# The Co-regulation Data Harvester for *Tetrahymena thermophila*: automated high-throughput gene annotation and functional inference in a microbial eukaryote

Lev M. Tsypin, Aaron P. Turkewitz*

*Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago IL, 60637*

## Abstract

Identifying co-regulated genes can provide a useful approach for defining pathway-specific machinery in an organism. To be efficient, this approach relies on thorough genome annotation, which is not available for most organisms with sequenced genomes. Studies in *Tetrahymena thermophila*, the most experimentally accessible ciliate, have generated a rich transcriptomic database covering many well-defined physiological states. Genes that are involved in the same pathway show significant co-regulation, and screens based on gene co-regulation have identified novel factors in specific pathways, for example in membrane trafficking. However, a limitation has been the relatively sparse annotation of the *Tetrahymena* genome, making it impractical to approach genome-wide analyses. We have therefore developed an efficient approach to analyze both co-regulation and gene annotation, called the Co-regulation Data Harvester (CDH). The CDH automates identification of co-regulated genes by accessing the *Tetrahymena* transcriptome database, determines their orthologs in other organisms via reciprocal BLAST searches, and collates the annotations of those orthologs' functions. Inferences drawn from the CDH reproduce and expand upon experimental findings in *Tetrahymena*. The CDH, which is freely available, represents a powerful new tool for analyzing cell biological pathways in *Tetrahymena*. Moreover, to the extent that genes and pathways are conserved between organisms, the inferences obtained via the CDH should be relevant, and can be explored, in many other systems.

*Corresponding Author
*Email address:* apturkew@uchicago.edu (Aaron P. Turkewitz)

*Keywords:* Co-expression, Co-regulation, Transcriptomics, Genomics, Bioinformatics, Tetrahymena

## 1. Motivation and significance

*Tetrahymena thermophila* is a ciliate, one of the best-studied members of this large group of protists [1]. Its use as a model system led to the Nobel Prize-winning discoveries of telomerase and self-splicing RNA, as well as to other breakthroughs, including the isolation of dyneins and making the link between histone modification and transcriptional regulation [2, 3, 4, 5]. These contributions to our understanding of important cellular pathways made use of classical forward and reverse genetics, as well as biochemical approaches. More recently, genomic and transcriptomic data became available for *T. thermophila*, which have been used to infer functional gene networks [6, 7, 8, 9, 10, 11].

The *T. thermophila* genome has been sequenced and assembled [6], and is available online on the *Tetrahymena* Genome Database (TGD) [7]. While the TGD collates the sequence data along with available gene annotations and descriptions, the genome overall remains incompletely annotated. An extensive transcriptomic database, the *Tetrahymena* Functional Genomics Database or *Tetra*FGD, is also available for *T. thermophila* [10]. These data were collected over a well-established range of culture conditions in which *T. thermophila* undergoes large physiological changes [8, 9, 10, 11]. In addition to displaying individual expression profiles, the *Tetra*FGD can indicate the statistical strength of co-expression between any two genes, as calculated using the Context Likelihood of Relatedness (CLR) algorithm [9, 12, 13]. Co-expression in *T. thermophila*, as judged based on mRNA levels, can reveal functionally significant co-regulation. Gene regulation in this species appears to predominantly occur at the level of transcription [14], and so steady-state mRNA levels may explain the majority of steady-state protein levels, as reported in other systems [15]. In this report, we will refer to genes that are listed as co-expressed in the *Tetra*FGD as co-regulated.

A high-throughput analysis of *T. thermophila* gene expression profiles revealed that accurate gene networks can be inferred from co-regulation data [13], providing evidence that co-regulated genes tend to be functionally associated. This approach has been used in bacterial, mammalian, and apicomplexan systems [12, 16, 17]. There is also experimental evidence in *T. thermophila* to support the conclusion that co-regulation corresponds to functional association: Co-regulation data were used to successfully predict novel sorting factors and proteases involved in the biosynthesis of a class of secretory vesicles, called mucocysts [18, 19]. These results suggest that the *T. thermophila* transcriptome may be used to bioinformatically infer factors involved in an array of cellular pathways.

The CDH was designed to facilitate genome-wide analyses of gene co-regulation. The CDH automatically mines co-regulation data for genes of interest, and annotates the co-regulated genes *via* forward and reciprocal BLAST searches that identify orthologs in other model organisms. The CDH provides a systematic tool for gathering and annotating genomic information from public databases, and it can allow a researcher to quickly develop a robust hypothesis about the cellular pathways or structures in which a gene of interest may be acting, based upon the genes with which it is co-regulated.

## 2. Software description

### *2.1. Software Architecture*

The CDH was developed for Python 2.7, along with the following packages: `sys`, `os`, `platform`, `logging`, `re`, `dill`, `difflib`, `csv`, `pdb`, `shutil`, `xml`, `win32com.shell`, `requests`, `BeautifulSoup4`, and `Biopython` [21]. Executables for Windows (x64) and MacOS (10.6+) were made using the Pyinstaller library. The CDH gathers available data for a set of co-regulated genes from publicly available databases, and uses these data to predict possible gene functions (Figure 1). The gathered information includes the co-regulation data from the *Tetra*FGD, and the gene names, sequences, and annotations from the TGD. The available annotations come from a combination of experimental results and inferences from homology [7]. The CDH predicts annotations for genes based on the annotation of their respective orthologs, which are themselves identified by a series of forward and reciprocal BLAST searches *via* the National Center for Biotechnology Information (NCBI). These predicted annotations are generated by using the Ratcliff-Obershelp algorithm [22], as implemented in the python difflib library, to identify common phrases in the orthologs' annotations.

### *2.2. Software Functionalities*

The basic functions of the CDH are to gather available co-regulation and annotation data, perform forward and reciprocal BLAST searches and predict gene annotations, and report this gathered information in a human-readable format. The CDH interface first asks the user to enter the ID for the gene whose co-regulated factors are of interest (Figure 2). Next, the user defines: how many of the co-regulated genes should be interrogated *via* BLAST; how to process data files that had been previously generated and are relevant to the current query; whether to use the BLASTp or BLASTx algorithm;

and in which taxa to run the forward BLAST searches. The results of the CDH analysis are saved as a Comma Separated Values (.csv) file in the user's "Documents" folder: `/Documents/CoregulationDataHarvester/csvFiles`.

## 3. Illustrative Examples

### *3.1. A CDH analysis of a factor required for programmed genome rearrangement returns the vast majority of experimentally-verified genes involved in the pathway*

Programmed genome rearrangement is a tightly regulated process that occurs during the formation of the new somatic nucleus in conjugating *Tetrahymena* [23, 24]. This process is well-studied and known to be driven by a special adaptation of RNA interference, utilizing Dicer- and Piwi-like proteins, among other factors [25, 26]. *TWI1* encodes a Piwi-like protein that plays a central role in programmed genome rearrangement [27, 26]. When *TWI1* is entered as the query for the CDH, the CDH retrieves a large number of DNA and RNA-processing factors, as well as chromodomain proteins (Supplementary File 1). Importantly, these include the key factors known to be involved in programmed genome rearrangement (Table 1). The CDH report for this *TWI1* query is attached as Supplementary File 2. Within this report, we have highlighted the cases in which the CDH matched or expanded upon existing annotations.

It is also notable that specific homologs of Dicer that are not involved in programmed genome rearrangement, namely *DCR1* and *DCR2* [26], are not present in the list of genes co-regulated with *TWI1*. Similarly, while *TPB2* is a known genome rearrangement factor and is present in the CDH output [26], its paralog *TPB1* is neither involved in this process nor identified as co-regulated with *TWI1*. Thus, the CDH is a useful tool for focusing on pathway-specific paralogs within gene families.

### *3.2. A CDH analysis of a mucocyst biogenesis factor enriches for mucocyst cargo and maturation factors*

Mucocysts in *Tetrahymena* are secretory organelles. Mucocysts undergo a maturation process that requires the catalyzed cleavage of cargo proteins, called GRLs [28]. The *T. thermophila* genome encodes approximately 480 predicted proteases [6], but only five of these are co-regulated with GRLs, as revealed by a manual inspection of expression profiles on the *Tetra*FGD [19]. Two of these proteases, called *CTH3* (cathepsin 3) and *CTH4* (cathepsin 4), were subsequently shown to represent key enzymes for GRL cleavage [19, 29].

Using *CTH3* as a query for the CDH results in a list that includes a large number of genes known to be involved in mucocyst biogenesis (Table

2), and is enriched in membrane-trafficking factors and proteins with as-yet unknown functions in this organism (Supplementary File 3). Among the latter are a subunit of the *AP3* complex and a syntaxin in the *STX7* subfamily. Subsequent functional analysis of these genes showed that they are both essential for mucocyst formation, providing the best evidence to date that mucocysts are lysosome-related organelles (Kaur et al., submitted). The CDH report for this *CTH3* query is attached as Supplementary File 4. This report is also edited to indicate the cases when the CDH matched or expanded upon existing gene annotations.

## 4. Impact

The CDH reproduces existing annotations with high accuracy, and provides a large number of new annotations and expansions upon existing ones (Table 3; Supplementary Files 2 and 4). Effectively, the CDH increased the annotation coverage of the genes co-regulated with *TWI1* from 46% to 60%, and the annotation coverage of the genes co-regulated with *CTH3* from 41% to 57%. Specifying the BLAST parameters allows the user to discover the most informative functional predictions for their genes and pathways of interest. Limiting the CDH search to lineages outside of the ciliates is more likely to retrieve previously annotated orthologs, but runs the increased risk that weak homologs will generate spurious results. For some processes, such as programmed genome rearrangement in which *TWI1* is involved, the most informative BLAST searches may be those restricted to the ciliates. In our trials, the effectiveness of the CDH is maintained regardless of which taxa the BLAST searches are run against.

In addition to providing a means of quickly gathering available data about a set of co-regulated genes and inferring their functions, the CDH data can be extended to to investigate the potential overlap between components of different cellular pathways. For example, *NUP50* encodes a gene that functions both in nuclear import at the nuclear pore complex and as part of a complex involved in transcription [30, 31]. Accordingly, the genes co-regulated with *NUP50* show extensive overlap with genes co-regulated with an import factor (Importin$\beta$) and with a gene involved in transcription (*RPB81*, an RNA polymerase II subunit), among other factors involved in both processes (Figure 3, A).

It is informative to compare the overlap of co-regulated genes in different pathways. The co-regulated gene sets for three factors involved in genome rearrangement, *TWI1*, *GIW1*, and *DCL1*, show almost complete overlap, suggesting that they may be involved in a single common process (Figure 3, B). In contrast to the case of programmed genome rearrangement, the co-regulated gene sets for three factors required in mucocyst formation, *CTH3*, *SOR4*, and *APM3*, show partial overlap, hinting that one or more of these factors may also play roles unrelated to mucocysts (Figure 3, C). Consistent with this idea, *CTH3* is an essential gene, while mucocysts themselves are dispensable for cell viability in the laboratory [19]. Importantly, there is very little overlap between the co-regulated gene sets defined for the three different cellular processes (nuclear import/transcription, genome remodeling, and mucocyst formation) (Figure 3, D). The overlap is smallest between the genes co-regulated with mucocyst biogenesis factors and genes co-regulated with either nuclear import, transcriptional regulation, or programmed genome rearrangement. The somewhat greater sharing of genes co-regulated with nuclear import, transcriptional regulation, and programmed genome rearrangement may reflect the fact that these pathways all take place in the nucleus and are intrinsically linked to the cell cycle. Given the ease of assembling sets of co-regulated genes using the CDH, this type of overlap analysis can be extended to many cellular pathways.

## 5. Conclusions

Protists constitute the majority of eukaryotic diversity, meaning that this group needs to be included in evolutionary analyses of cellular processes, but this diversity is largely overlooked in the standard collection of model eukaryotes [20]. We present the Co-regulation Data Harvester for *T. thermophila* (CDH) as a tool that expedites analyses of *T. thermophila* genome, transcriptome, and cellular biology in an evolutionary context. The CDH is freely available and provides a systematic framework for genome annotation. It quickly gathers information from disparate databases and, by optionally reusing BLAST results that had been stored during previous queries, can increase in speed with successive uses. In providing a new means to analyze transcriptomic data, the CDH makes clear the potential for using the rapidly growing amount of genomic and transcriptomic data in many organisms, to facilitate functional analysis in poorly annotated or emerging model systems.

Users of the CDH should keep in mind that its reports are necessarily limited by pre-existing data from the TGD, *Tetra*FGD, and the NCBI. For example, the *Tetra*FGD does not provide co-expression data for genes whose expression level falls below a set threshold. Because of this limit, some *T. thermophila* genes may be overlooked by the CDH. Executable files for the program can be found at `http://ciliate.org/index.php/show/CDH`. A manual with detailed instructions and usage examples is provided in Supplementary File 5.

## Acknowledgements

We would like to thank Stefano Allesina for providing the initial impetus for this project, excellent advice, and review of this manuscript. Thanks to Chad Pearson and Alexander Stemm-Wolf for their feedback on the CDH and this manuscript. We thank Naomi Stover and Wei Miao for granting the Ciliate community access to the TGD and *Tetra*FGD, without which none of this work would be possible. We thank Tokuko Haraguchi for helpful discussion about NUP50. We thank Daniela Sparvoli and Harsimran Kaur for helpful discussion and good company. This work was supported by NIH GM-105783 to APT.

[1] G. Witzany, M. Nowacki (Eds.), Biocommunication of Ciliates, Springer, 2016.

[2] C. W. Greider, E. H. Blackburn, Identification of a specific telomere terminal transferase activity in tetrahymena extracts, Cell 43 (2) (1985) 405 – 413. doi:http://dx.doi.org/10.1016/0092-8674(85)90170-9.

[3] K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, T. R. Cech, Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena, Cell 31 (1) (1982) 147–157. doi:10.1016/0092-8674(82)90414-7.

[4] I. Gibbons, A. Rowe, Dynein: a protein with adenosine triphosphatase activity from cilia, Science 149 (3682) (1965) 424–426.

[5] J. E. Brownell, J. Zhou, T. Ranalli, R. Kobayashi, D. G. Edmondson, S. Y. Roth, C. Allis, Tetrahymena histone acetyltransferase a: A homolog to yeast gcn5p linking histone acetylation to gene activation, Cell 84 (6) (1996) 843 – 851. doi:http://dx.doi.org/10.1016/S0092-8674(00)81063-6.

[6] J. A. Eisen, R. S. Coyne, M. Wu, D. Wu, M. Thiagarajan, J. R. Wortman, J. H. Badger, Q. Ren, P. Amedeo, K. M. Jones, L. J. Tallon, A. L. Delcher, S. L. Salzberg, J. C. Silva, B. J. Haas, W. H. Majoros, M. Farzad, J. M. Carlton, R. K. Smith Jr., J. Garg, R. E. Pearlman, K. M. Karrer, L. Sun, G. Manning, N. C. Elde, A. P. Turkewitz, D. J. Asai, D. E. Wilkes, Y. Wang, H. Cai, K. Collins, B. A. Stewart, S. R. Lee, K. Wilamowska, Z. Weinberg, W. L. Ruzzo, D. Wloga, J. Gaertig, J. Frankel, C.-C. Tsao, M. A. Gorovsky, P. J. Keeling, R. F. Waller, N. J. Patron, J. M. Cherry, N. A. Stover, C. J. Krieger, C. del

Toro, H. F. Ryder, S. C. Williamson, R. A. Barbeau, E. P. Hamilton, E. Orias, Macronuclear Genome Sequence of the Ciliate Tetrahymena thermophila, a Model Eukaryote, PLoS Biol 4 (9) (2006) e286. doi:10.1371/journal.pbio.0040286.

[7] N. A. Stover, C. J. Krieger, G. Binkley, Q. Dong, D. G. Fisk, R. Nash, A. Sethuraman, S. Weng, J. M. Cherry, Tetrahymena Genome Database (TGD): a new genomic resource for Tetrahymena thermophila research, Nucleic Acids Res 34 (suppl 1) (2006) D500–D503. doi:10.1093/nar/gkj054.

[8] W. Miao, J. Xiong, J. Bowen, W. Wang, Y. Liu, O. Braguinets, J. Grigull, R. E. Pearlman, E. Orias, M. A. Gorovsky, Microarray Analyses of Gene Expression during the Tetrahymena thermophila Life Cycle, PLoS ONE 4 (2) (2009) e4429. doi:10.1371/journal.pone.0004429.

[9] J. Xiong, X. Y. Lu, Y. M. Lu, H. H. Zeng, D. X. Yuan, L. F. Feng, Y. Chang, J. Bowen, M. Gorovsky, C. J. Fu, W. Miao, Tetrahymena Gene Expression Database (TGED): A resource of microarray data and co-expression analyses for Tetrahymena, Science China Life Sciences 54 (1) (2011) 65–67. doi:10.1007/s11427-010-4114-1.

[10] J. Xiong, Y. Lu, J. Feng, D. Yuan, M. Tian, Y. Chang, C. Fu, G. Wang, H. Zeng, W. Miao, Tetrahymena functional genomics database (TetraFGD): An integrated resource for Tetrahymena functional genomics, Database 2013 (2013) 6–11. doi:10.1093/database/bat008.

[11] J. Xiong, X. Lu, Z. Zhou, Y. Chang, D. Yuan, M. Tian, Z. Zhou, L. Wang, C. Fu, E. Orias, W. Miao, Transcriptome Analysis of the Model Protozoan, Tetrahymena thermophila, Using Deep RNA Sequencing, PLoS One 7 (2) (2012) e30630. doi:10.1371/journal.pone.0030630.

[12] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, T. S. Gardner, Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles, PLoS Biol 5 (1) (2007) 1–13. doi:10.1371/journal.pbio.0050008.

[13] J. Xiong, D. Yuan, J. S. Fillingham, J. Garg, X. Lu, Y. Chang, Y. Liu, C. Fu, R. E. Pearlman, W. Miao, Others, Gene network landscape of the ciliate Tetrahymena thermophila, PLoS One 6 (5) (2011) e20124.

[14] L. A. Stargell, K. M. Karrer, M. A. Gorovsky, Transcriptional regulation of gene expression in Tetrahymena thermophila, Nucleic Acids Res 18 (22) (1990) 6637–6639. doi:10.1093/nar/18.22.6637.

[15] G. Csárdi, A. Franks, D. S. Choi, E. M. Airoldi, D. A. Drummond, Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast, PLoS Genet 11 (5) (2015) e1005206. doi:10.1371/journal.pgen.1005206.

[16] C. Gurkan, H. Lapp, C. Alory, A. I. Su, J. B. Hogenesch, W. E. Balch, Large-scale profiling of Rab GTPase trafficking networks: the membrome, Mol Biol Cell 16 (8) (2005) 3847–3864.
URL `http://www.molbiolcell.org/content/16/8/3847.full`

[17] M. S. Behnke, J. C. Wootton, M. M. Lehmann, J. B. Radke, O. Lucas, J. Nawas, L. D. Sibley, M. W. White, Coordinated progression through two subtranscriptomes underlies the tachyzoite cycle of toxoplasma gondii, PLoS One 5 (8) (2010) 1–20. doi:10.1371/journal.pone.0012354.

[18] J. S. Briguglio, S. Kumar, A. P. Turkewitz, Lysosomal sorting receptors are essential for secretory granule biogenesis in Tetrahymena, J Cell Biol 203 (3) (2013) 537–550.

[19] S. Kumar, J. S. Briguglio, A. P. Turkewitz, An aspartyl cathepsin, CTH3, is essential for proprotein processing during secretory granule maturation in Tetrahymena thermophila., Mol Biol Cell 25 (16) (2014) 2444–60. doi:10.1091/mbc.E14-03-0833.

[20] M. Lynch, M. C. Field, H. V. Goodson, H. S. Malik, J. B. Pereira-Leal, D. S. Roos, A. P. Turkewitz, S. Sazer, Evolutionary cell biology: two origins, one objective, Proc Natl Acad Sci U S A 111 (48) (2014) 16990–16994.

[21] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M. J. L. De Hoon, Biopython: Freely available Python tools for computational molecular biology and bioinformatics, Bioinformatics 25 (11) (2009) 1422–1423. doi:10.1093/bioinformatics/btp163.

[22] J. W. Ratcliff, D. E. Metzener, Pattern-matching-the gestalt approach, Dr Dobbs Journal 13 (7) (1988) 46.

[23] E. H. Blackburn, K. M. Karrer, Genomic reorganization in ciliated protozoans, Annu Rev Genet 20 (1) (1986) 501–521.

[24] K. M. Karrer, Nuclear dualism, Methods Cell Biol 109 (2012) 29–52.

[25] M.-C. Yao, J.-L. Chao, RNA-Guided DNA Deletion in Tetrahymena: An RNAi-Based Mechanism for Programmed Genome Rearrangements, Annu Rev Genet 39 (1) (2005) 537–559. doi:10.1146/annurev.genet.39.073003.095906.

[26] M.-C. Yao, J.-L. Chao, C.-Y. Cheng, Programmed Genome Rearrangements in Tetrahymena, Microbiology Spectrum 2 (6).

[27] K. Mochizuki, N. A. Fine, T. Fujisawa, M. A. Gorovsky, Analysis of a piwi-Related Gene Implicates Small RNAs in Genome Rearrangement in Tetrahymena, Cell 110 (6) (2002) 689–699. doi:http://dx.doi.org/10.1016/S0092-8674(02)00909-1.

[28] A. T. Cowan, G. R. Bowman, K. F. Edwards, J. J. Emerson, A. P. Turkewitz, Genetic, Genomic, and Functional Analysis of the Granule Lattice Proteins in Tetrahymena Secretory Granules, Mol Biol Cell 16 (9) (2005) 4046–4060. doi:10.1091/mbc.E05-01-0028.

[29] S. Kumar, J. S. Briguglio, A. P. Turkewitz, Secretion of Polypeptide Crystals from Tetrahymena thermophila Secretory Organelles (Mucocysts) Depends on Processing by a Cysteine Cathepsin, Cth4p, Eukaryot Cell 14 (8) (2015) 817–833. doi:10.1128/EC.00058-15.

[30] M. E. Lindsay, K. Plafker, A. E. Smith, B. E. Clurman, I. G. Macara, Npap60/Nup50 is a tri-stable switch that stimulates importin-alpha:beta-mediated nuclear protein import., Cell 110 (3) (2002) 349–360. doi:S009286740200836X [pii].

[31] A. L. Buchwalter, Y. Liang, M. W. Hetzer, Nup50 is required for cell differentiation and exhibits transcription-dependent dynamics., Mol Biol Cell 25 (16) (2014) 2472–84. doi:10.1091/mbc.E14-04-0865.

[32] M. A. Nikiforov, J. F. Smothers, M. A. Gorovsky, C. D. Allis, Excision of micronuclear-specific DNA requires parental expression of Pdd2p and occurs independently from DNA replication in Tetrahymena thermophila, Genes Dev 13 (21) (1999) 2852–2862. doi:10.1101/gad.13.21.2852.

[33] M. A. Nikiforov, M. A. Gorovsky, C. D. Allis, A novel chromodomain protein, pdd3p, associates with internal eliminated sequences during macronuclear development in Tetrahymena thermophila, Mol Cell Biol 20 (11) (2000) 4128–4134. doi:10.1128/MCB.20.11.4128-4134.2000.

[34] R. M. Schwope, D. L. Chalker, Mutations in Pdd1 reveal distinct requirements for its chromodomain and chromoshadow domain in directing histone methylation and heterochromatin elimination, Eukaryot Cell 13 (2) (2014) 190–201. doi:10.1128/EC.00219-13.

[35] C. D. Malone, A. M. Anderson, J. A. Motl, C. H. Rexer, D. L. Chalker, Germ Line Transcripts Are Processed by a Dicer-Like Protein That Is Essential for Developmentally Programmed Genome Rearrangements of Tetrahymena thermophila, Mol Cell Biol 25 (20) (2005) 9151–9164. doi:10.1128/MCB.25.20.9151-9164.2005.

[36] K. Mochizuki, M. A. Gorovsky, A Dicer-like protein in Tetrahymena has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase, Genes Dev 19 (1) (2005) 77–89.

[37] T. Noto, H. M. Kurth, K. Kataoka, L. Aronica, L. V. DeSouza, K. W. M. Siu, R. E. Pearlman, M. A. Gorovsky, K. Mochizuki, The Tetrahymena Argonaute-Binding Protein Giw1p Directs a Mature Argonaute-siRNA Complex to the Nucleus, Cell 140 (5) (2010) 692–703. arXiv:NIHMS150003, doi:10.1016/j.cell.2010.02.010.

[38] C. M. Carle, H. S. Zaher, D. L. Chalker, A Parallel G Quadruplex-Binding Protein Regulates the Boundaries of DNA Elimination Events of Tetrahymena thermophila, PLoS Genet 12 (3) (2016) 1–22. doi:10.1371/journal.pgen.1005842.

[39] C. H. Rexer, D. L. Chalker, Lia1p, a novel protein required during nuclear differentiation for genome-wide DNA rearrangements in Tetrahymena thermophila, Eukaryot Cell 6 (8) (2007) 1320–1329. doi:10.1128/EC.00157-07.

[40] A. W. Y. Shieh, D. L. Chalker, LIA5 Is Required for Nuclear Reorganization and Programmed DNA Rearrangements Occurring during Tetrahymena Macronuclear Differentiation, PLoS One 8 (9) (2013) 1–15. doi:10.1371/journal.pone.0075337.

[41] M.-C. Yao, C.-H. Yao, L. M. Halasz, P. Fuller, C. H. Rexer, S. H. Wang, R. Jain, R. S. Coyne, D. L. Chalker, Identification of novel

chromatin-associated proteins involved in programmed genome rearrangements in Tetrahymena., J Cell Sci 120 (Pt 12) (2007) 1978–1989. doi:10.1242/jcs.006502.

[42] J. Bednenko, T. Noto, L. V. DeSouza, K. W. M. Siu, R. E. Pearlman, K. Mochizuki, M. a. Gorovsky, Two GW repeat proteins interact with Tetrahymena thermophila argonaute and promote genome rearrangement., Mol Cell Biol 29 (18) (2009) 5020–30. doi:10.1128/MCB.00076-09.

[43] L. Aronica, J. Bednenko, T. Noto, L. V. DeSouza, K. W. M. Siu, J. Loidl, R. E. Pearlman, M. A. Gorovsky, K. Mochizuki, Study of an RNA helicase implicates small RNA-noncoding RNA interactions in programmed DNA elimination in Tetrahymena, Genes and Development 22 (16) (2008) 2228–2241. doi:10.1101/gad.481908.

[44] A. Vogt, K. Mochizuki, A Domesticated PiggyBac Transposase Interacts with Heterochromatin and Catalyzes Reproducible DNA Elimination in Tetrahymena, PLoS Genet 9 (12). doi:10.1371/journal.pgen.1004032.

[45] I.-T. Lin, J.-L. Chao, M.-C. Yao, An essential role for the DNA breakage-repair protein Ku80 in programmed DNA rearrangements in Tetrahymena thermophila., Mol Biol Cell 23 (11) (2012) 2213–25. doi:10.1091/mbc.E11-11-0952.

[46] Y. Liu, S. D. Taverna, T. L. Muratore, J. Shabanowitz, D. F. Hunt, C. D. Allis, RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in Tetrahymena, Genes Dev 21 (12) (2007) 1530–1545. doi:10.1101/gad.1544207.

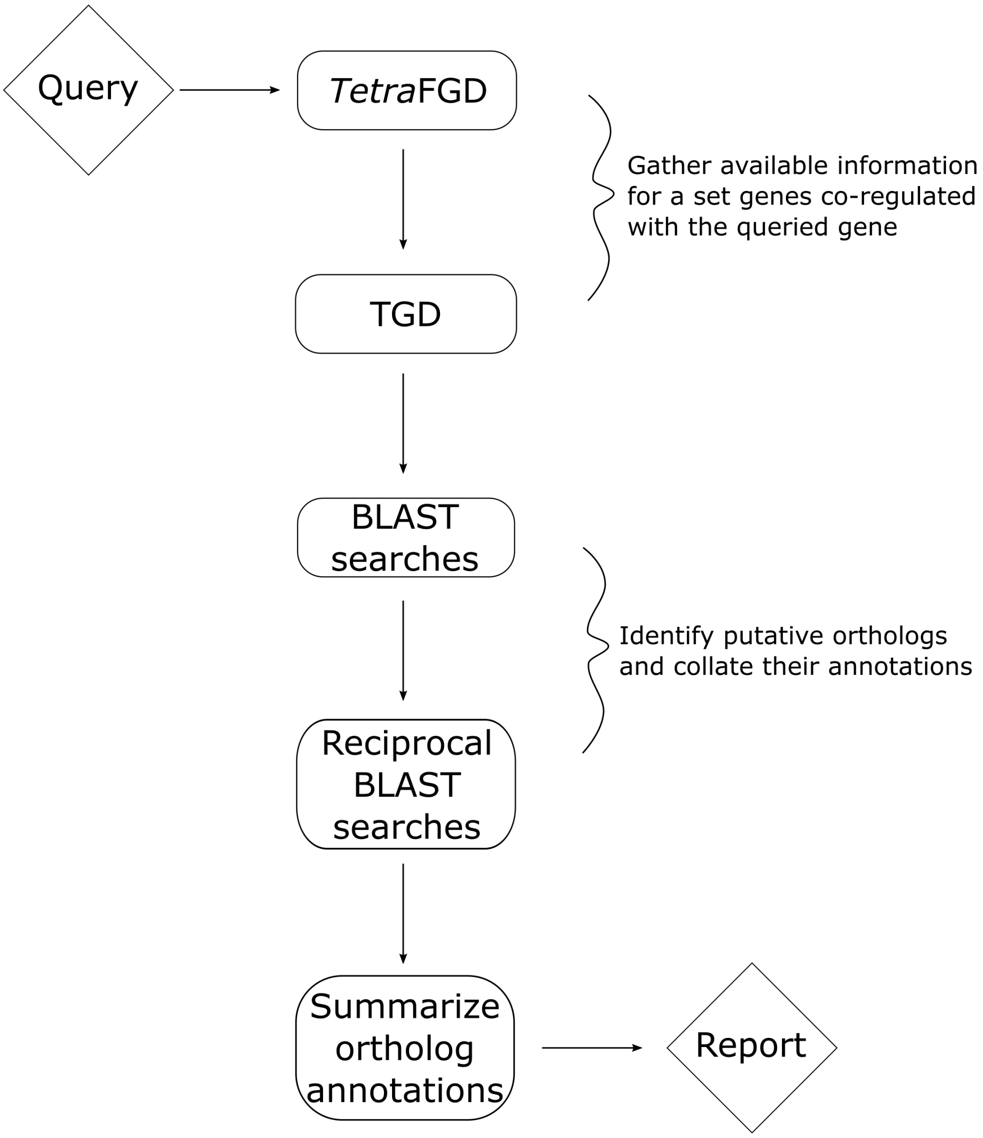
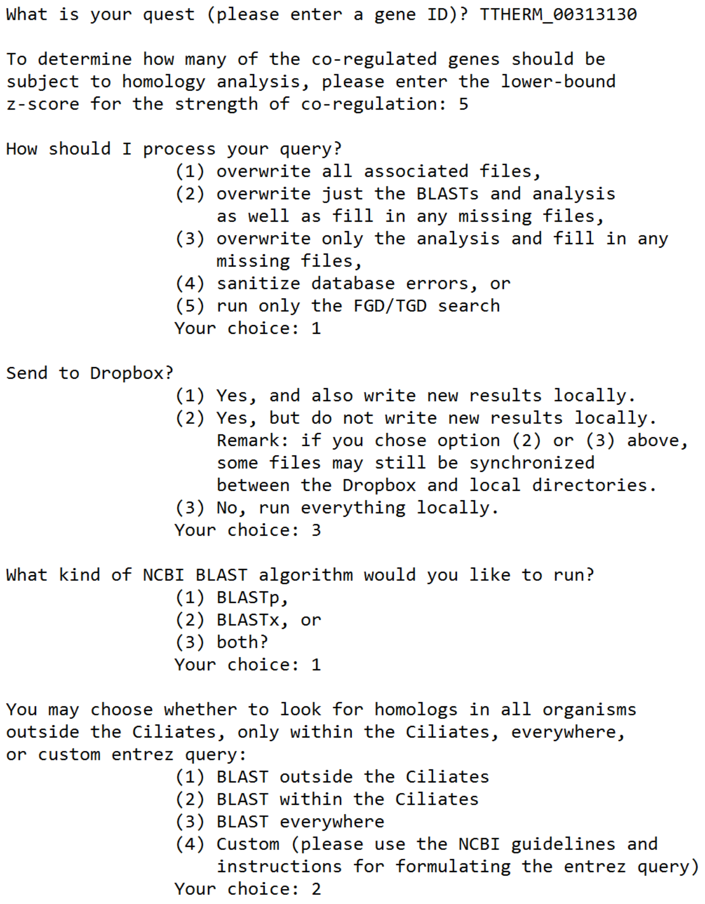

Query
*Tetra*FGD
TGD
Gather available information for a set genes co-regulated with the queried gene
BLAST searches
Reciprocal BLAST searches
Identify putative orthologs and collate their annotations
Summarize ortholog annotations
Report

Figure 1: CDH architecture. Beginning with a single *T. thermophila* gene as a query, the CDH identifies all genes that are co-regulated with it, via the TetraFGD. Next, the CDH uses the TGD to gather the annotation and sequence data for each gene in the co-regulated set. For each gene in the co-regulated set, the CDH then runs forward and reciprocal BLAST searches, through the NCBI and TGD, to identify likely orthologs. A phrase matching algorithm, based on the Ratcliff-Obershelp algorithm [22], as implemented by the python difflib library, is then used to summarize the annotations of the putative orthologs for each *T. thermophila* gene in the co-regulated set. These summaries, which provide predictions about the function (e.g., relevant biological pathway) of the *T. thermophila* gene query, are presented along with the other data gathered, in the final report.



```
What is your quest (please enter a gene ID)? TTHERM_00313130

To determine how many of the co-regulated genes should be
subject to homology analysis, please enter the lower-bound
z-score for the strength of co-regulation: 5

How should I process your query?
                (1) overwrite all associated files,
                (2) overwrite just the BLASTs and analysis
                    as well as fill in any missing files,
                (3) overwrite only the analysis and fill in any
                    missing files,
                (4) sanitize database errors, or
                (5) run only the FGD/TGD search
                Your choice: 1

Send to Dropbox?
                (1) Yes, and also write new results locally.
                (2) Yes, but do not write new results locally.
                    Remark: if you chose option (2) or (3) above,
                    some files may still be synchronized
                    between the Dropbox and local directories.
                (3) No, run everything locally.
                Your choice: 3

What kind of NCBI BLAST algorithm would you like to run?
                (1) BLASTp,
                (2) BLASTx, or
                (3) both?
                Your choice: 1

You may choose whether to look for homologs in all organisms
outside the Ciliates, only within the Ciliates, everywhere,
or custom entrez query:
                (1) BLAST outside the Ciliates
                (2) BLAST within the Ciliates
                (3) BLAST everywhere
                (4) Custom (please use the NCBI guidelines and
                    instructions for formulating the entrez query)
                Your choice: 2
```

Figure 2: Setting CDH search parameters. The CDH is run through the terminal. The CDH prompts the user to define several parameters. These are: 1) the initial gene, i.e., the query; 2) the z-score threshold to be applied as cutoff for strength of co-regulation, which determines how many of the co-regulated genes will be subject to analysis *via* homology; 3) the extent to which data gathered in prior searches should be used; 4) whether results should be stored in Dropbox; 5) whether to run BLAST searches with cDNA or protein sequences; and 6) in which taxa to run the BLAST searches. For (2), the z-score threshold determines how many co-regulated genes will be included. For example, raising the threshold increases the stringency of the requirement for strength of co-regulation, so results in fewer co-regulated genes that are subsequently analyzed via BLAST, etc. For (3), the available options are: a) to run the search from scratch, overwriting any files associated with the queried gene; b) to re-use existing data for co-regulation, annotations, and sequences, but to run all of the BLAST searches from scratch; c) to re-use any existing data that are pertinent to the given query; d) to clear NCBI database errors from a previously run search and redo the associated BLAST searches; e) to only run the search for the co-regulation, annotation, and sequence. The example query in this screenshot is set to run a CDH search for the gene TTHERM_00313130 (Sortilin 4); to consider genes that are co-regulated with it with a z-score of 5 or higher; to gather all of the data *de novo*; to save all of the data locally; and to run the BLASTp searches only within the Ciliates.

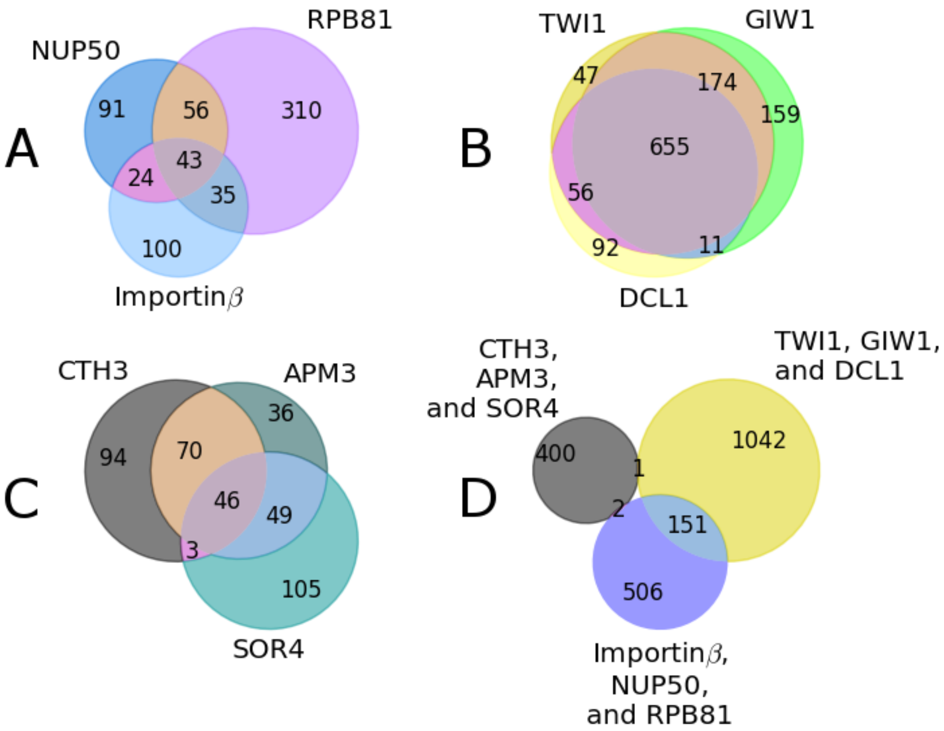


Figure 3: Using CDH outputs to assess overlap in gene function. Panels A, B, and C illustrate the overlaps in co-regulated genes for three different cellular pathways: A) nuclear import and transcriptional regulation; B) programmed genome rearrangement during cell conjugation; and C) mucocyst biogenesis. Each circle in the Venn diagrams corresponds to the full set of genes, as reported by the *Tetra*FGD, that are co-regulated with the gene indicated at the periphery of the circle. (A) *NUP50* (Nucleoporin 50) plays roles both in nuclear import and in gene transcription. The dual role of *NUP50* is reflected in the overlap of genes co-regulated with Importin$\beta$ (an import factor) and with *RPB81* (RNA Pol II subunit), a transcription factor. *NUP50*, *RPB81*, and Importin$\beta$ are mutually co-regulated. The CDH identifies 214 genes co-regulated with the nucleoporin *NUP50*, 444 genes co-regulated with *RPB81*, and 200 genes co-regulated with Importin$\beta$. (B) *TWI1* (Tetrahymena Piwi 1), *GIW1* (Gentleman in Waiting 1) and *DCL1* (Dicer-like 1) are all required for programmed genome rearrangement, and are mutually co-regulated. The CDH identifies 932 genes co-regulated with *TWI1*, 999 genes co-regulated with *GIW1*, and 814 genes co-regulated with *DCL1*. (C) *CTH3* (cathepsin 3), *APM3* ($\mu$ subunit of the adaptin 3 complex), and *SOR4* (sortilin 4) are all required for formation of mucocysts, and are mutually co-regulated. These genes also appear to have distinct cellular functions in addition to their roles mucocyst formation. For example, mucocysts are non-essential organelles, yet *CTH3* is an essential gene. The CDH identifies 213 genes co-regulated with *CTH3*, 201 genes co-regulated with *APM3*, and 203 genes co-regulated with *SOR4*. (D) Pooling all of the genes represented in A, B, and C demonstrates that there is no overlap in co-regulated genes between A and B or C, and limited overlap between B and C.

| **Gene Group** | **Genes** | **Experimental Reports in *T. thermophila*** |
|---|---|---|
| Chromo-domain | *PDD1*,<br>*PDD2*,<br>*PDD3* | *PDD1* and *PDD2* are essential for programmed genome rearrangement; *PDD3* is reported to co-localize with *PDD1*, *PDD2*, and other necessary factors [32, 33, 34]. |
| Dicer-like | *DCL1* | *DCL1* is essential for programmed genome rearrangement [35, 36]. |
| Piwi-Associated | *GIW1* | *GIW1* is essential for programmed genome rearrangement and has no known conserved function [37]. |
| Localized in nuclear anlagen | *LIA1*,<br>*LIA2*,<br>*LIA3*,<br>*LIA4*,<br>*LIA5*,<br>*LIA6* | *LIA* proteins co-localize with *PDD1* at DNA rearrangement foci; *LIA1* and *LIA5* have been shown to be necessary for genome rearrangement, and *LIA3* is required for precise excision of eliminated sequences [38, 39, 40, 41]. |
| Zinc knuckle | *cnjB* | A double-knockout of *cnjB* and *WAG1* inhibits the formation of DNA elimination structures [42]. |
| Nucleic acid helicase | *EMA1* | *EMA1* is necessary for the association of *TWI1* with chromatin [43]. |
| Transposase | *TPB2* | *TPB2* catalyzes DNA elimination [44]. |
| Ku70/Ku80 beta-barrel domain | *TKU80* | *TKU80* is necessary for both *PDD1* complex assembly and DNA repair after programmed DNA elimination [45, 26] |
| Histone lysine methyl transferase | *EZL1* | *EZL1* is necessary for H3K27 methylation and programmed DNA elimination [46]. |

Table 1: A subset of the programmed genome rearrangement factors that were identified by a CDH query for *TWI1*.

| **Gene Group** | **Genes** | **Experimental Reports in *T. thermophila*** |
|---|---|---|
| Sortilins | *SOR1*, *SOR2*, *SOR4* | *SOR2* and *SOR4* are essential for mucocyst biogenesis [18]. |
| Cathepsins | *CTH1*, *CTH2*, *CTH4* | *CTH1* and *CTH2* are involved in mucocyst biogenesis [19]. |
| Granule cargo | *GRL1*, *GRL2*, *GRL3*, *GRL4*, *GRL6*, *GRL7*, *GRL9*, *GRT1*, *IGR6*, *IGR7* | These genes, belonging to two gene families, encode mucocyst contents [28]. |
| Adaptin complex subunits | *AP3* $\mu$ and $\beta$ subunits | *AP3* $\mu$ subunit (*APM3*) is necessary for mucocyst biogenesis (Kaur et al., submitted). |
| Syntaxins | *STX7L1* | *STX7L1* is essential for mucocyst biogenesis (Kaur et al., submitted). |

Table 2: Mucocyst biogenesis and cargo factors that were identified by a CDH query for *CTH3*.

| **Queried Gene** | ***TWI1*** | ***CTH3*** |
|---|---|---|
| **Number co-regulated genes** | 932 | 213 |
| **Taxa BLASTed** | Only Ciliates | Excluding Ciliates |
| **Number previously annotated** | 430 | 88 |
| **Annotations matched** | 299 | 43 |
| **Annotations expanded upon** | 19 | 19 |
| **Novel annotations** | 126 | 34 |

Table 3: The CDH accurately reproduces existing annotations and provides new annotations at a high rate.

## Required Metadata

## Current code version

## Current executable software version

| **Nr.** | **Code metadata description** | **Please fill in this column** |
|---|---|---|
| C1 | Current code version | v1.1.0 |
| C2 | Permanent link to code/repository used for this code version | *https* : *//bitbucket.org/ltsypin/cdhproject/* |
| C3 | Legal Code License | GNU GPL v3 |
| C4 | Code versioning system used | git |
| C5 | Software code languages, tools, and services used | python |
| C6 | Compilation requirements, operating environments & dependencies | any system that can run python 2.7 |
| C7 | If available Link to developer documentation/manual | *http* : *//ciliate.org/index.php/show/CDH* |
| C8 | Support email for questions | coregulationdataharvester@gmail.com |

Table 4: Code metadata (mandatory)

| **Nr.** | **(Executable) software metadata description** | **Please fill in this column** |
|---|---|---|
| S1 | Current software version | 1.1.0 |
| S2 | Permanent link to executables of this version | *http* : *//ciliate.org/index.php/show/CDH* |
| S3 | Legal Software License | GNU GPL v3 |
| S4 | Computing platforms/Operating Systems | OS X (10.6+) and Windows (x64) Vista or later |
| S5 | Installation requirements & dependencies | |
| S6 | If available, link to user manual - if formally published include a reference to the publication in the reference list | *http* : *//ciliate.org/index.php/show/CDH* |
| S7 | Support email for questions | coregulationdataharvester@gmail.com |

Table 5: Software metadata (optional)