## Abstract

Somatic cellular differentiation plays a critical role in the transition from unicellular to multicellular life, but the evolution of its genetic basis remains poorly understood. By definition somatic cells do not reproduce to pass on genes and so constitute an extreme form of altruistic behavior. The volvocine green algae provide an excellent model system to study the evolution of multicellularity and somatic differentiation. In *Volvox carteri*, somatic cell differentiation is controlled by the *regA* gene, which is part of a tandem duplication of genes known as the *reg* cluster. While previous work found the *reg* cluster in divergent *Volvox* species, its origin and distribution in the broader group of volvocine algae has not been known. Here we show that the *reg* cluster is present in many species without somatic cells, indicating the genetic basis for soma arose before the phenotype. We hypothesize the ancestral function was involved in regulating reproduction in response to stress and that this function was later co-opted to produce soma. Determining that the *reg* cluster was co-opted to control somatic cell development provides insight into how cellular differentiation, and with it greater levels of complexity and individuality, evolves.


## Introduction

The evolutionary transition from unicellular to multicellular life involves an increase in organismal complexity and a shift in individuality from the level of the cell to the level of the multicellular organism. A key step in this transition is the evolution of cellular differentiation, in particular the division of labor between non-reproductive somatic cells and reproductive germ cells (Buss, 1987; Michod, 1999; Grosberg & Strathmann, 2007; Simpson, 2012). The evolution of altruistic somatic cells is a major step in transferring the level of selection, as well as

individuality, from the level of the cell to the level of the multicellular organism (Queller, 2000; Michod, 2005; Folse III & Roughgarden, 2013). This transition also represents an increase in complexity as measured by the hierarchical nestedness of the organism (from single-celled to multicellular) and the number of cell types present (Maynard Smith & Szathmáry, 1995; Bell & Mooers, 1997; Marcot & Mcshea, 2007; Niklas *et al*., 2014). In organisms with somatic differentiation, somatic cells give up reproductive capacity to specialize on maintenance and survival-related functions of the multicellular organism, whereas germ cells specialize on reproduction and have reduced viability at the cell level (were they to exist outside the context of the group). Consequently, germ and somatic cells must work together as a team to ensure high fitness for the multicellular organism (Michod, 2005).

The volvocine green algae, along with their unicellular relative *Chlamydomonas reinhardtii*, span a gradient in complexity from single-celled organisms, to undifferentiated multicellular groups, to species with thousands of cells and germ-soma differentiation. They are photosynthetic, facultatively sexual, haploid eukaryotes in the chlorophycean order Volvocales which includes three families: the Tetrabaenaceae, the Goniaceae, and the Volvocaceae (Figure 1). The Tetrabaenaceae contain two genera, *Tetrabaena* and *Basichlamys*, which are made up of four undifferentiated cells held together by extracellular matrix. The Goniaceae contain the genera *Astrephomene*, 32 – 64 celled spheroidal colonies with 2 – 4 sterile somatic cells in the posterior pole of the colony; and *Gonium*, colonies shaped like a slightly curved plate made up of 8 – 32 undifferentiated cells. The Volvocaceae are a diverse family containing genera with undifferentiated 8 – 32 celled colonies (such as *Pandorina*, *Volvulina*, *Platydorina*, *Yamagishiella*, and *Eudorina*), 64 – 128 celled *Pleodorina* with specialized somatic cells in the anterior portion of the colony, and *Volvox* with thousands of cells arranged in a sphere and

47 complete germ-soma cellular differentiation (Figure 1). Many of these genera were originally
48 described morphologically, but molecular phylogenies have revealed many to be polyphyletic,
49 including *Volvox* and *Eudorina* (Figure 1) (Nozaki *et al.*, 2002; Herron & Michod, 2008;
50 Coleman, 2012).

51 Multicellularity arose relatively recently in the volvocine green algae (~240 million
52 years ago) compared to embryophytes (~748 – 872 million years ago) and metazoa (~574 – 852
53 million years ago) (Herron *et al.*, 2009; Sharpe *et al.*, 2014). In addition, ancestral character state
54 reconstructions based on molecular phylogenies have predicted that numerous multicellular
55 traits, including the evolution of somatic cells, have been gained multiple times in this group
56 (Figure 5) (Herron and Michod 2008). Thus, these algae provide a uniquely detailed and recent
57 timeline of the stages involved in the evolutionary transition from unicellular to multicellular
58 individuals (Kirk, 2005; Herron & Michod, 2008). In addition, multiple genomes (Merchant *et*
59 *al.*, 2007; Prochnik *et al.*, 2010; Hanschen *et al.*, 2016) and available molecular techniques (Kirk,
60 1998; Umen & Olson, 2012), make this group particularly well suited for studying the evolution
61 of the genetic basis for cellular differentiation and individuality (Kirk, 2005; Michod, 2007;
62 Coleman, 2012).

63 Mutants of the differentiated multicellular alga *Volvox carteri* lacking terminally
64 differentiated somatic cells were first described by Starr (1970). These mutants are known as
65 "regenerator" mutants because colonies first develop seemingly normally, but later the somatic
66 cells develop into reproductive germ cells. The locus found to be responsible for the regenerator
67 phenotype, *regA,* encodes a putative transcription factor (Huskey & Griffin, 1979; Kirk *et al.*,
68 1999). *regA* is part of a tandem duplication of several paralogous genes known as the *reg* cluster
69 (Figure 2) (Duncan *et al.*, 2007; Hanschen *et al.*, 2014). All *reg* cluster genes encode a DNA-

binding SAND domain, also known as the VARL (volvocine algae *regA*-like) domain. During development in *V. carteri* the *regA* gene is turned on in soma-progenitor cells where it is thought to down regulate chloroplast biogenesis. This is thought to prevent somatic cells from being able to grow large enough for cell division, thus keeping them in a non-reproductive somatic state (Kirk, 2001). Preliminary work has shown that two other *reg* cluster genes, *rlsB* and *rlsC*, are co-expressed with *regA* during development suggesting that the other members of the *reg* cluster also play a role in somatic cell differentiation (Harryman, 2012).

The *reg* cluster is absent from the genomes of *Chlamydomonas reinhardtii* (Duncan *et al.*, 2007; Merchant *et al.*, 2007) and *Gonium pectorale* (Hanschen *et al.*, 2016), but present in divergent *Volvox* species including *V. carteri*, *V. ferrisii*, and *V. gigas* suggesting that the *reg* cluster evolved at the origin of the Volvocaceae (Figures 1 and 3) (Hanschen *et al.*, 2014). The most closely related homolog to the *reg* cluster found in *C. reinhardtii* and *G. pectorale* is *RLS1*, which is orthologous to *rlsD* from *V. carteri* and *V. ferrisii* (Duncan *et al.*, 2007; Hanschen *et al.*, 2014). It is thought that the *reg* cluster arose via the duplication of the *RLS1* gene after the Goniaceae diverged from the rest of the Volvocaceae (Hanschen *et al*., 2014; Hanschen *et al*., 2016). However, the possibility that the *V. carteri* and *V. ferrisii* lineages gained the *reg* cluster through independent duplication events could not be rejected (Hanschen *et al.*, 2014). The single duplication hypothesis predicts that the *reg* cluster is ancestral to all descendants of the last common ancestor of *V. carteri* and *V. ferrisii* including genera without somatic cells such as *Pandorina*, *Eudorina*, *Yamagishiella*, and *Platydorina.* Alternatively, independent evolutions of the *reg* cluster may correlate with the independent evolutions of somatic cells in *V. ferrisii* and the *V. carteri/V. gigas* group. If so, the *reg* cluster would not be found in undifferentiated genera such as *Pandorina* and *Yamagishiella* (Figure 1).

In order to understand the origin and evolution of *regA*, a key component of the genetic basis for somatic differentiation in *V. carteri*, we investigate for the first time the presence of the *reg* cluster in non-*Volvox* volvocine algae species representing multiple clades and levels of morphological complexity. We report the discovery of the *reg* cluster in *Pandorina morum*, *Platydorina caudata*, *Yamagishiella unicocca*, *Eudorina elegans* UTEX 1212, and *Pleodorina californica* (Figure 1). The phylogenetic position and morphology of these species indicate that the *reg* cluster arose in an undifferentiated ancestor early in the diversification of the volvocine lineage. Thus, there is a disparity between the known function of the *reg* cluster (*i.e.*, somatic differentiation in *V. carteri*) and the phenotype observed in species with the *reg* cluster but lacking somatic cells, suggesting the cluster originally served a different function and was later co-opted to produce somatic cells.

## Methods

### *Cultures and Cosmid Construction*

We grew *Pan. morum* (UTEX 1727), *Pla. caudata* (UTEX 1658)*, E. elegans* UTEX 1212 (NEIS 721), and a wild isolate of *Ple*. *californica* (see Supplemental methods for identification details) in standard *Volvox* medium (SVM) at 25℃ with a 16:8 hour light:dark cycle (~35 µmol photons/$m^2$/s). Genomic DNA was extracted from cultures using a previously described protocol (Miller & Kirk, 1999). Degenerate PCR was then performed as described in Hanschen *et al.* (2014), using two forward (regF2 and regF3) and two reverse primers (regR and DregR) designed to amplify the VARL domains of *regA* and related *reg* cluster genes (Supplemental Table 1). All reactions were performed using 2X Phusion HF Master Mix (Thermo Scientific, Waltham, MA) with 3% DMSO and 25 ng of template DNA. Cycling

conditions used for reactions were as follows: an initial melting step of 98˚C for 3 minutes; followed by 35 cycles of 98˚C for 10 seconds, 63-68˚C for 30 seconds, and 72˚C for 60 seconds; then a final 5 minute extension step at 72˚C. For sequences that required extension beyond the degenerate PCR amplified region TAIL-PCR was used based on methods described in Dent *et al.* (2005) using 2X Taq Master Mix (New England BioLabs). Cosmid library construction was performed as in Hanschen *et al.* (2014) using Epicentre's pWEB-TNC Cosmid Cloning Kit. Libraries were screened as described in Hanschen *et al.* (2014) using probes designed from sequences obtained through degenerate PCR (Supplemental Table 1). Cosmids found to contain genes of interest were then isolated for further analysis. See Supplemental Methods for detailed description of cosmid isolation in each species.

***Sequencing and Assembly***

DNA sequencing of PCR products, plasmids, and cosmid segments was performed by the University of Arizona Genetics Core service using Applied Biosystems 3730 DNA Analyzers (Waltham, MA). All PCR product sequences were derived from at least two reactions pooled together. Preparation of plasmid and cosmid DNA was performed using the Qiagen Plasmid Midi Kit or Qiagen Spin Miniprep Kit according to manufacturer's instructions. DNA sequences were assembled into contigs using the CLC Main Workbench or Geneious (version 6.1.7) software.

Full cosmids were sequenced using Illumina® sequencing. Illumina libraries were prepared by shearing cosmids with a Covaris® S220 using settings optimized for cosmids then generating libraries with the Illumina TruSeq® HT kit. All size selection steps during library preparation were for an average insert size of ~550 base pairs. Each cosmid library was

individually barcoded and paired-end sequenced with a MiSeq instrument using version 3 reagents for 600 cycles. Library sequences were de-multiplexed on the BaseSpace platform, resultant fastq files were then quality trimmed with Sickle, and assembled with Abyss 1.5.2 (Simps*on et al*., 2009) using a sequential *k*-mer optimization. Following assembly, gaps in cosmid sequences were manually edited to close gaps in Geneious (version 6.1.7). After masking the cosmid vector, genes were annotated with Augustus version 2.7 (Stanke *et al.*, 2008) with settings optimized for *C. reinhardtii*, followed by manual curation.

We used the *regA* VARL domain from *V. carteri f. nagariensis* to search the preliminary genome assembly of a *Y. unicocca plus* strain, and found that scaffold 42 contains the *reg* cluster. This region also contained several gaps. Forward and reverse PCR primers were designed flanking the gaps and used for PCR with genomic DNA, followed by sequencing the PCR product of successful reactions. Three gaps could be filled: one in the intergenic region between *ackB* and *rlsD*, and one near the C-terminal end of both *rlsB* and *rlsO*. Sequencing of genomic PCR products was used to confirm the nucleotide sequence of all *reg* cluster VARL domains. Due to the extensive sequence homology present in the non-VARL portions of *regA*, *rlsB*, and *rlsO*, these sequences were also confirmed by genomic PCR.

***Annotation***

Gene predictions for all reported genomic sequences, including *Y. unicocca*, were created using Augustus version 3.1 (Stanke & Morgenstern, 2005). Augustus was selected as its algorithm has been tuned to predict high GC content genomes such as *C. reinhardtii* and *V. carteri*. Annotations were created using partial gene models based on *C. reinhardtii*. Based on previous volvocine genome annotations (Hanschen *et al*., 2014; Hanschen *et al.*, 2016), the UTR

option was turned off. Models were manually modified to increase interspecific homology when possible. Gene identification was determined by phylogenetic relationship (Figure 3), syntenic position (Figure 2), characteristic protein homologies (Figure 4), and intron positions. Supplemental Table 3 summarizes support for all VARL gene annotations and additional annotation details can be found in the supplemental material.

The presence of SAND domains was predicted using both Pfam (version 28, (Finn *et al*., 2014)) and SMART (version 7, (Letunic *et al.*, 2012)) databases. Annotations for both Pfam and SMART were obtained using direct submission via Perl script. E-values of SAND domains (PF01342, SM00258) are reported in Supplemental Table 2. We also searched the conserved motifs highlighted in Figure 4 against the PROSITE motif database but found no significant hits.

***Phylogenetic and Protein Similarity Analyses***

VARL domain protein sequences were defined following the N-terminal extension and core VARL domain structure of Duncan *et al.* (2007). Sequences were aligned with MAFFT (version 7.213) using the accurate L-INS-i option (Katoh *et al*., 2005). Maximum likelihood gene phylogenies were generated using RAxML (version 8.1.12) with a rapid bootstrapping analysis with an automatically selected 1,000 bootstrap replicates (Stamatakis, 2014). The LG+G substitution model was automatically determined as the best substitution model by RAxML using the corrected Akaike information criterion. Stabilizing selection was tested for by calculating dN/dS ratios using codeml in PAML (Yang, 2007).

***Species tree construction***

We generated a species tree (Figure 1, Figure 5) using Bayesian Markov chain Monte

Carlo (MCMC) implemented in MrBayes version 3.2.2 (Ronquist *et al.*, 2012) using default parameters except as described below. Sequences for 80 volvocine operational taxonomic units (OTUs) and six non-volvocine green algae, consisting of five chloroplast genes (ATP synthase beta-subunit, *atpB*; P700 chlorophyll a-apoprotein A1, *psaA*; P700 chlorophyll a-apoprotein A2, *psaB*; photosystem II CP43 apoprotein, *psbC*; and the large subunit of Rubisco, *rbcL*; Supplemental Table 6) were included (excess OTUs were trimmed for Figure 1 and outgroup taxa were trimmed for Figure 5) and a codon partitioning scheme was used following Herron and Michod (2008). Four independent Bayesian analyses of four chains (three heated chains and one cold chain) were run for $2.5\text{x}10^{7}$ generations with a burn-in of $5\text{x}10^{6}$ generations. Trees were sampled every 100 generations and assembled to construct a majority rule consensus phylogram. Posterior probabilities for nodes were calculated using all post burn-in trees. We considered the run to have adequately sampled the solution space as the standard deviation of split frequencies was below $5\text{x}10^{-3}$. Statistical support for this species tree was further estimated using maximum likelihood bootstrapping. jModelTest version 2.1.10 (Darriba *et al.*, 2012) was used to select GTR+Γ+I as the best-fit model (ΔAICc=135.26). RAxML version 7.2.8 (Stamatakis, 2006) with a GTR+Γ+I model was used to draw 200 topologies onto the Bayesian consensus tree. The resulting species tree (Figure 1, Figure 5) is consistent with recently published phylogenetic analyses (Herron & Michod, 2008; Nozaki *et al.*, 2014). Specifically, there are no statistically supported differences between our topology and other published tree topologies (Herron & Michod, 2008).

***Ancestral character state reconstruction***

The presence of obligate somatic cells in each species and strain was compiled from published reports (Supplemental Table 5). Ancestral character states were reconstructed using maximum likelihood and Bayesian methods.

For the maximum likelihood analysis, the consensus species tree was analyzed in R (R Core Team, 2013) using the *diversitree* package (Fitzjohn, 2012). An ultrametric tree was calculated using a penalized likelihood function in the *ape* package (Paradis *et al.*, 2004). Two evolutionary models were evaluated, one constraining the rate of gain and loss of somatic cells to be equal and one with these two parameters unequal. Comparisons of these models using AICc (Akaike, 1974) with a small sample size correction (Burnham & Anderson, 2002), revealed the constrained, equal transition rates model as the best compromise between parameter number and model fit (AICc-equal=49.32, AICc-unequal=50.35, $\Delta$AICc=1.03). Given this low value ($\Delta$AICc $< 2$), the equal rates model is presented in Figure 5 and the unequal rates model is presented in Supplemental Figure 13. All nodes in the phylogeny were set to the state (presence of somatic cells or not) with the highest probability. A node was determined to be significantly supported if it was at least 7.39 times (if the natural logarithm of the ratio of two likelihoods is greater than 2) more likely than the alternative state (Pagel, 1999).

For the Bayesian analysis, phylogenetic uncertainty was explicitly taken into account by analyzing a sample of trees from the MCMC runs. The systematic subsample included every 1,000$^{th}$ tree from the four independent MrBayes runs (Ronquist *et al.*, 2012) for a total of 800 trees. Unicellular outgroup taxa were not pruned from these trees for the Bayesian ancestral character state reconstructions. As with the maximum likelihood analysis, both equal and unequal transition rate models were analyzed. A Bayes Factor (BF) was estimated based on twice the difference between the highest harmonic mean log likelihood from five independent

MCMC runs for each model. As with the maximum likelihood analysis, the equal rates model is weakly favored (Bayes Factor of -0.3) (Kass & Raftery, 1995). For specific nodes of interest, the presence of somatic cells was tested using hypotheses tests, estimating a Bayes Factor from three independent MCMC runs in which the node in question was constrained to one state or the other. Uniform priors, maximum likelihood priors, and gamma-distributed hyperpriors seeded from a uniform distribution were used. All MCMC runs included 5,500,000 generations with a burn-in period of 500,000 generations.

**Results**

We used degenerate PCR followed by cosmid cloning and sequencing to determine the sequence of the *reg* cluster from undifferentiated *Pan. morum*, *Pla. caudata*, and *E. elegans* UTEX 1212, as well as soma-differentiated *Ple. californica*. We also obtained sequences of the *reg* cluster and other VARL genes from a preliminary genome assembly of the undifferentiated *Y. unicocca*. Inferring the presence of the *reg* cluster in these species is based on several results. First, the syntenic structure of the *reg* cluster and nearby markers is conserved among all species studied here except for a translocation of *rlsD* and *ackB* in *V. carteri* (Duncan *et al*., 2007; Hanschen *et al*., 2014) and an inversion in *Pla. caudata* (Figure 2). Second, all *reg* cluster VARL domains fall into a single phylogenetic clade, though most relationships show little statistical support (Figure 3). Third, we identified several conserved amino acid motifs outside of the VARL domain that are shared between genes in the same syntenic positions (Figure 4). The most notable of these motifs ("*Pandorina's* Box") is always found in the first *reg* cluster gene (*rlsA*) and is relatively long (~40 amino acids) and highly conserved (Figure 4). Finally, all *reg*

cluster genes contain the same intron within the VARL domain classified as intron 4 by Duncan *et al.* (2007) (Supplemental Table 3).

***Phylogenetic relationships of VARL genes***

A maximum likelihood phylogeny (Figure 3) was constructed for all known VARL domains (Duncan *et al*., 2007; Merchant *et al*., 2007; Prochnik *et al*., 2010; Hanschen *et al*., 2014, Hanschen *et al.*, 2016) using the core VARL and N-terminal extension boundaries described by Duncan *et al.* (2007). Due to the short sequence length of the VARL domain (86 amino acids) most nodes have very low bootstrap support, although several features are worth noting.

First, all *reg* cluster genes fall within a single clade with *RLS1*/*rlsD* as an outgroup, consistent with the hypothesis that the *reg* cluster arose through the duplication of *RLS1*/*rlsD* (Hanschen *et al.*, 2014; Hanschen *et al.*, 2016). Statistical support is especially low among relationships within the *reg* cluster clade compared to relationships between non-*reg* cluster VARL domains. This may be due to the presence of many more domains within the *reg* cluster clade, which can lower bootstrap support by oversampling (Sanderson & Wojciechowski, 2000).

Second, several non-*reg* cluster VARL domains form clades with highly supported nodes such as the *rlsE*, *rlsF*, *rlsG*, *rlsH*, and *rlsL* clades (Figure 3). This is consistent with previous analyses comparing the VARL gene content of *C. reinhardtii* and *V. carteri* but with an additional clade; *rlsF,* which has no ortholog from *C. reinhardtii.* Also of note is that the *rlsG* and *rlsE* clades lack a domain from *G. pectorale*.

Third, a gene modeled on scaffold 300 of the *Y. unicocca* draft genome (*sc300.g10*) falls within the *reg* cluster clade, sister to the second VARL domain in the *rlsN* gene of *V. ferrisii*,

275 though with poor bootstrap support (Figure 3). This gene contains a VARL domain with a single
276 intron at position 4 as seen in *reg* cluster genes and other VARL genes including *rlsE*, *rlsF*, *rlsG*,
277 *rlsI*, and *rlsM*; but, it does not lie in the *reg* cluster itself suggesting a spurious relationship to the
278 *reg* cluster.

279 We also found a VARL domain lacking an N-terminal extension on scaffold 343 of the *Y.*
280 *unicocca* draft genome. The VARL domain contains an intron at splice site 4 (Duncan *et al*.,
281 2007) but appears to have a frameshift in its C-terminal end making its last 14 residues appear in
282 a different frame (data not shown). To ensure that this odd configuration was not due to
283 assembly error we confirmed the sequence using genomic PCR and sequencing of the product.
284 Given the frameshift in the VARL domain we suspect that this is a pseudogene and therefore
285 have omitted it from Figure 3.

286

287 ***Non-VARL domain homology***

288 Previous analyses of the VARL gene family have found that, except for *rlsN* in *V. ferrisii*,
289 VARL genes are characterized as having a single SAND domain with generally low protein
290 homology outside of this region (Duncan *et al*., 2007; Hanschen *et al*., 2014). Our Pfam and
291 SMART gene annotations support the presence of only a single SAND domain in the genes
292 reported here (Supplemental Table 2). However, we identified several smaller regions of
293 homology in the predicted protein sequences of *reg* cluster genes and *RLS1*/*rlsD* that were
294 previously unknown or underappreciated. Almost all *reg* cluster VARL domains are followed by
295 an acidic 5 amino acid sequence (labeled DSGDE in Supplemental Figures 1-5) which is lacking
296 in *RLS1*/*rlsD* (Duncan *et al.*, 2007). One particularly notable region is found upstream of the
297 VARL domain in *rlsA*, which we've named "*Pandorina*'*s* Box" (Figure 4, Supplemental Figure

1). Other homologous regions of note include a conserved "PRL" motif downstream of the VARL domain in *rlsA*; two conserved motifs found in *regA*, *rlsB*, and *rlsO* (the "LALRP" motif found upstream of the VARL domain, and the "FLQ" motif found downstream of the VARL domain); and an "EQ" motif downstream of the VARL domain in *RLS1*/*rlsD* (Figure 4, Supplemental Figures 1-6). We searched the Pfam, SMART and PROSITE databases for matches to these conserved regions but no significant matches were found. The three central genes (*regA*, *rlsB*, and *rlsO*) of the *reg* cluster in *Y. unicocca* have a higher degree of protein similarity to each other than to their putative counterparts in other species (Supplemental Figure 7). We are unable to computationally predict functions for these conserved regions.

Intriguingly, the "FLQ" motif we identified lies within a ~1,200 base-pair region known to contain a *cis*-regulatory repressor in the *regA* gene of *V. carteri* (Stark *et al*., 2001). It is possible then that the conservation of the "FLQ" motif is due to selection to maintain a shared *cis*-regulatory nucleotide sequence rather than selection operating on protein function. However, inspection of nucleotide alignments of the "FLQ" region show saturation of synonymous site mutations (average dS = 1.29, 1.35, and 0.95 for *regA*, *rlsB*, and *rlsO*, respectively) suggesting that purifying selection is acting at the amino acid level rather than the nucleotide level for this region.

***Synteny***

We analyzed the synteny of the *reg* cluster from *Pan. morum*, *Ple. californica*, *Pla. caudata*, *Y. unicocca*, and *E. elegans* UTEX 1212 (Figure 2, Supplemental Figure 8). In *V. carteri* the *reg* cluster is bordered by gene model *91984* upstream of *rlsA*, a relationship we observe in *Ple. californica*, *Y. unicocca*, and *Pan. morum* (Figure 2, Supplemental Figure 8).

Likewise, the *reg* cluster is bordered by *ackB* followed by *rlsD* in *Pan*. *morum*, *Y*. *unicocca*, and *V*. *ferrisii* (Figure 2) (Hanschen *et al*., 2014). The close syntenic linkage between *ackB* and *rlsD* is also observed in *V*. *carteri* and *Pla*. *caudata* though in *V*. *carteri* it appears that *rlsD* and *ackB* have been translocated away from the *reg* cluster (Figure 2) (Duncan *et al.*, 2007; Hanschen *et al.*, 2014). The *reg* cluster of *Pla*. *caudata* has undergone an inversion, as its *reg* cluster is bordered by *91984* and *ackB,* as observed in *Y*. *unicocca* and *Pan*. *morum*, but in the reverse order (Figure 2, Supplemental Figure 8). We were unable to clone syntenic markers for *E*. *elegans* UTEX 1212, but the relative order of *reg* cluster genes is the same as in *Ple*. *californica*, *V*. *gigas*, and *V*. *carteri*.

***reg* cluster nomenclature**

The namesake of the *reg* cluster is the *regA* gene which was named for its role in the regenerator mutant phenotype of *V*. *carteri* that results in spheroids whose somatic cells regenerate into gonidial cells (Huskey & Griffin, 1979). *Pandorina, Platydorina, Yamagishiella,* and *Eudorina* do not have somatic cells, however, and thus are incapable of becoming regenerator mutants. Thus, the *regA* gene terminology is somewhat misleading with respect to the effect of a knockout mutation in undifferentiated species. Nevertheless, we choose to use the *Volvox* nomenclature for species with and without soma due to the absence of the *reg* cluster in *C*. *reinhardtii* and *G*. *pectorale* and to keep the nomenclature consistent with previous work (Duncan *et al*., 2007; Hanschen *et al*., 2014), with one modification. The *reg* clusters of *Pan*. *morum, Y*. *unicocca* and *Pla*. *caudata* all contain five VARL domain genes. Hanschen *et al.* (2014) found that *V*. *ferrisii* also has a five gene *reg* cluster but the fourth gene in the *reg* cluster, *rlsN*, is unique due to having two VARL domains. The fourth *reg* cluster genes from *Pan*.

344 *morum, Pla. caudata,* and *Y. unicocca* only have a single VARL domain and possess the LALRP
345 and FLQ motifs which *rlsN* lacks, suggesting that they are not orthologous to *rlsN* in *V. ferrisii*.
346 For these reasons, we decided to name the fourth gene of the *reg* cluster in *Pan. morum*, *Pla.*
347 *caudata,* and *Y. unicocca* as *rlsO*. Naming conventions and justifications used for all VARL
348 genes described in the work can be found in Supplemental Table 3.
349
350 ***Purifying selection suggests function***
351 We calculated dN, dS, and dN/dS for interspecific comparisons of *rlsA*, *regA*, *rlsB*, *rlsO*,
352 *rlsC*, and *RLS1*/*rlsD* VARL domains from *C. reinhardtii*, *G. pectorale*, *Pan. morum*, *Pla.*
353 *caudata*, *Ple. californica*, *Y. unicocca*, *E. elegans* UTEX 1212, *V. ferrisii*, *V. gigas*, and *V.*
354 *carteri*. Average dN/dS ratios ranged between 0.16-0.27 for each *reg* cluster gene and *RLS1*/*rlsD*
355 (Supplemental Figure 9). While relatively high for between species comparisons, these dN/dS
356 values are reflective of the long divergence times between the species compared (Herron *et al*.,
357 2009) and are consistent with the genome-wide dN/dS values found between *C. reinhardtii*, *G.*
358 *pectorale*, and *V. carteri* (Hanschen *et al.*, 2016). Thus, our results suggest purifying selection is
359 operating on the VARL domain implying functional conservation. Given their high level of
360 protein similarity, we also calculated dN, dS, and dN/dS values for pairwise comparisons of
361 *regA*, *rlsB*, and *rlsO* coding regions in *Y. unicocca* and found dS values of 0.43-0.74 but dN
362 values of 0.17-0.31, resulting in dN/dS values of 0.40-0.46 (Supplemental Table 4).
363
364 **Ancestral character state reconstruction**
365 We constructed a Bayesian species tree using methods from Herron & Michod (2008)
366 and inferred ancestral character states using maximum likelihood and Bayesian methods (Figure

5). Both methods provide novel statistical support for inferring that the common ancestor of all species with a *reg* cluster, the ancestor of the *Volvocaceae*, was undifferentiated (Figure 5). Furthermore, both methods statistically significantly infer that the presence of somatic cells in *Volvox* section *Volvox* (i.e., *V. ferrisii*) and the group including *Pleodorina* and the remaining *Volvox* (i.e., *V. carteri*, *V. gigas*, and *Ple. californica*) represent independent evolutions. In addition, there is a statistically supported third independent evolution of somatic cells in *Astrephomene* (Figure 5).

This analysis was repeated with two models of character evolution, equal (Figure 5) and unequal transition rates (Supplemental Figure 13). The equal transition rate model was found to be marginally statistically preferred ($\Delta$AICc = 1.03). Both models of evolution support the results above, but differ in their inference regarding the ancestral state of the lineage of *E. elegans* UTEX 1212. Under the marginally-unfavored unequal rates model, *E. elegans* UTEX 1212 has a soma differentiated ancestor (Supplemental Figure 13), but under the equal rates model the ancestor of *E. elegans* UTEX 1212 lacked cellular differentiation (Figure 5).

## Discussion

We show using phylogenetic, syntenic, and sequence analysis that the *reg* cluster, a key component of the genetic basis for soma in *V. carteri*, is present in many species without somatic cells. All *reg* cluster VARL domains form a single phylogenetic clade (Figure 3), are found in the same relative order in the genome (Figure 2), and most *reg* cluster proteins contain regions of conservation outside of the VARL domain across species (Figure 4). These results and the phylogenetic position of the species examined (Figure 1) indicate that the *reg* cluster arose in the

common ancestor of the Volvocaceae following the groups divergence from the Goniaceae. In addition, we performed ancestral character state reconstruction analyses and inferred that the Volvocaceae ancestor lacked terminally differentiated somatic cells, and that soma has evolved at least twice independently in the Volvocaceae (Figure 5, Supplemental Figure 13). Thus, we conclude that the *reg* cluster evolved in an undifferentiated organism and was later co-opted to control somatic differentiation in *V. carteri*.

***Origin of the reg cluster***

Previous work has shown that the *reg* cluster, a key component of the genetic basis for somatic cell differentiation in *V. carteri*, arose through the duplication of the *RLS1*/*rlsD* gene, and is present in divergent *Volvox* species that evolved soma independently (Hansch*en et al*., 2014; Hanschen *et al*., 2016). However, it was uncertain if the *reg* cluster arose once, or twice coinciding with the two independent evolutions of soma in the Volvocaceae (Figure 1) (Hanschen *et al.*, 2014). The presence of the *reg* cluster in *Pan*. *morum*, *Pla*. *caudata*, and *Y. unicocca*, but absence in *G*. *pectorale* (Hanschen *et al.*, 2016), unambiguously demonstrates the *reg* cluster arose at the origin of the Volvocaceae. Thus, we show that the *reg* cluster evolved once, at the origin of the Volvocaceae following this group's separation from the Goniaceae.

***VARL gene evolution in the volvocine green algae***

The *reg* clusters of *Y. unicocca* and *V*. *ferrisii* indicate that *reg* cluster genes may not be strictly orthologous between species. The fourth gene in *V. ferrisii*'*s reg* cluster, *rlsN*, has a unique two domain structure which may have arisen via domain duplication (Li, 1997). This gene structure was not observed in any of the other species examined in this study, although

Hanschen *et al.* (2014) found that *Volvox rousseletii*, a close relative of *V. ferrisii*, also has an *rlsN* gene with two domains suggesting that this feature is unique to the *V. ferrisii* branch of *Volvox* species, often referred to as the "Euvolvox (section *Volvox*)" (Coleman, 2012). Furthermore, the *reg* cluster of *Y. unicocca* is also unique because the predicted protein sequences of *regA*, *rlsB*, and *rlsO* have high similarity (Supplemental Figure 7). This could be due to mechanisms of concerted evolution, such as gene conversion, unequal crossing over, or purifying selection acting on a gene birth-death process (Nei & Rooney, 2005). Pairwise nucleotide alignments of *regA*, *rlsB*, and *rlsO* in *Y. unicocca* reveal high nucleotide similarity in some coding regions (Supplemental Figures 10-12) but pairwise dS values averaged across the entirety of *Y. unicocca regA*, *rlsB*, and *rlsO* reading frames are much higher than corresponding dN values (Supplemental Table 4) indicating that gene-birth death with purifying selection is the most likely explanation for their high protein similarity. Thus, our results indicate that *reg* cluster evolution may be more complicated than previously thought with genes evolving via a birth-death process and at least one instance of domain duplication.

We also investigated the evolution of non-*reg* cluster VARL genes. Previous analyses of VARL genes in the volvocine green algae were limited to the genomes of *C. reinhardtii* and *V. carteri*, limiting their inferential power (Duncan *et al*., 2007). In this analysis, we included VARL genes from two additional genomes, *G. pectorale* (Hanschen *et al.*, 2016) and *Y. unicocca*. We found that *RLS10/rlsL* and *RLS2/rlsI* form distinct clades containing orthologs from all four genomes (Figure 3, Supplemental Table 3). In contrast, the *RLS11/rlsG* and *RLS8/rlsE* clades lack orthologs from *G. pectorale* which suggests lineage specific gene loss. Similarly, there is no *rlsF* homolog from *C. reinhardtii* suggesting this gene arose following the divergence of the multicellular volvocine green algae from *C. reinhardtii* or gene loss along the

435 *C. reinhardtii* lineage. Finally, the clade formed by *RLS6* from *C. reinhardtii* and *rlsK* from *V.*
436 *carteri* shows very similar gene content between *V. carteri* and *Y. unicocca* but also many genes
437 unique to *C. reinhardtii* and *G. pectorale* (Figure 3). Taken together, these results suggest little
438 change in non-*reg* cluster VARL gene content within the Volvocaceae but lineage specific
439 evolution in more distantly related groups.
440
441 ***The reg cluster arose in an undifferentiated organism***
442 We performed ancestral character state reconstructions to infer when the *reg* cluster arose
443 relative to the evolution of somatic cells in the volvocine green algae. Ancestral character states
444 in the volvocine algae have been previously inferred (Herron & Michod, 2008); however, with
445 the species available at that time, statistically significant inferences regarding the ancient
446 evolution of soma in the volvocine algae were not possible. Fortunately, numerous species have
447 been discovered and described since this work (Nozaki *et al.*, 2006, 2014; Nozaki & Coleman,
448 2011; Isaka *et al.*, 2012). Therefore, we repeated this analysis including additional species and
449 utilizing updated methods.
450 We infer, with statistical significance using both maximum likelihood and Bayesian
451 approaches, that the ancestor of the Volvocaceae was undifferentiated (Figure 5, Supplemental
452 Figure 13). This result is consistent under two models of trait evolution, equal (Figure 5) and
453 unequal (Supplemental Figure 13) transition rates between evolution of soma and loss of soma.
454 Interestingly, the undifferentiated *E. elegans* UTEX 1212 is inferred to have an ancestor
455 with somatic cells under the marginally-unpreferred unequal rates model (Supplemental Figure
456 13) and the maximum likelihood reconstructions under the marginally-preferred equal rates
457 model are ambiguous as to whether the ancestor of *E. elegans* UTEX 1212 had somatic cells

458 (Figure 5). Although, Bayesian ancestral state reconstructions show positive support for the
459 ancestor of *E. elegans* UTEX 1212 having somatic cells under the equal transition rate model.
460 Taken together, it is unclear if the ancestor of *E. elegans* UTEX 1212 had somatic cells. Yet the
461 *reg* cluster is intact in *E. elegans* UTEX 1212 (Figures 2 and 3). If the ancestor did have soma,
462 then *E. elegans* UTEX 1212 would represent a loss of soma indicating that loss of somatic cells
463 may occur without loss of the *reg* cluster.
464
465 ***Hypotheses for reg cluster functions in undifferentiated species***
466 The presence of the *reg* cluster, the genetic basis for soma in *V. carteri*, in organisms
467 lacking somatic cells presents a puzzle. What are these genes doing there? Here we propose three
468 hypotheses for the function of the *reg* cluster in species without somatic cells. It should be noted
469 that the three hypotheses below are not mutually exclusive and different members of the *reg*
470 cluster may have different functions within or between undifferentiated species.
471 First, the nearest homolog of the *reg* cluster in *C. reinhardtii*, *RLS1*/*rlsD*, is expressed in
472 response to environmental stressors that disfavor photosynthesis and cell growth, such as light,
473 sulfur, or phosphorus deprivation, when expression of chloroplast proteins is down-regulated
474 (Figure 6A) (Nedelcu & Michod, 2006; Nedelcu, 2009). In the somatic cells of *V. carteri*, *regA*
475 is thought to down-regulate the expression of chloroplast proteins, keeping somatic cells in a
476 starved state and thereby preventing their growth and reproduction (Kirk, 2001). These
477 observations suggest the hypothesis that, following duplication, the ancestral function of
478 *RLS1*/*rlsD* (down regulating cell growth in response to environmental stress) was co-opted by
479 *regA* to down regulate somatic cell growth in a developmental context (Figure 6E) (Nedelcu &
480 Michod, 2006; Nedelcu, 2009). Therefore, we hypothesize that the function of down-regulating

cell growth, and likely reproduction, in response to stress is maintained by the *reg* cluster in undifferentiated species (Figure 6B).

Second, volvocine species differ in their commitment to living in a group. Colonies of *G. pectorale* are known to dissociate under stressful conditions (Graves Jr *et al*., 1961), but to our knowledge this has never been reported in any members of the Volvocaceae. This may be because *G. pectorale* cells are held together via intercellular wall connections between cells and an outer ECM capsule while the Volvocaceae possess a colony boundary wall (Nozaki, 1990). Given that the *reg* cluster is considered to be a key regulator of cell growth during development in *V. carteri*, the *reg* cluster may regulate cell growth in species that are undifferentiated but committed to group living to prevent premature or spurious cleavage (Figure 6C).

Lastly, the Volvocaceae exhibit a gradient in eyespot and cell size along their anterior-posterior axis (AP axis) (Coleman, 2012). Such a gradient is not seen in the Tetrabaenaceae but *G. pectorale* (Goniaceae) does have a central-to-peripheral axis in flagella basal body rotation (Kirk, 2005; Coleman, 2012). Since the *reg* cluster is thought to control cell growth we hypothesize that it may be involved in forming the AP axis seen in the Volvocaceae by regulating cell growth with respect to the cell's position with in the colony (Figure 6D).

***Implications for evolution of multicellularity and individuality***

The evolution of multicellularity in the volvocine algae is thought to have arisen through three major phases: the evolution of cell cycle regulation, the evolution of increased body size, and the evolution of cellular differentiation (Hansch*en et al*., 2016). The genes underlying the first phase, the evolution of cell cycle regulation, evolved early in the evolution of the volvocine algae, probably to establish a life cycle at the group level (Hanschen *et al.*, 2016). Similarly, our

results show that the *reg* cluster, the genetic basis for somatic cells in *V. carteri*, evolved relatively early. There appears to be a common theme of early evolution and co-option of the genetic basis for traits important for multicellularity in the volvocine algae (Hanschen *et al.*, 2014; Hanschen *et al.*, 2016; Olson & Nedelcu, 2016), though when the genetic basis for the remaining phase, increased body size, evolved is currently unknown.

We have demonstrated that the *reg* cluster is present in undifferentiated species (*Pan*. *morum*, *Pla*. *caudata*, and *Y*. *unicocca*) and arose at the origin of the Volvocaceae in an ancestor inferred to be undifferentiated (Figure 5, Supplemental Figure 13) (Herron & Michod, 2008). We proposed three hypotheses for the function of the *reg* cluster in species without soma including response to environmental stress, adapting to a commitment to group living, and forming the anterior-posterior gradient (Figure 6). Whatever the function of the *reg* cluster is in undifferentiated species, the origin of the *reg* cluster predating the evolution of somatic cells demonstrates that the *reg* cluster must have been secondarily co-opted to control somatic cell development in *V. carteri* (Figure 6). Understanding the function of the *reg* cluster in undifferentiated species and what changes underlie its co-option to control somatic cell development will provide insight into how cellular differentiation, and with it greater levels of complexity and individuality, evolves.

## Literature Cited

Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Trans. Automat. Contr.* **19**: 716–723.

Bell, G. & Mooers, A.O. 1997. Size and complexity among multicellular organisms. *Biol. J. Linn. Soc.* **60**: 345–363.

Burnham, K.P. & Anderson, D.R. 2002. *Model selection and multi-model inference: a practical information-theoretic approach*. Springer-Verlag, New York.

Buss, L. 1987. *The Evolution of Individuality*. Princeton University Press, Princeton, NJ.

Coleman, A.W. 2012. A Comparative Analysis of the Volvocaceae (Chlorophyta). *J. Phycol.* **48**: 491–513.

Darriba, D., Taboada, G.L., Doallo, R. & Posada, D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**: 772–772. Nature Publishing Group.

Dent, R.M., Haglund, C.M., Chin, B.L., Kobayashi, M.C. & Niyogi, K.K. 2005. Functional genomics of eukaryotic photosynthesis using insertional mutagenesis of Chlamydomonas reinhardtii. *Plant Physiol.* **137**: 545–556.

Duncan, L., Nishii, I., Harryman, A., Buckley, S., Howard, A., Friedman, N.R., *et al.* 2007. The VARL gene family and the evolutionary origins of the master cell-type regulatory gene, regA, in Volvox carteri. *J. Mol. Evol.* **65**: 1–11.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., *et al.* 2014. Pfam: the protein families database. *Nucleic Acids Res.* **42**: D222–D230.

541 Fitzjohn, R.G. 2012. Diversitree: Comparative phylogenetic analyses of diversification in R.
542 *Methods Ecol. Evol.* **3**: 1084–1092.

543 Folse III, H.J. & Roughgarden, J. 2013. What is an individual organism? A multilevel selection
544 perspective. *Q. Rev. Biol.* **85**: 447–472.

545 Graves Jr, L.B., Kostir, W.J., Lynn, B. & Jerome, W. 1961. Some factors affecting the formation
546 of colonies in Gonium pectorale. *Ohio J. Sci.* **61**: 321.

547 Grosberg, R.K. & Strathmann, R.R. 2007. The Evolution of Multicellularity: A Minor Major
548 Transition? *Annu. Rev. Ecol. Evol. Syst.* **38**: 621–654.

549 Hanschen, E.R., Ferris, P.J. & Michod, R.E. 2014. Early Evolution of the Genetic Basis for
550 Soma in the Volvocaceae. *Evolution.* **68**: 2014–2025.

551 Hanschen, E.R., Marriage, T.N., Ferris, P.J., Hamaji, T., Toyoda, A., Fujiyama, A., *et al.* 2016.
552 The Gonium pectorale genome demonstrates co-option of cell cycle regulation during the
553 evolution of multicellularity. *Nat. Commun.* **7**: 11370.

554 Harryman, A. 2012. Investigating the roles of regA and related genes in the evolution of
555 multicellularity in the volvocine green algae. University of Maryland, Baltimore County.

556 Herron, M.D., Hackett, J.D., Aylward, F.O. & Michod, R.E. 2009. Triassic origin and early
557 radiation of multicellular volvocine algae. *Proc. Natl. Acad. Sci. U. S. A.* **106**: 3254–8.

558 Herron, M.D. & Michod, R.E. 2008. Evolution of complexity in the volvocine algae: transitions
559 in individuality through Darwin's eye. *Evolution* **62**: 436–51.

560 Huskey, R.J. & Griffin, B.E. 1979. Genetic Control of Somatic Cell Differentiation in Volvox.

*Dev. Biol.* **72**: 226–235.

Isaka, N., Kawai-Toyooka, H., Matsuzaki, R., Nakada, T. & Nozaki, H. 2012. Description of two new monoecious species of volvox sect. volvox (volvocaceae, chlorophyceae), based on comparative morphology and molecular phylogeny of cultured material. *J. Phycol.* **48**: 759–767.

Kass, R.E. & Raftery, A.E. 1995. Bayes Factors. *J. Am. Stat. Assoc.* **90**: 773–795.

Katoh, K., Kuma, K.I., Toh, H. & Miyata, T. 2005. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**: 511–518.

Kirk, D.L. 2005. A twelve-step program for evolving multicellularity and a division of labor. *Bioessays* **27**: 299–310.

Kirk, D.L. 2001. Germ-soma differentiation in Volvox. *Dev. Biol.* **238**: 213–23.

Kirk, D.L. 1998. *Volvox: Molecular-Genetic Origins of Multicellularity and Cellular Differentiation*. Cambridge University Press, Cambridge, UK.

Kirk, M.M., Stark, K., Miller, S.M., Müller, W., Taillon, B.E., Gruber, H., *et al.* 1999. regA, a Volvox gene that plays a central role in germ-soma differentiation, encodes a novel regulatory protein. *Development* **126**: 639–47.

Letunic, I., Doerks, T. & Bork, P. 2012. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* **40**: D302–D305.

Li, W.H. 1997. *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Marcot, J.D. & Mcshea, D.W. 2007. Increasing hierarchical complexity throughout the history of

life: phylogenetic tests of trend mechanisms. *Paleo* **33**: 182–200.

Maynard Smith, J. & Szathmáry, E. 1995. *The Major Transitions in Evolution*. Oxford University Press.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., *et al.* 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318**: 245–50.

Michod, R.E. 1999. *Darwinian Dynamics*. Princeton University Press, Princeton, NJ.

Michod, R.E. 2007. Evolution of individuality during the transition from unicellular to multicellular life. *Proc. Natl. Acad. Sci. U. S. A.* **104**: 8613–8.

Michod, R.E. 2005. On the transfer of fitness from the cell to the multicellular organism. *Biol. Philos.* **20**: 967–987.

Miller, S.M. & Kirk, D.L. 1999. glsA, a Volvox gene required for asymmetric division and germ cell specification, encodes a chaperone-like protein. *Development* **126**: 649–658.

Nedelcu, A.M. 2009. Environmentally induced responses co-opted for reproductive altruism. *Biol. Lett.* **5**: 805–8.

Nedelcu, A.M. & Michod, R.E. 2006. The evolutionary origin of an altruistic gene. *Mol. Biol. Evol.* **23**: 1460–4.

Nei, M. & Rooney, A.P. 2005. Concerted and Birth-and-Death Evolution of Multigene Families. *Annu. Rev. Genet.* **39**: 121–152.

Niklas, K.J., Cobb, E.D. & Dunker, A.K. 2014. The number of cell types, information content,

and the evolution of complex multicellularity. *Acta Soc. Bot. Pol.* **83**: 337–347.

Nozaki, H. 1990. Ultrastructure of the extracellular matrix of *Gonium* (Volvocales, Chlorophyta). *Phycologia* **29**: 1–8.

Nozaki, H. & Coleman, A.W. 2011. A New Species of Volvox Sect. Merrillosphaera (Volvocaceae, Chlorophyceae) From Texas. *J. Phycol.* **47**: 673–679.

Nozaki, H., Ott, F.D. & Coleman, A.W. 2006. Morphology, Molecular Phylogeny and Taxonomy of Two New Species of Pleodorina (Volvoceae, Chlorophyceae). *J. Phycol.* **42**: 1072–1080.

Nozaki, H., Takahara, M., Nakazawa, A., Kita, Y., Yamada, T., Takano, H., *et al.* 2002. Evolution of rbcL group IA introns and intron open reading frames within the colonial Volvocales (Chlorophyceae). *Mol. Phylogenet. Evol.* **23**: 326–38.

Nozaki, H., Yamada, T.K., Takahashi, F., Matsuzaki, R. & Nakada, T. 2014. New "missing link" genus of the colonial volvocine green algae gives insights into the evolution of oogamy. *BMC Evol. Biol.* **14**: 37–47.

Olson, B.J.S.C. & Nedelcu, A.M. 2016. Co-option during the evolution of multicellularity and developmental complexity in the volvocine green algae. *Curr. Opin. Genet. Dev.* **39**: 107–115. Elsevier Ltd.

Pagel, M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* **48**: 612–622.

Paradis, E., Claude, J. & Strimmer, K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289–290.

622 Prochnik, S.E., Umen, J., Nedelcu, A.M., Hallmann, A., Miller, S.M., Nishii, I., *et al.* 2010.
623 Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri.
624 *Science* **329**: 223–6.

625 Queller, D.C. 2000. Relatedness and the fraternal major transitions. *Philos. Trans. R. Soc. Lond.*
626 *B. Biol. Sci.* **355**: 1647–55.

627 R Core Team. 2013. R: A Language and Environment for Statistical Computing. Vienna.

628 Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., *et al.* 2012.
629 MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large
630 Model Space. *Syst. Biol.* **61**: 539–542.

631 Sanderson, M.J. & Wojciechowski, M.F. 2000. Improved Bootstrap Confidence Limits in Large-
632 Scale Phylogenies, with an Example from Neo-Astragalus (Leguminosae). *Syst. Biol.* **49**:
633 671–685.

634 Sharpe, S.C., Eme, L., Brown, M.W. & Roger, A.J. 2014. Timing the Origins of Multicellular
635 Eukaryotes Through Phylogenomics and Relaxed Molecular Clock Analyses. In:
636 *Evolutionary Transitions to Multicellular Life* (A. M. Nedelcu & I. Ruiz-Trillo, eds), pp. 3–
637 30. Springer.

638 Simpson, C. 2012. The evolutionary history of division of labour. *Proc. Biol. Sci.* **279**: 116–21.

639 Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. & Birol, I. 2009. ABySS: a
640 parallel assembler for short read sequence data. *Genome Res.* **19**: 1117–1123.

641 Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
642 thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–90.

643 Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of
644 large phylogenies. *Bioinformatics* **30**: 1312–1313.

645 Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. 2008. Using native and syntenically
646 mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**: 637–644.

647 Stanke, M. & Morgenstern, B. 2005. AUGUSTUS: a web server for gene prediction in
648 eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**: W465–W467.

649 Stark, K., Kirk, D.L. & Schmitt, R. 2001. Two enhancers and one silencer located in the introns
650 of *regA* control somatic cell differentiation in *Volvox carteri*. *Genes Dev.* **15**: 1449–1460.

651 Starr, R.C. 1970. Control of differentiation in Volvox. *Dev. Biol.* **4**: 59–100.

652 Umen, J.G. & Olson, B.J.S.C. 2012. Genomics of Volvocine Algae. In: *Genomic Insights into*
653 *the Biology of Algae* (G. Piganeau, ed), pp. 185–243. Elsevier Ltd: Academic Press.

654 Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**:
655 1586–1591.

656

657

## Figure Legends

**Figure 1.** Species phylogeny and micrographs of exemplar species of volvocine algae. A. Bayesian species tree, consistent with previously published species trees. *Volvox* (germ and soma differentiated) species for which the *reg* cluster has been previously sequenced are shown in black. Color of species without (*Chlamydomonas reinhardtii*, red; *Gonium pectorale*, orange) and with the *reg* cluster (undifferentiated *Pandorina morum*, *Platydorina caudata*, *Yamagishiella unicocca*, *Eudorina elegans* UTEX 1212, green; soma differentiated *Pleodorina californica*, blue) correspond to other figures, species in grey are not included in this analysis. Note that numbers following *Eudorina elegans* species refer to UTEX strain numbers. Inferred origin of the *reg* cluster based on our results is denoted. See Figure 5 for maximum likelihood and Bayesian support values. B. *Chlamydomonas reinhardtii* (scale bar, 10 μm); C. *Gonium pectorale* (10 μm); D. *Platydorina caudata* (25μm); E. *Volvox ferrisii* (50 μm); F. *Pandorina morum* (10 μm); G. *Yamagishiella unicocca* (20 μm); H. *Eudorina elegans* UTEX 1212 (10 μm); I. *Volvox carteri* f. *nagariensis* (50 μm); J. *Pleodorina californica* (25 μm).

**Figure 2.** Gene synteny near the *reg* cluster and closely related *regA*-like genes (bold). Synteny of *Chlamydomonas reinhardtii* (red), *Gonium pectorale* (orange), *Pandorina morum* (green), *Platydorina caudata* (green), *Yamagishiella unicocca* (green), *Eudorina elegans* UTEX 1212 (green), *Pleodorina californica* (blue), and *Volvox carteri* (black) is shown. All available data from *Pan. morum*, *Pla. caudata*, *E. elegans* UTEX 1212, and *Ple. californica* are shown, while representative genomic regions from *C. reinhardtii*, *G. pectorale*, *Y. unicocca*, and *V. carteri* are shown. Scaffold or chromosome numbers are indicated for *C. reinhardtii*, *G. pectorale*, *Y.*

*unicocca*, and *V. carteri*. Putative *reg* cluster and *RLS1*/*rlsD* orthologs are connected with black lines, syntenic genes are connected with gray lines.

**Figure 3.** Maximum likelihood VARL domain tree. Color of species without (*Chlamydomonas reinhardtii*, red; *Gonium pectorale*, orange) and with the *reg* cluster (*Pandorina morum*, *Platydorina caudata*, *Yamagishiella unicocca*, and *Eudorina elegans* UTEX 1212, green; *Pleodorina californica*, blue; *Volvox carteri*, *Volvox ferrisii*, and *Volvox gigas*, black) correspond to other figures. Nodes with 80% or higher bootstrap support values are labeled with support values; unlabeled nodes have less than 80% bootstrap support.

**Figure 4.** Protein similarity plots for all *reg* cluster and *RLS1/rlsD* proteins based on syntenic position (*rlsA*, *regA*, *rlsB*, *rlsO*, *rlsC*, and *rlsD*). Regions showing high similarity are highlighted with gray boxes. The two peaks in the shaded VARL region represent the N-terminal extension and core VARL domain separated by the less conserved linker region (Duncan et al. 2007).

**Figure 5.** Ancestral character state reconstruction of somatic differentiation. Branch color refers to undifferentiated (gray) or somatic differentiation (black) inferred by maximum likelihood methods using the equal transition rates model. Dashed branches indicate an ambiguous maximum likelihood reconstruction. Large font numbers at selected nodes indicate Bayes Factors using the equal rates model; negative, support for undifferentiated; positive, support for somatic differentiated. Bayes Factors are interpreted following Kass and Raftery (1995): 0-2 weak evidence, 2-6 positive evidence, 6-10 strong evidence, >10 very strong evidence. Small font numbers along branches indicate Bayesian posterior probabilities (left of slash) and

703 maximum likelihood bootstrap values (right of slash). Unlabeled nodes are supported with 1.00
704 PP and 100% bootstrap values. Bayes Factors and support values are colored consistent with the
705 reconstructed state at that node.

706

707 **Figure 6.** Conceptual schematic of hypothesized VARL gene expression patterns in unicellular,
708 undifferentiated, and differentiated species. A. Temporal expression of *RLS1* (dashed) in *C.*
709 *reinhardtii* in response to environmental change. B. The *reg* cluster maintains expression in
710 response to environmental change following its origin from the duplication of *rlsD* (dashed). C.
711 The *reg* cluster is developmentally co-opted to control cell division throughout the life cycle. D.
712 The *reg* cluster is developmentally co-opted to control the AP axis. Panels B, C, and D are not
713 mutually exclusive as different hypotheses may apply to different undifferentiated species and a
714 single hypothesis may not uniformly apply to all genes within the *reg* cluster of a given species.
715 E. The *reg* cluster is co-opted to regulate somatic differentiation in *V. carteri*. For all panels,
716 darker shading represents higher gene expression.

717

718