Short title: Chlamydomonas zygote development

Title: "Gene regulatory networks for the haploid-to-diploid transition of Chlamydomonas reinhardtii".

Sunjoo Joo, Yoshiki Nishimura, Evan Cronmiller, Ran Ha Hong, Thamali Kariyawasam, Ming Hsiu Wang, Na Chun Shao, Saif-El-Din El Akkad, Takamasa Suzuki, Tetsuya Higashiyama, Eonseon Jin, and Jae-Hyeok Lee.

a. Department of Botany, University of British Columbia, 6270 University Blvd., Vancouver, British Columbia V6T1Z4, CANADA
b. Department of Botany, Graduate School of Science, Kyoto University, Oiwake-cho, Kita-Shirakawa, Sakyo-ku, Kyoto 606-8502, JAPAN
c. ERATO, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8602, JAPAN
d. Department Life Sciences, Research Institute for Natural Sciences, Hanyang University, 222 Wangsipri-ro, Sungdong-gu, Seoul 133-791, Republic of KOREA

*. Corresponding author email: jae-hyeok.lee@botany.ubc.ca

One sentence summary

The zygotic wall program is executed by the heterodimeric homeobox transcription factors GSM1/GSP1 via transcriptional activation and post-transcriptional suppression in Chlamydomonas.

Author Contributions:
Funding:
This work was supported by grant 2016M1A8A1925345 from the Korea CCS R&D Center (KCRC), Korean Ministry of Science, and by Discovery Grant 418471-12 from the Natural Sciences and Engineering Research Council (NSERC) of Canada to J.-H. Lee. Y. Nishimura was supported by NEXT program: GS015. Sequencing was supported by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research program under Award Number DE-FC02-02ER63421 to Sabeeha Merchant.

Correspondence author: Jae-Hyeok Lee
Department of Botany, University of British Columbia,
6270 University Blvd., Vancouver, BC V6T1Z4, Canada
Phone: 1-604-827-5973
E-mail: jae-hyeok.lee@botany.ubc.ca
Abstract

The sexual cycle of the unicellular *Chlamydomonas reinhardtii* culminates in the formation of diploid zygotes which differentiate into dormant spores that eventually undergo meiosis. Mating between gametes induces rapid cell wall shedding via the enzyme g-lysin; cell fusion is followed by heterodimerization of sex-specific homeobox transcription factors, GSM1 and GSP1, and initiation of zygote-specific gene expression. To investigate the genetic underpinnings of the zygote developmental pathway, we performed comparative transcriptome analysis of both pre- and post-fertilization samples. We identified 253 transcripts specifically enriched in early zygotes, 82% of which were not up-regulated in *gsp1* null zygotes. We also found that the GSM1/GSP1 heterodimer negatively regulates the vegetative wall program at the post-transcriptional level, enabling prompt transition from vegetative wall to zygotic wall assembly. Annotation of the g-lysin-induced and early zygote genes reveals distinct vegetative and zygotic wall programs, supported by concerted up-regulation of genes encoding cell wall-modifying enzymes and proteins involved in nucleotide-sugar metabolism. The haploid-to-diploid transition in *Chlamydomonas* is masterfully controlled by the GSM1/GSP1 heterodimer, translating fertilization and gamete coalescence into a bona fide differentiation program. The fertilization-triggered integration of genes required to make related, but structurally and functionally distinct organelles -- the vegetative vs zygote cell wall -- presents a likely scenario for the evolution of complex developmental gene regulatory networks.
Introduction

Metazoa, Embryophyta, and Fungi, the so-called crown groups, independently evolved developmental programs that give rise to complex multicellular organisms. Their explosive morphological diversity is largely driven via modification or reuse of existing gene regulatory networks (GRNs) consisting of transcriptional regulators and signaling pathways (Carroll 2008; Davidson and Erwin 2006; Pires and Dolan 2012; Rudel and Sommer 2003). Therefore, inquiry about the origins of the crown groups has been focused on analyzing the evolutionary precursors of the GRNs that underlie their key developmental strategies.

Large-scale comparative genomics studies and more focused analyses of ancestral lineages closely related to the Metazoa (e.g. choanoflagellates) and Embryophyta (e.g. charophytes) have indeed reported that many GRN components predate the origins of the crown groups (Fairclough et al. 2013; Hori et al. 2014; King et al. 2008; Mendoza et al. 2013; Wickett et al. 2014; Worden et al. 2009). These findings encourage inquiry into the roles of pan-eukaryotic GRN modules in the differentiation repertoires of the unicellular eukaryotes that share common ancestry with crown-group organisms. Research on the unicellular green alga *Chlamydomonas reinhardtii* has been particularly useful for gaining insights into the origins of basal GRNs of the Embryophytes, since large numbers of gene families that are not found in fungi or animals are shared within Viridiplantae (Merchant et al. 2007; Worden et al. 2009).

Given that the most recent eukaryotic common ancestor engaged in sexual reproduction (Goodenough and Heitman 2014), an obvious focus for these inquiries has been the strategies and proteins involved with gametic differentiation, mate recognition, zygote/spore formation, and meiosis. Particularly fruitful have been studies of these processes in unicellular yeasts, where transcriptional and signaling cascades are found to be conserved in the
sexual reproduction of mushrooms and beyond (reviewed in Honigberg and Purnapatre 2003; Neiman 2011; Mata et al. 2002). Sexual development of *Chlamydomonas*, detailed below, uses several such strategies and proteins. As an example, both *Chlamydomonas* and yeast gametes contain a transcription factor that forms a heterodimer in the diploid zygote that is involved in diploid development (Goutte and Johnson, 1988; Lee et al. 2008). Furthermore, the *Chlamydomonas* heterodimeric transcription factors GSM1/GSP1 in the zygote are structural and functional homologs to the KNOX/BELL homeobox heterodimers that are also involved in the diploid development of land plants (Sakakibara et al. 2008; 2013; Horst et al., 2015). Such deeply rooted conservation of sexual GRNs suggests their fundamental importance for complex multicellular evolution.

The proteins involved in *Chlamydomonas* sexual development also exemplify the deep ancestry of sexual processes (Speijer et al., 2015). GEX1, for example, has been identified as a gamete-expressed nuclear envelope fusion protein that is required for nuclear fusion in protists, fungi, plants, and many vertebrates (except those, including humans and mice, whose pronuclei do not fuse in the zygote) (Ning et al., 2013). And the primordial gamete fusogen HAP2 likely was present at the origins of sexual reproduction in eukaryotes (Liu et al., 2008). Recently, HAP2 was shown to be a class II fusion protein, in the same family as dengue and Zika virus fusion proteins (Fedry et al., 2017).

Molecular understanding of sexual development of *Chlamydomonas* has been driven by molecular genetics studies (reviewed in Goodenough et al. 2007). The *Chlamydomonas* sexual cycle is invoked by nitrogen starvation, and entails expression of sex-specific proteins involved in pre- and post-fertilization events during gamete interactions (mating) and zygote formation (Figure 1). Sex-specific *(plus and minus)* gametogenesis programs are governed by the bi-allelic mating type loci called MTL+ and MTL- (Ferris et al. 2002; De Hoff et al. 2013). The *FUS1* gene, exclusive to MTL+, encodes a glycoprotein required for the *plus*
gametes to bind and fuse to minus gametes (Ferris et al. 1996), and the MID gene, exclusive to MTL-, encodes a transcription factor in the RWP-RK family that induces the minus sex program and suppresses the plus sex program (Lin and Goodenough 2007; Ferris and Goodenough 1997). Therefore, both MTL+ and MTL- are necessary for a successful progression through the sexual cycle.

Minus and plus sexual programs employ a series of key players acting in pairs that are expressed in a sex-specific manner: agglutinins (SAD1 and SAG1) on the flagellar membrane (Ferris et al. 2005; Goodenough et al. 1985), fusion-enabling factors (HAP2 and FUS1) on the plasma membrane (Misamore et al. 2003; Liu et al. 2015; 2008), and heterodimeric transcription factors in the
cytosol, GSM1 and GSP1 (Lee et al. 2008; Zhao et al. 2001). SAD1 and SAG1, large hydroxyproline-rich glycoproteins (HRGPs), serve as both ligands and receptors for the initial gamete-recognition event (Ferris et al. 2005), which triggers downstream events collectively called the mating reaction that involves cyclic AMP-mediated signal transduction via a cyclic GMP-activated protein kinase (reviewed in Snell and Goodenough 2009). Responses to cyclic AMP elevation include activation of a periplasmic pro-enzyme called gametolysin (g-lysin) (Buchanan et al. 1989) that removes cell walls, promotes translocation of additional agglutinins to flagella (Cao et al. 2015), and activates the sex-specific mating structures where HAP2 and FUS1 are localized (Goodenough and Jurivich 1978; Misamore et al. 2003; Liu et al. 2008). During gamete differentiation, the homeobox transcription factors GSM1 and GSP1 are synthesized and stored in the minus and plus cytoplasms; upon gametic cell fusion they heterodimerize and translocate to the nuclei, turning on zygote developmental pathway and preparing fused cells for eventual meiotic germination (Lee et al. 2008). Therefore, the MID-controlled minus and plus programs prepare minus and plus gametes for a series of events, occurring within a few minutes in a step-wise manner, that initiates the diploid phase of the life cycle (Figure 1).

The transition from gametes to zygotes (the haploid-to-diploid transition) represents a key step in the Chlamydomonas life cycle, analogous to the embryogenesis following fertilization in plants without entailing mitotic growth. Zygote development entails signature events largely completed within the first 12 hr, such as zygote-specific wall assembly (Minami and Goodenough 1978; Woessner and Goodenough 1989; Suzuki et al. 2000), selective destruction of chloroplast nucleoids derived from minus gametes and of mitochondrial DNAs from plus gametes (Kuroiwa et al. 1982; Aoyama et al. 2006), flagellar resorption and basal-body disassembly (Cavalier-Smith 1974; Pan and Snell 2005), and nuclear and chloroplastic fusion (Cavalier-Smith 1976). Such a rapid differentiation process would require robust regulatory networks to prevent
promiscuous execution, networks that combine intra- and perhaps intercellular
cues and coordinate a multitude of events.

The walls of vegetative and gametic cells serve two core functions. 1) They must
protect cells from mechanical and osmotic stresses, and 2) they must be easily
removed. Vegetative cells shed the so-called mother cell wall after their mitotic
progeny have formed their own cell walls, and upon encountering a gamete of
the opposite mating type, gametes must be able to rapidly shed their walls to
allow cell fusion. In both vegetative cells and gametes, transcripts for cell wall
genes are upregulated when their walls are experimentally released by
incubating cells in g-lysin (Kurvari, 1997; Hoffmann and Beck, 2005). One
notable difference is that only the vegetative wall needs to support large
increases in cell volume, which requires continuous remodeling of the existing
wall layers.

In contrast, the zygote wall has only a single function, which is to provide
physical and chemical enclosure that allows *Chlamydomonas* to survive loss of
nutrient sources and extreme environmental conditions, including dessication.
The zygote cell wall is unaffected by g-lysin (Schlosser, 1976), but is removed by
the zygote hatching enzyme (z-lysin) whose biochemical nature is unknown.

The zygote wall-specific features include secretion and assembly of zygote-
specific extracellular materials such as (1-3)β-D-glucans into 4-8 complex layers.
These are distinct in structure and composition from the vegetative wall (Catt
1979; Cavalier-Smith 1976; Grief et al. 1987), which is organized into seven
layers entirely made of HRGPs (Roberts et al. 1972; Goodenough and Heuser,
1985). In liquid cultures, zygote walls connect neighboring zygotes into extensive
sheets called pellicle (Minami and Goodenough 1978; Suzuki et al. 2000).

Two previous reports document that the haploid-to-diploid transition of *C.*
reinhardtii is initiated by the GSM1/GSP1 heterodimer: 1) ectopic expression of
GSM1/GSP1 heterodimers in vegetative cells can induce early zygote genes, pellicle formation, and ultimately commitment to meiosis (Lee et al. 2008); and 2) zygotes of gsp1 null (biparental31) plus gametes and wild type minus gametes cannot activate early zygotic genes nor degrade minus-derived chloroplast DNA, and they undergo extra fusions (Nishimura et al. 2012).

The molecular details of the zygote differentiation program have been explored by a transcriptome analysis (Lopez et al., 2015), which reported ~ 600 zygote-specific genes. However, further details of the regulatory networks for the early-zygote developmental pathway, and the functional contexts of the zygote-specific genes for zygote differentiation, are largely unexplored.

The present study analyzes the regulatory network controlling the majority of the early zygote (EZ) transcriptome by considering the mating-induced signaling events and contribution of the GSM1/GSP1 heterodimer. Our results define the predominant role of the GSM1/GSP1 heterodimer, and provide molecular insights concerning how various cellular machineries participate in zygote differentiation.
Results

Experimental design for identifying early zygote-specific genes

To identify which mRNAs derive from bona fide EZ-specific genes, we considered the following in the sample choice for transcriptome sequencing. Firstly, we included haploid and diploid strains in N-replete samples, called TAP1/2 in order to minimize mating-type-dependent or -limited or ploidy-dependent expression biases. Secondly, genetically diverse strains were used in the two early-zygote samples combining 45 min and 2 hr sets after mating as EZ1/2 in order to minimize strain-dependent biases. We compared our samples with the following published samples representing gene regulatory programs (given in parenthesis) that function prior to zygote development: N-replete and N-starved samples (program NS) (Miller et al. 2010); plus and minus gametes (program GAM), some being mating-type specific (programs PG and MG) (Ning et al. 2013); and gamete samples following dibutyryl cyclic AMP-treatment (program MR, > 350 genes) or subjected to g-lysin treatment (program gL, > 100 genes) (Ning et al. 2013). In the end, the zygote-enriched transcripts that are induced by the NS, GAM, MR or gL programs were excluded to collect bona fide transcripts controlled by the fusion-dependent program (program FD).

To assess the contribution of the GSM1/GSP1 heterodimer to the FD program, we compared zygote samples in a gsp1 null background (bp31) and its complemented strain (bp31C) (Nishimura et al., 2012). Finally, gamete samples ectopically expressing both GSM1 and GSP1 were included in the above multi-way comparison as pseudo-zygotes (PZ condition); these cells exhibit zygotic traits such as pellicle formation without undergoing gamete fusion (Lee et al. 2008). Hence, the entire dataset included 15 conditions in 11 genetically diverse strains (Table 1 and Supplemental Table S1).
Our dataset includes field isolates that are highly polymorphic. Therefore, during the RNA-seq mapping, we allowed 3 mismatches after considering average synonymous polymorphism among the field isolates (nucleotide diversity at synonymous sites, dS=0.0317, Flowers et al. 2015). As a result, mapping rates range between 82.1% and 98.2% of the total reads (Supplemental Table S2). Further details of data processing and analysis are given in Supplemental Note S1. Following read mapping, gene expression values normalized as FPKM (Fragments Per Kilobase transcript per Million reads) were calculated using the V5.3 genome annotation available at Phytozome (http://www.phytozome.net/, Supplemental Table S3).

Clustering analysis identifies early-zygote specific genes in two distinct pools

Since defining zygote-specific genes relies on the cross-referencing of all other conditions, we performed un-guided clustering analysis with 13 conditions, excluding two zygote samples of the gsp1 null strain, averaged from 33 independent samples following FPKM calculations. Prior to the analysis, 5656 (out of 18823) gene models were discarded as Cluster 0 by low expression (having less than 4 FPKM in more than 31 samples), likely representing either pseudo-genes or genes specifically expressed under conditions not included in our dataset. The rest were normalized per model by converting FPKM into percentage values of expression over the sum of average FPKM values under 13 conditions, so that condition-specific expression patterns, rather than absolute expression or fold-change, could be highlighted. The resulting percentage values were used to generate 50 clusters of similar expression patterns by Ward’s hierarchical clustering algorithm implemented in the Genomics workbench software v4.

Among the 50 clusters, C33, C43, and C50 show highly up-regulated or nearly exclusive expression in the EZ condition (Box plots, Figure 2A-C), and these are
Joo et al. Figure 2.

Figure 2. Identification of co-regulated genes in the early zygote transcriptome by hierarchical clustering analysis. Relative expression patterns among 13 conditions are shown for selected gene clusters representing the early zygote-specific clusters (C33, C43, C50, and C10) and the g-lysin-induced clusters (C44 and C24). Box plots show distribution of percentage expression values per condition, calculated relative to the sum of FPKM values in all 13 conditions per gene. Boxes show the median and the 25th and 75th percentile. Whiskers indicate the 10th and 90th percentiles. Details pertaining to the remaining clusters are found in Supplemental Tables S4 and S5.

designated EZ-specific genes. C10 also shows significant up-regulation in EZ, but relatively strong basal expression in all the other conditions as well, and was therefore excluded from the EZ-specific set (Figure 2D). Of special interest is C44, where similar enrichment is found between EZ and g-lysin-treated gametes (PL and ML conditions), in contrast to C24 where g-lysin-induction is strong but EZ-enrichment is absent (Figure 2E-F). Hereafter these two gene clusters are referred to as gL+EZ (C44) and gL-EZ (C24). Detailed FPKM statistics and description of the 50 clusters are available in Supplemental Table S4 and S5.
Out of 444 candidate genes from the three EZ-specific clusters (after discarding redundant gene models), we further refined the EZ set by removing genes with less than a 4-fold induction between EZ and the vegetative condition (TAP) (98 genes) or with maximum expression of less than 12 FPKM in EZ (54 genes). We further eliminated 39 genes whose zygote-specific expression is only evident in one of the two biological duplicates (EZ1 and EZ2 in Supplemental Table S1). The remaining 253 genes are defined as the core EZ genes (EZ-core, Supplemental Table S6).

We cross-examined the EZ-core with previously reported early-zygote genes (Supplemental Table S7 for EZ-core annotation). EZY1, ZYS1-4, and ZSP1-2 are found in the EZ-core, while EZY2 is absent due to its low expression. Genes EZY3-23 were identified in a microarray study by Kubo et al. (2008), where EZY genes were divided into two classes based on g-lysin-inducibility. Most of their non-g-lysin class genes (EZY4, 7, 8, 9, 15, 17, 18, 19, 20, 22) are found in our EZ-core, while their g-lysin-induced genes are mostly found in gL+EZ (EZY11, 12, 13, 14) and in gL-EZ (EZY5, 23). EZY3 and 10 in their non-g-lysin class are found in gL+EZ, with clear g-lysin-inducibility both in our data and in Figure 1 of Kubo et al. (2008). EZY6, 16, and 21 in their g-lysin class show strong expression elsewhere in our dataset and therefore fall in different clusters (EZY6, 16 in C38, and EZY21 in C13). This comparison confirms that our clustering approach collected the majority of EZY genes and precisely distinguished those that are g-lysin-induced.

Another EZ transcriptome study by Lopez et al. (2015) reported a total of 627 zygote-specific gene based on 4-fold up-regulation by RNA-seq, which is more than twice as large as our EZ-core. We found that the difference is mainly due to the inclusion of low-expressed genes (144 models with < 2 FPKM), g-lysin induced genes (60 models), and multiple models for the same gene (30 cases). Of the remaining 393 genes, 82.7% (325 genes) are found among our EZ-specific clusters and C10 (showing EZ up-regulation with strong basal...
expression). Overall, the comparison to the Lopez dataset shows that our clustering method can sort the complex EZ transcriptome into sub-classes and is sensitive enough to collect the majority of differentially expressed genes.

Expression of EZ-core genes is >94% controlled by the GSP1/GSM1 heterodimer

To learn which genes in the EZ-core are under the control of GSM1/GSP1 heterodimers, we analyzed differential gene expression between the zygotes generated from the gsp1 null and its complemented strain (bp31 and bp31C zygotes; Nishimura et al. 2012) using a bioconductor package, DESeq2 (Love et al. 2014). A significant increase in bp31C zygotes compared to bp31 zygotes is interpreted to indicate GSM1/GSP1 dependency. In total, 186 genes show > 4-fold increased expression (484 genes at > 2-fold increase) and only 11 genes show > 4-fold decreased expression (115 genes at > 2-fold decrease) in bp31C zygotes over bp31 zygotes (Supplemental Table S8, S9 and S10). Those up-regulated genes are mostly from the EZ-specific clusters (166/186). Of the 253 EZ-core genes, 207 (81.8%) show significant > 2-fold up-regulation in bp31C zygotes over bp31 zygotes (FDR <0.05, Supplemental Table S6). In contrast, a > 2-fold enrichment of bp31C zygotes was found for only 158 out of 13,185 genes (1.2%) from the non-EZ clusters (excluding C10/33/43/50, Supplemental Table S11). Of the remaining 46 EZ-core genes showing no significant up-regulation in bp31C zygotes, 34 genes are either poorly expressed in bp31C zygotes or found to be g-lysin-inducible, meaning that nearly all (207/219, 94.5%) of the EZ-core genes are dependent on GSP1 for EZ-expression.

We also examined the differential expression of EZ-core in the pseudozygotes (PZ condition) where the zygote program is activated by ectopic expression of GSM1/GSP1 without gamete fusion. We found that 146 of the 253 EZ-core genes show >2-fold induction (Supplemental Table S6, column Z) over the TAP condition, supporting the dominant role of the GSM1/GSP1 heterodimer in
activating EZ-specific genes. The level of up-regulation in the PZ condition is consistently lower than in the EZ condition, likely due to the fact that the zygotic program is non-synchronously induced in the pseudozygotes following N-starvation (data not shown).

To check the reproducibility of our analysis, 10 of the EZ-core genes at high (>200 FPKM), medium (100-200), and low (<100) expression levels (Supplemental Table S12A) were analyzed by quantitative RT-PCR (qRT-PCR) in three biological replicates of 30 min, 1 hr and 2 hr zygote samples and compared to plus and minus gametes (Figure 3A and Supplemental Figure S1). All show early induction at 30 min and continued expression until 2 hr following gametic fusion, in a GSP1-dependent manner.

The EZ-core is regulated at the transcriptional level via TGAC motif.

We next asked whether the GSM1/GSP1-dependent gene activation is regulated at the level of transcription given that RNA-seq cannot distinguish between transcriptional and post-transcriptional controls. We constructed reporter transgenes with luciferase coupled to the promoter sequences of two EZ-core genes -- g15252(ZSP2A) with very clean EZ-specific expression and Cre06.g256800(SND1A) with additional up-regulation in the activated gametes (MA/PA conditions) -- and analyzed their expression in wild type and gsp1 null zygotes (Figure 3B). The ZSP2A promoter drove luciferase activity from 1 hr up to 4 hr after zygote formation (~ 6 fold increase), whereas no activation was found in gsp1 null zygotes. The SND1A promoter showed ~ 100-fold increase in luciferase activity early in wild type zygotes, whereas a delayed but significant ~40-fold increase was found in gsp1 null zygotes. A similar delayed up-regulation of SND1A in gsp1 null zygotes was observed in qRT-PCR analysis (Figure 3A), indicating that the SND1A promoter is controlled by the mating reaction in addition to the GSM1/GSP1 heterodimer. In the EZ-core, we have found 11/253 additional mating reaction-induced genes based on the average FPKM ratio of...
However, they all show greater expression in the EZ condition, suggesting that their functions are mainly focused on zygote development.

With the promoter-reporter assay results demonstrating the action of GSM1/GSP1 heterodimer at promoters, we performed in silico analysis to identify putative cis-acting elements in the pool of EZ-specific genes by Amadeus software (Linhart et al. 2008). The analysis identified three 10-base motifs (Supplemental Figure S2A). The most significant hit, "gtGACacGAC," shares the same GACnnGAC consensus with the previously reported ZYgote Responsive cis-acting Element candidates (ZYRE; Lee et al. 2008; Hamaji et al. 2016). The second and third motifs also contain in common a single TGAC sequence,

Figure 3. GSM1/GSP1-dependent activation of EZ-core genes. (A) Relative abundance of transcripts in early stage zygotes to minus gametes was analyzed using qRT-PCR. Error bars: standard deviation of three biological replicates. g+: plus gametes, g-: minus gametes, Z.5, Z.1, and Z.2. 0.5 hr, 1 hr, and 2 hr after mixing gametes. Black bar: CC-621 (mt-) X CC-125 (mt+); Gray bar: CC-621 (mt-) X bp31 (mt+). (B) Promoter-luciferase reporter assay of two EZ-core promoters. Fold-difference was calculated from cumulative luciferase based promoter activity during the early zygote development relative to samples without mating. Error bars: standard deviation of three biological replicates.

the MA/PA conditions over the MG/PG conditions (Supplemental Table S6).
possibly representing the core motif recognized by the GSM1/GSP1 heterodimer, which is also the motif repeatedly identified for TALE-class homeobox transcription factors with "WFI50N" (also shared by GSM1) (Knoepfler et al. 1997; Krusell et al. 1997).

Distribution of the three ZYRE motifs was examined to find additional genes under the direct regulation of GSM1/GSP1 heterodimer. Overall, the three motifs are found 6,024 times in 4,468 out of 18,458 genes (24%, Supplemental Table S13). Per cluster distribution shows significant increase in the three EZ-specific clusters - C33, C43, and C50 - as expected (the highest occurrence being 57% in C43), but also in C10 and C44 (gL+EZ) (Chi-square test, p<0.001, Supplemental Figure S2B, Supplemental Table S14).

We also asked independently whether C10 (showing strong basal expression throughout the examined samples) and C44 (gL+EZ) genes are directly regulated by the GSM1/GSP1 heterodimer using the differential expression analysis between bp31C and bp31 zygotes. C10 includes 74/475 genes that are significantly up-regulated in bp31C zygotes, and C44 includes 9/162 such up-regulated genes (> 2-fold, FDR <0.05, Supplemental Table S11). Comparing bp31C/bp31 ratio distribution per our co-expressed gene cluster indicates that both C10 and C44 are indeed significantly higher in the bp31C/bp31 ratio than the average ratio distribution of all the other clusters combined (Wilcoxon-test, p<0.001, Supplemental Table S11). In conclusion, the two statistical analyses suggest that at least some of the C10 and C44 genes are directly up-regulated by the GSM1/GSP1 heterodimer.

A subset of g-lysinducible genes is selectively suppressed by the zygote program
g-lysind is released during the mating reaction to remove cell walls and allow gametes to fuse. It has long been known that if the supernatant from mating cells
is harvested and applied to vegetative cells or unmated gametes, their walls are also removed, and the cells proceed to assemble new walls via activating genes encoding wall proteins and proteins of unknown functions (Hoffmann and Beck, 2005; Kubo et al., 2008; Ning et al., 2013). The walls of vegetative and gametic cells are apparently identical (defined as the VG-wall in contrast to the zygotic wall [Z-wall]; Monk, 1988). Therefore, g-lysin-treated unmated gametes, and newly fused gametes exposed to g-lysin during a natural mating, are a priori expected to up-regulate the same VG-wall transcripts that are up-regulated in g-lysin-treated vegetative cells.

As noted earlier, two clusters containing the g-lysin-induced genes present an interesting anomaly: the C44 genes (gL+EZ) exhibit EZ-enrichment, whereas the C24 genes (gL-EZ) do not show any sign of EZ-induction. Moreover, the bp31C/bp31 zygote expression ratio is significantly less than one for the gL-EZ set (Supplemental Figure S3, Supplemental Table S11). This suggests that the gL+EZ subset of g-lysin-induced transcripts persists in early zygotes, whereas the gL-EZ subset does not, either because the transcription of the gL-EZ set is switched off and/or because their transcripts are actively targeted for degradation or decay.

To explore the differential regulation of gL+EZ and gL-EZ genes, we analyzed several g-lysin-inducible genes in wild-type and bp31 (gsp1 null) zygotes by qRT-PCR and the promoter-reporter assay used for the EZ-core analysis. Three gL+EZ genes (RHM1, AraGT1, RRA2) and five gL-EZ genes (SEC61G, VSP3, PHC19, GAS28, P4H1) of various expression levels were selected for the analysis (Supplemental Table S12).

qRT-PCR results show that all three gL+EZ genes, and also SEC61G, were up-regulated 3~6 fold in gsp1 null zygotes and 5~8 fold in wild-type zygotes (Figure 4A). In contrast, the other four gL-EZ genes show significant up-regulation in gsp1 null zygotes but almost no changes or decrease in expression in wild-type
zygotes (Figure 4A). Hence our qRT-PCR results document that the differential regulation of the two gene sets is GSM1/GSP1-dependent.

The exceptional SEC61G gene actually shows comparable RNA-seq data to our qRT-PCR results, where its expression in bp31C zygotes is 1.5 fold higher than in bp31 zygotes, similar to other gL+EZ genes (Supplemental Table S12). We
manually inspected RNA-seq data of the gL-EZ set (172 genes) and identified 16 genes showing comparable expression between the EZ and ML/PL conditions (less than two-fold difference) as potential gL+EZ genes (noted in the column K of Supplemental Table S17), which contains five out of seven ER-residing secretion-related genes including SEC61G in the original gL-EZ set.

We next sought to learn more about the g-lysin program. We first used our promoter-reporter assay to test if the g-lysin program operates at the transcriptional level. We constructed transgenic strains harboring promoter-reporter constructs for AraGT1, RHM1, and SEC61G (representing gL+EZ) and for PHC19 (representing gL-EZ); all showed consistent inducible reporter expression when subjected to exogenous g-lysin exposure, indicating that the g-lysin program regulates both sets via transcriptional activation (Supplemental Figure S4 for the promoter cloning schemes and Supplemental Figure S5 for g-lysin inducibility test). We then asked how the g-lysin program is affected during zygote development by examining those g-lysin inducible promoters. The RHM1, SEC61G, and PHC19 promoters showed comparable up-regulation during the first four hours after gametic fusion in both wild-type and bp31 (gsp1-null) zygotes (10~15-fold in 3 hr zygotes, Figure 4B). The AraGT1 promoter showed stronger up-regulation in wild-type zygotes, although weak but clear up-regulation occurred in gsp1 null zygotes (3-fold in 3 hr zygotes, Figure 4B).

The promoter assay results document that these test genes, and by extension the g-lysin program, operates normally during zygote development and hence is not inhibited by the GSM1/GSP1-dependent program. Therefore, the observed paucity of transcripts produced by gL-EZ genes in wild type zygotes suggests the existence of a post-transcriptional mechanism for their down-regulation, one that does not affect gL+EZ genes.

Gamete-specific genes are not apparently suppressed by the GSM1/GSP1 heterodimer during early zygote development.
The zygote developmental program is expected to be accompanied by a negative control of gametic features such as agglutinins and mating structures to prevent additional gametic fusion. We examined expression of known gamete/sex-specific genes in our RNA-seq dataset (Supplemental Table S15). Most known minus-specific genes in C18, including SAD1, GSM1, HAP2, and MTD1 are expressed at comparable levels in gametes and early zygotes. The same trend is true for the plus-specific genes residing in MTL+ such as FUS1 and MTA1. This trend indicates that the gamete-specific programs are not apparently suppressed by the zygote program. In contrast, the plus-specific genes residing outside MTL+ such as GSP1 and SAG1 showed rapid down-regulation as early as 30 min zygotes. Earlier examination of GSP1 and SAG1 expressions by qRT-PCR also showed rapid turn-off in early zygotes (Nishimura et al. 2012). Interestingly, SAG1 down-regulation was found equally in both bp31C and bp31 zygotes, suggesting that the down-regulation of the plus-specific genes may be triggered by cellular fusion rather than the GSM1/GSP1 heterodimer.

**Prediction of functional contexts of the early zygote transcriptome**

Our analysis has revealed that EZ-core genes and the gL+EZ set of g-lysin-induced genes constitute the main body of the early-stage zygote transcriptome dependent on the GSM1/GSP1 heterodimer. Although 203/253 members of the EZ-core were previously identified as zygote-specific genes by Lopez et al (2015) and 30/159 of the gL+EZ as g-lysin induced genes by Ning et al. (2013), these earlier studies offered annotations but not detailed analyses. Therefore, we have gone on to analyze these co-expressed gene clusters, focusing on which functional systems are enriched in the EZ transcriptome and assessing the differences among EZ-core/gL+EZ/gL-EZ gene contents.
To analyze molecular details, we performed exhaustive manual annotation as described in Materials and Methods, and categorized the genes into seven major functional groups relevant to the zygote differentiation processes: cell wall synthesis and assembly, organelle-targeted, miscellaneous metabolism including transporters, putative gene regulation, signal transduction, secretion/endomembrane/cytoskeleton system, and ubiquitin system (Detailed annotation results in Supplemental Tables S7, S16 and S17). Figure 5 and Supplemental Tables S18 summarize the functional categories among the three co-regulated gene sets. In the following, we focus on Z-wall assembly where our transcriptome has provided new insights and putative molecular players.

**Molecular components of the zygotic cell wall assembly**

The largest annotated functional category associated with EZ-core/gL+EZ/gL-EZ is cell wall-related genes (46/253, 25/159, and 49/172, respectively), including 1) HRGP-encoding, 2) HRGP-modifying enzymes, and 3) nucleotide-sugar metabolism-related, including sugar-converting enzymes and nucleotide-sugar transporters (Figure 5).

**Cell wall constituents: HRGPs and products of glycosyl transferases**

Previous studies (Minami and Goodenough 1978; Grief et al. 1987) have documented that the zygotic wall and vegetative wall consist of distinct sets of glycoproteins, primarily HRGPs, recognized by domains dominated by proline (P) and serine (S), often in repeated motifs. While the gL+EZ contains no HRGP-encoding genes, the EZ-core and gL-EZ include unique sets of HRGP-encoding genes. Of particular interest is the presence of a C-terminal hydrophobic domain in six out of the nine HRGPs in the EZ-core (Supplemental Table S7). Such a hydrophobic tail serves as either a transmembrane domain or a signal for GPI-like lipid-modification. *In silico* tools predicting GPI-like lipid modification signals (PredGPI and BIG-PI; Pierleoni et al. 2008; Borner et al. 2003) showed that five of the six candidates likely function as GPI-anchor signal (Supplemental Table
A GPI-anchor motif is often associated with plant AGP (arabinogalactan protein)-type HRGPs but has not yet been experimentally characterized in *Chlamydomonas*. GPI-type lipid-modification was predicted for a fasciclin-like HRGP, Algal-CAM, in the *Volvox* vegetative wall (Huber and Sumper 1994), introducing the idea that GPI-anchors or similar structures are likely present in algal cell walls. In support of this idea, most of the GPI-anchor biosynthesis genes are found in the *Chlamydomonas* genome, of which two enzymes...
functioning at rate-limiting steps, *PIG-O* and *PGAP1*, are up-regulated in early zygotes (C33) (Supplemental Table S20).

We found a novel six-Cys-containing homology domain among all the EZ-core HRGPs with C-terminal anchors except Zsp1, which we named MAW (Membrane Associated Wall protein) (Figure 6C for the conserved amino acids; Supplemental Figure S6 for the domain sequence alignment). We searched for additional MAW family members by HMMER3 (Mistry et al. 2013), and identified 11 *Chlamydomonas* members, five members in *Volvox carteri*, and seven members in *Gonium pectorale* (Supplemental Table S19). We could not find MAW homologs in any other published green algal genome sequences.

*Chlamydomonas* MAW proteins have up to three MAW domains, and the majority (10/11) contain Pro-rich domains of PPSPX, PPX, and/or PS repeats, characteristics of HRGP shafts (Figure 6A; Lee et al. 2007). Intriguingly, all MAW members identified are found to possess either one or two trans-membrane domains or a putative GPI-anchor signal at the C-terminus (Supplemental Table S19), indicating that the MAW domain associates with plasma membrane-anchored cell wall proteins in the Volvocales.

Phylogenetic analysis of the MAW family reveals six well-supported clades among Volvocalean species (Figure 6B). Clades I and II are distinguished as having a putative GPI-anchor signal at the C-terminus (except CreMAW10). Since all *Chlamydomonas* genes in Clade I and II are exclusively expressed in early zygotes, we propose that they may serve as key organizers of Z-wall development in Volvocacean algae. Clades IV and V share triple MAW domain configurations, and have diverse expression patterns in *Chlamydomonas* (Figure 6D).

The absence of HRGP-encoding genes in the gL+EZ set that contribute both to the VG- and Z-walls suggests that VG-wall and Z-wall may be assembled with exclusive sets of materials. We therefore next asked what features are unique to the VG-wall, taking the gL-EZ set as its representative. The gL-EZ includes 36
putative HRGP-encoding genes, 20 of which are pherophorins, the largest known HRGP family in the Volvocales. The *Chlamydomonas* genome encodes 73 pherophorin genes, 28 of which are up-regulated in g-lysin-enriched clusters.
(C24/8/34), while 10 are found in gamete- and mating reaction-induced clusters (C22/26/41) but none are in EZ-specific clusters (C10/33/44/44/50) (Supplemental Table S21). Hence we propose that pherophorins are hallmark components of the VG-wall.

**HRGP modifying enzymes**

HRGPs are made from nascent polypeptides that undergo extensive O-glycosylation on serines and hydroxylated prolines (HYPs) within proline-rich segments during ER-Golgi trafficking. Therefore, production of HRGPs requires prolyl 4-hydroxylases (P4Hs) and various glycosyltransferases. In land plants, distinct glycosyltransferase families are known to build either unbranched O-glycans found in extensins or complex O-glycans found in AGPs (reviewed in Hijazi et al. 2014).

Both serine- and HYP-glycosyltransferases have been recently identified in plants from studies using plant extensins as substrates: SGT (Peptidyl Serine alpha-Galactosyltransferase) and HPAT (hydroxyproline O-arabinosyltransferases) belong to the GT8 family and catalyze initial O-glycosylation on serine or HYP (Saito et al. 2014; Ogawa-Ohnishi et al. 2013), and subsequent arabinosylation is catalyzed by members of two GT77 subfamilies, RRA and xyloglucanase-113 (Gille et al. 2009; Egelund et al. 2007).

To learn whether HRGP modification capacity or specificity is regulated during wall assembly, we prepared an extensive catalog of the *Chlamydomonas* homologs for the known HRGP modifying enzymes (Supplemental Table S22). A majority of the HRGP-modification-related genes turned out to be up-regulated by the g-lysin- or EZ-program (6/17 P4Hs, and 15/16 HRGP-O-glycosylation enzymes), of which four P4Hs and four O-glycosylation genes are specifically up-regulated during zygote development. In contrast, we did not find a concerted up-regulation of N-glycosylation-related genes by the g-lysin- nor EZ-program (6/31 are up-regulated, Supplemental Table S23).
*Chlamydomonas* HRGPs are also decorated by branched glycans (ex. GP1, Ferris et al. 2001), whose core structure consists of Hyp₁-Ara₁₋₃ that are mono- or di-galatosylated. By contrast, the complex glycans associated with AGPs in land plants consist of a Hyp₁-Gal₁₋₃ core, perhaps explaining the absence of any *Chlamydomonas* homologs to currently identified glycosyltransferases associated with AGP-glycosylation that belong to the GT14, GT29 and GT31 families (Liang et al. 2013; Geshi et al. 2013; Nguema-Ona 2014; Ogawa-Ohnishi and Matsubayashi 2015).

EZ-specific and g-lysin-induced clusters include many putative glycosyltransferases, mainly of GT47 and GT90 members. Characterized GT47 members of *Arabidopsis* are reported to add sugars to glycans during pectin and hemicellulose biosynthesis (Geshi et al. 2011); however, pectin or hemicellulose has not been found in the *Chlamydomonas* vegetative wall. Phylogenetic analysis of GT47 family including 47 *Chlamydomonas* and 39 *Arabidopsis* members shows a peculiar dichotomy wherein all *Chlamydomonas* members are related to one *Arabidopsis* member, At3g57630, which carries an EGF domain, and the remaining *Arabidopsis* members form separate clades (Supplemental Figure S7). Six GT47 genes are exclusively expressed in early zygotes, and three are found in g-lysin-inducible clusters (C8 and C24).

GT90 proteins are known as multi-functional xylosyltransferases on mannans when involved in capsule formation in fungi, and as serine-O-glucosyltransferases when involved in Notch signaling in Metazoa (Acar et al. 2008; Reilly et al. 2009), whereas the role of GT90 in plants has not been elucidated. GT90 phylogeny shows great diversity and *Chlamydomonas*-specific expansion in a number of clades in comparison to *Volvox* proteins (Supplemental Figure S8). Both EZ- and g-lysin-induced programs include at least one member of each clade (in red and in blue), suggesting the involvement of GT90 members in both VG- and Z-wall assemblies (Supplemental Figure S8).
Pellicle formation

Pellicle formation in liquid culture begins with wall-to-wall adherence by early-stage zygotes, followed by extensive build-up of HRGP fibers (Minami and Goodenough 1978; Suzuki et al. 2000). Sulfated polysaccharides or glycoproteins are known to be involved in cell-cell adhesion and to be critical for ECM structural integrity in Metazoans and multicellular Chlorophytes, including *Volvox carteri* (Misevic et al. 2004; Domozych and Domozych 2014). The EZ-core contains three genes encoding sulfotransferases with distant homology to Metazoan proteins involved in chondroitin and heparan sulfate biosynthesis; we therefore named them CSR (Carbohydrate Sulfotransferase-Related; Supplemental Table S22E). The CSR phylogeny shows seven clades conserved in Volvocacean algae, and two of the three zygote-specific clades are specifically expanded in the multicellular *Volvox*, possibly participating in cell-cell adhesion (Supplemental Figure S9).

The zygote wall components become cross-linked into a matrix that resists most solubilizing agents (e.g. SDS), and evidence for H$_2$O$_2$-mediated crosslinking and isodityrosine (IdT)-linkages in the *Chlamydomonas* vegetative and zygote wall has been presented in Waffenschmidt et al. (1993), where Tyr-Gly-Gly (YGG) was proposed as a IdT motif. In the EZ-core, we found many YGG motifs in MAW10 with a C-terminal transmembrane domain. A member of the candidate protein family involved in H$_2$O$_2$-mediated cross-linking, glyoxal/galactose oxidase (GOX), is among the most highly expressed genes in the EZ-core. This family has been implicated in cross-linking of plant cell wall constituents such as pectin (K. Sõla, E. Gilchrist; M-C Ralet, C. S, Mansfield and G. Haughn, personal communication) and in direct defense against pathogens via its expected catalysis of H$_2$O$_2$ generation (Zhao et al. 2013). GOX genes are present in *Chlamydomonas* as a large gene family of 19 members, five of which are up-regulated by the g-lysin- or EZ-program, suggesting their general involvement in cell wall assembly. An interesting feature unique to the Volvocacean GOX is the presence of N-terminal Pro-rich domains (15/19) with repeat patterns such as
PPSPX, PS, or PX, indicating that these GOX proteins are likely part of the cell wall network (Supplemental Table S22F). Phylogenetic analysis of the GOX family shows a dichotomy similar to the GT47 phylogeny, where *Arabidopsis* and *Chlamydomonas* have undergone independent expansion (Supplemental Figure S10).

**Support of metabolic demands of glycosylation precursors**

To support increased HRGP-secretion for zygote wall formation and reported extracellular callose accumulation (Catt 1979; Grief et al. 1987), the cytosolic hexose pool would be in great demand during zygote wall assembly. Hexoses must be activated via UDP or GMP nucleotidylation, and the resultant nucleotide sugars then feed glycosyltransferases (Seifert 2004). A total of 20 genes involved in nucleotide-sugar biosynthesis and 38 putative sugar transporters, mostly in triose-phosphate transporter/nucleotide-sugar transporter (TPT/NST) family, are found in the *Chlamydomonas* genome (Supplemental Tables S24-S25), allowing us to propose a model for the flow of sugars for glycosylation reactions (Figure 7). A notable difference between the *Chlamydomonas* model and its plant counterpart is the absence of GMD/GER and AXS to produce GDP-L-Fuc and UDP-D-Api, since homologs of these genes have not been identified (Supplemental Table S24). According to our model, the sugar flow via UDP-D-Glc up to UDP-L-Ara would become highly activated during early zygote development by up-regulation of both NT-sugar transporters (13/34) at non-plastidic locations and sugar-converting enzymes for UDP-L-Ara (red arrows in Figure 7). Of the NST clades, we noticed size expansion in the clades D, E, and H in both *Chlamydomonas* and *Arabidopsis*, where 12 genes (out of 22) are found to be up-regulated in the EZ condition and four genes are up-regulated by g-lysin treatment, suggesting that high demands of glycosylation for zygote wall formation may have driven the expansion of NST family in *Chlamydomonas* (Supplemental Table S25, Supplemental Figure S11).
In addition to the flow of hexose into HRGPs via glycosylation in the Golgi, *Chlamydomonas* zygotes have another major sink for the hexose pool, callose deposition, contributing up to 30% of all extractable sugars in the zygotic wall (Catt 1979). Two callose synthases and one callose synthase-like gene (GT48) that are exclusively expressed in early zygotes are likely enzymes for such a...
huge extracellular build-up (Supplemental Table S26). So far, only trace amounts of glucose have been detected in the *Chlamydomonas* VG-wall (Catt 1979).
Discussion

Our transcriptome analysis comparing wild type and gsp1 null zygotes shows that the EZ transcriptome is mostly, if not entirely, under the regulation of the GSM1/GSP1 homeobox heterodimer, positively at the transcriptional level for the zygote-specific genes, and negatively at the post-transcriptional level for the VG-wall specific genes. The combination of the two regulatory modules achieves differential regulation of the EZ-core/gL+EZ/gL-EZ gene sets, whose annotation points to molecular differences of the cell-type specific wall assemblies as a hallmark of cellular differentiation.

Non-guided clustering reveals crosstalk between the EZ and g-lysin-induced programs.

Our study design explicitly considers strain variability and multi-way comparison by hierarchical clustering of 13 different conditions and diverse genetic backgrounds. Since zygotes are naturally of mixed genetic make-ups, their study needs to consider genetic variation that may bias the analysis. Therefore, we aimed to identify reliable and robust sets of genes (such as the EZ-core) that warrant more thorough examination. Non-guided clustering provided opportunities to discover expected and unexpected crosstalk among the five transcriptional programs: plus and minus-specific gametogenesis, mating reaction-induced, g-lysin-induced, and EZ-specific. Such a blinded approach has documented that the EZ transcriptome includes a major contribution from the g-lysin-induced program but not the other mating-associated programs.

Regulatory network for the assembly of two wall types.

The VG- and Z-wall assembly programs expectedly share factors involved in HRGP-modification and secretory routes for HRGPs, since both walls consist mainly of HRGPs. Such shared factors are clearly represented by the gL+EZ genes that are expressed equally well during the VG- and Z-wall assembly. Following our prediction, the gL+EZ cluster includes 42 ER/secretion-related
genes and 26 HRGP-modifying- and nucleotide-sugar metabolic enzyme-
encoding genes, but no obvious HRGP-encoding genes (Figure 5, Supplemental
Table S16).

Our study defines the EZ-core and gL-EZ genes that are exclusively expressed
during the Z-wall or VG-wall assembly respectively, reflecting the fact that the VG
wall and the Z wall are very different in their organization and properties.
Accordingly, the EZ-core and gL-EZ genes include distinct sets of HRGP-
encoding genes -- for example, GPI-anchored HRGPs for the EZ-core and
pherophorins for the gL-EZ -- indicating that they are indeed the Z- and VG-wall-
specific cohorts.

Interaction between EZ- and g-lysin programs is highly selective, affecting about
half of the g-lysin-induced genes. Follow-up analysis of three g-lysin-inducible
promoters provided compelling evidence for persistent expression from these
promoters in early zygotes; we also document a subsequent post-transcriptional
repression of > 150 genes that is GSM1/GSP1-dependent. Hence the early
zygote appears to be programmed to discard the transcripts encoding VG-wall
structural proteins that might interfere with the Z-wall assembly, but to maintain
production of enzymes for glycosylation and other modifications for cell wall
proteins employed in both VG-wall and Z-wall construction.

The post-transcriptional mechanisms employed by the early zygote to recognize
and discard VG-wall-specific transcripts remain to be investigated. One
possibility is suggested by the presence of Tudor Staphylococcal Nuclease
(TSN) in the gL+EZ set (Supplemental Table S19). The TSN homolog in
_Arabidopsis_ has been reported to be involved in regulating the stability of the
mRNAs that are targeted to the ER for co-translational translocation (Sundström
et al. 2009); more recently, its regulatory action has been found to depend on its
nuclease domain and on selective decapping of mRNAs at the stress granule
(Gutierrez-Beltran et al. 2015), where TSN operates together with a decapping
enzyme like Dcp1 and a 5'→3' exonuclease like Xm1, proteins that are also conserved in *Chlamydomonas*.

**Departure from unicellular contexts**

In flowering plants, the haploid-to-diploid transition is followed by sporophyte development, giving rise to multiple tissue and organ types, whereas the haploid phase is dramatically shortened, entailing only a few mitotic divisions. In the moss *Physcomitrella patens*, key players involved in the haploid-to-diploid transition include KNOX-family and their putative binding partners BELL-family transcription factors, necessary for spore maturation and suppression of the haploid program, and Clf/Fie as a component of the PRC2 that suppresses the diploid program (Sakakibara et al. 2008; 2013; Mosquina et al., 2009; Okano et al. 2009; Horst et al., 2015). Since GSM1 is the only KNOX-family transcription factor in *Chlamydomonas* (Lee et al. 2008), it is of great interest to understand how the GSM1/GSP1-dependent program evolved into the KNOX-dependent regulation of land-plant haploid-to-diploid transition. A recent study has shown that the two divergent KNOX subfamilies in land plants function in antagonistic roles: type II promotes differentiation programs whereas type I counteracts type II by promoting the undifferentiated state of stem cells (Furumizu et al. 2015). A plausible scenario, therefore, is that the type II KNOX retains the ancestral differentiation-triggering function of GSM1 and that type I evolved to delay the differentiation program leading to meiosis, thereby allowing the emergence of a multicellular diploid phase. Recently, Frangedakis et al. (2016) reported that the type I and II KNOX already diversified in charophyte algae before the evolution of multicellular diploid generation. The dual-action mechanisms of the GSM1/GSP1 heterodimer described in this study, controlling both positive gene expression and post-transcriptional mRNA degradation, may have opened up an evolutionary path for the KNOX-based GRN to become bifurcated by sub-functionalization following gene duplication.

**Materials and Methods**
**Strains and culture conditions**

Details of strains used in RNA-seq are found in Supplemental Table S1. CC-numbered strains are available in the Chlamydomonas Resource Center at University of Minnesota (www.chlamycollection.org). JL28 (nic7; mt-) was generated by mating between CC-125 and CC-2663 (nic7; mt-). Pseudo-zygote strains are described in Lee et al. (2008). bp31/bp31C strains are described in Nishimura et al. (2012). Plate cultures were maintained in continuous light on Tris-acetate-phosphate (TAP) medium (Harris, 1989) solidified with 1.5% Bacto agar. Liquid cultures were inoculated at 5 X 10^4 cells/mL concentration and maintained in continuous light on TAP medium until harvest at 5 X 10^6 cells/mL.

**Sample preparation and RNA-seq library construction**

Vegetative cells in logarithmic growth were obtained in TAP-liquid culture at 22 °C in constant light under 40 μmoles/m^2/sec, and harvested at 1 X 10^6 cells/mL. Early zygotes were obtained from mating mixtures of fully differentiated gametes prepared from 7-day old TAP-plate culture, and incubated in NF-TAP for 5 hr until more than 90% of cells became active gametes. Once plus and minus gametes were mixed in equal number (1E+7 cells), samples were harvested at 45 min and 2 hr after mixing and combined during the RNA-seq library preparation considering slightly low mating efficiencies (50-80% at 1 hr). Pseudozygotes from JL-N7RM-110 (T-GSM1, mt+[mt-;imp11]) and JL-M2MP-10 (T-GSM1, T-GSP1, mt-;mid-2) were obtained from 7-day old TAP-plate culture, incubated in NF-TAP for 24 hr until the majority of cells have formed large aggregates (pellicle) in the medium.

**RNA extraction and sequencing**

Total and poly(A)+ RNAs were extracted by methods described earlier (Lee et al., 2008). Poly(A)+ RNAs were reverse-transcribed using random primers and dT(15) primers. Resulting cDNAs were sheared into 300- to 400-bp fragments, and used to prepare six mRNA sequencing libraries according to the
manufacturer's protocol (cat# RS-930-1001, Illumina). Libraries for TAP/EZ/PZ samples were sequenced on the Illumina GAIIx for 76 bp single-end cycle except TAP duplicate #2 for 36 bp single reads, following the manufacturer's standard cluster generation and sequencing protocol. bp31/bp31C zygote samples were sequenced for 36 bp single reads.

**RNA-seq analysis**

All the new and published RNA-seq reads (Rachel et al., 2010, Ning et al., 2013) were mapped with Tophat2 (Kim et al. 2013) to the v5 genome combined with the mating-type minus sequence obtained from CC-2290 (genbank # GU814015, Ferris et al., 2010) with the allowance of 3 mismatches. The mapping details are given in the Supplemental Table S2. Normalized FPKM values were calculated accordingly to our modified V5.3 annotation that contains extra mating-type minus genes defined in GU814015 by cufflinks ver. 2.11 (cufflinks -l 2000 -no-faux -G, Trapnell et al., 2013), which performs correction for multihits by dividing a pool of multihits into corresponding genes according to their unique hit distribution. Following our analysis, an updated annotation (V5.5) was released and also used to calculate FPKMs of our mapped reads; however we found no significant effect in our zygotic transcriptome analysis and decided to utilize V5.3 information in this study (see Supplemental Note S2 for details.)

**Hierarchical clustering analysis**

Following FPKM calculation, we combined 33 samples of FPKM data (excluding bp31/bp31C zygote samples) representing 13 conditions in triplicates or duplicates (for 6 conditions), and performed clustering analysis following the method described in Ning et al. (2013). Prior to the analysis, 5656/18823 (as Cluster 0) gene models were discarded by the low-expression criterion (32/33 with less than 4 FPKM). The rest with 33 data points were averaged per 13 conditions; and normalized by converting the average FPKM values into percentage values of expression over the sum of 13 average FPKM values. The resulting percentage values were used to generate 50 clusters by Ward's
hierarchical clustering algorithm applied to Euclidean distances using Genomics Workbench v4 (see Supplemental Table S3 for clustering results, Supplemental Table S4 for clustering statistics, and Supplemental Table S5 for cluster descriptions based on their bias patterns).

In silico promoter motif analysis
Sequences combining 1500 bp upstream and 500 bp downstream to the transcription start site were compiled using a perl script (http://github.com/minillinim/gff2fasta), and assigned to 51 bins matching their co-expression clusters. Statistical enrichment of 10-base motifs was analyzed by AmadeusPBM 1.0 (Linhart et al., 2008) with default parameters.

Gametogenesis and mating
Gametogenesis was induced as described (Ferris and Goodenough, 1997). Unless otherwise stated, cultures were grown on TAP plates for 6 days, resuspended in TAP, and aliquoted into two tubes, one for gametogenesis and the other for control. Gamete samples were washed once and resuspended in nitrogen free-TAP (NF-TAP) to a final 5 \times 10^7 cells/mL, while control samples were washed and resuspended in TAP to the same concentration. The resulting liquid cultures were incubated under continuous light in 24-well plates. Four hr later, the cells in NF-TAP displayed vibrant agglutinating activity when mixed with tester gametes of opposite mating type, the sign of gametic differentiation, but not the cells in TAP medium. Mating efficiency was calculated based on QFC formation as previously described (Goodenough et al., 1993). Zygotes were produced by mating equal numbers of plus and minus gametes.

Transformation by the glass bead method
To introduce promoter-reporter constructs, nic7 mutant cells (\textit{mt}-) were grown in liquid TAP containing 4 \mu g/mL nicotinamide until they reached a density of 5 \times 10^6 cells/mL. Approximately 1 \times 10^8 cells were harvested and incubated with 2.5 mL g-lysin for 1-2 hr until >90% had lost cell walls, as monitored by the sensitivity
of cells to lysis in 0.02% NP-40. Protoplasts were harvested in a 10-mL glass tube and resuspended in 300 μL TAP, to which were added 0.3 g glass beads (0.5 mm) and 100 μL 20% PEG-8000 containing 0.5 - 2 μg linearized transforming DNA derived from pNic7.9 (Ferris, 1995). Cells were vortexed for 25 sec at the top setting of a Vortex-Genie mixer, washed twice with 10 mL TAP, resuspended in 300 μL TAP, mixed with 2.5 mL 0.5% Top Agar in TAP (melted at 45°C), and poured onto TAP supplemented with 15 ppm of 3-acetylpyridine (Sigma), which is toxic to nicotinamide requiring cells (Harris, 1989). Colonies of Nic+ transformants became visible after 3-5 days. g-lysin, which removes *Chlamydomonas* cell walls (Kinoshita et al., 1992), was prepared as described in Lee et al. (2008).

**Real-time qRT-PCR**

To analyze the relative abundance of transcripts using qRT-PCR, DNase-treated 5 μg of total RNA was reverse-transcribed with Superscript II RT-PCR kit (Invitrogen). Primers were designed to produce PCR amplicon lengths of 100–150 bp. qRT–PCR was performed in optical 96-well plates using Bio-Rad CFX96 Real-Time PCR systems. Reactions were performed in a final volume of 20 μL containing 10 μL of 23 SYBR Green Master Mix, 0.5 μM each primer, and 10 ng of cDNA. PCR conditions were as follows: 30 sec at 95°C, followed by 45 cycles of 5 sec at 95°C and 30 sec at 60°C. Fluorescence threshold data (Ct) were analyzed using Bio-Rad CFX Manager software (version 2.0). Quantitative reactions were done in duplicate and averaged. Relative expression levels in each cDNA sample were normalized to the *RACK1* reference gene under the same conditions. Three biological replicates were averaged for quantitative analysis. Primers used in Real-time qPCR are found in the Supplemental Table S12.

**Cloning of Promoters**

To isolate promoters of the genes of interests, genomic DNAs of CC-125 strain were used in PCR amplification. 5' primers are tagged by XbaI and 3' primers are...
tagged by EcoRI. PCR amplicons were cloned into pGEM-T Easy vector and fully sequenced for verification. Resulting XbaI and EcoRI fragments were ligated into XbaI and EcoRI of linearized pNIC7-GL that contains a Gaussian Luciferase ORF derived from pHsp70A/RbcS2-cgLuc (Fuhrman et al., 2004) in pNic7.9 (Ferris, 1995). The resulting constructs were used for transformation into nic7 mutant strains.

**Promoter-luciferase assays**

Promoter constructs were transformed into mt- Chlamydomonas cells (nic7, mt-, JL28). Randomly picked 24 colonies were screened for secreted luciferase activity with or without g-lysin treatment for C24 and C44 promoters and with or without mating to plus gametes for C43 and C50 promoters. Two representatives were chosen for further studies. Luciferase assays were done under the following conditions to test their effects on promoter activation. Luciferase-containing transformants were prepared as gametes and mixed with CC-125 or bp31 as mating partners. Hourly accumulation of luciferase activity in fresh culture media was measured along with unmated gametes.

**Annotation of EZ-specific and g-lysin-inducible genes**

Collected protein sequences were initially analyzed by three approaches: 1) reciprocal best hits with Arabidopsis proteins for retrieving available functional annotation from Arabidopsis orthologs (TAIR v10); 2) Interpro (http://www.ebi.ac.uk/interpro/, Finn et al., 2017) search for collecting functional/homology domain information; and 3) PredAlgo (https://giavap-genomes.ibpc.fr/cgi-bin/predalgodb.perl?page=main, Tardif et al., 2012) and/or TargetP (http://www.cbs.dtu.dk/services/TargetP/, Emanuelsson et al., 2007) search for N-terminal leader sequences. This initial analysis allowed us to choose five major categories for further investigation: Nucleotide-sugar biosynthesis/transport (Seifert 2004), Cell wall protein-modifying enzymes (Kim and Brandizzi, 2016; Hijazi et al., 2014; Nguema-Ona et al., 2014; Atmodjo et al., 2013; McCarthy et al. 2014), ER-translocon/ERAD/N-glycosylation (Ruiz-May et
al., 2012; Mathieu-Rivet et al., 2013), Organellar genome replication/recombination/repair (Day and Madesis, 2007), and Secretion/endomembrane system (Vernoud et al., 2003; Richter et al., 2007; Sanderfoot 2007; Fujimoto and Ueda, 2012; Fendrych et al., 2010). For each category, we compiled the lists of known molecular players from recent primary research papers and reviews, mostly focused on Arabidopsis/plant systems. The known Arabidopsis proteins were used to identify Chlamydomonas homologs by reciprocal best hits. If a best-hit Chlamydomonas protein showed an E-value of larger than 1E-10, or many proteins showed similar homology, phylogenetic analysis of all the homologs with similar domain structure was performed to define evolutionarily conserved clades. In case of multigene families, homologous sequences that belong to a family were collected by Hmmer3 search of the key homology domain against the non-redundant proteomes in relevant species. The final genome-wide collection of Chlamydomonas genes related to cell wall is given in the Supplemental Tables S19, S20, S21, S22, S24, S25, and S26. Supplemental Table S31 is made available for cross-referencing of EZ-core, g-lysin-induced genes, and gamete-specific genes with the annotation and the compiled transcriptome data.

Phylogenetic tree construction

For each target family, family member protein sequences were collected by a genome-wide search by HMMer3 using homology domain motifs specific to the target families retrieved from the Pfam 3.0 database or generated from our own domain alignments. Family searches were done in non-redundant predicted protein sets of the selected species listed in Supplemental Table S27. Collected protein sequences were aligned using MAFFT v7.017 (Katoh et al., 2002).

Resulting alignments were manually inspected for refinement. A consensus Neighbor-Joining tree was initially constructed with 500 bootstrapping. If clade structure was not satisfactory in the N-J tree, MrBayes V3.2.6 (Huelsenbeck and Ronquist, 2001) was used for constructing a second phylogeny with a best-fit model searched by Prottest v.3.4.2 (Darriba et al., 2011) or with the WAG model of protein evolution. The BMCMC was run mostly for $10^6$ generations, sampling
every 100 generations and discarding the first 2,000 samples as burn-in. The remaining 8000 samples were used to construct the 25% majority-rule consensus phylogeny.

Accession numbers
RNA-seq reads were submitted to NCBI with accession# GEO91400 except bp31/bp31C zygotes data available at http://dandelion.liveholonics.com/pothos/Nishimura_GSP1/ for download. Additional data accession numbers for the published datasets and sequencing statistics are shown in Supplemental Table S2.

Supplemental Data
Supplemental Note S1. Gene model consideration between V5.3 and V5.5
Supplemental Note S2. Cautionary tales in FPKM counting
Supplemental Table S1. Details of the samples used in the transcriptome analysis.
Supplemental Table S2. RNA-seq mapping statistics by Tophat2.
Supplemental Table S3. Average FPKM values and cluster ID of 13167 gene models analyzed by hierarchical clustering.
Supplemental Table S4. Statistics of FPKM estimates among the 50 hierarchical clusters collected in this study.
Supplemental Table S5. Description of the co-expression patterns of the 50 final clusters.
Supplemental Table S6. FPKM estimates and differential expression analysis of the 253 EZ-core genes.
Supplemental Table S7. Detailed annotation of the EZ-core genes.
Supplemental Table S8. DESeq2 results for differential expression analysis between the bp31C- and bp31C-zygotes.
Supplemental Table S9. Significantly up-regulated genes (>2 fold) in bp31C- over bp31C-zygotes by DESeq2 analysis (cutoff FDR<0.05).
Supplemental Table S10. Significantly down-regulated genes (>2 fold) in bp31C-
over bp31-zygotes by DESeq2 analysis (cutoff FDR<0.05).

Supplemental Table S11. Nonparametric Wilcoxon test for the difference in the bp31C/bp31 ratio distribution among clusters

Supplemental Table S12. Primers used for qPCR and promoter cloning.

Supplemental Table S13. List of genes having at least one of the three ZYRE motifs identified in this study.

Supplemental Table S14. Statistics of TGAC motif occurrence among the clusters

Supplemental Table S15. FPKM estimates of the known gamete-specific genes

Supplemental Table S16. Detailed annotation of the L+EZ (C44) genes.

Supplemental Table S17. Detailed annotation of the L-EZ (C24) genes.

Supplemental Table S18. Distribution of functional categories among EZ-core, L+EZ, and L-EZ genes.

Supplemental Table S19. Details of the MAW gene family.

Supplemental Table S20. Chlamydomonas genes involved in GPI-anchor biosynthesis.

Supplemental Table S21. Pherophorin-encoding genes in Chlamydomonas.

Supplemental Table S22. Chlamydomonas genes involved in cell wall protein maturation.

Supplemental Table S23. Chlamydomonas genes involved in N-glycosylation.

Supplemental Table S24. Chlamydomonas genes involved in nucleotide-sugar metabolism for cell wall biosynthesis.

Supplemental Table S25. Chlamydomonas genes encoding nucleotide-sugar transporters.

Supplemental Table S26. Chlamydomonas genes encoding plasma membrane-localized glycosyl transferase.

Supplemental Table S27. Genomic resources used for the phylogenetic analysis of the target gene families.

Supplemental Table S28. Comparison of the gene models, version 5.3 and version 5.5.

Supplemental Table S29. Analysis of correlation in FPKM estimates between
V5.3 model- and V5.5 model-based analysis.

Supplemental Table S30. Comparison of the >4 fold up-regulated genes in EZ/TAP using V5.5 models-based FPKM estimates to the results using V5.3 models.

Supplemental Table S31. The Chlamydomonas haploid-to-diploid transition (HDT) transcriptome

Supplemental Figure S1. qRT-PCR analysis of EZ-core gene expression.

Supplemental Figure S2. TGAC motifs are most significantly enriched among the early zygote-specific genes.

Supplemental Figure S3. Distribution of bp31C/bp31 FPKM ratio among the EZ-specific (C33, C43, and C50) and g-lysin-induced (C24 and C44) gene clusters.

Supplemental Figure S4. Structure of the genes that were analyzed by the promoter-reporter assay.

Supplemental Figure S5. Analysis of g-lysin inducible luciferase expression of the selected transgenic strains harboring a g-lysin-inducible promoter-luciferase construct.

Supplemental Figure S6. MAW domain sequence alignment.

Supplemental Figure S7. A phylogenetic analysis of the Exostosin (GT47) family in Arabidopsis and Chlamydomonas.

Supplemental Figure S8. A phylogenetic analysis of the GT90 family in green plants.

Supplemental Figure S9. A phylogenetic analysis of the CSR (Carbohydrate Sulfurtransferase Related) family.

Supplemental Figure S10. A phylogenetic analysis of the GOX (glyoxal oxidase) family by the Neighbor-Joining method.

Supplemental Figure S11. A phylogenetic analysis of the TPT/NST family using Arabidopsis and Chlamydomonas members.

Supplemental Figure S12. RNA-seq results based on V5.3 models show little difference from the results based on V5.5 models.
Acknowledgements

We are grateful to Sean Gallaher and Sabeeha Merchant for supporting RNA-seq reads and data-handling in her laboratory. We thank David Casero and Matteo Pellegrini for their support for next-gen sequencing and sequence analysis. We thank William Snell for sharing his RNA-seq reads prior to the publication. We thank Masaki Odahara for technical assistance in cDNA sequencing library preparation. Also tremendous thanks to JGI for providing V5.3/V5.5 assembly and annotation. We also thank George Haughn for reading manuscripts and sharing unpublished results.
Figure 1. A gene-regulatory network model of sexual development in *Chlamydomonas reinhardtii* consisting of at least six regulatory programs; gray bars depict the duration of each program. The upper bar shows physiological/cellular events providing cues for the programs. Arrows denote the consequence of specific events/factors as induction or repression. Known molecular players are shown as boxes, blue for *plus*-specific and red for *minus*-specific. NS/GAM: N-starvation induced/gametogenesis program; PG/MG: *plus*- and *minus*-specific gamete programs; MR: mating reaction-induced program (dibutyryl-cyclic AMP-inducible); gL: g-lysin-induced program; FD/HB: fusion-dependent/homeobox transcription factor-dependent program.

Figure 2. Identification of co-regulated genes in the early zygote transcriptome by hierarchical clustering analysis. Relative expression patterns among 13 conditions are shown for selected gene clusters representing the early zygote-specific clusters (*C*33, *C*43, *C*50, and *C*10) and the g-lysin-induced clusters (*C*44 and *C*24). Box plots show distribution of percentage expression values per condition, calculated relative to the sum of FPKM values in all 13 conditions per gene. Boxes show the median and the 25th and 75th percentile. Whiskers indicate the 10th and 90th percentiles. Details pertaining to the remaining clusters are found in Supplemental Tables S4 and S5.

Figure 3. GSM1/GSP1-dependent activation of EZ-core genes. (A) Relative abundance of transcripts in early stage zygotes to *minus* gametes was analyzed using qRT-PCR. Error bars: standard deviation of three biological replicates. g+: *plus* gametes; g-: *minus* gametes; Z.5, Z1, and Z2: 0.5 hr, 1 hr, and 2 hr after mixing gametes. Black bar: CC-621 (*mt*-) X CC-125 (*mt*+); Gray bar: CC-621 (*mt*-) X *bp31* (*mt*+). (B) Promoter-luciferase reporter assay of two EZ-core promoters. Fold-difference was calculated from cumulative luciferase-based promoter activity during the early zygote development.
Figure 4. Differential activation of g-lysin-induced genes during the early zygote development. (A) Relative abundance of transcripts in early stage zygotes to minus gametes was analyzed using qRT-PCR. Error bars: standard deviation of three biological replicates. g+: plus gametes; g-: minus gametes; Z.5, Z1, and Z2: 0.5 hr, 1 hr, and 2 hr after mixing gametes. Black bar: CC-621 (mt-) X CC-125 (mt+); Gray bar: CC-621 (mt-) X bp31 (mt+). (B) Promoter-luciferase reporter assay of two EZ-core promoters. Fold-difference was calculated from cumulative luciferase-based promoter activity during the early zygote development relative to samples without mating. Error bars: standard deviation of three biological replicates.

Figure 5. Distribution of annotated EZ-core/gL+EZ/gL-EZ genes in functional categories related to cellular differentiation (excluding the unknowns). The categories listed in the legend represent, in order, the categories in the pie charts starting from midnight and proceeding clockwise. Cell wall-related categories are marked by thick outlines. Left: EZ-core (n=204), Middle: gL+EZ (C44) (n=129), Right: gL-EZ (C24) (n=113). Number of genes in the top two categories are shown in pie charts. Annotation details are found in Supplemental tables S18, S19, and S20.

Figure 6. Identification of the MAW (Membrane-Anchored cell Wall protein) family encoding HRGPs with a C-terminal hydrophobic domain. (A) Domain structure of MAW family proteins in Chlamydomonas. (B) Bayesian phylogeny of MAW domains among Volvocacean MAW proteins. Predicted GPI-anchored MAW proteins are found in clades I and II. (C) Cys-rich MAW domain consensus sequence. Domain sequence alignment is found in Supplemental Figure S6. (D) Differential expression of MAW family genes in Chlamydomonas based on average FPKM values among nine conditions. Upper graph: EZ-specific genes. Lower graph: Non-EZ-specific genes. Due
to very low expression, CreMAW5 and CreMAW11 are not included in the graph.

Figure 7. Predicted flow of cytosolic hexoses into nucleotide-sugar metabolism in support of glycosyl transferases up-regulated during the zygote development. Cytosolic hexose pool is supplied by gluconeogenesis and export from the chloroplast. Arrows show directional or bi-directional conversions. Red arrows: predicted sugar flow changes in early zygotes, resulting in a large increase in available nucleotide-sugars both in the cytosol and in Golgi via NST-family transporters. Blue arrows: predicted sugar flow increase into GDP-D-Mannose and GDP-L-Galactose in the cytosol. Red metabolites: forms usable by glycosyl transferases. Shaded symbols represent genes. Brown: early zygote-specific or up-regulated (EZ or gL+EZ). Green: g-lysin induced (gL-EZ). Gray: found in other clusters. Details of these genes are found in Supplemental Tables S24 and S25. Putative nucleotide-sugar transporters in ER/Golgi are found in six classes (B-I excluding C and D classes known to be localized to the plasma membrane; Supplemental Table S26). Frc: Fructose, Man: Mannose, Glc: Glucose, Gal: Galactose, Rhm: Rhamnose, GlcA: Glucuronic acid, GalA: Galacturonic acid, Xyl: Xylose, Arap: Arabinose-pyranose, Araf: Arabinose-furanose.


Egelund J, Obel N, Ulvskov P, Geshi N, Pauly M, Bacic A, Petersen BL. (2007). Molecular characterization of two Arabidopsis thaliana glycosyltransferase mutants caf1 and caf2, which have reduced residual arabinose content in a polymer tightly


Minami SA, Goodenough UW. (1978). Novel glycopolyptide synthesis induced by gametic cell fusion in Chlamydomonas


