www.sciencemag.org/cgi/content/full/329/5988/223/DC1

Supporting Online Material for

Genomic Analysis of Organismal Complexity in the Multicellular Green Alga *Volvox carteri*

Simon E. Prochnik, James Umen,† Aurora Nedelcu, Armin Hallmann, Stephen M. Miller, Ichiro Nishii, Patrick Ferris, Alan Kuo, Therese Mitros, Lillian K. Fritz-Laylin, Uffe Hellsten, Jarrod Chapman, Oleg Simakov, Stefan A. Rensing, Astrid Terry, Jasmyn Pangilinan, Vladimir Kapitonov, Jerzy Jurka, Asaf Salamov, Harris Shapiro, Jeremy Schmutz, Jane Grimwood, Erika Lindquist, Susan Lucas, Igor V. Grigoriev, Rüdiger Schmitt, David Kirk, Daniel S. Rokhsar†

†To whom correspondence should be addressed. E-mail: umen@salk.edu (J.U.); dsrokhsar@gmail.com (D.S.R.)

Published 9 July 2010, *Science* **329**, 223 (2010)
DOI: 10.1126/science.1188800

This PDF file includes:

Supplemental Online Material for the Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*

Table of Contents

1) MATERIALS AND METHODS

A. Nuclear genome sequencing and assembly

We prepared high quality genomic DNA from a vegetative culture of female *Volvox carteri f. nagariensis,* Eve (S*1*), a subclone of HK10 (S*2, 3*), which is a standard female lab strain of *Volvox carteri* (hereafter *Volvox*) that was originally isolated in 1965 by Richard Starr from a pond associated with a rice paddy near Kobe, Japan. The genomic DNA was prepared by a standard protocol involving CsCl gradient banding to separate it from RNAs (S*1*), but it could not be separated from chloroplast and mitochondrial DNA. The genome sequences of these two organelles have already been determined (S*4*).

Paired-end whole-genome shotgun (WGS) sequencing (S*5*) of three libraries with insert sizes of 2-3 kb (AOBN); 6-8 kb (ABSY) and 35-40 kb (AOBO) generated 1,430,397, 1,269,395 and 230,112 reads respectively, covering 1,361, 1,310 and 235 Mb raw sequence respectively, together totaling 2,906 Mb of raw sequence. The reads were screened for vector sequence using Cross_match (S*6*) and trimmed for vector and low quality sequences. Reads shorter than 100 bases after trimming were excluded from the assembly leaving 1,343,753 2-3 kb insert reads (94%, 836 Mb of sequence); 1,207,057 6-8 kb insert reads (95% 760 Mb) and 224,372 35-40 kb insert reads (98%, 113 Mb).

The filtered and trimmed read sequences were assembled using JAZZ 1.0.3 (S*7*). A word size of 14 was used for seeding alignments between reads. The 'unhashability threshold' parameter was set to 40, meaning that words present over 40 times in the data set were not used to seed alignments. A mismatch penalty of -30.0 was used that generally allows assembly of sequences that are more than ~97% identical.

The initial assembly contained 147.4 Mb of scaffold sequence, of which 12.5 Mb (8.5%) was gaps. There were 7,391 scaffolds, with a scaffold N50/L50 of 35/1.41 Mb, and a contig N50/L50 of 795/42.7 kb Scaffolds < 1 kb long as well as redundant scaffolds (those scaffolds shorter than 5kb long with >80% identity to another scaffold whose length was greater than 5kb) were removed from the assembly. This left 141.5 Mb of scaffold sequence, of which 12.4 Mb (8.8%) was gaps. The filtered assembly contained 1,327 scaffolds, with a scaffold N50/L50 of 33/1.50 Mb, and a contig N50/L50 of 729/45.4 kb. The sequence depth derived from the assembly was 11.1 ± 0.2.

To estimate the completeness of the assembly with respect to transcribed genes, 72 *Volvox* mRNAs that were known prior to the genome project were downloaded from the nr database at NCBI (S*8*) and aligned to the assembly using BLAT (S*9*) with default parameters. All 72 mRNAs had hits to the assembly with >97% identity over most of their lengths.

As a second test of completeness relative to transcribed loci, we considered 129,528 dideoxy-sequenced ESTs that had ≤40% of unmasked sequence after removal of low complexity and simple repeat regions (see below). Of these ESTs, 127,056 (98.0%) aligned to the assembly with BLAT (S*9*) (>90% identity over > 50% of their length). The 2,472 filtered ESTs that did not align to the genome were examined further. Approximately 1/3 (857) had hits with BLASTX (S*10*) (E-value < 1e-10) to known proteins from the UniProt database (S*11*). These included 408 ESTs (48% of unmapped ESTs with hits) with best hits to proteins annotated as "ribosomal protein" and 93 ESTs (11% of unmapped ESTs with hits) so annotated as related to chlorophyll binding. We do not rule out the possibility that these and other unmapped ESTs are derived from loci not included in the genome assembly because they are embedded in repetitive sequence. Overall, we can conservatively estimate that the completeness of the *Volvox* genome assembly with respect to transcribed loci captured by ESTs is likely better than 98%.

B. Comparison and annotation of repeats in *Volvox* and *Chlamydomonas*

B1. Overview of repeat analysis

The *Volvox* genome assembly is 19,621,448 bp longer than that of *Chlamydomonas reinhardtii* (hereafter *Chlamydomonas*) (Table 1). We compared the repeat content of the two genomes to determine the contribution made by repeats to the difference in genome size. To do this, we built and annotated a custom repeat library for each algal genome and ran RepeatMasker (S*12*) on each assembly with the appropriate custom repeat library and the '-gccalc' option (Table S1).

The custom *Volvox* repeat library was assembled from five component libraries:

i) 45 *Volvox carteri*-specific and 72 *Chlamydomonas*-specific repeat sequences from RepBase (20080611 update) (S*13*);

ii) 147 sequences that had been generated by analysis of the *Chlamydomonas* genome (S*14*);

iii) 33 repeat elements from the *Volvox* assembly that were generated using the same approach as had been used previously for the *Chlamydomonas* genome (S*14*). (We estimate the curated set from the *Volvox* genome is 20-25% complete);

iv) a library of 1,704 satellite repeat sequences (with lengths ranging from 20 to 1,162 bp) built by searching the whole genome shotgun reads for over-represented 16-mers and assembling overlapping 16-mers (as described below), and

v) 1,511 repeats identified by RepeatScout (S*15*) (Table S16). The repeat sets were annotated and filtered leaving 1,449 sequences (as described below).

In parallel, a custom *Chlamydomonas* repeat library was assembled from:

i) *Volvox*- and *Chlamydomonas*-specific repeat sequences from RepBase (20080611 update) (S*13*));

ii) 147 *Chlamydomonas* repeat sequences identified as in ii) above;

iii) 33 *Volvox* repeats identified as in iii) above;

iv) a library of 100 satellite sequences (with lengths 25, 92, 107, 181 or 184 bp) (see below) and

v) 1,057 repeats identified by RepeatScout. After filtering the library contained 1,013 repeats (Table S16 and see below).

B2. Analysis of satellite repeats in *Volvox* and *Chlamydomonas*

A library of all 16 nt long sequences (16-mers) that occur at least 500 times was generated from approximately half the WGS reads (all reads from the AOBN library). 16-mers that overlap each other were assembled into longer sequences by repeatedly looking for 15 nt overlaps and extending by a single nucleotide overhang until either no further extensions were possible (in which case extensions in the opposite direction were explored) or the sequence looped back on itself. Both the sequences that could not be extended further and the circular sequences were added to the library of putative satellite sequences as long as they were at least 20 nt long.

B3. Generation annotation and filtering of RepeatScout Libraries

Generation of libraries of repeats with RepeatScout (S*15*) and their subsequent filtering and annotation was accomplished as follows. First, RepeatScout was run on the *Volvox* assembly. This produced a library of 1,511 repeat sequences (Table S16). Next RepeatScout was run on the *Chlamydomonas* assembly, generating 1,057 sequences. The repeat sequences in these two libraries were classified as described in the set of rules below.

To annotate and filter repeat sequences in the RepeatScout libraries generated from the *Volvox* and *Chlamydomonas* genomes, we first masked the *Volvox* genome with the 1,511 sequence RepeatScout library using RepeatMasker (S*12*) with the '-gccalc' option. We then counted the number of times each repeat sequence hit the genome. We also counted the percentage of repeat instances in the genome that also overlapped gene models and ESTs by two criteria: ≥200 nt length and ≥80% of the length of the repeat. Sequences in the repeat library were assigned Pfam domains by running HMMPFAM, part of the HMMER package (S*16*), on the library with an E-value cutoff of 1E-5. Repeat sequences with Pfam domain assignments were sub-divided into those with a TE-associated Pfam (PF00075, PF00078, PF00665, PF03372, PF03732, PF07727, PF01527) and those with non-TE associated Pfam domains (all other Pfam domains). We also ran tRNAScan-SE (S*17*) on the repeat sequences.

To assign TE classes to sequences in the *Volvox* RepeatScout library that have homology to known TE classes, we ran RepeatMasker on the *Volvox* RepeatScout library with each of two repeat libraries (as these two libraries contain partially overlapping sequences): in the first run, the custom library of repeats that we had curated manually (see above) was used to mask the RepeatScout repeat library; in the second run, RepeatMasker was run with the option '-species chlamydomonodales' to use the volvocine algae repeat sequences in the 20080611 release of RepBase Update (S*13*). In cases where the longest repeat that masked a RepeatScout library sequence was in the class 'Simple_repeat' or 'Low_complexity', this annotation was ignored as RepeatMasker has dedicated algorithms for finding repeats of these two classes that are based solely on sequence composition, rather than homology to known TEs. In cases where the longest annotation in the RepeatScout repeat sequence was not a Simple_repeat or non Low_complexity-repeat, the repeat sequence was assigned the class 'Complex_repeat'. If the RepBase Update library found a complex repeat and our curated library did not, then the complex repeat that was found was used for the classification.

For all the sequences still without a 'Complex_repeat' classification, in which either RepeatMasker detected a tRNA in the sequence or tRNAScan-SE predicted a tRNA with score > 22 and the length of the repeat < 120 nt, the repeat was given the classification 'tRNA'.

Sequences were classified as 'Satellite' or 'rRNA' if RepeatMasker assigned either of these classifications to a sequence.

Sequences that still had not been given a classification and also had Pfam domains were classified 'non_TE_PFAM' if the Pfam domain is not associated with TEs or 'TE_associated_PFAM' if the Pfam domain is associated with TEs (see above).

122 repeat sequence that still had not been classified met all of the following three criteria and we therefore reasoned that these repeat may be novel and classified them as 'Putative_novel' (Table S16). The three criteria were:

> i) either there were no instances of the repeat sequence in the genome that overlapped an EST by at least 200 bp or no instances in the genome that overlapped an EST by at least 80% of the length of the repeat sequence;

ii) either there were no instances of the repeat sequence in the genome that overlapped a gene model by at least 200 bp or no instances in the genome that overlapped a gene model by at least 80% of the length of the repeat sequence; and

iii) the length of the repeat was over 500 nt.

The remaining 911 sequences were classified 'Unknown'. To see if these unknown repeats could be classified further, InterProScan (S*18*) was run on the 911 sequences to assign Pfam domains using specific gathering thresholds for each HMM. This is more accurate than using a single E-value cutoff for all domains. Hits were manually inspected and 62 sequences with Pfams that are not associated with TEs were deleted from the RepeatScout library. This left 1,449 (Table S16).

A parallel analysis in *Chlamydomonas* starting with a RepeatScout library of 1,057 sequences produced a filtered and annotated set of 1,103 sequences (Table S16)

C. Analysis of repeat expansions

The *Volvox* genome was masked with RepeatMasker using the RepeatScout library (see above), which was annotated as described above. All repeat sequences in the *Volvox* genome longer than 500 nt and belonging to a known class of TE were collected and their Jukes-Cantor distance, corrected for multiple substitutions ($K=-3/4 \times \ln(1-4i/3)$, where i is percent nucleotide dissimilarity from the repeat consensus) from the RepeatScout consensus repeat sequence were plotted in a histogram (Fig. S4A-C). A parallel analysis was performed for *Chlamydomonas* (Fig. S4D-E).

Bursts of TE expansion appear as secondary peaks in the histogram to the right of the descending curve that starts at a Jukes-Cantor distance of zero. No secondary peaks are apparent in the total repeat histograms for *Volvox* or *Chlamydomonas* (Fig S4A,4D), but they are present in plots for specific TE families such as Gypsy and Copia in *Chlamydomonas* (Figs. S4E,4F).

D. Calculation of corrected 4-fold degenerate transversion (4DTV) distances

The frequency of transversions at the third position of four-fold degenerate codons (4DTV) can be used to measure the rate of neutral evolution as these transversions do not change the amino acid that is encoded. We calculated 4DTV distances between orthologous protein sequences in pairs of genomes using a previously described method (S*19*). Briefly, we identified a set of mutual best BLASTP hits (MBH) between all predicted proteins in each pair of species and used them to align coding regions. The number of transversions at conserved four-fold degenerate sites divided by the total number of four-fold degenerate sites gives the 4DTV frequency. This raw calculation is then corrected for multiple substitutions using the formula $4DTV_C = -1/2\ln(1-2 \times 4DTV_U)$, where $4DTV_C$ is the corrected 4DTV and $4DTV_U$ the uncorrected 4DTV.

E1. Synteny and genomic rearrangements

Synteny dotplots for *Volvox*-*Chlamydomonas* and human-chicken are shown in Fig. S5 and reveal the extent of conserved gene order.

We used the updated *Volvox* v2 assembly (http://genome.jgi-psf.org/Volca1/Volca1.download.ftp.html) and the *Chlamydomonas* v4 assembly (http://www.phytozome.net/chlamy) for the following analysis of synteny between *Volvox* and *Chlamydomonas*. The *Chlamydomonas* v4 assembly has 17 chromosomes and 61 minor scaffolds; the *Volvox* v2 assembly has 434 scaffolds (compared to 1,265 for v1).

At the time of analysis, neither the Volvox v2 assembly nor the *Chlamydomonas* v4 assembly had been annotated with gene model annotations so we mapped *Volvox* v1 and *Chlamydomonas* v3.1 transcripts to their respective updated assemblies using blat (S*9*) with default parameters and taking the best hit to the assembly. After mapping and filtering (see below), 4,349 of the 4,804 (91%) *Volvox* gene models were on scaffolds containing 25 or more genes, permitting useful synteny analysis.

Syntenic segments were constructed between pairs of genomes as follows. We only considered the longest gene model at any locus because the commonest problem with gene prediction for a genome with incomplete EST coverage is truncation.

1) Gene models whose translations did not have a WU-BLASTP (S*10*) hit to the other proteome (E-value < 1E-10) were removed.

2) Tandem expansions were collapsed: if two or more neighboring genes encode similar proteins (WU-BLASTP E-value < 1E-10) and had no more than 2 intervening genes, only the longest gene model of the two or more similar, neighboring, genes was retained as a representative of the duplication.

3) Gene models whose best hit to the other proteome had a C-score (see below) less than 0.8 were removed.

4) Gene models with more than 10 hits (E-value < 1E-10) to the other proteome were removed because large gene families can seed false syntenic blocks in many different genomic locations.

5) The remaining gene models were ordered along chromosomes (or scaffolds in the case of the *Volvox* assembly). The chromosomes/scaffolds in Fig. S5 were arranged in decreasing order of the numbers of gene models contained. In multiple iterations, the gene models were used to seed syntenic blocks (defined as containing two or more genes with conserved gene order) in each genome with different numbers of intervening genes in the range zero to ten being picked in each iteration (data from zero to four intervening genes are shown in Table S4).

6) As the number of intervening genes allowed between two genes in a syntenic block increases, so does the chance of finding such blocks by chance. In order to establish a "null" model for each condition the order of the filtered genes was scrambled and the number of syntenic blocks formed with different number of intervening genes was determined (Table S4).

The number of genes remaining in syntenic blocks after this filtering process is shown in Table S3. A comparison of the synteny dotplots of *Volvox* vs. *Chlamydomonas* (Fig. S5A) and human vs. chicken (Fig. S5B) shows that the human-chicken genes tend to lie on longer (up to whole chromosome arm) syntenic segments than in the two algae. Furthermore, the syntenic blocks that are present in the algae are broken up by micro-inversions to a greater extent. Overall, there has been less overall rearrangement in vertebrates (Fig. S5B) than in *Volvox-Chlamydomonas* Fig. S5A.

Where whole genome duplication (WGD) has taken place, it is visible in plots of this type as repeated diagonal stretches in a row or column. There is no evidence of WGD in *Volvox*, *Chlamydomonas*, or their common ancestor (Fig. S5), unlike yeasts, higher plants and metazoans (S*20*) where WGDs have played a significant role in genome evolution.

E2. Definition of C-score

We used the metric C-score as a measure of similarity between a protein from one predicted proteome and the proteins from a second predicted proteome. The C-score for protein X in one species and protein Y in a second species (C_{XY}) is defined as the BLAST score of X against Y divided by the best BLAST score for protein X against all of the proteins in species Y. The C-score can be used to detect the presence of both orthologs (defined as mutual best BLAST hits) as well as potential paralogs. If X and Y are mutual best hits, then C_{XY} and C_{YX} will both equal 1. Recent paralogs of X will have a C-score of slightly less than 1 relative to Y; similarly, recent paralogs of Y will have a C-score of slightly less than 1 relative to X.

F. Loss of synteny through genomic rearrangements

To quantify the amount of rearrangement on the gene by gene scale, we used the following metric: we calculated the fraction of all pairs of neighboring syntenic orthologs from each set of two genomes (ascertained in the previous section) that were not adjacent to each other in the other genome in the pair, reasoning that this would have been caused by a rearrangement since the two genomes diverged (Table S2).

G. cDNA library construction and EST sequencing

We extracted total RNA from *Volvox carteri f. nagariensis* female strains Eve and Eve10 and male strain 69-1b. For Eve and 69-1b, we extracted RNA from samples 1.5, 10, 24, 48 hours after sexual-induction and pooled the samples. For Eve10, we extracted RNA from 2-4 and 32-128 cell stages and pooled the samples. Poly A^+ RNA was isolated from total RNA using the Absolutely mRNA Purification kit and manufacturer's instructions (Stratagene, La Jolla, CA). cDNA synthesis and cloning used a modified procedure based on the "SuperScript plasmid system with Gateway technology for cDNA synthesis and cloning" (Invitrogen, Carlsbad, CA). 1-2 µg of poly A^+ RNA, SuperScript II reverse transcriptase (Invitrogen) and oligo dT-*Not*I primer (5' GACTAGTTCTAGATCGCGAGCGGCCGCCCT$_{15}$VN 3' ,where V is any nucleotide except T and N is any nucleotide) were used to synthesize first strand cDNA. Second strand synthesis was performed with *E. coli* DNA polymerase I, DNA ligase, and RNaseH followed by end repair using T4 DNA polymerase. An adaptor including the overhanging pre-cut *Sal*I site at the 5' end (5' TCGACCCACGCGTCCG 3' and 5' CGGACGCGTGGG 3') was ligated to the cDNA that was then digested with *Not*I (New England BioLabs, Ipswich, MA), and size selected by gel electrophoresis (1.1% agarose). The cDNA inserts were ligated into the *Sal*I and *Not*I digested vector pCMVsport6 (Invitrogen). The ligation was transformed into ElectroMAX T1 DH10B cells (Invitrogen). In total, five cDNA libraries were constructed.

Library quality was assessed in two ways. First we ensured that the number of clones without inserts was less than 10% by randomly selecting 24 clones and PCR amplifying the cDNA inserts with the primers M13-F (5' GTAAAACGACGGCCAGT 3') and M13-R (5' AGGAAACAGCTATGACCAT 3'). Second, a test production run of a single 384-well plate was undertaken (as described below) and sequence quality, diversity and length were investigated. For the main production run, cells from each library were plated onto agarose plates (254 mm plates from Teknova, Hollister, CA) at a density of approximately 1,000 per plate. Plates were grown at 37°C for 18 hours then individual colonies were picked and each used to inoculate a well containing LB media with appropriate antibiotic in a 384 well plate (Nunc, Rochester, NY). Clones were grown in selective media in 384 well plates and plasmid DNA for sequencing was produced by rolling circle amplification (S*21*) (Templiphi, GE Healthcare, Piscataway, NJ). Inserts were sequenced from both ends using primers complimentary to the flanking vector sequence with the following sequences: Fwd: 5' ATTTAGGTGACACTATAGAA and Rev: 5' TAATACGACTCACTATAGGG) and Big Dye terminator chemistry on ABI 3730 DNA Analyzers (ABI, Foster City, CA). We generated pairs of reads (from both 5' and 3' ends of each cDNA clone), generating 42,240, 51,456 and 72,192 reads from Eve, Eve10 and 69-1b respectively, giving a grand total of 165,888 ESTs (Expressed Sequence Tags).

All 165,888 ESTs were processed through the JGI EST pipeline. Phred (S*6, 22*) was used to call bases and generate quality scores. Vector, linker, adapter, poly-A/T, and other artifact sequences were removed using the Cross_match software (S*6, 22*) and an internally-developed short pattern finder. Low quality regions of the read were identified using internally-developed software, masking regions with a combined quality score of less than 15. The longest high quality region of each read was considered to be the sequence of the EST. ESTs shorter than 150 bp as well as those containing common contaminating sequences from e.g. *E. coli*, common vectors, and sequencing standards were removed from the data set. After these filtering steps, 132,038 ESTs were left (33,407, 37,354, and 61,277 from Eve, Eve10 and 69-1b respectively. An additional 2,510 ESTs were not included in the analysis of assembly completeness (see above) due to their having > 40% low complexity and repetitive sequence as determined by mdust (S*23*) run with the '-v 20' setting. This left 129,528 ESTs for consideration in analysis of assembly completeness.

Clustering the EST sequences involved first generating all-by-all pairwise alignments between the 132,0338 filtered reads. ESTs sharing an alignment of at least 98% identity were then assigned to the same cluster. In addition, ESTs not sharing alignments but derived from opposite ends of the same cDNA clone were assigned to the same cluster. Clusters of ESTs were assembled into consensus sequences, contigs or singlets using CAP3 (S*24*). A total of 16,569 assembled consensus sequences were generated.

H. Prediction of gene models

The 1,265 *Volvox* v.1 scaffolds were masked using RepeatMasker (http://www.repeatmasker.org/) and a library of 1,015 transposable elements (TEs), including manually curated *Volvox* and *Chlamydomonas* TEs (http://www.girinst.org/).

After masking, the JGI annotation pipeline was used to generate gene models. This pipeline employs gene prediction programs that are based on a variety of methods, as follows:

1) *ab initio* methods (FGENESH; http://www.softberry.com/);

2) homology-based methods (FGENESH+ and Genewise; http://www.ebi.ac.uk/Wise2/) seeded by Blastx alignments against sequences of nr, IPI (http://www.ebi.ac.uk/IPI/), and JGI *Chlamydomonas* annotation v3 (http://www.jgi.doe.gov/chlamy/);

3) cDNA-based methods (EST_map; http://www.softberry.com/) seeded by 13,722 EST cluster consensus sequences derived from 87,866 *Volvox* ESTs. At the time the JGI annotation pipeline was run, 87,593 seqeuences had already been sequenced by the JGI (see above). The remaining 273 EST sequences were downloaded from the nr database at GenBank (S*8*);

4) synteny-based methods (FGENESH-2; http://www.softberry.com/) using the JGI *Chlamydomonas* assembly and annotation (http://www.jgi.doe.gov/chlamy/).

Genewise models were completed using scaffold data to find start and stop codons. EST clusters were used to extend, verify, and complete the predicted gene models. The resulting set of models was then filtered for the "best" models, based on criteria of completeness, length, EST support, and homology support, to produce a non-redundant representative set. This representative set was subject to protein functional analysis and manual curation, as described in the next sections.

The function of the translations of the predicted gene models was predicted using TMHMM (http://www.cbs.dtu.dk/services/TMHMM/), InterProScan (http://www.ebi.ac.uk/interpro/), and hardware-accelerated double-affine Smith-Waterman alignments (http://www.timelogic.com/decypher_sw.html) against SwissProt (http://www.expasy.org/sprot/), KEGG (http://www.genome.jp/kegg/), and KOG (http://www.ncbi.nlm.nih.gov/COG/). Finally, KEGG hits were used to map EC numbers (http://www.expasy.org/enzyme/), and Interpro and SwissProt hits were used to map GO terms (http://www.geneontology.org/).

We initially predicted 15,544 gene models in the genome of *Volvox*. 23% of these gene models were seeded by alignments of proteins in nr against the *Volvox* genome, while 67% were predicted *ab initio* and 10% were seeded using synteny with *Chlamydomonas reinhardtii* gene models (Table S5). Complete models with start and stop codons comprise 85% of the 15,544 initial gene predictions; 34% are consistent with ESTs and 70% align with proteins in Swissprot (http://www.expasy.org/sprot/) (Table S6).

The average *Volvox* gene is 5.27 kb long, the average gene density is 113 genes/Mb, and the average transcript has 7.78 exons (Table S7). The average protein length is 558 aa. We predicted that 4% of the proteins possess at least one transmembrane domain, 30% possess a signal peptide, and 2% possess both. We assigned 1,757 distinct GO terms to 4,566 proteins (30%), and we assigned 3,062 proteins (20%) to KEGG pathways, totaling 625 distinct EC numbers. We assigned 9,889 proteins (64%) to 3,145 distinct KOGs.

Knowing that the repeat masking was incomplete, as a last step, we filtered the initial set of 15,544 gene models, removing all those that endcoded proteins with homology to transposable elements or were assigned TE-associated Pfam domains by InterProScan (S*18*). 1,103 protein models were removed from the set of 15,544, leaving 14,520 (Table S6).

Web-based editing tools available at the JGI genome portal were used to examine and improve predicted gene structures, and to record textual annotations and protein function. As of December 15, 2009, 1,628 genes (11%) have been manually curated. All annotations, both automatic and manual, may be viewed at a dedicated JGI portal (http://www.jgi.doe.gov/volvox/).

I. *Volvox* has longer introns than *Chlamydomonas*

The median intron size in *Volvox* is about twice that of *Chlamydomonas* (358 bp vs. 174 bp; Table 1, Fig. S6). (Mean length and S.D. in *Volvox* are 491 bp and 749 respectively, and 371 bp and 527 in *Chlamydomonas* respectively (Table S7)) This differential accounts for 10.5 Mb of the longer assembly in *Volvox*. 3.5Mb of the *Volvox* introns are made up of repeats, the composition of which reflects the overall repeat class composition of the genome (see above). The length of introns at conserved positions between orthologous exons in *Volvox* vs. *Chlamydomonas* divide into three subpopulations (Fig. S6), each of which has a mean that is significantly different from the others (Welch's t-test, $p < 2.2E-16$). The majority (93%) orthologous introns are >100 bp long and show no size correlation between the two species (Pearson's $r^2 = 0.0044$), though the mean length in *Volvox* (440 bp) is significantly longer than in *Chlamydomonas* (313 bp) (Welch's t-test, $p<2.2E-16$). A small (3%) subset of introns are short (~60-100 bp) in both species lie near the diagonal (although only weakly correlated, Pearson's correlation coefficient = 0.39) suggesting the existence of a common yet unknown selective mechanism. The third small subset of *Volvox* introns (4%) are around 60-100 bp long but have an uncorrelated length in *Chlamydomonas* (Pearson's $r^2 = -0.0096$) and appear as a horizontal distribution across the bottom of the plot.

J. Pfam protein domain assignments

To assign Pfam domains to proteins in a predicted proteome, we made a set of the longest protein sequence at each locus and ran the HMMPFAM module within InterProScan (S*18*) with Pfam v20 on these sequences. This algorithm assigns Pfam domains based on the gathering threshold specific to each HMM rather than using the same E-value for every domain.

K. Pfam domain combinations unique to Chlamydomonodales

The last common ancestor of *Volvox* and *Chlamydomonas* is represented by the clade Chlamydomonodales (taxonomy ID 3042) in the NCBI taxonomy (S*8*). To compare the protein domains found in species inside this clade to those found outside, we took our Pfam annotations in *Volvox* and *Chlamydomonas* (see above) and added all Pfam domain annotations in Uniprot (ftp://ftp.pir.georgetown.edu/databases/iproclass/; release date 9/3/08) from all other species that are descended from the Chlamydomondales node.

To date, no protein domains unique to the Chlamydomondales have been deposited in the Pfam database. 2,650 domains are found in species within and outside the Chlamydomonodales while 7,690 are only found in species outside this group.

L. Pfam domain combinations specific to *Volvox* or specific to *Chlamydomonas* or both

We counted the number of different pairwise domain combinations in various species (Table S9), considering only unique pairs of protein domain types, regardless of how many times any domain occurs in a protein. In a search for Pfam domain combinations that are present in the volvocine algae (the clade represented by descendants of the last common ancestor of *Volvox* and *Chlamydomonas*), but not in other species, we found only a single domain combination in *Volvox* or *Chlamydomonas* and not other species in uniprot (ftp://ftp.pir.georgetown.edu/databases/iproclass/; release date 9/3/08). After this analysis, the JGI released a genome portal for another species in Chlorophyta, *Chlorella* sp. NC64A (http://genome.jgi-psf.org/ChlNC64A_1/ChlNC64A_1.home.html). The domain combination is found in *Chlorella* too. From this, we conclude that there are no volvocine algae-specific domain combinations.

We also found 199 domain combinations that are present in *Volvox* but not *Chlamydomonas* or other species and, conversely, 122 that are present in *Chlamydomonas* but not *Volvox* or other species. The majority of the gene models in these two sets have no EST support across their lengths and are on short, poorly assembled scaffolds that often include only one WGS read's length of sequence at each end and an internal gap several kb in length, suggesting that the gene models may span more than one genetic locus. This suggests there are few Pfam domain combinations found in one alga and not the other.

M. Construction of protein families

We compared the reference set of 14,520 predicted proteins from *Volvox* and 14,516 predicted proteins from *Chlamydomonas* to each other and to proteins from twenty other organisms spread across the entire tree of life, including animals, plants, fungi, amoebae, chromalveolates and bacteria (Table S10). [In addition to these species, the recently-published predicted proteomes of two *Micromonas* species (S*25*) were used in the analysis of protein families specific to the Volvocine algae (see below)]. Protein comparisons were performed using WU-BLASTP 2.0MP-WashU [04-May-2006] (S*10*) with filtering from low-complexity sequences and simple repeats and Smith-Waterman post-processing. (To determine the cutoff for protein family construction, we manually examined BLASTP alignments at different E-values. We found that a cutoff of E-value < 1E-10 included proteins with distinct regions of homology compared to E-values ≥ 1E-10 that had scattered regions of similarity in the alignments that appeared to be present by chance.) Mutual best hits (E-value < 1E-10) between a protein in *Volvox* and a protein in any of the 21 other species including *Chlamydomonas* (as well as mutual best hits between a protein in *Chlamydomonas* and a protein in any of the 21 other species including *Volvox*) were used to establish orthology. Paralogs were added according to empirically-determined criteria that include in-paralogs. In a final step, proteins that were not in families were pledged to a family if their best hit (E-value < 1E-20, coverage >50%) was in a family, another good hit (E-value < 1E-20, coverage > 50%) was in the same family, and the family had 50 or fewer proteins in it before pledging. This E-value and coverage cutoffs were determined by chosing a few dozen families and comparing the range of E-values and coverages of proteins within families to those of proteins that had similarity, yet had not been included in the protein familes, making them candidates for pledging.

There are 7,612 mutual best hit relationships between *Volvox* and *Chlamydomonas* proteins. These, together with 168 mutual best hits between another species and either *Volvox* or *Chlamydomonas* form the backbone of 7,780 families (with the latter 168 families lacking proteins from either *Volvox* or *Chlamydomonas*). After addition of paralogs 7,293 contain 9,311 (64%) *Volvox* proteins and 7,233 contain 9,189 (64%) *Chlamydomonas* proteins. We found that 3,683 families (containing 3,809 *Volvox* proteins) are also conserved in moss (5,765 proteins) and 3,204 families (containing 3,309 *Volvox* proteins) are also conserved in Arabidopsis (4,141 proteins).

Notably, 10 of these families have a single member in *Chlamydomonas* and more than five members in *Volvox* whereas only two families have a single *Volvox* member and more than five *Chlamydomonas* members (Table S11). There are only 80 families (1.1%) with over 5 proteins from *Chlamydomonas* and/or *Volvox*. 295 families contain a single *Volvox* protein and 2-5 *Chlamydomonas* proteins, while 282 families contain a single *Chlamydomonas* protein and 2-5 *Volvox* proteins.

N. *Volvox*-specific genes

We were interested in identifying how many novel protein coding genes had appeared in the *Volvox* lineage since divergence from *Chlamydomonas*, since these proteins could encode *Volvox*-specific functions. From a starting set of 5,209 *Volvox* proteins that had not been placed into a protein family (see above), we identified 142 putative potentially *Volvox*-specific proteins based on the following three criteria: these proteins had no TBLASTN hit to the

Chlamydomonas genome assembly (E-value < 1E-10); at least one splice site supported by EST evidence and no BLASTP hit (E-value < 1E-10) to any protein from any of the proteomes we had used to make the protein families (Table S10 and see above).

We found 84 of the 142 proteins had BLASTP homology (E-value < 1E-10) to at least one other protein in the set, suggesting they are part of a protein family; the remaining 58 were singletons (Table S12). The quality of each of the 142 putative *Volvox*-specific gene models was inspected manually on the JGI genome browser at http://www.jgi.doe.gov/volvox. Many of these models were short and/or based solely on *ab initio* gene modelling and/or had no EST evidence or conflicted with EST evidence. Nonetheless, 25 gene models were completely consistent with EST evidence, and a further 11 gene models have partial EST support (Table S12). When we searched these 36 gene models against the protein sequences from the two *Micromonas* genomes (S*25*) using BLASTP (E-value < 1E-5) we found no detectable homology.

Intriguingly, none of the known *Volvox* developmental regulators was in this set of *Volvox*-specific proteins. Our analyses suggest that there are a small number of *Volvox*-specific proteins, despite substantial differences in developmental complexity between *Volvox* and *Chlamydomonas*.

O. *Chlamydomonas*-specific genes

In a parallel analysis to that performed for *Volvox*-specific genes, we identified 757 putative *Chlamydomonas*-specific genes from a starting set of 5,327 proteins that we were not able to place in a protein family. The larger number of *Chlamydomonas*-specific proteins compared to the number of *Volvox*-specific proteins may in part be due to deeper EST coverage in *Chlamydomonas*.

We found 238 of the 757 proteins had BLASTP homology (E-value < 1E-10) to at least one other protein in the set, suggesting they belong to a *Chlamydomonas*-specific protein family; the remaining 519 were singletons (Table S13). We chose a random sample of 50 putative *Chlamydomonas*-specific gene models from each of the above classes and examined the gene models manually at http://genome.jgi-psf.org/Chlre3/Chlre3.home.html and hence estimate that 32% and 60% of the models respectively are completely consistent with EST data (Table S13). We extrapolate this analysis to suggest that *Chlamydomonas* may have up to 400 novel proteins.

P. Volvocine algae-specific protein families

We investigated three classes of proteins that are only found in volvocine algae (defined as the group of organisms that includes *Volvox* and *Chlamydomonas*, as well as other species, such as *Gonium*, *Pandorina*, *Eudorina* and *Pleodorina* for which genome sequences are not yet available (Fig. S2) and see below). We discuss the results in this section and the next two sections, where presence or absence of a protein was based on the protein families described above. The first class of proteins is those found in both *Volvox* and *Chlamydomonas* but not other organisms. The second class consists of proteins that are only found in *Volvox*, and the third class consists of proteins that are only found in *Chlamydomonas*. These last two classes of proteins (together with various changes in regulation) might be associated with specific developmental and ecological adaptations in each species (see below).

We found 1,835 volvocine-specific protein families out of the total of 7,780 (Fig. 2B). To perform this analysis, we included data from the genomes of two *Micromonas* species that have been published recently (S*25*). These prasinophytes are substantially less reduced than the related *Ostreococcus* species that we had used in constructing protein families. We re-examined the 2,018 volvocine-specific families from our protein families in the light of this new data. We compared all *Volvox* and *Chlamydomonas* proteins in these families to all proteins in the predicted proteomes of *Micromonas pusilla* CCMP1545 v2.0 (http://genome.jgi-psf.org/MicpuC2/MicpuC2.home.html) and *Micromonas pusilla* sp. Rcc299 v3.0 (http://genome.jgi-psf.org/MicpuN3/MicpuN3.home.html) using WU-BLASTP (S*10*) (E-value < 1E-10). We removed 183 families containing a *Volvox* and/or *Chlamydomonas* protein that had a mutual best blast hit to a *Micromonas* protein. This left 1,835 volvocine-specific families. (Fig. 2B,D). Although these families have not be extensively characterized, they are expected to function in processes that are specific to volvocine algae and indeed, they include families of extracellular matrix proteins that participate in formation of the cell wall and ECM (Fig. 3A, S8).

Q. Analysis of Transcription Associated Proteins

Transcription associated proteins (TAPs) include transcription factors (TFs, proteins that bind to *cis*-regulatory elements enhancing or repressing gene transcription) and transcriptional regulators (TRs, proteins with indirect regulatory functions, such as the assembly of the RNA polymerase II complex, functioning as scaffold proteins in enhancer/repressor complexes or controlling chromatin structure by modifying histones or the DNA methylation).

To identify the TAPs in *Volvox* and *Chlamydomonas*, we combined three sets of TAP classification rules for plants, PlantTFDB (S*26*), PlnTFDB (S*27*) and PlanTAPDB (S*28*), and expanded them to yield a set of classification rules for 111 families. Conflicts between the initial three sources were manually evaluated and resolved based on an analysis of the scientific literature. The resulting set was then expanded by adding recently defined families or subfamilies from published sources. The rule set for each family consists of at least one entry defining a "should" rule, i.e. a mandatory domain for that particular family. Additional entries may define further "should" or "should not" (forbidden) domains. All domains relevant for classifying the TAPs were represented by a full length, global (termed “ls”) HMM. If available, the HMMs were retrieved directly from the ‘PFAM_ls’ database (S*29*). For the remaining domains, HMMs were custom-made using multiple sequence alignments (MSAs) to identify the conserved domain(s) of interest. The MSAs used for creating the custom HMMs were downloaded from PlnTFDB (S*30*). For domains not represented in this database, MSAs were created as follows. BLAST searches with a protein query containing the respective domain yielded homologous hits defined by having at least 30% sequence identity with the query over a minimum length of 80 amino acids. Those hits were aligned using MAFFT (S*31*) and manually curated using Jalview (S*32*). The conserved domain of interest was extracted and the HMM calculated with HMMER 2.0 (http://hmmer.janelia.org/) using ‘hmmbuild’ with the default parameters to generate ls HMMs and subsequently ‘hmmcalibrate’ with the option ‘--seed 0’ which sets the random starting seed to a constant value and hence obtains reproducible results during the calibration process.

Gathering cutoff (GA) values were defined for each custom HMM. The GA was set as the lowest score of a domain-containing protein (true positive) after a ‘hmmpfam’ search (using an E-value cutoff of 1E-5) against the full proteome sets of several different species and considering the alignments of all hits. In order to avoid sampling bias, only fully sequenced genomes were used in this study. For each organism, the complete set of proteins derived by conceptual translation of the nuclear gene models (using the filtered/selected model per locus) was combined with the proteins encoded by the respective mitochondrial and plastid genome, if available. All proteins can be unambiguously identified via their fasta id. We used a unique five letter code for each organism followed by “mt” (mitochondrial) or “pt” (plastid), if applicable, and the accession number of the gene model.

Using all proteins of the investigated organisms as query, ‘hmmpfam’ searches were performed against an HMM library containing all 129 domains necessary for the TAP classification. The GA was used during this procedure to minimize the number of false positive hits, with GA values either provided with the Pfam HMMs or defined as described above. The classification rules were subsequently applied to all proteins for which at least one significant domain hit was found. In cases where the domain composition of a protein matched more than one classification rule, the ‘should’ rule with the highest score determined the family into which the protein was categorized.

Highly similar domains which are often found in the same or overlapping regions of a protein were treated in similar fashion, i.e., the domain with the lowest E-value/highest score was used for the subsequent classification. This procedure was necessary for four sets of domains, namely i) Myb_DNA-binding and G2-like_Domain, ii) NF-YB, NF-YC and CCAAT-Dr1_Domain, iii) PHD and Alfin-like and iv) GATA and zf-Dof. In addition, a Boolean OR rule was applied to three families. In these cases one out of two domains was found to be necessary and sufficient for a protein to be classified into the corresponding family. This rule was applied to the bZIP, HD-Zip and GARP_ARR-B families. Whenever the presence of a combination of domains led to more than one possible family classification, TF was favored over TR or PT (putative TAPs). This situation was encountered in 14 cases.

In *Volvox*, the proportion of all proteins that are transcription factors is 347/14,520; in *Chlamydomonas* it is 297/14,516 (Table S15). This proportion is not significantly higher (p=0.02831, one-tailed Fisher Exact test) in *Volvox* compared to *Chlamydomonas*. A scatter plot of the number of *Volvox vs. Chlamydomonas* proteins in each TAP family (Fig. S7) shows that most families lie on or near the diagonal, with the larger families showing slight over-representation of *Volvox* proteins.

R. Annotation of genes associated with developmental biological processes in *Volvox*

Membrane trafficking proteins

We started with a set of SNARE and Rab GTPase proteins from *Chlamydomonas* (S*14, 33*) and searched for appropriate gene models in homologous regions in the *Volvox* genome using TBLASTN (E-value < 10). Reciprocal searches were conducted to identify the mutual best hit pairs between the two species. The NCBI nr protein database (S*8*) was also queried with each protein from *Volvox* to identify the best hit in another species such as human, Arabidopsis and yeast (S*34*) which were then used as the query protein in searches against the *Volvox* and *Chlamydomonas* genomes. Finally, to assign a family name to each protein, we performed phylogenetic analysis for Rab proteins (aligning proteins and building 1,000 bootstrap neighbor-joining trees using CLUSTAL X 1.82 (S*35*)) and Syp proteins (aligning proteins with CLUSTAL X 1.82 (S*35*) and MUSCLE (S*36*), and building 100 boostrap maximum parsimony trees with PAUP* 4.0 beta 10 (S*37*)) using *Volvox* and *Chlamydomonas* proteins and *Arabidopsis* and human homologs found by BLAST searches at the nr database at GenBank (S*8*).

Cell cycle proteins

We started with a set of cell cycle proteins from *Chlamydomonas* (S*38*) and searched for homologs in *Volvox* using BLASTP and TBLASTN (E-value < 10). Reciprocal searches were conducted to identify mutual best hit pairs between the two species. The NCBI nr protein database (S*8*) was also queried with each predicted cell cycle protein from *Volvox* and *Chlamydomonas* to identify the best hit in another species, which was then used as the query protein in searches against the *Volvox* and *Chlamydomonas* genomes and predicted proteomes. This process was iterated until all significant BLAST hits between cell cycle proteins and gene models in *Chlamydomonas* and *Volvox* had been identified. For each cell cycle gene model identified in *Volvox*, flanking genes were used to identify synteny with the putative orthologous model in *Chlamydomonas*. In all cases the synteny was in agreement with orthology assignments based on mutual best hits. In addition, an identical approach was used to identify *Volvox* and *Chlamydomonas* orthologs, this time starting with all *Volvox* proteins with PFAM or KOG domain assignments specific to cell cycle regulation.

Cytoskeletal proteins

We searched the GenBank nr database (S*8*) for members of known cytoskeletal protein families (S*39*), and used sequences identified from Arabidopsis (or *Drosophila* when Arabidopsis hits were not found) as queries in TBLASTN searches (E-value < 0.01) against the *Volvox* and *Chlamydomonas* genome assemblies at the *Volvox* or *Chlamydomonas* genome portal at the JGI. We assumed proteins without a hit were not encoded in the algal genome. The best gene model from the hit results was chosen, based on E-value, EST evidence and homology to the other algal genome and generally gave best hit E-values < 1E-7 when queried back against GenBank by BLASTP. Each protein model obtained in this way was next used as query in a second TBLASTN (E-value < 1E-5) against both algal genomes to identify additional homologs. This process was repeated until all members of the family were identified. Orthology between *Volvox* and *Chlamydomonas* proteins was inferred when the candidates were mutual best hits in TBLASTN searches and the Vista track at the JGI browser showed significant conservation at the DNA level.

Cell wall and extracellular matrix proteins

We started with a set of known extracellular proteins/ECM proteins /cell wall proteins from *Volvox* (S*40, 41*).

We made TBLASTN searches (E-value < 1E-7) with the protein sequences against both, the *Volvox* and *Chlamydomonas* genomes. All hits were searched reciprocally against the other algal genome, also using TBLASTN.

Whenever TBLASTN hits corresponded to an existing gene model, the model was used, or the model was edited or a new model was generated using the JGI portal.

Phylogenetic analyses

The following describes the phylogenetic analyses used to generate the trees in Fig. 3. Homologous protein sequences were aligned with MUSCLE (S*36*). Poorly-aligning end regions were trimmed and the sequences were realigned. The process was repeated until no further improvements could be made. Positions with gaps were removed prior to construction of phylogenies. ProtTest (S*42*) was used to select the best model of protein evolution for each set of proteins. Maximum likelihood trees were constructed using PhyML 3.0 (S*43, 44*) under the following parameters: 100 bootstrap replicates; four-category gamma distribution; proportion of variable sites estimated from the data.

2) SUPPORTING TEXT

A. Volvocine algae as a model for the evolution of multicellularity

In addition to its strengths as a developmental-genetic model, *Volvox*, together with its relatives in the "volvocine lineage" (Fig. S2), provides an unrivalled opportunity to explore the details of a pathway by which multicellular organisms with differentiated cell types evolved from a unicellular ancestor – one of the most complex and interesting steps in the evolution of higher organisms (S*45*). Formation of a multicellular body of predictable shape and size has usually required the invention of novel morphogenetic mechanisms, while differentiation of two or more distinct cell types within such a body has required elaboration of novel spatial patterns of gene expression. This is likely true in *Volvox* too. Multicellularity has evolved not just once but repeatedly and independently in a highly diverse array of taxa (S*46-48*). However, in most cases the transition to multicellularity has occurred so long ago (more than 500 MYA in many cases (S*47, 49, 50*) that most details of the molecular genetic changes leading to multicellularity have diverged so much they can no longer be studied.

It has long been suggested, however, that the volvocine algae provide an interesting exception to the preceding generalization. The volvocine lineage comprises several genera of green flagellates that can be arranged in a conceptual series according to increasing complexity (Fig. S2) –*Chlamydomonas*, *Gonium*, *Pandorina*, *Eudorina*, *Pleodorina*, *Volvox* – within which there are progressive increases in cell number, size of adult organisms, volume of ECM per cell, and the tendency to produce sterile, terminally differentiated somatic cells. Recent molecular-phylogenetic analyses not only indicate that these algae constitute a coherent, monophyletic group that began its radiation within the last ~200 MYA (S*51*) (S*52*), but also that the sequence indicated above serves as a reasonable first approximation of the historical sequence in which members of the group evolved (S*53*). Furthermore the allure of volvocine algae as an evolutionary model system is significantly enhanced by the finding that the kind of germ-soma division of labor that has traditionally earned an alga membership in the genus *Volvox* has arisen independently on at least four separate branches of the volvocine family tree (S*54*)

Molecular-genetic studies of *Volvox* embryogenesis have already indicated that different aspects of the evolution of *Volvox* from a *Chlamydomonas*-like ancestor have involved qualitatively different amounts of genetic change. For example, the *glsA* gene (whose product is required for the asymmetric divisions that set apart the germ and somatic cell lineages of *Volvox* embryos) (S*55*) obviously was adopted for this novel function with no significant changes, because the orthologous *GAR1* gene of *Chlamydomonas* is fully capable of substituting for it (S*56*), even though there is no known asymmetric division in the *Chlamydomonas* life cycle. Similarly, the *invA* gene (whose product is a kinesin that the *Volvox* embryo requires for inversion at the end of embryogenesis) can be replaced by its *Chlamydomonas* ortholog, *IAR1* (S*57*), indicating that this gene was also adopted to play an entirely novel morphogenetic role without any significant evolutionary modification of the protein that it encodes. In marked contrast, *Chlamydomonas* lacks any recognizable ortholog of the *regA* gene of *Volvox* that plays a central role in differentiation and programmed death of somatic cells (apparently by repressing chloroplast biogenesis; (S*58*) (S*59*)): *regA* encodes an entirely novel combination of pre-existing and new protein domains, of which only the sequence of the presumed DNA-binding domain can be traced back to its Chlamydomonad ancestry (S*60*).

The attractiveness of *Volvox* as a developmental and evolutionary model is enhanced by the availability of several important molecular tools, including a variety of selectable markers (S*61-64*), a transposon-tagging system (S*65, 66*), a nuclear transformation system (S*67*), and a reporter gene (S*68*).

B. The Volvox vegetative life cycle

Volvox has two cell types: ~2,000 small, biflagellate *Chlamydomonas*-like somatic cells that are embedded in the surface of a transparent sphere of glycoprotein-rich extracellular matrix (ECM), and ~16 large reproductive cells (termed gonidia) that lie just below the somatic cell monolayer (S*69*).Each gonidium grows, divides and undergoes morphogenesis to produce the next generation (Fig. S1B). Asymmetric cell divisions during embryogenesis determine the germ-line precursors. Following cleavage, the embryo turns inside-out in a process called inversion; inverted juveniles expand by the deposition of ECM, and finally hatch out of the mother colony to complete the life cycle.

C. The Volvox sexual cycle

Sexual development in *Volvox* and *Chlamydomonas* is controlled by a large, multigenic, haploid mating locus (*MT*) that segregates as a single Mendelian trait. *MT* occupies the same chromosome in both species, but is five times larger in *Volvox* relative to *Chlamydomonas* (S*70*). Both sexes of *Volvox* have the same vegetative developmental cycle that is described in the preceding section. However, in response to a diffusible sex inducer protein *Volvox* males and females undergo modified developmental programs to produce sperm packets and eggs, respectively (S*71*). This developmental response to sex inducer involves changes in the timing of asymmetric cell division, altered gametic gene expression (S*72, 73*), and male germ cell divisions into sperm packets.

D. The Volvox ECM

The *Volvox* ECM comprises up to 99% of the spheroid volume (reached in the adult shortly before release of daughter spheroids) and provides a highly organized substrate that compartmentalizes its interior space (Fig. S3). It is likely to be involved in intercellular signaling and nutrient transport (S*40*).

Evolutionarily, the *Volvox* ECM can be understood as a massive elaboration of the cell wall of *Chlamydomonas*. In both organisms the cell wall and ECM are composed of hydroxyproline-rich glycoproteins (HRGPs) that form rod-like structures and which often have additional globular domains at each end (S*74*). In *Volvox*, individual pherophorin subtypes are associated with distinct regions of the ECM, and each subtype is likely to be involved in the assembly and/or specific function of these ECM subdomains (also referred to as ECM subzones) (S*74*).

E. Environmental adaptations of *Volvox* and *Chlamydomonas*

Volvox and *Chlamydomonas* are cosmopolitan species and occupy overlapping habitats (S*75*). *Chlamydomonas* can proliferate in more transient bodies of water than *Volvox*, thanks to its faster generation time and smaller size. While *Chlamydomonas* is often portrayed as a soil alga, it is usually collected from soil in the form of environmentally resistant and dormant zygospores which can travel long distances and last for years under unfavorable conditions (S*75, 76*). Thus, the places where *Chlamydomonas* is collected provide only a partial indication of the environment to which it was adapted and in which it proliferates. On the other hand, *Volvox* and other multicellular volvocine algae are generally collected as live specimens from permanent or semi-permanent bodies of water; but such collection sites are biased against transient and unreliable locations. Thus, the specific environmental adaptations that may have arisen in the two species have not been systematically examined.

3) SUPPLEMENTAL FIGURES

Fig. S1: *Volvox* phylogeny, morphology and development.

(A) The phylogenetic position of *Volvox* and *Chlamydomonas* (within Chlorophyceae, green) is shown in an unrooted schematic cladogram of the eukaryotic tree of life (Sfrom *47*); open and filled triangles denote clades consisting of solely unicellular lineages, and clades comprising both unicellular and multicellular lineages, respectively. (B) The asexual life cycle of *Volvox* with photomicrographs, taken as described in (S*66*), of a newly-hatched adult (top left, bar = 200 µm) and of an adult *Volvox* (lower right, bar = 500µm) as well as scanning electron micrographs of an embryo after the first asymmetric cell division at cleavage cycle 6 (top right inset) and of a post-cleavage embryo during inversion (middle right inset). Each micrograph is placed near its corresponding developmental stage in the schematic diagram.

Fig. S1

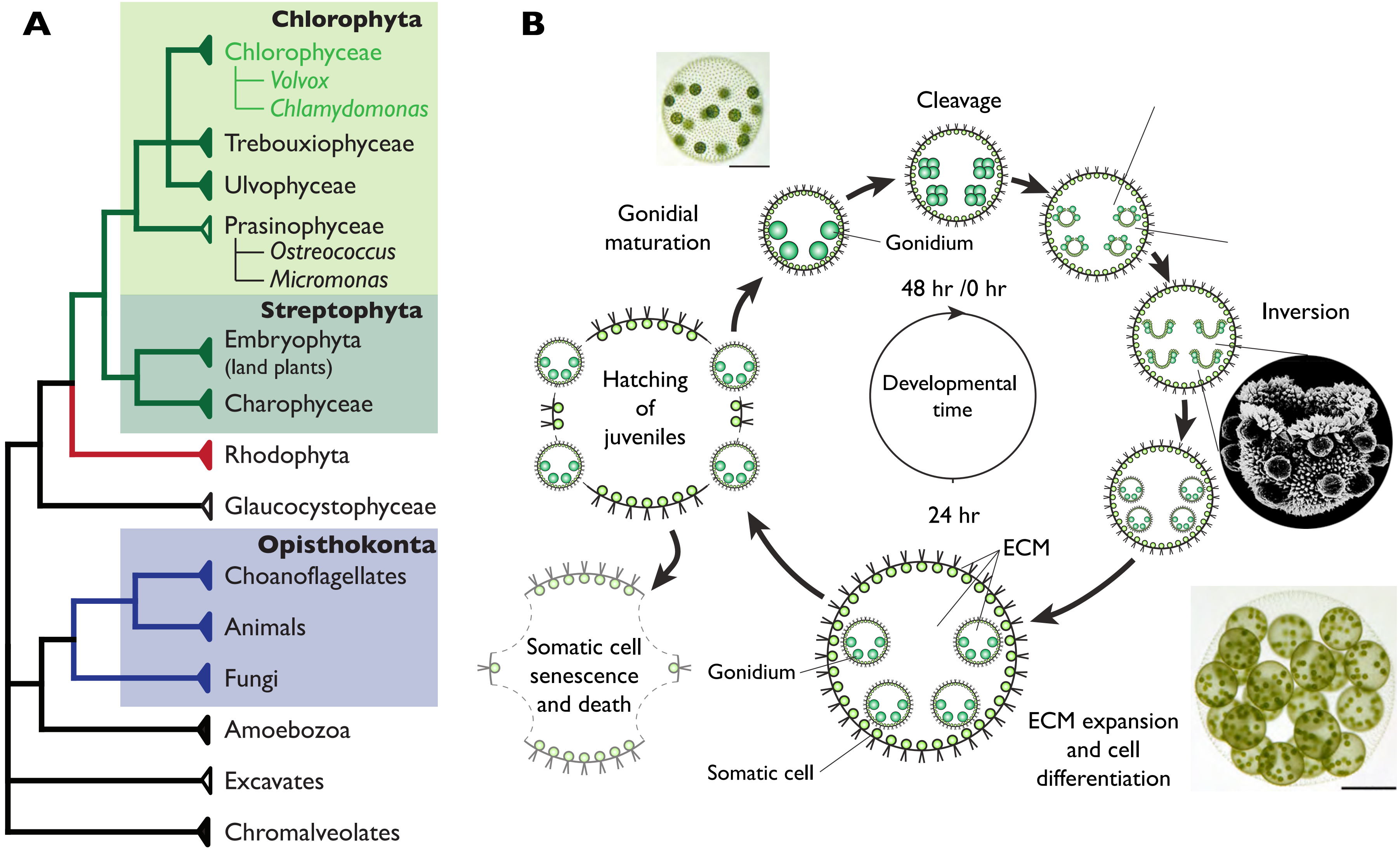

A
Chlorophyta
Chlorophyceae
Volvox
Chlamydomonas
Trebouxiophyceae
Ulvophyceae
Prasinophyceae
Ostreococcus
Micromonas
Streptophyta
Embryophyta
(land plants)
Charophyceae
Rhodophyta
Glaucocystophyceae
Opisthokonta
Choanoflagellates
Animals
Fungi
Amoebozoa
Excavates
Chromalveolates
B
Gonidial maturation
Gonidium
Cleavage
48 hr /0 hr
Inversion
Developmental time
24 hr
Hatching of juveniles
ECM
Somatic cell senescence and death
Gonidium
Somatic cell
ECM expansion and cell differentiation

Fig. S2: The algae of the volvocine lineage

The volvocine algae comprise dozens of species that range in complexity from *Chlamydomonas* through colonial forms that have evolved different types of developmental traits. Photomicrographs of representative species from key genera are arranged along the top of the chart. The presence of the developmental traits listed along the left side is indicated in the grid by a ‘X’. The photomicrographs were taken as described in (S*66*). Strains used are as follows: *Chlamydomonas reinhardtii* strain c-239+ (S*77*); *Gonium pectorale* strain kaneko3 (S*78*); *Pandorina morum* strain NIES-877; *Eudorina elegans* strain NIES-721; *Pleodorina starii* (female) (S*79*)

Fig. S2

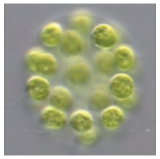
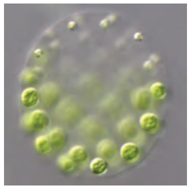
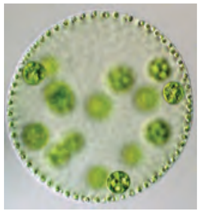
Developmental trait(s)	 *Chlamydomonas reinhardtii*	 *Gonium pectorale*	 *Pandorina morum*	 *Eudorina elegans*	 *Pleodorina starrii*	 *Volvox carteri*
Unicellular	X					
Cell sheets Partial inversion		X				
Spherical colonies Full inversion Incomplete cytokinesis			X	X	X	X
Expansion of ECM				X	X	X
Anisogamy				X	X	X
Partial division of labor					X	
Complete division of labor Asymmetric cell division Bifurcated cell division program						X

Fig. S3: Schematic diagram of ECM in *Volvox*

In this schematic cross-section of a *Volvox* adult (redrawn from (S*71, 75, 80*)), the elaboration of the ECM into deep, cellular and boundary and flagellar zones is shown, with the three subzones of the cellular and boundary zones surrounding a single zoomed in somatic cell. Fibrous cellular zone 1 is attached to the somatic cell body plasmalemma, cellular zone 2 is relatively amorphous; fibrous cellular zone 3 forms compartments around the somatic cells. The boundary zone is continuous except where interrupted by flagella, the dense fibrous boundary zones 1 and 3 flank the tripartite boundary zone 2. Deep zone 1 is an band of filaments and surrounds the amorphous deep zone 2 (S*75*).

Fig. S3

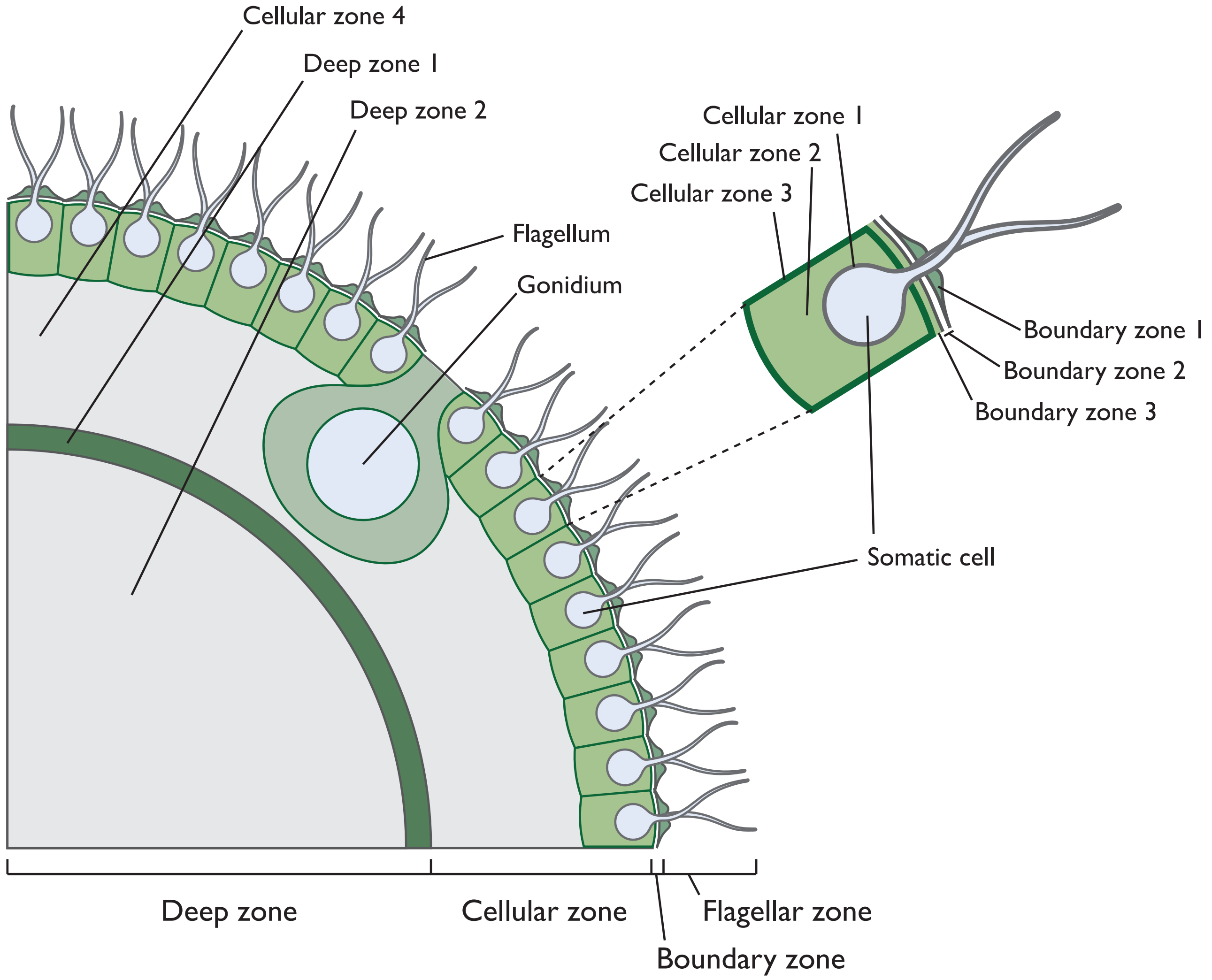

Cellular zone 4
Deep zone 1
Deep zone 2
Flagellum
Gonidium
Cellular zone 1
Cellular zone 2
Cellular zone 3
Boundary zone 1
Boundary zone 2
Boundary zone 3
Somatic cell
Deep zone
Cellular zone
Flagellar zone
Boundary zone

Fig. S4: Histograms of Jukes-Cantor distance between repeats

Histograms plot the distribution of sequence divergence (as measured by Jukes-Cantor distance) between repeats within *Volvox* (top row, A-C) and *Chlamydomonas* (bottom row, D-F). They show the distances between all repeats identified by Repeat Scout (A,D), Copia elements only (B,E) and Gypsy elements only (C,F).

Fig. S4

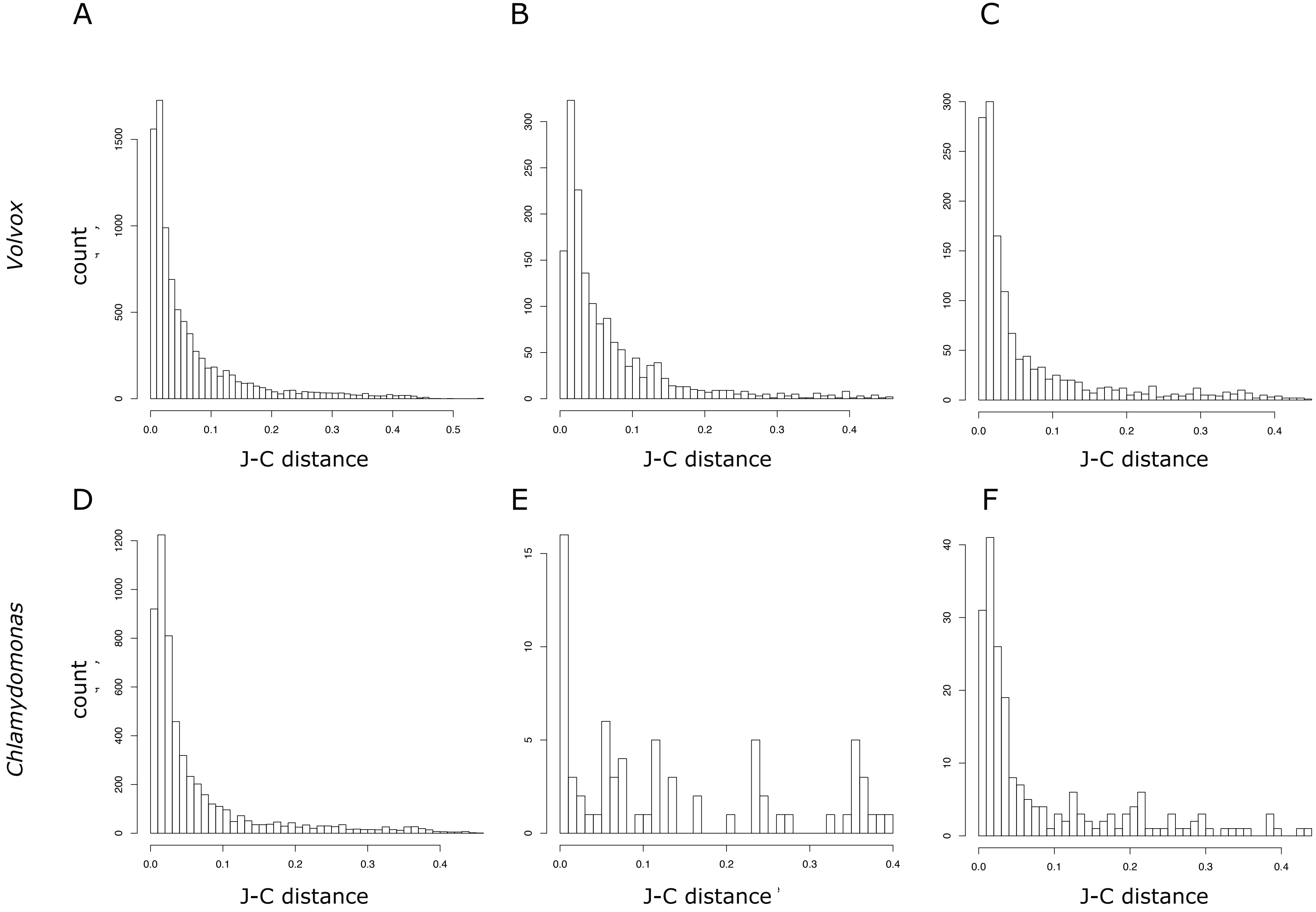

A
B
C
Volvox
count
1500
1000
500
0
0.0
0.1
0.2
0.3
0.4
0.5
J-C distance
300
250
200
150
100
50
0
0.0
0.1
0.2
0.3
0.4
J-C distance
300
250
200
150
100
50
0
0.0
0.1
0.2
0.3
0.4
J-C distance
D
E
F
Chlamydomonas
count
1200
1000
800
600
400
200
0
0.0
0.1
0.2
0.3
0.4
J-C distance
15
10
5
0
0.0
0.1
0.2
0.3
0.4
J-C distance
40
30
20
10
0
0.0
0.1
0.2
0.3
0.4
J-C distance

Fig. S5: Synteny dotplot between *Volvox* and *Chlamydomonas* genomes

Conserved gene order plots for (A) *Volvox-Chlamydomonas* and (B) human-chicken, showing locations of syntenic orthologs (max 2 intervening genes, segment size 2 or more genes). Syntenic genes lie along the two axes. These are arbitrarily numbered as follows: syntenically orthologs are numbered along scaffolds (*Volvox*) or chromosomes (*Chlamydomonas* v4 assembly, human and chicken) from largest to smallest, arbitrarily starting at 1 for the x-axes and 100,000 for the y-axes. unmap, chicken scaffolds that have not been mapped to a chromosome.

Fig. S5

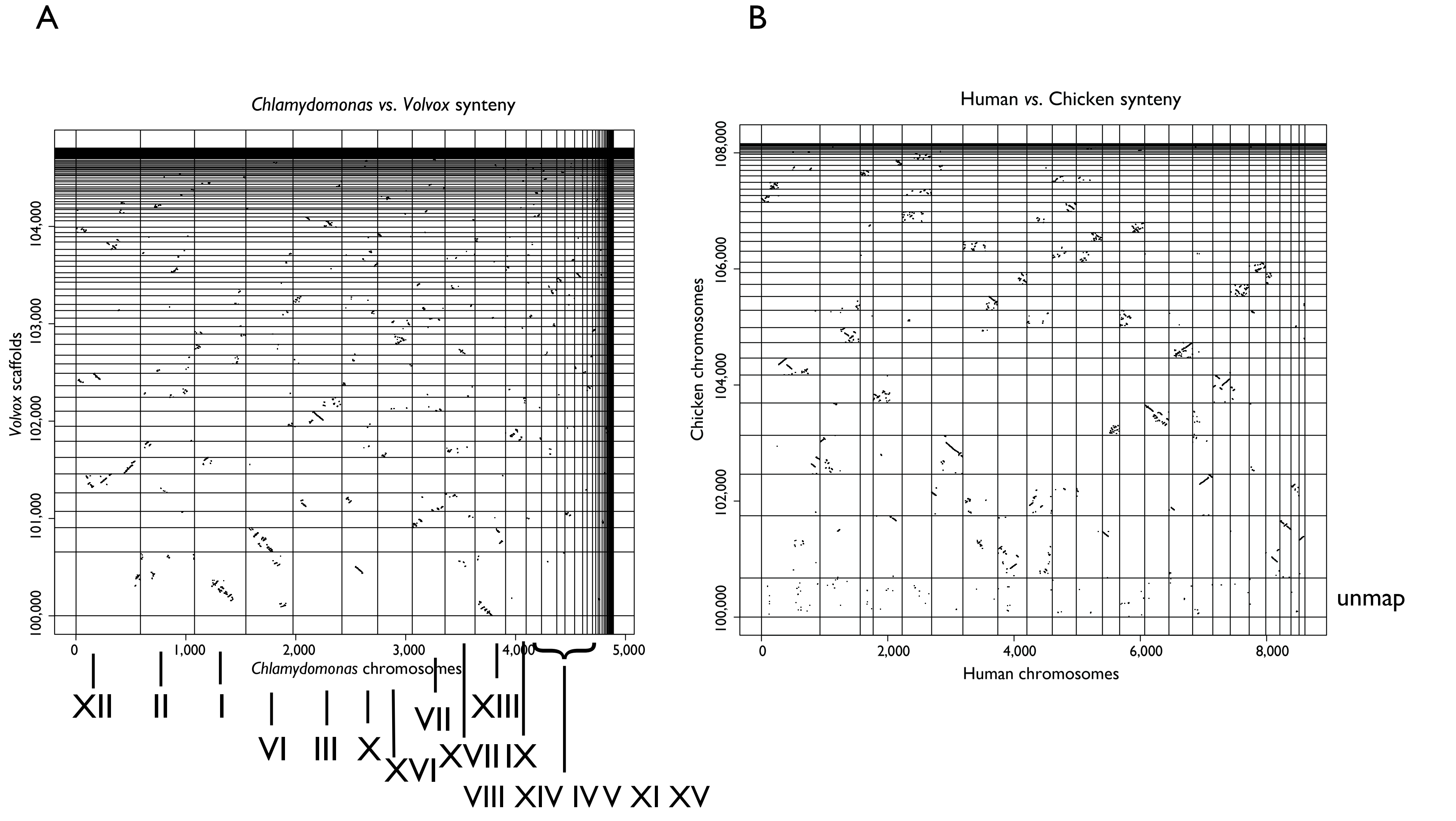

A
Chlamydomonas vs. Volvox synteny
Volvox scaffolds
100,000
101,000
102,000
103,000
104,000
0
1,000
2,000
3,000
4,000
5,000
Chlamydomonas chromosomes
XII
II
I
VI
III
X
XVI
VII
XVII
XIII
IX
VIII XIV IV V XI XV
B
Human vs. Chicken synteny
Chicken chromosomes
100,000
102,000
104,000
106,000
108,000
0
2,000
4,000
6,000
8,000
Human chromosomes
unmap

Fig. S6: Intron lengths in *Volvox* and *Chlamydomonas*

The length of introns at conserved positions between orthologous exons in *Volvox* vs. *Chlamydomonas* is shown in a scatter plot (see above). Three subpopulations are evident (boxes). The majority (93%) orthologous introns are >100 nt long, longer in *Volvox* and show no size correlation between the two species. A small (3%) subset of introns are short (≤100 nt) in both species and when plotted, lie near the diagonal meaning that they have similar sizes in the two species. Finally, a third small subset of *Volvox* introns (4%) are around 60-100 nt long but vary over a wide size range in *Chlamydomonas* and appear as a faint horizontal smear across the bottom of the plot.

Fig. S6

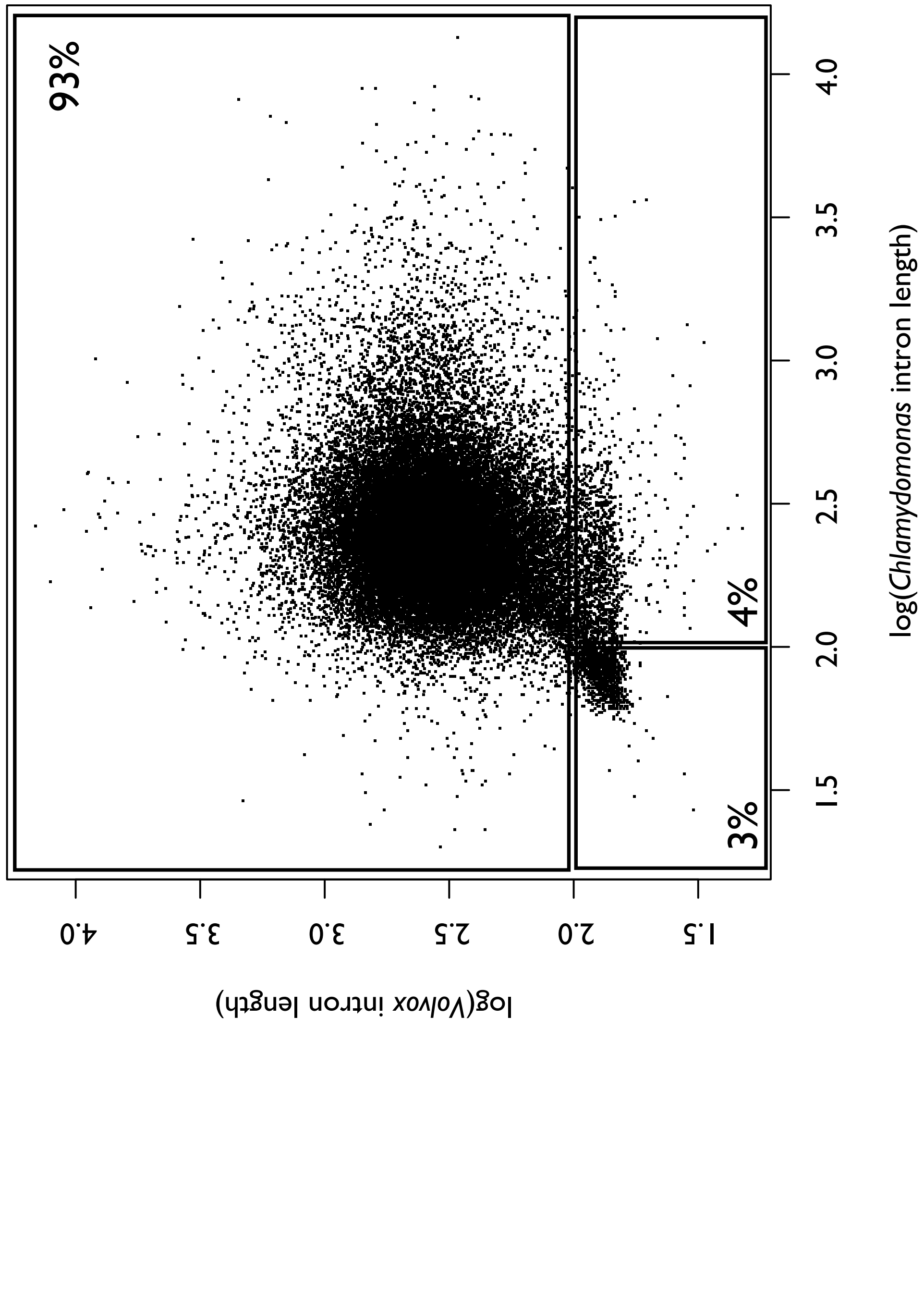
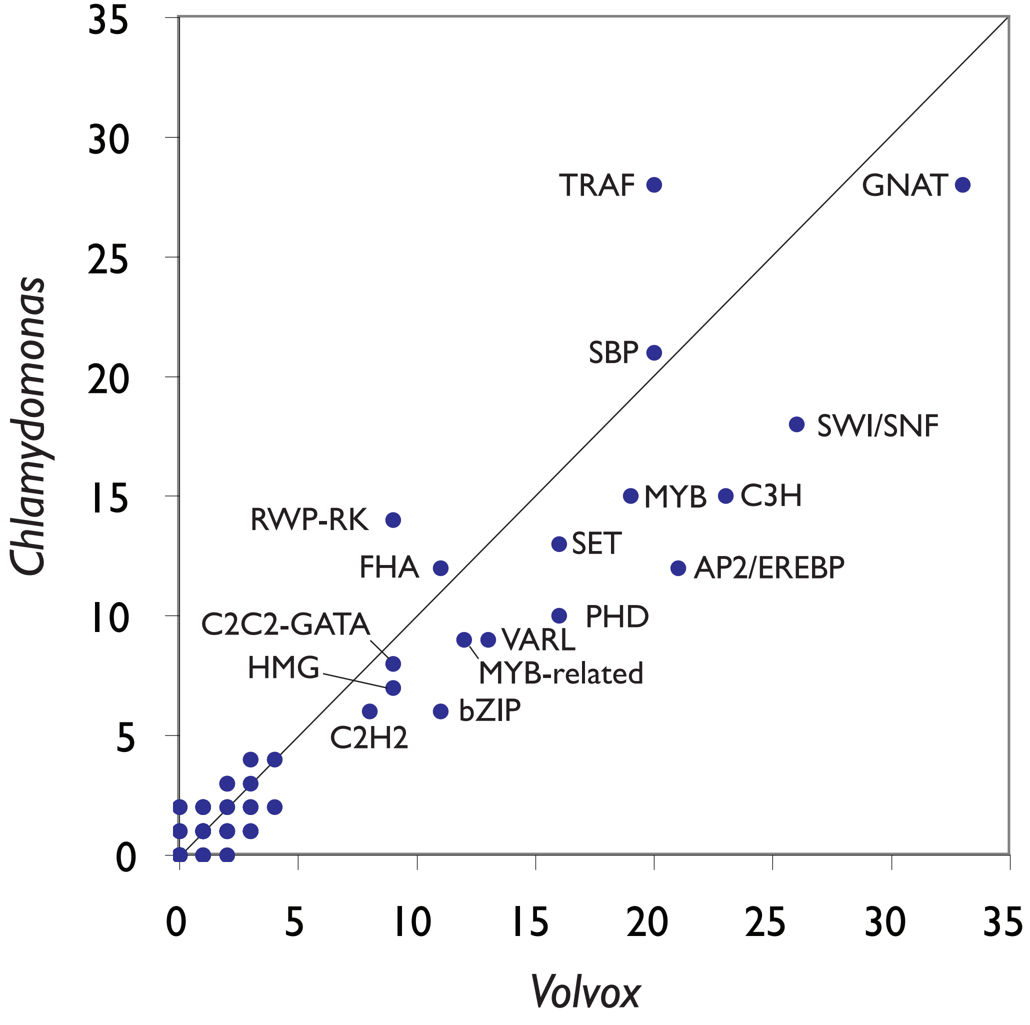

93%
4%
3%
1.5
2.0
2.5
3.0
3.5
4.0
log(*Chlamydomonas* intron length)
1.5
2.0
2.5
3.0
3.5
4.0
log(*Volvox* intron length)

Fig. S7: Scatter plot of family size in transcription associated proteins

The number of *Volvox* proteins in a transcription-associated protein family is plotted against the number of *Chlamydomonas* proteins in the same family. The diagonal line marks the positions of families with equal numbers of proteins from each species. The names of families with more than five members from each species are indicated.

Fig. S7

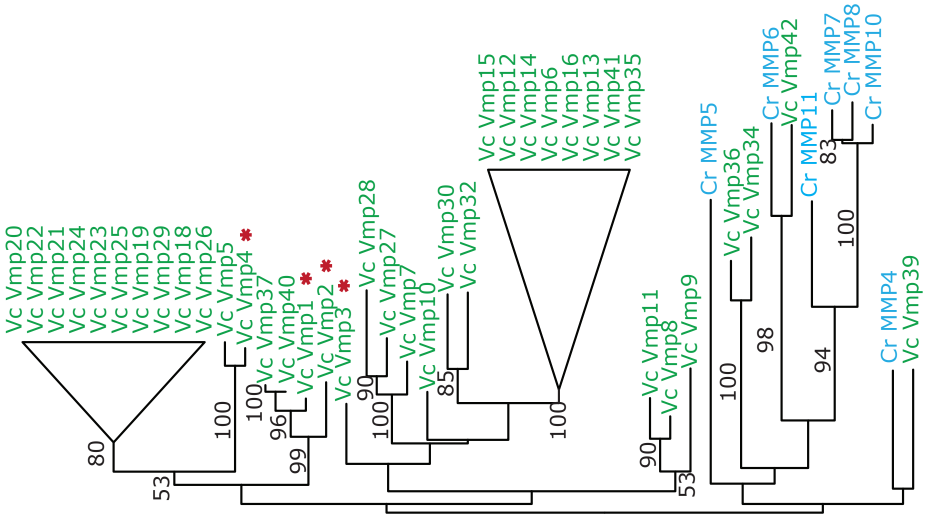


Chlamydomonas
35
30
25
20
15
10
5
0
Volvox
0
5
10
15
20
25
30
35
TRAF
GNAT
SBP
SWI/SNF
MYB
C3H
RWP-RK
SET
FHA
AP2/EREBP
PHD
C2C2-GATA
VARL
HMG
MYB-related
bZIP
C2H2

Fig. S8: Diversification of *Volvox* matrix metalloprotease family.

Unrooted maximum likelihood tree of *Volvox* matrix metalloproteases. Protein sequences are from *Volvox* (Vc; green) and *Chlamydomonas* (Cr; blue). Incomplete gene models were not included ; *Volvox*-specific clades with poorly-resolved branches are collapsed into triangles; bootstrap support ≥ 50% is indicated on branches. Red asterisks indicate proteins whose mRNA levels are up-regulated by sex inducer.

Fig. S8


Vc Vmp20
Vc Vmp22
Vc Vmp21
Vc Vmp24
Vc Vmp23
Vc Vmp25
Vc Vmp19
Vc Vmp29
Vc Vmp18
Vc Vmp26
Vc Vmp5
Vc Vmp4
Vc Vmp37
Vc Vmp40
Vc Vmp1
Vc Vmp2
Vc Vmp3
Vc Vmp28
Vc Vmp27
Vc Vmp7
Vc Vmp10
Vc Vmp30
Vc Vmp32
Vc Vmp15
Vc Vmp12
Vc Vmp14
Vc Vmp6
Vc Vmp16
Vc Vmp13
Vc Vmp41
Vc Vmp35
Vc Vmp11
Vc Vmp8
Vc Vmp9
Cr MMP5
Vc Vmp36
Vc Vmp34
Cr MMP6
Vc Vmp42
Cr MMP11
Cr MMP7
Cr MMP8
Cr MMP10
Cr MMP4
Vc Vmp39
80
53
100
100
96
99
90
100
85
100
90
53
100
98
94
100
83

4) SUPPLEMENTAL TABLES

Table S1: Summary of repeats in *Volvox* and *Chlamydomonas* genomes

The extent (and percentage in parentheses) of the *Volvox* and *Chlamydomonas* genomes that are masked by different classes of repeat family/subfamily and simple repeats are shown. Repeat masking was performed with RepeatMasker and the custom library that we had built for the genome.

Repeat family/subfamily	***Volvox* assembly**	***Chlamydomonas assembly***
SINEs	298,781 (0.22%)	125,738 (0.11%)
LINEs	2,681,727 (1.95%)	4,544,976 (3.84%)
LTR elements	5,067,964 (3.68%)	890,315 (0.75%)
LTR/Copia	218,136	89,130
LTR/Gypsy	675,620	403,108
DNA elements	1,861,025 (1.35%)	2,003,374 (1.69%)
Jordan	152,065	0
Unclassified	18,267,323 (13.25%)	7,206,766 (6.10%)
Total Interspersed Repeats	**28,176,820 (20.44%)**	**14,771,169 (12.50%)**
Satellites	145,736 (0.11%)	489,348 (0.41%)
Simple Repeats	4,561,091 (3.31%)	6,184,379 (5.23%)
Low Complexity	1,246,389 (0.90%)	1,799,865 (1.52%)
Total non-Interspersed Repeats	**5,953,216 (4.32%)**	**8,473,592 (7.17%)**
Total Repeats	**34,130,036 (24.76%)**	**23,244,761 (19.66%)**

Table S2: Genome evolution in green algae, animals, plants and diatoms

We compare neutral nucleotide substitutions (4DTV), species divergence time, genome rearrangements and protein evolution (mutual best hits) for selected species pairs. N.D. not determined because at least three whole genome duplications since speciation prevented clear assignment of orthologs.

Species pair	Corrected 4DTV	Time since divergence (Myr)	Neighbor rearrangements (%)	Median distance between rearrangements in genome 1 / genome 2 (kb)	Similarity between mutual best BLAST hits (%)
Volvox/*Chlamydomonas* (green algae)	0.71	~ 220 (S*51*)	34	6 / 6	73.8
human/chicken (vertebrates)	0.57	~310 (S*81*)	15	113 / 40	75.9
human/frog (vertebrates)	0.80	~350 (S*82*)	14	83 / 49	71.8
Arabidopsis/*Populus* (angiosperms)	0.68	~110 (S*83*)	N.D.	N.D.	72.0
Thalassiosira/*Phaeodactylum* (diatoms)	1.94	~90 (S*84*)	45	2 / 2	54.9

Table S3: Counts of filtered genes that were used to build syntenic blocks

The numbers of genes in the table correspond to syntenic orthologs after tandem duplicates and high-copy gene were removed.

Species	Filtered genes
Human	8,612
Chicken	8,159
Frog	8,158
Chlamydomonas	4,890
Volvox	4,804

Table S4: Counts of *Volvox* and *Chlamydomonas* genes making up syntenic blocks with selected numbers of intervening genes allowed for real and scrambled gene order

This table shows the number of syntenic orthologs that are part of syntenic blocks that were generated when a range of zero to four intervening genes were allowed between syntenic orthologs for both the real gene order, and randomized gene order.

Maximum no. intervening genes	Real gene order	Scrambled gene order
0	2,839	8
1	3,363	24
2	3,589	40
3	3,712	58
4	3,775	84

Table S5: Counts of gene models predicted in *Volvox* by initial automated annotation, classified by method

The number of gene models that were generated with the automated JGI gene annotation pipeline are shown partitioned into the different methods that generated them. The gene models shown here are the raw output before genes with homology to Transposable Elements were filtered.

Method used to generate gene model	Number of gene models
Based on homology to proteins in nr database at GenBank	3,645 (23%)
Ab initio gene prediction	10,217 (67%)
Based on EST cluster consensuses	143 (1%)
Based on synteny with *C. reinhardtii*	1,539 (10%)
Total initial models	15,544 (100%)

Table S6. EST and homology evidence supporting initial *Volvox* and *Chlamydomonas* gene models

The numbers of gene models in the initial predictions that are complete from the start to the stop, have EST support or homology to a protein in Swissprot are shown. The models included in this table are those that were the output of the automated JGI annotation pipeline for Volvox and the frozen GeneCatalog (S*14*) that was submitted to GenBank (S*8*) (Accession ABCN00000000).

Evidence	*Volvox*	*Chlamydomonas*
Complete models	13,134 (85%)	8,919 (58%)
Models with EST alignment	5,356 (34%)	7,894 (51%)
Models with Swissprot alignment	10,947 (70%)	10,760 (71%)

Table S7: Gene structure statistics of *Volvox* and *Chlamydomonas* gene models

A variety of statistics are shown for the set of *Volvox* gene models after removing those with homology to Transposable Elements and the manually-curated set of *Chlamydomonas* gene models that were submitted to GenBank (S*8*) (Accession ABCN01000000).

	Volvox	*Chlamydomonas*
Protein-coding loci	14,520	14,516
Mean gene span (nt)	5,269	4,375
Total length of spliced transcripts (nt)	27,126,224	23,675,605
Mean transcript length (nt)	1,833	1,631
Mean protein length (aa)	568	454
Mean exon length (nt)	194	232
Mean intron length (nt)[1]	491	371
Mean no. exons	7.78	8.42

[1] Introns less than 20 nt long are ignored

* This is the set of gene models that was submitted to GenBank under the Accession ACJH00000000

Table S8: Comparison of *Volvox* genome statistics to selected other genomes.

Group	Species	Genome Size (Mb)	Number of chromo-somes	%GC	Protein coding loci	% coding	% genes with introns	Introns per gene	Median intron length (bp)
CHLOROPHYTA	*Volvox carteri*	138	14*	56	14,520	18.0	92	7.05	358
	Chlamydomonas reinhardtii	118	17	64	14,516	16.3	91	7.4	174
STREPTOPHYTA	*Physcomitrella patens*	480	27	34	35,938	17.9	86	3.9	205
	Arabidopsis thaliana	140.1	5	36	26,541	23.7	80	4.4	55
OPISTHOKONTA	*Homo sapiens*	2851	23	41	23,328	1.2	83	7.8	20,383
	Nematostella vectensis	450	15	40	27,273	6.0	68	4.3	290
	Monosiga brevicollis	42	N.A.	55	9,196	39.4	89	6.6	135
	Neurospora crassa	40	7	54	10,107	36.4	80	1.7	72
AMOEBOZOA	*Dictyostelium discoideum*	34	6	22	13,574	62.2	68	1.3	236
CHROMALVEOLATA	*Thalassiosira pseudonana*	34.5	24	47	11,390	49.4	60	1.5	57

* see (S75)

N.A. not available

Table S9: Pfam domain counts and combinations in *Volvox* and *Chlamydomonas* compared to selected other species

	Volvox	*Chlamydomonas*	Arabidopsis	*Monosiga*	sea anemone	human
Total number of domains in proteome	10,318	10,168	38,887	11,786	30,535	42,057
No. different PFAM domains	2,431	2,354	3,028	2,232	3,078	3,832
No. different pairwise combinations	1,392	1,219	1,838	2,128	2,723	4,038
No. proteins with 1 domain	5,368	5,437	15,547	4,154	12,843	11,570
No. proteins with 2 domain types	989	880	3,639	1,157	2,456	3,543
No. proteins > 2 domain types	287	267	1,193	494	797	1,799

Table S10: Complete predicted protein sets used to build protein families

The genus and species, together with abbreviations used in e.g. Table S12 as well as their version and notes are shown for all proteomes used to make protein families (see above).

Species name	Abbreviation	Version and Notes
Cyanidioschyzon merolae 10D	Cme	release Apr 8, 2004; http://merolae.biol.s.u-tokyo.ac.jp/download
Synechocystis sp. PCC 6803	Syn	complete genome - 0..3573470 GenBank Accession NC_000911
Pseudomonas aeruginosa PA01	Pae	complete genome - 0..6264403 GenBank Accession NC_002516
Staphylococcus aureus subsp. aureus N315	Sau	complete genome - 0..2814816 GenBank Accession NC_002745
Dictyostelium discoideum	Ddi	dictyBase.org; Full Chromosomes made 10/05/2004; Primary Features made 7/11/2005
Tetrahymena thermophila SB210	Tth	*Tetrahymena* Genome Database (TIGR) Aug 2004
Phytophthora ramorum	Pra	JGI v.1 http://genome.jgi-psf.org/Phyra1_1/Phyra1_1.home.html
Phytophthora sojae	Pso	JGI v.1 http://genome.jgi-psf.org/sojae1/sojae1.home.htmlsojae1

Neurospora crassa	Ncr	http://fungal.genome.duke.edu, genome neurospora_crassa.20020212.nt.gz
Prochlorococcus marinus str. MIT9313	Pma	2003 JGI/ORNL http://genome.jgi-psf.org/prom9/prom9.home.html
Arabidopsis thaliana	Ath	TAIR6, updated 11.2005 from NCBI ftp://ftp.ncbi.nih.gov/genomes/Arabidopsis_thaliana
Homo sapiens	Hsa	NCBI 36 from ensembl build 38
Caenorhabditis elegans	Cel	WS 150 from ensembl build 38
Ostreococcus tauri	Ota	JGI v2.0 http://genome.jgi-psf.org/Ostta4/Ostta4.home.html
Ostreococcus lucimarinus	Olu	(*O. pacifica*; *Ostreococcus* CCE9901) JGI v. 2.0 http://genome.jgi-psf.org/Ost9901_3/Ost9901_3.home.html
Physcomitrella patens	Ppa	JGI v. 1 http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.download.ftp.html
Monosiga brevicollis	Mbr	JGI v. 1 http://genome.jgi-psf.org/Monbr1/Monbr1.home.html
Thalassiosira pseudonana	Tps	JGI v. 3.0 http://genome.jgi-psf.org/Thaps3/Thaps3.home.html
Naegleria gruberi	Ngr	JGI v.1 http://genome.jgi-psf.org/Naegr1/Naegr1.home.html
Paramecium tetraurelia	Pte	peptides from macronuclear genome downloaded from Paramecium DB release date 28-MCH-2007
Chlamydomonas reinhardtii	Cre	JGI v.3.1 freeze for GenBank submission 9/13/2007 from http://genome.jgi-psf.org/Chlre4/Chlre4.download.ftp.html
Volvox carteri	Vca	JGI v1 freeze from http://genome.jgi-psf.org/Volca1/Volca1.download.ftp.html

Table S11: Protein family size distribution in *Volvox* and *Chlamydomonas*

The number of protein families containing 1, 2-5 or more than 5 proteins from *Chlamydomonas* (columns across) and *Volvox* (rows down) are shown.

	Chlamydomonas proteins in family		
Volvox proteins in family	1	2-5	>5
1	5,423	295	2
2-5	282	669	13
>5	10	19	33

Table S12: *Volvox*-specific gene models with EST evidence

Presence of EST support and its quality is shown as counts of putative *Volvox*-specific genes that either have homology to another putative *Volvox*-specific gene (left column) or do not have such homology (right column), suggesting that these proteins might belong to *Volvox*-specific families, or might represent singleton *Volvox*-specific proteins respectively.

Does protein have a hit to another putative *Volvox*-specific protein?	yes	no
Full-length EST support	16	9
EST support over part of the gene model	11	0
Problem with EST support	57	47
Total	84	58

Table S13: *Chlamydomonas*-specific gene models with EST support

Presence of EST support and its quality is shown as fractions of a random sample of putative *Chlamydomonas*-specific genes that either have homology to another putative *Chlamydomonas*-specific gene (left column) or do not have such homology (right column), suggesting that these proteins might belong to *Chlamydomonas*-specific families, or might represent singleton *Chlamydomonas*-specific proteins respectively.

Does protein have a hit to another putative *Chlamydomonas*-specific protein?	Yes	No
Consistent EST support	32 %	60 %
EST probably supports gene model	30 %	16 %
Problem with EST support	38 %	24%

Table S14: Proteins involved in processes that are associated with increased developmental complexity in *Volvox* relative to *Chlamydomonas*

In the table, the names given are gene symbols, with synonyms given after a forward slash. Symbols of paralogs/co-orthologs are separated by semi-colons. The JGI protein ID and defline are given in the next two columns. The following columns show information for *Chlamydomonas* (co-)orthologs. Where there is no gene symbol and protein ID in a column, a homolog could not be found. The ID of the protein family the proteins belong to is shown after the *Chlamydomonas* defline and is followed by abbreviations of all the species that have a member in that protein family. If the protein does not belong to a family, this columns shows ‘unclustered’. The abbreviations used are as follows: Cme, *Cyanidioschyzon merolae*; Syn, *Synechocystis* sp.; Pae, *Pseudomonas aeruginosa*; Sau, *Staphylococcus aureus*; Ddi, *Dictyostelium discoideum*; Tth, *Tetrahymena thermophila*; Pra, *Phytophthora ramorum*; Pso, *Phytophthora sojae*; Ncr, *Neurospora crassa*; Pma, *Prochlorococcus marinus*; Ath, *Arabidopsis thaliana*; Hsa, *Homo sapiens*; Cel, *Caenorhabditis elegans*; Ota, *Ostreococcus tauri*; Olu, *Ostreococcus lucimarinus*; Ppa, *Physcomitrella patens*; Mbr, *Monosiga brevicollis*; Tps, *Thalassiosira pseudonana;* Ngr, *Naegleria gruberi*; Pte, *Paramecium tetraurelia*; Cre, *Chlamydomonas reinhardtii*; Vca, *Volvox carteri.*

Table S12

Volvox protein name, synonyms and paralogs	Volvox v1 JGI protein ID(s)	Volvox Defline from JGI	Chlamydomonas protein (and paralogs)	homolog v3.1 JGI protein ID(s)	Chlamydomonas homolog v4 JGI protein ID(s)	Phytozome v5.0/Augustus u9 transcript ID	Chlamydomonas JGI defline	Protein family ID	species in protein family
Protein secretion and membrane trafficking									
RabA	60379	Rab-related GTPase Rab-related GTPase RabA/Rab11	RABA1	195519	195519	Au9.Cre03.g189250	small Rab-related GTPase	6572769	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
YptV4/RabB	106534	small G-protein (yptV4) RabB/Rab2	RABB1	148836	148836	Au9.Cre02.g126100	small Rab-related GTPase	6572135	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
RabC1	66052	Rab-related GTPase RabC/Rab18	RABC1	195518	195518	Au9.Cre09.g386900	small Rab-related GTPase	6571340	Cre Vca
RabC2	78758	Rab-related GTPase RabC/Rab18	RABC2	195872	195872	Au9.Cre13.g595450	small Rab-related GTPase	6574558	Cme Pra Pso Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Cre Vca
YptV3/RabC3	108490	small G-protein (yptV3) RabC/Rab18	RABC3	24345		Au9.Cre11.g482900	small Rab-related GTPase	6574558	Cme Pra Pso Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Cre Vca
YptV1/RabD	104295	small G-protein (yptV1) RabD/Rab1	RABD1	60490	60490	Au9.Cre12.g560150	small Rab-related GTPase	6576078	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
YptV2/RabE	83299	small G-protein (yptV2) RabE/Rab8	RABE1	195520	195520	Au9.Cre15.g641800	small Rab-related GTPase	6571469	Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Tps Ngr Pte Cre Vca
RabF	108963	RabF/Rab5	RABF1	81259	81259	Au9.Cre12.g517400	small Rab-related GTPase	6572121	Ddi Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Cre Vca
YptV5/RabG	79428	small G-protein (yptV5) RabG/Rab7	RABG1	195521			RABG1, small Rab-related GTPase; small Rab-related GTPase	6577001	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
RabH	73845	Rab-related GTPase RabH/Rab6	RABH1	195522	195522	Au9.Cre01.g047550	small Rab-related GTPase	6575479	Cme Ddi Tth Pra Pso Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Rabl2A	102687	Rab-related GTPase Rab-related GTPase	RABL1/RABL2A	192441	192441	Au9.Cre17.g722350	small Rab-related GTPase	6574820	Tth Pra Pso Hsa Mbr Tps Ngr Pte Cre Vca
Rab23	115722	Rab-related GTPase Rab23	Rab23	195517			RAB23, small Rab-related GTPase; small Rab related GTPase	6574017	Pra Pso Hsa Ppa Mbr Ngr Cre Vca
Fap156	80291	Rab-related GTPase	FAP156	129193	129193	Au9.Cre01.g047950	small Rab-related GTPase	6572642	Tth Pso Hsa Mbr Tps Ngr Pte Cre Vca
Rab28	62839	Rab-related GTPase Rab28-like	RAB28	195523	195523	Au9.Cre17.g715050	small Rab-related GTPase	6571252	Cre Vca
Syp1	89085	Qa-SNARE Sso1/Syntaxin1 (PM)-type	SYP1	195969	195969	Au9.Cre13.g588550	Qa-SNARE Sso1/Syntaxin1 PM-type	6576792	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Syp3	81752	Qa-SNARE Sed5/Syntaxin5-family	SYP3	195412	195412	Au9.Cre16.g692050	Qa-SNARE protein Sed5/Syntaxin5-family	6570757	Cme Ddi Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Cre Vca
Syp4	103767	Qa-SNARE Tlg2/Syntaxin16-family	SYP4	195401	195401	Au9.Cre12.g507450	Qa-SNARE protein Tlg2/Syntaxin16-family	6575295	Cme Ddi Pra Pso Ncr Ath Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Syp5	91993	Qc-SNARE Syn8/Syntaxin8-family	SYP5	189016	189016	Au9.Cre17.g709350	Qc-SNARE protein Syn8/Syntaxin8-family	6577946	Ath Hsa Ppa Mbr Ngr Cre Vca
Syp6	55748	Qc-SNARE Tlg1/Syntaxin 6-family	SYP6	130559	130559	Au9.Cre06.g290100	Qc-SNARE protein Tlg1/Syntaxin 6-family	6576050	Pra Pso Ath Hsa Ota Olu Ppa Mbr Cre Vca
Syp71	91067	Qc-SNARE SYP7-family	SYP71 SYP72	195405 195406	195405 195406	Au9.Cre02.g099050; Au9.Cre02.g098950	Qc-SNARE, SYP7-family; Qc-SNARE protein, SYP7-family Qc-SNARE protein SYP7-family	6576050	Pra Pso Ath Hsa Ota Olu Ppa Mbr Cre Vca
Syp8	96826		SYP8	195411	195411	Au9.Cre17.g711450	Qa-SNARE protein Ufe1/Syntaxin 18 family	6572660	Ath Olu Ppa Cre Vca
Syp6L	55748	Qc-SNARE Tlg1/Syntaxin 6-family	SYP6L	160025	160025	Au9.Cre06.g293100	Qc-SNARE SYP6-like protein	6576050	Pra Pso Ath Hsa Ota Olu Ppa Mbr Cre Vca
Syp2	86761		SYP2	195407	406247	Au9.Cre01.g010700	Qa-SNARE protein Pep12/Syntaxin 7-family	6573981	Ath Ota Olu Ppa Cre
Vamp75	54976		VAM5/VAMP75	195409	195409	Au9.Cre13.g596900	R-SNARE protein VAMP71-family	6571705	Cme Ddi Tth Pra Pso Ncr Ath Hsa Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Sec22	80612	R-SNARE Sec22-family	SEC22	57076	57076	Au9.Cre12.g554700	R-SNARE protein Sec22-family	6572658	Cme Ddi Tth Pra Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Ykt6	75062		YKT6	195465	195465	Au9.Cre17.g728150	R-SNARE protein YKT6-family	6571909	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Vamp71	70315		VAM1/VAMP71	128777	128777	Au9.Cre04.g224750	R-SNARE protein VAMP72-family	6571254	Cre Vca
Vamp72	70298		VAM2/VAMP72	195403	195403	Au9.Cre04.g225900	R-SNARE protein VAMP72-family	6571705	Cme Ddi Tth Pra Pso Ncr Ath Hsa Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Vamp73	78158		VAM3/VAMP73	195404	195404	Au9.Cre04.g225850	R-SNARE protein VAMP72-family	6571254	Cre Vca
			VAM4/VAMP74	136188	136188	Au9.Cre04.g224800	R-SNARE protein VAMP72-family	6571254	Cre Vca
Bet1	120875	Qc-SNARE Bet1/mBET1 family	BET1	183904	183904	Au9.Cre03.g198100	Qc-SNARE protein Bet1/mBET1 family	6571125	Ath Ppa Mbr Cre Vca
Bet3.1 Bet3.2	76384 76381		BET3	184613	184613	Au9.Cre01.g059600		6573174	Cme Ddi Tth Pra Ncr Ath Hsa Cel Olu Ppa Mbr Tps Ngr Pte Cre Vca
Tml1	86534	R-SNARE Tomsyn-like family	SNR7/TML1	195468	195468	Au9.Cre12.g558250	R-SNARE protein Tomsyn-like family	6573056	Ath Ppa Cre Vca
Snap34	104786		SNAP34	195408	195408	Au9.Cre02.g100250	Qb+c-SNARE, SNAP25-family; Qb+c-SNARE protein SNAP25-family	6575462	Hsa Cel Cre Vca
Memb1	104271	Qb-SNARE Bos1/Membrin family	MEMB1	195462	195462	Au9.Cre12.g554200	Qb-SNARE protein Bos1/Membrin family	6572992	Cme Ddi Pso Ath Hsa Cel Ota Olu Ppa Tps Cre Vca
SnapG1	95241		SNAPG1	173710	173710	Au9.Cre08.g385700	gamma SNAP	6574387	Ddi Pra Pso Ath Hsa Cel Olu Ppa Ngr Cre Vca
SnapA1	81196		SNAPA1	23974	23974	Au9.Cre02.g099150	alpha-SNAP	6577758	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Cdc48	78972		CDC48	134171	134171	Au9.Cre06.g269950	Protein involved in ubiquitin-dependent degradation of ER-bound substrates	6574344	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Vps45	77223		VPS45	195471	195471	Au9.Cre07.g333950		6571169	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Vps33	103690		VPS33	195470	195470	Au9.Cre12.g515550		6570478	Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Ngr Pte Cre Vca
Actin cytoskeleton									
MyoA	82838	type XI myosin heavy chain MyoA	MYO1	185104	185104	Au9.Cre16.g658650	myosin heavy chain class XI	6572947	Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
MyoB	57247	type XI myosin heavy chain MyoB	MYO2	119317	119317	Au9.Cre13.g563800	myosin heavy chain class XI	6572947	Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
MyoC	107085	type VIII myosin heavy chain MyoC	MYO3	113760	113760	Au9.Cre09.g416250	myosin heavy chain class VIII	6572947	Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
MyoD	103833	class XI myosin heavy chain MyoD						unclustered	unclustered
MyoE	90836	putative class XI myosin MyoE	MYO5	190489	377863	Au9.Cre03.g197800	myosin heavy chain class XI	6571332	Cre Vca
MyoF	95387	type XI myosin MyoF						unclustered	unclustered
ActA	109972	actin	IDA5	24392	24392	Au9.Cre13.g603700		6571555	Cme ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Nap1	127374	novel actin-like protein	NAP1	168932				6570522	Cme Tth Pra Pso Ncr Cel Mbr Tps Ngr Cre Vca
Arp2	107669	actin-related protein Arp2	ARP2	24114	24114		Actin-related protein	6575022	Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Ngr Pte Cre Vca
Arp3	101864	actin-related protein Arp3	ARP3	118745	118745	Au9.Cre16.g676050	Actin-related protein	6572066	Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Ngr Pte Cre Vca
Arp4	69220	actin-related protein Arp4	ARP5	112378	112378	Au9.Cre06.g249300	Actin-related protein	6576600	Cme Ddi Tth Pra Pso Ath Ota Olu Ppa Mbr Ngr Cre Vca
Arp6	121135	actin-related protein Arp6					Actin-related protein	6577015	Pso Olu Ppa Vca
Arp7	94505	actin-related protein Arp7	ARP7	103067	103067	Au9.Cre12.g545000	Actin-related protein	6573348	Ath Ota Olu Ppa Cre Vca
ArpC1	58274	actin-related protein ArpC1	ARPC1	177203	177203	Au9.Cre16.g648100	ARP2/3 (actin-related protein) complex	6575904	Ddi Tth Pra Pso Ncr Ath Hsa Cel Ppa Mbr Ngr Pte Cre Vca
ArpC2	119936	actin-related protein ArpC2	ARPC2		518332		Actin-related protein	6575960	Ath Ota Olu Ppa Vca
ArpC3	65632	actin-related protein ArpC3	ARPC3	65632	399563		Actin-related protein	6575290	Ddi Tth Pra Pso Ncr Ath Hsa Cel Ppa Mbr Ngr Pte Cre Vca
ArpC4	72447	actin-related protein ArpC4	ARPC4	115207	400848	Au9.Cre06.g260800	Actin-related protein	6574535	Ddi Tth Ncr Ath Cel Ota Olu Ppa Mbr Ngr Pte Cre Vca
ForA	107471	putative protein containing a FH2 domain formin	FOR1	144027	144027	Au9.Cre03.g166700	formin	6570922	Cme Ddi Tth Pra Pso Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
AdfA	68164	actin-depolymerizing factor AdfA		183437	378877	Au9.Cre07.g339050	actin-depolymerizing factor	6577211	Cme Ddi Tth Pra Pso Ncr Ath Hsa Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
gelsolin	108308	actin-binding protein Gelsolin	gelsolin	190925	190925	Au9.Cre12.g555600	gelsolin	6570979	Cme Ddi Pra Pso Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Cre Vca
PrfA	106260	profilin	PRO1	181887	181887	Au9.Cre10.g427250	profilin	6571558	Cme Ddi Ncr Ath Cel Olu Ppa Mbr Cre Vca
CapA	79641	actin monomer-binding protein Cap/Srv2p	CAP1	115172	402056	Au9.Cre06.g304100	actin monomer-binding protein Cap/Srv2p	6572424	Cme Syn Pae Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
VilA	97468	villin-like protein	VIL1	206215	206215	Au9.Cre12.g493700	f-actin-binding Viilin-like protein	6571606	Cre Vca
Microtubule cytoskeleton									
TubA1 TubA2	77526 109786	alpha tubulin (tubA2)	TUA1 TUA2	128523 186023	128523 186023	Au9.Cre03.g190950; Au9.Cre04.g216850	alpha tubulin 1 alpha tubulin 2	6575816	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca

TubB1 TubB2	75910 77081	beta tubulin (tubB2)	TUB1 TUB2	129876 129868	129876 129868	Au9.Cre12.g542250; Au9.Cre12.g549550	beta tubulin 1 beta tubulin 2	6571720	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
TubG	80553	Gamma tubulin	TUG1	188933	188933	Au9.Cre06.g299300	Gamma tubulin was TUG	6576302	Cme Ddi Tth Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
TubD	60395	Delta tubulin	UNI3	136082	136082	Au9.Cre03.g187350	delta tubulin	6572439	Tth Pra Hsa Ppa Mbr Ngr Pte Cre Vca
TubH	116596	Eta tubulin	TUH1	154376	154376	Au9.Cre12.g513450	Eta tubulin was TUH	6574884	Ngr Cre Vca
TubE	56250	Epsilon tubulin	TUE1	188195	188195	Au9.Cre03.g172650	Epsilon tubulin was TUE	6574637	Tth Pra Pso Hsa Ppa Mbr Ngr Pte Cre Vca
centrin CnrA	109845 66997	putative centrin	VFL2/CEN1	159554	159554	Au9.Cre11.g468450		6577172	Tth Pra Pso Ath Hsa Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
KatA	82654	katanin catalytic subunit 60 kDa	KAT1	53314	53314	Au9.Cre10.g427600	katanin catalytic subunit 60 kDa	6574921	Tth Pra Pso Ath Hsa Cel Ppa Mbr Ngr Pte Cre Vca
KatB	76999	katanin p60 catalytic subunit	VPS4	98650	98650	Au9.Cre10.g446400	katanin p60 catalytic subunit was KAT2	6572037	Tth Pra Pso Ath Hsa Ppa Mbr Tps Ngr Pte Cre Vca
KatC	94919	microtubule severing protein katanin p80 subunit	PF15	80954	80954	Au9.Cre03.g160450		6573852	Pra Pso Ath Hsa Ppa Mbr Ngr Cre Vca
Map1	99065	microtubule associated protein (MAP) 1a						unclustered	unclustered
MorA	121331	microtubule organizing protein MorA	TOG1	175143	175143	Au9.Cre03.g149800	Microtubule associated protein	6572022	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
map65	120144	microtubule-associated protein MAP65	MAP65	193464	394462	Au9.Cre14.g614050		6575396	Ath Ppa Cre Vca
EB1	107043	microtubule plus-end binding protein EB1	EBP1/EB1	194209	194209	Au9.Cre17.g741200	microtubule plus-end binding protein	6573189	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Clasp	120833	CLIP-associating protein	CLASP	206206	206206	Au9.Cre09.g415700	CLIP-associating protein	6575258	Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Ngr Cre Vca
Spr1	103536	putative cortical microtubule associated protein SPIRAL1	SPR1	191409	416188	Au9.Cre02.g130150		6571789	Pra Pso Ota Olu Ppa Cre Vca
Gcp2	105911	gamma tubulin interacting protein	GCP2	150712	150712	Au9.Cre12.g525500	Gamma tubulin interacting protein	6576345	Ddi Tth Pra Pso Ncr Ath Hsa Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Gcp3	97867	gamma tubulin interacting protein	GCP3	152585	152585	Au9.Cre22.g764400	Gamma tubulin interacting protein	6576345	Ddi Tth Pra Pso Ncr Ath Hsa Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Gcp4	118030	gamma tubulin interacting protein	GCP4	146323	146323	Au9.Cre01.g019150	Gamma tubulin interacting protein	6573454	Tth Pra Pso Ath Hsa Ota Olu Ppa Mbr Ngr Pte Cre Vca
PldA	99461	putative MT associated signaling protein phospholipase D	PLD1	190403	190403	Au9.Cre13.g591900	putative MT associated signaling protein	6573958	Pae Tth Pra Pso Hsa Olu Mbr Pte Cre Vca
D1bLIC	92778	Cytoplasmic dynein 1b light intermediate chain D1bLIC		130394	130394	Au9.Cre02.g135900	Cytoplasmic dynein 1b light intermediate chain D1bLIC	6572451	Tth Pra Pso Hsa Cel Ngr Cre Vca
D1bHC	64869	Cytoplasmic dynein 1b heavy chain	DHC12/DHC1B	24009	24009	Au9.Cre06.g250300		6571949	Ddi Tth Pra Pso Ncr Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
Asp	121726	microtubule-associated protein Asp	ASP	174686	306828	Au9.Cre06.g281150	abnormal spindle protein	6573164	Pra Pso Hsa Olu Ppa Mbr Tps Ngr Cre Vca
	79641			115172	402056	Au9.Cre06.g304100		6572424	Cme Ssp Pae Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
TtlA	59059	tubulin tyrosine ligase	TTL1	170153	401790	Au9.Cre06.g300250	Tubulin tyrosine ligase	6575448	Tth Pra Pso Ath Tps Pte Cre Vca
TtlB	30444	tubulin tyrosine ligase	TTL2/FAP267	100760	100760	Au9.Cre17.g699500	Tubulin tyrosine ligase	6573843	Ddi Tth Pra Pso Hsa Cel Ota Olu Ppa Mbr Ngr Pte Cre Vca
TtlC	65051		TTL3	119250	119250	Au9.Cre01.g059200	Tubulin tyrosine ligase	6572934	Tth Pra Pso Hsa Mbr Pte Cre Vca
TtlD	105441							6576582	Tth Hsa Pte Vca
TtlE	107180		TTL5	190829	513852	Au9.Cre12.g547700	Tubulin tyrosine ligase	6574391	Cre Vca
TtlF	69077		TTL6	146893	146893	Au9.Cre01.g050450	Tubulin tyrosine ligase	6571683	Pra Mbr Cre Vca
TtlG	99465		CYG40/TTL7	190398	190398	Au9.Cre13.g591700	Tubulin tyrosine ligase	unclustered	unclustered
TtlH	108470		TTL8	176529	417267	Au9.Cre02.g120750	Tubulin tyrosine ligase	6570722	Tth Pso Hsa Cel Mbr Tps Ngr Pte Cre Vca
Basal Body Proteins									
bbs5	78967	Bardet-Biedl syndrome 5	BBS5	182299	182299	Au9.Cre06.g267550	Bardet-Biedl syndrome 5 protein	6570754	Tth Pra Pso Hsa Cel Mbr Ngr Pte Cre Vca
SFA	84701	SF-assemblin	SFA	127995	127995	Au9.Cre07.g332950	SF-assemblin	6577509	Pra Pso Ota Olu Ngr Cre Vca
bbs8	78311	putative TRP protein for flaggelar function	BBS8	140113	140113	Au9.Cre16.g666500	TRP protein for ciliary function	6571806	Tth Pra Pso Hsa Cel Mbr Ngr Pte Cre Vca
			BBS7	190054	190054	Au9.Cre01.g043750	Bardet-Biedl syndrome 7 protein	6574722	Tth Pra Pso Hsa Cel Mbr Ngr Pte Cre
Bbs4	66874	Bardet-Biedl syndrome protein 4	BBS4	129948	129948	Au9.Cre12.g548650	Bardet Biedl syndrome 4 protein	6577091	Ddi Tth Pra Pso Ncr Hsa Cel Mbr Ngr Pte Cre Vca
arf8	70270	small Arf-related GTPase	BBS3B	24475	24475	Au9.Cre16.g664500	Bardet-Biedl syndrome protein 3B	6572906	Tth Pra Pso Hsa Cel Ngr Pte Cre Vca
Bbs2	121664	Bardet-Biedl syndrome protein 2	BBS2	126758	126758	Au9.Cre06.g257250	Bardet-Biedl syndrome protein 2	6578098	Pra Pso Hsa Cel Mbr Ngr Pte Cre
Bbs1	84205	Bardet-Biedl syndrome protein 1	BBS1	132537	132537	Au9.Cre17.g741950	Bardet-Biedl syndrome protein 1	6576237	Tth Pra Pso Cel Mbr Ngr Pte Cre Vca
Ofd1	90133	basal body protein	OFD1	31640	31640	Au9.Cre17.g703600	basal body protein	6576652	Tth Ppa Pte Cre Vca
Vfl3	84510	protein conserved only in organisms with motile cilia	VFL3	130542	130542	Au9.Cre06.g279900	protein required for templated centriole assembly	6577987	Tth Pra Pso Hsa Ppa Mbr Pte Cre Vca
bldJ	57967	basal body protein	BLD10	166062	166062	Au9.Cre10.g418250	basal body protein	6573326	Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Tps Ngr Pte Cre Vca
bbs9	119731	Bardet-Biedl syndrome 9	BBS9	101137	101137		Bardet-Biedl syndrome 9; Bardet-Biedl syndrome 9 protein	6572522	Tth Pra Pso Hsa Mbr Ngr Pte Cre Vca
Kinesin Motor Proteins									
	93532	kinesin-like protein		191502				unclustered	unclustered
flaJ	107307	Kinesin-II Motor Protein	FLA10	185750	185750	Au9.Cre17.g730950		6573133	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
flaH/klpA	103736	kinesin-like protein	FLA8	150766	150766	Au9.Cre12.g522550	Kinesin-II motor subunit	6573133	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
	97023	putative Kif3C kinesin		175469	345354	Au9.Cre09.g415450		6574543	Tth Pra Pso Hsa Cel Ota Olu Ppa Ngr Pte Cre Vca
	58470	Kif9 type kinesin similar to C. reinhardtii KLP1	KLP1	186414	186414	Au9.Cre02.g073750	Kinesin-like protein	6575085	Tth Pra Pso Hsa Ppa Mbr Ngr Pte Cre Vca
	94697			146648				unclustered	unclustered
	64266	Kif6 type kinesin-like protein						6573711	Tth Pra Pso Hsa Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
	76107	putative Kif9 kinesin		186275				unclustered	unclustered
invA	127192	kinesin invA	IAR1	126081	126081	Au9.Cre10.g418950	kinesin family protein	6576834	Cre Vca
	127420	kinesin-like protein		143399				unclustered	unclustered
	127421	kinesin-like protein		149713	149713	Au9.Cre09.g386700	kinesin motor family protein	unclustered	unclustered
	127422	kinesin-like protein		13743	13743	Au9.Cre03.g202000	Kinesin family member heavy chain	6573133	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
	127423	kinesin-like protein		187696				unclustered	unclustered
	105173	kinesin-like protein		188286				unclustered	unclustered
	99650	subfamily 14A kinesin		147592				unclustered	unclustered
	95520	kinesin-like protein		149888				unclustered	unclustered
	127417	kinesin-like protein		193028				6570717	Cme Ddi Tth Pra Pso Ncr Ath Hsa Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
	95391	kinesin related to Arabidopsis ATK5		194730				unclustered	unclustered
	127355	Kar3 member kinesin-like protein		131637				6575670	Tth Pra Pso Ath Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
	127424	kinesin-like protein						unclustered	unclustered
	93886	kinesin-like protein		194730				unclustered	unclustered
	127418	kinesin-like protein		192784				unclustered	unclustered
	106268	kinesin-like protein		141042				6577255	Cre Vca
	94482	similar to kinesin FAP125	FAP125	190816	190816	Au9.Cre12.g546100	kinesin-like protein	6573456	Ddi Tth Ath Hsa Pte Cre Vca
	93168	appears to contain kinesin motor domain		143755				6576941	Pra Pso Ppa Vca
	96903	kinesin-like protein		143993				unclustered	unclustered
	98132	putative Kif21-type kinesin		167999				6571365	Cre Vca
	127414	kinesin-like protein		143714				6577520	Pra Cre Vca
	31481			106906	408759	Au9.Cre17.g735200		6570716	Ath Ppa Cre Vca
	127427			180630				6576263	Tth Pra Pso Ota Olu Ppa Mbr Ngr Pte Cre Vca
	44127	Kif4A type kinesin		142871				6575075	Pso Ppa Cre Vca
	127415	kinesin-like protein						unclustered	unclustered
Cell wall and Extracellular Matrix									
***Volvox* pherophorin homologs**									
ssgA	127165	pherophorin-like ECM-glycoprotein SSG185						6574635	Cre Vca
phI	104381	extracellular matrix glycoprotein pherophorin I						6577727	Cre Vca

phII	85077	extracellular matrix glycoprotein pherophorin II						6574211	Cre Vca
phIII	83859	extracellular matrix glycoprotein pherophorin III						unclustered	unclustered
phS	104453	extracellular matrix glycoprotein pherophorin-S						unclustered	unclustered
phDZ1	77905	extracellular matrix glycoprotein pherophorin-DZ1						unclustered	unclustered
phDZ2	90188	extracellular matrix glycoprotein pherophorin-DZ2						unclustered	unclustered
phV1	77332	extracellular matrix glycoprotein pherophorin-V1						unclustered	unclustered
phV2	108351	extracellular matrix glycoprotein pherophorin-V2						unclustered	unclustered
phV3	110102	extracellular matrix glycoprotein pherophorin-V3						unclustered	unclustered
phV4	77335	extracellular matrix glycoprotein pherophorin-V4						unclustered	unclustered
phV5	77338	extracellular matrix glycoprotein pherophorin-V5						unclustered	unclustered
phV6	107455	extracellular matrix glycoprotein pherophorin-V6						unclustered	unclustered
phV7	108791	extracellular matrix glycoprotein pherophorin-V7						unclustered	unclustered
phV8	63512	extracellular matrix glycoprotein pherophorin-V8						unclustered	unclustered
phV9	75691	extracellular matrix glycoprotein pherophorin-V9 (phV9/sef5)						unclustered	unclustered
phV10	104451	extracellular matrix glycoprotein pherophorin-V10						unclustered	unclustered
phV11	104452	extracellular matrix glycoprotein pherophorin-V11						6577727	Cre Vca
phV12	59555	extracellular matrix glycoprotein pherophorin-V12						unclustered	unclustered
phV13	127206	extracellular matrix glycoprotein pherophorin-V13						unclustered	unclustered
phV14	80812	extracellular matrix glycoprotein pherophorin-V14						unclustered	unclustered
phV15	80731	extracellular matrix glycoprotein pherophorin-V15						6577727	Cre Vca
phV16	78132	extracellular matrix glycoprotein pherophorin-V16						unclustered	unclustered
phV17	100178	extracellular matrix glycoprotein pherophorin-V17						unclustered	unclustered
phV18	80872	extracellular matrix glycoprotein pherophorin-V18						6574211	Cre Vca
phV19	83847	extracellular matrix glycoprotein pherophorin-V19 (phV19/sef6)						unclustered	unclustered
phV20	42640	extracellular matrix glycoprotein pherophorin-V20						6574211	Cre Vca
phV21	90431	extracellular matrix glycoprotein pherophorin-V21						6574211	Cre Vca
phV22	94616	extracellular matrix glycoprotein pherophorin-V22						6577727	Cre Vca
phV23	94560	extracellular matrix glycoprotein pherophorin-V23						unclustered	unclustered
phV24	127208	extracellular matrix glycoprotein pherophorin-V24						unclustered	unclustered
phV25	101351	extracellular matrix glycoprotein pherophorin-V25						unclustered	unclustered
phV26	107546	extracellular matrix glycoprotein pherophorin-V26 (similar to Chlamydomonas GAS30)						6577070	Cre Vca
phV27	67890	extracellular matrix glycoprotein pherophorin-V27 (similar to Chlamydomonas GAS30)						6577070	Cre Vca
phV28	77438	extracellular matrix glycoprotein pherophorin-V28 (similar to Chlamydomonas GAS28)						6576940	Cre Vca
phV29	67883	extracellular matrix glycoprotein pherophorin-V29 (similar to Chlamydomonas GAS30)						6577070	Cre Vca
phV30	67897	extracellular matrix glycoprotein pherophorin-V30 (similar to Chlamydomonas GAS30)						6577070	Cre Vca
phV31	104151	extracellular matrix glycoprotein pherophorin-V31 (similar to Chlamydomonas GAS31)						6576285	Cre Vca
phV32	104390	extracellular matrix glycoprotein pherophorin-V32 (similar to Chlamydomonas GAS30)						6571970	Cre Vca
phV33	95611	extracellular matrix glycoprotein pherophorin-V33						6570639	Ath Ota Olu Ppa Cre Vca
phV34	107649	extracellular matrix glycoprotein pherophorin-V34						unclustered	unclustered
phV35	127252	extracellular matrix glycoprotein pherophorin-V35						6574700	Cre Vca
phV36	89832	extracellular matrix glycoprotein pherophorin-V36						6577727	Cre Vca
phV37	127212	extracellular matrix glycoprotein pherophorin-V37						unclustered	unclustered
phV38	55073	extracellular matrix glycoprotein pherophorin-V38						6574211	Cre Vca
phV39	100779	extracellular matrix glycoprotein pherophorin-V39						6574211	Cre Vca
phV40	106281	extracellular matrix glycoprotein pherophorin-V40						6577727	Cre Vca
phV41	97824	extracellular matrix glycoprotein pherophorin-V41						6575769	Cre Vca
phV42	90414	extracellular matrix glycoprotein pherophorin-V42						6574211	Cre Vca
***Chlamydomonas* Pherophorin homologs**									
			PHC1	196399	196399	Au9.Cre17.g717900	cell wall protein pherophorin-C1	6577727	Cre Vca
			PHC2	196402	196402	Au9.Cre14.g620600	cell wall protein pherophorin-C2	6577727	Cre Vca
			PHC3	196403	196403	Au9.Cre13.g596450	cell wall protein pherophorin-C3	6574635	Cre Vca
			PHC4	196405	196405	Au9.Cre12.g549000	cell wall protein pherophorin-C4	6577727	Cre Vca
			PHC5	196406	196406	Au9.Cre05.g238650	cell wall protein pherophorin-C5	unclustered	unclustered
			PHC6	196407	196407	Au9.Cre17.g718000	cell wall protein pherophorin-C6	unclustered	unclustered
			PHC7	195997	195997		cell wall protein pherophorin-C7	6577727	Cre Vca
			PHC8	196005	196005	Au9.Cre17.g717850	cell wall protein pherophorin-C8	6577727	Cre Vca
			PHC9	196020	196020	Au9.Cre06.g299150	cell wall protein pherophorin-C9	6574211	Cre Vca
			PHC10	196024	196024	Au9.Cre09.g404150	cell wall protein pherophorin-C10	unclustered	unclustered
			PHC11	196027	196027	Au9.Cre07.g331100	cell wall protein pherophorin-C11	unclustered	unclustered
			PHC12	194264	394232	Au9.Cre11.g472250		6575769	Cre Vca
			PHC13	196029	196029	Au9.Cre14.g620650	cell wall protein pherophorin-C13	6577727	Cre Vca
			PHC14	536129				unclustered	unclustered
			PHC15	148333	148333	Au9.Cre09.g396100		6570645	Cme Ddi Pra Pso Ncr Ath Hsa Ota Olu Ppa Mbr Tps Ngr Cre Vca
			PHC16	141291	141291	Au9.Cre02.g078800		unclustered	unclustered
			PHC17	164137	164137	Au9.Cre05.g238850		6575036	Cre Vca
			PHC18	522381	522381			unclustered	unclustered
			PHC19	189163	189163	Au9.Cre17.g696500		6574700	Cre Vca
			PHC20	196022	196022	Au9.Cre09.g388250	cell wall protein pherophorin-C20	6570639	Ath Ota Olu Ppa Cre Vca
			PHC21	93464	93464	Au9.Cre02.g094450		unclustered	unclustered
			PHC22	94393	94393	Au9.Cre17.g696700		6574700	Cre Vca
			PHC23	196019			cell wall protein pherophorin-C23	unclustered	unclustered
			PHC24	196025			cell wall protein pherophorin-C24	6577727	Cre Vca
			GAS28	192908	192908	Au9.Cre11.g481600	hydroxyproline-rich glycoprotein, stress-induced	6576940	Cre Vca
			GAS30	195828	195828	Au9.Cre11.g481750	hydroxyproline-rich glycoprotein, stress-induced	6577070	Cre Vca
			GAS31	193780	193780	Au9.Cre45.g788350	cell wall protein pherophorin	6576285	Cre Vca
Sex-inducer									
sex1	83768	sex-inducer, sex-inducing pheromone						unclustered	unclustered

sex2	67483	sex-inducer sex-inducing pheromone						unclustered	unclustered
sex3	67432	sex-inducer sex-inducing pheromone						unclustered	unclustered
***Volvox* VMPs**									
vmp1	104262	VMP1						6575762	Cre Vca
vmp2	84562	VMP2						6575762	Cre Vca
vmp3	103843	VMP3						6575762	Cre Vca
vmp4	127191	VMP4						unclustered	unclustered
vmp5	62107	VMP5						6575762	Cre Vca
vmp6	61971	VMP6						6575762	Cre Vca
vmp7	94099	VMP7						6575762	Cre Vca
vmp8	88651	VMP8						6575762	Cre Vca
vmp9	103842	VMP9						6575762	Cre Vca
vmp10	41832	VMP10						6575762	Cre Vca
vmp11	40166	VMP11						6575762	Cre Vca
vmp12	66557	VMP12						6575762	Cre Vca
vmp13	66578	VMP13						6575762	Cre Vca
vmp14	127215	VMP14						6575762	Cre Vca
vmp15	127216	VMP15						unclustered	unclustered
vmp16	127218	VMP1						6575762	Cre Vca
vmp17	127219	VMP17						6575762	Cre Vca
vmp18	69762	VMP18						unclustered	unclustered
vmp19	69756	VMP19						6575762	Cre Vca
vmp20	42295	VMP20						6575762	Cre Vca
vmp21	41341	VMP21						6575762	Cre Vca
vmp22	60514	VMP22						6575762	Cre Vca
vmp23	60687	VMP23						6575762	Cre Vca
vmp24	41306	VMP24						6575762	Cre Vca
vmp25	60653	VMP2						6575762	Cre Vca
vmp26	82178	VMP2						unclustered	unclustered
vmp27	94088	VMP27						6575762	Cre Vca
vmp28	63421	VMP28						6575762	Cre Vca
vmp29	127220	VMP29						unclustered	unclustered
vmp30	66400	VMP30						6575762	Cre Vca
vmp31	127221	VMP31						unclustered	unclustered
vmp32	127222	VMP32						unclustered	unclustered
vmp33	127223	VMP33						unclustered	unclustered
vmp34	41348	VMP34						unclustered	unclustered
vmp35	41924	VMP3						unclustered	unclustered
vmp36	56439	VMP36						unclustered	unclustered
vmp37	127224	VMP37						6575762	Cre Vca
vmp38	127225	VMP38						6572204	Cre Vca
vmp39	82108	VMP39						6575762	Cre Vca
vmp40	127226	VMP40						unclustered	unclustered
vmp41	101266	VMP41						unclustered	unclustered
vmp42	104131	VMP42						6572204	Cre Vca
***Chlamydomonas* VMP homologs**									
			MMP4	196035			metalloproteinase, cell wall protein; metalloproteinase, cell wall protein, homolog of Volvox VMPs; metalloproteinase, cell wall protein metalloproteinase of VMP family	6575762	Cre Vca
			MMP5	196036	196036	Au9.Cre22.g763950	metalloproteinase of VMP family	6575762	Cre Vca
			MMP6	178030			metalloproteinase, cell wall protein, homolog of Volvox VMPs; metalloproteinase, cell wall protein metalloproteinase of VMP family	6572204	Cre Vca
			MMP7	194576	194576	Au9.Cre07.g353600	metalloproteinase of VMP family	unclustered	unclustered
			MMP8	177780	177780	Au9.Cre19.g752600	metalloproteinase of VMP family	unclustered	unclustered
			MMP9	151617	151617	Au9.Cre16.g652200	metalloproteinase of VMP family	6575762	Cre Vca
			MMP10	194578	194578	Au9.Cre07.g353750	metalloproteinase of VMP family	unclustered	unclustered
			MMP11	148193	148193	Au9.Cre09.g388350	metalloproteinase of VMP family	unclustered	unclustered
Cell cycle									
cdka1	127504	cyclin dependent kinase	CDKA1	127285	127285	Au9.Cre10.g465900		6571732	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
cdkb1	103386	plant specific cyclin dependent kinase	CDKB1	59842	59842	Au9.Cre08.g372550		6575306	Cme Ath Cel Ota Olu Ppa Cre Vca
cdkc1	82776	cyclin dependent kinase	CDKC1	148395	148395	Au9.Cre08.g385850		6574232	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Tps Ngr Pte Cre Vca
cdkd1	65162		CDKD1	137457	137457	Au9.Cre09.g388000		6571911	Cme Ddi Pra Pso Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Cre Vca
cdke1	68336	cyclin dependent kinase	CDKE1	120881	120881	Au9.Cre04.g213850		6570970	Ddi Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Ngr Pte Cre Vca
cdkg1	127266	Cyclin dependent kinase G1 subfamily	CDKG1	126776	126776	Au9.Cre06.g271100	cyclin dependent kinase	6572812	Cel Cre Vca
cdkg2	127318	cyclin dependent kinase.	CDKG2	139908	139908	Au9.Cre17.g742250	cyclin dependent kinase	unclustered	unclustered
cdkh1	83876	cyclin dependent kinase	CDKH1	153970	153970	Au9.Cre07.g355400	cyclin dependent kinase	6570488	Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
cdki1/lf2	119542	cyclin dependent kinase	CDKI1	195781	195781	Au9.Cre12.g494500	cyclin dependent kinase	6575423	Cre Vca
cyca1	127267	A type cyclin.	CYCA1	147453	147453	Au9.Cre03.g207900	A-type cyclin	6573447	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
cycb1	127276	B type mitotic cyclin	CYCB1	206115	206115	Au9.Cre06.g284350	B-type cyclin	6573447	Cme Ddi Tth Pra Pso Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
cycab1	127293	Related to A and B type cyclins.	CYCAB1	206112	206112			6576048	Tps Cre Vca
cycc1	127505	C type cyclin	CYCC1	206655				unclustered	unclustered
cycd1.1; cycd1.2; cycd1.3 cycd1.4	127281; 127284; 127282 127283	D type cyclin.	CYCD1	195780	195780	Au9.Cre22.g763200	D-type cyclin	6574018	Ath Cre Vca
cycd2	127277	D type cyclin	CYCD2	191762	191762	Au9.Cre06.g289750	D-type cyclin	unclustered	unclustered
cycd3	127287	D type cyclin	CYCD3	206110	206110	Au9.Cre06.g298750	D-type cyclin	6573500	Cre Vca
cycd4	127321	D type cyclin	CYCD4	206166	206166	Au9.Cre06.g259500	D-type cyclin	unclustered	unclustered
cycl1	127508	L type cyclin	CYCL1	206656				unclustered	unclustered
cycm1	107638	cyclin	CYCM1	206658				6576965	Cre Vca
cycu1	127509	cyclin	CYCU1	206659			conserved expressed protein of unknown function	unclustered	unclustered
cyct1	120142		CYCT1	193461	193461	Au9.Cre14.g613900		6573203	Ddi Tth Pra Pso Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca

wee1	127274	wee1 kinase ortholog	WEE1	194589	194589	Au9.Cre07.g355250	CDK inhibitory kinase	6575995	Cme Ddi Pra Ncr Ath Hsa Olu Ppa Mbr Tps Ngr Cre Vca
cks1	127315	CKS1 homolog	CKS1	182779	182779	Au9.Cre03.g180350		6574039	Cme Tth Pra Ncr Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
mat3	127376		MAT3	187248	187248	Au9.Cre06.g255450	retinoblastoma protein	6574768	Cme Ddi Pso Ath Hsa Ota Olu Ppa Ngr Cre Vca
e2f1	127253	E2F transcription factor family homolog.	E2F1	206364				6577131	Cme Ddi Tth Pra Pso Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
dp1	121369	putative DP transcription factor	DP1	206363				6577455	Cme Ddi Tth Pra Pso Ath Hsa Cel Ota Olu Ppa Mbr Tps Ngr Pte Cre Vca
e2fr1	127270	related to E2F and DP transcription factors	E2FR1	168563	168563	Au9.Cre13.g573000	related to E2F and DP transcription factors, Chlamydomonas specific; transcription factor E2F and DP-related	unclustered	unclustered

Table S15: Predicted numbers of TAPs

The number of proteins that were predicted in each Transcription Associated Protein (TAP) family in *Volvox* and *Chlamydomonas* are shown.

TAP	***Volvox***	***Chlamydomonas***
ABI3/VP1	1	1
Alfin-like	1	1
AP2/EREBP	21	12
ARF	0	0
Argonaute	2	3
ARID	2	3
AS2/LOB	0	0
Aux/IAA	0	0
BBR/BPC	0	0
BES1	0	0
bHLH	2	3
bHSH	0	0
BSD domain containing	3	1
bZIP	11	6
C2C2_CO-like	2	1
C2C2_Dof	1	1
C2C2_GATA	9	8
C2C2_YABBY	0	0
C2H2	8	6
C3H	23	15
CAMTA	0	0
CCAAT_Dr1	0	2
CCAAT_HAP2	0	0
CCAAT_HAP3	3	1
CCAAT_HAP5	2	2
Coactivator p15	1	1
CPP	2	1
CSD	2	1

CudA	0	0
DBP	0	0
DDT	0	0
Dicer	0	0
DUF246 domain containing	0	0
DUF296 domain containing	0	0
DUF547 domain containing	1	1
DUF632 domain containing	0	0
DUF833 domain containing	0	0
E2F/DP	3	3
EIL	0	0
FHA	11	12
GARP_G2-like	4	4
GARP_ARR-B	1	1
GeBP	0	0
GIF	1	1
GNAT	33	28
GRAS	0	0
GRF	0	0
HB	0	1
HB_KNOX	0	0
HD-Zip	0	0
HMG	9	7
HRT	0	0
HSF	2	2
IWS1	1	1
Jumonji	0	0
LFY	0	0
LIM	0	0
LUG	0	0
MADS	1	2
MBF1	1	1
MED6	0	1
MED7	1	0

mTERF	3	1
MYB-related	12	9
MYB	19	15
NAC	0	0
NZZ	0	0
OFP	0	0
PcG_EZ	0	0
PcG_FIE	1	1
PcG_VEFS	0	0
PHD	16	10
PLATZ	3	4
Pseudo ARR-B	0	2
RB	0	1
Rcd1-like	1	2
Rel	0	0
RF-X	0	0
RRN3	1	0
Runt	0	0
RWP-RK	9	14
S1Fa-like	0	0
SAP	0	0
SBP	20	21
SET	16	13
Sigma70-like	1	1
Sin3	1	1
Sir2	3	2
SOH1	1	0
SRS	0	0
SWI/SNF_BAF60b	1	2
SWI/SNF_SNF2	26	18
SWI/SNF_SWI3	0	0
TAZ	4	2
TCP	0	0
TEA	0	0

TFb2	0	1
TRAF	20	28
Trihelix	0	0
TUB	3	2
ULT	0	0
VARL	13	9
VOZ	0	0
Whirly	1	1
WRKY	2	1
zf_HD	0	0
tify	0	0
Zinc finger, AN1 and A20 type	2	1
Zinc finger, MIZ type	2	0
Zinc finger, ZPR1	1	1
Zn_clus	0	0
Total	**347**	**297**

Table S16: Summary of RepeatScout libraries

A summary of the number of repeat sequences (and their mean length) generated from running RepeatScout on the *Volvox* and *Chlamydomonas* assemblies is shown.

	Volvox	*Chlamydomonas*
Sequences in raw repeat library	1,511	1,057
Putative novel repeat sequences	122	58
Mean repeat sequence length	919	595
No. sequences left after removing unknown sequences with non-TE Pfam domains	1,449	1,013

5) SUPPLEMENTAL REFERENCES

S1. C. R. Adams *et al.*, *Curr Genet* **18**, 141 (1990).

S2. R. C. Starr, *Arch Protistenkd* **111**, 204 (1969).

S3. R. C. Starr, *Dev Biol Suppl* **4**, 59 (1970).

S4. D. R. Smith, R. W. Lee, *BMC Genomics* **10**, 132 (2009).

S5. J. L. Weber, E. W. Myers, *Genome Res* **7**, 401 (May, 1997).

S6. B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res* **8**, 175 (Mar, 1998).

S7. S. Aparicio *et al.*, *Science* **297**, 1301 (Aug 23, 2002).

S8. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers, *Nucleic Acids Res* **37**, D26 (Jan, 2009).

S9. W. J. Kent, *Genome Res* **12**, 656 (Apr, 2002).

S10. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J Mol Biol* **215**, 403 (Oct 5, 1990).

S11. T. U. Consortium, *Nucleic Acids Res* **38**, D142 (Jan, 2010).

S12. A. F. A. Smit, R. Hubley, P. Green, *http://www.repeatmasker.org*, (1996-2004).

S13. J. Jurka *et al.*, *Cytogenet Genome Res* **110**, 462 (2005).

S14. S. S. Merchant *et al.*, *Science* **318**, 245 (Oct 12, 2007).

S15. A. L. Price, N. C. Jones, P. A. Pevzner, *Bioinformatics* **21 Suppl 1**, i351 (Jun, 2005).

S16. S. R. Eddy, *Bioinformatics* **14**, 755 (1998).

S17. T. M. Lowe, S. R. Eddy, *Nucleic Acids Res* **25**, 955 (Mar 1, 1997).

S18. E. Quevillon *et al.*, *Nucleic Acids Res* **33**, W116 (Jul 1, 2005).

S19. G. A. Tuskan *et al.*, *Science* **313**, 1596 (Sep 15, 2006).

S20. A. Coghlan, E. E. Eichler, S. G. Oliver, A. H. Paterson, L. Stein, *Trends Genet* **21**, 673 (Dec, 2005).

S21. J. C. Detter *et al.*, *Genomics* **80**, 691 (Dec, 2002).

S22. B. Ewing, P. Green, *Genome Res* **8**, 186 (Mar, 1998).

S23. T. G. I. Project.

S24. X. Huang, A. Madan, *Genome Res* **9**, 868 (Sep, 1999).

S25. A. Z. Worden *et al.*, *Science* **324**, 268 (Apr 10, 2009).

S26. A. Y. Guo *et al.*, *Nucleic Acids Res* **36**, D966 (Jan, 2008).

S27. D. M. Riano-Pachon, S. Ruzicic, I. Dreyer, B. Mueller-Roeber, *BMC Bioinformatics* **8**, 42 (2007).

S28. S. Richardt, D. Lang, R. Reski, W. Frank, S. A. Rensing, *Plant Physiol* **143**, 1452 (Apr, 2007).

S29. R. D. Finn *et al.*, *Nucleic Acids Res* **36**, D281 (Jan, 2008).

S30. P. Perez-Rodriguez *et al.*, *Nucleic Acids Res* **38**, D822 (Jan, 2009).

S31. K. Katoh, K. Kuma, H. Toh, T. Miyata, *Nucleic Acids Res* **33**, 511 (2005).

S32. M. Clamp, J. Cuff, S. M. Searle, G. J. Barton, *Bioinformatics* **20**, 426 (Feb 12, 2004).

S33. A. Sanderfoot, *Plant Physiol* **144**, 6 (May, 2007).

S34. S. Rutherford, I. Moore, *Curr Opin Plant Biol* **5**, 518 (Dec, 2002).

S35. J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, D. G. Higgins, *Nucleic Acids Res* **25**, 4876 (Dec 15, 1997).

S36. R. C. Edgar, *Nucleic Acids Res* **32**, 1792 (2004).

S37. D. Swofford. (Sinauer Associates, Sunderland, MA, 2003).

S38. K. Bisova, D. M. Krylov, J. G. Umen, *Plant Physiol* **137**, 475 (Feb, 2005).

S39. M. Berriman *et al.*, *Science* **309**, 416 (Jul 15, 2005).

S40. A. Hallmann, *Int Rev Cytol* **227**, 131 (2003).

S41. E. H. Harris, Stern, D.B., and Witman, G.B., *The Chlamydomonas Sourcebook.* (Academic Press, 2009).

S42. F. Abascal, R. Zardoya, D. Posada, *Bioinformatics* **21**, 2104 (May 1, 2005).

S43. S. Guindon, O. Gascuel, *Syst Biol* **52**, 696 (Oct, 2003).

S44. S. Guindon, F. Lethiec, P. Duroux, O. Gascuel, *Nucleic Acids Res* **33**, W557 (Jul 1, 2005).

S45. E. Szathmary, J. M. Smith, *Nature* **374**, 227 (Mar 16, 1995).

S46. J. T. Bonner, *Integr Biol* **1**, 27 (1998).

S47. S. L. Baldauf, *Science* **300**, 1703 (Jun 13, 2003).

S48. R. K. Grosberg, R. Strathmann, *Annu Rev Ecol Evol Syst* **38**, 621 (2007).

S49. G. A. Wray, *Genome Biol* **3**, REVIEWS0001 (2002).

S50. H. S. Yoon, J. D. Hackett, C. Ciniglia, G. Pinto, D. Bhattacharya, *Mol Biol Evol* **21**, 809 (May, 2004).

S51. M. D. Herron, J. D. Hackett, F. O. Aylward, R. E. Michod, *Proc Natl Acad Sci U S A* **106**, 3254 (Mar 3, 2009).

S52. H. Rausch, N. Larsen, R. Schmitt, *J Mol Evol* **29**, 255 (Sep, 1989).

S53. T. Nakada, K. Misawa, H. Nozaki, *Mol Phylogenet Evol* **48**, 281 (Jul, 2008).

S54. H. Nozaki, *Biologia, Bratislava* **58**, 425 (2003).

S55. S. M. Miller, D. L. Kirk, *Development* **126**, 649 (Feb, 1999).

S56. Q. Cheng, R. Fowler, L. W. Tam, L. Edwards, S. M. Miller, *Dev Genes Evol* **213**, 328 (Jul, 2003).

S57. I. Nishii, S. Ogihara, D. L. Kirk, *Cell* **113**, 743 (Jun 13, 2003).

S58. M. M. Kirk *et al.*, *Development* **126**, 639 (Feb, 1999).

S59. M. Meissner, K. Stark, B. Cresnar, D. L. Kirk, R. Schmitt, *Curr Genet* **36**, 363 (Dec, 1999).

S60. L. Duncan *et al.*, *J Mol Evol* **65**, 1 (Jul, 2007).

S61. H. Gruber, S. D. Goetinck, D. L. Kirk, R. Schmitt, *Gene* **120**, 75 (Oct 12, 1992).

S62. A. Hallmann, A. Rappel, *Plant J* **17**, 99 (Jan, 1999).

S63. T. Jakobiak *et al.*, *Protist* **155**, 381 (Dec, 2004).

S64. A. Hallmann, S. Wodniok, *Plant Cell Rep* **25**, 582 (Jun, 2006).

S65. S. M. Miller, R. Schmitt, D. L. Kirk, *Plant Cell* **5**, 1125 (Sep, 1993).

S66. N. Ueki, I. Nishii, *Genetics* **180**, 1343 (Nov, 2008).

S67. B. Schiedlmeier *et al.*, *Proc Natl Acad Sci U S A* **91**, 5080 (May 24, 1994).

S68. A. Hallmann, M. Sumper, *Eur J Biochem* **221**, 143 (Apr 1, 1994).

S69. D. L. Kirk, *Bioessays* **27**, 299 (Mar, 2005).
S70. P. Ferris *et al.*, *Science* **328**, 351 (Apr 16, 2010).
S71. A. Hallmann, K. Godl, S. Wenzl, M. Sumper, *Trends Microbiol* **6**, 185 (May, 1998).
S72. A. Hallmann, *Planta* **226**, 719 (Aug, 2007).
S73. N. Aono, T. Inoue, H. Shiraishi, *J Biochem* **138**, 375 (Oct, 2005).
S74. A. Hallmann, *Plant J* **45**, 292 (Jan, 2006).
S75. D. Kirk, *Volvox: Molecular-genetic origins of multicellularity and cellular differentiation.* (Cambridge University Press, Cambridge, 1998).
S76. D. Stern, G. Witman, E. Harris, (Jan 1, 2009).
S77. T. Kubo, J. Abe, T. Saito, Y. Matsuda, *Curr Genet* **41**, 115 (May, 2002).
S78. T. Hamaji *et al.*, *Genetics* **178**, 283 (Jan, 2008).
S79. H. Nozaki, T. Mori, O. Misumi, S. Matsunaga, T. Kuroiwa, *Curr Biol* **16**, R1018 (Dec 19, 2006).
S80. D. L. Kirk, R. Birchem, N. King, *J Cell Sci* **80**, 207 (Feb, 1986).
S81. R. R. Reisz, J. Muller, *Trends Genet* **20**, 237 (May, 2004).
S82. M. Goodman, G. W. Moore, G. Matsuda, *Nature* **253**, 603 (Feb 20, 1975).
S83. H. Wang *et al.*, *Proc Natl Acad Sci U S A* **106**, 3853 (Mar 10, 2009).
S84. C. Bowler *et al.*, *Nature* **456**, 239 (Nov 13, 2008).