OPEN ACCESS Freely available online

PLoS one

Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes

Yubo Hou, Senjie Lin*

Department of Marine Sciences, University of Connecticut, Groton, Connecticut, United States of America

Abstract

The ability to predict gene content is highly desirable for characterization of not-yet sequenced genomes like those of dinoflagellates. Using data from completely sequenced and annotated genomes from phylogenetically diverse lineages, we investigated the relationship between gene content and genome size using regression analyses. Distinct relationships between $\log_{10}$-transformed protein-coding gene number (Y′) versus $\log_{10}$-transformed genome size (X′, genome size in kbp) were found for eukaryotes and non-eukaryotes. Eukaryotes best fit a logarithmic model, $Y' = \ln(-46.200+22.678X'$, whereas non-eukaryotes a linear model, $Y' = 0.045+0.977X'$, both with high significance ($p<0.001$, $R^2>0.91$). Total gene number shows similar trends in both groups to their respective protein coding regressions. The distinct correlations reflect lower and decreasing gene-coding percentages as genome size increases in eukaryotes (82%–1%) compared to higher and relatively stable percentages in prokaryotes and viruses (97%–47%). The eukaryotic regression models project that the smallest dinoflagellate genome (3×10^6 kbp) contains 38,188 protein-coding (40,086 total) genes and the largest (245×10^6 kbp) 87,688 protein-coding (92,013 total) genes, corresponding to 1.8% and 0.05% gene-coding percentages. These estimates do not likely represent extraordinarily high functional diversity of the encoded proteome but rather highly redundant genomes as evidenced by high gene copy numbers documented for various dinoflagellate species.

Citation: 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

Editor: Rosemary Jeanne Redfield, University of British Columbia, Canada

Received May 19, 2009; **Accepted** August 14, 2009; **Published** September 14, 2009

Funding: This work was supported by the NSF grants OCE-0452780 and EF-0626678. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: senjie.lin@uconn.edu

Introduction

An increasing amount of evidence supports a general positive correlation between gene content and genome size in prokaryotes and small eukaryotes, but whether this trend applies to all eukaryotes has been questioned and remains to be investigated [1–3]. As genome size can be measured easily, a robust correlation between gene content and genome size would provide a simple tool for predicting gene contents of not-yet sequenced genomes such as those of dinoflagellates. Dinoflagellates are one of the largest algal groups in the ocean, contributing significantly to oceanic primary production and coral reef building. Dinoflagellates are ecologically and economically important also because many of them form harmful algal blooms and even produce toxins. Among many unique characteristics, dinoflagellates possess unusually large genomes [4]. Although smaller genomes may occur in some yet unrecognized dinoflagellates [5], the typical dinoflagellate genomes are larger than most eukaryotes examined to date. The smallest documented dinoflagellate genomes are found in the coral reef symbiont *Symbiodinium* spp., ranging from 1.5 to 4.8 (average ~3) pg DNA per haploid genome [6], while the largest (250 pg DNA per haploid genome) is found in *Prorocentrum micans* [7]. Equivalent to $3–245\times10^6$ kbp per haploid genome, dinoflagellate genomes are about 1–77 fold that of the human haploid genome, and greater than any other algal groups (~$13–200\times10^3$ kbp) by a factor of hundreds to thousands [6–10]. It has been suggested that the large fraction of the dinoflagellate genomes are nonfunctional repeated DNA sequences [9,11–15]. How many genes are encoded in the genomes of these unicellular and seemingly simple organisms remains a question, which potentially bears significance on eukaryotic genome evolution. Information on gene contents of dinoflagellate genomes will allow researchers to gain understanding on how the large genomes favor or disfavor these organisms in their wide range of habitats.

Unfortunately, the infeasibility of sequencing these gigantic genomes with the current technology has hindered the progress in understanding dinoflagellate gene content. The next generation technologies such as 454, Solexa, or SOLiD™ are promising in reducing the enormous costs needed to sequence a dinoflagellate genome. However, the challenge in assembling the relatively short fragments is still insurmountable especially because in dinoflagellates many genes occur in numerous highly similar copies [16,17]. Predictably, it will not be so soon before a dinoflagellate genome can be completely sequenced and accurately assembled to give a correct gene count. Any indirect approach to provide gene content estimate is desirable presently.

Taking advantage of the rapidly growing genome sequence dataset, we analyzed the relationship between gene content and genome size in all sequenced life forms. We then used the resultant eukaryotic regression equations to estimate gene content for dinoflagellate genomes. In light of high gene copy numbers reported for various dinoflagellates, implications of the high gene

numbers and possible evolutionary mechanisms giving rise to the enormous genomes in this phylum is discussed.

Methods

Data collection

Data up to date by February 2009 were retrieved from the Reference Sequence (RefSeq) collection in the National Center for Biotechnology information (NCBI; http://www.ncbi.nlm.nih.gov), the Integrated Microbial Genomes (IMG) system in DOE Joint Genome Institute (JGI; http://img.jgi.doe.gov), and peer-reviewed publications (Supplemental Table S1). Dataset included total number of nucleotide base pairs (i.e. genome size), number of protein-coding genes, and total number of genes (including protein-coding, rRNA, and tRNA), gene-coding percentage (percent of DNA bases that codes for genes in a genome) for 55 completely sequenced eukaryotic genomes and 1055 non-eukaryotic genomes including prokaryotes (478 from bacteria and 60 from archaea), viruses (260), and organelles (231 from mitochondria and 26 from chloroplasts). For gene-coding percentage, only data published in peer-reviewed articles were used in the analysis as data from JGI included introns and other untranslated regions and significantly overestimated gene-coding percentage in large eukaryotic genomes (Supplemental Table S1). Incomplete or draft genome sequence data were excluded from this study to avoid potential errors.

Regression analyses and dinoflagellate gene content prediction

The genome size and gene number datasets were subject to Shapiro-Wilk and Kolmogorov-Smirnov normality tests using SPSS 15. When normality was violated, data were logarithmic-transformed. Regression analyses for logarithmic-transformed protein-coding (or total) gene number (dependent variables) versus log genome size (independent variable) were conducted using linear, logarithmic, and power regression models in SPSS 15. The intention was to seek an overall correlation for all genomes, but if it failed, to seek separate correlations for separate groups of genomes (e.g. eukaryotes and others). The different regression models were compared based on significance level and R^2, and the best-fit model was selected. The established regression models were then used to predict dinoflagellate gene number based on documented genome size data ($3–245\times10^6$ kbp). Dinoflagellate gene-coding percentages were estimated based on this formula: (total gene number x average gene length/genome size)×100%, where average gene length was approximated as 1.346 kbp, a value previously found highly conserved in eukaryots [18].

Results

Distinct correlations between genome size and gene content for eukaryotes and non-eukaryotes

In the dataset we collected, the sequenced eukaryotic genomes ranged from 373 to 3,175,581 thousand base pairs (kbp) in size, while the genomes of non-eukaryotes (including bacteria, archaea, viruses, mitochondria, and chloroplasts) were substantially smaller, i.e., 2.4–9949.9 kbp (or kilobases in the case of single-stranded viral DNA or RNA) (Figure 1A). Correspondingly, total gene numbers were higher in eukaryotes than in non-eukaryotes (Figure 1A). The Shapiro-Wilk and Kolmogorov-Smirnov normality tests showed that the eukaryotic and non-eukaryotic genome sizes and total gene number were not of normal distribution. Thus, logarithmic-transformed data were used in further analysis.

When the $\log_{10}$-transformed data of gene number were plotted against $\log_{10}$ genome size, two distinct relations appeared: eukaryotes in one and non-eukaryotes in the other, with markedly different slopes emerging from initial linear regressions (Fig. 2A). Therefore, further multi-model analyses were performed separately for these two groups. For non-eukaryotes, the linear regression model was best fit ($p<0.001$, highest R^2) among all the different models examined (Table 1). For eukaryotes, the $\log_{10}$-transformed data best fit a natural logarithmic (ln) regression model (Table 1, Figure 3). As the protein-coding gene number was generally very close to the total gene number in each genome, similar significant positive correlations were found for total gene numbers in both eukaryotic and non-eukaryotic genomes (Table 1), although only the protein-coding gene number is shown in the figures (Figure 2A, 3).

On the contrary, the gene-coding fraction of the genome, i.e., gene-coding percentage, showed a different trend against genome size than the gene number trend (Figure 1B, 2B). In eukaryotes, the gene-coding percentage declined from 81.6% to 1.2% as the genome size increased (Figure 2B, Supplemental Table S1). The gene-coding percentage in non-eukaryotes was generally higher (97%–47%) and varied markedly less with genome size (Figure 1B, 2B) than in eukaryotes. The only exceptions were the organellar genomes, which exhibited a substantially lower gene-coding percentage than prokaryotes and viruses, indicating disproportionate loss of coding sequences during organellar genome reduction.

Dinoflagellate gene content estimation

The high R^2 and low p values (<0.001) in the $\log_{10}$ gene number versus $\log_{10}$ genome size regression models (Table 1) suggested that the empirically derived correlations were highly significant and could be used to make valid predictions of gene numbers. As the smallest recognized dinoflagellate genome (3×10^6 kbp, in *Symbiodinium* spp.) falls within the range of genome sizes used to derive the eukaryotic correlation, the regression equation can be applied directly, which gave 38,188 protein-coding (40,086 total) genes per genome. For the largest documented dinoflagellate genome (245×10^6 kbp, in *P. micans*), the empirical regression equation needed to be extrapolated with the assumption that the same correlation holds for larger genomes. As a result, the gene number estimate was 87,688 protein-coding (92,013 total) genes (Figure 3). Based on the previously reported average eukaryotic gene length, 1.346 kbp [18], these gene number estimates corresponded to 1.80% and 0.05% respectively for the smallest and the largest dinoflagellate genomes (Figure 2B).

Discussion

Distinction and robustness of regression models

Statistical analyses on up-to-date sequenced genome data show the lack of a universal correlation covering all life forms, in agreement with previous studies [1–3]. Our results further present evidence, for the first time, of an overall correlation in eukaryotic genomes between $\log_{10}$ gene number and $\log_{10}$ genome size. The best-fit regression model for $\log_{10}$-transformed eukaryote data is a $\log_e$ function and that for $\log_{10}$-transformed non-eukaryote data is a linear function, two distinct relationships. This indicates that as genome size increases the number of genes increases at a disproportionately slower rate in eukaryotes than in non-eukaryotes. In another word, the proportion of non-coding DNA increases with genome size faster in eukaryotes than in non-eukaryotes. This is consistent with the previous findings that the vast majority of nuclear DNA in eukaryotes is non-gene-coding

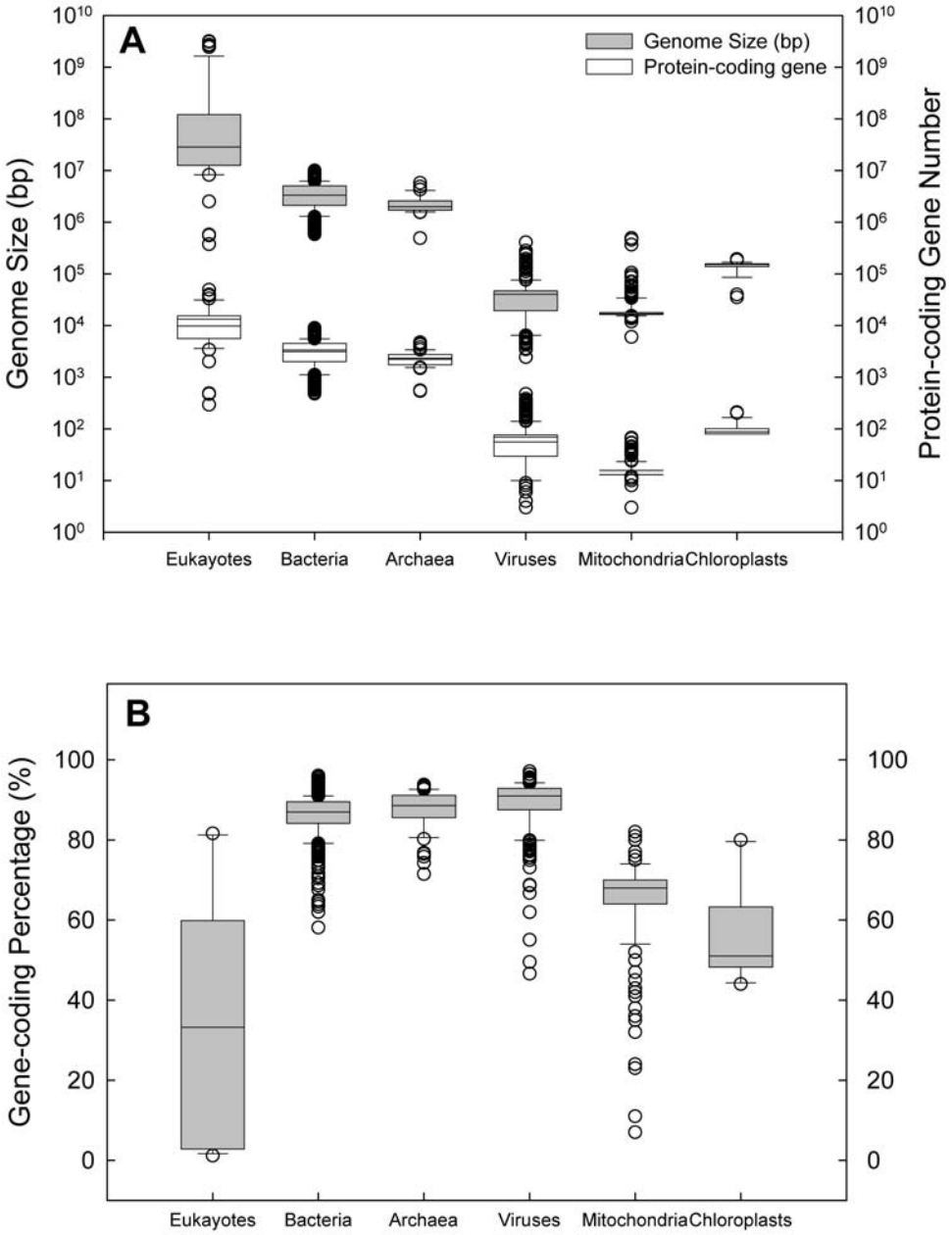

Figure 1. Genome sizes, protein-coding gene numbers, and gene-coding percentages of eukaryotic, bacterial, archaea, viral, and organellar genomes. (A) Genome size (shaded boxes) and number of protein-coding genes (open boxes). Total gene number is very close to protein-coding gene number and is not shown here. (B) Genome gene-coding percentage (fraction of DNA that constitutes genes). The lower and upper boundaries of the box indicate the first and third quartiles (or 25th and 75th percentiles) of each dataset, and the middle line in the box indicates the median value. The whiskers above and below the box indicate the 90th and 10th percentiles.
doi:10.1371/journal.pone.0006978.g001

elements including introns, pseudogenes, and transposable elements whereas prokaryotic, viral, and organellar genomes are mostly composed of gene-coding sequences [1,3].

The smallest eukaryotic genomes collected in this study are from the nucleomorphs of *Bigelowiella natans* (373 kbp), *Guillardia theta* (551 kbp), and *Hemiselmis andersenii* (572 kbp) followed by the parasitic fungus *Encephalitozoon cuniculi* (2,500 kbp). Their gene numbers and genome sizes are comparable to some bacteria (Figure 2). The nucleomorph is a remnant nucleus of the secondary endosymbiont that has evolved to a chloroplast in the host crytophyte and chlorarachniophyte algae [19]. While the counterparts in other lineages of algae have been completely lost, nucleomorphs in these two lineages remain, but the sizes of their genomes have remarkably reduced. For *E. cuniculi*, its small genome may be a result of selection for a minimal genome size in parasitism evolution. Gene numbers of these small eukaryotic genomes appear to also fit on the non-eukaryotic regression lines (Fig. 2A), suggesting that nuclear genome reduction during chloroplast and parasitism evolution has resulted in elevated gene density. This is the reverse of genome expansion that results from disproportionate increase of non-gene-coding DNA [1,3]. The two largest eukaryotic genomes analyzed were about 3,175,581 kbp in

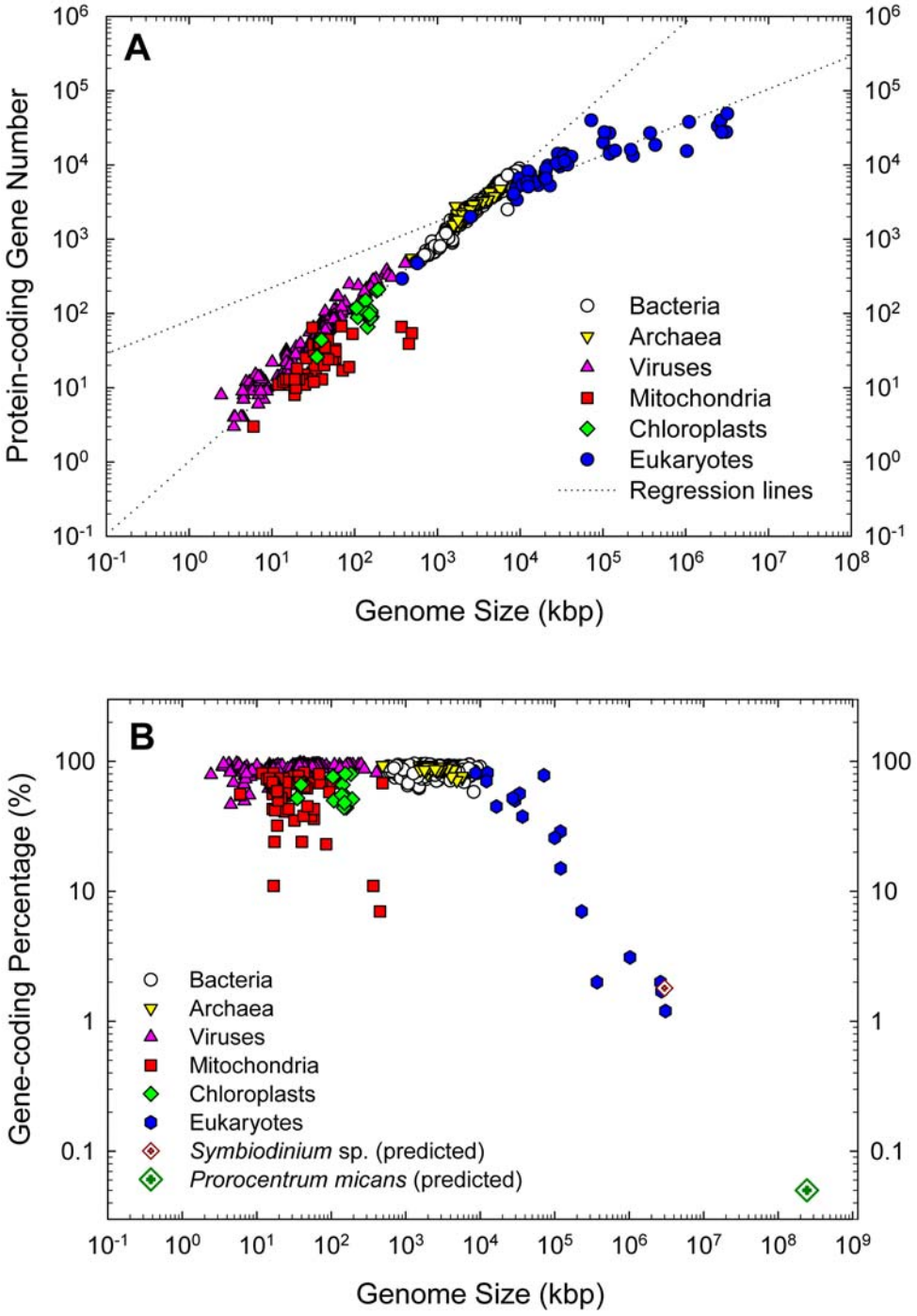

Figure 2. Distinct relationships between genome features in sequenced eukaryotes and non-eukaryotes. All correlations were highly significant ($p<0.001$). (A) Protein-coding gene number vs. genome size regression lines on log scale. Separate regression lines were yielded for eukaryotes (blue circles) and the non-eukaryotes (prokaryotes, viruses, and organelles; other symbols). (B) Gene-coding percentage vs. genome size on log scale. Note the negative trend for the eukaryotic genomes. The projected gene-coding percentage for the smallest (*Symbiodinium* sp., 1.80%) and largest dinoflagellate (*Prorocentrum micans*, 0.05%) genomes calculated based on reported average eukaryotic gene length (1.346 kbp) are shown for comparison. The trend for the non-eukaryotes is almost horizontal except for the outliers from some organelles.
doi:10.1371/journal.pone.0006978.g002

the primate *Pan troglodytes* and 3,080,436 kbp in humans, 8,514 times larger than the smallest (*B. natans* nucleomorph). Genome sequencing probably has biased toward relatively small genomes, as indicated by limited number of sequenced genomes larger than humans'; however, the current dataset cover a wide genome size, phylogenetic, and ecological ranges. The high statistical significance and R^2 value of the $\log_{10}$ gene number- $\log_{10}$ genome size correlation derived from this dataset suggests that the resultant regression equation should provide reliable predictions on gene numbers for many species.

Table 1. Summary of regression models with best fit models for each group italicized.

Model[a]	Regression equation[b]	R^2	Estimated gene	dinoflagellate numbers
			Smallest (3×10^6 kbp)	Largest (245×10^6 kbp)
Eukaryotes				
Protein coding genes (n = 55)				
linear	$y' = 1.902+0.445x'$	0.795		
logarithmic	*$y' = ln(-46.20+22.22x')$*	*0.919*	*38188*	*87688*
power	$\ln(y') = 1.629+0.583\ln(x')$	0.853		
Total genes (n = 48)				
linear	$y' = 1.802+0.470x'$	0.857		
logarithmic	*$y' = ln(-47.28+22.74x')$*	*0.924*	*40086*	*92013*
power	$\ln(y') = 1.602+0.597\ln(x')$	0.900		
Non-eukaryotes				
Protein coding genes (n = 1051)				
linear	*$y' = 0.045+0.977x'$*	*0.984*		
logarithmic	$y' = 0.840+2.051\ln(x')$	0.954		
power	$\ln(y') = 1.012+0.980\ln(x')$	0.963		
Total genes (n = 1051)				
linear	*$y' = 0.379+0.884x'$*	*0.987*		
logarithmic	$y' = 1.096+1.855\ln(x')$	0.958		
power	$\ln(y') = 1.227+0.828\ln(x')$	0.968		

[a]$p<0.001$ for all models; n = sample size.
[b]$y' = \log_{10}$ gene number (y, protein coding or total gene number) and $x' = \log_{10}$ genome size (x, in kbp).
doi:10.1371/journal.pone.0006978.t001

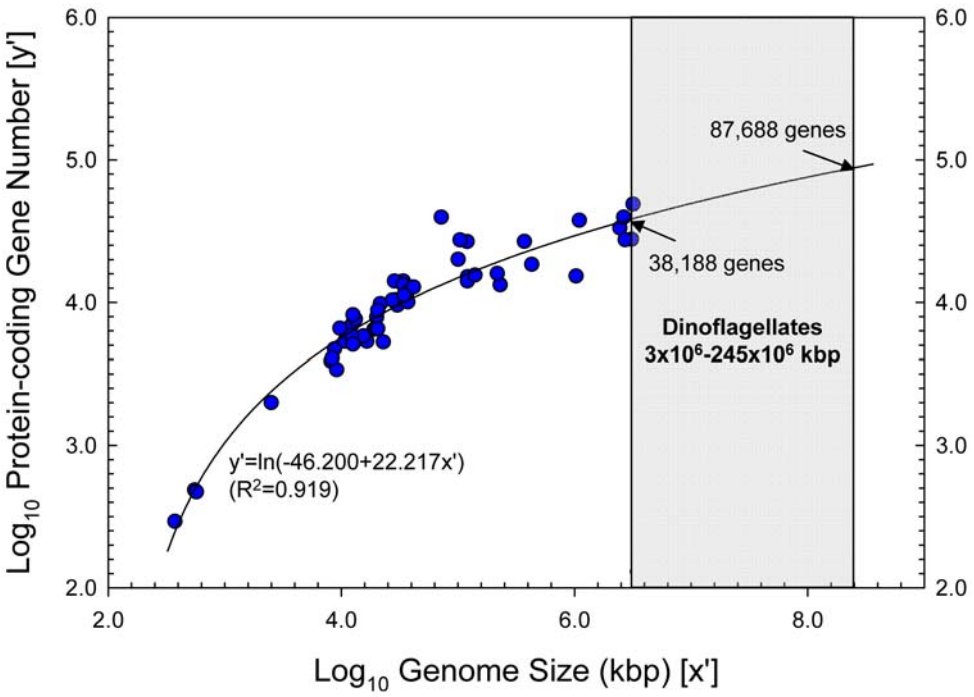

Figure 3. Logarithmic regression model for $\log_{10}$-transformed eukaryotic gene number (y') versus $\log_{10}$-transformed genome size (x'). Range of dinoflagellate genome size (3×10^6–245×10^6 kbp) is indicated by the shaded areas. The predicted gene numbers for the recognized smallest (38,188) and largest (87,688) dinoflagellate genomes correspond to their gene-coding percentages shown in Fig. 2B.
doi:10.1371/journal.pone.0006978.g003

Predicting power of the eukaryotic regression model for dinoflagellate genomes

A question about applying the eukaryotic regression model to dinoflagellate genomes stems from potential effects of distinct dinoflagellate genome organization on the $\log_{10}$ gene number-$\log_{10}$ genome size correlation. Unique among eukaryotes, dinoflagellate genomes have a few to over 200 chromosomes, which are permanently condensed, and not organized by nucleosomes [20]. The condensed chromosomes show a striating banding pattern under electron microscope that result from liquid cholesteric DNA crystal, which are formed by stacked disks of parallel bundles of DNA filaments that make a continuous left-handed twist along the chromosome's longitudinal axis [21]. Histone-like basic DNA-binding proteins are probably involved in stabilizing this structure by neutralizing local electronegative charges that would result from tightly compacted DNA filaments [22]. While most of this DNA is believed to be transcriptionally inactive, at the periphery of these disks are loops of DNA that are less tightly compacted and actively transcribed [23,24]. As mentioned earlier, most of the dinoflagellate genes studied so far are organized in tandem repeats, not so commonly seen in eukaryotes. Dinoflagellate genomes also host complex molecular machinery of mRNA editing [25] and spliced leader (SL) *trans*-splicing [26 and ref therein].

While no information is available to prove whether these genomic features will lead to alteration of the $\log_{10}$ gene number - $\log_{10}$ genome size relationship, an examination on organisms sharing similar genomic features may provide some clue. Genomes of the kinetoplastids, which are phylogenetically distinct from dinoflagellates, share with dinoflagellates many of the unique genomic features, such as permanently condensed chromosomes, gene tandem repeat organizations, mRNA editing, and SL *trans*-splicing of transcripts [27]. Genomes of two kinetoplastid species, *Leishmania major* (32,800 kbp) and *Trypanosoma brucei* (26,000 kbp), have been sequenced, but data were not used in the regression analyses because the sequence annotation had not been finished at time of our data collection. The total gene numbers based on the draft genome sequences are 9,183 for *L. major* and 9,068 for *T. brucei* [28,29], which are similar to what our eukaryotic regression model predicts (10,301 and 9,346, respectively). This comparison result indicates that the unique genome structures in this lineage will not cause significant deviation of genome features from the eukaryotic $\log_{10}$ gene number- $\log_{10}$ genome size relationship we have derived. It suggests that the relationship very likely holds for dinoflagellate genomes, particularly those of *Symbiodinium* spp. ($\sim 3\times10^6$ kbp), which are within the genome size range sampled in this study. The genomes of *Symbiodinium* spp. and some other modern dinoflagellates are shown to be haploid [30–35]. If polypoidy occurs in some dinoflagellates and accounts for their large nuclear genomes (see next section), practically gene contents in these species can also be estimated with their factored-down "haploid" genome sizes (if $\leq 3\times10^6$ kbp) using the regression equation developed here and the gene number estimate can then be factored up to the actual genome size. The equation can also be used to estimate the gene numbers for those having smaller genome size than *Symbidinium* spp. but yet to be identified [5].

Extrapolation of the regression model to accommodate genomes larger than sampled will have risk of overestimating or underestimating gene numbers, because the trend of the regression may possibly shift for large genomes like those of dinoflagellates. However, compared to a linear regression, the logarithmic regression we derived for eukaryotes inherently predicts a slower increase of gene number, and hence a progressively lower gene-coding percentage, as genome size increases. In fact, the predicted gene-coding percentages for the smallest and the largest dinoflagellate genome, 1.80% and 0.05% respectively, are remarkably lower than those for most other eukaryotes (1%–82%). Therefore, further leveling off of the regression line may not be so likely. A recent small-scale survey of *Heterocapsa triquetra* nuclear genome [36] is worth noting. Out of a 230 kbp sequence analyzed, 89.5% was non-repeated sequences with no similarity to any known genes but a 546-bp gene was identified. Applying the one per 230 kbp DNA gene density to the entire genome would yield about 91,500 genes for the $18.6–23.6\times10^6$ (21.1×10^6 on average) kbp *H. triquetra* nuclear genome. Alternatively, if we assume that the gene-coding percentage of this 230-kbp DNA (0.2%) and the previously reported eukaryotic average gene length (1.346 kbp) apply to this genome, the gene number would be 31,352. Our model-predicted 60,128 gene number for this species lies in the middle of the two extremes. Therefore, it seems unlikely that the eukaryote regression model we derived will seriously if at all overestimate gene numbers for large dinoflagellate genomes.

Dinoflagellate gene contents and their implications in genome evolution

While all the available information point to a reasonable accuracy, or at least no overestimation, the model-predicted gene numbers for dinoflagellates (38,188–87,688 or about 1-3 fold as many as that in a human genome) are exceedingly high for these unicellular and therefore relatively "simple" organisms. However, these gene number estimates may not really represent an extraordinarily high functional diversity of the encoded proteome. A survey of literature reveals that previously examined dinoflagellate genes occur in 30–5,000 copies per genome (Table 2), indicating that high gene copy number is a widespread phenomenon in dinoflagellate genomes. The sequences of these gene copies may be identical in some cases like the rRNA locus but slightly different from each other in most cases. Regardless, the widespread gene duplicates may offset the high total protein-coding gene numbers, giving a reasonable number of unique genes compared to what is expected of a typical unicellular eukaryote.

While little genomic data are available to support this proposition, some insights can be obtained from EST data that have been generated for several dinoflagellate species. Typically in these studies EST sequences in each species were clustered at an identity cutoff around 95%, which is expected to group cDNA copies into unique (or semi-unique) transcripts. In *Alexandrium tamarense* (genome size 200×10^6 kbp), 6,723 unique transcripts were identified out of a 11,171-EST dataset [37]; in *Heterocapsa triquetra* (about 20×10^6 kbp), 2,022 unique clusters were assembled out of 6,765 sequenced ESTs [38]; in *Karenia brevis* (about 100×10^6 kbp), 11,937 unique out of 25,000 ESTs [39]; in *K. veneficum* (formerly *K. micrum*; 5×10^6 kbp), 11,903 unique out of 16,544 [40]; in *Oxyrrhis marina* (genome size unknown), 9,876 unique out of 18,012 [41]. True unique-gene numbers of these species likely are higher than these unique-transcript numbers because an EST dataset does not include genes not expressed at time of sampling, and furthermore, as the sequencing scales in these projects were relatively small the data likely only account for a fraction of the expressed gene pool missing those expressed at lower levels. Nevertheless, these incomplete EST data reveal a minimum of nearly 12,000 unique genes even for the relatively small dinoflagellate genome of *K. veneficium* ($\sim 5\times10^6$ kbp). In this case, if the average gene copy number is 3, the 42,770 protein-coding genes predicted by our regression model would represent a collection of 14,257 unique genes, a number close to the EST-based unique gene estimate (>12,000).

Table 2. Dinoflagellate gene copy numbers documented to date.

Gene	Species	Copy number per genome**	Reference
Actin*	*Amphidinium carterae*	≥113	[49]
Protein kinase gene	*Lingulodinium polyedrum*	~30	[50]
Form II Rubisco gene*	*Prorocentrum minimum*	148	[16]
Luciferase gene*	*Alexandrium affine*	60	[51]
	Alexandrium tamarense	126	[51]
	Lingulodinium polyedrum	146	[51]
	Pyrocystis fusiformis	44	[51]
	Pyrocystis lunula	160	[51]
	Pyrocystis noctiluca	110	[51]
	Protoceratium reticulatum	48	[51]
Luciferin-binding protein gene	*Lingulodinium polyedrum*	~1,000	[52]
Mitotic cyclin gene	*Lingulodinium polyedrum*	~5,000	[53]
Peridinin-chlorophyll *a* binding protein gene*	*Lingulodinium polyedrum*	~5,000	[54]
	Symbiodinium sp. 203	36	[55]
Proliferating cell nuclear antigen gene*	*Pfiesteria piscicida*	41	[17]

*arranged in tandem repeats.
**A. *carterae* actin copy number was based on cloning and sequencing (Figure 4 in [49]); all other gene copy numbers here were based on probe hybridization or quantitative PCR.
doi:10.1371/journal.pone.0006978.t002

Many questions remain regarding dinoflagellate genome composition and its evolution. As the gene-coding percentage is very low, the large and widely ranged dinoflagellage genome sizes are clearly not due to the high gene numbers we predicted here. Non-coding DNA (e.g. repetitive sequences, introns, transposons) dominates the genomes as in any large eukaryote genomes, attested to by the abundant transposable elements found in a small fraction of *H. triquetra* genomic DNA [36]. On the contrary, the high gene numbers, especially high gene copy numbers, is likely the result of genome expansion. It is believed that dinoflagellate genomes have been subject to duplications of individual genes or segmental to whole genome duplication [5,39], or combinations of these mechanisms. Tandem-repeated genes, like those that have been studied in dinoflagellates (Table 2), are more likely to have resulted from successive gene duplications through unequal cross-over of chromosomes [16]. In addition, it is possible that dinoflagellate genomes can take up and incorporate cDNAs, resulting in multiplication of genes such as that coding for SL [42]. However, location of gene copies on separate chromosomes is evident at least in the case of Rubisco in *Prorocentrum minimum*, suggesting possible duplication at chromosomal level or higher [16]. Whole genome duplications by autopolyploidy or allopolyploidy events are the most efficient mechanism to introduce extra genetic material and significantly expand the genomes [43], and have been well documented for animals, plants and protists such as the budding yeast *Saccharomyces cerevisiae* and the ciliate *Paramecium tetraurelia* [44–47]. Given the widespread gene repetition in dinoflagellates, genome duplication is very possible. In fact, ancient polyploidy has been suggested as a mechanism of speciation in the dinoflagellate *Heterocapsa pygmaea* [48]. Because usually most gene duplicates are eventually lost or diverged to different genes after genome duplication, the retention of the numerous copies of genes in dinoflagellates may indicate an evolutionary driving force associated with functional requirements imposed on dinoflagellates for adaptation to a wide range of habitats. In support of this, highly expressed genes tend to occur in tandem-repeated copies [16,49]. The predicted high gene numbers can be a result of gene and genome duplication followed by differential gene loss and diversification. Ultimate verification of actual gene number and genome duplication as a potential causative mechanism would require sequencing of one or more dinoflagellate genomes, which will also further validate the eukaryotic log gene number-log genome size correlation empirically derived in this study.

Supporting Information

Table S1 Genome size, protein-coding gene number, total gene number, and gene-coding percentage for the sequenced genomes of eukaryotes, bacteria, archaea, viruses, mitochondria, and chloroplasts estimated based on genome sequences.
Found at: doi:10.1371/journal.pone.0006978.s001 (1.97 MB DOC)

Acknowledgments

We thank Xue Feng Liu for assistance in compiling an initial dataset. Comments from the two reviewers helped to improve the manuscript significantly.

Author Contributions

Conceived and designed the experiments: SL. Performed the experiments: YH. Analyzed the data: YH SL. Wrote the paper: YH SL.

References

1. Lynch M, Conery JS (2003) The origins of genome complexity. Science 302: 1401–1404.
2. Konstantinidis KT, Tiedje JM (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci U S A 101: 3160–3165.
3. Gregory TR (2005) Synergy between sequence and size in large-scale genomics. Nature Rev Genet 6: 699–708.
4. Hackett JD, Anderson DM, Erdner DL, Bhattacharya D (2004) Dinoflagellates: a remarkable evolutionary experiment. Am J Bot 91: 1523–1534.
5. Lin S (2006) The smallest dinoflagellate genome is yet to be found: a comment on LaJeunesse, et al. "*Symbiodinium* (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates". J Phycol 42: 746–748.
6. LaJeunesse TC, Lambert G, Andersen RA, Coffroth MA, Galbraith DW (2005) *Symbiodinium* (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates. J Phycol 41: 880–886.
7. Veldhuis MJW, Cucci TL, Sieracki ME (1997) Cellular DNA content of marine phytoplankton using two new fluorochromes: taxonomic and ecological implications. J Phycol 33: 527–541.
8. Holm-Hansen O (1969) Algae: amounts of DNA and organic carbon in single cells. Science 163: 87–88.
9. Rizzo PJ (1987) Biochemistry of the dinoflagellate nucleus. In: Taylor FJR, ed. The biology of dinoflagellates. Oxford: Blackwell Science Inc. pp 143–173.
10. Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, et al. (2007) Eukaryotic genome size databases. Nucleic Acids Res 35: D332–D338.

11. Allen JR, Roberts TM, Loeblich I, Alfred R, Klotz LC (1975) Characterization of the DNA from the dinoflagellate *Crypthecodinium cohnii* and implications for nuclear organization. Cell 6: 161–169.
12. Hinnebusch AG, Klotz LC, Immergut E, Loeblich ARI (1980) Deoxyribonucleic acid sequence organization in the genome of the dinoflagellate *Crypthecodinium cohnii*. Biochem 19: 1744–1755.
13. Steel RE (1980) Aspects of the composition and organization of dinoflagellate DNA [Ph.D. thesis]. New Haven: Yale University.
14. Anderson DM, Grabher A, Herzog M (1992) Separation of coding sequences from structural DNA in the dinoflagellate *Crypthecodinium cohnii*. Mol Mar Biol Biotechnol 2: 89–96.
15. Moreau H, Geraud ML, Bhaud Y, Soyer-Gobillard MO (1998) Cloning, characterization and chromosomal localization of a repeated sequence in *Crypthecodinium cohnii*, a marine dinoflagellate. Int Microbiol 1: 35–43.
16. Zhang H, Lin S (2003) Complex gene structure of the form II RUBISCO in the dinoflagellate *Prorocentrum minimum* (Dinophyceae). J Phycol 39: 1160–1171.
17. Zhang H, Hou Y, Lin S (2006) Isolation and characterization of proliferating cell nuclear antigen from the dinoflagellate *Pfiesteria piscicida*. J Euk Microbiol 53: 142–150.
18. Xu L, Chen H, Hu X, Zhang R, Zhang Z, et al. (2006) Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. Mol Biol Evol 23: 1107–1108.
19. Archibald J (2007) Nucleomorph genomes: structure, function, origin and evolution. BioEssays 29: 392–402.
20. Spector DL (1984) Dinoflagellate nuclei. In: Spector DL, ed. Dinoflagellates. New York: Academic Press. pp 107–147.
21. Bouligand Y, Norris V (2001) Chromosome separation and segregation in dinoflagellates and bacteria may depend on liquid crystalline states. Biochimie 83: 187–192.
22. Chan YH, Wong JTY (2007) Concentration-dependent organization of DNA by the dinoflagellate histone-like protein HCc3. Nucleic Acids Res 35: 2573–2583.
23. Sigee DC (1984) Structural DNA and genetically active DNA in dinoflagellate chromosomes. Biosystems 16: 203–210.
24. Bhaud Y, Guillebault D, Lennon JF, Defacque H, Soyer-Gobillard MO, et al. (2000) Morphology and behaviour of dinoflagellate chromosomes during the cell cycle and mitosis. J Cell Sci 113: 1231–1239.
25. Lin S, Zhang H, Gray MW (2008) RNA editing in dinoflagellates and its implications for the evolutionary history of the editing machinery. In: Smith H, ed. RNA and DNA editing: molecular mechanisms and their integration into biological systems. Hoboken, NJ: John Wiley & Sons, Inc. pp 280–309.
26. Zhang H, Campbell DA, Sturm NR, Lin S (2009) Dinoflagellate spliced leader RNA genes display a variety of sequences and genomic arrangements. Mol Biol Evol 26: 1757–1771.
27. Lukes J, Leander BS, Keeling PJ (2009) Cascades of convergent evolution: the corresponding evolutionary histories of euglenozoans and dinoflagellates. Proc Natl Acad Sci U S A 106: 9963–9970.
28. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, et al. (2005) The genome of the kinetoplastid parasite, *Leishmania major*. Science 309: 436–442.
29. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, et al. (2005) The genome of the African trypanosome *Trypanosoma brucei*. Science 309: 416–422.
30. Rizzo PJ, Nooden LD (1973) Isolation and chemical composition of dinoflagellate nuclei. J Euk Microbiol 20: 666–672.
31. Roberts TM, Tuttle RC, Allen JR, Loeblich AR, Klotz LC (1974) New genetic and physicochemical data on structure of dinoflagellate chromosomes. Nature 248: 446–447.
32. Blank RJ (1987) Cell architecture of the dinoflagellate *Symbiodinium* sp. inhabiting the Hawaiian stony coral *Montipora verrucosa*. Mar Bio 94: 143–155.
33. Pfiester L, Anderson DM (1987) Dinoflagellate reproduction. In: Taylor FJR, ed. The Biology of dinoflagellates. Oxford: Blackwell Scientific Inc. pp 611–648.
34. Coats DW (2002) Dinoflagellate life-cycle complexities. J Phycol 38: 417–419.
35. Santos SR, Coffroth MA (2003) Molecular genetic evidence that dinoflagellates belonging to the genus *Symbiodinium* Freudenthal are haploid. Biol Bull 204: 10–20.
36. McEwan M, Humayun R, Slamovits CH, Keeling PJ (2008) Nuclear genome sequence survey of the dinoflagellate *Heterocapsa triquetra*. J Euk Microbiol 55: 530–535.
37. Hackett JD, Scheetz TE, Yoon HS, Soares MB, Bonaldo MF, et al. (2005) Insights into a dinoflagellate genome through expressed sequence tag analysis. BMC Genomics 6.
38. Patron NJ, Waller RF, Archibald JM, Keeling PJ (2005) Complex protein targeting to dinoflagellate plastids. J Mol Biol 348: 1015–1024.
39. Van Dolah FM, Lidie KB, Monroe EA, Bhattacharya D, Campbell L, et al. (2009) The Florida red tide dinoflagellate *Karenia brevis*: new insights into cellular and molecular processes underlying bloom dynamics. Harmful Algae 8: 562–572.
40. Patron NJ, Waller RF, Keeling PJ (2006) A tertiary plastid uses genes from two endosymbionts. J Mol Biol 357: 1373–1382.
41. Slamovits CH, Keeling PJ (2008) Plastid-derived genes in the nonphotosynthetic alveolate *Oxyrrhis marina*. Mol Biol Evol 25: 1297–1306.
42. Slamovits CH, Keeling PJ (2008) Widespread recycling of processed cDNAs in dinoflagellates. Curr Biol 18: R550–R552.
43. Lynch M (2007) Genomic expansion by gene duplication. In: Lynch M, ed. The origins of genome architecture. Sunderland, MA: Sinauer Associates. pp 193–235.
44. Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708–713.
45. Ramsey J, Schemske DW (1998) Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu Rev Ecol Syst 29: 467–501.
46. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, et al. (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444: 171–178.
47. Gregory TR, Mable BK (2005) Polyploidy in animals. In: Gregory TR, ed. The evolution of the genome. San Diego, CA: Elsevier. pp 427–517.
48. Loeblich AR, III, Schmidt RJ, Sherley JL (1981) Scanning electron microscopy of *Heterocapsa pygmaea* sp. nov., and evidence for polyploidy as a speciation mechanism in dinoflagellates. J Plankton Res 3: 67–79.
49. Bachvaroff TR, Place AR (2008) From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate *Amphidinium carterae*. PLoS ONE 3: e2929.
50. Salois P, Morse D (1997) Characterization and molecular phylogeny of a protein kinase cDNA from the dinoflagellate *gonyaulax* (Dinophyceae). J Phycol 33: 1063–1072.
51. Liu LY, Hastings JW (2006) Novel and rapidly diverging intergenic sequences between tandem repeats of the luciferase genes in seven dinoflagellate species. J Phycol 42: 96–103.
52. Lee D, Mittag M, Sczekan S, Morse D, Hastings JW (1993) Molecular cloning and genomic organization of a gene for luciferin-binding protein from the dinoflagellate *Gonyaulax polyedra*. J Biol Chem 268: 8842–8850.
53. Bertomeu T, Morse D (2004) Isolation of a dinoflagellate mitotic cyclin by functional complementation in yeast. Biochem Biophys Res Commun 323: 1172–1183.
54. Le QH, Markovic P, Hastings JW, Jovine RVM, Morse D (1997) Structure and organization of the peridinin chlorophyll a binding protein gene in *Gonyaulax polyedra*. Mol General Genet 255: 595–604.
55. Reichman J, Wilcox T, Vize P (2003) PCP gene family in *Symbiodinium* from *Hippopus hippopus*: low level of concerted evolution, isoform diversity and spectral tuning of chromophores. Mol Biol Evol 20: 2143–2154.