end reads are respectively represented as red ×'s and green +'s that form an asterisk when they. All fragments were sequenced from both ends, but some reads were orphaned during quality control. These single-end reads are represented as blue ○'s.

The trimmed reads were assembled by organism (or organism combination) into a total of over a million and a half contiguous sequences ("contigs") by Trinity, which attempts to reconstruct complete mRNA splice isoform sequences from short-read RNA-Seq data. Trinity organizes these assembled isoform sequences into inferred genes based on the structure of the assembly graphs. In our experience, these Trinity genes do not always match the biological genes, especially when the biological genes share very high sequence identity, but they do offer a conservative clustering of very closely related sequences, including partially assembled isoform fragments that can be very difficult to cluster using traditional pairwise sequence alignments.

The default minimum length for Trinity isoforms is 200 nucleotides, which is 97 nucleotides shorter than the 99-codon minimum contig length required for the TransDecoder program packaged with Trinity to identify coding sequences (CDS). This study is focused on coding genes, so the four hundred thousand Trinity isoforms between 200 and 296 nucleotides were excluded from most analyses, leaving just over a million sequences organized into nearly 900k Trinity genes (Table 1).

Table 1 - Assembly statistics

	Trinity		Lengths			
			Combined	Median	Mean	Max
Assembly	Genes	Isoforms	(Mbp)	(nt)	(nt)	(nt)
Po.glac	58,733	84,519	55	511	656	10,782
AmxAsSp	36,613	46,624	79	867	1,688	33,758
AmxAs48	135,948	163,554	204	825	1,245	19,729
AmxKven*	44,453	50,487	82	856	1,624	23,599
Ka.vene*	104,718	116,807	120	723	1,028	26,832
Am.cart	45,996	54,776	68	981	1,232	14,160
Pr.mini	94,059	120,554	116	706	961	10,772
Pr.3122	85,735	109,873	101	712	919	13,451
Pr.mica	94,036	139,598	156	927	1,115	8,938
Pr.hoff	54,838	73,607	88	1,007	1,199	12,279
Gy.inst	118,166	193,837	249	993	1,286	15,631
Totals	873,295	1,154,236	1,317	-	-	-
Averages	79,390	104,931	120	828	1,178	17,266

**Amoebophrya* sp. ex *K. veneficum* and its host were not sequenced and assembled separately. Instead, sequences from the combined assembly were separated based on average per-gene GC% (see Methods)

One of the few analyses to include the short 200-296 nt sequences was the evaluation of Transcriptome assembly completeness using the Conserved Eukaryotic Genes Mapping Approach (CEGMA) (Table 2). CEGMA estimates completeness of *de novo* assemblies by searching for a set of 248 "ultra-conserved" CEGs. The Trinity assemblies for most of the free-living dinoflagellate transcriptomes contain full-length isoforms for 74-88% of the CEGs, and 85-91% when partial sequences were included. The *Prorocentrum* sp. CCMP3122 assembly was a bit worse than the others, containing full-length contigs for only 67% of the CEGs, although even this outlier still performed considerably better than the *P. glacialis* assembly, whose numbers were closer to those for the parasitic *Amoebophrya*.

Table 2 - CEGMA coverage

	Complete		Partial	
Assembly	Count	Percent	Count	Percent
Po.glac	140	56%	189	76%
AmxAsSp	96	39%	113	46%
AmxAs48	219	88%	225	91%
AmxKven*	135	54%	161	65%
Ka.vene*	204	82%	218	88%
Am.cart	211	85%	215	87%
Pr.mini	204	82%	215	87%
Pr.3122	165	67%	200	81%
Pr.mica	190	77%	210	85%
Pr.hoff	184	74%	204	82%
Gy.inst	217	88%	222	90%
Averages	179	72%	198	80%

Note: The CEGMA analysis preceded the TransDecoder predictions and included contigs (isoforms) as short as 200 nt.
**Amoebophrya* sp. ex *K. veneficum* and its host were not sequenced and assembled separately. Instead, sequences from the combined assembly were separated based on average per-gene GC% (see Methods)

Intracellular parasites often undergo genome reduction and gene loss (Katinka et al., 2001; Keeling, 2004; Keeling et al., 2010; McCutcheon & Moran, 2011; Moran, 2002; Slamovits, Fast, Law, & Keeling, 2004), so the observation that the *Amoebophrya* assemblies lack many CEGs is not necessarily indicative of deficiencies with the data. In the case of *P. glacialis*, however, which is free-living and for which the sequencing coverage should have been deeper than for many of the other free-living dinoflagellates, the poor CEGMA report is likely indicative of an inferior assembly, presumably caused by the shorter GAIIx reads. This suspicion is further supported by looking at the contig lengths for the assemblies (Table 1), including the fractions lost by removal of the short 200-296 nt sequences (Table S2).

2.2.2 Coding sequence identification and annotation

Paralogy detection typically begins with pairwise alignments of all sequences obtained from one or more species, often after identifying and translating the coding regions. This conversion from nucleotide to protein sequences is useful because protein sequences are shorter and have larger alphabets that contain more information within each character, which are desirable features from an algorithmic perspective. It is also useful because protein sequences tend to be more highly conserved than mRNA over evolutionary time scales. The scores from pairwise sequence alignments are then analyzed to infer clusters of homologous sequences. There are many programs available for these respective tasks, such as the BLAST and MCL combination popularized by OrthoMCL (F Chen, Mackey, Stoeckert, & Roos, 2006; L. Li, Stoeckert, & Roos, 2003).

Coding sequences were identified within the *de novo* Trinity assemblies using the TransDecoder program packaged with Trinity. In addition to considering the length and hexamer frequencies within each ORF, TransDecoder also calls HMMer3 to identify Pfam-A domains. Table S3 shows the numbers of Trinity genes and isoforms in which TransDecoder identified at least one CDS, the numbers of Trinity genes and isoforms in which at least one CDS was annotated with at least one Pfam-A domain, and the numbers of Trinity genes and isoforms in which at least one Pfam-A annotated CDS was subsequently mapped to one or more GO terms using the Pfam2GO mapping provided by InterPro.

After removal of contigs shorter than 297 nucleotides, the total number of remaining Trinity genes in which TransDecoder identified at least one CDS within at least one isoform was above 75%, with an average frequency just over one CDS per coding isoform. A little more than a third of the translated CDSs were annotated with at least one

Pfam-A domain, and roughly half of these Pfam-A domains were mapped to GO terms. For comparison, the proteomes of the seven model eukaryotes used to create the KOG database were run through the same protein sequence annotation pipeline (Table S4). Annotation statistics for the novel dinoflagellate transcriptomes have been summarized in the figures by Trinity gene in an effort to reduce over- or under-counting stemming from misassembly. Paralogs with long stretches of high sequence conservation are particularly problematic for assembly programs, especially when the conserved regions extend beyond the lengths of the sequenced fragments. Figure S1 illustrates the highly variable numbers of coding genes across all organisms, but Figure 1 reveals that the fractions of Pfam-A annotated genes were pretty consistent across the transcriptomes of the free-living dinoflagellates, as were the fractions of Pfam-A annotated protein sequences across the model organisms. Figure 1 also shows that the relative annotation rates were nearly twice as high in the model organisms as in the novel dinoflagellate transcriptomes.

The lower annotation rates in the dinoflagellates are certainly due in part to them having diverged from model organisms beyond the degree to which model organisms have diverged from each other, causing their scores to drop below the internal cutoffs for the Pfam-A HMMs. It could also be indicative of transcripts for certain homologous genes being much less abundant within the cells, to the point of being so outnumbered that their coverage ended up not being deep enough to facilitate *de novo* assembly. A fundamental limitation of transcriptome analysis in the absence of a corresponding reference genome is that genes that are absent from the genome are indistinguishable from genes that are simply not being actively expressed, or that were expressed in undetectably small amounts when the RNA was extracted. It is of course also possible that a significant

fraction represent genes that are unique to dinoflagellates, or at least that are absent from model plants and opisthokonts. Distinguishing between divergent and novel sequences will become easier as dinoflagellate transcriptome and genome data become more abundant, and as their quality improves.

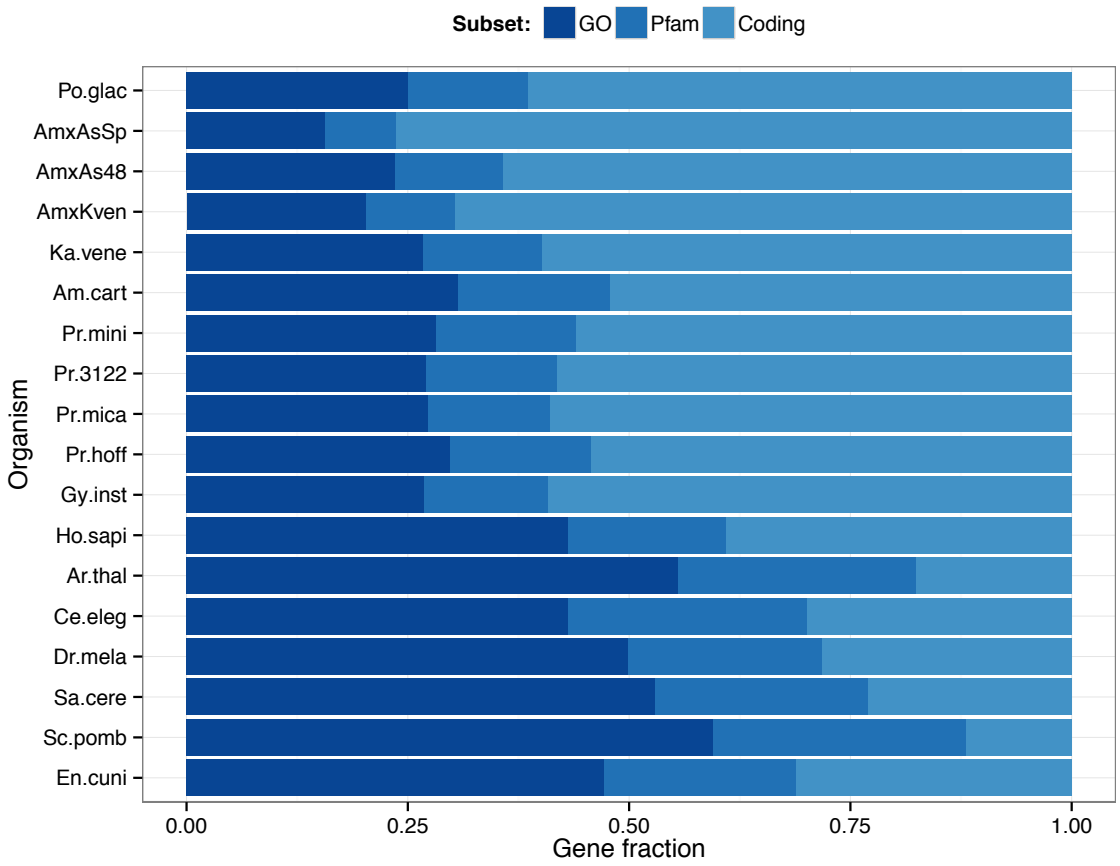


Figure 2 – Relative annotation statistics for coding genes

Bars indicate 100% of either, for the novel dinoflagellate transcriptomes, the Trinity genes in which TransDecoder identified at least one coding isoform, or, for the seven KOG organisms, the protein sequences contained within the proteomes that were released with the KOG database. Both are referred to here as genes. Colors indicate the fractions of these genes that contained at least one translated CDS sequence that was annotated by TransDecoder to contain at least one Pfam-A domain (Pfam), at least one such Pfam-A domain that was mapped to at least one GO term by the Pfam2GO mapping released by the GO consortium (GO), or if no translated CDSs were annotated with any Pfam-A domains (Coding). (Note: *Amoebophrya* sp. ex *K. veneficum* and its host were not sequenced and assembled separately. Instead, sequences from the combined assembly were separated based on average per-gene GC% (see Methods))

2.2.3 Identifying divergent paralogs

The translated coding sequences identified within the *de novo* assemblies were clustered together using BLASTp and MCL (T. Gibbons, 2015). More than 70% of the protein sequences were assigned to a cluster. For 90% of the Trinity genes contributing a clustered sequence, all of the sequences from that Trinity gene were contained within a single MCL cluster. Conversely, the MCL clusters spanned more than three Trinity genes apiece, on average. The distribution exponentially decayed, however, such that half of the MCL clusters contained exactly one complete Trinity gene, while on the other extreme, a few MCL clusters spanned over a thousand Trinity genes.

To characterize the most highly duplicated genes, the Pfam- and GO-annotated Trinity genes were ranked according to the highest number of MCL-identified paralogs for any translated CDS originating from them, and the top 10% of genes were analyzed for GO term enrichment relative to the overall set. Many terms were enriched, including several complete GO subtrees consistent with genes that are known to be highly duplicated within certain dinoflagellates, such as photosystems I & II, carbon fixation, and cytoskeleton assembly (T. Gibbons, 2015). Two previous studies of gene duplication in dinoflagellates focused on organisms that were also included in this study, providing opportunities to directly evaluate the success of using standard tools to identify highly duplicated genes within *de novo* assemblies of dinoflagellate transcriptomes from short-read HTS data.

2.2.3.1 Comparison with published copy numbers of actin in *A. carterae*

Bachvaroff & Place (2008) used targeted Sanger sequencing of cDNA and corresponding genomic regions to identify at least 24 distinct copies of actin in *A. carterae*. Using this collection of partial and full-length reference sequences as tBLASTn queries, the

corresponding Trinity isoforms were identified within the *A. carterae* assembly. Among these, there were two nearly complete copies, respectively stretching 1,326 and 1,383 nucleotides, and containing the complete canonical actin CDS, flanked on either side by complete or partial untranslated regions (UTRs) (Figure 3).

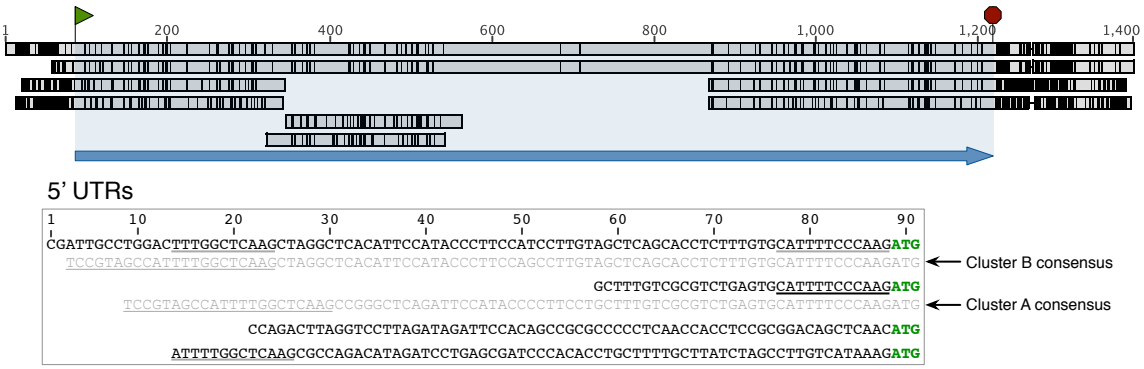


Figure 3 - Multiple alignment of actin sequences identified in the *A. carterae* assembly

Multiple alignment of two full-length and six partial actin sequences from the *A. carterae* assembly. Blue box and arrow indicate the coding region. Black bars indicated either disagreement with a consensus determined by simple majority voting, or lack of a consensus. Inset shows the four 5' UTR sequences, two of which matched those identified by Bachvaroff & Place (2008) as belonging to the two most abundant families of actin, termed "cluster A" and "cluster B". DinoSL, partial-SL, and SL-like sequences are underlined. [Images were modified from MAFFT alignment viewed in Geneious]

The 5' UTRs of the full-length sequences match those reported by Bachvaroff and Place for the two groups of highly expressed actin genes they referred to as clusters A & B. Both contain the distinctive SL-like sequence immediately upstream of the start codon. The 5' UTR matching cluster A appears to have been truncated less than half way to the 5' cap, while the 5' UTR matching cluster B extends past both known splice acceptor sites, suggesting the presence of an additional site even further upstream. Both coding sequences end in the TGA stops found only in the cluster B sequences, however, and both contain identical 156-nt 3' UTRs, followed by short 7-nt poly(A) tails. This suggests that the Trinity isoform with the 5' UTR from cluster A is actually a chimera created by the assembler.

Three pairs of fragmented sequences were also found to have very high identity to the reference protein sequence, and to be consistent with the actin gene structure. One pair included alternate 5' UTR sequences and about 85 codons of the CDS. Neither of these matched the 5' UTRs for clusters A or B and neither contained the distinctive SL-like sequence upstream of the start codon, although one ended in a 13-nt DinoSL suffix at its 5' end. The second pair of fragments began roughly where the first pair ended, and offered alternate sequences for the following stretch of 70+ internal codons. The final pair of sequences each included the last 116 codons, TGA stops, and about 165 nt of 3' UTR sequences, which diverged from both each other and the 3' UTR sequence shared by the full-length copies. Only one of the 3' fragments ended in a short poly(A), although all four 3' UTRs were very similar in length, and the one lacking a poly(A) was a bit shorter than the others. Whether these partial sequences correspond to only two additional variants that could not be completely pieced together by the assembler, or six individual copies, the total number of observed sequences is well short of the 24+ unique sequences identified using targeted sequencing.

2.2.3.2 Comparison with published copy numbers of form II rubisco in *P. minimum*

Direct evidence for at least 10 unique transcribed units (TU) encoding form II (nuclear encoded) rubisco was described in *P. minimum* by H. Zhang & Lin (2003). Each transcriptional unit sequenced by Zhang and Lin contained four coding units (CU), similar to earlier work by Rowan et al. (1996) that found three coding units per transcriptional unit in *Symbiodinium*. Using quantitative PCR to analyze serially diluted genomic DNA, Zhang and Lin estimated that the complete genome contains a total of 148±16 CUs, arranged into 37±4 TUs. Protein reference sequences for form II rubisco

were downloaded from GenBank from both this study, and a similar study from the same lab repeated later in *Prorocentrum donghaiense* (Shi et al., 2013). These were queried against the *P. minimum* assembly using tBLASTn. Three sequences, each about 1700 nt, aligned to multiple reference sequences with 96% amino acid identity from their 5' ends to stops located 508 codons in, covering all but about 200 nt of 3' UTR in each Trinity isoform (Figure 4). One UTR ended in a short poly(A) of 7 nt, the other two did not. No 5' UTRs could be identified.

In both *Symbiodinium* and *P. minimum*, each transcript is translated into a single polyprotein spanning all coding units, connected by linker peptides that are subsequently cleaved to yield mature proteins (Rowan et al., 1996; H. Zhang & Lin, 2003). Close inspection of the Trinity isoforms revealed that each CDS contained exactly the 21-codon linker and the 487-codon sequence of the terminal (3') coding unit, which includes a 3' terminal LS* motif not found in the non-terminal coding units (H. Zhang & Lin, 2003). Trinity assembled three variants of the terminal coding unit, and it appears these contigs absorbed the reads from the other non-terminal coding units. It is unclear why Trinity was unable to assemble even a single 5' UTR when it seemed to have had no trouble assembling multiple variants of the 3' UTR.

An additional Trinity isoform was found to have 92-93% identity to a variety of reference sequences over its full length, but it spanned only 343 internal codons near the 5' end of a single coding unit. BLASTn was unable to identify any regions of significant similarity between this sequence and the other three, making it unclear if this sequence actually represents an additional gene encoding form II rubisco. Even with the inclusion of this

sequence, however, the *P. minimum* assembly contains only a small fraction of the diversity that was expected based on targeted sequencing and qPCR.

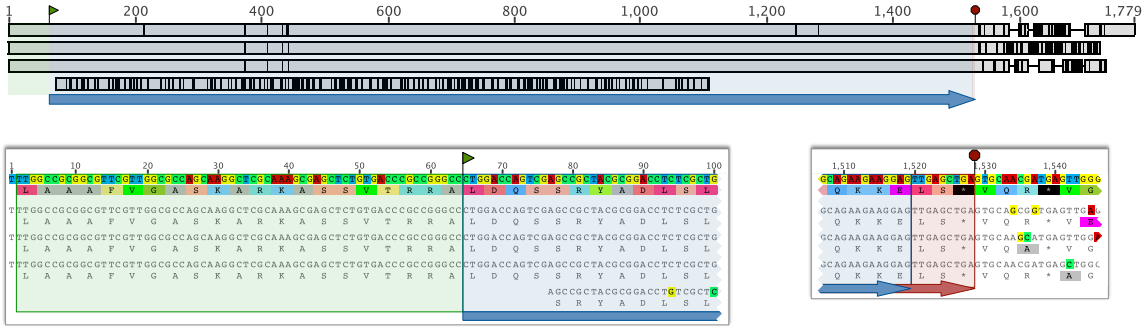


Figure 4 - Multiple alignment of form II rubisco sequences identified in the *P. minimum* assembly

Multiple alignment of one partial and three full-length form II rubisco sequences from the *P. minimum* assembly. Black bars indicated either disagreement with a consensus determined by simple majority voting, or lack of a consensus. It is believed that the dinoflagellate form II rubisco is expressed in transcriptional units that contain four coding units, which are translated together into a single polyprotein (see Figs. 1 & 2 of Zhang and Lin (2003) [11]). Blue box and arrow indicate a single form II rubisco coding region. Green box indicates the linker sequence separating coding regions. Red box indicates unique terminus of the 3' coding unit. Left inset shows that the 5' ends of all three full-length sequences began with the exact linker sequence identified by Zhang and Lin. Right inset shows that all three full-length sequences contain the extra LS* found only in terminal coding units. No sequences containing the 5' UTR of the overall transcriptional unit could be identified. Instead, it appears that all coding units from all transcriptional units were collapsed into three full-length versions of the terminal coding unit with 3' UTRs of about 200 nt each. [Images were modified from MAFFT alignment viewed in Geneious]

In both cases, all sequences identified for each gene within each organism were co-clustered by MCL. The MCL cluster containing actin from the *A. carterae* assembly included a total of 238 partial or full-length translated TransDecoder CDSs and spanned all 11 assemblies. The cluster containing the four *P. minimum* form II rubisco sequences lacked any sequences from *K. veneficum* or the associated *Amoebophrya* parasite, but spanned all nine of the other assemblies. Subsequent investigation of the assemblies from *K. veneficum* and the associated *Amoebophrya* parasite confirmed the absence of form II rubisco transcripts from both assemblies.

Overall, the assembled sequences are consistent with what has been previously shown using targeted sequencing, but in both cases the number of identified sequences fell well short of capturing the diversity of sequences that are known to be present within these organisms based on direct evidence. Each targeted sequencing study also provided strong evidence for the presence of additional variants that have not yet been sequenced, so a truly complete transcriptome should contain novel variants in addition to all of the variants sequenced in the smaller studies.

2.2.4 Identifying conserved paralogs

One likely source of undercounting is that some of the sequences reported by Trinity do not actually represent individual transcripts, but instead represent families of highly conserved paralogs that have been collapsed into consensus sequences during assembly. Figure 5 illustrates a hypothetical scenario in which a gene has been repeatedly duplicated into many functional copies whose transcripts were then sequenced using short-read HTS and partially reconstructed by *de novo* assembly. In the illustration, the resulting set of contigs accurately represent a direct subset of the original sequences, although chimeric sequences can also be created during assembly, as apparently happened with actin in *A. carterae*. Either way, reads from paralogs that are similar enough to be collapsed during assembly should map back to the corresponding consensus sequences, where the minor differences between them would appear as SNPs and small InDels.

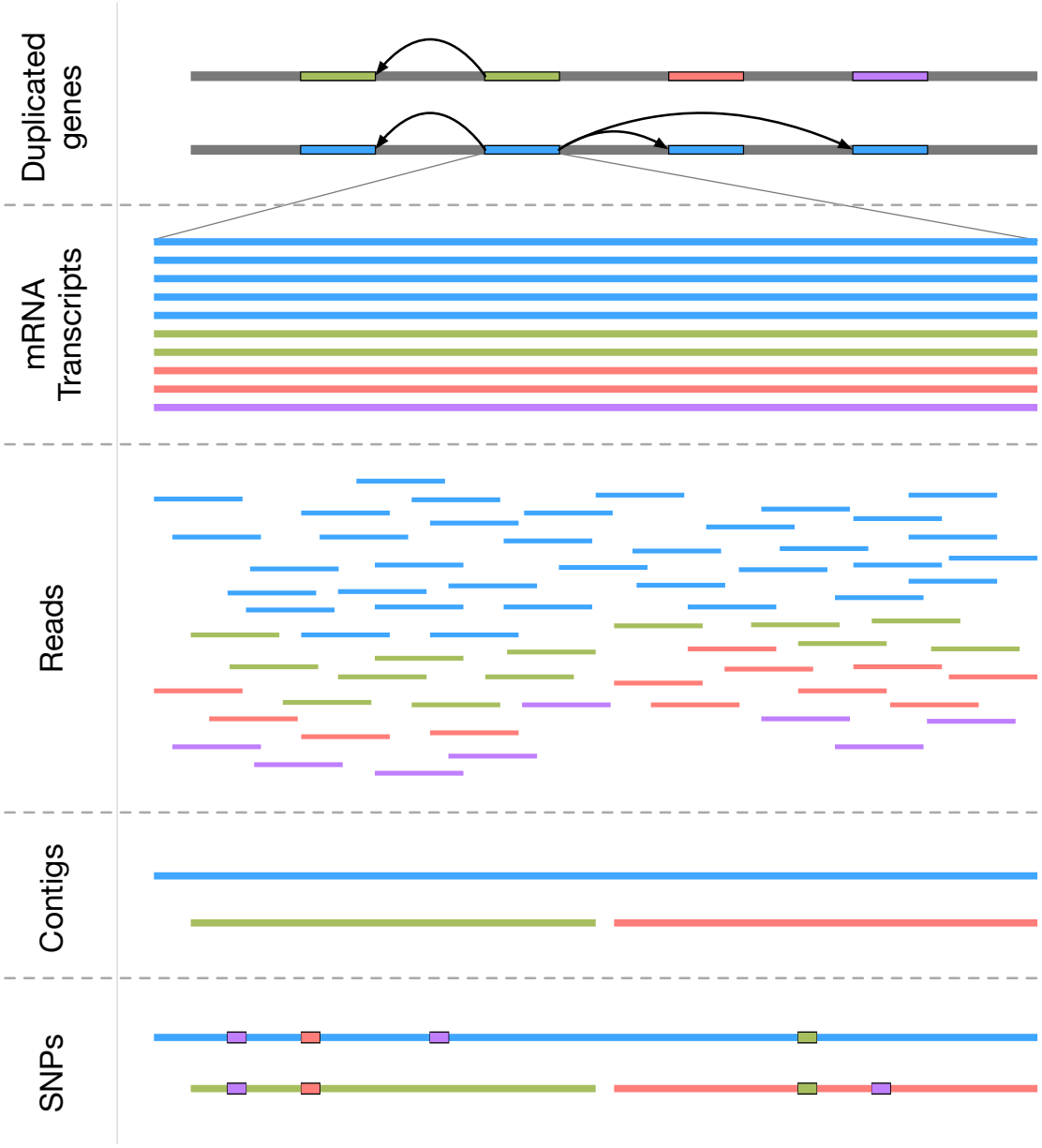


Figure 5 - Identifying consensus sequences from hidden paralogs

This five-part cartoon illustrates the detection of unassembled variants from a group of conserved paralogs using only short-read RNA-Seq data. The first box illustrates a couple of genome segments containing eight different copies of a single gene that has undergone multiple rounds of duplication. Colors indicate relationships from the most recent round of duplication, illustrated with arrows. The second box illustrates the relative abundances of the transcripts within the extracted RNA. The degree to which dinoflagellates regulate transcript abundances is still largely unknown, but their relative abundances should be roughly reflected in the Illumina reads, as illustrated in the third box. The fourth box illustrates a set of Trinity contigs that partially reconstruct some of the transcript sequences, favoring those with deeper sequencing coverage. The final box shows the presence of the unassembled variants being detected based on the presence of many SNPs.

To investigate this hypothesis, the reads from this study were mapped back to their corresponding assemblies and inspected for signs of unassembled variation. Figure 5 shows screenshots taken from the Integrative Genomics Viewer (IGV) (Robinson et al., 2011; Thorvaldsdottir, Robinson, & Mesirov, 2013) for the two full-length actin transcripts from the *A. carterae* assembly. Positions where at least 5% of the reads support at least one variant base are colored by the read support for each base. The histograms have been aligned by their respective coding sequences, revealing imperfectly overlapping patterns of variation. Overlapping patterns of variation occur when reads from unassembled transcripts map equally well to both Trinity isoforms, resulting in the unassembled reads being split between the contigs.

A strong bias for the 3rd codon position was observed within the coding regions of both sequences (Figure 6 inset). In total, the upper and lower sequences respectively contain 66 & 69 variant positions within their coding regions, of which only five from each did not occur in the 3rd position of their respective codons. These five SNPs were instead found in the 1st positions of codons 28, 95, 261, 293, & 346 in both coding sequences, where they synonymously mutated either arginine or leucine codons. Such strong biases are consistent with the purifying selection that one would expect for real variation within coding sequences.

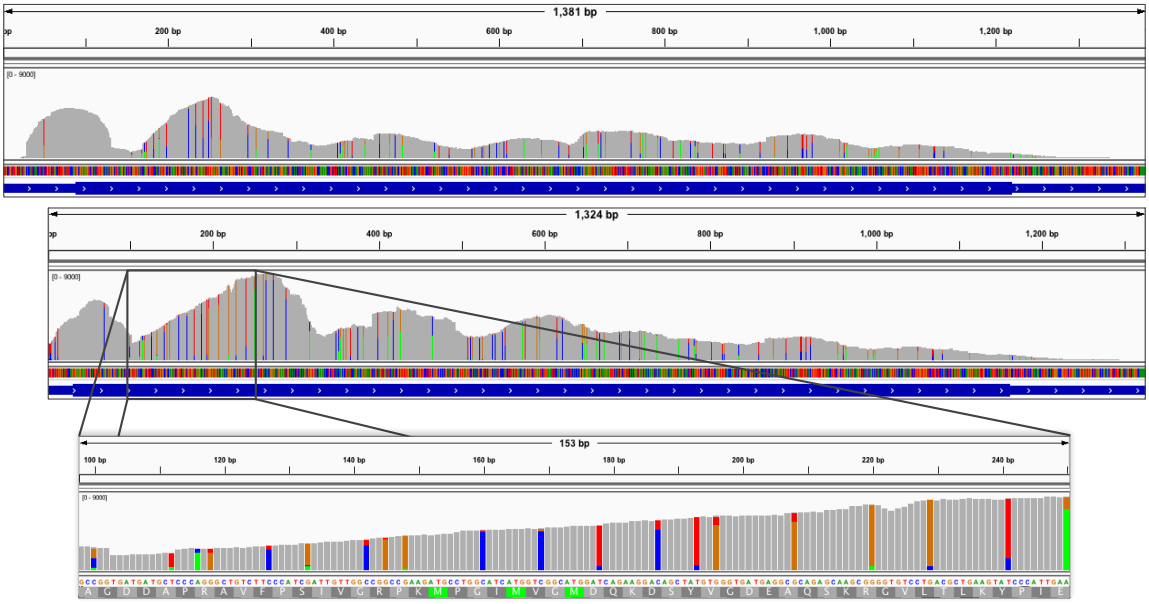


Figure 6 - Sequencing coverage and SNP distribution of full-length actin isoform sequences in the *A. carterae* assembly

Read coverage, ranging from 0-9k, for the two Trinity isoforms from the *A. carterae* assembly that contained complete coding units of actin. Colors indicate relative read support for each base at positions where more than 5% of the reads support at least one variant (green=A, red=T, blue=C, orange=G). The histogram images have been aligned by the coding regions of the sequences (thick portions of blue bars underneath each histogram) to show regions of high identity where the reads from unassembled variants appear to have been split between the sequences, resulting in shared patterns of variation. Inset shows the strong bias observed for SNPs to occur in the third codon position. [Images were modified from Bowtie2 read mappings viewed in IGV]

The observed SNP densities were even higher for the Trinity isoforms containing complete coding units for form II rubisco from the *P. minimum* assembly (Figure 7), with an average of 133 SNPs per kilobase (SPK) of assembled sequence, compared to 63 for the two nearly complete actin transcripts from *A. carterae*. It is tempting to speculate that higher SNP density is indicative of higher copy number, and there may indeed be some correlation, but not all genes, or even positions within individual genes, evolve under the same evolutionary constraints. Attempting to estimate gene copy number from SNP density is therefore unlikely to be very precise or reliable on a genomic scale. In this case, the exceptionally high SNP density of the form II rubisco genes in *P. minimum* could be a consequence of the two-dimensional collapse of the coding units.

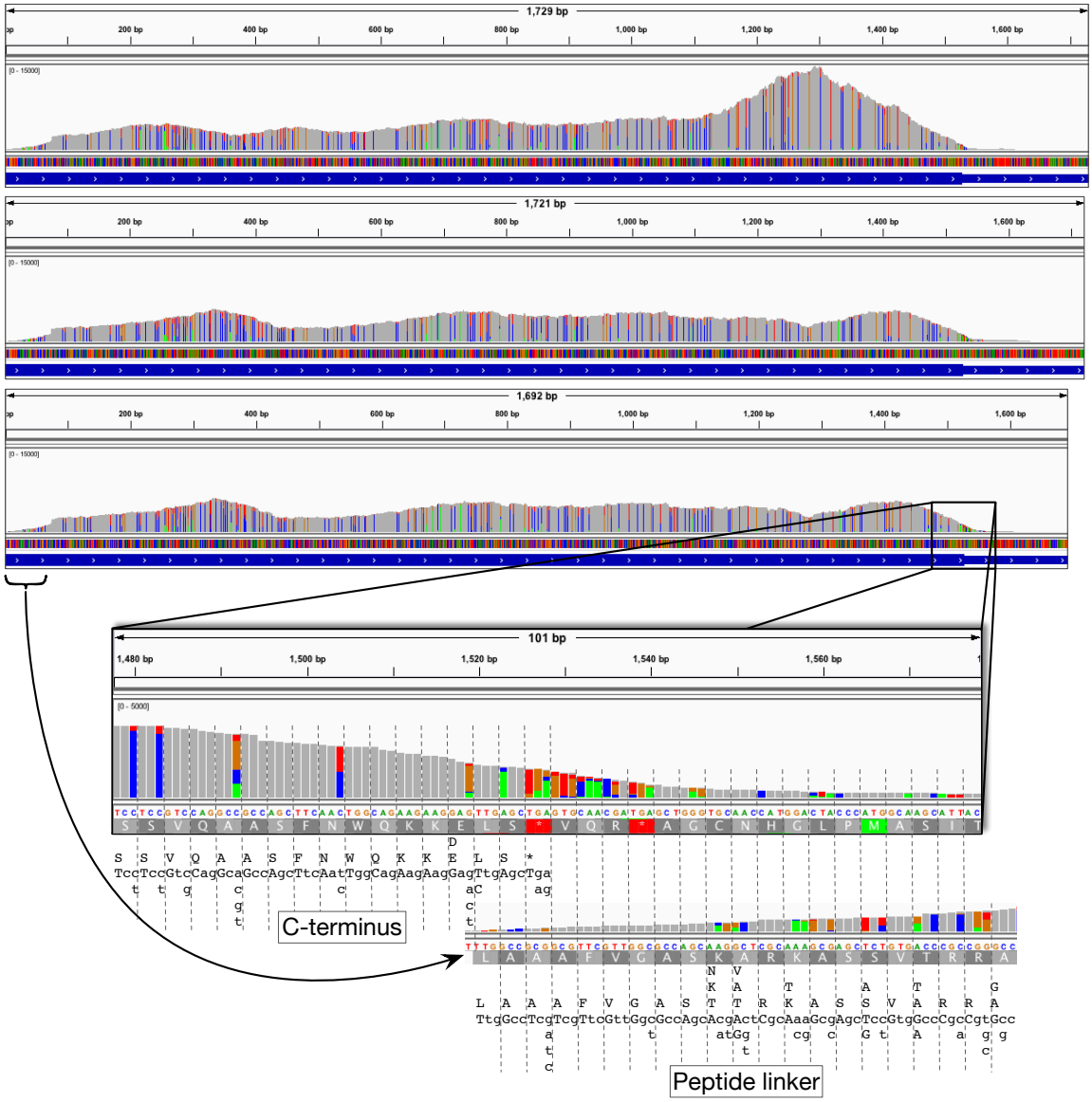


Figure 7 - Sequencing coverage and SNP distribution of *P. minimum* contigs containing complete coding units for form II rubisco

Read coverage, ranging from 0-15k, for the three Trinity isoforms from the *P. minimum* assembly that contained complete coding units of form II rubisco. Colors indicate relative read support for each base at positions where more than 5% of the reads support at least one variant (green=A, red=T, blue=C, orange=G). The coverage maps were aligned by the coding regions of the sequences, which revealed regions of high identity where reads from unassembled variants appear to have been split, resulting in shared patterns of variation. Inset shows agreement with of patterns of variation reported by Zhang & Lin (2003), who also found that transcriptional units encoding form II rubisco in *P. minimum* tend to contain four coding units separated by 21-codon linker sequences. They found that only the terminal CU contains a stop, which is preceded by two extra codons for lysine and serine not observed in the non-terminal CUs. Inset also shows spike in variation immediately following the stop, which was observed in all three sequences. It was suspected that the variation might match the linker sequence observed at the 5' ends of each contig (aligned underneath), but a convincing correlation could not be identified, possibly because any signal was overwhelmed by variability in the 3' UTRs observed in both this assembly and by Zhang & Lin. [Images were modified from Bowtie2 read mappings viewed in IGV]

Little is known about how the putative polyprotein is processed into functional form II rubisco. It's possible that coding units in different positions along each transcriptional unit are under different evolutionary constraints. If Trinity did indeed collapse not only dozens of transcriptional units, but also all non-terminal coding units together into these three consensus sequences matching the structure of the terminal CU, then the aligned SNP profiles could complement each other to cover an exceptionally large number of positions. In this case, the 3rd positions of roughly a third of all codons contained at least one well-supported variant.

A conspicuous spike in variation was observed within and immediately following the stop positions. To determine if this was the result of competing signals between the linker sequence into which all non-terminal coding units should extend and the terminal LS* and 3' UTRs found in all three Trinity isoforms, the corresponding regions of the histograms were aligned (Fig. 6 inset). No apparent correlation between the linker sequence and the variant bases could be identified. Instead, it appears that the linker sequence was not similar enough for the corresponding reads to even map to this region, and that the spike in variation is better explained by the diversity of stops and 3' UTRs reported in Figure 3 of (H. Zhang & Lin, 2003). Reads from the non-terminal coding units not mapping into the 3' UTR could also explain why the coverage rapidly falls to a relatively low plateau after the stop. Likewise, the sudden jump in coverage just past the 5' linker sequences of the Trinity isoforms could be explained by reads from within the leading coding unit mapping to the assembled coding unit, with the exception of the reads that extended into the 5' UTRs.

The large amount of variation observed within the raw sequence data associated with the most completely assembled contigs from both actin and form II rubisco provides strong evidence for both families having been collapsed into small numbers of consensus sequences during assembly. SNP density for the six actin fragments ranged from 9 to 31 SNPs per kilobase. The reads mapping to the long, divergent fragment form II rubisco were essentially uniform, however, highlighting the need to consider SNP density alongside traditional clustering results. Whether the sequence is a divergent paralog or a contaminant sequence, traditional clustering based on alignment of translated coding sequences is the only way to learn that it contains an ORF that conspicuously encodes 343 amino acids that are identical to the corresponding region of a protein encoded by a family of highly conserved paralogs.

2.2.5 Combining complementary paralogy detection methods

To scale up the investigation to the rest of the assembled sequences, a custom SNP caller was written and used (see Methods for details). As illustrated in Figure 5, reads from a single unassembled variant can potentially map to multiple contigs, which can lead to the same SNP being called on each of these contigs. Ambiguously mapping reads are typically undesirable for genotyping and tend to be filtered out by standard SNP calling software. In the scenario depicted in Figure 5, however, it is correct to say that all three contigs are associated with a family of underrepresented conserved paralogs, and so for this purpose it is desirable to call the same SNP within any sequence that is homologous to the region in which the SNP occurs.

The custom SNP caller written for this application ensures the desired behavior by considering only the per-base quality and the coverage, eschewing many of the more

complex filters incorporated into popular modern SNP-calling software. This raised concerns that, in an effort to correct the dramatic false negative problem encountered with standard genotyping software (data not shown), the relatively simple SNP-calling procedure used here would introduce an equally serious false positive problem. Encouragingly, the SNPs agreed very well with the apparently real variation observed in the two cases detailed above. To evaluate its overall performance on the rest of the sequences, the SNPs that fell within TransDecoder coding sequences were counted by codon position (Figure 8, Table S5).

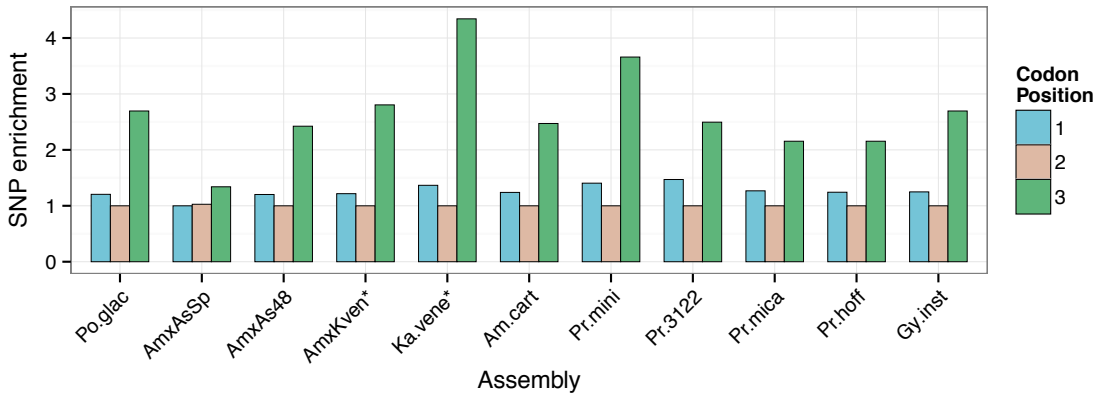


Figure 8 - SNP enrichment by codon position in the custom SNP-calling procedure

Enrichment of SNP counts by codon position for SNPs occurring within TransDecoder-identified coding sequences. SNP counts are normalized by the count for the position with the lowest count, which in all cases was the second position except the dinospores-enriched *Amoebophrya* sp. ex *A. sanguinea* assembly, for which the first position had a slightly lower count.
(Note: *Amoebophrya* sp. ex *K. veneficum* and its host were not sequenced and assembled separately. Instead, sequences from the combined assembly were separated based on average per-gene GC% (see Methods))

Due to the non-random degeneracy of the genetic code (Osawa, 1995), mutations in the 3rd codon position are much more likely to be synonymous than mutations in either the 1st or 2nd. Between the first two positions, mutations in the 1st position are slightly more likely to be synonymous, although with redundancy nowhere near that of the 3rd

position. The SNP distribution ratio across all free-living dinoflagellates included in this study was 1.3 : 1 : 2.5, which reassuringly agrees with the anticipated biases. The only assembly for which this pattern did not hold was from the parasite-enriched *Amoebophrya* sp. ex *A. sanguinea* library. Both *Amoebophrya* assemblies contained lower numbers of TransDecoder-identified CDSs and lower fractions of functionally annotated translated coding sequences. The consistently aberrant results from *Amoebophrya* may reflect unique genomic structure relative to the Dinophyceae, presumably owing to their parasitic lifestyle. For all free-living dinoflagellates, the results provide strong support not only for the SNPs, but also for the TransDecoder CDS annotations.

Figure 9 shows the differing gene duplication pictures provided by SNP density compared to traditional paralogy detection based solely on assembled sequences, and that the collapse of conserved paralogs is a widespread problem in dinoflagellate transcriptome assembly from short-read HTS. Actin and form II rubisco sequences have been highlighted to demonstrate the value of using a combined approach for paralogy detection in dinoflagellate transcriptomes. In both cases, MCL correctly identified all assembled copies of each gene, but the assembler had under-reported the actual number and diversity of paralogs, leading to very low levels of inferred paralogy along the MCL axis. By also considering the SNP density, it is obvious that most of these sequences are related to one or more unassembled variants. A more complete picture of paralogy is therefore obtained by combining the SNP density information with clustering results based on pairwise sequence alignments.

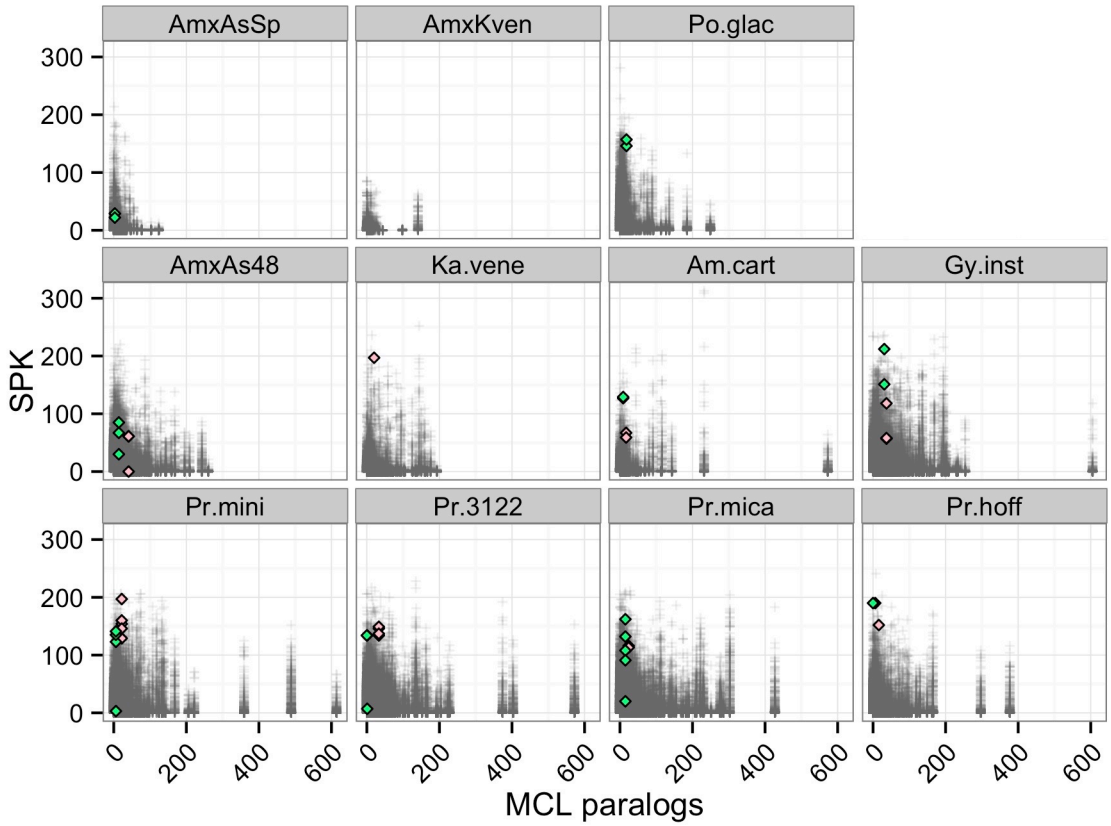


Figure 9 - SNP density combined with traditional clustering offers a more complete description of paralogy

SNPs per Kilobase (SPK) reveal hidden paralogy that is undetectable using traditional clustering methods. Small grey +'s represent translated TransDecoder ORFs, plotted according to two different methods for inferring the presence of paralogs. The x-axis indicates the number of intra-organism sequences that were co-clustered using a popular clustering method based on BLASTp and MCL. The y-axis shows SNP densities, reported as SPK. Green diamonds indicate the positions of sequences with ≥98% identity to one or more actin protein references sequences over at least 300 aa of aligned sequence. Lightish red diamonds indicate the positions of sequences with ≥90% identity to one or more form II rubisco protein references sequences over at least 200 aa of aligned sequence. These cutoffs were selected to include at least one reference sequence from as many organisms as possible without including sequences that were known to be spurious respectively to actin with *A. carterae* or form II rubisco within *P. minimum*.
(Note: *Amoebophrya* sp. ex *K. veneficum* and its host were not sequenced and assembled separately. Instead, sequences from the combined assembly were separated based on average per-gene GC% (see Methods))

Corresponding sequences from the other assemblies were also annotated by filtering tBLASTn results with fixed thresholds that included the largest number of mostly complete sequences from each assembly, without introducing spurious hits for either actin in *A. carterae* or form II rubisco in *P. minimum*. In most cases, the same pattern of

low MCL paralogy and high SNP density is observed, although a few sequences, like the questionable partial RuBisCO sequence from the *P. minimum* assembly, had both low MCL paralogy and low SNP density.

The overall distributions from both methods decay exponentially with similar relative rates, indicating that many sequences belong to either divergent families of paralogs that Trinity assembled into its constituent sequences, or conserved families of paralogs that Trinity collapsed into a small number of consensus sequences, but that relatively few sequences belong to families of paralogs from which Trinity was able to assemble many full-length consensus sequences that each represent unassembled sub-families of conserved paralogs. Two thirds of the CDS-containing Trinity isoforms had either no high-confidence SNPs or no MCL-detected paralogs, while an overlapping two thirds had either at least one MCL-detected paralog and/or at least a handful of SNPs per kilobase. Over 200k coding isoforms had no SNPs and no MCL-identified paralogs. The average length of these sequences was about 1,000 nt, which is long enough to encompass complete coding regions and both paired-end reads from many sequencing fragments. Despite this, the average estimated transcript abundance for these sequences was about a third lower than for the overall set. While it isn't possible to distinguish between highly expressed genes, and highly duplicated genes that are each expressed in low levels using only RNA-Seq data, it is conceivable that if these sequences are indeed from single-copy genes, that their transcripts really are less abundant, on average, than transcripts from families of highly duplicated genes. These genes therefore provide an excellent starting place for researchers aiming to identify single-copy genes suitable for a variety of applications.

All eleven assemblies for the novel transcriptomes that were sequenced for this study have been modified to include both MCL paralogy and SNP density information within the FASTA headers, and have been uploaded to Figshare (T. Gibbons, 2015). Researchers studying species for which a transcriptome has not yet been sequenced can query these assemblies with sequences of interest to obtain an overview of paralogy for that gene in other dinoflagellates. For those with their own transcriptomic data, the SNP-calling software is available on GitHub, along with the other custom scripts using for my thesis (see 2.5 Availability of supporting data and custom software below).

2.3 Discussion

There has been an explosion of short-read HTS transcriptome data from a wide variety of dinoflagellates in the past few years. Assemblies from these data are conspicuously deficient in some of the hallmark features of dinoflagellate genomes, such as large families of paralogs. Characterization of dinoflagellate genomes will become more precise and more comprehensive as long-read HTS data provide full-length transcript sequences and improved reference genome sequences, but it will takes years for such data to catch up with the wealth of short-read HTS data that is already available. In the meantime, unassembled short reads can be used to augment assembled sequences in order to evaluate long-standing hypotheses about dinoflagellate genes on a genomic scale.

In this study, I have shown that large, conserved, transcriptionally active families of paralogs are abundant in a diverse set of free-living dinoflagellates, contrary to the evidence available from analysis of only the assembled transcript sequences. Combining SNP densities obtained from read mapping with traditional clustering based on pairwise alignment of assembled sequences provides a more complete picture of gene duplication,

without the aid of reference genomes or improved sequencing technologies. The SNP profiles of individual genes provide valuable information about the evolutionary constraints under which they are evolving. Comparing profiles between more divergent paralogs and between homologs from different species could be used to identify exceptionally highly conserved regions that are likely to be important for gene function. Such regions are also useful for primer design and to anchor multiple alignments.

In other cases, it is more convenient to simply avoid duplicated genes. A popular method for characterizing gene function is to somehow inactivate the gene of interest and observe the resulting phenotypic changes. When the target gene is actually a large family of genes, inactivating them all can be especially challenging. Assembled transcripts that lack any SNPs are not guaranteed to be from single-copy genes, although a lack of SNPs does indicate extremely high sequence conservation among any sequences that may have been merged during assembly. The identification of such highly conserved sequences could be used to design sequence-specific protocols such as microRNA in a way that suppressed the products from all copies of a gene.

Phylogenetic analyses are also made easier with the use of single-copy genes because it simplifies the process of identifying orthologous sequences. If suitable single-copy genes spanning all lineages cannot be identified, comparison of SNP profiles could potentially be used to infer whether two families of conserved paralogs began expanding before or after speciation. Co-orthologs may not be ideal for phylogenetic analysis, although they do not violate the fundamental assumptions of most methods. Sophisticated investigations of this nature could be used to identify lineage-specific expansion of gene families.

An alternative to SNP-based analyses would of course be to optimize the assemblies in the hopes of successfully assembling the hidden paralogs, however, parameter optimization for assembly of short-read HTS data is extremely expensive in terms of both time and computational resources. With emerging long-read HTS technologies like PacBio's Sequel platform poised to negate the need for transcriptome assembly, that time and money would be better spent normalizing transcript abundances for resequencing. The procedures used in this study rely on open source software, are intuitive, and provide additional valuable information using only data that are already publicly available, allowing the dinoflagellate research community to benefit now from information that seems to have been largely regarded as inaccessible.

2.4 Methods

2.4.1 Cell culture

Cultures of *Amphidinium carterae*, *Gyrodinium instriatum*, *Karlodinium veneficum*, *Polarella glacialis*, *Prorocentrum hoffmannianum*, *Prorocentrum micans*, and *Prorocentrum* sp. CCMP3122 were ordered from the Culture Collection of Marine Phytoplankton (now the National Center for Marine Algae and Microbiota). *Prorocentrum minimum* was isolated from an active algal bloom in Baltimore's Inner Harbor (Cooper et al., 2014). Two strains of *Amoebophrya*, obligate intracellular parasites, were previously isolated from the Chesapeake Bay alongside their respective dinoflagellate hosts, *Akashiwo sanguinea* and *K. veneficum* (Coats & Park, 2002). Cultures of *A. sanguinea* and both *Amoebophrya* strains were donated by the Coats lab for this study. See Table 3 for a summary.

Table 3 - Sample sources

Sample	Source
Polarella glacialis	CCMP1383
Amoebophrya sp. ex *Akashiwo sanguinea**	Coats lab
Amoebophrya sp. ex *Karlodinium veneficum**	Coats lab
Karlodinium veneficum	CCMP1974
Amphidinium carterae	CCMP1314
Prorocentrum minimum	Balt. Bloom
Prorocentrum sp. CCMP3122	CCMP3122
Prorocentrum micans	CCMP1589
Prorocentrum hoffmannianum	CCMP683
Gyrodinium instriatum	CCMP3173

* Co-cultured with host on a near-synchronous infection cycle

All cultures were grown in either 15ppt or 32ppt sterilized L1 medium (minus Na_2SiO_3. (Guillard & Ryther, 1962)). *P. glacialis* was kept at 4ºC on an 8:16 light:dark cycle. The rest were kept at 18ºC on a 16:8 light:dark cycle.

2.4.2 *Amoebophrya* pre-extraction conditions and timing

The two *Amoebophrya* strains were co-cultured on a near-synchronous infection cycle with their respective hosts, following methods described in (Bachvaroff, Place, & Coats, 2009; Coats & Park, 2002). RNA was extracted from one *Amoebophrya sp. ex A. sanguinea* co-culture 48 hours post infection (HPI), a mid-point in the standard infection cycle. The other *Amoebophrya sp. ex A. sanguinea* extraction and both *Amoebophrya sp. ex K. veneficum* extractions were timed within about 6hrs after most of the parasites started releasing dinospores from their host cells, maximizing the relative abundance of the *Amoebophrya* cells. These dinospores were then further enriched by gravity filtration using Nucleopore filters (8-µm for *A. sanguinea*, 5-µm for *K. veneficum*; Whatman, Piscataway, NJ). The filtration proved to be much more effective for *A. sanguinea* than for *K. veneficum*, most likely because the *K. veneficum*

cells are much closer to the pore size (Coats & Park, 2002; Garcés et al., 2006; Menden-Deuer & Lessard, 2000).

2.4.3 *G. instriatum* pre-extraction conditions and timing

Twenty-four hours prior to RNA extraction, three *G. instriatum* subcultures were transferred into different experimental conditions, while three remained in the control conditions described above. One of the experimental cultures was cooled to 10ºC with an 8:16 light:dark cycle, one was warmed to 25ºC with a 14:10 light:dark cycle, and one was placed under a bright BTEX lamp, which remained on for the entire 24hrs. RNA was extracted from all three of these experimental cultures, as well as two of the cultures from the control room, around noon the following day. RNA was not extracted from the final control culture until 3am the following morning in order to simulate night.

2.4.4 RNA extraction

G. instriatum cells were pelleted by centrifugation and frozen in liquid nitrogen, then resuspended in RNAlater. *G. instriatum* and *Prorocentrum* cells were disrupted in a Mini-Beadbeater-1 (Biospec Products, Bartlesville, OK) for one minute. RNA for all libraries was then extracted using either the Nucleospin Plant Total RNA kit (Machery-Nagel, Düren, Germany), or the RNAqueous kit (Ambion, Carlsbad, CA).

2.4.5 Sequencing

RNA from *A. carterae*, *Amoebophrya* sp. ex *A. sanguinea*, *Amoebophrya* sp. ex *K. veneficum*, *P. glacialis*, *P. hoffmannianum*, *P. micans*, and *Prorocentrum* sp. CCMP3122 was shipped to Macrogen (Seoul, Republic of Korea). *P. glacialis* was sequenced for 78 cycles from each "paired" cDNA fragment end on an Illumina GAIIx. The rest were sequenced for 101 cycles from each fragment end on Illumina HiSeq

instruments. The two *Amoebophrya* sp. ex *A. sanguinea* libraries were multiplexed together in one lane, *A. carterae* was multiplexed with both *Amoebophrya* sp. ex *K. veneficum* libraries in a second lane, and the three *Prorocentrum* libraries were multiplexed together in a third lane. *P. glacialis* was sequenced in a lane by itself. The *P. minimum* and all six *G. instriatum* samples were sent to the University of Maryland Institute of Bioscience and Biotechnology Research sequencing center (College Park, Maryland), where they were all sequenced for 101 cycles from each fragment end on an Illumina HiSeq 1000. The six *G. instriatum* libraries were multiplexed together in a single lane. *P. minimum* was multiplexed with green algal RNA, as described in Cooper et al. (2014).

All untrimmed reads have been uploaded to the Short Read Archive under project SRP046737.

2.4.6 Adapter removal

Adapter sequences were removed from the reads using TagCleaner version 0.12 (Schmieder, Lim, Rohwer, & Edwards, 2010). Up to 16 mismatched and/or missing bases were allowed from the 5' end (-mm5 16) for all reads and adapter sequences. This number was chosen after reviewing output from the '-stats' option (data not shown). The adapter sequences specified to be removed from the 5' end (-tag5 <adapter_seq>) varied with sequencing platform and read orientation. Using grep and TagCleaner's '-stats' option, it was observed that the forward reads of all libraries sequenced on Illumina HiSeq machines contained significant numbers of complete or partial TruSeq barcode sequences

(GATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCC

GTCTTCTGCTTG). *Polarella* was sequenced on an Illumina GAIIX and the forward reads instead contained large numbers of the reverse-complemented Illumina paired-end PCR primer (v2.0) sequence (AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG). All reverse reads appeared contaminated with the reverse complement of the same universal adapter sequence (AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT), regardless of sequencing platform.

2.4.7 Quality trimming

Low-quality bases were removed using PRINSEQ-lite version 0.20.2 (Schmieder & Edwards, 2011). Forward and reverse reads were processed in pairs using both the '-fastq <forward.fastq>' and '-fastq2 <reverse.fastq>' options. The FASTQ files for *A. carterae* and *P. glacialis* libraries, as well as all five pairs of FASTQ files for the various *Amoebophrya* cultures, were all generated using a version of CASAVA that predated 1.8, and therefore required the '-phred64' flag to adjust the quality values into the correct range. The other read files were output by CASAVA v1.8 or later. PRINSEQ allows the specification of a tiered trimming scheme. Each end of each read was first trimmed until 10 consecutive base calls with quality scores of 20 or better were encountered ('-trim_qual_left 20 -trim_quality_right 20 -trim_qual_window 10'). Any remaining reads shorter than 25 base pairs ('-min_len 25') or with average quality scores below 20 ('-min_qual_mean 20') were then removed.

2.4.8 *De novo* transcript assembly

Reads were assembled using Trinity version r20140413p1 (Grabherr et al., 2011) with 16 CPU threads ('--CPU 16'). Paired-end reads that passes quality control (QC) were specified in pairs with the '--left <forward.fastq' and '--right <reverse.fastq>' options. Orphaned reads passing QC were combined into a single file and provided with the '--single <orphans.fastq>' flag. Both *Amoebophrya* sp. ex. *K. veneficum* libraries were assembled together, as were all six *G. instriatum* libraries. For the combined *G. instriatum* assembly, the '--normalize_reads' option was specified to accommodate the exceptionally deep coverage.

Contigs from the combined *Amoebophrya* sp. ex. *K. veneficum* were aligned against a reassembly of *Karlodinium micrum* reads released by the Moore Foundation (Bentlage, 2014; Grabherr et al., 2011; Keeling et al., 2014) using BLASTn version 2.2.29+ (Camacho et al., 2009) with an E-value threshold of 1e-5 ('-evalue 1e-5'). Based on these results (Figure 10), the contigs were sorted into separate host and parasite files by GC content using a custom Python program splitFastaByGC.py.

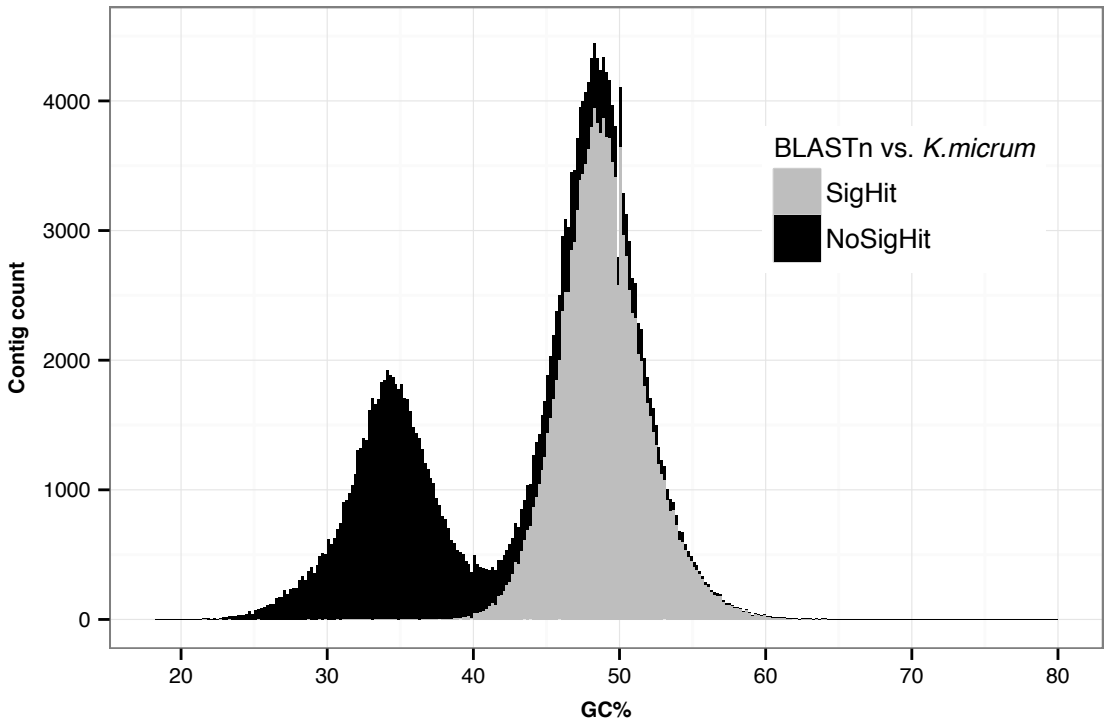


Figure 10 - Bimodal GC% distribution of combined *Amoebophrya* sp. ex *K. veneficum* assembly

Histogram of contigs from the *Amoebophrya* sp. ex *Karlodinium veneficum* combined assembly by GC content, showing whether or not they had at least one significant (E-value ≤ 1e-5) BLASTn alignment to a contig in a reassembly of *Karlodinium micrum* data released by the Moore Foundation. Bin widths on the histogram are 0.2%.

A set of short, non-standard abbreviations is used throughout this manuscript to identify each assembly in the various tables and figures (Table 4). *Prorocentrum hoffmannianum*, *micans*, *minimum* and unnamed strain CCMP3122 have been respectively abbreviated Pr.hoff, Pr.mica, Pr.mini, and Pr.3122. *Polarella glacialis*, *Karlodinium veneficum*, *Amphidinium carterae*, and *Gyrodinium instriatum* have been respectively abbreviated Po.glac, Ka.vene, Am.cart, and Gy.inst (see below for further information about the *K. veneficum assembly*). The two *Amoebophrya* sp. ex *Akashiwo sanguinea* assemblies were named based on the extraction conditions, with AmxAsSp indicating the *Amoebophrya*

dinospores, and AmxAs48 indicating the 48 HPI (hours post infection) assembly. The single *Amoebophrya* sp. ex *K. veneficum* assembly was simply abbreviated AmxKven.

Table 4 - Assembly abbreviations

Assembly	Abbreviation
Polarella glacialis	Po.glac
Amoebophrya sp. ex Akashiwo sanguinea (Dinospores)	AmxAsSp
Amoebophrya sp. ex Akashiwo sanguinea (48 HPI)	AmxAs48
*Amoebophrya sp. ex Karlodinium veneficum**	AmxKven
*Karlodinium veneficum**	Ka.vene
Amphidinium carterae	Am.cart
Prorocentrum minimum	Pr.mini
Prorocentrum sp. CCMP3122	Pr.3122
Prorocentrum micans	Pr.mica
Prorocentrum hoffmannianum	Pr.hoff
Gyrodinium instriatum	Gy.inst

*Amoebophrya sp. ex K. veneficum and its host were not sequenced and assembled separately. Instead, sequences from the combined assembly were separated based on average per-gene GC% (see below, S1 Fig)

2.4.9 CDS identification and annotation

Coding sequences (CDSs) were identified within the contigs using the TransDecoder program distributed with Trinity version v20140417 (Haas et al., 2013), which uses length, hexamer composition, and HMMer-annotated Pfam-A domains to score and filter ORFs. TransDecoder was run with 16 CPU threads ('--MPI --CPU 16'). HMMer version 3.0 (Eddy, 2011) was installed and made available to TransDecoder along with the Pfam-A HMM database (v27.0) (Finn et al., 2014) with the '--search_pfam' option. Using this option, TransDecoder also provided Pfam annotations for the translated protein sequences in HMMer domtblout format.

Gene Ontology (GO) categories were assigned to ORFs with a custom Python program pfam2go.py that parses both a HMMer domtblout '.dat' output file produced by TransDecoder and the Pfam2GO mapping released by InterPro (Hunter et al., 2009).

2.4.10 Clustering

The translated CDSs from all eleven assemblies were combined into a single database containing over a million sequences and aligned against itself using BLASTp (v2.2.29+) (Camacho et al., 2009) with an E-value threshold of 1e-5 (‘-evalue 1e-5'), filtering low complexity sequences only in the lookup table (‘-softmasking true'), increasing the limit on the number of target sequences (‘-max_target_seqs 10000'), and printing in a modified tab-delimited format (‘-outfmt "6 std qlen slen positive"'). An MCL-readable ‘abc' graph file was created from the tab-delimited BLASTp results using Porthos (T. R. Gibbons, Mount, Cooper, & Delwiche, 2015), combining any non-overlapping hits (‘--combine') between each pair of sequences and requiring the sum of identical amino acids to be at least 50% of the shorter sequence length (‘--min_cov 0.5'). The database was clustered using MCL (Enright, Van Dongen, & Ouzounis, 2002; van Dongen, 2000) with eight threads (‘-te 8') and an inflation parameter of 1.1 (‘-I 1.1').

2.4.11 Transcript quantification

Transcript abundances were estimated using Salmon (v0.4.2) (Patro, Duggal, & Kingsford, 2015). salmon index was called with default parameters ‘-t FASTA -i INDEX’. A simple Python script transcript_to_gene_map.py was used to extract isoform to gene mappings from the Trinity FASTA headers so that the transcripts could be quantified both by Trinity isoform and by Trinity gene. A wrapper script called salmon_quant.py was used to run salmon quant with the parameters ‘--index <idx_dir> --libType IU --unmatedReads <orphans.fasta> --mates1 <forward.fastq> --mates2 <reverse.fastq> --threads 8 --extraSensitive --output <out_dir> --geneMap <gene.map> --mappingCacheMemoryLimit 1600000’.

2.4.12 Read mapping

Each library of trimmed Illumina RNA-Seq reads was individually mapped back to the Trinity assembly to which it contributed using Bowtie2 (v2.2.4) (Langmead & Salzberg, 2012) in full-length "very sensitive" mode ('--very-sensitive' is equivalent to '-D 20 -R 3 -N 0 -L 20 -i S,1,0.50'), allowing insert sizes up to 1,000 base pairs ('--maxins 1000'), using eight threads ('--threads 8'), and providing not only the paired-end reads, but also the singleton reads that were orphaned during quality control. Without the '-k' or '-a', Bowtie2 reports only the single best alignment for each read (pair). In cases where a read (pair) maps multiple places equally well, Bowtie2 reports one of the best alignments and reduces the mapping quality to zero.

Reads from *A. carterae*, *P. glacialis*, and the four libraries from the *Amoebophrya* host-parasite systems were all generated before the CASAVA 1.8 update and required the '--phred64' flag. The rest were run with the '--phred33' flag. Each of the resulting Sequence Alignment/Map (SAM) formatted files was compressed into binary BAM files using SAMtools view (v1.2) (H. Li, 2011; H. Li et al., 2009) with eight threads ('-@ 8'), then sorted using SAMtools sort (v1.2) with eight threads ('-@ 8') and the best available level of compression ('-l 9').

2.4.13 Variant calling and filtering

The SAMtools suite suffered from a serious false-negative problem stemming from apparent assumptions that multiply-mapping reads are fundamentally unreliable for SNP calling. This is a responsible assumption for identifying genomic haplotypes, but was inappropriate for this application. A Python program snps_from_mpileup.py was therefore written to implement a relatively simple custom SNP calling procedure that operates on mpileup files.

Each Trinity assembly was indexed using samtools faidx (v1.2) with default parameters and combined with the sorted BAM files to create mpileup files using samtools mpileup (v1.2) with the '--fasta-ref FASTA' option. It was not necessary to specify Phred versions because Bowtie2 converted all of the quality scores to Phred+33. The mpileup files provide, for each position in each reference, the reference base, the coverage, and a stack of bases with corresponding Phred qualities, extracted from the mapped reads that overlap that position. The snps_from_mpileup.py program identified SNPs by analyzing each stack of bases and base qualities using the following procedure: For each position, the total coverage was noted before removal of all bases with Phred quality values below 20. Of the remaining bases, the greater of either two bases or 5% of the total coverage were required to support the reference base. After that, all variants supported by the greater of either two bases or 5% of the total coverage were reported as SNPs.

2.5 Availability of supporting data and custom software

All custom software used in this study has been uploaded to a dedicated GitHub repository: https://github.com/trgibbons/DissertationScripts. Assemblies and other supplemental files are available on Figshare (T. Gibbons, 2015).

2.6 Contributions

AToL: An Integrated Approach to the Phylogeny of Dinoflagellates grant (NSF award 0629624) was awarded to Professor Charles F. Delwiche (CFD), who authored the original project proposal, and the direction of the project evolved through many discussions with Professors Tsvetan R. Bachvaroff (TRB), CFD, and Stephen M. Mount (SMM), Doctors Gregory T. Concepcion, Endymion D. Cooper (EDC), and Bastian Bentlage, as well as Gregory S. Mendez (GSM) and Travis S. Rogers (TSR). Organisms

were selected and obtained by CFD, TRB, GTC, and GSM. Cultures were maintained and RNA was extracted by TRB, GTC, EDC, GSM, TSR, and Jacob A. Hyman. All software was selected, evaluated, and executed by Theodore R. Gibbons (TRG) except the CEGMA pipeline, which was run by GSM. All custom software was written by TRG.

2.7 Acknowledgements

We would like to thank the staff at NCGAS for providing not only computational resources, but also friendly and responsive tech support (NSF Award #1062432 - *ABI Development: National Center for Genome Analysis Support*). Additional funding was provided by NSF awards 1036506 and 1046075 to CFD. TRG would also like to thank Computomics and the members of Professor Daniel Huson's lab at the University of Tübingen for hosting him and providing a productive work environment while this project was completed.

3 Estimating genome-scale patterns of trans-splicing in eleven species of dinoflagellates

3.1 Background

Dinoflagellates are one of a handful of disparate eukaryotic lineages that trans-splice short spliced leader (SL) RNA mini-exons onto the 5' ends of their mRNA transcripts during pre-mRNA processing. In all cases, the dinoflagellate trans-splicing spliced leader (DinoSL) sequence has been described as a 22-nt sequence that varies only at the first position (5'-NCCGTAGCCATTTTGGCTCAAG) (Lidie & van Dolah, 2007; H. Zhang & Lin, 2009; H. Zhang et al., 2007). SL trans-splicing was first discovered in *Trypanosoma brucei* over 30 years ago (Boothroyd & Cross, 1982; Van der Ploeg et al., 1982) and has been extensively studied in trypanosomes, where it contributes a 5' hypermethylated cap to each mature mRNA as it is spliced from a polycistronic pre-mRNA [for reviews, see Günzl (2010) & Hastings (2005)]. The discovery of SL trans-splicing in dinoflagellates is relatively recent and the process is relatively poorly understood.

Like trypanosomes, dinoflagellates are distantly diverged from the model plants and opisthokonts whose genomes have been most thoroughly studied. The discovery of such a rare feature within both groups inspired many to wonder if trypanosomes might offer a better template for dinoflagellate genomics. In particular, it has been hypothesized that dinoflagellate tandem array genes might be expressed into polycistronic transcripts, which are then trans-spliced as they are cleaved apart in a process similar to what has been characterized in trypanosomes (Preußer et al., 2012; H. Zhang et al., 2007). While this hypothesis remains unconfirmed (Beauchemin et al., 2012), the discovery of transcripts encoding polyproteins (Rowan et al., 1996; Shi et al., 2013; H. Zhang & Lin,

2003), and a scarcity of strand-switch regions between genes in the *Symbiodinium minutum* genome (Beauchemin et al., 2012), similar to what has been observed in trypanosomes, continues to motivate researchers to seek further similarities between dinoflagellate and trypanosome genomics, as well as links between trans-splicing and dinoflagellate tandem-array genes.

Shortly after the discovery and characterization of the DinoSL, degraded DinoSLs were found embedded within the 5' UTRs of hundreds of dinoflagellate transcripts (Slamovits & Keeling, 2008). Not only were these internal SLs degraded, but many were missing the first seven nucleotides as a result of additional spliced leaders evidently using an internal AG at positions six and seven as a splice acceptor site. In many cases, several SL suffixes were chained together, with each subsequent suffix being more degraded than the one before it until, by the fourth copy, few were clearly recognizable. Using targeted sequencing of the corresponding genes, they showed that the degraded SLs were relict copies from mature mRNA transcripts that had evidently been reinserted back into the genome, rather than the result of repeated splicing of the same transcript. The mechanism of gene duplication within dinoflagellates is unknown, although the repeated discovery of paralogs arranged into tandem genomic arrays, combined with evidence of transcript recycling, has intriguing implications for the existence of a lineage-specific mechanism of gene duplication through tandem transcript recycling.

Further investigation of this phenomenon, as well as of dinoflagellate genomics in general, would greatly benefit from confirmation that the trans-splicing is ubiquitous within dinoflagellates, or at least that the vast majority of mature transcripts have been trans-spliced. A conserved dinoflagellate-specific sequence at the 5' ends of all

transcripts would provide an extraordinarily useful primer target for not only obtaining full-length dinoflagellate transcript sequences, but for selectively amplifying them from more natural environments containing both bacteria and other eukaryotes.

Multiple attempts have been made to determine whether all dinoflagellate transcripts are trans-spliced. Using DinoSL primers, the Lin lab has demonstrated that trans-splicing is widespread in terms of both dinoflagellate species diversity and gene function (Senjie Lin, Zhang, Zhuang, Tran, & Gill, 2010; H. Zhang, Campbell, Sturm, & Lin, 2009; H. Zhang et al., 2013). While useful, confirmation of ubiquitous trans-splicing will require analysis of complete transcriptomes that were not amplified with DinoSL primers. Many such transcriptomes have recently been sequenced using random hexamer priming (Beauchemin et al., 2012; Cooper et al., 2014; Keeling et al., 2014; Meyer et al., 2015; Ryan et al., 2014; H. Zhang et al., 2013; S. Zhang et al., 2014; Y. Zhang et al., 2014). The groups that reported numbers of trans-spliced transcripts, provided counts that were shockingly low.

I show here that the absence of DinoSLs within *de novo* transcriptomes assembled from short-read high-throughput sequencing (HTS) data is the result of incomplete coverage of the 5' ends. Leveraging the statistical power available from these high-throughput experiments, I show that the vast majority of 5'-anchored DinoSL suffixes of at least six nucleotides are real, and that they provide a more accurate count of trans-spliced transcripts within dinoflagellate transcriptomes. The presence of DinoSL suffixes is correlated with other metrics for assembly quality, making them useful sequence-specific markers for assessing the completeness of transcript sequences. Ubiquity could not be confirmed, but improved counting shows that the DinoSL is abundant within the

transcriptomes of a diverse set of dinoflagellates. Internal and serial SLs were also found within all assemblies, expanding the diversity of dinoflagellates suspected of recycling mature transcripts back into their genomes and supporting a hypothesis that transcript recycling is a common mechanism for gene family expansion within the dinoflagellates.

3.2 Results and Conclusions

3.2.1 5' DinoSL abundances

Over a million and a half assembled transcript sequences from eleven species of dinoflagellates were analyzed for this study. Among these, only 242 were found with full-length 22-nucleotide DinoSLs anchored to their 5' ends (Table 5). Nearly three times as many transcripts began with the last 21 nucleotides of the DinoSL, and the numbers rapidly grew as the DinoSL suffix was shortened even further (Table S6). This suggested that many contigs extended near to, but did not reach, the 5' ends of the corresponding transcripts.

Table 5 - DinoSL abundances

	Trinity		DinoSL abundances		
Assembly	Isoforms	Genes	22 nt	21 nt	99%
Po.glac	138,262	100,326	10	17	349
AmxAsSp	63,010	51,895	33	85	4,199
AmxAs48	204,082	173,753	50	89	44,159
AmxKven*	65,001	58,520	13	36	4,212
Ka.vene*	153,056	139,051	20	52	41,173
Am.cart	67,161	57,183	14	92	24,133
Pr.mini	157,617	126,400	8	38	18,733
Pr.3122	168,400	139,299	23	43	24,802
Pr.mica	182,219	131,539	17	50	34,216
Pr.hoff	90,791	70,051	16	53	32,160
Gy.inst	265,280	184,025	38	101	36,235
Totals:	1,554,879	1,232,042	242	656	264,371

*Amoebophrya sp. ex K. veneficum and its host were not sequenced and assembled separately. Instead, sequences from the combined assembly were separated based on average per-gene GC% (see chapter 2)

The chances of a particular 15 or 20 nt sequence occurring randomly are quite low (10^{-9}-10^{-12}), but the DinoSL suffix abundances don't even break one thousand sequences per assembly until the suffix is shortened to 13 nt. The concern of course, is that at some point, the suffix becomes so short that the DinoSL signal is indistinguishable from background noise. Fortunately, the hypothesis in question is an extreme case of the DinoSL being ubiquitously trans-spliced onto all dinoflagellate transcripts, or at least the vast majority. If this hypothesis is true, and if large fractions of the assembled contigs extend very near to the 5' ends of the corresponding transcripts, then the DinoSL suffix sequences should be highly enriched. 5' enrichment of DinoSL suffixes was evaluated two ways.

To evaluate enrichment over other 5'-anchored sequences of equal length, 5'-anchored *k*mers ranging in length from 22 to 4 nucleotides were counted in all of the assemblies and ranked by abundance. Figure 11 summarizes this information by indicating the position of each DinoSL suffix within the corresponding distribution of 5'-anchored *k*mers for each assembly. For all assemblies from 101-bp HiSeq reads, all suffixes of at least 6 nt were either the most abundant, or within the top 1% of most abundant *k*mers. Visual inspection of the top-ranked *k*mers revealed that most of the other sequences within the top percentile were subsequences of the DinoSL of varying lengths. The only assembly for which suffixes of at least 6-nt fell below to top percentile was the *P. glacialis* assembly from 78-bp GAIIx reads. Even in that case, the 6-nt suffixes were still in the 97th percentile, the 7 and 13-nt suffixes were in the 98th percentile, and the rest were all in the 99th percentile.

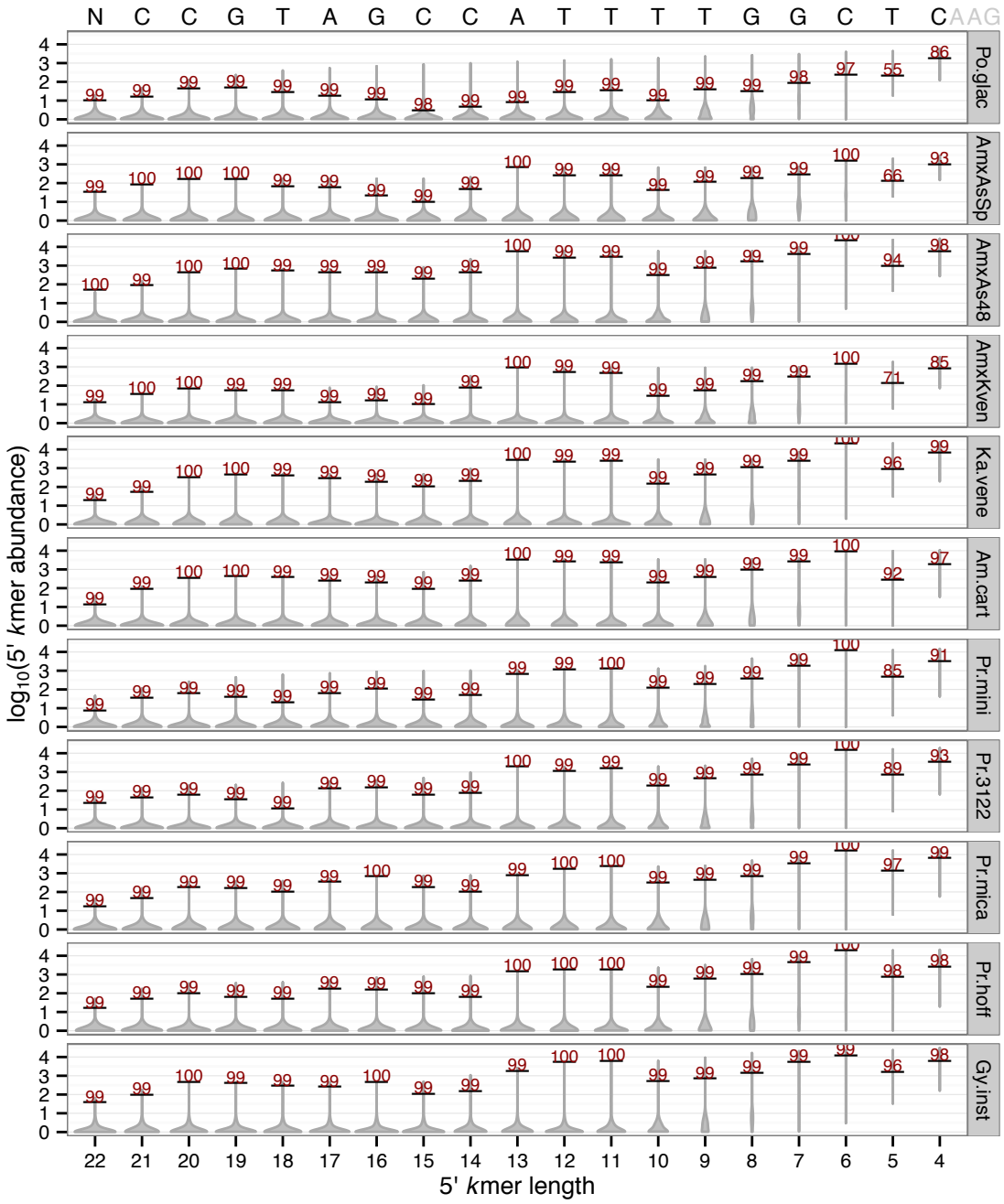


Figure 11 –DinoSL suffix abundances relative to other 5'-anchored *k*mers

The $\log_{10}$ abundances of 5'-anchored *k*mers ranging in length from 22 to 4 nucleotides are shown for each assembly as violin plots. Abundances of the DinoSL suffixes are indicated with perpendicular horizontal lines and labeled by the percentiles into which they fell. The DinoSL sequence is provided across the top and is arranged such that the letter aligned with a particular set of violin plots represents the left-most position of the suffix corresponding to that vertical stack of distributions. Counts for the four variants of the 22-nt DinoSL were combined.

The 5'-anchored abundances of the DinoSL were also compared against expected values estimated from their unanchored abundances (see Methods for details). All suffixes of at least six nucleotides were enriched above expected values and most were enriched over 100 times above their expected values (T. Gibbons, 2015). The least enriched sequence was the 8-nt suffix within *P. glacialis*, which was only 2.7 times as abundant as expected. The most enriched was the 13-nt suffix within *Amoebophrya* sp. ex *K. veneficum*, which was over 800 times more abundant than expected. That particular suffix occurred over 1,400 times within that assembly, although many of the other large enrichment values were inflated by very small abundances of unanchored sequences.

Recounting trans-spliced transcripts using a cutoff of 6 nt brought the numbers up into the tens of thousands for all of the free-living dinoflagellates except *P. glacialis* (Table 5). As discussed in chapter 2, the *P. glacialis* transcriptome was sequenced with shorter GAIIx reads, and the assembly seems to have greatly suffered as a result, making it unlikely that the low *P. glacialis* counts reflect actual differences in the underlying biology.

The lower counts within the *Amoebophrya* were expected due to their smaller genomes, although the lower fractions of detectably trans-spliced transcripts, despite having some of the deepest sequencing coverage of any assembly in this study, could reflect reduced trans-splicing within these parasites. For comparison, the other two assemblies containing fewer than 100k contigs, those corresponding to *A. carterae* and *P. hoffmannianum*, had the two highest fractions of contigs that appear to be trans-spliced among all of the assemblies considered in this study, presumably owing to deeper coverage. Furthermore, the SL-abundance counts within both *Amoebophrya* assemblies have likely been inflated

by contamination from their hosts, which had the largest numbers of detectably trans-spliced sequences among all of the assemblies.

Overall, these results show that the DinoSL is abundant within a diverse set of free-living dinoflagellates, contrary to what has previously been reported, and consistent with the popular hypotheses that the DinoSL is trans-spliced on to the 5' ends of most or all transcripts in more or all dinoflagellates. The observed scarcity of full-length DinoSLs within *de novo* transcriptomes assembled from short-read HTS data is the result of incomplete reconstruction of 5' transcript ends, rather than a reflection of the true underlying biology.

3.2.2 DinoSL suffixes as quality indicators

DinoSL suffix length was not tightly linearly correlated with either transcript length or estimated transcript abundance (measured as TPM or transcripts per million reads mapped, see chapter 2), as revealed by very weak Pearson correlation coefficients of only 0.15 with isoform length and 0.04 with TPM. Other factors, such as sequence-specific biases in the sequencing technology likely also play a role (Aird et al., 2011; K. D. Hansen, Brenner, & Dudoit, 2010), which might explain the relative maxima observed in Figure 11 for 13-11 nucleotide suffixes starting in the AT rich region of bases 10-14, although the peak is also quite close to the recently discovered TTTG promoter that appears to be active within relict SLs (S. Lin et al., 2015).

Despite the lack of a strong direct linear correlation, the detectably trans-spliced sequences were 400 nt longer, on average, and had TPM values that were nearly twice as high, indicating that 5'-anchored suffixes as short as six nucleotides are useful sequence-specific indicators of assembly quality. A correlated implication of this observation is

that sequences without detectable SLs are less complete, providing further evidence that SL trans-splicing is extremely abundant within free-living dinoflagellates.

3.2.3 Internal spliced leaders

While computing the expected abundances of the anchored DinoSL suffixes, it was discovered that full-length DinoSLs were several times more abundant within the contigs than they were at the 5' ends (T. Gibbons, 2015). Figure 12 shows the distribution of conserved 21-nt DinoSL sequences by relative starting position within the contigs. A strong bias towards the 5' ends was observed, with the first three bins respectively containing 590, 264, and 88 DinoSLs, which together represented 942 of the 1337 contigs that contained the complete canonical DinoSL sequence. Figure 13 provides a more detailed picture of DinoSL abundances at or near the 5' ends without scaling positions by contig length.

Past the first 30 nucleotides, a second cluster of SLs was observed near the 3' ends, and dozens of individual SLs were scattered through the contigs in which they occurred (Figure 12). Many of these sequences also contained multiple TransDecoder-identified coding sequences (see chapter 2). Sequences with internal SLs were investigated in IGV and by querying the NCBI non-redundant protein sequence database, but no cases were identified in which misassembly was not strongly suspected, or where the contigs aligned to reference sequences that were known to be polycistronic.

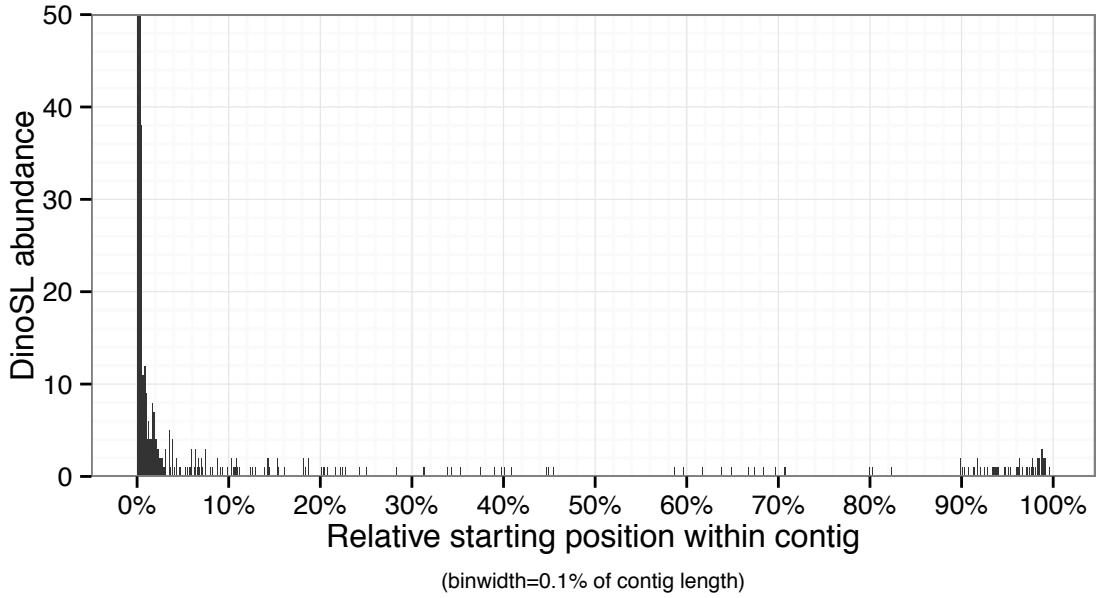


Figure 12 - Canonical DinoSL abundances by relative starting position

Abundances of the 21 conserved nucleotides of the DinoSL, plotted by relative starting position within the contigs (starting position divided by contig length). The results from all assemblies were combined because the individual counts were so low. The range was truncated at y=50 for readability. Only the first three bins exceeded this cap, with 590, 264, and 88 sequences, respectively.

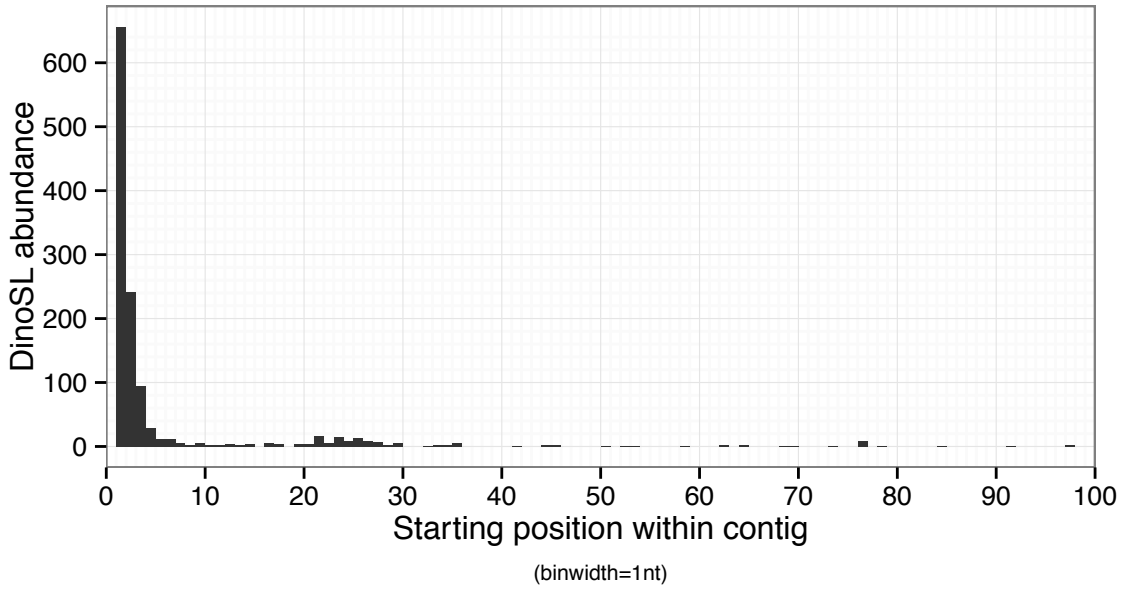


Figure 13 - Canonical DinoSL abundances by absolute starting position

Abundances of the 21 conserved nucleotides of the DinoSL, plotted by absolute starting position within the first 100 nucleotides of the contigs from all eleven transcriptomes.

To circumvent potential complications arising from misassembly, the raw reads were queried for exact matches to the canonical DinoSL (Figure 14). The reads were too short to confirm the presence of DinoSLs embedded within 3' UTRs or between coding regions of polycistronic transcripts, yet thousands of reads were identified that provide direct evidence of full-length canonical DinoSLs embedded up to 80 nucleotides from the 5' ends of their respective transcripts.

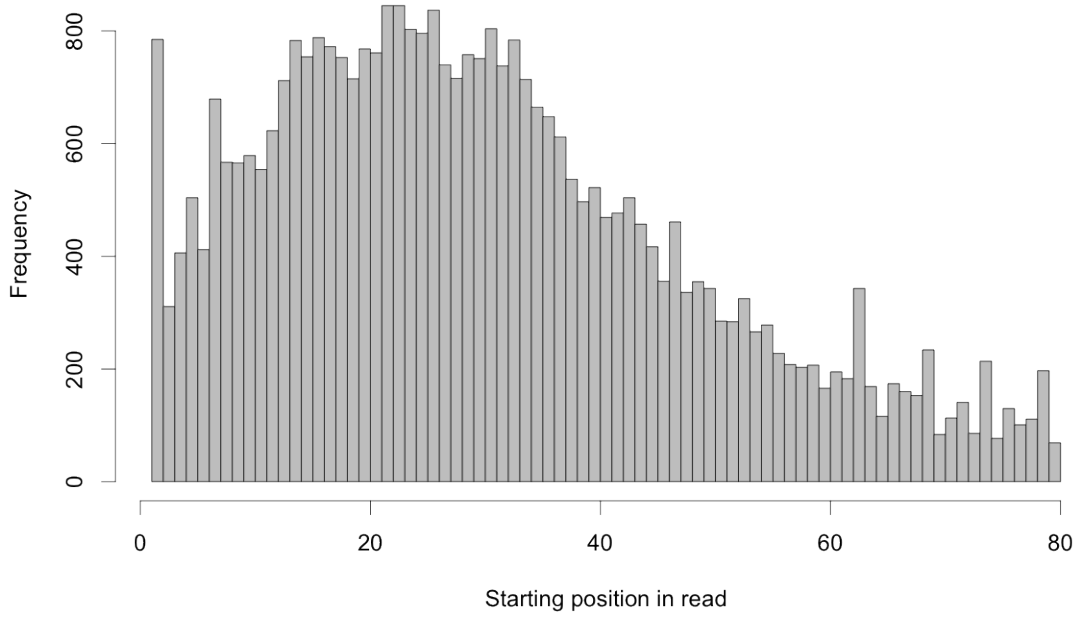


Figure 14 - Assembly-free evidence of internal canonical DinoSLs

Histogram of starting positions of the 21 conserved nucleotides of the canonical DinoSL within the trimmed reads, combined across all data included in this study.

Slamovits & Keeling (2008) found that relict SLs were quickly degraded after apparent reinsertion back into the genome. These perfectly conserved internal SLs could be the result of transcripts from genes that were recently duplicated via reinsertion of mature transcripts back into the genome and have not yet acquired any mutations within the relict SLs. Slamovits and Keeling showed that relict SLs were often partially or completely removed during subsequent trans-splicing because the canonical DinoSL sequence offers

two AG splice acceptor sites. New findings also show that in *Symbiodinium kawagutii*, the TTTG from positions 12-15 of relict SLs often become conserved promoters, while the more common TTTT promoter tends to occur upstream of relict SLs, despite also being present in the canonical DinoSL (S. Lin et al., 2015). In such cases, there is a chance that a new active SL will be spliced into an AG acceptor upstream of the relict SLs, leaving them intact. The occurrence of such an event within a recently recycled transcript could explain the presence of perfectly conserved, full-length DinoSLs embedded within the 5' UTRs of some sequences, and the rarity of such sequences hints at the speed with which relict SLs are degraded after being recycled back into the genome.

3.2.4 Serial spliced leaders

Identifying the degraded relict SLs described by Slamovits and Keeling required a modified strategy based on fuzzy-string matching (see Methods for details). Exact matches to full-length DinoSLs were rare, so the search was restricted to multiples of just the 15-nt suffix that follows the internal splice acceptor site at positions six and seven. Figure 15 shows the abundances of tandem spliced leader suffixes by their relative starting positions within the contigs, colored by similarity to the reference sequences.

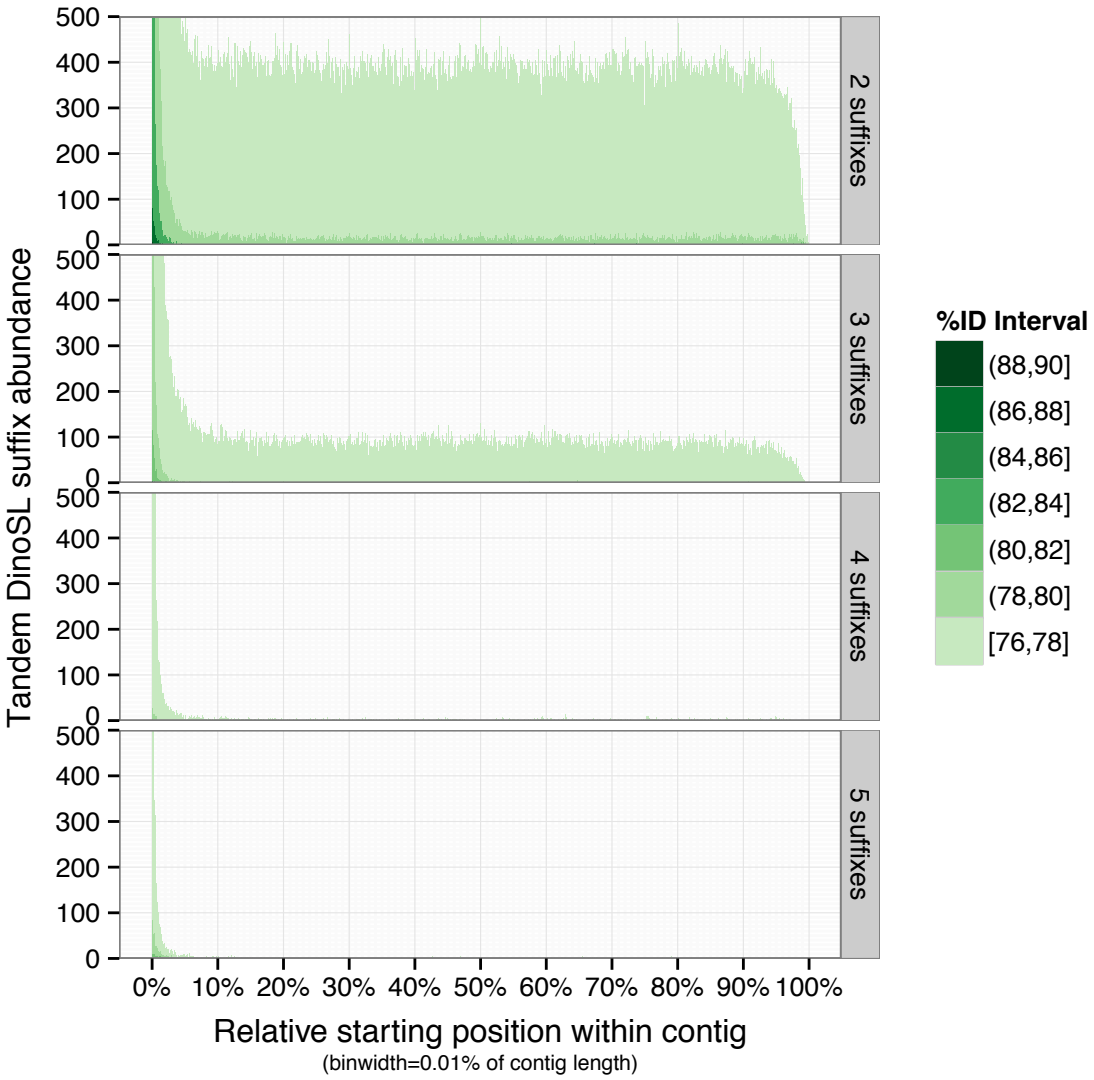


Figure 15 - Tandem DinoSL suffix abundances by relative starting position

Abundances of multiples of the 15-nt DinoSL suffix described by Slamovits and Keeling, plotted by relative starting position within the contigs from all eleven transcriptomes. Plots are faceted by the number of repeated suffix sequences (gray bars) and are filled by the percent of matching characters to the canonical sequence. The range was truncated at y=500 for readability.

The results were complicated by an obvious abundance of spurious matches, particularly from sequences matching 2 and 3-suffix repeats with no more than 78% identity. Nevertheless, an equally obvious signal comes through in the extreme bias towards the 5' ends of the transcripts. The increase near the 5' end was so great in fact, that it was nearly impossible to observe most of the distribution without truncating the counts at 500.

Therefore, as with the internal spliced leaders, a second set of plots was generated to show the distributions by absolute position near the 5' ends (Figure 16).

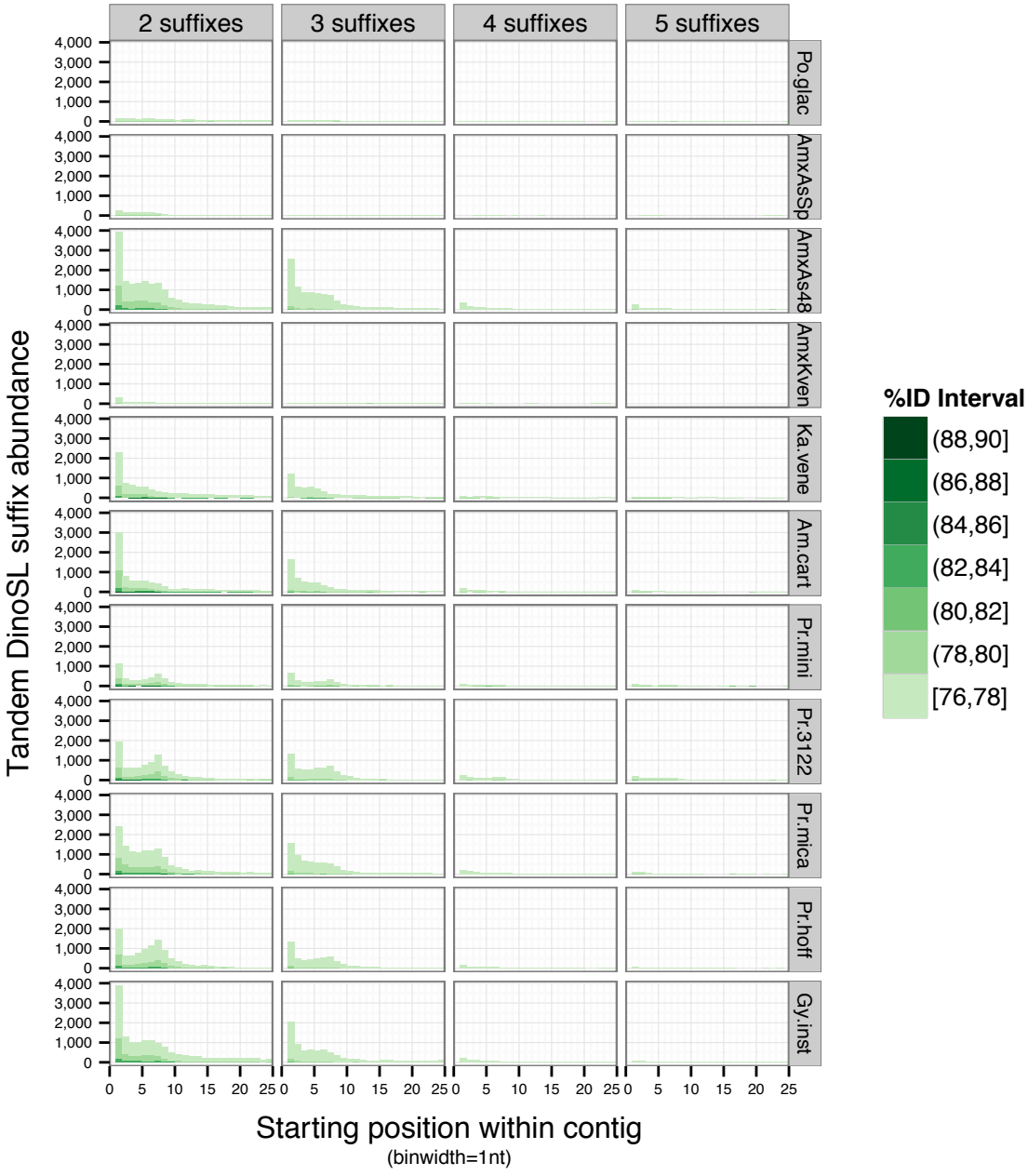


Figure 16 – 5' Tandem DinoSL suffix abundances by assembly

Abundances of multiples of the 15-nt DinoSL suffix described by Slamovits and Keeling, plotted by absolute starting position within the first 25 nucleotides of the contigs. Plots are faceted horizontally by the number of repeated suffix sequences and vertically by assembly. The histograms are filled by the percent of matching characters to the canonical sequence.

Counting only matches with >75% sequence identity to multiples of at least two tandem suffixes of the SL, beginning within the first 15 nucleotides of the contigs, the number of evidently recycled transcripts within the free-living dinoflagellates ranged from 12,501 in *Karlodinium veneficum*, to 26,871 in the combined assembly of *Amoebophrya* sp. ex *Akashiwo sanguinea* with its host from mRNA extracted 48 hours after infection. *P. glacialis* continued to be a conspicuous outlier, containing fewer than 2,500 evidently recycled transcripts. The assemblies enriched for *Amoebophrya* dinospores had numbers even lower than the flawed *P. glacialis* assembly, suggesting that this mechanism of transcript recycling might be dramatically reduced, or even absent from the *Amoebophrya*.

The local maxima seen in many of the plots around 8-nt are the result of additional upstream suffixes that were not assembled completely enough to meet the >75% identity threshold. In all cases, conserved, leading SL suffixes that reached the 5' end are indistinguishable from functional SLs, but all reported cases contained at least one additional SL suffix following any potentially functional SLs.

3.3 Discussion

The canonical DinoSL is known to be trans-spliced onto the 5' ends of a diverse set of transcripts within a diverse set of dinoflagellates (Senjie Lin et al., 2010) and is suspected of being ubiquitously spliced onto the 5' ends of all mature dinoflagellate mRNA transcripts (H. Zhang et al., 2009). This hypothesis cannot be confirmed using libraries that have been selectively amplified using primers based on the DinoSL, but all such studies to date have reported suspiciously few transcripts containing full-length 5' DinoSLs (Beauchemin et al., 2012; Cooper et al., 2014; Keeling et al., 2014; Meyer et

al., 2015; Ryan et al., 2014; H. Zhang et al., 2013; S. Zhang et al., 2014; Y. Zhang et al., 2014). By leveraging the statistical power of high-throughput short-read sequencing, I have shown that the paucity of full-length spliced leaders is an artifact of the technologies used to sequence and assemble them, and that suffixes as short six nucleotides are sufficient to positively identify a trans-spliced isoform.

Using this cutoff, tens of thousands of trans-spliced transcripts were identified within eight transcriptomes from a diverse set of free-living dinoflagellates. Even using this more sensitive threshold, relatively few trans-spliced transcripts could be identified within the transcriptomes of three additional dinoflagellates. The low levels of detectable DinoSLs in the *P. glacialis* assembly are presumed to be an artifact of technical limitations, but the relatively low DinoSL abundances in the two *Amoebophrya* assemblies appear to reflect underlying biological differences.

Amoebophrya are obligate intracellular parasites belonging to the Syndiniales, a major clade of dinoflagellates that is typically placed adjacent to the "core" free-living Dinophyceae. This classification has been based in part on a lack of other characteristic dinoflagellate features, such as enlarged, permanently condensed "dinokaryon" genomes. These results indicate that reduced prevalence of trans-splicing may be yet another feature that distinguishes this clade from the core dinoflagellates, and could be interpreted as evidence that the *Amoebophrya* possess an alternative means to cap their mRNA transcripts.

Ultimately, the hypothesis that the DinoSL is ubiquitously trans-spliced onto all dinoflagellate transcripts could not be confirmed, although these results correct the dramatic undercounting of trans-spliced transcripts identified within dinoflagellate

transcriptomes sequenced from short-read HTS, showing that they are indeed consistent with the hypothesis, contrary to what had previously been reported. Additionally, the presence of conserved internal SLs and degraded tandem SL suffixes near the 5' ends of thousands of transcripts is consistent with the hypothesis of widespread recycling of dinoflagellate mRNA transcripts back into the genome leading to lineage-specific expansion of gene families within the dinoflagellates (Slamovits & Keeling, 2008). Degraded tandem SLs were originally identified within a few hundred transcripts, and a few corresponding genes, collected from a diverse set of dinoflagellates. The genome-wide abundance of relict SLs has now been confirmed within *Symbiodinium kawagutii* (S. Lin et al., 2015) where nearly 5,600 (~15%) of the identified genes were found to have complete or partial relict SLs embedded within their UTRs. The genomes of the free-living dinoflagellates included in this study have genomes ranging from 3 to 300+ times the size of the *S. kawagutii* genome, yet have only 2-5 times as many transcripts that contain tandem SL suffixes. Poor coverage of the 5' ends is one likely factor, as it was with the conserved DinoSLs. Evidence for this can be seen in the peak at 8 nt in Figure 16, corresponding to partial SL suffixes that were not reconstructed completely enough to match an additional copy with sequence identity above 75%. Slamovits and Keeling also showed that any AG occurring within the vicinity of relict SLs can act as a splice acceptor site, including the AG at the 3' end of relict SLs, in which case the relict copies are completely removed and replaced by the new SL. There is no way to determine the number of recycled transcripts whose degraded SLs were either removed by subsequent splicing or simply not reconstructed during sequencing and assembly, or to know if the 15% recycling rate observed within the *S. kawagutii* genome is a reasonable

rough estimate for larger species. What is clear is that a growing body of evidence suggests that gene duplication through transcript recycling is a major source of genome expansion in dinoflagellates.

3.4 Methods

3.4.1 Counting *k*mers and startmers

Jellyfish version 2.1.4 (Marçais & Kingsford, 2011) was used to count all *k*mers from 4-22 nucleotides ('-k <k> -s 100000 -t 8' for k=4..22) in all eleven assemblies for the novel dinoflagellate transcriptomes described in chapter 2. The libraries were not strand-specific, so both forward and reverse-complemented sequences were counted together ('--canonical'). Position-independent DinoSL suffix abundances were extracted from the resulting Jellyfish databases using the 'jellyfish query' command, which requires no options other than the desired *k*mer sequence and the path to the database. With the '--canonical' flag, Jellyfish considers reverse-complemented sequences as duplicates and stores both counts as a single entry. The 'query' command accounts for this by searching for both orientations of each query sequence, making it unnecessary to explicitly retrieve and merge counts for the forward and reverse-complemented version of each DinoSL suffix.

We define "startmers" as the subset of *k*mers occurring at the 5' end of a sequence. Startmers were counted using startmers.py, a custom Python program written for this study. As with the *k*mers, counts from the 5' ends of both the forward and reverse-complement of each sequence were combined for this study.

3.4.2 Estimating DinoSL suffix enrichment over expected abundances

Assuming an even distribution of starting positions for a given *k*mer, the expected abundance of that *k*mer at a particular position (a_i) within a set of sequences can be estimated using the length of the *k*mer (k), the position-independent abundance ($\sum_{i=1}^{N} a_i$) of the *k*mer within the sequences, the number of sequences (N), their mean length ($\bar{l}$), with the following equation:

$$a_i = \frac{N \times \sum_{i=1}^{N} a_i}{N \times (\bar{l} \times (k-1))}$$

Estimates of enrichment were calculated by dividing the observed startmer abundances by the corresponding predicted abundances (a_0).

3.4.3 Identifying degenerate serial DinoSLs

A custom Python program (dinosl_abundances_by_position.py) that had been written to identify exact DinoSL sequences was modified to search for fuzzy matches of serial DinoSL suffixes using the fuzzywuzzy library (https://github.com/seatgeek/fuzzywuzzy). With full-length DinoSL sequences being so rare, the Trinity isoforms were searched for multiples of the 15-nt suffix, rather than multiples of the 15-nt suffix following a full-length DinoSL. The fuzz.ratio function reports the percentage of matching characters as the sequence length minus the Levenshtein edit distance, divided by the full sequence length, multiplied by 100. Each 75-nt substring within each assembly was compared to five multiples of the 15-nt DinoSL suffix. If a match exceeding 75% was found, it was reported and the algorithm moved onto the next position. If a match was not found, the substring was shortened by 15 nt and compared against four multiples of the DinoSL suffix using the same 75% sequence identity threshold. This was repeated down a single multiple of the DinoSL suffix.

3.5 Availability of supporting data and custom software

All custom software used in this study has been uploaded to a dedicated GitHub repository: https://github.com/trgibbons/DissertationScripts. Assemblies and other supplemental files are available on Figshare (T. Gibbons, 2015).

3.6 Contributions

This project was conceived and evolved through many discussions with Professors Tsvetan R. Bachvaroff (TRB), Charles F. Delwiche (CFD), and Stephen M. Mount (SMM), Doctors Gregory T. Concepcion, Endymion D. Cooper (EDC), and Bastian Bentlage, as well as Gregory S. Mendez (GSM) and Travis S. Rogers (TSR). All software was selected, evaluated, and executed or developed by Theodore R. Gibbons.

3.7 Acknowledgements

We would like to thank the staff at NCGAS for providing not only computational resources, but also friendly and responsive tech support (NSF Award #1062432 - *ABI Development: National Center for Genome Analysis Support*). Funding was provided by NSF awards 0629624, 1036506 and 1046075 to CFD.

4 Evaluation of BLAST-based edge-weighting metrics used for homology inference with the Markov Clustering algorithm

4.1 Background

Clustering protein sequences by inferred homology (descent from a common ancestral sequence) is a fundamental step for many analyses involving the growing number of large sequence data sets. Functional and structural predictions may be greatly accelerated by analyzing only a single representative sequence from each cluster and then transferring these annotations to the other cluster members. Other methods, such as phylogenetic inference, can only be applied to individual homologous groups. In either case, errors introduced in this critical early step can propagate throughout downstream analyses. It is therefore imperative that the sources and causes of these errors are understood so that they can be avoided, or at least mitigated.

Perhaps the most popular approach for identifying homologous sequences shared between multiple genomes is to generate all against all pairwise alignments using a program such as BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990; Camacho et al., 2009), followed by the application of fixed filtering thresholds or the Reciprocal Best Hits (RBH) algorithm (Rivera, Jain, Moore, & Lake, 1998). Fixed filtering thresholds can be an efficient way to remove large numbers of very poor hits, but are ineffective for most other uses (Paccanaro, Casbon, & Saqi, 2006). RBH works well for analyzing pairs of proteomes containing few inparalogs or other recently duplicated protein-coding genes, but is not easily extended to more complex cases that violate these conditions (we follow here the molecular homology terminology of Sonnhammer and Koonin (Heidelberg et al., 2002; Koonin, 2005; Remm, Storm, & Sonnhammer, 2001)). Some

attempts have been made to extend RBH for specific situations (Remm et al., 2001; R. L. Tatusov, Koonin, & Lipman, 1997), but the algorithm is not easily generalizable.

The Markov Cluster (MCL) algorithm (and corresponding program of the same name) is a robust and widely used alternative to RBH (Ekseth, Kuiper, & Mironov, 2014; Enright et al., 2002; L. Li et al., 2003). It was designed to handle data from an arbitrary number of organisms whose proteins share arbitrarily complex evolutionary histories. MCL operates on an abstracted graph-representation of a set of BLAST hits in which each sequence is stored as a node, and each BLAST hit is stored as a weighted edge connecting a pair of sequences. At the time of its release, MCL was unable to directly read a table of BLAST hits, and so the TribeMCL Perl module was published alongside the original MCL program to handle this task (Enright et al., 2002). The following year, the OrthoMCL suite of Perl scripts extended the TribeMCL method by normalizing the edge weights using inter-organism averages, while simultaneously circumventing memory limitations by interfacing with a MySQL relational database (L. Li et al., 2003). It has now been over a decade since the original publication of these programs, and the typical computer workstation contains multiple processing cores and has significantly more memory. These improvements in hardware recently motivated another group to reimplement OrthoMCL as a standalone, multithreaded C++ program orthAgogue (Ekseth et al., 2014). With it, Ekseth et al. also introduced an option to use the bit score as an alternative to the negative common log ($-\log_{10}$) of the expectation value (NLE) edge-weighting metric used by both TribeMCL and OrthoMCL. While the authors described several practical benefits of switching to the bit score, they offered no demonstration of its performance compared to the traditional NLE.

Our study provides the first direct performance comparison between the NLE and BS. To avoid potential confounding effects introduced by heuristic discrepancies between different implementations, we wrote our own custom version of OrthoMCL in Python. Our program stores and processes all metrics identically, and outputs a set of graphs that differ only by their edge weights. We have also included two additional metrics that have not previously been used for MCL-based clustering: the bit score divided by either the self bit score (termed BLAST score ratio (Rasko, Myers, & Ravel, 2005; Sahl, Caporaso, Rasko, & Keim, 2014) or bit score ratio; BSR) or the anchored alignment length (bit score over anchored alignment length; BAL). These latter two metrics were included to address fragmented and partial sequences that often result from short-read *de novo* DNA and RNA sequencing projects. Bit scores and E-values from alignments between these fragmented sequences can easily fall into the range of spurious hits between very distantly or unrelated sequences, causing the corresponding edges to be removed by MCL and thus leading to unwanted cluster fragmentation.

We compare here the performance of these four edge-weighting metrics over a range of MCL inflation parameter values, and consider different sequence fragmentation scenarios varying from all sequences being intact, to some or all being split into two or three subsequences. Contrary to our expectations, we observed that the bit score matched or exceeded the performance of all other edge-weighting metrics in each scenario. This suggests that the performance of the popular pipeline is actually improved by switching to the relatively simple bit score as an edge-weighting metric.

4.2 Results and Conclusions

4.2.1 Test database creation

Evaluation of inference methods requires a reference data set for which the correct solutions are known. It is not possible to go back and directly observe the evolution of extant organisms, and there is no single, universally recognized reference dataset for the problem of clustering protein sequences based on inferred homology, although the manually curated Eukaryotic Orthologous Groups (KOG) database is a popular choice (Roman L Tatusov et al., 2003). For this study, we used an extension of the Conserved Eukaryotic Genes Mapping Approach (CEGMA) database (Parra, Bradnam, & Korf, 2007), which is a subset of the KOG database.

The KOG database was derived from the proteomes of seven eukaryotes whose genomes had been sequenced, annotated, and published by 2003 (*Saccharomyces cerevisiae* (Goffeau et al., 1996), *Caenorhabditis elegans* (Consortium, 1998), *Arabidopsis thaliana* (Initiative, 2000), *Drosophila melanogaster* (Adams et al., 2000), *Encephalitozoon cuniculi* (Katinka et al., 2001), *Homo sapiens* (Lander et al., 2001), and *Schizosaccharomyces pombe* (Wood et al., 2002)). Each KOG (cluster) is an assertion that the sequences within it share a more recent common ancestor with each other than with sequences in any other KOG, guaranteeing neither that a KOG is free of outparalogs, nor that it contains all members of a particular group of (co)orthologs. This is partially due to the consideration of functional data by the human curators, and because only 50-75% of the predicted proteins from each organism were included. Despite this reduction from its potential size, the 60,758 KOG-annotated protein sequences still proved to be inconveniently large for the dozens of all *vs.* all BLASTP jobs required for this study.

The CEGMA database is a more computationally tractable alternative, containing sequences from only 458 KOGs that span all six free-living eukaryotes (Parra et al., 2007). From these KOGs, however, the CEGMA developers removed all sequences from the parasite *E. cuniculi,* as well as inparalogous sequences, which we observed to be a major source of false negative errors when clustering sequences based on inferred homology. Restoring these sequences to the 458 CEGMA KOGs also restored the false negative errors observed when analyzing the complete KOG database (Figure S2). The addition of sequences from four other eukaryotes (*Anopheles gambiae* (Holt et al., 2002), *Ciona intestinalis* (Dehal et al., 2002), *Chlamydomonas reinhardtii* (Grossman et al., 2003), and *Toxoplasma gondii* (Chance, Saronio, & Leigh, 1975)) for which the CEGMA developers have provided annotations led to a test data set we refer to as the Expanded CEGMA KOGs (ECK) database. ECK has been used for all tests in this study, except where otherwise noted. Statistics for all three databases are shown in Table 6.

Table 6 - Statistics for KOG, CEGMA, and ECK databases

	KOG		CEGMA		ECK	
	Seqs	KOGs	Seqs	CEGs	Seqs	ECKs
Homo sapiens	19,039	4,597	458	458	1,350	458
Arabidopsis thaliana	13,744	3,285	458	458	1,175	458
Caenorhabditis elegans	10,581	4,235	458	458	635	458
Drosophila melanogaster	8,445	4,351	458	458	611	458
Saccharomyces cerevisiae	4,003	2,668	458	458	606	458
Schizosaccharomyces pombe	3,728	2,762	458	458	557	458
Encephalitozoon cuniculi	1,218	1,073	-	-	311	291
Anopheles gambiae	-	-	-	-	453	453
Ciona intestinalis	-	-	-	-	432	432
Chlamydomonas reinhardtii	-	-	-	-	407	407
Toxoplasma gondii	-	-	-	-	303	303
Totals:	60,758	4,852	2,748	458	6,840	458

4.2.2 Software implementation

Each implementation of the TribeMCL BLAST-graph clustering algorithm uses custom software to convert a table of BLASTP hits into a MCL-readable graph. Complex software inevitably contains unique elements that can lead to performance differences between implementations of the same algorithm. Most of these differences will be slight and arise from seemingly minor decisions, such as whether to use the larger score from a pair of reciprocal BLAST hits (which are typically very similar, but occasionally unequal), or to average the two. It is difficult to catalogue these differences across published implementations, and harder still to gauge their impact. To remove such confounding factors from our comparison of the performance of different edge-weighting metrics, we wrote our own implementation, which stores all competing metrics within a single data structure and acts on them identically. It then prints a set of topologically identical graphs, differing only by edge weights.

Recent publications have focused on improving the computational performance in the graph creation (Ekseth et al., 2014) and clustering steps (Szilágyi & Szilágyi, 2014). However, in our tests, the runtime of every approach has been dominated by the shared BLASTP step. Even our unoptimized Python implementation required only a small fraction of the CPU cycles of BLASTP, and used only a few hundred megabytes of memory for the complete 60k+ sequence KOG database. My implementation, along will all other scripts used in our analysis pipeline, are publicly available at https://github.com/trgibbons/BlastGraphMetrics. This software was not intended for use beyond this study, so I also developed a more user-friendly implementation of the blast2graphs.py program called Porthos, which includes a basic version relying only on standard Python libraries. My hope is that Porthos will alleviate many challenges

associated with software installation, and that the code may serve as a helpful guide for future reimplementations. Porthos can be found at https://github.com/trgibbons/porthos.

4.2.3 BLAST graph topology and E-value cutoff

All edge-weighting metrics used in this study are derived from only the top scoring hit between each pair of sequences within a given test database. The theoretical limit for the number of such top hits is the square of the number of sequences in the database, although in practice the number of BLAST hits that pass any reasonably stringent E-value threshold will be much less than this theoretical limit. Consequently, each BLAST graph passed to MCL will be sparse, comprised of connected components that share no edges between them. Each of these connected components can be evaluated by the same criteria used to evaluate the clusters output by MCL, providing a baseline for the performance of MCL using any metric for weighting the edges of the adjacency graph.

Figure 17 shows the topology of the BLAST graph for the ECK database using an E-value cutoff of 1e-5. Each connected component can be considered a BLAST cluster. Before subsequent clustering with MCL, 237/458 ECKs have been perfectly reconstructed (contain all members of exactly one ECK) using BLASTP alone. Thirty one ECKs were split into multiple BLAST clusters and therefore cannot be rescued by MCL, which will divide connected components, but never combine them. In the end, only 54 BLAST clusters stand to be improved by MCL. The goal at this stage is to choose an edge-weighting metric that will give MCL the greatest chance to resolve or improve these multi-ECK BLAST clusters without (further) incorrectly splitting the 291 single-ECK BLAST clusters.

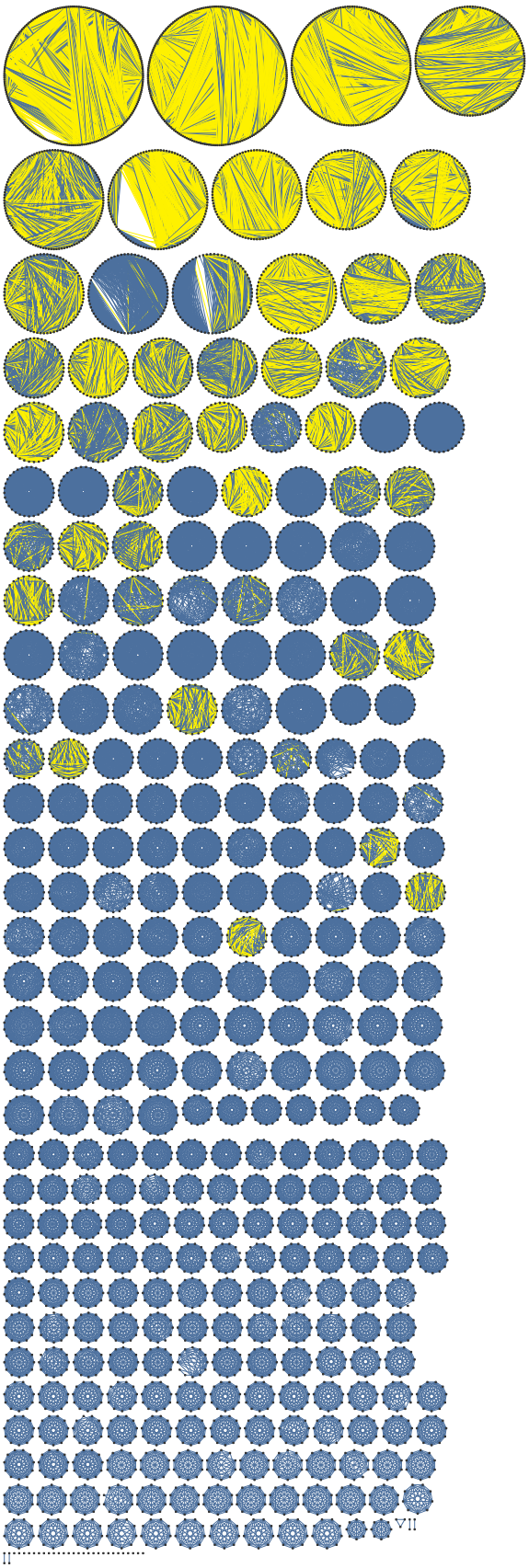


Figure 17 - Pre-clustering BLAST graph for ECK database with E-value ≤ 1e-5

BLAST graph representing sequences as nodes that have been arranged into rings corresponding to the connected components. Blue edges connect sequences from the same KOG. Yellow edges connect sequences from different KOGs. The green circles represent single-KOG clusters that have been successfully resolved using only BLAST's E-value threshold option (1e-5 in this case).

Predictably, a cutoff of 1e-3 decreased the number of fragmented ECKs and increased the average size of multi-ECK clusters, although it had little effect on the number of perfectly resolved ECKs (239 vs. 237). The effects of the E-value threshold on sequence clustering have previously been addressed in some detail (Apeltsin, Morris, Babbitt, & Ferrin, 2011; Feng Chen, Mackey, Vermunt, & Roos, 2007), so we chose not to explore this further and instead settled on a fixed threshold of 1e-5.

4.2.4 Simulated sequence fragmentation

Part of the motivation for this study was the problem posed by fragmented sequences resulting from *de novo* transcript assembly. It is not possible to generate alignments that span the full-length of a protein sequence when part of the sequence is missing. Previous studies have not addressed the impact of this on either the weighted adjacency matrix or the resulting clusters. This study explored the effects of fragmentation on clustering by splitting portions of the sequences in the ECK database into two or three subsequences before performing the all-vs-all BLASTP alignments. To determine if sequences from different organisms played equivalent roles in cluster formation, each fragmentation scheme was first applied along organismal lines (Figure 18, a & c), then randomly, such that a sequence's organism of origin did not affect its likelihood of being fragmented into a particular number of subsequences (Figure 18, b & d).

The number of sequences contributed from a particular organism varies within each ECK. To ensure that the relative proportions of sequence fragmentation remained constant within a given ECK across test scenarios, fragmentation labels were applied to each sequence within each ECK according to organismal origin, prior to fragmentation. One set of test sequences was generated by fragmenting them according to these labels

(i.e. sequences labeled with a “2” were split once, those labeled with a “3” were split twice, and sequences labeled with a “1” were left intact), then the labels were shuffled and reapplied to obtain a second set of fragmented sequences.

Within these “ordered” and “shuffled” test sets, the effects of applying evenly spaced breakpoints (Figure 18, a & b) were tested against the application of randomly spaced breakpoints (Figure 18, c & d), resulting in a total of four test data sets for each fragmentation scheme. No differences were observed between the “ordered” and “shuffled” test sets for any fragmentation scenario (Figure S3Figure S4). In contrast, the distributions of break points within each sequence, and therefore often the alignment of break points across homologous sequences, had a dramatic effect in every test case (see following section).

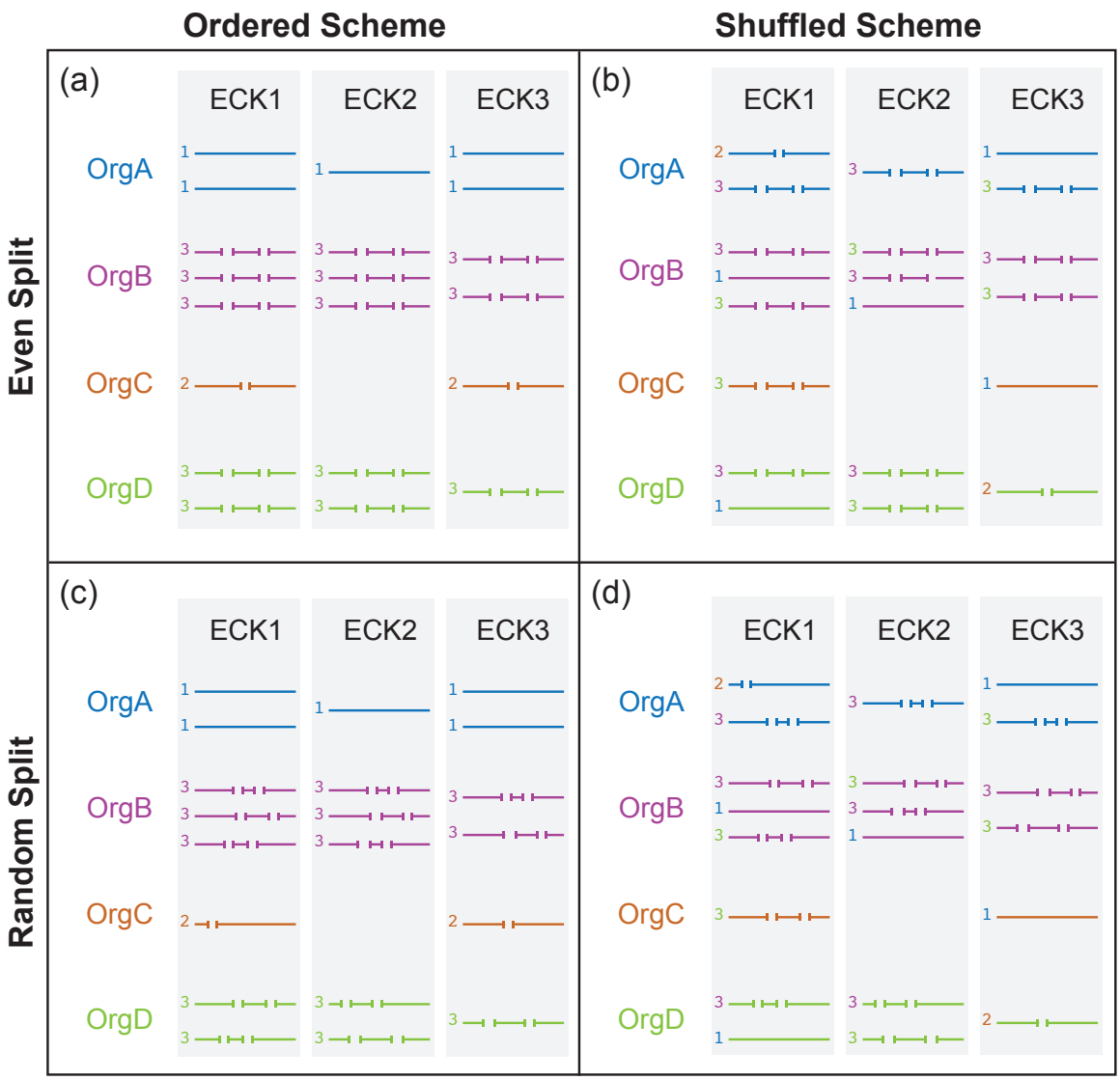


Figure 18 - Illustration of simulated fragmentation

Illustration of simulated sequence fragmentation. This example illustrates the four different ways in which the fragmentation scheme 1323 would be applied to a toy input test database with only three clusters containing sequences from only four organisms. The four resulting test sets represent a cross of two variables, arranged here into rows and columns. The sequences in the top row (a & b) have been split into even subsequences. The sequences in the lower row (c & d) have been randomly fragmented into uneven subsequences. In the left column (a & c), the user-defined integer assigned to each organism directly determines the number of subsequences into which each sequence is split. In the right column (b & d), these integers are first mapped to all sequences within a cluster, but are then shuffled within that cluster before fragmentation

4.2.5 Edge-weighting metrics and sequence fragmentation

The transformation of a set of BLAST hits into an MCL-readable graph encompasses the majority of the steps that distinguish the popular software packages from one another and is at the heart of this study. TribeMCL, OrthoMCL, and later versions of MCL itself, all use the NLE from the top hit between each pair of sequences as the basis for the edge weights in the BLAST graph. In addition to the NLE, the recently published orthAgogue program also offers a sum of bit scores from all non-overlapping hits between two sequences as an alternative edge-weighting metric.

There are several practical benefits gained by using the bit score in place of the NLE (Ekseth et al., 2014), most notably the ability to combine scores from non-overlapping hits using simple addition. Another important benefit is that, while BLAST (v2.2.28+) rounds E-values below 1e-180 to zero, all bit scores are accurately reported. The log of zero is undefined and, so these "missing" values must be heuristically supplemented. Even alignments between protein sequences of modest length (ca. 200-300 amino acids) can generate E-values exceeded this rounding threshold, so this is not merely a theoretical concern. In practice, many of the strongest NLE edges in the graph end up being weighted by heuristics, rather than by the rigorous statistical metrics produced by BLAST. Using bitscores to weight the edges avoids this complication.

Use of the bit score introduces a different problem, however, which the BLAST E-value was meant to solve. The bit score is linearly correlated with alignment length. In fact, by definition the bit score increases, on average, by two *bits* of information for every pair of aligned amino acids. This means that long sequences have the ability to produce large bit scores from relatively poor alignments, while short sequences may not be able to generate large bit scores from even perfect, full-length alignments. The E-value attempts to

account for these potential differences in sequence length (Equation 1), but (for the purposes of MCL-based clustering) does so at the cost of the other practical limitations mentioned above.

$$\text{E-value} = \frac{mN}{2^{BS}} \quad [1]$$

m = query sequence length

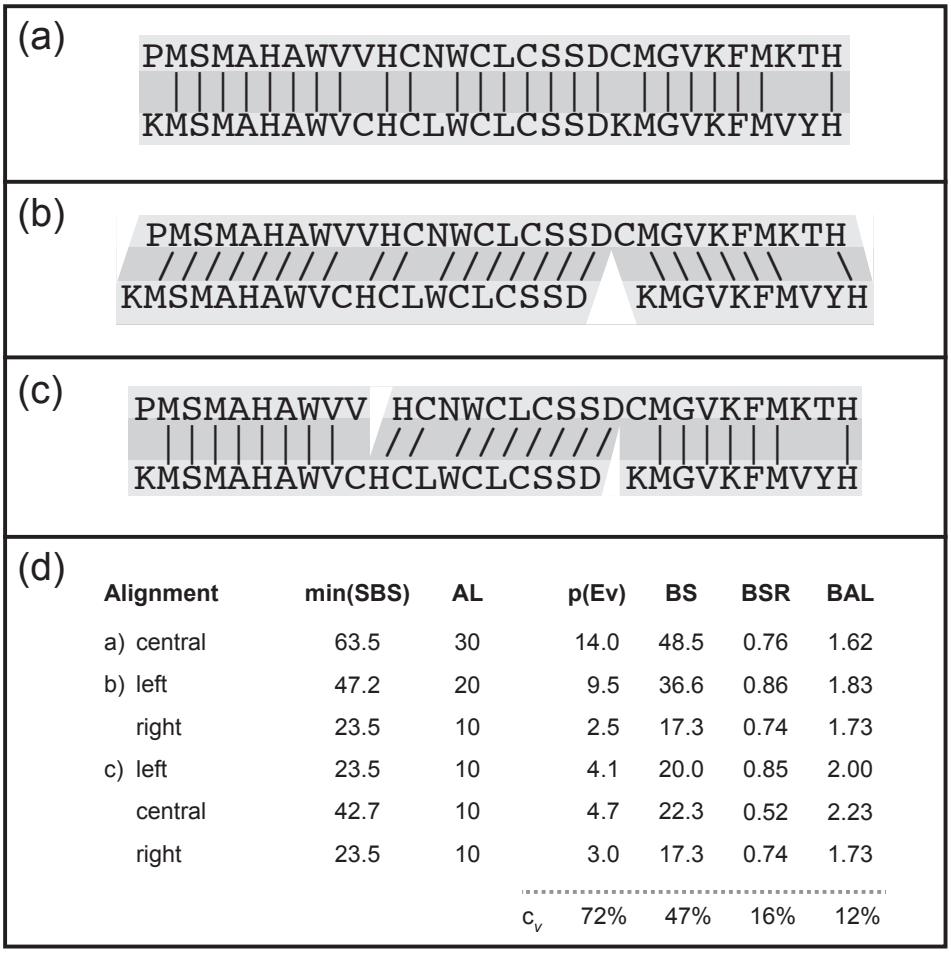
N = concatenated sequence database length

BS = bit score

In their original 2002 paper, Enright et al. mentioned that alternative metrics using length normalization might outperform the NLE. Until recently, this area has not received much attention in the literature. One reason may be that high quality alignments between pairs of full-length sequences are often (nearly) end-to-end. As long as detectable homologs are of similar lengths and are not fragmented, competing alignments will have both similar scores and lengths, negating any effects from length-based normalization. These assumptions are violated by partial and fragmented sequences, which are increasingly common as the products of high-throughput short-read *de novo* sequencing projects. It therefore seemed worth considering such metrics alongside the NLE and the BS.

Of the novel metrics considered for this study, the simplest was the bit score divided by the length of the shorter of the two sequences. This metric penalizes partial alignments between full-length sequences that share only a small conserved domain, while aiming to give equivalent weight to alignments between homologous sequences, whether they are both full-length (Figure 19a), similarly fragmented, or one has been fragmented into a subsequence of the other (Figure 19b). To work as intended, this metric assumes a direct linear relationship between alignment length and bit score. As stated above, the bit score

between two perfectly aligned sequences will be twice the alignment length, on average, but the underlying distribution could be broad and/or skewed.



Alignment	min(SBS)	AL	p(Ev)	BS	BSR	BAL
a) central	63.5	30	14.0	48.5	0.76	1.62
b) left	47.2	20	9.5	36.6	0.86	1.83
right	23.5	10	2.5	17.3	0.74	1.73
c) left	23.5	10	4.1	20.0	0.85	2.00
central	42.7	10	4.7	22.3	0.52	2.23
right	23.5	10	3.0	17.3	0.74	1.73
			c_v 72%	47%	16%	12%

Figure 19 - Illustration of edge-weighting metrics

Toy example demonstrating performance similarities and differences between the graph-weighting metrics in three scenarios simulating different fragmentation scenarios: (a) alignment between two full-length sequences, (b) alignment between one full-length sequence and one unevenly fragmented sequence, and (c) alignment between two unevenly fragmented sequences. Section (d) lists information about each alignment, including the minimum self bit score (SBS), the anchored alignment length (AL), and each of the four edge-weighting metrics. The coefficients of variation ($c_v = \sigma/\mu$) summarize the variation relative to the respective means for each metric.

To observe a real distribution, we plotted bit scores from full-length self-alignments (self bit score; SBS) against sequence lengths for the complete KOG database (Figure 20). The correlation turned out to be remarkably tight and close to the theoretical relationship (blue line). Furthermore, preliminary results showed no appreciable difference between

this simple bit score/length metric and a modified version using a ratio of the alignment bit score to the smaller of the two SBSs (bit score ratio; BSR), so only the results from the BSR are included here. Despite the similar performance, the BSR is preferable because it does not require modification of the tab-delimited BLAST output, which does not contain the sequence lengths by default, and because it remains theoretically possible for a sequence to generate a SBS substantially smaller or larger than twice its length.

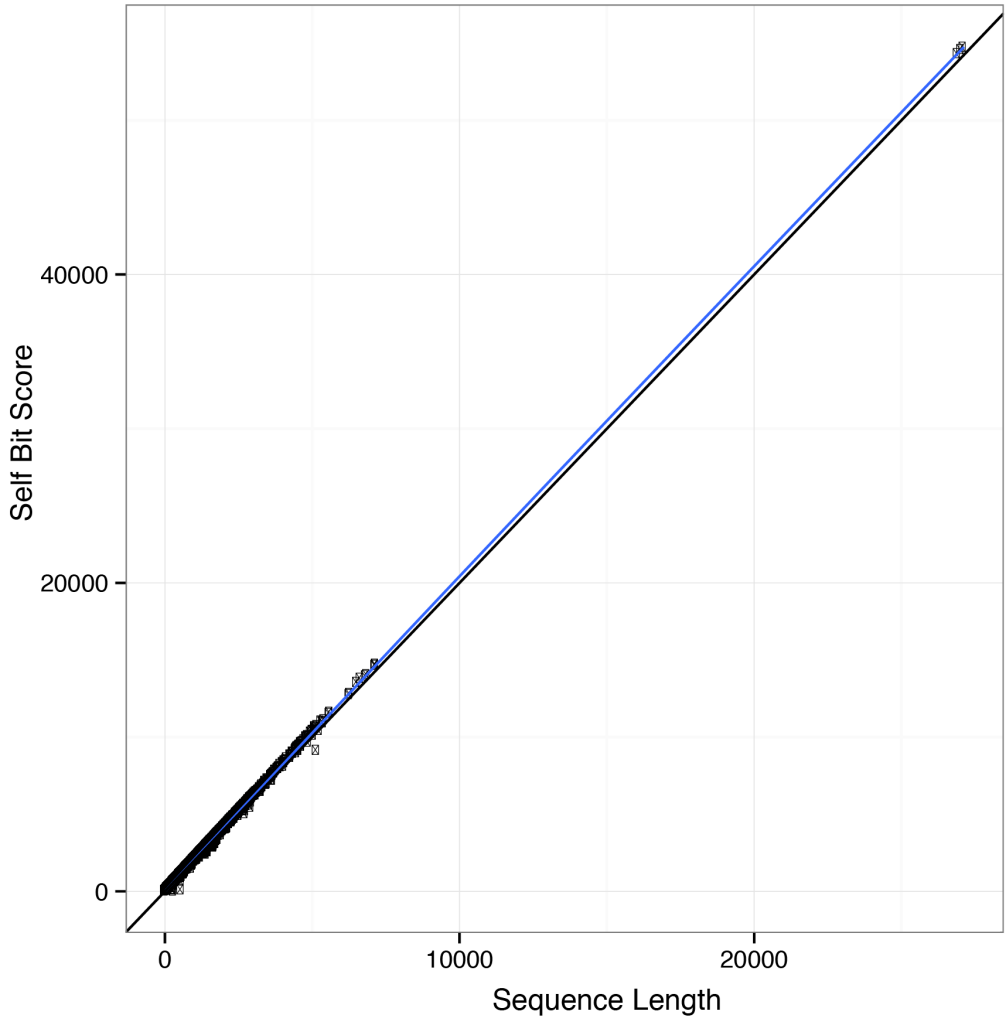


Figure 20 - Self bit score vs. sequence length

Scatterplot of self bit score vs. sequence length for all 60k+ protein sequences in the KOG database, plotted against the theoretical bit score = 2x sequence length (black line). The blue line is a smoothed line of best fit. 95% confidence intervals were plotted, but are not visible due to the tightness of the distribution.

As with simple length normalization, the BSR is designed to rescue edges between homologous sequences when one has been fragmented into a subsequence of the other, and it likewise falters when both sequences are incomplete and overlap on opposing ends (Figure 10c, middle alignment). In these cases, dividing by the full length of the shorter of the two sequences (or the smaller of the two SBSs) penalizes alignments for not extending beyond the homologous region shared between the two sequences, and in this way does not faithfully accomplish what is desired with such normalization.

Unfortunately, while bit scores can be easily combined, they are not easily split. Computing a SBS for just the alignable region would require re-running BLASTP after identifying a set of alignable homologous regions. It is therefore convenient that sequence length turned out to be a reasonable proxy for the SBS in most cases, as it is much more easily manipulated. These observations inspired the final metric considered in this study.

For the BAL metric, the aligned sequences are first anchored relative to each other based on the coordinates of the top BLASTP hit. The bit score from each alignment is then divided by the sum of the alignment length and the lengths of the shorter overhanging sequences extending from either side of the aligned region (Figure 19). In other words, the bit score from each top hit is normalized by the length of the maximum alignable region anchored by that hit. This inflates edge weights between fragmented sequences, whether or not one is a subsequence of (a homolog of) the other, while simultaneously deflating edge weights between full-length sequences that share only a small conserved domain.

In practice, all three scenarios shown in Figure 19 are likely to be encountered when clustering data from short-read *de novo* sequencing projects, and edges with relatively small weights will be eliminated in favor of those with larger weights. The coefficients of variation (a ratio of the standard deviation to the mean) illustrate how dramatically the BSR and BAL metrics can reduce the range of scores between competing high-quality alignments, increasing the likelihood that alignments between overhanging ends will persist and connect subclusters within a group of fragmented homologous sequences.

An initial performance comparison was made across all metrics using full-length sequences. Performance was measured in two ways that respectively evaluate the sensitivity (Figure 21) and specificity (Figure 22) for each metric across a range of MCL inflation parameter values. The MCL inflation parameter affects the granularity of the clusters, with larger values leading to smaller clusters. Popular values for inferring sequence homology are around 1.5, although any value larger than 1.0 should lead to convergence. Sensitivity was measured as the number of clusters into which the members of a particular ECK have been split, and specificity as the number of unique ECKs to which the members of a particular MCL cluster belong. There are a total of 458 ECKs, so a perfect score by either metric is 458 MCL clusters, each containing all sequences for a single ECK. When all sequences are intact, the performance of the various metrics is nearly identical (Figure 21 &Figure 22, top rows). Close inspection reveals that the NLE is slightly less sensitive than the other metrics for this data set, although the difference does not seem significant. In contrast, dramatic differences emerged once some of the sequences were split into subsequences (Figure 21 &Figure 22, lower rows).

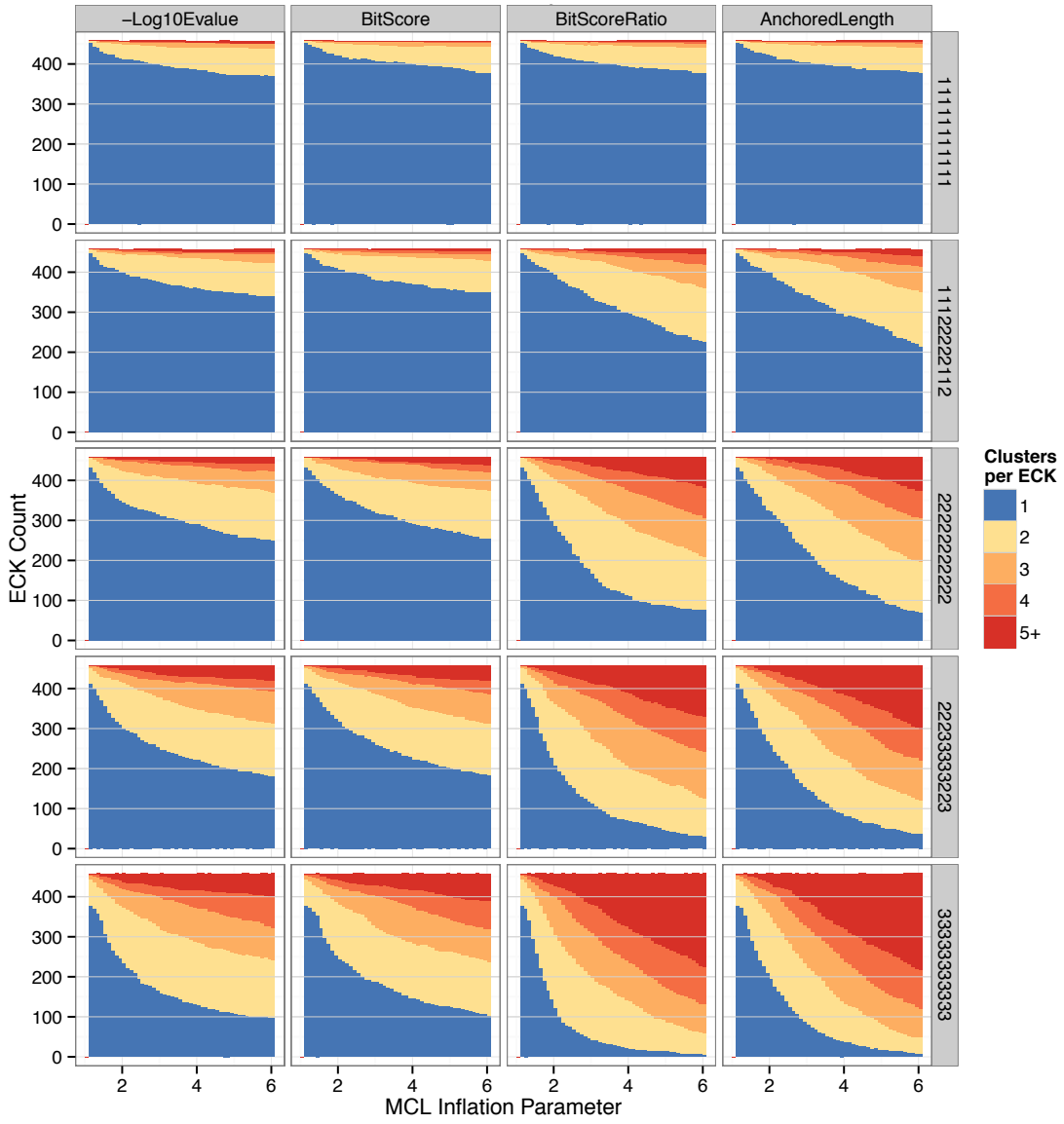


Figure 21 - Sensitivity performance comparison for each edge-weighting metric and fragmentation scenario

Sensitivity performance of MCL on graphs weighted using each of the four metrics (columns) over a range of inflation parameter values (x-axes) in a variety of fragmentation scenarios (rows). Vertically stacked bars indicate the number of clusters into which the members of a particular ECK have been split. Blue segments represent ECKs that were completely contained within a single MCL cluster. Other segments represent ECKs that were split into two or more MCL clusters, with redder color indicating higher degrees of fragmentation. The number of ECKs is fixed, so each stack sums to exactly 458. Five different simulated fragmentation scenarios are displayed as faceted rows: 11111111111 – all sequences are intact; 11122222112 – approximately half of the sequences have been split into two pieces; 22222222222 – all sequences have all been split into two pieces; 33311222111 – approximately one third of the sequences have been split into two pieces, one third have been split into three pieces, and the remaining third were left intact; 33333333333 – all sequences were split into three pieces.

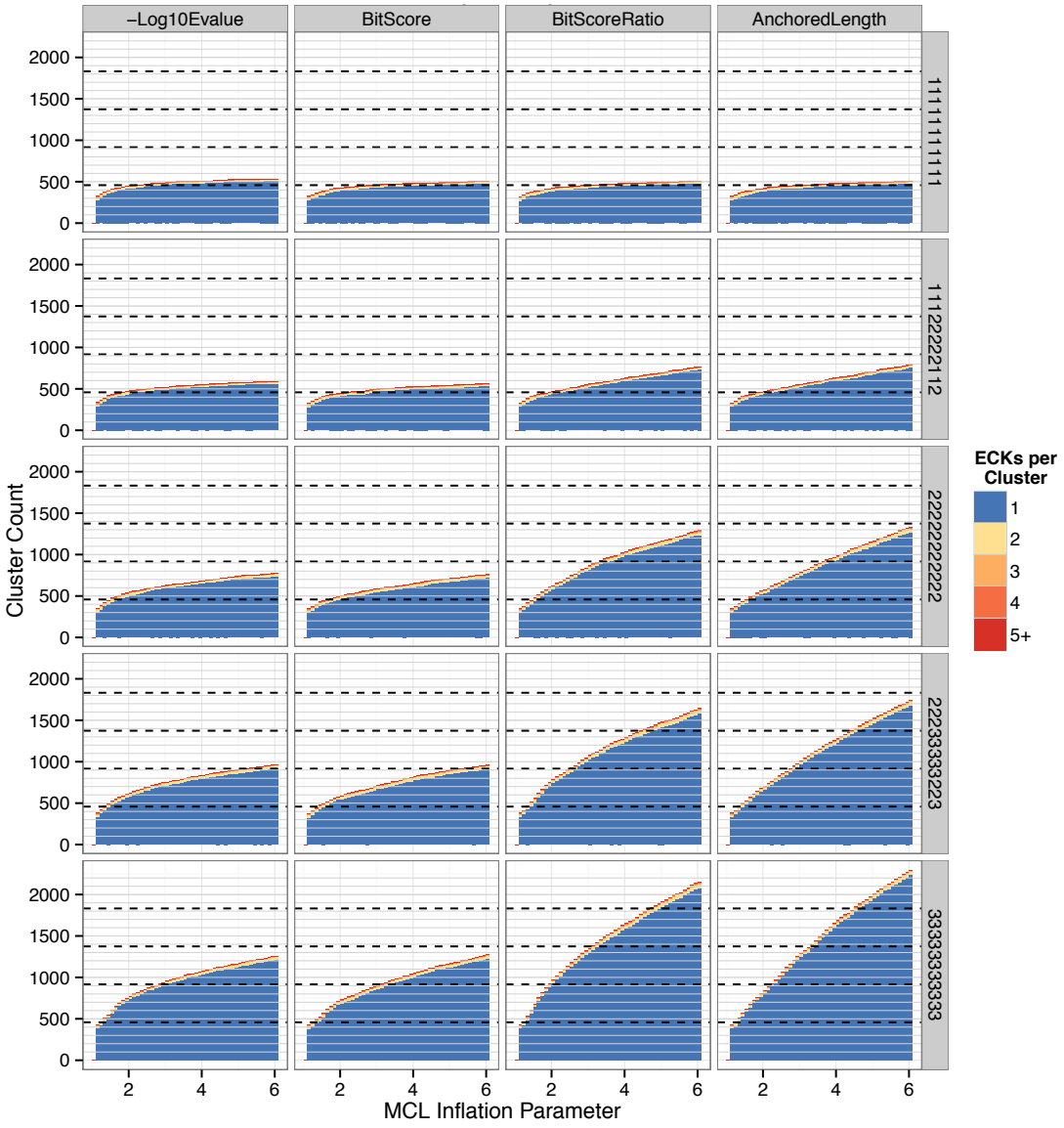


Figure 22 - Specificity performance for each edge-weighting metric and fragmentation scenario

Specificity performance of MCL on graphs weighted using each of the four metrics (columns) over a range of inflation parameter values (x-axes) in a variety of fragmentation scenarios (rows). Vertically stacked bars indicate the number of unique ECKs from which the members of a particular MCL cluster originated. Blue segments represent MCL clusters that contain sequences from only a single ECK. Other segments represent MCL clusters containing sequences from two or more ECKs, with redder color indicating higher degrees of contamination. A large number of "pure" clusters containing only a small portion of a particular ECK can appear deceptively good, so multiples of the desired 458 total clusters are indicated with dashed horizontal lines. Simulated fragmentation scenarios are as described in Figure 21.

The beneficial effects of the BAL normalization on clustering sensitivity can be seen when all of the sequences are split into fragments of equal size (Figure 23, Figure S5).

For instance, when the sequences are all split into halves, many of the break points line up, preventing any alignable overlaps and ensuring undesirable subclustering. Due to the sequence length variation within some ECKs (Figure S6), however, a few BLASTP hits do connect otherwise disparate subclusters, and the BAL metric was able to successfully preserve many of these critical edges. Unfortunately, when the breakpoints are randomized and longer overlaps become more common, the benefits from these few rescued edges are quickly overshadowed by an unintended negative side-effect of the normalization (Figure 21 & Figure 22, lower rows).

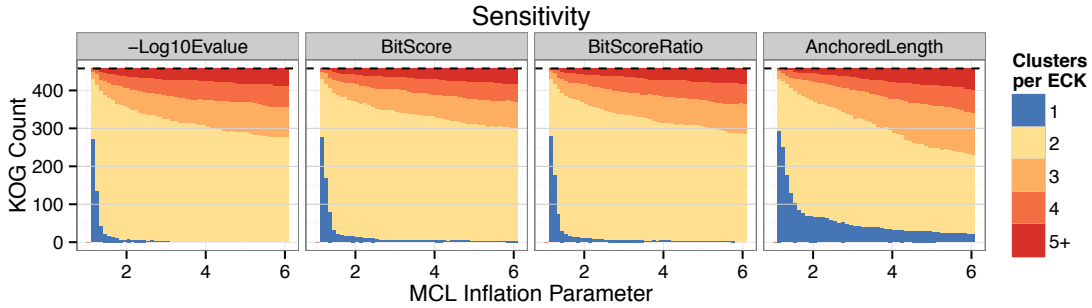


Figure 23 - Clustering performance comparison when all ECK sequences were split into even halves

Performance on the ECK 22222222222 "even" data set, in which all sequences were split into even halves. Plots are otherwise as described in Figure 21. Bit scores generated by high-quality alignments between full-length sequences are commonly 1-2 orders or magnitude greater than high-quality alignments between small sequence fragments or short overlapping regions shared by long sequences. A comparable differential in bit scores can also be seen between high- and low-quality full-length alignments, and this dynamic range turns out to be critical to the success of MCL-based homology inference. While the BSR and BAL do help to differentiate between the distributions of high-quality short alignments and low-quality long alignments, they also tighten the overall distribution by down-weighting the heaviest edges (Figure 24). In doing so, these metrics make it less obvious to MCL that these exceptionally good alignments should be kept.

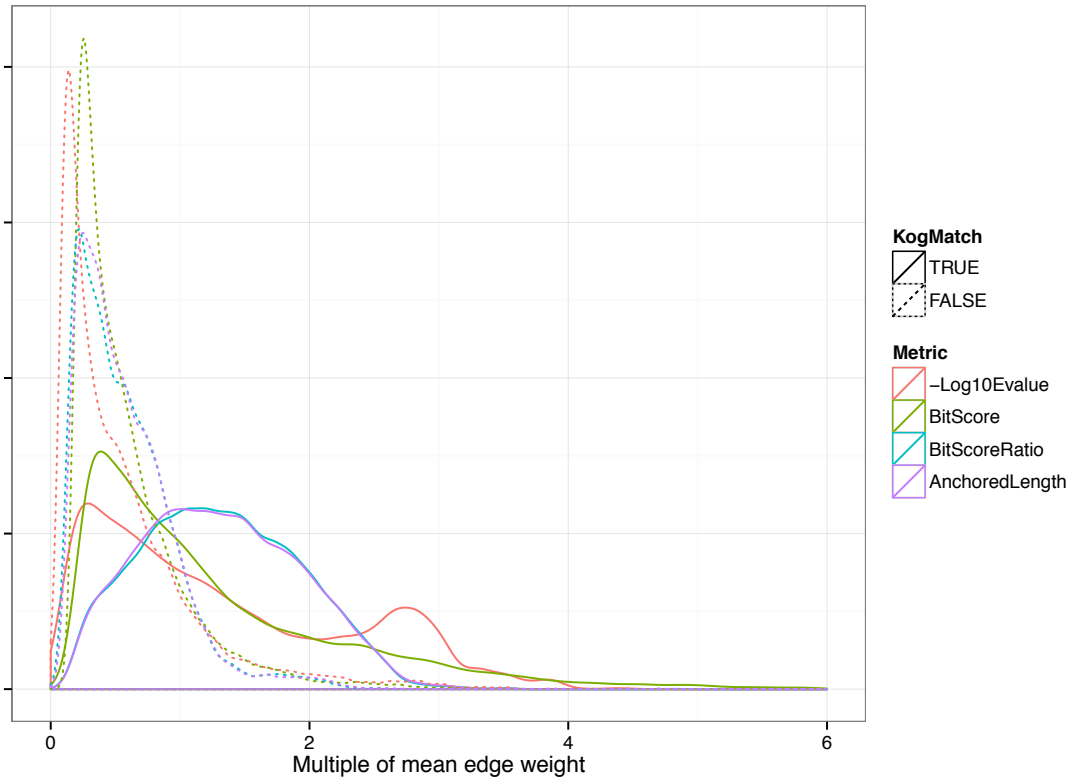


Figure 24 - Distribution of intra- and inter-ECK edge weights by metric

Probability density plots for all four metrics, scaled by their respective mean edge weights. Each distribution has been split into intra- and inter-ECK distributions.

The most important discovery from this study is the observation that the bit score performs as well or better than all other metrics in all conditions we tested. Considering the practical benefits of using the bit score (Ekseth et al., 2014), it appears to be the best choice between the two metrics currently in popular use, and not improved by either of the other two normalization methods (BSR, BAL) considered here.

4.2.6 Inter-organism normalization

One of the principle improvements introduced by OrthoMCL over TribeMCL was inter-species normalization. After a graph has been created and all missing E-values have been replaced using their heuristic, OrthoMCL computes the average edge weight over the entire graph, and also for the set of edges between each pair of organisms. It then

multiplies each edge by a ratio of the average edge weight between the two corresponding organisms over the average edge weight for the entire graph. This has the effect of increasing edge weights between organisms whose sequences are more divergent in sequence space, while decreasing edge weights between organisms that are relatively close.

For each test data set, we generated a complete set of graphs both before and after inter-organism normalization. The effect was positive but minimal in all test cases evaluated for this study. Supplemental Figure S7 & Figure S8 show the effects for full-length sequence clustering. The decision to include the additional organisms annotated by the CEGMA developers was based primarily on a desire to better demonstrate the effect of inter-organism normalization. It is possible that such normalization could significantly improve performance in certain cases, although our results demonstrate that it's not always worth the trouble.

In cases where normalization is desired, it is not necessary to use complex or highly optimized software. We demonstrate this with our own Python program Porthos, which uses only standard modules and accomplishes the task with fewer than 100 lines of code, including the help menu. Despite this simplicity, Porthos is able to cluster all 60,758 protein sequences from the complete KOG database (2,572,291 best BLASTP hits) in less than one minute and requiring less than 300MB of RAM. We present Porthos as both a simple, portable alternative to some of the more complex programs and as a heavily-commented example for anyone seeking to incorporate similar functionality into their own projects.

4.3 Discussion

In every scenario evaluated in this study, the bit score performed as well or better than the other three edge-weighting metrics for MCL-based inference of protein sequence homology. The significance of this finding lies in the practical benefits of using the bit score over the alternative metrics, which all require extension and/or post-processing of tab-delimited BLAST results. We further observed little benefit from inter-organism normalization, indicating that an MCL-readable graph file created by simply extracting the sequence identifier and bit score columns from a standard tab-delimited BLAST output file could produce results comparable to those obtained from the popular OrthoMCL program.

4.4 Methods

4.4.1 Test database creation

The 458 Eukaryotic Orthologous Groups (KOGs) used to create v2.5 of the Conserved Eukaryotic Genes Mapping Approach (CEGMA) clusters (CEGs) were extracted in their entirety from the complete KOG database in order to recover the inparalogs removed by the CEGMA developers. Protein sequences from four additional organisms (*Anopheles gambiae*, *Chlamydomonas reinhardtii*, *Ciona intestinalis* and *Toxoplasma gondii*) that were annotated for an unpublished version of the CEGMA database (http://korflab.ucdavis.edu/datasets/cegma/) were then added to these 458 KOGs to increase the taxonomic diversity. None of these four organisms contributed more than a single sequence to any cluster, and none contributed sequences to all 458 clusters. We refer to the resulting 458 clusters as Expanded CEGMA KOGs (ECKs). Nearly all analyses in this study were carried out with these ECKs, although a few were repeated using the CEGMA and/or KOG databases.

4.4.2 Analysis pipeline

A central Bash script called eckPipeline.sh was developed to streamline our analysis pipeline. The pipeline has 5 major steps: 1) sequence fragmentation, 2) sequence alignment, 3) graph creation, 4) sequence clustering, 5) generation of supplemental files and summary statistics.

4.4.2.1 Sequence fragmentation

A pre-formatted sequence database and user-defined fragmentation scheme are passed to the eckTestData.py Python program, which applies the scheme to the input database to generate a set of test databases that simulate fragmentation of the sequences. Each fragmentation scheme is encoded as an integer, with each digit being mapped to an organism in alphabetical order. If the integer does not contain at least one digit for each organism represented in the sequence database, it is repeated until it reaches or exceeds this threshold, then truncated as needed. The fragmentation schemes used in this study were encoded as eleven-digit integers because the ECK database contains sequences from eleven organisms (Table 6).

Each organism contributed a different number of sequences to the ECK database, so in order to simulate half of the sequences in the database being fragmented into a certain number of pieces, it was not sufficient to simply use the first or last half of the organisms. No combination of organisms perfectly divides the database into equal halves containing exactly 3,420 sequences in each, although a combination of the sequences from *A. gambiae*, *A. thaliana*, *C. elegans*, *S. cerevisiae*, and *S. pombe* gets very close with a total 3,426. Thus, the fragmentation scheme that simulates half of the sequences being split into two pieces was encoded as 11122222112, and the scheme that simulates half

being split into two pieces and the other half split into three was encoded as 22233333223.

From this mapping, the pipeline generates two pairs (a total of four) test data sets. In the "ordered" pair, these digits directly determine the number of fragments into which the sequences from a particular organism will be split. In the "shuffled" pair, the integer labels are first mapped to the sequences within each ECK cluster, but then they are shuffled before fragmentation (Figure 9). Within each pair of data sets, the "even" set uses evenly distributed breakpoints within each sequence, while in the "random" set they are randomly distributed.

The format of the sequence database is described in the GitHub wiki.

4.4.2.2 Sequence Alignment

Each of the four data sets generated in the first step are formatted as BLASTP databases and then aligned against themselves using BLASTP v.2.2.28+ with an E-value cutoff of 1e-5 (-evalue 1e-5) and soft masking turned on (-soft_masking true). The output is formatted as a tab-delimited table with headers and two extra columns for the query and subject sequence lengths (-outfmt '7 std qlen slen').

4.4.2.3 Graph Creation

Graph creation is accomplished with the blast2graphs.py Python program, which converts a table of BLAST hits, with or without header lines, into a set of eight graphs corresponding to the four edge-weighting metrics used in this study, both before and after inter-organism normalization. All metrics use only the best hit for each pair of aligned sequences.

Normalization is accomplished by calculating, for each metric, the average edge weight for the entire graph, and the average weight for all edges connecting sequences from each

pair of organisms. Each edge is then multiplied by a ratio of the average graph edge weight over the average weight between the corresponding organisms.

4.4.2.4 Sequence Clustering

For each graph, MCL (v12-068) is used to generate clusters with inflation parameter values ranging from 1.1 to 6.0, creating a total of 50 clusterings per graph.

4.5 Supplemental Files and Summary Statistics

The mcl2rtab.py Python program is used to generate a pair of files containing summary statistics for all 400 clusterings corresponding to the eight graphs generated from a single BLASTP file. These files are then converted into stacked bar charts using ggplot2 in R with the barcharts.R program. One last custom Python program called graphs2gml.py is used to generate annotated representations of the graphs and clusterings in a variety of popular file formats that can be read into interactive graph visualization software, such as Cytoscape (Shannon et al., 2003; Smoot, Ono, Ruscheinski, Wang, & Ideker, 2011).

4.6 Availability of supporting data and custom software

All custom software used in this study can be cloned from a dedicated GitHub repository: https://github.com/trgibbons/BlastGraphMetrics.git. Instructions for installation and execution are included in the project wiki.

The ECK test database was derived from the publicly available CEGMA and KOG databases. The CEGMA database can be downloaded from the Korf lab website: http://korflab.ucdavis.edu/datasets/cegma/. The KOG database can be downloaded from the National Center for Biotechnology Information website: ftp://ftp.ncbi.nih.gov/pub/COG/KOG. The downloadEckDatabase.py program can also be used to automatically fetch and format all three databases from their respective websites.

Our simple Python orthology inference program Porthos can be cloned from a separate dedicated GitHub repository: https://github.com/trgibbons/porthos.git.

4.7 Contributions

TRG implemented and ran all software, and authored manuscript. SMM & TRG initially conceived of project. TRG, SMM, EDC, and CFD refined study, figures, and text.

4.8 Acknowledgements

We thank Prof. Carl Kingsford for his assistance with brainstorming and development of earlier variants of this project. We also thank Dr. Bastian Bentlage for engaging discussion that inspired the eventual redirection of the original project. TRG was primarily funded by NIH Ruth L. Kirschstein National Research Service Award Institutional Research Training Grant T32 GM080201. Additional funding was provided by NSF awards 0629624, 1036506, and 1046075 to CFD.

5 Concluding remarks

The massively parallel short-read sequencing technologies that have finally made it possible to begin characterizing dinoflagellate genomics on a genomic scale, have also introduced a great deal of additional complexity into the interpretation of the results. The low-throughput sequencing technologies used for decades were capable of sequencing 30-60% of the average gene or transcript with only a single read. Modern short-read HTS technologies have unique weaknesses that can make it difficult to directly test hypotheses developed using these older technologies. This had led to apparent disagreement with several important open hypotheses. By leveraging the statistical power of HTS data, I have identified and characterized several sources of systematic error that result from straightforward analysis of short-read transcriptome data using popular tools, and where possible, have developed ways to infer more accurate estimates of the underlying biology.

I have shown that large families of conserved paralogs are systematically underreported by assembly of dinoflagellate transcriptomes from short-read data, but that these collapsed families of paralogs are detectable and abundant within a diverse set of free-living dinoflagellates, consistent with the hypothesis that dinoflagellate genomes have primarily expanded through the duplication of individual genes. Additionally, I showed that these assemblies often do not include the 5' ends of the transcripts, leading to apparent disagreement with the hypothesis that most dinoflagellates trans-splice most or their mature mRNA transcripts. I showed that the presence of relatively short 5' suffixes of the DinoSL are useful markers for assembly quality, and can be used to positively identify trans-spliced transcripts. Using this more sensitive cutoff, I found broad support for the hypothesis that trans-splicing is ubiquitous in dinoflagellates. I also found that

relict spliced leaders, believed to result from recycling of mature transcripts back into the genome, are abundant within the free-living dinoflagellates, consistent with the hypothesis that one of the primary mechanisms of gene duplication within dinoflagellates involves mature mRNA as an intermediate.

Despite the numerous problems identified within the assemblies, many high quality transcript sequences were recovered, including many divergent paralogs and (co)orthologs. Identifying these relationships required the use of a highly efficient and scalable automated clustering method, but algorithmic efficiency is often achieved at the cost of accuracy. I investigated the systematic errors of the Markov Clustering (MCL) algorithm, once of the more popular methods for clustering large sets of protein sequences, and found that it was capable of achieving very high sensitivity with the right settings, but suffers from an apparently unavoidable problem with false-positives.

I also compared popular metrics for weighting the input graph for MCL against each other, and against some simple alternatives. I found that MCL works equally well, and in some cases even slightly better, using a relatively simple edge-weighting metric that has been largely overlooked by the community in favor of more complex alternatives

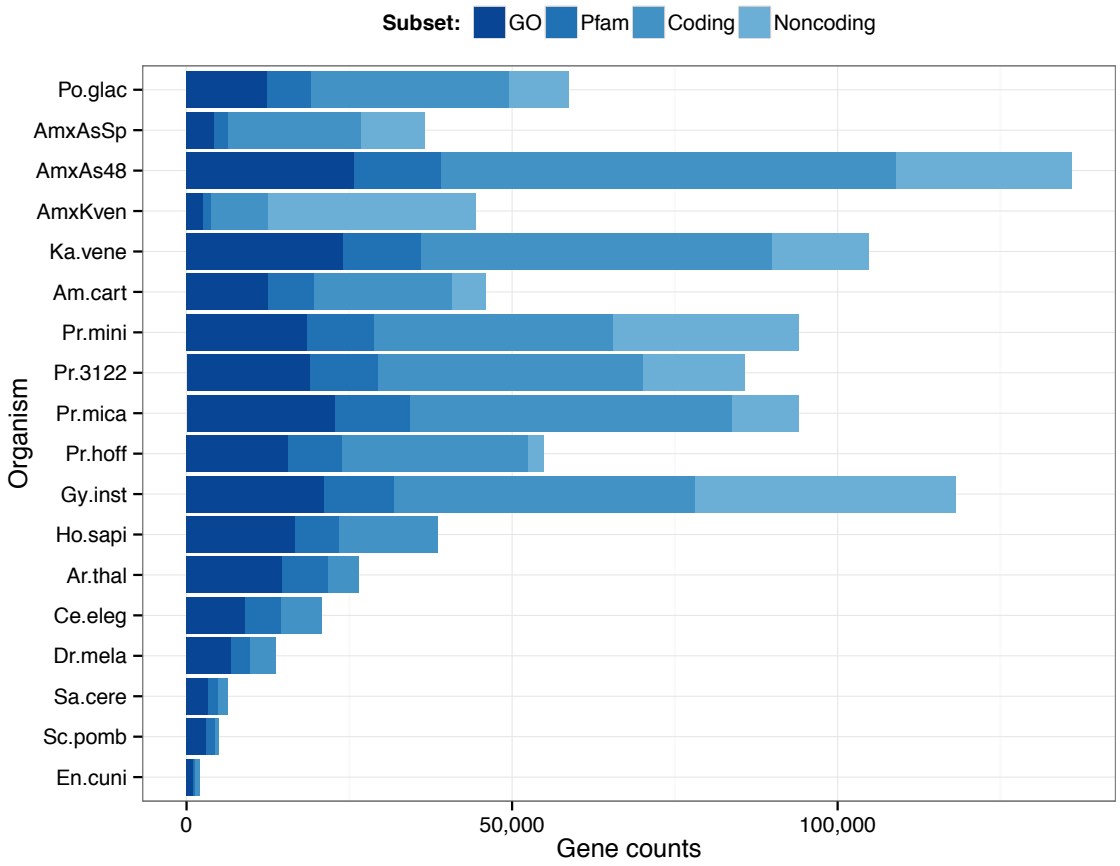
This is a very exciting time in genomics, and especially in dinoflagellate genomics. While genome sequencing and analysis of certain organisms has matured to the point of becoming routine, dinoflagellate genomics has been hindered by the unique and bizarre characteristics that also make them so fascinating. The first nearly complete dinoflagellate genome sequence has just been published. While it is from an endosymbiotic species and has a genome two orders of magnitude smaller than those of the larger free-living species, the genome does contain many of the hallmark features of

these larger organisms, such as rampant gene duplication and evidence of transcript recycling from the presence of relict DinoSL sequences upstream of 15% of the identified genes. This suggests that *S. kawagutii* could finally provide the community with a much-needed model for dinoflagellate genome structure and evolution.

Sequencing prices continue to fall, even as read lengths rapidly grow. The new Sequel instrument from Pacific Biosciences is capable of generating a million transcript-length reads in a single run, eliminating the need for transcriptome assembly for researchers able to invest a few thousand dollars per organism for most eukaryotes, including the smaller free-living dinoflagellates. As high-quality long-read HTS technologies become cheaper, it will become easier to focus more directly on the biology motivating these studies. Dinoflagellate genomics will likely move away from short-read HTS, and many transcriptomes will eventually be resequenced to eliminate any lingering complications from short-read assembly. Nevertheless, the results from these short-read studies are providing the first genome-wide glimpses of dinoflagellate genomic structure and evolution. As with the targeted Sanger and EST libraries of the past, the results of these short-read studies will shape dinoflagellate genomics research for years to come.

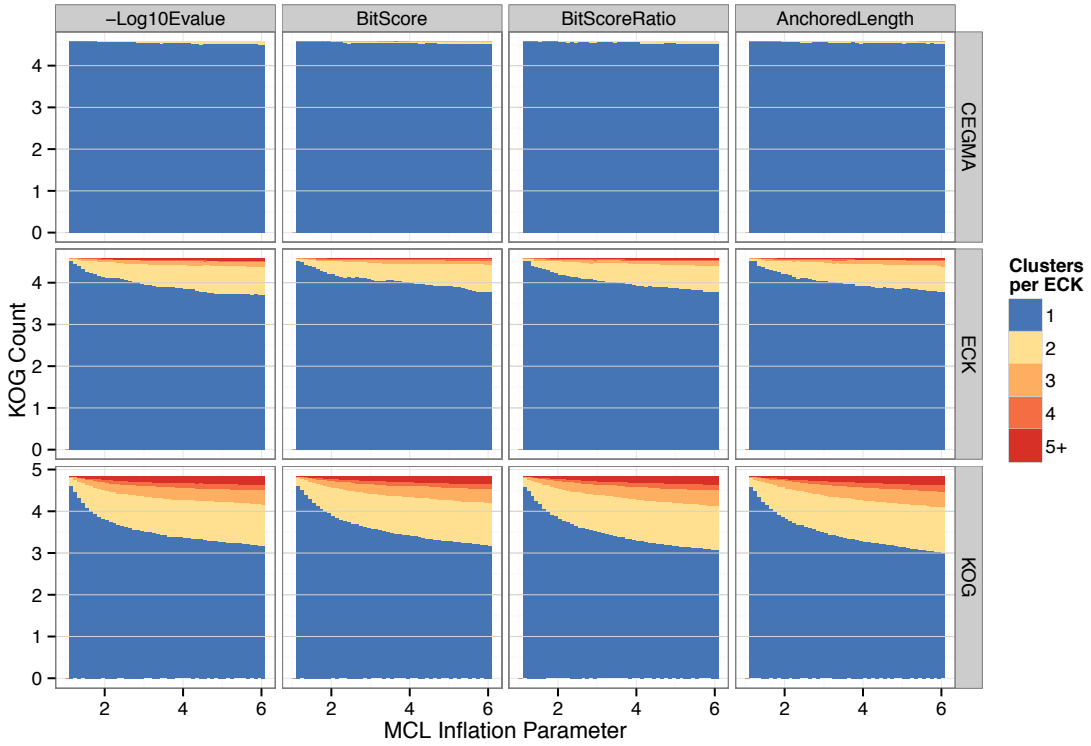
Appendix A – Supplemental figures

Figure S1



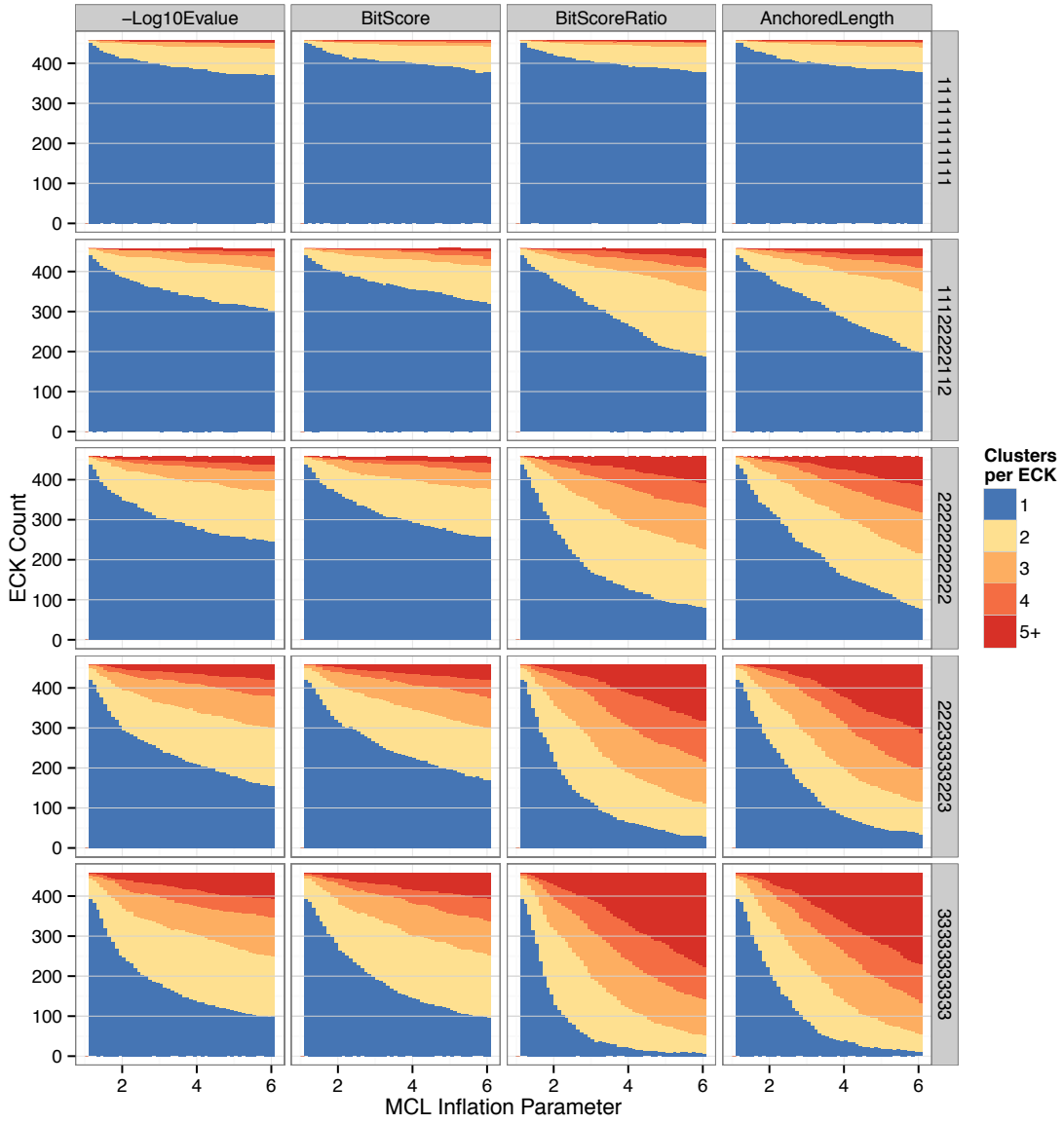
Gene counts by organism.
For the eleven novel dinoflagellate transcriptomes assembled using Trinity, bars indicate the number of genes predicted by Trinity. Colors indicate whether at least one isoform within the gene was predicted by TransDecoder to contain an ORF (Coding), at least one translated ORF sequence was annotated by TransDecoder to contain at least one Pfam-A domain (Pfam), at least one Pfam-A domain assigned to at least one ORF in at least one isoform was mapped to at least one GO term by the Pfam2GO mapping released by the GO consortium (GO), or if no isoforms within the gene were predicted to contain any ORFs (Noncoding). The well-curated proteomes of the seven KOG organisms were included to compare annotation statistics for protein sequences, so the proteins were used as a proxy for coding genes and no non-coding genes were included.

Figure S2



Sensitivity performance comparison for each test database.
Sensitivity performance on CEGMA (upper row), ECK (middle row), and KOG (lower row) databases. The sensitivity problems observed with the KOG database are not observed with the CEGMA database, but are restored with the ECK database. Plots are otherwise as described in Figure 21.

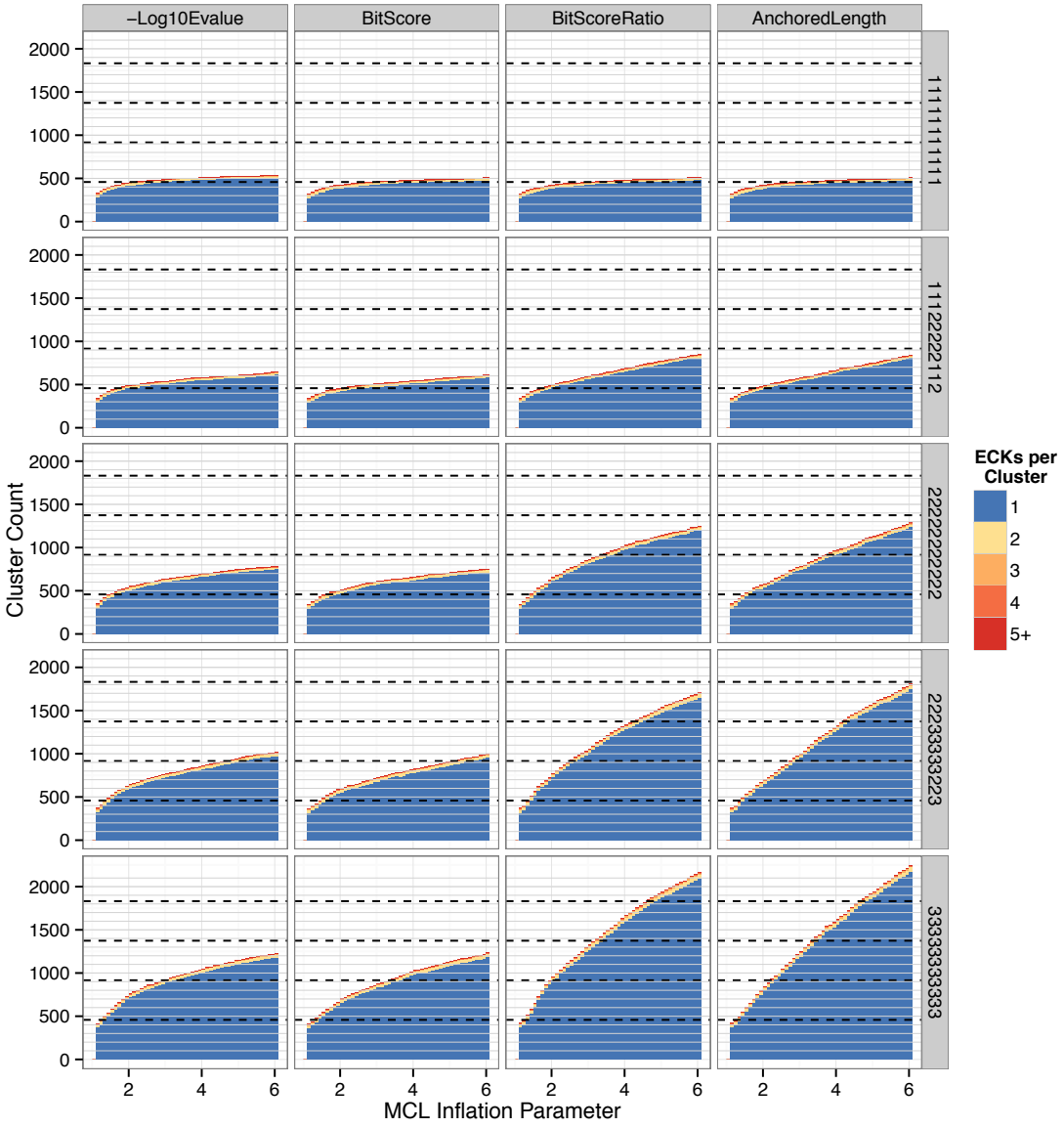
Figure S3



Sensitivity performance comparison with ordered application of the fragmentation scheme.

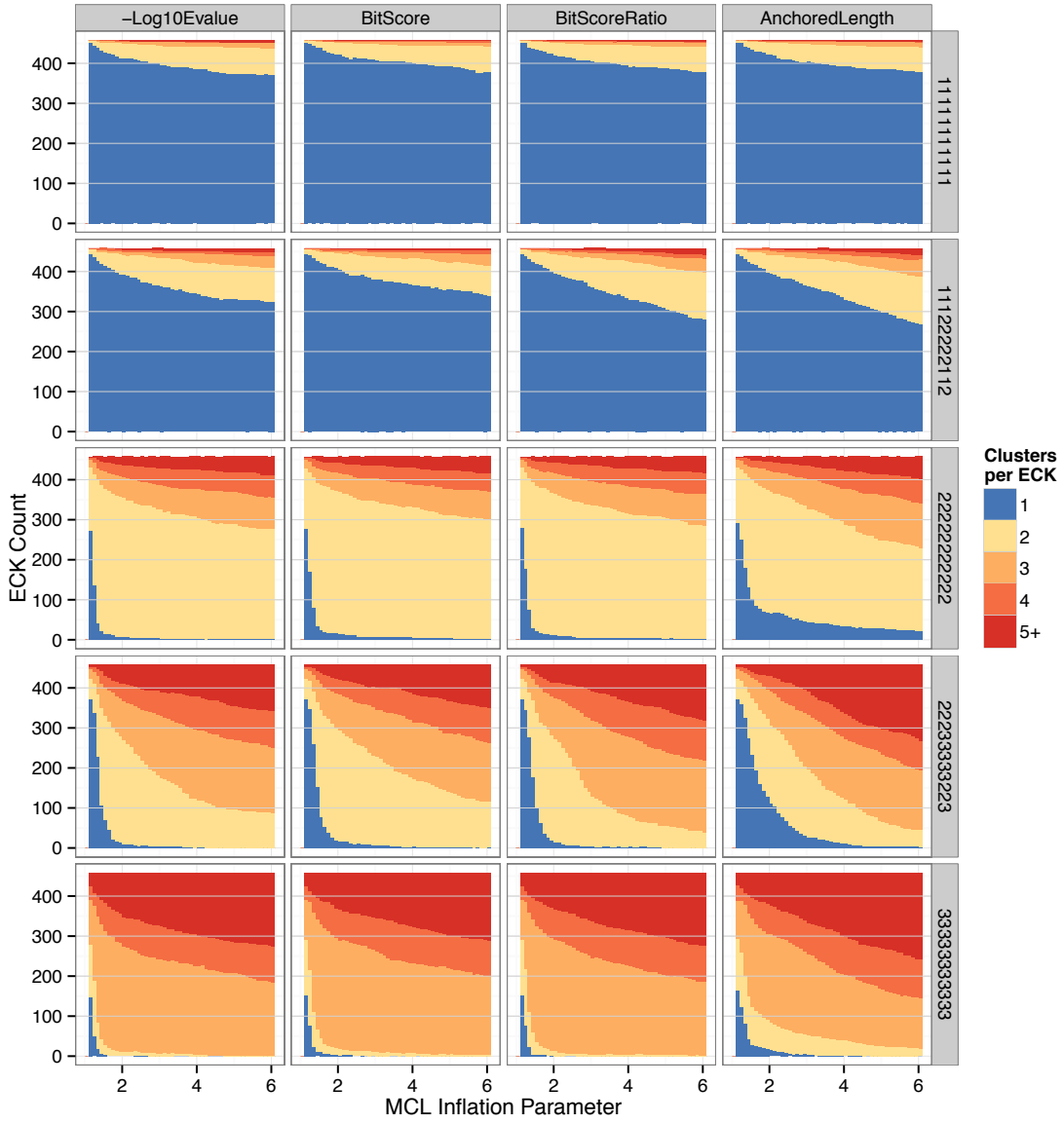
Plots were prepared identically to those in Figure 21, except that the fragmentation scheme was applied directly along organismal lines. Results are indistinguishable from those shown in Figure 21, indicating that no organism's sequences were more or less important to clustering.

Figure S4



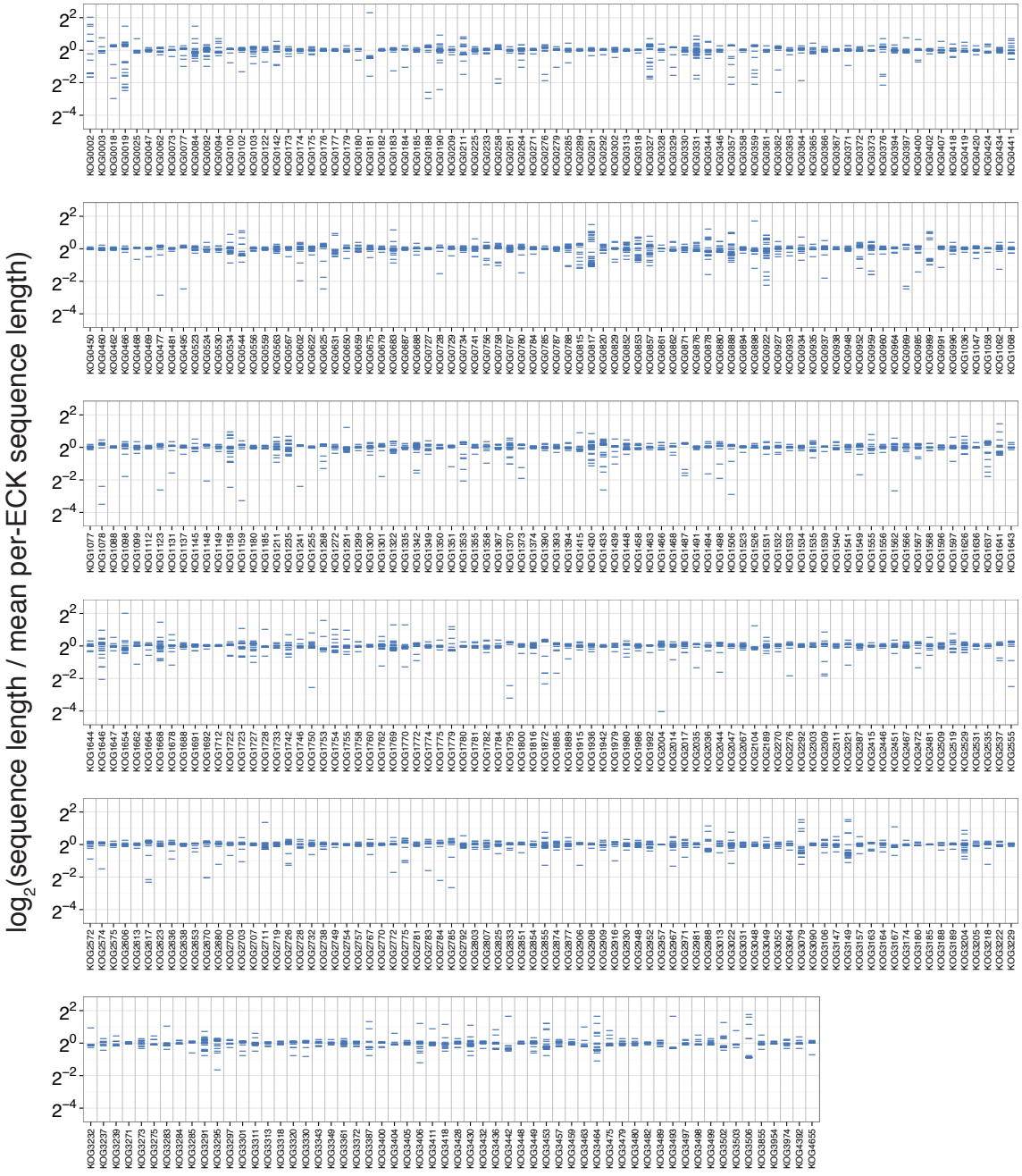
Specificity performance comparison with ordered application of the fragmentation scheme.
Plots were prepared identically to those in Figure 22, except that the fragmentation scheme was applied directly along organismal lines. Results are indistinguishable from those shown in Figure 22, indicating that no organism's sequences were more or less important to clustering.

Figure S5



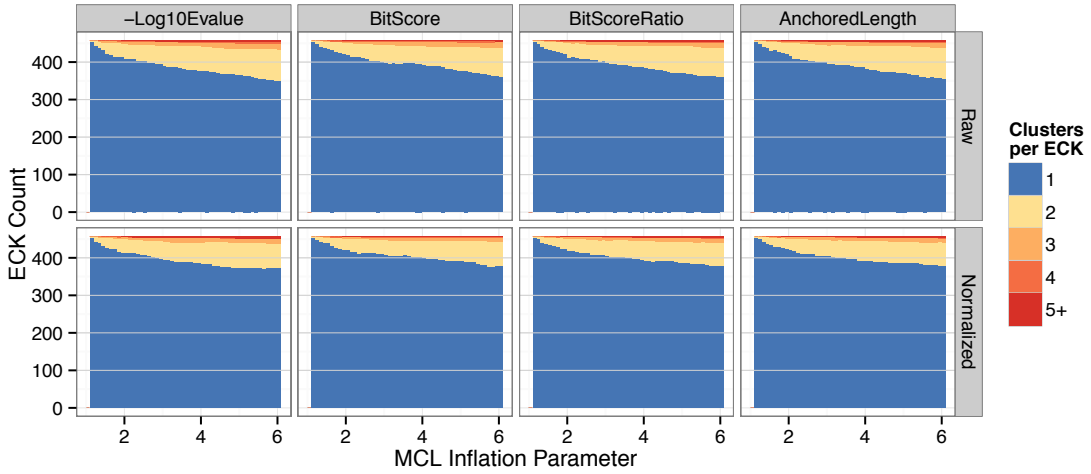
Sensitivity performance comparison with evenly fragmented sequences .
Plots were prepared identically to those in Figure 21, except that sequences were fragmented in equal pieces (i.e. halves or thirds). Results confirm that when all sequences are split into equal pieces, many breakpoints align, leaving minimal overlapping sequences. The consequent lack of high quality edges between halves or thirds of sequences makes it nearly impossible to recover complete clusters containing all fragments. The bit score over anchored alignment length (BAL) metric performs slightly better than they other metrics in this scenario as a result of successfully strengthening whatever weak edges do manage to cross these fragmentation boundaries.

Figure S6



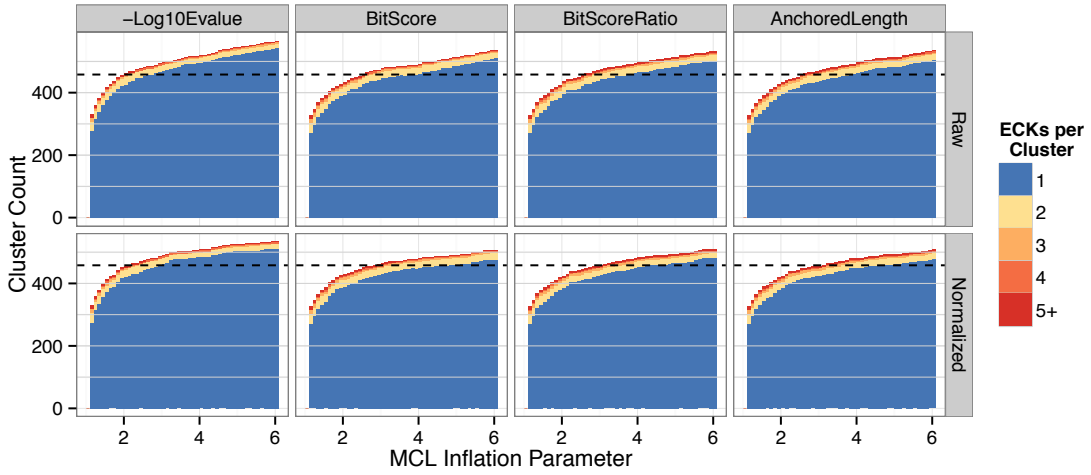
Sequence lengths relative to the mean for each ECK.
Sequence lengths divided by the mean length for each ECK (blue dashes). Sequences within the ECK database are organized using KOG identifiers, and are therefore sorted by these KOG identifiers along the x-axis. The y-axis has been transformed into a log2 scale to emphasize fold-changes relative to the mean sequence lengths. Most ECKs have very uniform sequence lengths, but some contain sequences with lengths varying from 1/8 to 4 times the corresponding mean. These differences provide an important variety of test cases when splitting the sequences evenly because they prevent BLAST-alignable overlaps within most, but not all, ECKs.

Figure S7



Sensitivity performance with and without inter-organism normalization. Sensitivity performance on the ECK database with all sequences intact before (upper row) and after (lower row) inter-organism normalization. Plots are otherwise as described in Figure 21. Close inspection reveals some small improvement from the normalization.

Figure S8



Specificity performance with and without inter-organism normalization. Specificity performance on the ECK database with all sequences intact before (upper row) and after (lower row) inter-organism normalization. Plots are otherwise as described in Figure 22. Close inspection reveals some small improvement from the normalization.

Appendix B – Supplemental tables

Table S1

Post-quality control Illumina sequencing statistics.

Sample	Read length	Fragments (m) PE*	Fragments (m) SE*	Data (Gbp)
Polarella glacialis†	78	31.2	2.6	4.7
Amoebophrya sp. ex *Akashiwo sanguinea* (dinospores)‡	101	46.6	2.4	9.0
Amoebophrya sp. ex *Akashiwo sanguinea* (48HPI)§	101	52.1	3.1	9.9
Amoebophrya sp. ex *Karlodinium veneficum*‖	101	70.2	4.9	13.3
Amphidinium carterae	101	25.9	3.4	4.5
Prorocentrum minimum	101	44.2	2.6	8.6
Prorocentrum sp. CCMP3122	101	48.9	4.7	8.6
Prorocentrum micans	101	45.0	4.5	7.9
Prorocentrum hoffmannianum	101	44.4	4.9	7.9
Gyrodinium instriatum¶	101	150.0	5.3	29.9
Totals	-	558.5	38.4	104.3

* All libraries were sequenced paired-end (PE), but the orphaned reads were treated as single-end (SE) reads after quality control (QC)

† Sequenced using the Illumina GAIIx platform instead of HiSeq

‡ Co-cultured with host; greatly enriched for parasite cells

§ Co-cultured with host

‖ Co-cultured with host; partially enriched for parasite cells; sequenced in two separate libraries

¶ Sequenced in six libraries after growth in five different conditions

Table S2

Numbers of Trinity isoforms and genes before and after filtering by TransDecoder minimum length.

	Trinity Isoforms				Trinity Genes			
Assembly	≥200nt	≥297nt	Lost	%Lost	≥200nt	≥297nt	Lost	%Lost
Po.glac	138,262	84,519	53,743	39	100,326	58,733	41,593	41
AmxAsSp	63,010	46,624	16,386	26	51,895	36,613	15,282	29
AmxAs48	204,082	163,554	40,528	20	173,753	135,948	37,805	22
AmxKven*	65,109	50,487	14,622	22	58,551	44,453	14,098	24
Ka.vene*	152,948	116,807	36,141	24	139,020	104,718	34,302	25
Am.cart	67,161	54,776	12,385	18	57,183	45,996	11,187	20
Pr.mini	157,617	120,554	37,063	24	126,400	94,059	32,341	26
Pr.3122	168,400	109,873	58,527	35	139,299	85,735	53,564	38
Pr.mica	182,219	139,598	42,621	23	131,539	94,036	37,503	29
Pr.hoff	90,791	73,607	17,184	19	70,051	54,838	15,213	22
Gy.inst	265,280	193,837	71,443	27	184,025	118,166	65,859	36
Totals	1,554,879	1,154,236	400,643	26	1,232,042	873,295	358,747	29

**Amoebophrya* sp. ex *K. veneficum* and its host were not sequenced and assembled separately. Instead, sequences from the combined assembly were separated based on average per-gene GC% (see chapter 2)

Table S3

CDS detection and annotation statistics for the DAToL database.

		Trinity			Pfam	
Assembly	Subset	Genes	Isoforms	Proteins	Domains	Unique
AmxAsSp	All	36,613	46,624	-	-	-
	Coding	26,819	34,454	39,953	-	-
	Pfam	6,340	8,524	9,011	17,988	2,217
	GO	4,201	5,702	5,950	9,722	1,118
AmxAs48	All	135,948	163,554	-	-	-
	Coding	109,058	132,053	144,759	-	-
	Pfam	39,041	48,043	48,901	110,593	3,960
	GO	25,726	31,905	32,292	56,154	1,879
AmxKven*	All	44,453	50,487	-	-	-
	Coding	12,505	14,839	18,998	-	-
	Pfam	3,800	4,508	5,022	9,054	1,817
	GO	2,542	3,025	3,324	5,065	949
Ka.vene*	All	104,718	116,807	-	-	-
	Coding	89,865	100,118	103,732	-	-
	Pfam	36,104	40,423	40,664	92,690	3,817
	GO	24,086	27,000	27,123	47,314	1,809
Am.cart	All	45,996	54,776	-	-	-
	Coding	40,837	48,308	51,044	-	-
	Pfam	19,562	22,850	23,005	53,760	3,481
	GO	12,537	14,531	14,590	25,840	1,706
Gy.inst	All	118,166	193,837	-	-	-
	Coding	78,220	140,439	175,148	-	-
	Pfam	31,942	58,766	61,915	145,785	4,092
	GO	21,057	38,620	40,258	70,268	1,929
Po.glac	All	58,733	84,519	-	-	-
	Coding	49,601	70,508	72,145	-	-
	Pfam	19,151	27,752	27,961	52,340	3,540
	GO	12,403	18,096	18,189	27,293	1,743
Pr.hoff	All	54,838	73,607	-	-	-
	Coding	52,465	70,195	78,467	-	-
	Pfam	23,983	31,355	31,704	75,777	3,610
	GO	15,637	20,472	20,619	36,150	1,742
Pr.mica	All	94,036	139,598	-	-	-
	Coding	83,791	123,113	142,658	-	-
	Pfam	34,424	50,418	50,910	114,127	3,788
	GO	22,874	33,914	34,146	58,815	1,799
Pr.mini	All	94,059	120,554	-	-	-
	Coding	65,538	81,863	84,070	-	-
	Pfam	28,856	36,834	37,083	84,303	3,747
	GO	18,516	23,818	23,923	40,782	1,788
Pr.3122	All	85,735	109,873	-	-	-
	Coding	70,171	89,810	102,380	-	-
	Pfam	29,347	37,691	37,969	80,595	3,551
	GO	19,030	24,625	24,725	40,949	1,720

Totals	All	873,295	1,154,236	-	-	-
	Coding	678,870	905,700	1,013,354	-	-
	Pfam	272,550	367,164	374,145	837,012	-
	GO	178,609	241,708	245,139	418,352	-

**Amoebophrya* sp. ex *K. veneficum* and its host were not sequenced and assembled separately. Instead, sequences from the combined assembly were separated based on average per-gene GC% (see chapter 2)

Table S4

Annotation statistics for the KOG proteomes.

			Pfam	
Assembly	Subset	Proteins	Domains	Unique
Ho.sapi	Coding	38,638	-	-
	Pfam	23,539	87,676	5,477
	GO	16,644	40,325	2,452
Ar.thal	Coding	26,406	-	-
	Pfam	21,755	62,009	3,998
	GO	14,655	30,760	1,929
Ce.eleg	Coding	20,751	-	-
	Pfam	14,557	34,944	3,791
	GO	8,949	16,805	1,836
Dr.mela	Coding	13,703	-	-
	Pfam	9,840	27,711	4,043
	GO	6,843	14,844	1,909
Sa.cere	Coding	6,387	-	-
	Pfam	4,917	9,818	3,101
	GO	3,381	5,734	1,548
Sc.pomb	Coding	5,035	-	-
	Pfam	4,434	9,294	3,088
	GO	2,997	5,244	1,522
En.cuni	Coding	2,000	-	-
	Pfam	1,377	2,719	1,126
	GO	945	1,675	668
Totals	Coding	112,920	-	-
	Pfam	80,419	234,171	-
	GO	54,414	115,387	-

**Amoebophrya* sp. ex *K. veneficum* and its host were not sequenced and assembled separately. Instead, sequences from the combined assembly were separated based on average per-gene GC% (see chapter 2)

Table S5

SNP counts by codon position within coding sequences.

	Codon position			
Assembly	1	2	3	Totals
Po.glac	101,911	84,571	227,956	414,438
AmxAsSp	33,024	33,933	44,251	111,208
AmxAs48	149,259	124,155	300,849	574,263
AmxKven*	4,830	3,969	11,131	19,930
Ka.vene*	25,948	18,989	82,469	127,406
Am.cart	28,023	22,611	55,888	106,522
Pr.mini	95,519	67,998	248,865	412,382
Pr.3122	227,112	154,500	385,564	767,176
Pr.mica	284,851	224,677	484,057	993,585
Pr.hoff	113,868	91,635	197,448	402,951
Gy.inst	140,511	112,518	303,286	556,315
Totals	1,204,856	939,556	2,341,764	4,486,176

**Amoebophrya* sp. ex *K. veneficum* and its host were not sequenced and assembled separately. Instead, sequences from the combined assembly were separated based on average per-gene GC% (see chapter 2)

Table S6

Counts of Trinity isoforms with detectable 5'-anchored DinoSL suffixes by suffix length and assembly.

	Suffix length																	
Assembly	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Po.glac	1,797	222	235	92	33	42	10	37	30	8	5	3	12	19	30	52	44	17
AmxAsSp	993	126	1649	305	182	114	42	275	277	685	49	10	21	61	71	173	163	85
AmxAs48	5,694	985	22158	4290	1,634	751	308	2,840	2,809	5,953	437	204	441	463	523	659	466	89
AmxKven*	849	140	1429	299	168	57	28	452	525	904	80	10	17	13	58	60	71	36
Ka.vene*	6,428	940	20146	2578	1,152	476	154	2,510	2,212	2,933	210	110	190	310	414	483	338	52
Am.cart	1,782	288	9203	2706	987	403	202	2,471	2,568	3,443	264	98	206	255	382	427	363	92
Pr.mini	3,239	501	12542	1906	398	194	125	1,266	1,151	711	51	30	107	63	21	41	62	38
Pr.3122	3,438	706	15629	2446	723	445	194	1,589	1,213	1,926	78	64	160	129	12	36	62	43
Pr.mica	6,245	1,335	16168	3450	745	437	305	2,314	1,703	773	108	193	677	341	101	167	174	50
Pr.hoff	2,466	775	19708	4371	1,124	620	228	1,855	1,831	1,557	63	100	163	186	50	66	103	53
Gy.inst	6,227	1,565	12107	5633	1,410	711	552	6,452	5,290	1,786	151	110	461	253	283	397	455	101
Totals	39,158	7,583	130,974	28,076	8,556	4,250	2,148	22,061	19,609	20,679	1,496	932	2,455	2,093	1,945	2,561	2,301	656

**Amoebophrya* sp. ex *K. veneficum* and its host were not sequenced and assembled separately. Instead, sequences from the combined assembly were separated based on average per-gene GC% (see chapter 2)

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., … Venter, J. C. (2000). The genome sequence of Drosophila melanogaster. *Science (New York, N.Y.)*, *287*(2185), 2185–95. doi:10.1126/science.287.5461.2185

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., … Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18. doi:10.1186/gb-2011-12-2-r18

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–10. doi:10.1016/S0022-2836(05)80360-2

Apeltsin, L., Morris, J. H., Babbitt, P. C., & Ferrin, T. E. (2011). Improving the quality of protein similarity network clustering algorithms using the network edge weight distribution. *Bioinformatics (Oxford, England)*, *27*(3), 326–33. doi:10.1093/bioinformatics/btq655

Bachvaroff, T. R., Gornik, S. G., Concepcion, G. T., Waller, R. F., Mendez, G. S., Lippmeier, J. C., & Delwiche, C. F. (2014). Dinoflagellate phylogeny revisited: using ribosomal proteins to resolve deep branching dinoflagellate clades. *Molecular Phylogenetics and Evolution*, *70*, 314–22. doi:10.1016/j.ympev.2013.10.007

Bachvaroff, T. R., & Place, A. R. (2008). From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate Amphidinium carterae. *PloS One*, *3*(8), e2929. doi:10.1371/journal.pone.0002929

Bachvaroff, T. R., Place, A. R., & Coats, D. W. (2009). Expressed sequence tags from Amoebophrya sp. infecting Karlodinium veneficum: comparing host and parasite sequences. *The Journal of Eukaryotic Microbiology*, *56*(6), 531–41. doi:10.1111/j.1550-7408.2009.00433.x

Band-Schmidt, C. J., Bustillos-Guzmán, J. J., López-Cortés, D. J., Gárate-Lizárraga, I., Núñez-Vázquez, E. J., & Hernández-Sandoval, F. E. (2010). Ecological and physiological studies of Gymnodinium catenatum in the Mexican Pacific: a review. *Marine Drugs*, *8*(6), 1935–61. doi:10.3390/md8061935

Beauchemin, M., Roy, S., Daoust, P., Dagenais-Bellefeuille, S., Bertomeu, T., Letourneau, L., … Morse, D. (2012). Dinoflagellate tandem array gene transcripts are highly conserved and not polycistronic. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(39), 15793–8. doi:10.1073/pnas.1206683109

Bentlage, B. (2014). Marine microbial eukaryotic transcriptomes. *Figshare*. doi:http://dx.doi.org/10.6084/m9.figshare.1152727

Bertomeu, T., & Morse, D. (2004). Isolation of a dinoflagellate mitotic cyclin by functional complementation in yeast. *Biochemical and Biophysical Research Communications*, *323*(4), 1172–83. doi:10.1016/j.bbrc.2004.09.008

Bhattacharya, D., Yoon, H. S., Hedges, S. B., & Hackett, J. D. (2009). Eukaryotes (Eukaryota). In S. B. Hedges & S. Kumar (Eds.), *The Timetree of Life* (pp. 116–120). Oxford Univ Press.

Boothroyd, J. C., & Cross, G. A. (1982). Transcripts coding for variant surface glycoproteins of Trypanosoma brucei have a short, identical exon at their 5' end. *Gene*, *20*(2), 281–9. doi:10.1016/0378-1119(82)90046-4

Bouligand, Y., & Norris, V. (2001). Chromosome separation and segregation in dinoflagellates and bacteria may depend on liquid crystalline states. *Biochimie*, *83*(2), 187–92. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11278068

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*, 421. doi:10.1186/1471-2105-10-421

Chance, B., Saronio, C., & Leigh, J. S. (1975). Functional intermediates in reaction of cytochrome oxidase with oxygen. *Proceedings of the National Academy of Sciences of the United States of America*, *72*(4), 1635–40. doi:10.1093/nar/gkg072

Chen, F., Mackey, A. J., Stoeckert, C. J., & Roos, D. S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, *34*(Database issue), D363–D368. Retrieved from papers://41b1743c-4b5e-4d8d-8de5-4d45b069b2db/Paper/p3219

Chen, F., Mackey, A. J., Vermunt, J. K., & Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS One*, *2*(4), e383. doi:10.1371/journal.pone.0000383

Coats, D. W., & Park, M. G. (2002). PARASITISM OF PHOTOSYNTHETIC DINOFLAGELLATES BY THREE STRAINS OF AMOEBOPHRYA (DINOPHYTA): PARASITE SURVIVAL, INFECTIVITY, GENERATION TIME, AND HOST SPECIFICITY. *Journal of Phycology*, *38*(3), 520–528. doi:10.1046/j.1529-8817.2002.01200.x

Consortium, C. elegans S. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science (New York, N.Y.)*, *282*(5396), 2012–8. doi:10.1126/science.282.5396.2012

Cooper, E. D., Bentlage, B., Gibbons, T. R., Bachvaroff, T. R., & Delwiche, C. F. (2014). Metatranscriptome profiling of a harmful algal bloom. *Harmful Algae*, *37*, 75–83. doi:10.1016/j.hal.2014.04.016

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., … Velayoudon, D. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1261605–1261605. doi:10.1126/science.1261605

Dehal, P., Satou, Y., Campbell, R. K., Chapman, J., Degnan, B., De Tomaso, A., … Rokhsar, D. S. (2002). The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science (New York, N.Y.)*, *298*(5601), 2157–67. doi:10.1126/science.1080049

Delwiche, C. (1999). Tracing the thread of plastid diversity through the tapestry of life. *American Naturalist*, *154*, 164–177. Retrieved from papers://41b1743c-4b5e-4d8d-8de5-4d45b069b2db/Paper/p3779

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195. doi:10.1371/journal.pcbi.1002195

Ekseth, O. K., Kuiper, M., & Mironov, V. (2014). orthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics (Oxford, England)*, *30*(5), 734–6. doi:10.1093/bioinformatics/btt582

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*(7), 1575–84. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=101833&tool=pmcentrez&rendertype=abstract

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, *42*(Database issue), D222–30. doi:10.1093/nar/gkt1223

Freudenthal, H. D., & Lee, J. J. (1963). Glenodinium halli n. sp. and Gyrodinium n. sp., dinoflagellates from New York Waters. *Journal of Protozoology*, *10*, 182–189.

Garcés, E., Fernandez, M., Penna, A., Van Lenning, K., Gutierrez, A., Camp, J., & Zapata, M. (2006). Characterization of nw mediterranean Karlodinium spp. (Dinophyceae) strains using morphological, molecular, chemical, and physiological methodologies. *Journal of Phycology*, *42*(5), 1096–1112. doi:10.1111/j.1529-8817.2006.00270.x

Gautier, A., Michel-Salamin, L., Tosi-Couture, E., McDowall, A. W., & Dubochet, J. (1986). Electron microscopy of the chromosomes of dinoflagellates in situ:

confirmation of Bouligand's liquid crystal hypothesis. *Journal of Ultrastructure and Molecular Structure Research*, *97*(1-3), 10–30. doi:10.1016/S0889-1605(86)80003-9

Gavrila, L. (1977). Cytogenetical Investigations in Mesokaryotic Algae I. the Nuclear Division, Chromosomes and the Tentative Karyotype. *Caryologia*, *30*(3), 273–288. doi:10.1080/00087114.1977.10796701

Gibbons, T. (2015). Dinoflagellate AToL transcriptomes - hidden paralogy. doi:doi:10.6084/m9.figshare.1600974

Gibbons, T. R., Mount, S. M., Cooper, E. D., & Delwiche, C. F. (2015). Evaluation of BLAST-based edge-weighting metrics used for homology inference with the Markov Clustering algorithm. *BMC Bioinformatics*, *16*(1), 218. doi:10.1186/s12859-015-0625-x

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., … Oliver, S. G. (1996). Life with 6000 genes. *Science (New York, N.Y.)*, *274*(5287), 546, 563–7. doi:10.1126/science.274.5287.546

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. a, Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–52. doi:10.1038/nbt.1883

Grossman, A. R., Harris, E. E., Hauser, C., Lefebvre, P. A., Martinez, D., Rokhsar, D., … Zhang, Z. (2003). Chlamydomonas reinhardtii at the crossroads of genomics. *Eukaryotic Cell*, *2*(6), 1137–50. doi:10.1128/EC.2.6.1137-1150.2003

Guillard, R. R. L., & Ryther, J. H. (1962). STUDIES OF MARINE PLANKTONIC DIATOMS: I. CYCLOTELLA NANA HUSTEDT, AND DETONULA CONFERVACEA (CLEVE) GRAN. *Canadian Journal of Microbiology*, *8*(2), 229–239. doi:10.1139/m62-029

Günzl, A. (2010). The pre-mRNA splicing machinery of trypanosomes: complex or simplified? *Eukaryotic Cell*, *9*(8), 1159–70. doi:10.1128/EC.00113-10

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–512. doi:10.1038/nprot.2013.084

Hackett, J. D., Anderson, D. M., Erdner, D. L., & Bhattacharya, D. (2004). DINOFLAGELLATES: A REMARKABLE EVOLUTIONARY EXPERIMENT. *American Journal of Botany*, *91*(10), 1523–1534. Retrieved from http://www.amjbot.org/content/91/10/1523.short

Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, *38*(12), e131. doi:10.1093/nar/gkq224

Hansen, P. J. (2011). The role of photosynthesis and food uptake for the growth of marine mixotrophic dinoflagellates. *The Journal of Eukaryotic Microbiology*, *58*(3), 203–14. doi:10.1111/j.1550-7408.2011.00537.x

Harrington, G. W., Beach, D. H., Dunham, J. E., & Holz, G. G. (1970). The Polyunsaturated Fatty Acids of Marine Dinoflagellates. *The Journal of Protozoology*, *17*(2), 213–219. doi:10.1111/j.1550-7408.1970.tb02359.x

Hastings, K. E. M. (2005). SL trans-splicing: easy come or easy go? *Trends in Genetics : TIG*, *21*(4), 240–7. doi:10.1016/j.tig.2005.02.005

Heidelberg, J. F., Paulsen, I. T., Nelson, K. E., Gaidos, E. J., Nelson, W. C., Read, T. D., … Fraser, C. M. (2002). Genome sequence of the dissimilatory metal ion-reducing bacterium Shewanella oneidensis. *Nature Biotechnology*, *20*(11), 1118–23. doi:10.1038/nbt749

Henderson, R. J., & Mackinley, E. E. (1991). Polyunsaturated fatty acid metabolism in the marine dinoflagellate Crypthecodinium cohnii. *Phytochemistry*, *30*(6), 1781–1787.

Herrera-Sepúlveda, A., Medlin, L. K., Murugan, G., Sierra-Beltrán, A. P., Cruz-Villacorta, A. a., & Hernández-Saavedra, N. Y. (2015). Are Prorocentrum hoffmannianum and Prorocentrum belizeanum (DINOPHYCEAE, PROROCENTRALES), the same species? An integration of morphological and molecular data. *Journal of Phycology*, *51*(1), 173–188. doi:10.1111/jpy.12265

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., … Hoffman, S. L. (2002). The genome sequence of the malaria mosquito Anopheles gambiae. *Science (New York, N.Y.)*, *298*(5591), 129–49. doi:10.1126/science.1076181

Hou, Y., & Lin, S. (2009). Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. *PloS One*, *4*(9), e6978. doi:10.1371/journal.pone.0006978

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., … Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Research*, *37*(Database issue), D211–5. doi:10.1093/nar/gkn785

Ignatiades, L., & Gotsis-Skretas, O. (2010). A review on toxic and harmful algae in Greek coastal waters (E. Mediterranean Sea). *Toxins*, *2*(5), 1019–37.

doi:10.3390/toxins2051019

Initiative, T. A. G. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796–815. doi:10.1038/35048692

Kalaitzis, J. a, Chau, R., Kohli, G. S., Murray, S. a, & Neilan, B. a. (2010). Biosynthesis of toxic naturally-occurring seafood contaminants. *Toxicon : Official Journal of the International Society on Toxinology*, *56*(2), 244–58. doi:10.1016/j.toxicon.2009.09.001

Katinka, M. D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., … Vivarès, C. P. (2001). Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature*, *414*(6862), 450–3. doi:10.1038/35106579

Keeling, P. J. (2004). Reduction and compaction in the genome of the apicomplexan parasite Cryptosporidium parvum. *Developmental Cell*, *6*(5), 614–616. doi:10.1016/S1534-5807(04)00135-2

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. a, … Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology*, *12*(6), e1001889. doi:10.1371/journal.pbio.1001889

Keeling, P. J., Corradi, N., Morrison, H. G., Haag, K. L., Ebert, D., Weiss, L. M., … Tzipori, S. (2010). The reduced genome of the parasitic microsporidian Enterocytozoon bieneusi lacks genes for core carbon metabolism. *Genome Biology and Evolution*, *2*(1), 304–309. doi:10.1093/gbe/evq022

Kingsford, C., Schatz, M. C., & Pop, M. (2010). Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformatics*, *11*, 21–32. Retrieved from papers://41b1743c-4b5e-4d8d-8de5-4d45b069b2db/Paper/p2411

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, *39*, 309–38. doi:10.1146/annurev.genet.39.073003.114725

LaJeunesse, T. C., Lambert, G., Andersen, R. A., Coffroth, M. A., & Galbraith, D. W. (2005). Symbiodinium (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates. *Journal of Phycology*, *41*, 880–886. doi:10.1111/j.1529-8817.2005.00111.x

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. doi:10.1038/35057062

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–9. doi:10.1038/nmeth.1923

Lee, D. H., Mittag, M., Sczekan, S., Morse, D., & Hastings, J. W. (1993). Molecular cloning and genomic organization of a gene for luciferin-binding protein from the dinoflagellate Gonyaulax polyedra. *The Journal of Biological Chemistry*, *268*(12), 8842–50. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8473328

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. doi:10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–9. doi:10.1093/bioinformatics/btp352

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–89. doi:10.1101/gr.1224503

Lidie, K. B., & van Dolah, F. M. (2007). Spliced leader RNA-mediated trans-splicing in a dinoflagellate, Karenia brevis. *The Journal of Eukaryotic Microbiology*, *54*(5), 427–35. doi:10.1111/j.1550-7408.2007.00282.x

Lin, S. (2011). Genomic understanding of dinoflagellates. *Research in Microbiology*, *162*(6), 551–69. doi:10.1016/j.resmic.2011.04.006

Lin, S., Cheng, S., Song, B., Zhong, X., Lin, X., Li, W., … Morse, D. (2015). The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis. *Science*, *350*(6261), 691–694. doi:10.1126/science.aad0408

Lin, S., Zhang, H., Zhuang, Y., Tran, B., & Gill, J. (2010). Spliced leader-based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(46), 20033–8. doi:10.1073/pnas.1007246107

Liu, L., & Hastings, J. W. (2006). NOVEL AND RAPIDLY DIVERGING INTERGENIC SEQUENCES BETWEEN TANDEM REPEATS OF THE LUCIFERASE GENES IN SEVEN DINOFLAGELLATE SPECIES. *Journal of Phycology*, *42*(1), 96–103. doi:10.1111/j.1529-8817.2005.00165.x

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England)*, *27*(6), 764–70. doi:10.1093/bioinformatics/btr011

McCutcheon, J. P., & Moran, N. a. (2011). Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, *10*(1), 13–26. doi:10.1038/nrmicro2670

Menden-Deuer, S., & Lessard, E. J. (2000). Carbon to volume relationships for dinoflagellates, diatoms, and other protist plankton. *Limnology and Oceanography*, *45*(3), 569–579. doi:10.4319/lo.2000.45.3.0569

Mendez, G. S., Delwiche, C. F., Apt, K. E., & Lippmeier, J. C. (2015). Dinoflagellate Gene Structure and Intron Splice Sites in a Genomic Tandem Array. *Journal of Eukaryotic Microbiology*, *62*(5), 679–687. doi:10.1111/jeu.12230

Meyer, J. M., Rödelsperger, C., Eichholz, K., Tillmann, U., Cembella, A., McGaughran, A., & John, U. (2015). Transcriptomic characterisation and genomic glimps into the toxigenic dinoflagellate Azadinium spinosum, with emphasis on polykeitde synthase genes. *BMC Genomics*, *16*(1), 1–14. doi:10.1186/s12864-014-1205-6

Moran, N. a. (2002). Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, *108*(5), 583–6. doi:10.1016/S0092-8674(02)00665-7

Neale, D. B., Wegrzyn, J. L., Stevens, K. a, Zimin, A. V, Puiu, D., Crepeau, M. W., … Langley, C. H. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, *15*(3), R59. doi:10.1186/gb-2014-15-3-r59

Osawa, S. (1995). *Evolution of the genetic code*. Oxford University Press on Demand.

Paccanaro, A., Casbon, J. A., & Saqi, M. A. S. (2006). Spectral clustering of protein sequences. *Nucleic Acids Research*, *34*(5), 1571–80. doi:10.1093/nar/gkj515

Park, M. G., Yih, W., & Coats, D. W. (2004). Parasites and phytoplankton, with special emphasis on dinoflagellate infections. *The Journal of Eukaryotic Microbiology*, *51*(2), 145–55. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15134249

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics (Oxford, England)*, *23*(9), 1061–7. doi:10.1093/bioinformatics/btm071

Patro, R., Duggal, G., & Kingsford, C. (2015). Salmon : Accurate , Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. *bioRxiv*, 1–18. doi:http://dx.doi.org/10.1101/021592

Preußer, C., Jaé, N., & Bindereif, A. (2012). mRNA splicing in trypanosomes. *International Journal of Medical Microbiology : IJMM*, *302*(4-5), 221–4. doi:10.1016/j.ijmm.2012.07.004

Rasko, D. A., Myers, G. S. A., & Ravel, J. (2005). Visualization of comparative genomic

analyses by BLAST score ratio. *BMC Bioinformatics*, *6*, 2. doi:10.1186/1471-2105-6-2

Reguera, B., Riobó, P., Rodríguez, F., Díaz, P. a, Pizarro, G., Paz, B., … Blanco, J. (2014). Dinophysis toxins: causative organisms, distribution and fate in shellfish. *Marine Drugs*, *12*(1), 394–461. doi:10.3390/md12010394

Remm, M., Storm, C. E., & Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, *314*(5), 1041–52. doi:10.1006/jmbi.2000.5197

Rivera, M. C., Jain, R., Moore, J. E., & Lake, J. A. (1998). Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(11), 6239–44. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=27643&tool=pmcentrez&rendertype=abstract

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–6. doi:10.1038/nbt.1754

Roth, M. S. (2014). The engine of the reef: photobiology of the coral-algal symbiosis. *Frontiers in Microbiology*, *5*(August), 422. doi:10.3389/fmicb.2014.00422

Rowan, R., Whitney, S. M., Fowler, A., & Yellowlees, D. (1996). Rubisco in marine symbiotic dinoflagellates: form II enzymes in eukaryotic oxygenic phototrophs encoded by a nuclear multigene family. *The Plant Cell*, *8*(3), 539–53. doi:10.1105/tpc.8.3.539

Ryan, D. E., Pepper, A. E., & Campbell, L. (2014). De novo assembly and characterization of the transcriptome of the toxic dinoflagellate Karenia brevis. *BMC Genomics*, *15*(1), 888. doi:10.1186/1471-2164-15-888

Sahl, J. W., Caporaso, J. G., Rasko, D. A., & Keim, P. (2014). The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. *PeerJ*, *2*, e332. doi:10.7717/peerj.332

Sano, J., & Kato, K. H. (2009). Localization and copy number of the protein-coding genes actin, alpha-tubulin, and HSP90 in the nucleus of a primitive dinoflagellate, Oxyrrhis marina. *Zoological Science*, *26*(11), 745–53. doi:10.2108/zsj.26.745

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)*, *27*(6), 863–4. doi:10.1093/bioinformatics/btr026

Schmieder, R., Lim, Y. W., Rohwer, F., & Edwards, R. (2010). TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics*, *11*, 341. doi:10.1186/1471-2105-11-341

Schnepf, E., & Elbrächter, M. (1992). Nutritional strategies in dinoflagellates: A review with emphasis on cell biological aspects. *European Journal of Protistology*, *28*(1), 3–24. doi:10.1016/S0932-4739(11)80315-9

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–504. doi:10.1101/gr.1239303

Shi, X., Zhang, H., & Lin, S. (2013). Tandem Repeats, High Copy Number and Remarkable Diel Expression Rhythm of Form II RuBisCO in Prorocentrum donghaiense (Dinophyceae). *PLoS ONE*, *8*(8), e71232. doi:10.1371/journal.pone.0071232

Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., … Satoh, N. (2013). Draft assembly of the Symbiodinium minutum nuclear genome reveals dinoflagellate gene structure. *Current Biology*, *23*(15), 1399–1408. doi:10.1016/j.cub.2013.05.062

Slamovits, C. H., Fast, N. M., Law, J. S., & Keeling, P. J. (2004). Genome Compaction and Stability in Microsporidian Intracellular Parasites. *Current Biology*, *14*(10), 891–896. doi:10.1016/j.cub.2004.04.041

Slamovits, C. H., & Keeling, P. J. (2008). Widespread recycling of processed cDNAs in dinoflagellates. *Current Biology : CB*, *18*(13), R550–2. doi:10.1016/j.cub.2008.04.054

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., & Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)*, *27*(3), 431–2. doi:10.1093/bioinformatics/btq675

Szilágyi, S. M., & Szilágyi, L. (2014). A fast hierarchical clustering algorithm for large-scale protein sequence data sets. *Computers in Biology and Medicine*, *48*, 94–101. doi:10.1016/j.compbiomed.2014.02.016

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V, … Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, *4*, 41. doi:10.1186/1471-2105-4-41

Tatusov, R. L., Koonin, E. V, & Lipman, D. J. (1997). A genomic perspective on protein families. *Science (New York, N.Y.)*, *278*(5338), 631–7. doi:10.1126/science.278.5338.631

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. doi:10.1093/bib/bbs017

Van der Ploeg, L. H., Liu, A. Y., Michels, P. A., De Lange, T., Borst, P., Majumder, H. K., … Van Boom, J. (1982). RNA splicing is required to make the messenger RNA for a variant surface antigen in trypanosomes. *Nucleic Acids Research*, *10*(12), 3591–604. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=320737&tool=pmcentrez&rendertype=abstract

van Dongen, S. M. (2000). Graph clustering by flow simulation. *PhD Dissertation*.

Veldhuis, M. J. W., Cucci, T. L., & Sieracki, M. E. (1997). Cellular Dna Content of Marine Phytoplankton Using Two New Fluorochromes: Taxonomic and Ecological Implications. *Journal of Phycology*, *33*(3), 527–541. doi:10.1111/j.0022-3646.1997.00527.x

Watkins, S. M., Reich, A., Fleming, L. E., & Hammond, R. (2008). Neurotoxic shellfish poisoning. *Marine Drugs*, *6*(3), 431–55. doi:10.3390/md20080021

Wetzel, J., Kingsford, C., & Pop, M. (2011). Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. *BMC Bioinformatics*, *12*(1), 95. doi:10.1186/1471-2105-12-95

Wilson, T., & Hastings, J. W. (1998). BIOLUMINESCENCE. *Annual Review of Cell and Developmental Biology*, *14*(1), 197–230. doi:10.1146/annurev.cellbio.14.1.197

Wisecaver, J. H., & Hackett, J. D. (2011). Dinoflagellate genome evolution. *Annual Review of Microbiology*, *65*, 369–87. doi:10.1146/annurev-micro-090110-102841

Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Lyne, R., Stewart, A., … Cerrutti, L. (2002). The genome sequence of Schizosaccharomyces pombe. *Nature*, *415*(6874), 871–80. doi:10.1038/nature724

Zhang, H., Campbell, D. a, Sturm, N. R., & Lin, S. (2009). Dinoflagellate spliced leader RNA genes display a variety of sequences and genomic arrangements. *Molecular Biology and Evolution*, *26*(8), 1757–71. doi:10.1093/molbev/msp083

Zhang, H., Hou, Y., & Lin, S. (2006). Isolation and characterization of proliferating cell nuclear antigen from the dinoflagellate Pfiesteria piscicida. *The Journal of Eukaryotic Microbiology*, *53*(2), 142–50. doi:10.1111/j.1550-7408.2005.00085.x

Zhang, H., Hou, Y., Miranda, L., Campbell, D. a, Sturm, N. R., Gaasterland, T., & Lin, S. (2007). Spliced leader RNA trans-splicing in dinoflagellates. *Proceedings of the*

National Academy of Sciences of the United States of America, *104*(11), 4618–23. doi:10.1073/pnas.0700258104

Zhang, H., & Lin, S. (2003). Complex gene structure of the form II RUBISCO in the dinoflagellate Prorocentrum minimum (Dinophyceae). *Journal of Phycology*, *39*(6), 1160–1171. doi:10.1111/j.0022-3646.2003.03-055.x

Zhang, H., & Lin, S. (2009). Retrieval of missing spliced leader in dinoflagellates. *PloS One*, *4*(1), e4129. doi:10.1371/journal.pone.0004129

Zhang, H., Zhuang, Y., Gill, J., & Lin, S. (2013). Proof that Dinoflagellate Spliced Leader (DinoSL) is a Useful Hook for Fishing Dinoflagellate Transcripts from Mixed Microbial Samples: Symbiodinium kawagutii as a Case Study. *Protist*, *164*(4), 510–527. doi:10.1016/j.protis.2013.04.002

Zhang, S., Sui, Z., Chang, L., Kang, K., Ma, J., Kong, F., … Ma, Q. (2014). Transcriptome de novo assembly sequencing and analysis of the toxic dinoflagellate Alexandrium catenella using the Illumina platform. *Gene*, *537*(2), 285–93. doi:10.1016/j.gene.2013.12.041

Zhang, Y., Zhang, S.-F., Lin, L., & Wang, D.-Z. (2014). Comparative Transcriptome Analysis of a Toxin-Producing Dinoflagellate Alexandrium catenella and Its Non-Toxic Mutant. *Marine Drugs*, *12*(11), 5698–718. doi:10.3390/md12115698