Article title: Evolution of heterodimerizing TALE homeobox transcription factors as a developmental mechanism for the haploid-to-diploid transition.
Authors: Sunjoo Joo, Ming Hsiu Wang, Gary Lui, Jenny Lee, Andrew Barnas, Eunsoo Kim, Sebastian Sudek, Alexandra Z. Worden, and Jae-Hyeok Lee.

Supplemental Methods S1-S4
Method S1. Collecting TALE homeobox protein sequences
Twenty six available algal genomes, and 16 additional transcriptomes from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) dataset (Keeling et al., 2014) (Details in Table S1), were searched by Hmmer for the HD/HD_KN (PF00046/PF05920) and by BLASTP/TBLASTN using published algal homeobox sequences to collect TALE homeobox. Homeodomain sequences were extracted from the collection and aligned using the MAFFT algorithm. Most TALE proteins were identified by the three amino-acid insertions between positions 23/24 in the 60 amino acid-long typical homeodomain (Bürglin, 1997). More divergent sequences with extensive gaps in the alignment could not be ascertained for their classes, therefore, we used secondary structure modeling against the 3D structure database at the SWISS-MODEL site (https://swissmodel.expasy.org/interactive) to evaluate whether the typical TALE homeodomain structure is retained and whether the introduced indels preserve or distort the known homeodomain structure (Biasini et al., 2014). After excluding 6 sequences of high similarity in order to prevent oversampling of a particular taxa, the resulting TALE alignment of 96 TALE homeodomains contained one conserved insertion in the TALE loop between sites b/c among the GLX and Mam-A sequences in Chlorophyta, and two singleton insertions elsewhere that were excluded for phylogenetic reconstructions in the final alignment (S1 Fig). The final alignment includes four Arabidopsis thaliana (STM, BP, KNAT3, and BEL1) and six Physcomitrella patens TALEs as land plant references, three Human (MEIS2, PKNOX1 and PBX1) and five protozoan TALEs from Guillardia theta, Acanthamoeba castellani, and Naegleria gruberi as references outside Archaeplastida (S2 Spreadsheet).

Method S2. Phylogenetic reconstruction
A phylogenetic reconstruction was performed using Bayesian (MrBayes; Ronquist et al., 2012) and Maximum-likelihood (PhyML, RAxML, and IQ-TREE; Guindon et al., 2010; Rokas, 2011; Minh et al., 2013) methods. IQ-TREE yielded a strongly supported consensus unrooted tree with Ultrafast bootstrap test score with ‘-bb 2000 -bi 500’ option, whereas the other methods produced trees with poorly resolved clade structure. The LG+R5 model was chosen according
to Bayesian Information Criteria (Luo et al., 2010). Shimodaira-Hasegawa-like (SH) and Bayesian approximate likelihood-ratio tests were performed by IQ-TREE with "-alrt 1000" and "abayes" options (Anisimova et al., 2011). A score of 95% in Ultrafast bootstrap and 80% in SH-alrt was considered highly significant recommended by the software authors. Classification relied on phylogenetic identification of clades with high bootstrap support in more than one tree topology test as well as ad hoc homology-motif searches to identify shared domain structure among the collected homeobox proteins.

**Method S3. Curation of gene models**

Many sequences lacked homology domains outside the homeodomain, possibly due to incomplete sequence information, lineage- or gene-specific divergence, or false phylogenetic association caused by limited sequence information. To locate possible missing or very divergent domains, we used alignment as a tool, generating extensive alignments combining ranging from three to 20 sequences. All the alignments were manually inspected for the identified motif/domains to be correctly aligned. Sequences with large insertions or deletions were compared with the original genome annotations to detect possible alternative splice sites. When a better gene model was identified than that deposited in GenBank, we used it. These sequences are denoted with "v2" in figures and the protein sequences are provided in S2 Spreadsheet. When necessary, we amended the homeodomain alignment and updated the phylogenetic analysis.

**Method S4. Cloning of Yeast-two-hybrid constructs**

Micco_62153 and Picsa_04684 contained a single intron, whereas all the other nine genes lacked an intron in the entire open reading frame. For cloning of Micco_62153, we synthesized the middle fragment lacking the intron and ligated them via XhoI and ClaI sites. For cloning of Picsa_4684, 5'- and 3'-side exons were amplified separately and combined using PCR. Using 5'-EcoRl or MfeI and 3'-BamHl, BglII or XhoI, cloned DNA fragments were placed in-frame to pGBKKT7 and pGADT7 vectors at the EcoRl/BamHl or EcoRl/SalI sites (Clonetech).

**References in the supporting information**


