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, Nai Chun Shao, Saif-El-Din El Akkad, Takamasa Suzuki, Tetsuya Higashiyama, Eonseon Jin, and Jae-Hyeok Lee

Supplemental Notes

**S1. Gene model consideration between V5.3 and V5.5**

According to Gallagher et al. (2015), gene models can significantly affect expression estimates mainly due to annotation accuracy; since non-concordant reads to the given reference will be excluded with -G option by cufflinks. Since the V5.5 gene model was released (Phytozome release 11) after our analysis had been completed using V5.3 gene models, we tested whether any significant difference was made by using different gene models (detailed V5.3 and V5.5 differences are provided in Supplemental Table S28).

**Our revised gene model from V5.3**  
1) 67 gene models residing in the MTL- locus were added. (Scaffold89 models)  
2) 15 gene models were discarded as a poorly supported model overlapped with other models.  
3) 3 ribosomal RNA genes were excluded in chromosome 4 near the telomere.  
4) The *PSBW* gene was added and labeled by Cre11.g475200, which is not annotated in both V5.3 and V5.5 models.
As a result, 18,756 (plus 67 for MTL-) gene models are present in our annotation file.

**V5.3 VS V5.5 models**
The major consolidation between V5.5 (17,741 transcripts) and V5.3 (18,756 transcripts) entailed merging overlapping gene models into a single locus as alternative transcripts and by splitting combined gene models into separate loci, affecting 1219 (discarded) and 204 (split), respectively.

**FPKM comparison**
A total of 34 samples were re-analyzed according to V5.5 models, and V5.5- FPKM estimates and TAP/EZ fold induction were compared. Overall correlation between V5.3- and V5.5-based FPKM values lies between 0.928-0.984 except MA samples between 0.837-0.913 (Supplemental Table S29). Based on >4 fold cut-off for EZ/TAP, 1030 gene models were selected using V5.5 FPKM values, corresponding to 1107 models in V5.3 gene models. Differences were found mostly in the overlapping/redundant models of V5.3 that were discarded in V5.5 (Supplemental Figure S12 and Table S30). Two trends were found; 1) In 42 pairs, one member showed the same up-regulation in both versions and the other showed very low FPKM values (except one pair with higher FPKM, indicating a better model); and 2) 58 cases exhibited FPKM values split into two or more gene models, although all the split V5.3 gene models showed >2 fold up-regulation. The remaining 1007 cases exhibited only an average 0.72% change in the EZ/TAP fold ratio and no case showed <4 fold in EZ/TAP. Since our analysis selected EZ-enriched genes if one of the overlapping gene models fits the criteria, it is very unlikely that our analysis based on V5.3 has missed any true EZ-enriched gene that V5.5-based analysis would select otherwise.

**S2. Cautionary tales in FPKM counting**
A recent transcriptome study by Lopez et al. (2015) reported a collection of 627 genes as early zygote-specific genes, which excluded only EZY15 and 17 among the Kubo set (EZY3-23) by finding significant expression at other times. Its relatively large number of EZ genes compared to our 253 EZ-core genes is explained by the following factors: 1) minimal expression cut-off (144 genes of <2 FPKM in their EZ condition), 2) whether to consider g-lys induction as part of the early-zygote transcriptome (49 and 11 found in two g-lys inducible clusters, C44 and C24, of our analysis, respectively), and 3) multiple transcripts for the same gene (30 genes). In addition, we noticed all three ZSP2 repeats in the Lopez set but our EZ-core contains only one. We therefore investigated whether systemic counting errors occur among near-identical genes. If so, it can potentially result in overestimating the number of differentially regulated genes.

Problems with cufflinks default parameters
When default parameters are used in alignments allowing multiple-hits (which is also default), a single read can be mapped onto multiple loci/transcript models. FPKM counting using such alignments may cause problems by counting all those mapped reads as individual transcripts analyzed independently. Optional correction tools implemented in the updated cufflinks (newer than 2.11) seem to manage those redundant reads into separate models, actually preventing over-estimation of FPKMs due to multiple counting.

ZSP2 copies exemplify an expected scenario. ZSP2 is found in three contiguous copies in head-to-head and tail-to-tail orientation. The g15252 gene and the identical g15253/g15256 pair have 98.5% identity in coding sequence. The table below shows the over-estimation that occurs when the default options were used.

<table>
<thead>
<tr>
<th>Cufflinks options</th>
<th>g15252</th>
<th>g15253</th>
<th>g15256</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>5211.66</td>
<td>3535.75</td>
<td>3620.61</td>
</tr>
<tr>
<td>-u (multi-mapped)</td>
<td>2329.11</td>
<td>31.08</td>
<td>0.03</td>
</tr>
<tr>
<td>-b (fragment-bias)</td>
<td>2136.09</td>
<td>28.85</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Since counting of multiple hits becomes a problem, the same problem is expected in counting alternative transcripts. When the provided annotation (e.g. V5.3) contains multiple transcript models per gene, under-expressed alleles are potentially overestimated in the default setting.

References


Supplemental Figures

Joo et al. Supplemental Figure S1

**Supplemental Figure S1.** qRT-PCR analysis of EZ-core gene expression. Relative abundance of transcripts in early stage zygotes to minus gametes was analyzed using qRT-PCR. Error bars: standard deviation of two 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+).
Supplemental Figure S2. TGAC motifs are most significantly enriched among the early zygote-specific genes. (A) Identification of 10-mer motifs significantly enriched among the promoters of C33, C43, and C50 genes in comparison to all the other genes using AmadeusPBM 1.0 (Linhart et al., 2008). The three motifs identified share [t]GAC sequences once or twice. (B) Location of the Motif1 in the 2000 bp promoter domains is significantly biased to between −750 and 0 in all the promoter pools from C33, C43, and C50. Graph on the right shows the distribution of locations. Yellow: C33; blue: C43; purple: C50. (C) Occurrence of the three motifs per gene that contains at least one of the motifs was compiled and the average occurrence per cluster was calculated to see if the motifs tend to be found in multiples. Significant differences were found in C10, C18, and C30 as well as in the EZ-specific clusters used in the motif identification. C10 is EZ-enriched, C18 is MA-exclusive, and C30 is PZ-exclusive. Tips of diamonds indicate the upper and lower 95% of the group. Red dotted line shows average 1.33 motifs per gene. Detailed statistics are provided in the Supplemental Table S35.
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. Red horizontal line shows 0.896, the median ratio of the 'others' collection (n=11,838).
Supplemental Figure S4. Diagram of gene structure for the genes analyzed by the promoter-reporter assay. Cloned promoter locations are indicated as light-green ribbons, which are either delimited by the neighboring gene or a 500 bp segment upstream from the predicted ATG start codon if the neighboring gene is within 500 bp distance. Scale in base is shown at the top of each diagram.
Supplemental Figure S5. Analysis of g-lys inducible luciferase expression of the selected transgenic strains harboring a g-lys-inducible promoter-luciferase construct. For each promoter construct, two independent transgenic lines were selected as representatives showing expected g-lys inducible expression. Luciferase activity in culture media was measured after one hr in the given conditions. Some of the strains were analyzed twice for verification purpose. PHC19 and SEC61G belong to C24, and AraGT1 and RHM1 belong to C44. The VSP3 promoter construct was transformed but we failed to obtain strains showing significant g-lys-inducible expression.
Supplemental Figure S6. MAW domain sequence alignment. 45 MAW domain sequences extracted from 23 MAW homologs were aligned by the MAFFT algorithm. The affix after the gene name indicates the order of occurrence in the protein. All members with predicted GPI-anchor motif contain single MAW domain. Additional information of the genes in the alignment is found in Supplemental Table S22.
Supplemental Figure S7. A phylogenetic analysis of the Exostosin (GT47) family in *Arabidopsis* and *Chlamydomonas*. Clade nomenclature of *Arabidopsis* members follows Geshi et al. (2011). **A**: MUR3 (deficient in fucose, xylosyl
galactotransferase on xyloglucan, no transmembrane domain), B: ARAD1 (deficient in arabinan/pectin), C: XGD1 (xylosyltransferase on xylogalacturonan), C': not characterized, C1: At4g38040 only, D: IRX10 (deficient in glucuronic acid of xylan/pectin) and IRX7 (deficient in glucuronic acid of glucuronoxylan/hemicellulose in secondary wall), E: At1g21480 only, and F: At3g57630 only. The consensus tree was constructed from 8000 MCMC sampled trees by MrBayes 3.3 with the WAG+G protein evolution model.
Supplemental Figure S8. A phylogenetic analysis of the GT90 family in green plants. Five groups (A-E) are tentatively defined, of which only the A group is
shared throughout Viridiplantae. The estimated number of conserved clades with at least one *Chlamydomonas* and one *Volvox* homologs is given in bracket. Bold face gene IDs in black and red indicate g-lysin-induced and EZ-specific members, respectively. The consensus tree was constructed from 1968 of the 4370 MCMC sampled trees by MrBayes 3.3 with the LG+G protein evolution model.
Supplemental Figure S9. A phylogenetic analysis of the CSR (Carbohydrate Sulfurtransferase Related) family. The CSR family is highly diversified in Volvocaceae but absent in most green algal and plant lineages, except a single *Micromonas* homolog (Mp109625). Volvocacean CSR family members are found in eight conserved clades that are combined into three groups (A - C). All three *Chlamydomonas* genes in the group A are exclusively expressed in early zygotes, suggesting their specialized role in the zygote wall, possibly for providing adhesiveness to the zygotic wall. In the group A, A1 and A3 show *Volvox*-specific expansion. Animal homologs (HsCarboSulfotrans15 from Human and DromeQ9V3L1 from Fly) are included as outgroups. The consensus tree was constructed from 1300 MCMC sampled trees by MrBayes 3.3 with the WAG+G protein evolution model.
Supplemental Figure S10. A phylogenetic analysis of the GOX (glyoxal oxidase) family by the Neighbor-Joining method. The percentages of replicate trees supporting the cluster of taxa in the bootstrap test (1000 replicates) are shown next to the branches. Volvocine algae have five conserved clades, whereas Arabidopsis, Klebsormidium and Chlorella species possess only one clade of their own (green, yellow, and orange) that is not related to any of the five volvocine clades.
Supplemental Figure S11. A phylogenetic analysis of the TPT/NST family using *Arabidopsis* and *Chlamydomonas* members. Six NST clades, a single TPT clade (A), and a single amino-acid transporter clade (D) are defined. In cases where the substrate-specificity of any NST homolog is known, it is given below the clade name. Clade name and predicted/characterized localization information are taken from Reyes and Orellana (2008) and Rautengarten et al. (2014). AT5G04160 in the clade B has been characterized as an UDP-GlcA transporter by Saez-Aguayo et al., (2017). The consensus tree was constructed from 3733 MCMC sampled trees by MrBayes 3.3 with the WAG+G protein evolution model.
Figure S12. RNA-seq results based on V5.3 show little difference from the results based on V5.5 models. The comparison points out that all the different results between the studies based on the V5.3 and V5.5 models are due to split or overlapping gene models. Further examination of those overlapping and split models indicates that the V5.3-based analysis is unlikely to miss EZ-upregulated genes selected only by V5.5-based analysis. A full description of these results is found in Supplemental Note S2. The dataset included in the graph is found in Supplemental Table S6.