Deep origins of eukaryotic multicellularity revealed by the *Acrasis kona* genome and developmental transcriptomes

Sheikh et al.

Supplemental Data

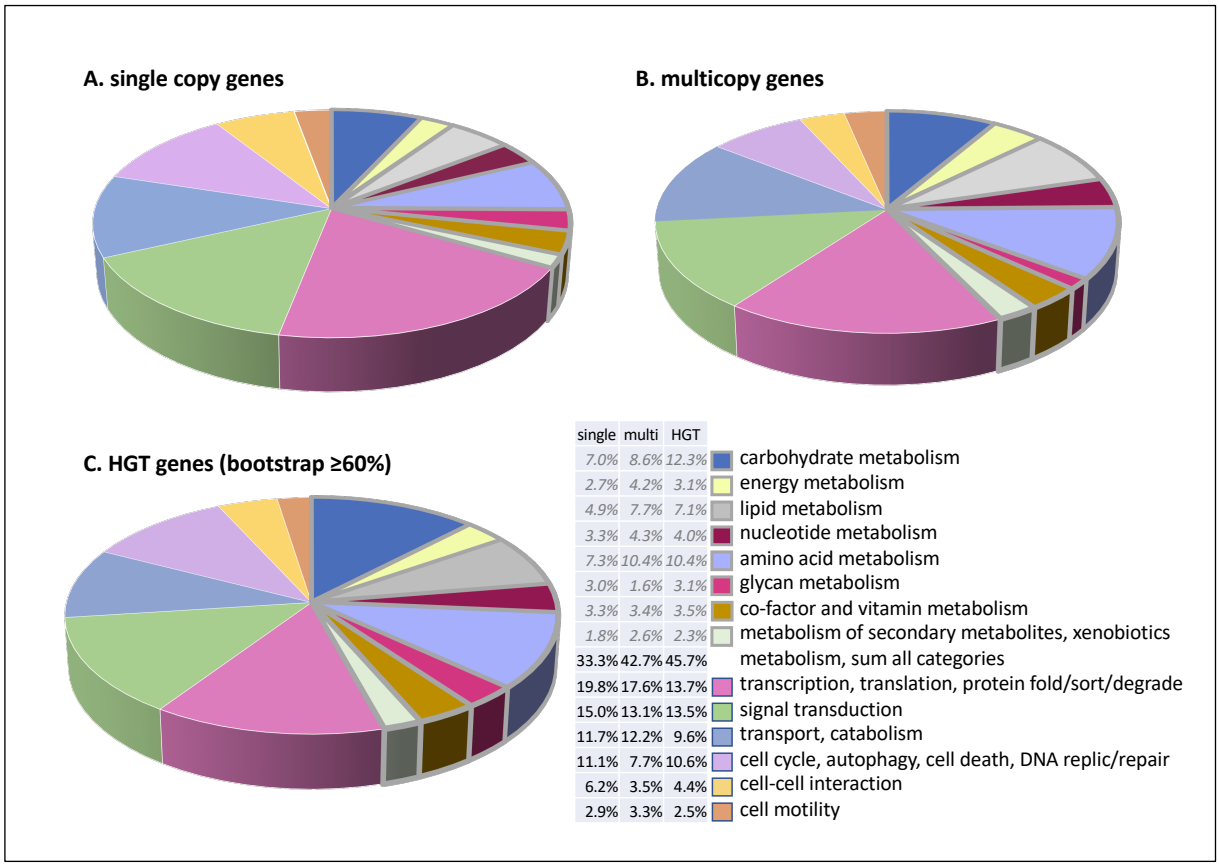

Figure S1: Predicted protein function for single-copy, multicopy and horizontal gene transfer (HGT) candidate accessions in *Acrasis kona*. Percentages of general functional categories are shown for *A. kona* predicted proteins for A) single copy, B) multicopy and 3) HGT candidate accessions. Functional categorization of predicted proteins was determined using GhostKOALA (Kanehisa et al. 2016).

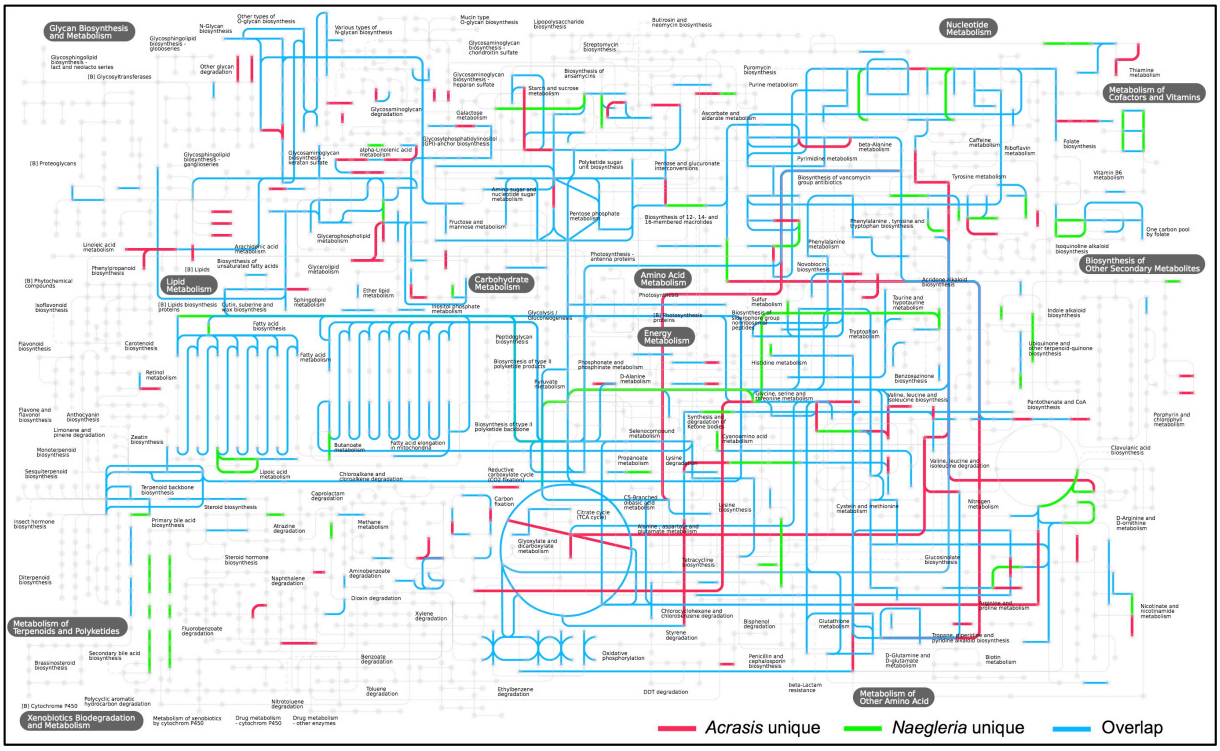

Figure S2. Metabolic capacity of *Acrasis kona* and *Naegleria gruberi*. Unique enzymatic pathways for *A. kona* and *N. gruberi* are shown in red and green, respectively, and the overlap between the two in blue. Pathways were predicted and graphically summarized using iPath (Letunic et al. 2008).

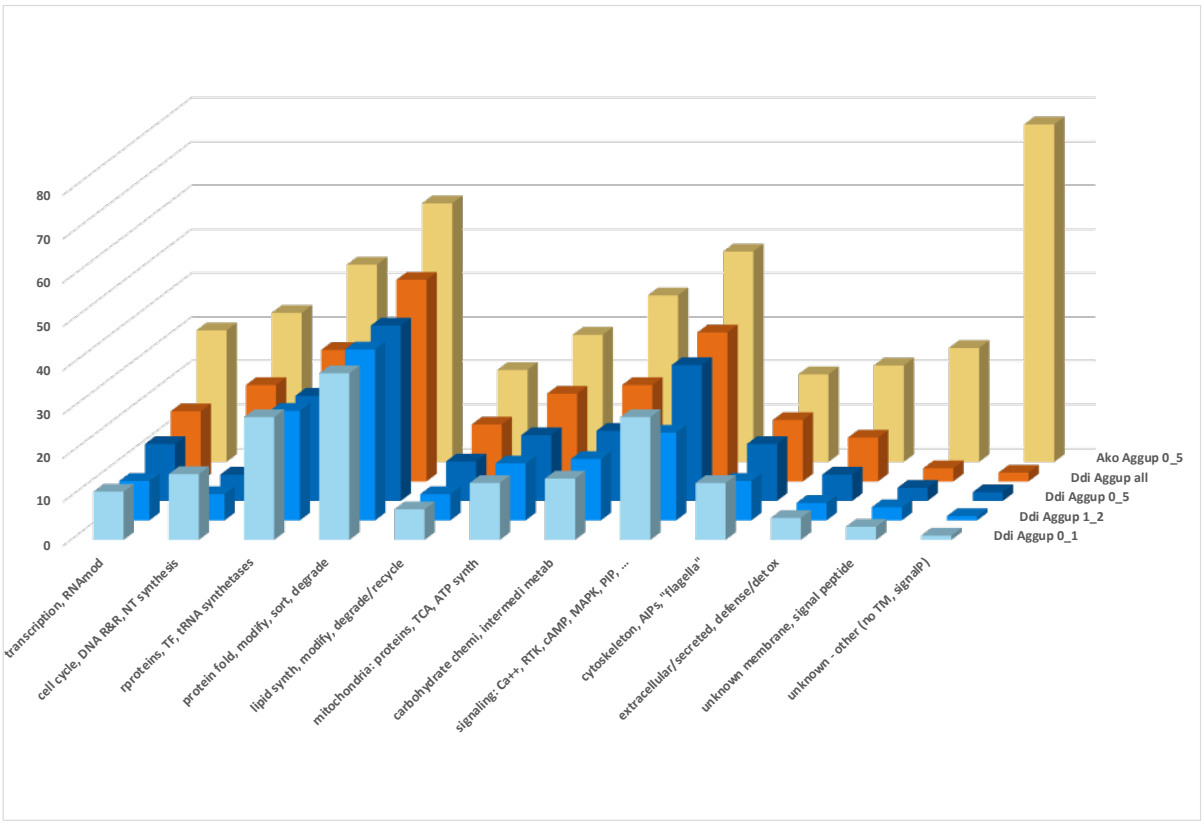

Figure S3. Variation by time point in protein function profile for *Dictyostelium discoideum* homologs of *Acrasis kona* aggregation proteins. *D. discoideum* AX4 (Ddi) proteins homologous to *Acrasis kona* (Ako) Agg_{up} accessions and with increased expression (DE >0.9) and/or highly expressed (RPKM >1000) during starvation are clustered into functional categories. Time point comparisons correspond to hours after food depletion: hours 0 vs 1 (0_1), 1 vs 2 (1_2), 0 vs 5 (0_5), all four comparisons combined (all), and all Ako Agg_{up} accessions with DE >0.9 and/or RPKM >1000 in growth versus aggregation (0_5). Bar height (y axis) corresponds to numbers of accessions in major function categories (Table S11).

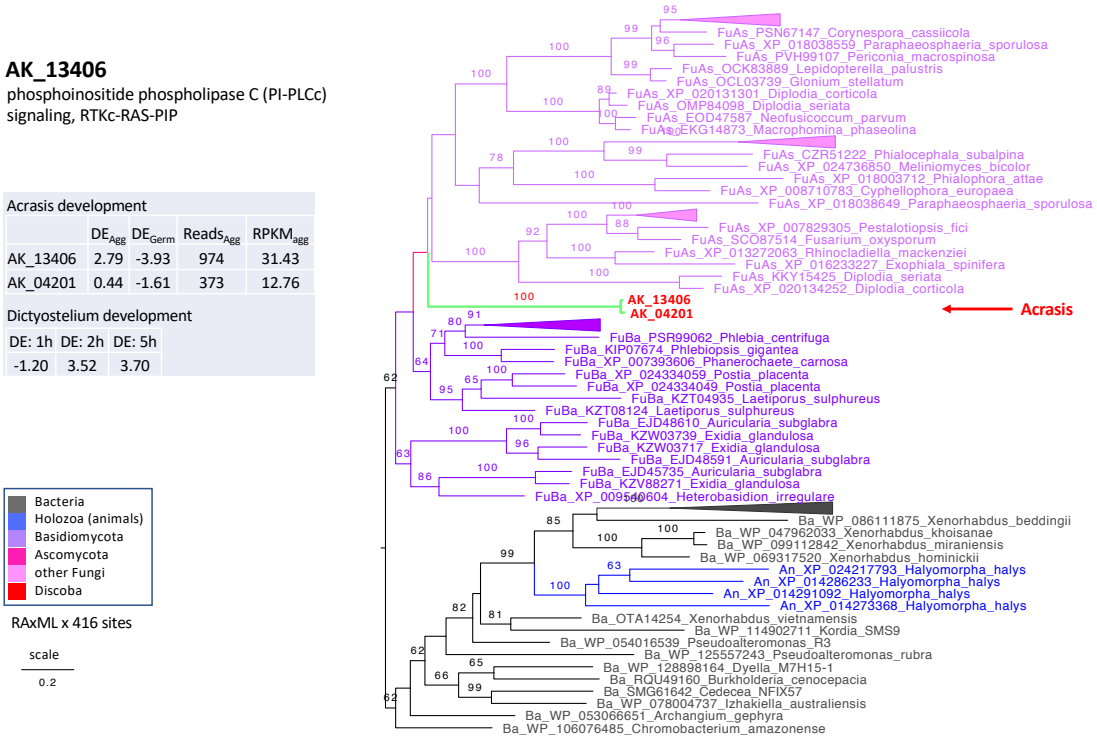

Figure S4. An *Acrasis kona* aggregation-specific phosphoinositide phospholipase C derived by horizontal transfer (accession AK_13406). The tree shown was derived from a 415 amino acid alignment using RAxML and the LG+gamma model (Stamatakis 2006). Only bootstrap values over 60% are shown and branch lengths are drawn to scale according to the scale bar at the lower left (substitutions per site). Sequence names are colored according to higher order taxonomy shown in the lower left key. Some large groups of closely related sequences are collapsed to proportional triangles to increase readability of more relevant parts of the trees. The table to the immediate left shows gene expression values for growth vs aggregation (DE_{Agg}), aggregation vs germination. (DE_{Germ}) and numbers of raw reads ($Read_{Agg}$) and/or length-corrected reads ($RPKM_{Agg}$) for aggregating cells (Table S10). DE_{Log2} values for Ddi AX4 homologs are given for growth (0 hours) versus 1, 2 and 5 hours starvation (Santhanam et al. 2015).

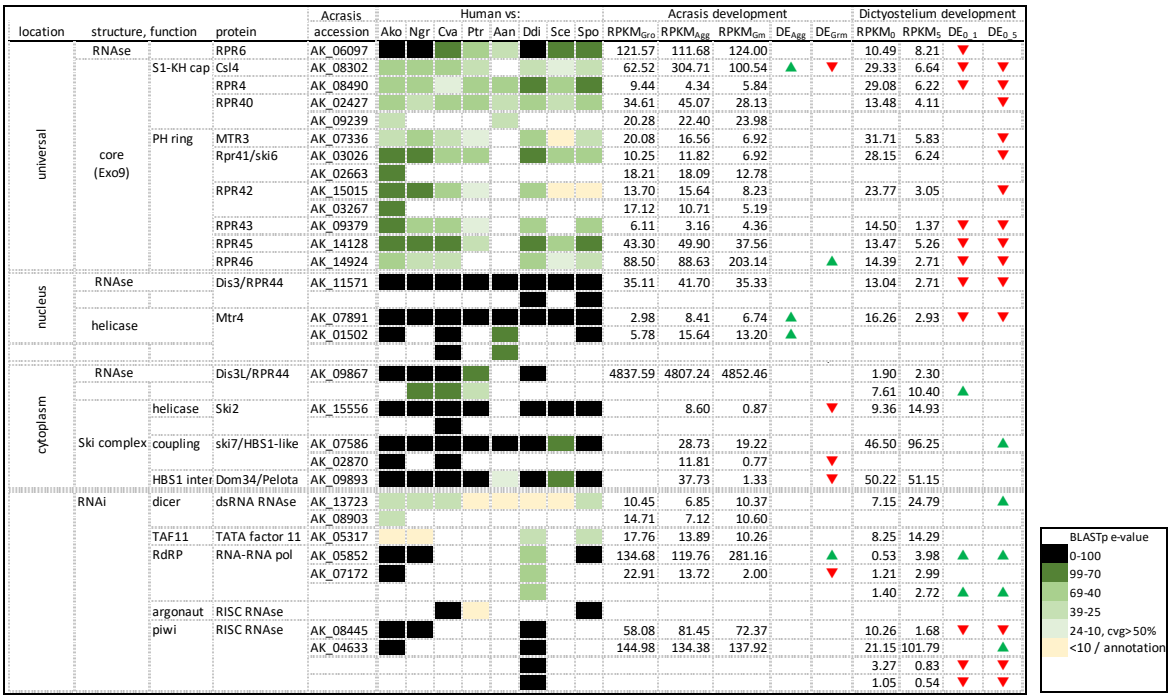

Figure S5. Sequence conservation and developmental expression of the *Acrasis kona* exosome. Homologs of exosomal proteins were identified by BLASTp using human queries. The strength of BLASTp hits (c-value) is indicated by color intensity (key at bottom right). BLASTp hits with high e-values (e^{-24} to e^{-10}) were scored as present only if their BLASTp alignments included >50% query coverage (cvg). Exosomal components annotated as present but with only very weak BLASTp scores (> e^{-10}) are indicated in pale yellow. RNAseq length-corrected read numbers (RPKM) are shown for actively growing ($RPKM_{Gro}$), aggregating ($RPKM_{Agg}$) and germinating ($RPKM_{Germ}$) *A. kona* cells and similarly for growth and aggregation in *Dictyostelium discoideum AX4* ($RPKM_0$ and $RPKM_5$, respectively). Green and red arrows indicate substantially increased or decreased (DE >0.9 or DE <-0.9, respectively) gene expression for growth vs aggregation (DE_{Agg}) and aggregation vs germination (DE_{Germ}) in *A. kona* and similar time points for Ddi AX4, *i.e.,* 0 vs 1 (DE_{0_1}) and 0 vs 5 (DE_{0_5}) hours of starvation. Accession numbers, BLASTp e-values and gene expression values are given in Table S14. Taxon names are abbreviated as follows: *A. kona* (Ako), *Aureococcus anophagefferens* (Aan), *Chlorella sp.* (Cva), *Ddi AX4* (Ddi), *Naegleria sp.* (Ngr), *Phaeodactylum tricornutum* (Ptr), *Saccharomyces cerevisiae S288c* (Sce) and *Schizosaccharomyces pombe* (Spo). Ddi AX4 gene expression values were derived from Santhanam et al. (2015).

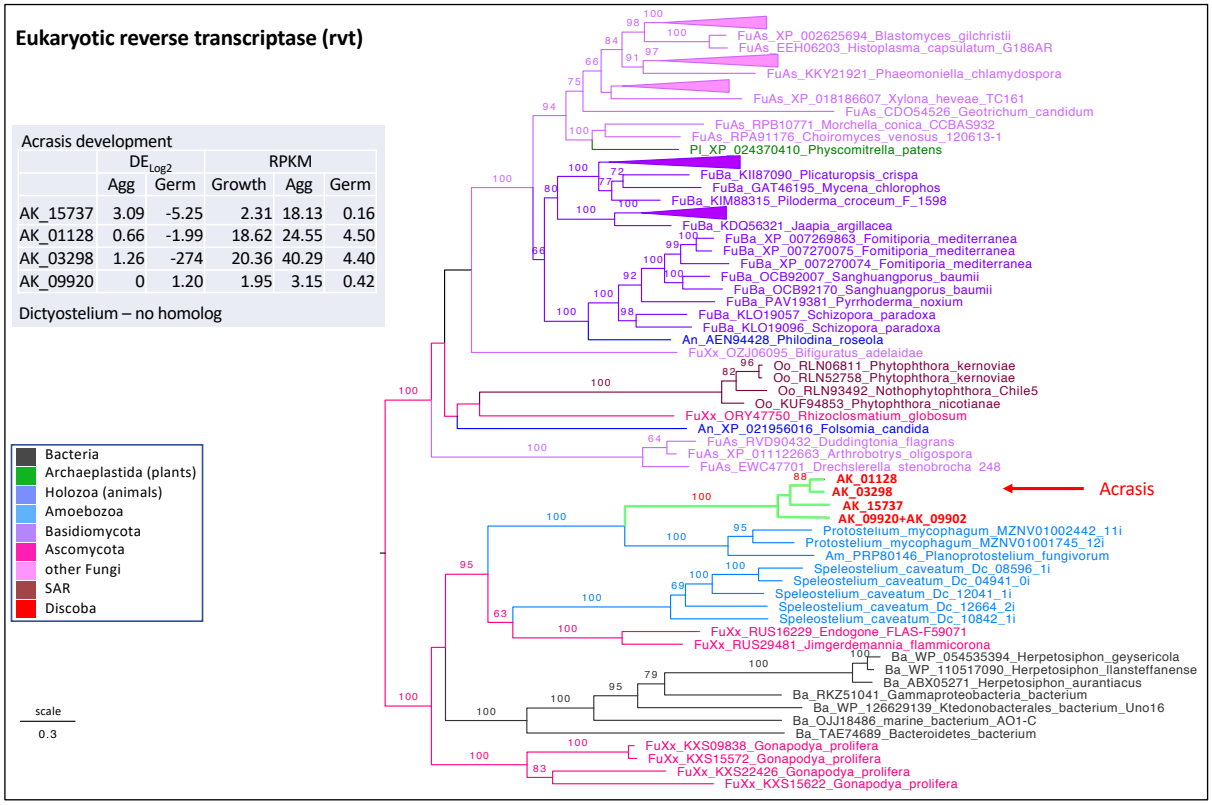

Figure S6. A family of reverse transcriptases in *Acrasis kona*. The tree shown was constructed from a 628 amino acid alignment using RAxML and the LG+gamma model (Stamatakis 2006). Only bootstrap values over 60% are shown and branch lengths are drawn to scale according to the scale bar at the bottom left (substitutions per site). Sequence names are colored according to higher order taxonomy according to the key to the lower left of the tree. Some large groups of closely related sequences are collapsed as triangles to increase readability of more relevant parts of the trees. The table to the upper left shows gene expression values for growth vs aggregation (DE_{Agg}), aggregation vs germination. (DE_{Germ}), and numbers of length corrected ($RPKM_{Agg}$) reads (Table S10). DE values for Ddi AX4 homologs are given for growth (0 hours) versus 1, 2 and 5 hours starvation (Santhanam et al. 2015).

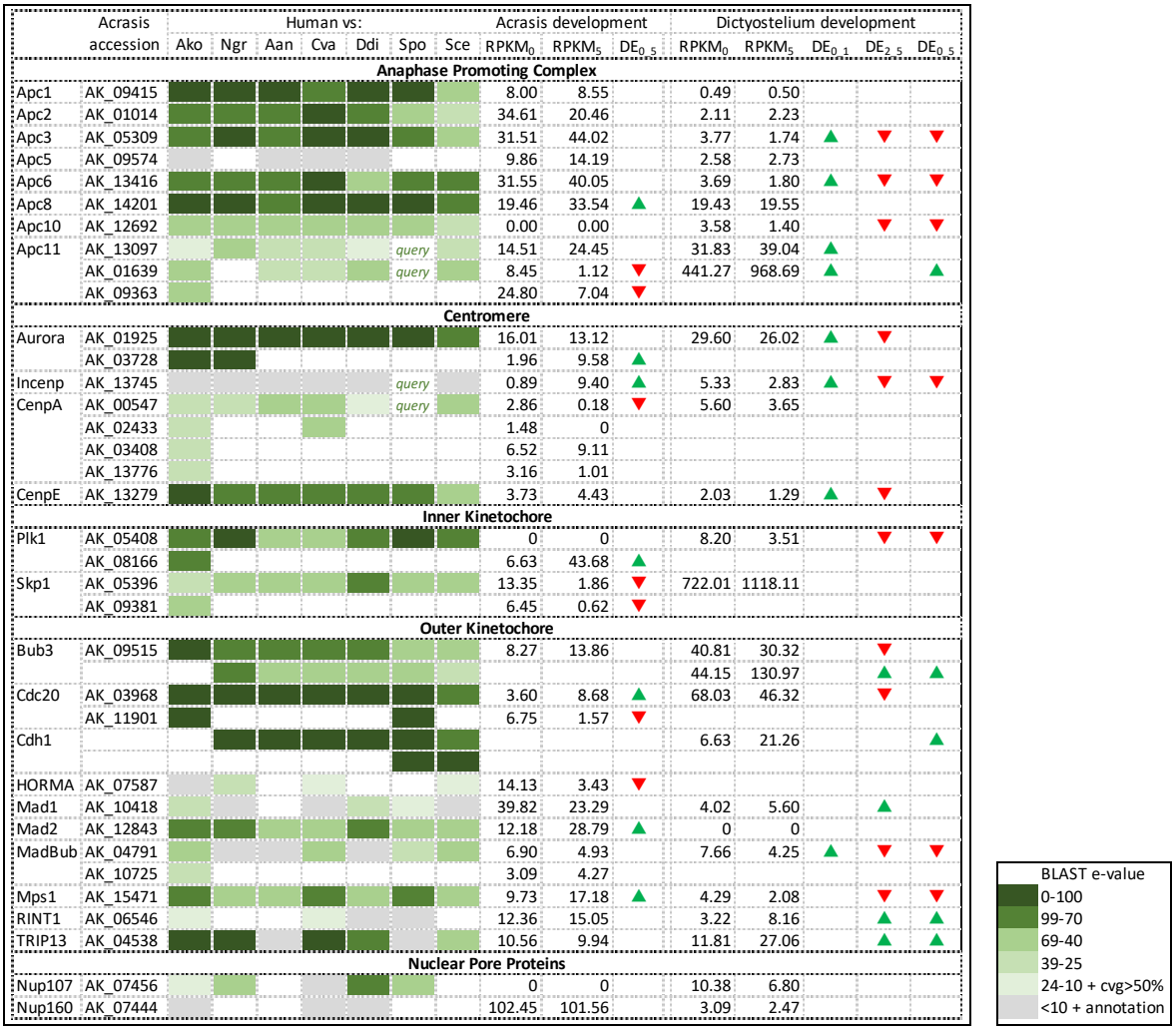

Figure S7. The *Acrasis kona* kinetochore and life cycle-stage specific expression patterns. Homologs of kinetochore proteins were identified by BLASTp using *Homo sapiens* sequences as queries. The strength of BLASTp hits (c-value) is indicated by color intensity according to the key at the bottom right. BLASTp hits with high e-values (e^{-24} to e^{-10}) were scored as present only if their BLASTp alignments included >50% query coverage (cvg). RNAseq length-corrected read numbers (RPKM) are shown for *A. kona* for actively growing ($RPKM_0$) and aggregating ($RPKM_5$) cells and for similar time points for *Dictyostelium discoideum AX4* (growth: $RPKM_0$, aggregation: $RPKM_5$). Green and red arrows indicate substantially increased or decreased gene expression (DE >0.9, DE <-0.9, respectively) for both *A. kona* and *D. discoideum* for growth vs aggregation (DE_{0_5}). Accession numbers, BLASTp e-values and gene expression values are given in Table S15. Taxon names are abbreviated as follows: *A. kona* (Ako), *Aureococcus anophagefferens* (Aan), *Chlorella variabilis* (Cva), *D. discoideum AX4* (Ddi), *Naegleria gruberi* (Ngr), and *Saccharomyces cerevisiae S288c* (Sce). *D. discoideum AX4* gene expression values were derived from Santhanam et al. (2015).

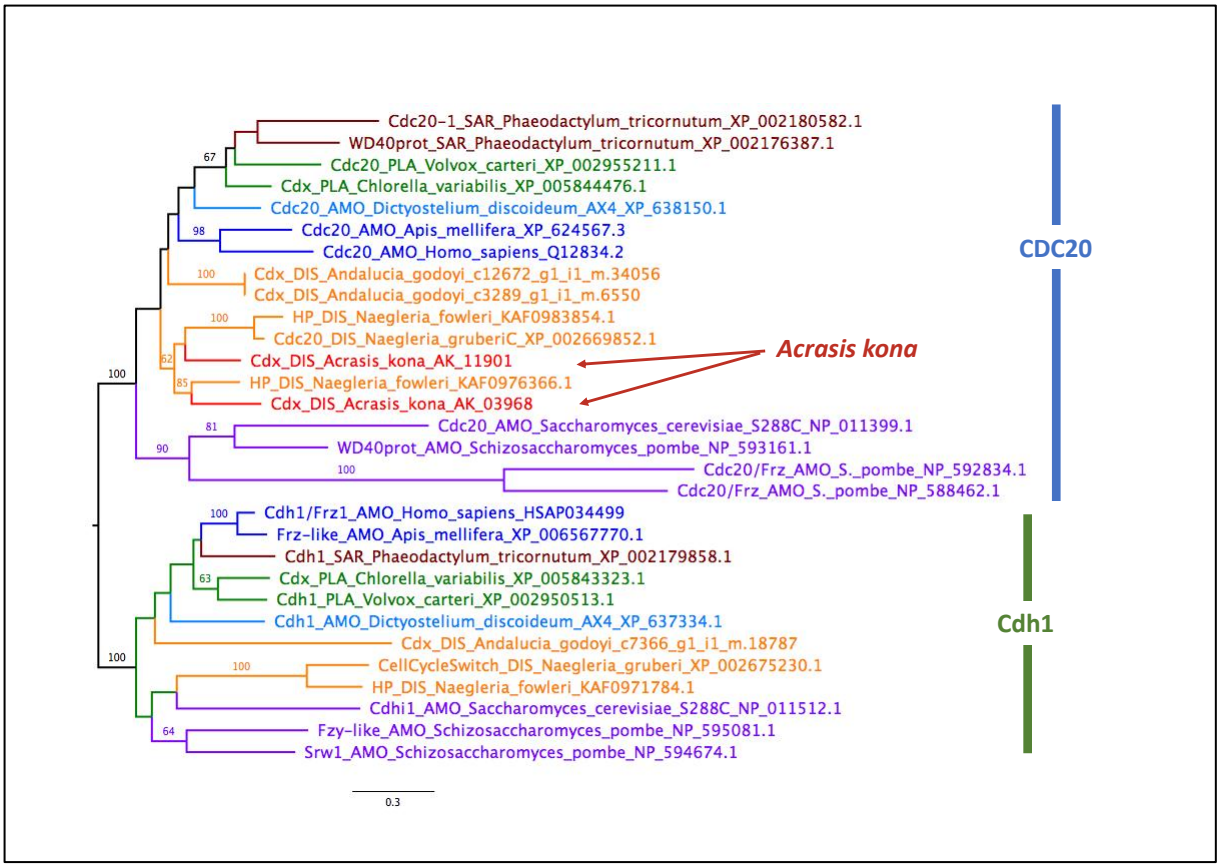

Figure S8. Both of the autophagy control proteins of *Acrasis kona* are of the CDC20-type. Protein sequences were identified by taxon-specific BLASTp using human protein sequence queries, and all hit sequences with e- values < e^{-35} were retrieved. Sequences were aligned and trimmed using the Auto-1 algorithm of Trim-A1 (Capella-Gutierrez et al. 2009). The tree was constructed from the resulting 306 aligned positions using RAxML with the LG+gamma amino acid substitution model and bootstrapping with automated stopping (Stamatakis 2006). Sequences are labeled according to their primary annotation (otherwise denoted as hypothetical proteins or HP), and major taxonomy as follows: SAR (Stramenopila, Alveolata and Rhizaria), PLA (Archaeplastida), DIS (Discoba), Amo (Amorphea). Numbers on branches show bootstrap values over 60%. The tree is drawn to scale according to the scale at the bottom (substitutions per site).

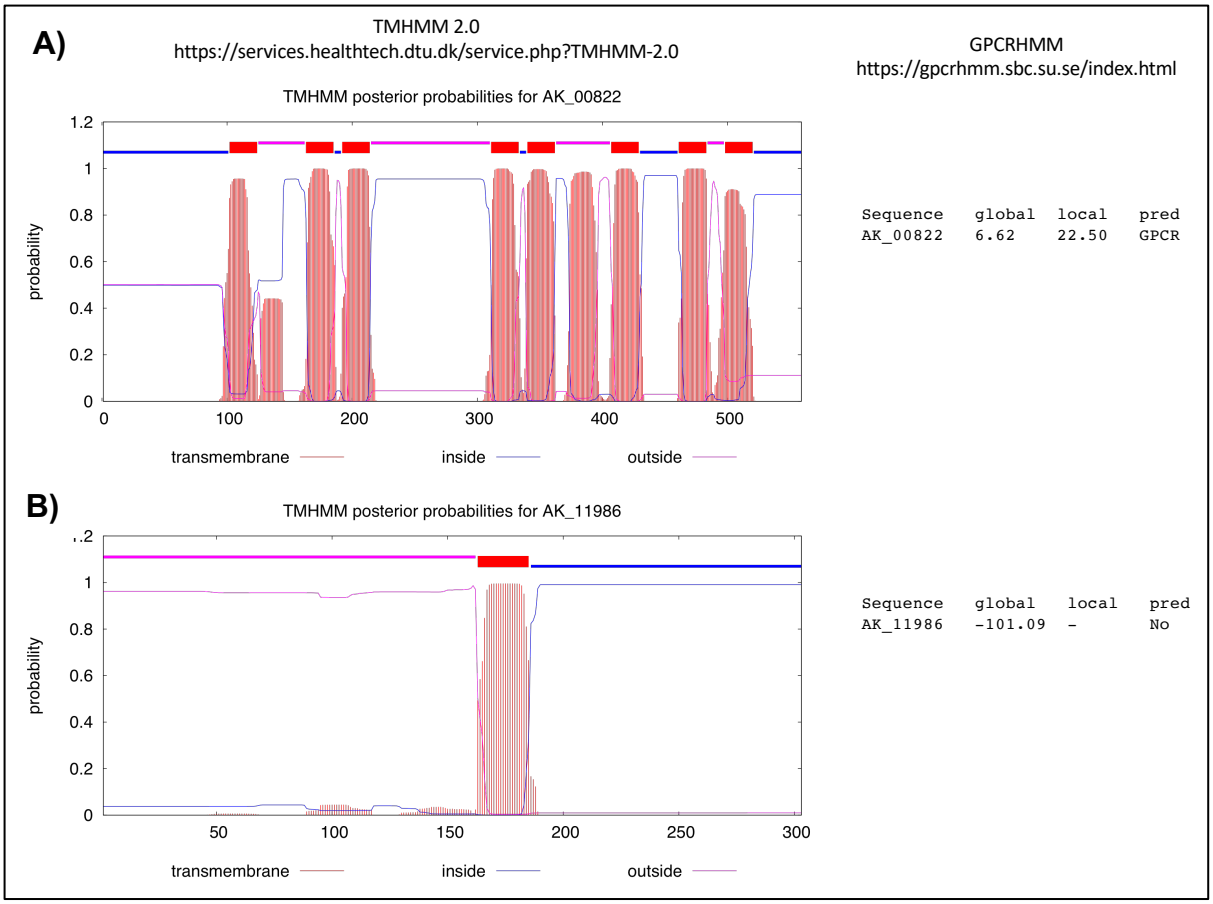

Figure S9. Two extracellular signaling proteins predicted from novel *Acrasis kona* aggregation accessions.. All unannotated *A. kona* aggregation proteins with predicted transmembrane domains (Table S3) were analyzed using the Hidden Markov Model based GPCR predictor TMHMM 2.0 (services.healthtech.dtu.dk/service.php?TMHMM-2.0; Wistrand et al. 2006). A) AK_00822 is strongly predicted to encode a classic GPCR with 8-9 transmembrane domains and large loops on both sides of the plasma membrane. B) AK_11986 is predicted to encode a membrane signaling or receptor protein with a single membrane spanning domain and large domains on both external and internal membrane surfaces.

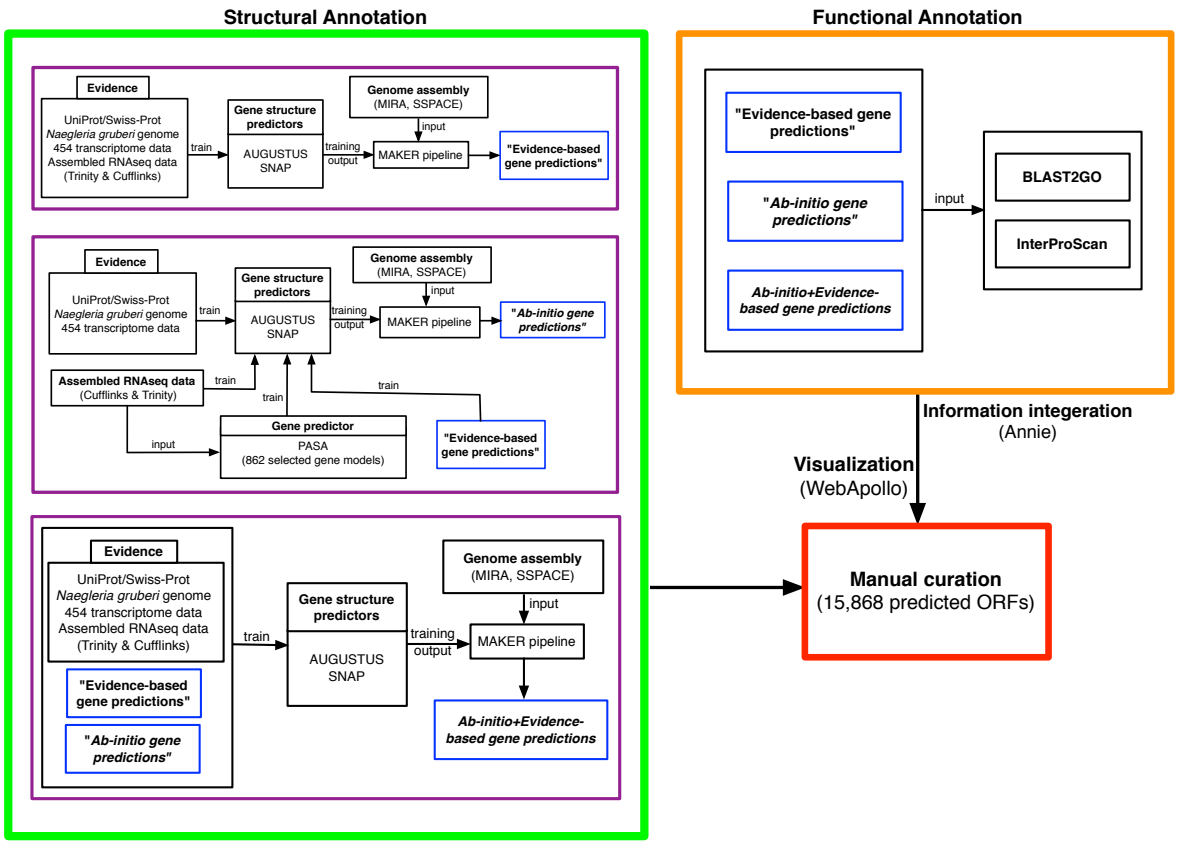

Figure S10. ***Acrasis kona*** **genome annotation pipeline.** The pipeline shown was used to perform structural (left) and functional (right) annotation followed by manual curation of the *A. kona* nuclear genome. Three different strategies were used to train gene predictors for structural annotation to produce evidence-based gene predictions, *ab-initio* gene predictions and evidence+*ab-initio* gene predictions.