Analysis of amino acid and codon usage inParamecium bursaria

Title Page and Abstract

Title

Analysis of amino acid and codon usage in *Paramecium bursaria*

Author names and affiliations

Hideo Dohra[a,b], Masahiro Fujishima[c,d], and Haruo Suzuki[c,*]

[a]Instrumental Research Support Office, Research Institute of Green Science and Technology, [b]Department of Biological Science, Graduate School of Science, Shizuoka University, 836 Ohya, Suruga-ku, Shizuoka 422-8529, Japan, [c]Department of Environmental Science and Engineering, Graduate School of Science and Engineering, Yamaguchi University, 1677-1 Yoshida, Yamaguchi 753-8512, Japan, [d]National Bio-Resource Project of Japan Agency for Medical Research and Development.

[*]Corresponding author. E-mail address: haruo@yamaguchi-u.ac.jp

Keywords

Paramecium bursaria; Transcriptome; Gene expression; Synonymous codon usage; Amino acid usage

Abbreviations

logCPM, log-counts-per-million; logFC, log2 Fold Change; WCA, within-group correspondence analysis; SCU, Synonymous codon usage; AAU, Amino acid usage.

Abstract

The ciliate *Paramecium bursaria* harbors the green-alga *Chlorella* symbionts. We reassembled the *P. bursaria* transcriptome to minimize falsely fused transcripts, and investigated amino acid and codon usage using the transcriptome data. Surface proteins

preferentially use smaller amino acid residues like cysteine. Unusual synonymous codon and amino acid usage in highly expressed genes can reflect a balance between translational selection and other factors. A correlation of gene expression level with synonymous codon or amino acid usage is emphasized in genes down-regulated in symbiont-bearing cells compared to symbiont-free cells. Our results imply that the selection is associated with *P. bursaria-Chlorella* symbiosis.

1. Introduction

The ciliate *Paramecium bursaria* harbors several hundred cells of the symbiotic algae *Chlorella* species in the cytoplasm. This symbiosis is a mutualistic relationship and seems to be a stable association in that the algal cells are retained through cell division as well as sexual reproduction of the host *P. bursaria* [1]. Algae-bearing cells of *P. bursaria* can grow faster than algae-free cells [2] and bacteria- or yeast-bearing cells [3]. Moreover, the division of algal cells in *P. bursaria* is dependent on the host cell cycle [4]. Algae-bearing *P. bursaria* also acquires resistance to high temperature [5], infection with bacteria and yeasts [3], photo-oxidative stress [6], and UV damage [7]. The symbiotic *Chlorella* also affect behaviors of the host *P. bursaria*, such as expression of circadian rhythms [8-10] and response to light [11-14]. Furthermore, because of the stable symbiotic relationship, it is very unusual for aposymbiotic *P. bursaria* to be collected from the natural environments, and an only one *P. bursaria* mutant lacking symbiotic algae has been reported by Tonooka and Watanabe [15,16]. The mutualistic relationship and the stable association between the host *P. bursaria* and the symbiotic *Chlorella*, and the physiological or behavioral changes in algae-bearing *P. bursaria* bring us to the question which some selective pressure on the host genome may be involved in establishment of the stable symbiotic relationship between *P. bursaria* and the symbiotic algae.

In many organisms, synonymous codons are not used with equal frequencies. This phenomenon known as ‘codon bias’ can reflect a balance between selection, mutation, and genetic drift [17,18]. In unicellular organisms such as the bacterium *Escherichia coli* and the yeast *Saccharomyces cerevisiae*, genes expressed at high levels preferentially use a subset of synonymous codons, which are best recognized by the most abundant tRNA species [19]. It is also reported that highly expressed genes have unusual codon usage in the ciliates *Tetrahymena thermophila* and *Paramecium tetraurelia* with alternative genetic codes [20-22]. The codon bias presumably reflects natural selection for efficient and accurate translation (a.k.a. ‘translational selection’).

Amino acid and codon usage bias have been reported in several symbionts and parasites, such as *Buchnera*, endosymbiotic bacteria of aphids [23], the nitrogen-fixing endosymbiont *Bradyrhizobium japonicum* [24], *G. lamblia* [25,26], *Plasmodium* species [27], and *Mycoplasma bovis*, a major pathogen of cattle [28]. Nevertheless, little is known about amino acid and codon usage bias in hosts associated

with symbionts. RNA-Seq de-novo transcriptome assembly of *P. bursaria* has been recently reported [29]. Here, we reassemble the *P. bursaria* transcriptome, and perform analyses of amino acid and codon usage in *P. bursaria* using the transcriptome data. Our goal was to gain insight into forces driving the architecture of genome, transcriptome, and proteome, which might provide the clue to understand the co-evolution and symbiosis of *P. bursaria* and *Chlorella*.

2. Materials and methods

2.1. Transcriptome data analysis

We reassembled transcriptome of *P. bursaria* strain Yad1g1N provided by Symbiosis Laboratory, Yamaguchi University with support in part by the National Bio-Resource Project of Japan Agency for Medical Research and Development. The read sequence data obtained from RNA-Seq analysis in the previous study [29] (http://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=DRA000907) were de novo assembled using the Trinity program version: trinityrnaseq_r20140717 with an option "-jaccard_clip" to prevent fusion of transcripts with the 3'-UTR overlap [30]. Possible contaminant sequences derived from symbiotic *Chlorella* and lowly expressed transcripts with log-counts-per-million (logCPM) < 0 were removed as described elsewhere [29].

OrfPredictor [31] was used to predict protein-coding sequences from the *P. bursaria* transcript sequences. We annotated the proteins based on BLASTP searches (E-value < 1e-5) [32] against the COG database including the eukaryotic orthologous groups (KOGs) [33], HMMER searches (http://hmmer.janelia.org/) against the Pfam database of protein families [34], and InterProScan [35] against the InterPro protein families database [36]. We also used G-Links [37] to collect information from different databases about the genes of interest. To minimize sampling errors, 14,252 proteins longer than 99 amino acids were used for the subsequent analysis of synonymous codon and amino acid usage. A comprehensive list of the genes are shown in (Table S1).

2.2. Analyses of synonymous codon and amino acid usage

All analyses were implemented on the G-language Genome Analysis Environment version 1.9.0, available at http://www.g-language.org [38-40].

Correspondence analysis combines multivariate data into a small number of

variables (axes) that explains most of the variation among the original variables (i.e. 61 codons or 20 amino acids for each gene), and yields the coordinates of each gene on each new axis [41-43]. The correspondence analysis was implemented using the 'ade4' library of R [44].

First, we analyzed a correlation between the axis scores and the following gene features: the relative frequency of aromatic amino acids (AROMA), the mean hydropathicity (GRAVY), and the mean molecular weight (MMW), calculated from the amino acid sequence [41,42]; the relative frequency of guanine and cytosine (GC3) and that of guanine and thymine (GT3) at the third codon positions, calculated from the nucleotide sequence. Pearson's product moment correlation coefficient (r) between each axis and each feature was calculated.

Second, we analyzed distributions of the axis scores for putatively highly expressed genes encoding ribosomal proteins. A mean standard score (z-score) for the ribosomal proteins was calculated. A z-score indicates how many standard deviations an element is from the mean.

3. Results

3.1. Transcriptome reassembly

We reassembled the *P. bursaria* transcriptome. The read sequence data obtained from RNA-Seq analysis in the previous study [29] were de novo assembled using the Trinity program version: trinityrnaseq_r20140717 with an option "-jaccard_clip" to prevent fusion of transcripts with the 3'-UTR overlap. The de novo assembly produced 57,890 genes and 72,480 transcripts. We removed transcript sequences derived from symbiotic *Chlorella*, ribosomal RNAs, and lowly expressed transcripts. This produced 19,323 transcript sequences containing isoforms, and we picked just one highest-covered isoform per gene. The resulting 15,005 unigenes were 4,448 more than that of the previous study [29]. We compared the new unigenes to the previous ones by BLASTN (E-value < 1e-100). Of the previous unigenes, 2,983 matched several (upto nine) of the new unigenes with different length, functional annotations, and values obtained by gene expression analysis such as log2 Fold Change (logFC) and log-counts-per-million (logCPM). Moreover, 2,060 of the new unigenes were judged to be differentially expressed in the opposite direction (either up- or down-regulation) compared to the previous unigenes (data not shown). These results

suggested that the previous unigene set contained artificially fused contigs and that our new assembly has reduced the artifacts and improved transcriptome data for more accurate gene expression analysis.

3.2. Gene expression

We compared gene expressions of symbiont-bearing and symbiont-free cells of *P. bursaria* as described elsewhere [29]. Of the 14,252 unigenes (>99 amino acids) obtained in this study, 9,142 (64.1%) were significantly differentially expressed between symbiont-bearing and symbiont-free cells with false discovery rates (FDR) < 0.05 (Figure 1). The positive and negative values of log2 Fold Change (logFC) indicate that the genes were up-regulated and down-regulated, respectively, in symbiont-bearing cells compared to symbiont-free cells. The parametric analysis of gene set enrichment (PAGE) [45] based on the logFC with FDR < 0.05 detected enrichment in Pfam protein families. The up-regulated protein families (logFC > 0) included 'Myb-like DNA-binding domain', 'von Willebrand factor type A domain', and 'NMDA receptor-regulated protein 1'. The down-regulated protein families (logFC < 0) included 'Glutathione S-transferase', 'Aminotransferase class I and II', 'Alcohol dehydrogenase GroES-like domain', 'Zinc-binding dehydrogenase', 'Eukaryotic translation initiation factor eIF2A', and '*Paramecium* surface antigen domain' (PF01508), which is a cysteine rich extracellular repeat found in the G surface protein of *Paramecium primaurelia* [46,47]. The down-regulation of the surface antigens in symbiont-bearing *P. bursaria* cells is reminiscent of the down-regulation of surface antigens after infection of *Holospora obtusa* to the host *Paramecium caudatum* [48,49].

3.3. Synonymous codon usage

Overall (summed) usage values of 63 sense codons for all the 14,252 protein-coding genes longer than 99 amino acids in *P. bursaria* are shown in Table 1. There is a strong bias toward AT-rich codons for all protein genes. This bias is smaller and GC-ending codons are more frequently used in highly expressed genes encoding ribosomal proteins. Among the four synonymous codons encoding glutamine (Q), the reassigned codons 'taa' and 'tag' are more frequently used than the canonical codons 'caa' and 'cag'. Similar trends were reported for other ciliates *T. thermophila* and *P. tetraurelia* [22].

We performed within-group correspondence analysis (WCA) [43,50] on

codon frequencies to identify major sources of variation in synonymous codon usage among the *P. bursaria* genes. The first and second axes (SCU1 and SCU2) obtained by WCA explained 8.3% and 6.1% of the total variance of codon usage data, respectively (Figure 2). The first axis (SCU1) was positively correlated with GC3 (r = 0.542), and negatively correlated with GT3 (r = -0.568). The second axis (SCU2) was strongly correlated with G+C content at the third codon position (r = 0.77).

We investigated a relationship between gene expression level and synonymous codon usage. The highly expressed genes encoding ribosomal proteins showed a mean z-score of 2.959 and were thus strongly deviated from the other genes on SCU1. Genes with high SCU1 scores (SCU1 > 5) included those expressed at high levels (logCPM > 5) encoding translation elongation factors, tubulins (alpha and beta), histones (H2A, H2B, H3 and H4), cathepsins (L and S), actin, molecular chaperones (HSP70 and HSP90), and glutathione S-transferase, in addition to ribosomal proteins (Table S1). The 14 genes encoding 'Cysteine proteinase Cathepsin L' (KOG1543O) containing 'Cathepsin propeptide inhibitor domain' (PF08246) tended to have high SCU1 scores (ranging from 0.05 to 7.14 with a median of 4.48) and high logCPM values (ranging from 3.15 to 12.4 with a median of 8.81). This is consistent with the observation that 'The ciliate *P. tetraurelia* secretes large amounts of a cysteine protease into the growth medium, presumably for extracellular food digestion.' [51] Many of the genes with high SCU1 scores, and thus predicted to be highly expressed, are uncharacterized based on homology and domain searches against sequence databases. The uncharacterized genes might be good targets for future experimental studies.

There is a correlation between logCPM and SCU1 (r = 0.412). To investigate the relationship between logCPM and SCU1 in genes with different logFC, the genes were classified into down-regulated (FDR < 0.05, logFC < -2, n = 203), up-regulated (FDR < 0.05, logFC > +2, n = 389), and non-differentially expressed genes (FDR > 0.05, n = 5110). The correlation between logCPM and SCU1 was strongest in the down-regulated genes (r = 0.58), weakest in the up-regulated genes (r = 0.22), and intermediate in the non-differentially expressed genes (r = 0.38), indicating that the correlation between gene expression level and synonymous codon usage is clearer in the down-regulated genes than in the up-regulated or unchanged genes.

3.4. Amino acid usage

Overall (summed) usage values of 20 amino acids for all the 14,252 proteins longer than 99 amino acids in *P. bursaria* are shown in Table 2. The top three most frequent amino acids are leucine (L), glutamine (Q), and isoleucine (I), with percentage contents of 9.8%, 9.1%, and 8.8%, respectively. The top three least frequent amino acids are tryptophan (W), histidine (H), and cysteine (C), with percentage contents of 0.78%, 1.72%, and 1.73%, respectively.

We performed correspondence analysis on amino acid frequencies to identify major sources of variation in amino acid usage among the *P. bursaria* proteins. The first two axes (AAU1 and AAU2) obtained by the correspondence analysis explained 26.2% and 14.4% of the total variance of amino acid usage data, respectively (Figure 3). The first axis (AAU1) was positively correlated with the mean molecular weight (MMW) of the amino acids (r = 0.693) and negatively correlated with the cysteine content (r = -0.48) in the protein. Proteins annotated as '*Paramecium* surface antigen domain' (PF01508) (n = 105), 'Subtilisin-like proprotein convertase' (KOG3525O) (n = 102), and 'Giardia variant-specific surface protein' (PF03302) (n = 13) showed negative AAU1 scores and low MMW values. Of the 13 proteins assigned to 'Giardia variant-specific surface protein', 12 were assigned concurrently to 'Subtilisin-like proprotein convertase'. The proteins annotated as 'Subtilisin-like proprotein convertase' or 'Giardia variant-specific surface protein' were also assigned to InterPro entry IPR009030 annotated as 'Insulin-like growth factor binding protein, N-terminal' (Table S1). A median value of percentage cysteine contents was higher for these anomalous proteins, i.e., '*Paramecium* surface antigen domain' (10.9%), 'Subtilisin-like proprotein convertase' (9.5%), and 'Giardia variant-specific surface protein' (13.3%), than for all the 14,252 proteins (1.3%) (Table 2).

We investigated a relationship between gene expression level and amino acid usage. The highly expressed ribosomal proteins showed a mean z-score of -2.429 and were thus deviated from the other proteins on the second axis (AAU2). There is a negative correlation between logCPM and AAU2 (r = -0.462). The correlation between logCPM and AAU2 was strongest in the down-regulated genes (r = -0.549), weakest in the up-regulated genes (r = -0.252), and intermediate in the non-differentially expressed genes (r = -0.478), indicating that the correlation between gene expression level and amino acid usage is clearer in the down-regulated genes than in the up-regulated or unchanged genes.

The third axis generated by the correspondence analysis (12.4% of the total variance of amino acid usage data) was correlated with the aromaticity (AROMA, r = -0.748) and hydropathy (GRAVY, r = -0.727) of each protein. Similar trends of amino acid usage in the proteins were reported for *Giardia lamblia* [25,26].

4. Discussion

We present a new transcriptome assembly of *P. bursaria*. We performed de novo transcriptome assembly using the Trinity program (version trinityrnaseq_r20140717) with an option "-jaccard_clip" to minimize falsely fused transcripts. To assess how the choice of assemblers affects our results, the SOAPdenovo-Trans version 1.04 [52] was run with the same k-mer length as Trinity (i.e. 25 bp). Basic statistics such as maximum, average, median, and N50 sequence lengths for the Trinity assembly were longer than those for the SOAPdenovo-Trans assembly (Table S2). Moreover, the predicted protein lengths tended to be longer when using Trinity than when using SOAPdenovo-Trans (Figure S1). Previous studies provide some evidence that Trinity performs better than the other assemblers [53,54].

We investigated synonymous codon usage and amino acid usage in *P. bursaria* using the new transcriptome data. In *P. bursaria*, surface proteins (those annotated as '*Paramecium* surface antigen domain', 'Subtilisin-like proprotein convertase', and 'Giardia variant-specific surface protein') preferentially use smaller amino acid residues such as cysteine. A similar trend of amino acid usage was reported for several protists, such as surface proteins known as an immobilization antigen of *P. tetraurelia* [55,56] and *T. thermophila* [57,58], variant surface proteins of *G. lamblia* [26], and variant-surface cysteine-rich proteins of the mitochondrion-lacking diplomonad fish parasite *Spironucleus salmonicida* [59]. Surface antigens of *P. tetraurelia* are clearly essential for its survival because they were estimated to represent 3.5% of total cellular protein [60]. Given the large amount of *Paramecium* surface antigens, their amino acid usage may be subject to a selection to reduce the synthetic cost [61,62]. We found that highly expressed genes (those encoding ribosomal proteins) are unusual in amino acid and codon usage. The correlation of gene expression level with amino acid or codon usage is emphasized in genes down-regulated in symbiont-bearing cells compared to symbiont-free cells, suggesting that the strength of translational selection in *P. bursaria* may be related to *P. bursaria-Chlorella* symbiosis.

It is important to note that the results remained similar when using different transcriptome assemblers; i.e. Trinity and SOAPdenovo-Trans (data not shown).

Sharp et al. (2005) reported that fast-growing bacteria tend to have stronger selection on codon usage, and that a lack of selection on codon usage in obligate intracellular parasites or endosymbionts may reflect parasitic lifestyle and low effective population sizes [63]. In the ciliates like *P. bursaria*, translational selection on amino acid and codon usage may reflect the presence of large amounts of macronuclear DNAs in the cells. During macronuclear development in *Paramecium* cells, macronuclear DNA molecules are amplified 10-20 fold in *P. bursaria*; for review see [64]. In *P. bursaria*, amino acid and codon usage may result from selection for efficient and accurate translation of mRNAs transcribed from such large amount (high copy number) of macronuclear DNAs.

Trinity reported potential isoforms from alternative splicing in the de novo transcriptome assembly. We are aware of the variation in gene expression and compositional features among the alternatively spliced isoforms (data not shown). There is currently an ongoing project of *P. bursaria* genome sequencing. The combined use of the reference genome and transcriptomes of *P. bursaria* will allow us to test the contribution of translational selection and splicing-related forces on codon usage [65].

A phylogeny of the genus *Paramecium* indicates that the most basal *Paramecium* lineage is *P. bursaria* [21]. The increasing number of genome sequences available for *Paramecium* species, especially *P. chlorelligerum* with symbiotic green algae [66] will help to generate hypotheses about lineages in which *Paramecium-Chlorella* symbiosis has taken place.

Acknowledgments:

Computational resources were provided by the Data Integration and Analysis Facility, National Institute for Basic Biology.

References

[1] Siegel, R.W. (1960). Hereditary endosymbiosis in *Paramecium bursaria*. Exp Cell Res 19, 239-52.

[2] Karakashian, S.J. (1963). Growth of *Paramecium bursaria* as influenced by the presence of algal symbionts. Physiol. Zool. 36, 52-68.

[3] Görtz, H.D. (1982). Infections of *Paramecium bursaria* with bacteria and yeasts. J Cell Sci 58, 445-53.

[4] Kadono, T., Kawano, T., Hosoya, H. and Kosaka, T. (2004). Flow cytometric studies of the host-regulated cell cycle in algae symbiotic with green paramecium. Protoplasma 223, 133-41.

[5] Iwatsuki, K., M., N. and Suehiro, K. (1998). Symbiotic *Chlorella* enhances the thermal tolerance in *Paramecium bursaria*. Comp Biochem Physiol Part A 121, 405-409.

[6] Hortnagl, P.H. and Sommaruga, R. (2007). Photo-oxidative stress in symbiotic and aposymbiotic strains of the ciliate *Paramecium bursaria*. Photochem Photobiol Sci 6, 842-7.

[7] Summerer, M., Sonntag, B., Hortnagl, P. and Sommaruga, R. (2009). Symbiotic ciliates receive protection against UV damage from their algae: a test with *Paramecium bursaria* and *Chlorella*. Protist 160, 233-43.

[8] Miwa, I., Fujimori, N. and Tanaka, M. (1996). Effects of symbiotic *Chlorella* on the period length and the phase shift of circadian rhythms in *Paramecium bursaria*. Eur J Protistol 32 (Suppl 1), 102-107.

[9] Tanaka, M. and Miwa, I. (1996). Significance of photosynthetic products of symbiotic *Chlorella* to establish the endosymbiosis and to express the mating reactivity rhythm in *Paramecium bursaria*. Zool Sci 13, 685-692.

[10] Miwa, I. (2009) Regulation of Circadian Rhythms of *Paramecium bursaria* by Symbiotic *Chlorella* Species. In Endosymbionts in *Paramecium*, Microbiology Monographs Volume 12 (Fujishima, M., ed.), pp. 83-110. Springer

[11] Saji, M. and Oosawa, F. (1974). Mechanism of photoaccumulation in *Paramecium bursaria*. J Protozool 21, 556-61.

[12] Niess, D., Reisser, W. and Wiessner, W. (1981). The role of endosymbiotic algae in photoaccumulation of green *Paramecium bursaria*. Planta 152, 268-71.

[13] Sommaruga, R. and Sonntag, B. (2009) Photobiological Aspects of the Mutualistic Association Between *Paramecium bursaria* and *Chlorella*. In Endosymbionts in *Paramecium*, Microbiology Monographs Volume 12 (Fujishima, M., ed.), pp. 111-130. Springer

[14] Iwatsuki, K. and Naitoh, Y. (1981). The role of symbiotic chlorella in photoresponses of *Paramecium bursaria*. Proc. Japan Acad 57(B), 318-323.

[15] Tonooka, Y. and Watanabe, T. (2002). A natural strain of *Paramecium bursaria* lacking symbiotic algae. Eur. J. Protistol. 38, 55-58.

[16] Tonooka, Y. and Watanabe, T. (2007). Genetics of the relationship between the ciliate *Paramecium bursaria* and its symbiotic algae. Inv. Biol. 126, 287-294.

[17] Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897-907.

[18] Shah, P. and Gilchrist, M.A. (2011). Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proc Natl Acad Sci U S A 108, 10231-6.

[19] Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2, 13-34.

[20] Eisen, J.A. et al. (2006). Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol 4, e286.

[21] Aury, J.M. et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444, 171-8.

[22] Salim, H.M., Ring, K.L. and Cavalcanti, A.R. (2008). Patterns of codon usage in two ciliates that reassign the genetic code: *Tetrahymena thermophila* and *Paramecium tetraurelia*. Protist 159, 283-98.

[23] Rispe, C., Delmotte, F., van Ham, R.C. and Moya, A. (2004). Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. Genome Res 14, 44-53.

[24] Das, S., Pan, A., Paul, S. and Dutta, C. (2005). Comparative analyses of codon and amino acid usage in symbiotic island and core genome in nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum*. J Biomol Struct Dyn 23, 221-32.

[25] Garat, B. and Musto, H. (2000). Trends of amino acid usage in the proteins from the unicellular parasite *Giardia lamblia*. Biochem Biophys Res Commun 279, 996-1000.

[26] Lafay, B. and Sharp, P.M. (1999). Synonymous codon usage variation among *Giardia lamblia* genes and isolates. Mol Biol Evol 16, 1484-95.

[27] Yadav, M.K. and Swati, D. (2012). Comparative genome analysis of six malarial parasites using codon usage bias based tools. Bioinformation 8, 1230-9.

[28] Zhou, J.H. et al. (2014). The effect of multiple evolutionary selections on

synonymous codon usage of genes in the *Mycoplasma bovis* genome. PLoS One 9, e108949.

[29] Kodama, Y., Suzuki, H., Dohra, H., Sugii, M., Kitazume, T., Yamaguchi, K., Shigenobu, S. and Fujishima, M. (2014). Comparison of gene expression of *Paramecium bursaria* with and without *Chlorella variabilis* symbionts. BMC Genomics 15, 183.

[30] Grabherr, M.G. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644-52.

[31] Min, X.J., Butler, G., Storms, R. and Tsang, A. (2005). OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Res 33, W677-80.

[32] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-402.

[33] Tatusov, R.L. et al. (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41.

[34] Finn, R.D. et al. (2014). Pfam: the protein families database. Nucleic Acids Res 42, D222-30.

[35] Jones, P. et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236-40.

[36] Mitchell, A. et al. (2015). The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res 43, D213-21.

[37] Oshita, K., Tomita, M. and Arakawa, K. (2014). G-Links: a gene-centric link acquisition service. F1000Research 3, 285.

[38] Arakawa, K., Mori, K., Ikeda, K., Matsuzaki, T., Kobayashi, Y. and Tomita, M. (2003). G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. Bioinformatics 19, 305-6.

[39] Arakawa, K. and Tomita, M. (2006). G-language System as a platform for large-scale analysis of high-throughput omics data. J Pesticide Sci 31, 282-288.

[40] Arakawa, K., Suzuki, H. and Tomita, M. (2008). Computational Genome Analysis Using The G-language System. Genes, Genomes and Genomics 2, 1-13.

[41] Lobry, J.R. and Gautier, C. (1994). Hydrophobicity, expressivity and aromaticity

are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res 22, 3174-80.

[42] Zavala, A., Naya, H., Romero, H. and Musto, H. (2002). Trends in codon and amino acid usage in *Thermotoga maritima*. J Mol Evol 54, 563-8.

[43] Suzuki, H., Brown, C.J., Forney, L.J. and Top, E.M. (2008). Comparison of correspondence analysis methods for synonymous codon usage in bacteria. DNA Res 15, 357-65.

[44] R_Core_Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[45] Kim, S.Y. and Volsky, D.J. (2005). PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics 6, 144.

[46] Prat, A., Katinka, M., Caron, F. and Meyer, E. (1986). Nucleotide sequence of the *Paramecium primaurelia* G surface protein. A huge protein with a highly periodic structure. J Mol Biol 189, 47-60.

[47] Prat, A. (1990). Conserved sequences flank variable tandem repeats in two alleles of the G surface protein of *Paramecium primaurelia*. J Mol Biol 211, 521-35.

[48] Nakamura, Y., Hori, M. and Fujishima, M. (2004). Endonuclear symbiotic bacterium *Holospora obtusa* reversibly changes types of surface antigens expressed in the host *Paramecium caudatum*. Jpn. J. Protozool. 37, 46-47.

[49] Nakamura, Y., Hori, M. and Fujishima, M. (2005). Endonuclear symbiotic bacterium *Holospora obtusa* reversibly changes types of surface antigens expressed in the host *Paramecium caudatum*. Jpn. J. Protozool. 38, 7-8.

[50] Charif, D., Thioulouse, J., Lobry, J.R. and Perriere, G. (2005). Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics 21, 545-7.

[51] Volkel, H., Kurz, U., Linder, J., Klumpp, S., Gnau, V., Jung, G. and Schultz, J.E. (1996). Cathepsin L is an intracellular and extracellular protease in *Paramecium tetraurelia*. Purification, cloning, sequencing and specific inhibition by its expressed propeptide. Eur J Biochem 238, 198-206.

[52] Xie, Y. et al. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30, 1660-6.

[53] Clarke, K., Yang, Y., Marsh, R., Xie, L. and Zhang, K.K. (2013). Comparative

analysis of de novo transcriptome assembly. Sci China Life Sci 56, 156-62.

[54] Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J.A., Stewart, R. and Dewey, C.N. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol 15, 553.

[55] Thai, K.Y. and Forney, J.D. (2000). Analysis of the conserved cysteine periodicity of *Paramecium* variable surface antigens. J Eukaryot Microbiol 47, 242-8.

[56] Scott, J., Leeck, C. and Forney, J. (1993). Molecular and genetic analyses of the B type surface protein gene from *Paramecium tetraurelia*. Genetics 134, 189-98.

[57] Deak, J.C. and Doerder, F.P. (1995). Sequence, codon usage and cysteine periodicity of the *SerH1* gene and in the encoded surface protein of *Tetrahymena thermophila*. Gene 164, 163-6.

[58] Doerder, F.P. (2000). Sequence and expression of the *Ser*J immobilization antigen gene of *Tetrahymena thermophila* regulated by dominant epistasis. Gene 257, 319-26.

[59] Baer, F.M., Smolarz, K., Jungehulsing, M., Theissen, P., Sechtem, U., Schicha, H. and Hilger, H.H. (1992). Feasibility of high-dose dipyridamole-magnetic resonance imaging for detection of coronary artery disease and comparison with coronary angiography. Am J Cardiol 69, 51-6.

[60] Macindoe, H. and Reisner, A.H. (1967). Adsorption titration as a specific semi-quantitative assay for soluble and bound *Paramecium* serotypic antigen. Aust J Biol Sci 20, 141-52.

[61] Smith, D.R. and Chapman, M.R. (2010). Economical evolution: microbes reduce the synthetic cost of extracellular proteins. MBio 1

[62] Krick, T., Verstraete, N., Alonso, L.G., Shub, D.A., Ferreiro, D.U., Shub, M. and Sanchez, I.E. (2014). Amino Acid metabolism conflicts with protein diversity. Mol Biol Evol 31, 2905-12.

[63] Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F. and Sockett, R.E. (2005). Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33, 1141-53.

[64] Freiburg, M. (1988) Organization and expression of the nuclear genome. In *Paramecium* (Görtz, H.D., ed.), pp. 141-154. Springer-Verlag, Berlin.

[65] Warnecke, T. and Hurst, L.D. (2007). Evidence for a trade-off between

translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. Mol Biol Evol 24, 2755-62.

[66] Kreutz, M., Stoeck, T. and Foissner, W. (2012). Morphological and molecular characterization of *Paramecium* (*Viridoparamecium* nov. subgen.) *chlorelligerum* Kahl (Ciliophora). J Eukaryot Microbiol 59, 548-63.

Figure legends

Figure 1. Scatter plot showing the log 2 Fold Change (logFC) in gene expression of symbiont-bearing cells relative to symbiont-free cells, plotted against the log counts per million (logCPM) for the 14,252 genes. The genes were classified into down-regulated genes (blue dots, FDR < 0.05, logFC < -2, n = 203), up-regulated genes (red dots, FDR < 0.05, logFC > +2, n = 389), non-differentially expressed genes (black dots, FDR > 0.05, n = 5110), and the remaining genes (grey dots, FDR < 0.05, -2 < logFC < +2, n = 8,550).

Figure 2. Plot of the first two axes (SCU1 and SCU2) generated by within-group correspondence analysis on codon frequencies for the *Paramecium bursaria* genes. Ribosomal protein genes are indicated by red circles, genes annotated as '*Paramecium* surface antigen domain' (PF01508) are indicated by green circles, 'Subtilisin-like proprotein convertase' (KOG3525O) are indicated by blue circles, and 'Giardia variant-specific surface protein' (PF03302) are indicated by light blue circles, and the remaining genes are indicated by black circles.

Figure 3. Plot of the first two axes (AAU1 and AAU2) generated by correspondence analysis on amino acid frequencies for the *Paramecium bursaria* proteins. Ribosomal proteins are indicated by red circles, proteins annotated as '*Paramecium* surface antigen domain' (PF01508) are indicated by green circles, 'Subtilisin-like proprotein convertase' (KOG3525O) are indicated by blue circles, and 'Giardia variant-specific surface protein' (PF03302) are indicated by light blue circles, and the remaining proteins are indicated by black circles.

Appendix A. Supplementary data

Table S1. Data for *Paramecium bursaria* transcripts. The columns are as follows: Trinity sequence name, the length in amino acids (Laa), the relative frequency of aromatic amino acids (AROMA), the mean hydropathicity (GRAVY), the mean molecular weight (MMW), the G+C content at the third codon positions (GC3), the G+T content at the third codon positions (GT3), the first two axes (SCU1 and SCU2) generated by within-group correspondence analysis on codon frequencies, the first two axes (AAU1 and AAU2) generated by correspondence analysis on amino acid

frequencies, functional annotations from COG (KOG), Pfam and InterPro databases, and statistics obtained by gene expression analysis (logFC, logCPM, PValue, and FDR). Transcripts are sorted by the value of SCU1.

Table S2. Basic statistics for three transcriptome assemblies obtained by Trinity (trinityrnaseq_r2012-04-27 and trinityrnaseq_r20140717) and SOAPdenovo-Trans (soapdenovo-trans-1.04).

Figure S1. Log-log plot of the predicted protein lengths (the numbers of amino acids) in the SOAPdenovo-Trans assembly (ordinate) versus those in the Trinity assembly (abscissa). The 38,183 proteins from the three transcriptome assemblies obtained by Trinity (trinityrnaseq_r2012-04-27 and trinityrnaseq_r20140717) and SOAPdenovo-Trans were classified into 6,224 protein groups using FastOrtho (http://enews.patricbrc.org/fastortho/). Of the 6,224 protein groups, 1,329 were present in a single assembly, 4,895 were present in two or more assemblies, of which 2,873 were shared by all the assemblies. For the 1,069 protein groups shared by all the assemblies and contained only a single copy from each assembly, we compared the corresponding protein lengths between Trinity (trinityrnaseq_r20140717) and SOAPdenovo-Trans. In 63% of cases, the Trinity assembly had a longer protein length than the SOAPdenovo-Trans assembly, in 10% of cases, the SOAPdenovo-Trans assembly had the longer protein length, and in 27% of cases did the two assemblies have the same protein length. The median value of the protein lengths in the Trinity assembly (317 aa) was higher than that in the SOAPdenovo-Trans assembly (175 aa). A Wilcoxon signed rank test, which compared the protein lengths for the two assemblies, was highly significant ($P < 2.2e-16$). Thus, in general, the protein lengths tended to be longer when using Trinity than when using SOAPdenovo-Trans.

Figures

Figure 1.

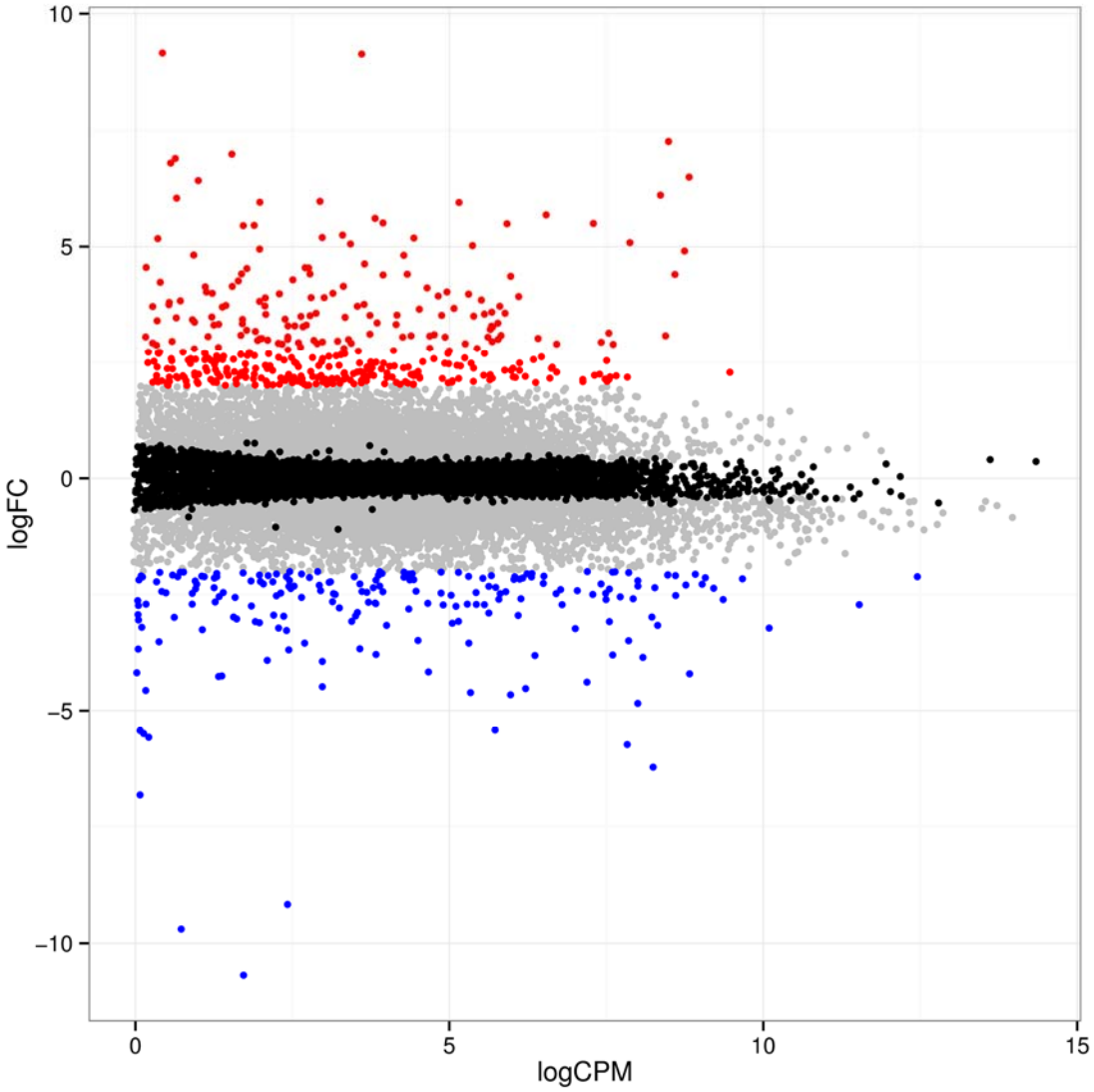


Figure 2.

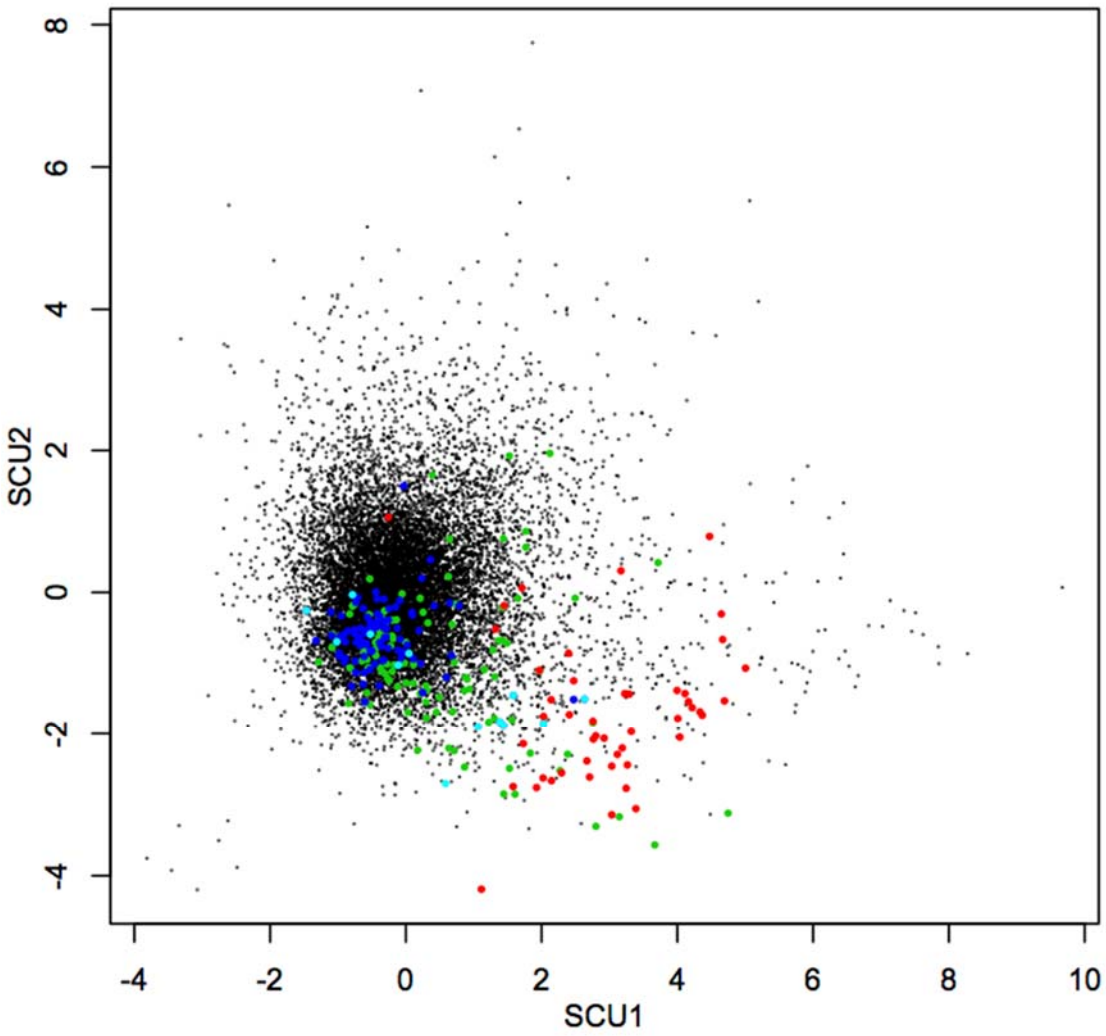
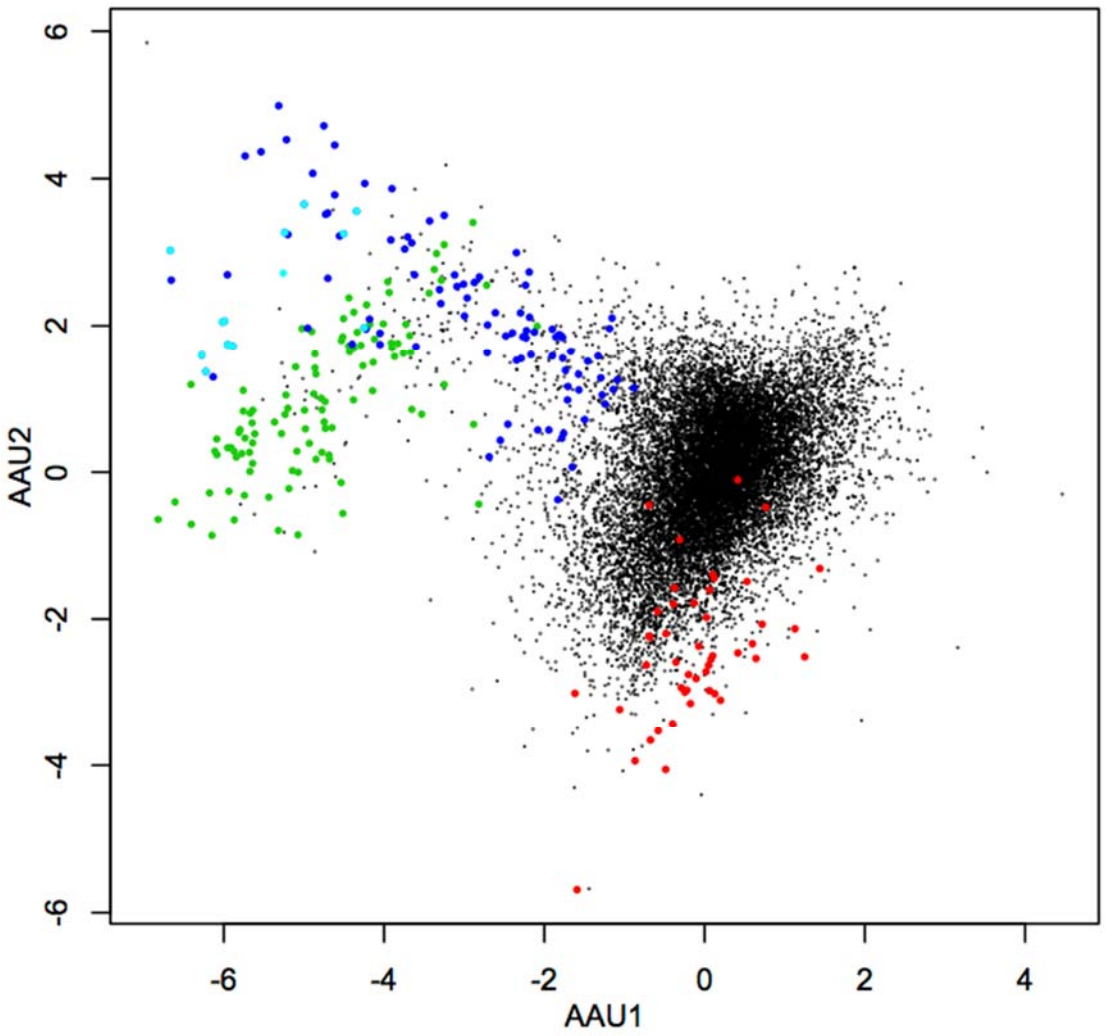


Figure 3.



Tables

Table 1. Codon usage summed for groups of genes in *Paramecium bursaria*.

Amino acid	Group of genes				
codon	All	Ribosomal	PF01508	PF03302	KOG3525O
Agca	103217	255	2374	294	2002
Agcc	32461	142	777	99	595
Agcg	11412	3	194	17	185
Agct	92291	308	2381	262	1779
Ctgc	42332	30	3100	635	4261
Ctgt	71106	24	4589	931	7225
Dgac	65061	77	556	50	1151
Dgat	283545	266	2806	254	5484
Egaa	282374	308	748	96	2732
Egag	134293	124	268	29	1020
Fttc	109735	177	565	131	1784
Fttt	224749	113	1270	189	4460
Ggga	131676	374	2936	527	3560
Gggc	25350	21	494	65	748
Gggg	27319	10	432	46	648
Gggt	65658	197	1328	204	1849
Hcac	27243	95	95	8	232
Hcat	85506	96	220	12	771
Iata	243333	76	1212	173	4499
Iatc	95606	191	527	75	1516
Iatt	240562	253	1377	191	4726
Kaaa	371510	538	2385	149	4123
Kaag	183755	541	745	45	1406
Lcta	89450	38	567	100	1573
Lctc	39596	84	234	33	553
Lctg	30698	3	141	15	486
Lctt	88497	127	547	90	1408

Ltta	257908	262	1775	325	5124
Lttg	138470	151	722	171	2227
Naac	97456	123	1046	191	2408
Naat	336907	247	3963	737	9191
Pcca	90287	201	950	219	1910
Pccc	23746	52	193	41	494
Pccg	8293	0	84	15	193
Pcct	77322	78	728	154	1764
Qcaa	135510	92	1040	186	2556
Qcag	36211	10	262	27	607
Qtaa	298511	309	2740	504	6468
Qtag	124248	74	778	158	2065
Raga	160067	675	661	94	1735
Ragg	43530	24	173	16	427
Rcga	19139	35	76	7	219
Rcgc	2791	2	9	0	25
Rcgg	2733	0	19	2	28
Rcgt	7076	12	27	1	66
Sagc	39579	40	551	110	1053
Sagt	88981	63	1420	257	2558
Stca	145204	199	3113	607	4795
Stcc	36967	49	554	84	990
Stcg	25506	7	355	55	718
Stct	91706	60	1768	270	2581
Taca	128630	253	4165	709	4549
Tacc	36613	50	917	173	1238
Tacg	15822	2	328	50	487
Tact	118110	110	3344	538	4321
Vgta	79611	97	792	143	1810
Vgtc	42722	188	356	87	740
Vgtg	50680	51	322	53	902
Vgtt	134943	328	1191	243	2683

Ytac	74579	71	659	172	1923
Ytat	224028	170	1978	514	6028

PF01508 = *Paramecium* surface antigen domain

PF03302 = Giardia variant-specific surface protein

KOG3525O = Subtilisin-like proprotein convertase

Table 2. Amino acid usage summed for groups of proteins in *Paramecium bursaria*.

	Group of proteins				
Amino acid	All	Ribosomal	PF01508	PF03302	KOG3525O
A	239381	708	5726	672	4561
C	113438	54	7689	1566	11486
D	348606	343	3362	304	6635
E	416667	432	1016	125	3752
F	334484	290	1835	320	6244
G	250003	602	5190	842	6805
H	112749	191	315	20	1003
I	579501	520	3116	439	10741
K	555265	1079	3130	194	5529
L	644619	665	3986	734	11371
M	125473	87	340	78	1641
N	434363	370	5009	928	11599
P	199648	331	1955	429	4361
Q	594480	485	4820	875	11696
R	235336	748	965	120	2500
S	427943	418	7761	1383	12695
T	299175	415	8754	1470	10595
V	307956	664	2661	526	6135
W	50970	68	666	46	691
Y	298607	241	2637	686	7951

PF01508 = *Paramecium* surface antigen domain

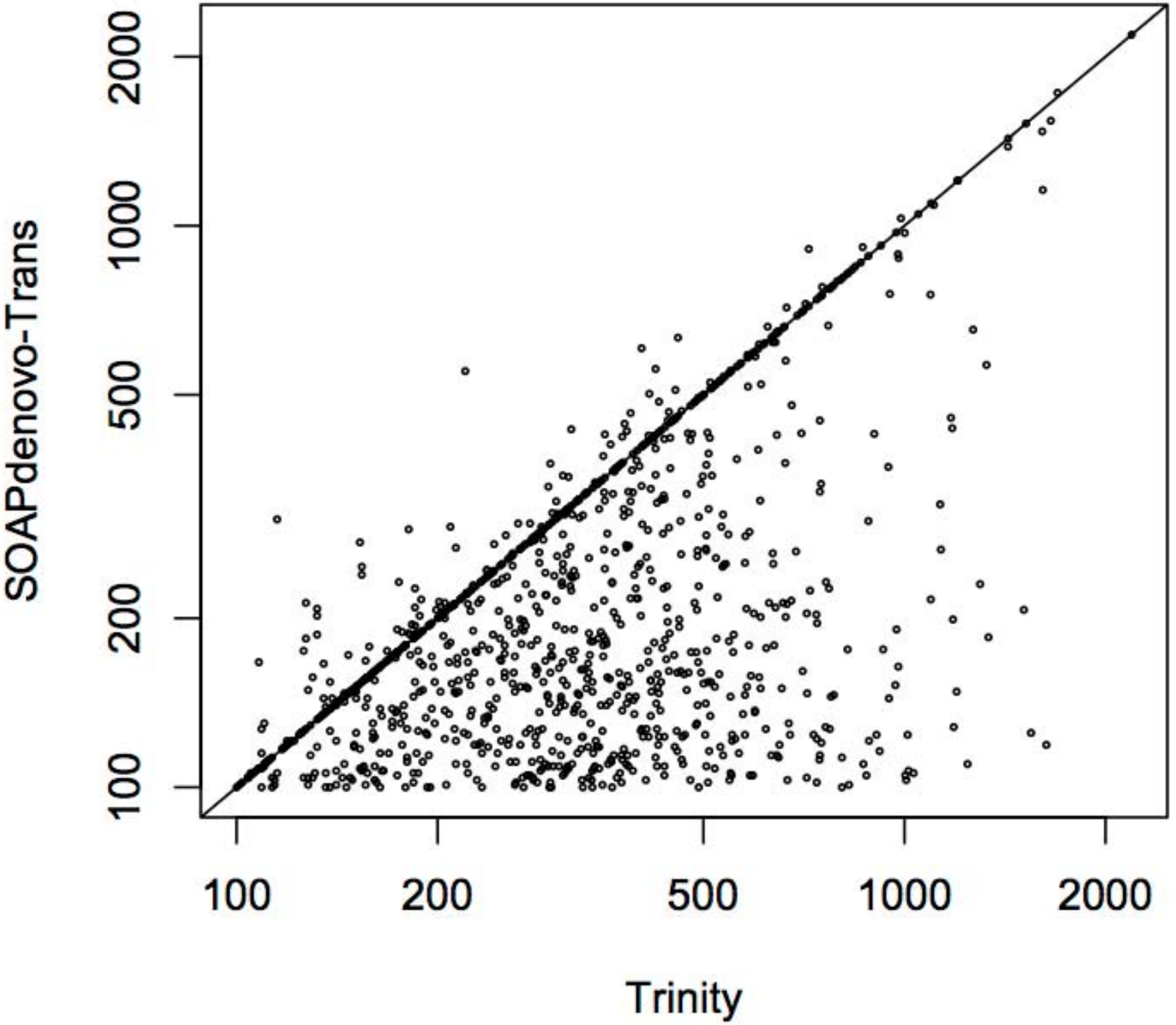
PF03302 = Giardia variant-specific surface protein

KOG3525O = Subtilisin-like proprotein convertase

Table S2. Basic statistics for three transcriptome assemblies obtained by Trinity (trinityrnaseq_r2012-04-27 and trinityrnaseq_r20140717) and SOAPdenovo-Trans (soapdenovo-trans-1.04).

	de novo transcriptome assembler		
	trinityrnaseq_r2012-04-27	trinityrnaseq_r20140717	soapdenovo-trans-1.04
Sequence number	40,805	57,890	72,291
Total bases of sequence	36,894,860	50,014,551	40,194,578
Maximum sequence leng	22,858	30,375	13,967
Average sequence lengtl	904	864	556
Median sequence length	406	517	361
N50 sequence length	1,832	1,350	689
Sequences filterd	10,557	15,005	21,861
aa > 99	10,134	14,252	13,771
FDR < 0.05	6,433	9,142	7,064
logFC ≧ 2	271	389	511
logFC ≦ -2	135	203	204
-2 < logFC < 2	6,027	8,550	6,349

* Trinity contigs used in this statistics were longest isoforms only to compare SOAPdenovo-Trans contigs.


SOAPdenovo-Trans
100
200
500
1000
2000
Trinity
100
200
500
1000
2000