Phylogenomics of the epigenetic toolkit reveals punctate retention of genes across eukaryotes

Agnes K.M. Weiner[1], Mario A. Cerón-Romero[1,2], Ying Yan[1], Laura A. Katz[1,2,*]

[1]Department of Biological Sciences, Smith College, 44 College Lane, Northampton, Massachusetts, 01063, USA

[2]Program in Organismic and Evolutionary Biology, University of Massachusetts Amherst, Amherst, Massachusetts, 01003, USA

*Corresponding author: Laura A. Katz, Department of Biological Sciences, Smith College, 44 College Lane, Northampton, Massachusetts, 01063, USA, lkatz@smith.edu, 413-585-3825

Keywords: epigenetics, chromatin modification, non-protein-coding RNA, macroevolution, LECA, protists

Author Contributions: AKMW and LAK designed the study, AKMW and YY produced the transcriptome data, AKMW, YY and MACR analyzed the data, all authors contributed to writing the manuscript and approved its final version.

Abstract

Epigenetic processes in eukaryotes play important roles through regulation of gene expression, chromatin structure and genome rearrangements. Mechanisms such as chromatin modification (e.g. DNA methylation, histone modification) and non-protein-coding RNAs (npc-RNAs) have been well studied in animals and plants. With the exception of a few model organisms (e.g. *Saccharomyces*, *Plasmodium*), much less is known about epigenetic toolkits across the remainder of the eukaryotic tree of life. Even with limited data, previous work suggested the existence of an ancient epigenetic toolkit in the last eukaryotic common ancestor (LECA). We use PhyloToL, our taxon-rich phylogenomic pipeline, to detect homologs of epigenetic genes and evaluate their macroevolutionary patterns among eukaryotes. In addition to data from GenBank, we increase taxon sampling from understudied clades of SAR (Stramenopila, Alveolata and Rhizaria) and Amoebozoa by adding new single-cell transcriptomes from ciliates, foraminifera and testate amoebae. We focus on 118 gene families, 94 involved in chromatin modification and 24 involved in npc-RNA processes based on the epigenetics literature. Our results indicate: 1) the presence of a large number of epigenetic gene families in LECA; 2) differential conservation among major eukaryotic clades, with a notable paucity of genes within Excavata; and 3) punctate distribution of epigenetic gene families between species consistent with rapid evolution leading to gene loss. Together these data demonstrate the power of taxon-rich phylogenomic studies for illuminating evolutionary patterns at scales of >1 billion years of evolution and suggest that macroevolutionary phenomena, such as genome conflict, have shaped the evolution of the eukaryotic epigenetic toolkit.

Significance statement

Eukaryotic organisms evolved complex epigenetic processes to orchestrate gene expression and genome dynamics. By applying a taxon-rich phylogenomic approach, including adding transcriptome data from several lineages of understudied microeukaryotes, we identify homologs of the epigenetic gene toolkit in diverse lineages across the eukaryotic tree of life. We show that gene families involved in chromatin modification and the processing of non-protein-coding RNAs originated in the last eukaryotic common ancestor (LECA). However, the distribution of epigenetic genes across eukaryotes now reflects a punctate pattern, with differential conservation of genes across taxonomic lineages and functional categories. This suggests that macroevolutionary phenomena, such as genome conflict and/or adaptations to diverse living styles, shaped the epigenetic toolkit in eukaryotes.

Introduction

Throughout the last decades it has become increasingly clear that epigenetic modifications play major roles in regulating the expression of the genotype in a wide range of eukaryotic taxa (e.g. Wolffe and Matzke 1999; Bird 2007; Goldberg, et al. 2007). The existence of epigenetic mechanisms expands upon the idea of a linear relationship between genotypes and phenotypes, and can challenge Mendelian inheritance of genes (e.g. Katz 2006). Epigenetics can modify gene expression, including completely silencing genes and mobile genetic elements, and also be responsible for altering genome structures (e.g. Bernstein and Allis 2005; Heard and Martienssen 2014). The effects of these epigenetic processes range from cell differentiation to genomic imprinting and, in cases where they malfunction, disease (e.g. Jiang, et al. 2004; Gluckman, et al. 2009; Handel, et al. 2010). Epigenetics also plays a role in shaping genome architectures through DNA rearrangement/elimination and polyploidization in diverse lineages of eukaryotes (e.g. Liu and Wendel 2003; Maurer-Alcalá and Katz 2015). In addition to impacting individual cells or organisms, epigenetics likely also acts across generations, influencing the evolution of populations and species (e.g. Smith and Ritchie 2013; Smith, et al. 2016) and may contribute to rapid adaptive responses (e.g. Rey, et al. 2016). Overall, its effects can be summed up as creating a variety of phenotypes from the same genotype.

The term “epigenetics” was first introduced by Waddington (1942) to refer broadly to the expression of the phenotype during development. Ever since, its definition has been subject to intense discussion (e.g. Haig 2004; Bird 2007; Goldberg, et al. 2007; Stotz and Griffiths 2016) and generally includes both well-known processes (i.e. histone modifications, DNA methylation) as well as a variety of poorly known genetic phenomena (i.e. paramutation, transgenerational effects). Today’s textbook definition is that epigenetics refers to heritable phenotypic changes that arise without change in the underlying DNA sequence (e.g. Tollefsbol 2017). However,

here we use Denise Barlow's broader definition of epigenetics as "all the weird and wonderful things that cannot [yet] be explained by genetics" (McVittie 2006).

The molecular processes of epigenetics can be roughly assigned to two classes: chromatin modifiers (e.g. DNA methylation, histone modifications; e.g. Razin and Riggs 1980; Ng and Bird 1999) and non-protein-coding RNAs (npc-RNAs, RNA interference: microRNAs, Piwi interacting RNAs and small interfering RNAs; e.g. Sharp 2001; Shabalina and Koonin 2008; Peng and Lin 2013; Bond and Baulcombe 2014). Of the two classes, chromatin modifiers are currently understood more deeply. Through mechanisms such as the addition or removal of methyl or acetyl groups to nucleotides or histones, chromatin modifiers can silence or activate genes by producing physical changes to chromatin accessibility (e.g. Fuks 2005). A large number of enzymes is known to be involved in these processes, including DNA and histone methyltransferases, histone acetyltransferases and deacetylases as well as the members of the Polycomb-group proteins (e.g. Fuks 2005; Zemach and Zilberman 2010; Maumus, et al. 2011; Di Croce and Helin 2013; Aravind, et al. 2014; Rastogi, et al. 2015; Vogt 2017).

In contrast, npc-RNAs, act through sequence-specific gene silencing and their targets include viral genes, transposons, and eukaryotic genes in both germline and somatic cells (Shabalina and Koonin 2008; Peng and Lin 2013). They have been argued to have originated in genome screening and defense (Obbard, et al. 2009). Based on previous analyses, the genes involved in generating npc-RNAs appear widespread across eukaryotes and the most prominent members include *ARGONAUTE*, *PIWI*, the RNases III *DROSHA* and *DICER* as well as RNA-dependent RNA polymerases (RdRps) and RNA helicases (Sharp 2001; Peng and Lin 2013; Li and Patel 2016).

Though epigenetic processes are best understood in plants and animals, many components of the epigenetic toolkit are also found in other lineages across the eukaryotic tree of life (e.g. Maurer-Alcalá and Katz 2015) and an extensive epigenetic machinery was likely present already in the last eukaryotic common ancestor (LECA) as key elements can be traced

back to prokaryotic systems of secondary metabolism and genome conflict (Iyer, et al. 2008; Aravind, et al. 2014). Authors such as Fedoroff (2012), Lisch (2009) and Klobutcher and Herrick (1997) have also hypothesized that epigenetic processes originally arose as a means to restrict the spreading of transposable elements within genomes and only later were their roles expanded to other dynamic genome processes. Despite the importance of epigenetics for the development and evolution of eukaryotic lineages, knowledge on these processes in non-model lineages remains scarce. Especially for many clades of microbial eukaryotes, including Rhizaria, Amoebozoa and diverse ciliates, details on epigenetic gene families remain unknown, even though these groups are known for complex genome dynamics that likely involve epigenetics (e.g. Parfrey, et al. 2008; Croken, et al. 2012).

The combination of advances in single-cell 'omics (e.g. Kolisko, et al. 2014; Saliba, et al. 2014), large-scale sequencing (e.g. Massana, et al. 2015), and phylogenomics (e.g. Ceron-Romero, et al. 2019) now allow for easy access and exploration of data from uncultivable microeukaryotes. Among the clades with the greatest paucity of data are Amoebozoa, Rhizaria, and Ciliophora (with the exception of models such as *Tetrahymena* and *Paramecium*; Maurer-Alcalá, et al. 2018), which are now included in this study. Though genomes are well-sampled for pathogens (e.g. *Acanthamoeba, Entamoeba)* and model lineages (e.g. *Physarum, Dictyostelium)* within Amoebozoa, clades such as the shell-building Arcellinida lack 'omics data. The situation is similar within the Rhizaria, where the lack of human parasites within this major eukaryotic clade likely contributes to the dearth of data (Grattepanche, et al. 2018).

To investigate macroevolutionary patterns of the epigenetic toolkit across the eukaryotic tree of life, we analyze epigenetic gene families using PhyloToL (Ceron-Romero, et al. 2019). PhyloTol was specifically designed for the investigation of the heterogenous evolutionary patterns in diverse eukaryotic clades, spanning 1.8 billion years of evolution. We combine PhyloToL with a taxon-rich dataset to assess homology and generate both multiple sequence alignments (MSA) and gene trees. PhyloToL (Ceron-Romero, et al. 2019) also allows for the

137 removal of contaminants that are frequent in ‘omics datasets. For our analyses, we included a
138 maximum of 278 transcriptomes and 182 genomes representing 460 species from all major
139 eukaryotic clades. We also include a limited set of 89 bacterial genomes and 25 archaeal
140 genomes. In addition to the genomes and transcriptomes obtained from publicly available
141 databases, such as GenBank and OrthoMCL, we added single-cell transcriptomes from diverse
142 clades of microbial eukaryotes for understudied taxa from Amoebozoa and SAR (Stramenopila,
143 Alveolata and Rhizaria) in order to improve taxonomic coverage. We analyzed a total of 118
144 epigenetic gene families that are involved in either chromatin modification or npc-RNAs. Our
145 intention is to characterize the distribution of the epigenetic toolkit across the eukaryotic tree of
146 life, especially targeting microbial eukaryotic clades that remain understudied.
147
148 **2. Results**
149 **2.1 Distribution of the epigenetic toolkit across major eukaryotic clades**
150 Based on the literature, we analyzed 179 genes in the eukaryotic epigenetic toolkit as
151 those that play major roles in either chromatin modification or npc-RNA processes. These 179
152 genes fall into 118 gene families as defined by the database OrthoMCL (Li, et al. 2003; Tables
153 1, S1), which is the starting point for gene family delineation in PhyloToL (Ceron-Romero, et al.
154 2019). This focal set of genes is both incomplete and biased as epigenetics has so far been
155 best studied in plants (e.g. Finnegan, et al. 1998; Rapp and Wendel 2005), animals (e.g.
156 Fazzari and Greally 2004; Glastad, et al. 2011) and only a few other eukaryotic lineages (e.g.
157 Grewal 2000; Aramayo and Selker 2013).
158 To evaluate the distribution of the epigenetic toolkit across eukaryotes, we analyzed the
159 presence/absence of the 118 gene families in up to 574 species sampled from all major
160 eukaryotic clades plus a limited number of bacteria and archaea (**Tables 2, S2**). The dataset
161 includes 69 newly-sequenced transcriptomes of six species of Arcellinida (Amoebozoa), three
162 species of Ciliophora (Alveolata) and 14 species of Rhizaria, which substantially increases

taxonomic coverage for these understudied clades (sequences available at GenBank SRA BioProject PRJNA637648). To assess the impact of taxon sampling on macroevolutionary patterns, we compared the results obtained for four different datasets: 1) ALL: all 574 taxa that passed the quality cut-off; 2) INFORMED: taxonomically-informed 'even' subsample across clades with 25 taxa each; 3) RANDOM: random subsample of 25 taxa per major eukaryotic clade; and 4) GENOME: only taxa for which we had whole genome data (232 total), which allowed us to rule out missing data in transcriptomes as a major driver of the observed patterns. All 118 gene trees were generated for the taxon sets ALL, INFORMED and GENOME. The RANDOM set, on the other hand, failed to produce a gene tree for the methyl-DNA binding protein MECP2 (OG5_140477), as this gene family had too few taxa for tree inference.

The sizes of the gene trees are highly variable (**Figure 1**), indicating complex patterns of distribution of the toolkit across eukaryotic lineages. Among the three larger datasets (ALL, INFORMED and RANDOM), we observe a consistent pattern of presence/absence of gene families across major clades. For example, all three datasets yielded similar numbers of gene families that seem to have existed already before the last eukaryotic common ancestor (pre-LECA; 20-28 gene families, defined as present in all but one major eukaryotic clade, bacteria and/or archaea) or that were present in the LECA (34-39 gene families, defined as present in all but one major eukaryotic clade, **Table S3**). This indicates that taxon choice did not have a substantial impact on our interpretation. The only exception is the GENOME dataset that generally shows lower values (**Table S3**), which corresponds to the low number of whole genomes available for some major clades (e.g. only two whole genomes were publicly available for Rhizaria and eight for Amoebozoa, **Table S2**). Given the overall similarity among datasets, we provide the results for all four datasets in the supplementary files (**Table S4**) and focus the rest of our study on results from the INFORMED subsample where the even distribution of species allows better comparisons across major clades.

Overall, patterns of conservation of epigenetic gene families are complex (**Figure 2**). As expected, given the relatively large number of studies, Opisthokonta (Op) and Archaeplastida (Pl) contain the highest number of gene families with 109 and 97 out of 118, respectively. We identified 86 gene families in Amoebozoa (Am), 85 in Rhizaria (Rh), 83 in Stramenopila (St), 76 in Alveolata (Al) and 84 among the non-monophyletic orphan lineages (i.e. EE "everything else"). A striking difference is that the 25 species within Excavata (Ex) contain only 53 gene families, the smallest number among all eukaryotic clades (**Figure 2**). Bacteria (Ba) and archaea (Za) only contain a few of the gene families analyzed, which is as expected given the eukaryotic focus of this study.

We identified three distinct patterns from the presence/absence analysis of gene families in major eukaryotic clades (**Figure 2**): i) Pre-LECA gene families that are present in six of the seven eukaryotic clades (Op, Pl, Al, St, Rh, Am and/or Ex) as well as in bacteria and/or archaea, ii) LECA gene families that are present in six of the seven major eukaryotic clades but absent in the sampled bacteria/archaea; and iii) the remaining gene families that are found in one to five of the eukaryotic clades. In total, 21 of the 118 gene families meet the pre-LECA criteria for the INFORMED taxon selection (**Figure 2**). Of these, 17 gene families are part of the 94 gene families involved in chromatin modification pathways and the remaining four are among the 24 gene families involved in npc-RNA processes. A total of 39 of the 118 gene families can be assigned to the LECA, of which 31 have functions related to chromatin modification and eight to npc-RNAs (**Figure 2, Tables S1, S4, S5**). The remaining 58 gene families have variable distributions among the major eukaryotic clades (49 of 58, >1 MC label **Figure 2; Tables S4, S5**) or are specific to a certain major clade (nine of 58, 1 MC label **Figure 2**). Of these, 46 gene families are involved in chromatin modification and 12 in npc-RNA processes.

We further assessed the relationship of gene function and patterns of conservation (**Figure 3**). Of the gene families belonging to chromatin modification pathways, the degree of conservation appears to depend on function: lysine deacetylases and acetyltransferases show a

high degree of conservation, as the majority of gene families in these categories are designated to pre-LECA/LECA (90% and 80%, respectively). Lysine demethylases, arginine methyltransferases and a group of other histone-modification proteins all have around 50% of their respective gene families likely present in pre-LECA/LECA. In contrast, lysine methyltransferases only have 45% pre-LECA/LECA gene families and the Polycomb-related gene families show the least degree of conservation among the chromatin modifiers with only 25% present in the LECA. Instead, 42% of the Polycomb-related gene families are in fewer than six but more than one major eukaryotic clade and 33% are even restricted to one clade (**Figure 3**). For the npc-RNA related gene families, 50% are conserved as they fall in the pre-LECA and LECA datasets, whereas DNA methylation gene families are a less conserved functional class with 26% of the gene families in pre-LECA/LECA, 53% in between one to five major eukaryotic clades and 20% in only one major clade.

2.2 Distribution of the epigenetic toolkit at the species level

To assess species-specific patterns of gene family presence/absence, we repeated the analysis on the 250 species in the INFORMED dataset and mapped the data onto a phylogeny generated from a concatenation of 391 housekeeping gene families (non-epigenetic genes that are widespread across eukaryotes and likely were present already in or before the LECA, **Figure 4**). First, we evaluated the quality of our data by assessing presence/absence of 118 housekeeping gene families (i.e. the same number as in our epigenetic set) that we chose randomly from among the 391 gene families used for the phylogenomic analysis (see methods). The housekeeping gene families are present in almost all species sampled here, demonstrating the overall good quality of data in our INFORMED dataset, which includes 121 transcriptomes among the 250 species (**Figure 4**). Though four of the 200 eukaryotic species contained none or only one of the 118 epigenetic gene families, some of the other species with only transcriptome data are among the samples with the greatest numbers of gene families (**Table**

S2). The INFORMED dataset contains the newly generated transcriptomes of five species of Arcellinida (Amoebozoa) and 10 species of Rhizaria, which is a subset of our newly added transcriptome data as described above. Orphan lineages like the Apusozoa and *Malawimonas* have both few epigenetic and few housekeeping gene families, suggesting data quality plays a role here.

At the species level, the same overall pattern emerges as for the level of major clades, with the greatest numbers of gene families found within species of Opisthokonta and/or Archaeplastida and the fewest among Excavata (**Figure 4**, **Table S2**). Among Opisthokonta, animal species show a high degree of similarity in the composition of their epigenetic toolkits (**Figure 4**). The same is true for the species of fungi, yet compared to animals their toolkit contains fewer gene families. Among Archaeplastida, the toolkit of green algae is homogeneous across species and can be differentiated from the toolkit of the red algae and glaucophytes (**Figure 4**, **Tables S4, S5**). The three SAR clades as well as the Amoebozoa appear similar in the composition of their toolkits and there are no obvious lineage specific patterns given our taxon sampling. As with the clade-based analyses, the size of the Excavata toolkit is overall smaller than in other eukaryotes, with Euglenozoa and the other Excavata showing a distinctive subset of gene families (**Figure 4**, **Tables S4, S5**).

2.3 Punctate distribution of many epigenetic gene families

We observe a punctate distribution pattern among eukaryotes for many epigenetic gene families. Here, punctate refers to gene families that are widespread across eukaryotic lineages (i.e. present in 3 or more major clades), and yet are found in only a small number of species per major clade. Among the pre-LECA/LECA gene families (i.e. those present in at least six and often all major clades) there are cases were gene families are retained in only 24 out of the 250 species (i.e. the gene family OG5_135026, RNA helicase). This punctate pattern can be seen in some individual gene trees (**Figure 1b**) as well as in the presence/absence data at the species

level (**Figure 4**). The punctate pattern is apparent when the presence/absence data for the epigenetic gene families are compared to the housekeeping gene families, which show a more homogeneous distribution across the same eukaryotic species (**Figure 4**).

Two possibilities to explain the punctate distribution of gene families include: 1) functional constraints are similar across lineages but gene loss is higher among epigenetic genes than housekeeping genes; and 2) punctate genes are evolving rapidly such that homologs now fail to meet the criteria for homology-assessment necessary to generate MSAs and gene trees. To distinguish between these possibilities, we calculated the average branch length for each of the gene trees for the epigenetic gene families and compared them to our housekeeping gene set. In the first scenario (i.e. change in pattern of gene loss), branch lengths from nodes to tips may not be significantly different, while in the second case (i.e. rapid evolution of epigenetic genes), branch lengths are expected to be longer. For this, we classified the epigenetic trees in three categories, big (>100 sequences), medium (26–100 sequences) and small (≤25 sequences). While the big trees and the housekeeping gene trees have similar branch lengths, the medium and small trees have increasingly longer average branch lengths (**Figure 5**).

To compare mean branch lengths across these trees, we used a parametric test, Welch's t-test. The data points of the three epigenetic categories showed a normal distribution according to a Shappiro Wilk test (big: $p > 0.5$ and $n = 31$, middle: $p > 0.8$ and $n = 60$, small: $p > 0.4$ and $n = 27$) and QQ plots (**Figure S1**). In contrast, the housekeeping gene families do not fit expectations for normal distribution ($p < 0.005$ and $n = 391$; **Figure S1**), which is likely due to the large number of data points that lead to a high sensitivity to deviations from normality. Under Welch's t-test, the means of each category (i.e. housekeeping, big, medium, small; **Figure 5**) are statistically significantly different from every other category (**Table S6**).

To address the possibility that we failed to include rapidly evolving members of smaller gene families, we used BLAST to identify additional sequences for three of the npc-RNA gene families (*DICER*, *PIWI*, *ARGONAUTE*), but found that few added genes survived Guidance analysis, the MSA tool we use to assess homology (see methods). For example, an alignment of sequences from 11 gene families that we identified as potential *DICER* homologs did not survive Guidance **(Table S7)**. When we forced the genes to align using MAFFT and checked the result by eye, we saw little evidence of homology, consistent with either rapid evolution or the independent origin of these genes. We saw a similar result for *PIWI* genes: combining eight potential homologs, only three survived Guidance and the resulting tree indicated deep divergence between gene families consistent with ancient paralogy rather than lost nested homologs (**Figure S2, Table S7**). The forced alignment of the potential *ARGONAUTE* homologs retained six out of eight gene families that fall into two clades in the tree (**Figure S2, Table S7**). However, each of the taxa present is also represented by the "main" *ARGONAUTE* gene family and so inclusion of the divergent genes would not have changed our assessment of presence/absence of this gene family. In sum, manually combining additional gene families does not add any further information to the macroevolutionary patterns of the epigenetic genes.

2.4 Paralogs

We find a trend towards higher numbers of sequences per species per gene family (i.e. paralogs) in the housekeeping genes than in the epigenetic genes, though the absolute number of paralogs is confounded here by observation of only highly expressed genes in the transcriptome data. We repeatedly subsampled 60 gene families (100 repetitions) from the housekeeping dataset and compared them to the 60 pre-LECA/LECA epigenetic gene families. The overall trend of more sequences in the housekeeping gene families was significant for 93 out of the 100 iterations of the analysis (Sign test, H_a: epigenetic < housekeeping, $p < 0.05$, **Table S8**). The major clades responsible for this trend are: Stramenopiles, Rhizaria,

317 Archaeplastida, Excavata and Amoebozoa (Mann-Whitney, H_a: epigenetic < housekeeping, $p <$
318 0.05 for more than 65/100 iterations). While Alveolata show no evident trend with high data
319 dispersion, Opisthokonta show the opposite trend with more sequences in the epigenetic genes
320 than in the housekeeping genes (Mann-Whitney, H_a: epigenetic > housekeeping, $p < 0.05$ for
321 more all 100 iterations; **Table S8**).
322
323 **3. Discussion**
324 Our taxon-rich analyses yield three main insights: 1) a rich epigenetic toolkit existed in
325 the LECA, containing genes for both chromatin modification and npc-RNA processes; 2) the
326 toolkit is differentially conserved among major eukaryotic clades with a notable paucity of genes
327 within Excavata; and 3) in contrast to the housekeeping gene families, many epigenetic gene
328 families show a punctate distribution in that they are widespread across eukaryotes but retained
329 in only a few species.
330
331 *Presence of the epigenetic toolkit in the LECA*
332 Since epigenetic processes play fundamental roles in many eukaryotes, several authors
333 have proposed the existence of a widespread, ancient epigenetic toolkit (e.g. Cerutti and Casas-
334 Mollano 2006; Parfrey, et al. 2008; Shabalina and Koonin 2008; Aravind, et al. 2014; Maurer-
335 Alcalá and Katz 2015). Previous analyses have largely focused on a narrow sampling of
336 lineages (e.g. animals and plants; Finnegan, et al. 1998; Fazzari and Greally 2004; Rapp and
337 Wendel 2005; Glastad, et al. 2011), leaving the majority of eukaryotic diversity understudied.
338 However, data from a limited sample of microeukaryotes and phylogenomic approaches
339 suggested that epigenetics is not restricted to multicellular organisms, but present in microbial
340 lineages as well and may indeed have been present already in the LECA (e.g. Aravind, et al.
341 2014). Epigenetic processes play a role in the complex genome dynamics of microbial lineages,
342 such as changes in ploidy level (up to thousand copies of the genome) in some lineages of

Rhizaria and Alveolata (Parfrey, et al. 2008) and/or separation of the genome into germline and soma within one cell (e.g. Ciliophora; Prescott 1994; Katz 2001). Other lineages have a parasitic lifestyle that involves frequent changes to their chromatin structures and gene expression profiles (e.g. Croken, et al. 2012), which have been shown to be influenced by epigenetic processes as well (e.g. Liu, et al. 2007; Cortes, et al. 2012; Croken, et al. 2012; Chalker, et al. 2013). Yet, for many microbial eukaryotic lineages it remained unclear if these processes and the underlying epigenetic genes correspond to gene families present in animals and/or plants, or if they evolved independently.

Our taxon-rich phylogenomic approach allows us to provide a more detailed depiction of the conservation of epigenetic processes across eukaryotes, and supports the hypothesis of a toolkit in the LECA as all major eukaryotic clades contain gene families of all functional categories as defined in this study (**Figure 6; Table S5**). Coupling PhyloToL (Ceron-Romero, et al. 2019), which allows rapid homology assessment and the generation of MSAs and gene trees, with single-cell transcriptome data of uncultivable microbial eukaryotes in Rhizaria, Amoebozoa and ciliates allowed us to provide additional detail to the evolution of eukaryotic epigenetic gene families.

Our analyses indicate that the retention of epigenetic genes varies by functional categories, with gene families related to histone modifications, especially acetylation and deacetylation, being overrepresented in pre-LECA/LECA while the Polycomb-group proteins and DNA methylation genes are retained in fewer lineages (e.g. **Figure 3**, **Table S5**). Gene families involved in processes like lysine acetylation/deacetylation are used in post-translational modifications in bacteria and archaea (e.g. Christensen, et al. 2019) and have been coopted to serve in chromatin modification in eukaryotes. The Polycomb-group proteins, on the other hand, appear to be a eukaryotic invention as members such as the protein SUZ, chromobox proteins (CBX), enhancer of zeste (EZH) and the Polycomb-group ring finger proteins (PCGF) are found only among eukaryotes (**Tables S1, S4**). Early work on Polycomb-group proteins demonstrated

their roles in cell differentiation and development and so they were originally assumed to be restricted to multicellular lineages (animals and plants; e.g. Kohler and Villar 2008). However, core components of the Polycomb Repressive Complex 2 (PRC2) also exist in unicellular eukaryotes, such as the green alga *Chlamydomonas* and the diatom *Thalassiosira* (Shaver, et al. 2010). Our analysis extends on this as we find PRC2 components (e.g. Nurf55, ESC, EZH; **Tables S1, S4**) in a wide range of unicellular lineages (e.g. especially among Stramenopila and Rhizaria). The most parsimonious explanation, therefore, is that a basic set of Polycomb-group proteins was already present in the LECA, and has been lost or has evolved rapidly and beyond recognition where they appear absent. Intriguingly, some have argued that Polycomb-group proteins originated as defense against mobile genetic elements and only later they took on the more specific roles in multicellular lineages (Shaver, et al. 2010). For DNA methylation systems it has been suggested that they may have been transferred from bacteria to eukaryotes several times independently and that some components may have been lost in individual lineages (Ponger and Li 2005; Iyer, et al. 2008; Zemach and Zilberman 2010). Our study supports this idea, since – despite much wider taxon sampling – we also observe the DNA methylation gene families to be less widespread across eukaryotes (**Figures 3, 6**).

Smaller toolkit size in the Excavata

Phylogenomic analyses demonstrate a notable paucity of genes among Excavata, despite the fact that complete genomes exist for many of these species (i.e. we can rule out failure to detect signal from incomplete transcriptome data; **Table S2**). Excavata lack the majority of Polycomb-group gene families, which are also sparse in other major eukaryotic clades (**Figure 6**). More surprising, most Excavata also lack gene families with conserved functions related to methylation (e.g. lysine methyltransferases and demethylases, DNA methylation; **Figure 6, Table S4**). The smaller toolkit size in Excavata could be due to several factors discussed in detail below: 1) Excavata exhibit unusual genome structures, suggesting

that their chromatin may be regulated differently; 2) the parasitic and thus often anoxic/microaerophilic lifestyle of many sampled Excavata may be incompatible with epigenetic processes involving methylation some of which require oxygen; or 3) if Excavata are at the root of the eukaryotic tree of life (He, et al. 2014), some functions of the epigenetic toolkit may have expanded after their divergence.

Unusual genome structures within Excavata may underlie the smaller number of epigenetic gene families. Among the Excavata, members of the Kinetoplastida exhibit an unusual genome organization, with protein-coding genes arranged in large polycistronic transcription units that are processed post-transcriptionally through trans-splicing (e.g. Belli 2000; El-Sayed, et al. 2005; Clayton 2019). In addition, histone sequences in Excavata, and especially of the Trypanosomatids, are highly divergent from those of other eukaryotes (Sullivan, et al. 2006). These structural peculiarities suggest that processes underlying chromatin modification in Excavata may also be divergent from other eukaryotes. Even though histone modifications governed by epigenetic processes exist within Excavata, the specific patterns of these marks, i.e. the “histone code”, differ from conserved eukaryotic patterns (Sullivan, et al. 2006; Croken, et al. 2012). Elias and Faria (2009) do report roles of npc-RNA processes in gene regulation in some Trypanosomatids. While we find support for the existence of some npc-RNA gene families in Excavata, some such as *ARGONAUTE* are represented by divergent “*ARGONAUTE*-like” gene family (OG5_149426) instead of the more widespread *ARGONAUTE* gene family (OG5_127240; **Table S4**). Together, these data suggest unusual genome structures may have led to divergent epigenetic strategies in Excavata.

A second possible explanation for the smaller set of epigenetic gene families within Excavata is that gene families underlying methylation processes (e.g. the DNA methylase DNMT and lysine demethylases KDM; **Table S4**) may have been reduced in parasites that can live in low-oxygen environments. For example, DNA methylation seems to be absent in the Excavata genus *Giardia* (Lagunas-Rangel and Bermudez-Cruz 2019), whereas histone

421 acetylation and npc-RNAs are important for its encystation and expression of surface proteins
422 for host immune evasion (Prucca, et al. 2008; Carranza, et al. 2016; Ortega-Pierres, et al.
423 2018). Similar patterns are found in other anaerobic parasites, such as *Trypanosoma gondii*
424 (Excavata), and even two Apicomplexans (Plasmodium and Cryptosporidium, Alveolata;
425 Croken, et al. 2012). In human tumor cells and germinating rice, low or anoxic conditions lead to
426 aberrant DNA methylation patterns, suggesting that these epigenetic processes require oxygen
427 as substrate (Bhandari, et al. 2017; Narsai, et al. 2017; Camuzi, et al. 2019). Together these
428 data suggest that the anaerobic life style of many Excavata may have an influence on the
429 composition of the epigenetic toolkit similar to how a microaerophilic lifestyle is thought to be
430 related to altered genome structures and gene expression in a range of human parasites
431 (Vanacova, et al. 2003).
432 Though the position of the root of the eukaryotic tree of life is still debated, one
433 hypothesis is that it lies within Excavata, and specifically between Discoba (i.e. Euglenozoa,
434 Heterolobosea, Tsukubea, Jakobea) and the rest of eukaryotes (He, et al. 2014). If this
435 hypothesis were true, the smaller epigenetic toolkit in Excavata could be an indicator that the
436 epigenetic functions expanded in the remainder of the eukaryotes after the divergence of the
437 Excavata. However, the position of the root within Excavata may be the result of phylogenetic
438 artefacts such as long-branch attraction, and alternative roots such as between Unikonta and
439 Bikonta (Stechmann and Cavalier-Smith 2003; Derelle, et al. 2015) and between Opisthokonta
440 and the other eukaryotes (Stechmann and Cavalier-Smith 2002; Katz, et al. 2012) are still valid
441 hypotheses (reviewed in: Burki, et al. 2020).
442
443 *The epigenetic toolkit shows a pattern of punctate distribution across eukaryotes*
444 We observe a punctate distribution pattern of many epigenetic gene families (**Figure 4**).
445 Most strikingly, gene families that we conservatively define as being present in pre-LECA/LECA
446 (i.e. those in more than five of seven major eukaryotic clades) are not present in many of the

447 sampled lineages, which stands in stark contrast with the high conservation of housekeeping
448 genes in the same dataset (**Figure 4**). We see a similar pattern among the more 'recent' gene
449 families as some are present in two or more major clades but only in a few of the species
450 sampled (**Figure 4**). Similarly, we see fewer paralogs among epigenetic gene families as
451 compared to housekeeping genes (**Table S8**). Two possible explanations for this punctate
452 pattern include: 1) genes may have been lost in some lineages; and/or 2) epigenetic genes
453 evolve rapidly in some lineages and are no longer detected as homologs in our phylogenomic
454 approach.
455 Distinguishing between these two explanations is challenging due to both data
456 availability and the definitions used for gene family membership. Though assessing cases of
457 gene loss especially is hampered by the lack of whole genome data from many eukaryotic
458 lineages, our analyses of the limited set of whole genome data show the same punctate
459 distribution of genes (**Table S4**). Consistent with the hypothesis of rapid evolution of epigenetic
460 gene family members, we did observe longer branch lengths (i.e. from tips to first node) in
461 smaller (i.e. more punctate) gene families as opposed to larger gene families (**Figure 5**), but
462 phylogenetic artefacts and biases likely contribute to this pattern. More fundamentally, gene
463 'loss" can occur in a continuum, from the accumulation of numerous mutations that impact
464 homology assessment to the complete elimination of genes from within genomes. Hence, some
465 'lost' members of epigenetic gene families may have changed sufficiently to be exclude as
466 members of their ancestral gene families.
467
468 *Macroevolutionary phenomena may underlie the distribution of epigenetic gene families among*
469 *eukaryotes*
470 We hypothesize that the punctate distribution pattern of genes in the epigenetic toolkit is
471 the result of genome conflict, either as a defense against mobile genetic elements and/or as a
472 regulator of germline/soma differentiation. Some epigenetic processes are believed to have

originated as mechanisms for defense against viruses and other mobile genetic elements (Fedoroff 2012), and the relatively-rapid rates of some epigenetic genes (e.g. those involved in processing npc-RNAs) may be the result of an arms race between host and intruder genomes (e.g. Obbard, et al. 2009). Epigenetic genes also play a role in germline-soma distinctions. For example, ciliates rely on complex epigenetic processes to drive germline/soma distinction and DNA elimination throughout their lifecycle (e.g. Liu, et al. 2007; Maurer-Alcalá and Katz 2015; Pilling, et al. 2017).

Another macroevolutionary pattern that may explain the punctate distribution of genes in the epigenetic toolkit is their potential role in differential adaptation and reproductive isolation. A growing number of studies find differences in epigenetic marks (e.g. methylomes) of populations that are exposed to different environmental conditions (e.g. Marsh and Pasqualone 2014; Johnson and Kelly 2020; Wogan, et al. 2020) and in some cases these differences seem to be correlated with reproductive isolation (e.g. Smith, et al. 2016; Blevins, et al. 2017). Further, by regulating gene expression, epigenetic modifications can produce phenotypic plasticity, upon which selection may act (Rey, et al. 2016), which in turn can lead to reproductive isolation and ultimately to speciation. Further, the possibility of intergenerational or transgenerational inheritance of epigenetic marks or npc-RNAs (as reviewed in: Boskovic and Rando 2018; Perez and Lehner 2019) may enhance the possibility of adaptation. Epigenetics, therefore, may allow for adaptation of species to changing environmental conditions (Rey, et al. 2016).

4. Material and Methods

All approaches taken for data acquisition and data analysis are summarized here, and we refer the reader to the online supplementary text for details on methods.

4.1 Data acquisition

We identified genes involved in epigenetic processes by delving into the literature describing the molecular basis of epigenetics (Fuks 2005; Anantharaman, et al. 2007; Peters and Meister 2007; Hollick 2008; Shaver, et al. 2010; Maumus, et al. 2011; Fedoroff 2012; Bond and Baulcombe 2014; Rastogi, et al. 2015; Li and Patel 2016; Vogt 2017) and searching databases such as Pfam (https://pfam.xfam.org/) and KEGG (www.genome.jp/kegg/; **Table S1**). We used the resulting list of genes to identify the corresponding OG (orthologous groups) numbers in the OrthoMCL database (Li, et al. 2003), which correspond to the gene families in the phylogenomic pipeline PhyloToL (Ceron-Romero, et al. 2019). In total, we identified 179 genes that group into 118 distinct gene families (**Table 1**) and we ran PhyloToL to search for homologs of these epigenetic gene families in all major eukaryotic clades, plus a limited number of bacteria and archaea.

In addition to the sequence data included in PhyloToL (retrieved from either GenBank, RefSeq or OrthoMCL; **Table S2**) we added 69 transcriptomes from understudied clades within SAR (Stramenopila, Alveolata and Rhizaria) and Amoebozoa that we generated to increase taxonomic sampling. Since these microbial eukaryotes are not currently cultivable, we used a single-cell whole transcriptome amplification approach and assessed the quality of the resulting data based on the presence of at least 100 of 391 housekeeping gene families (**Table S1**). This approach resulted in the final number of 574 taxa, 296 of which are represented by whole genomes and 278 by transcriptomes (**Tables 2, S2**). We subsampled these data in three different ways to test the robustness of our analyses to taxon selection (**Table S2**). We then used PhyloToL to produce MSAs and gene trees for each of the epigenetic gene families for all four taxon selections. We also repeated this analysis for the 391 housekeeping genes.

4.2 Data analysis

As described in detail in the supplemental text, we used custom Python scripts (**Github**) to count the number of species per major clade that appeared in each gene family tree as well

as their number of paralogs (**Table S4**). We repeated this analysis for all four taxon sets and used the resulting data to estimate which gene families were present in the LECA or even before (**Tables S3, S4**). We assessed the evolutionary history of gene families in relationship with their grouping into certain functional categories (**Figure 3**). We also calculated the branch length of each gene tree (**Figure 5, Tables S1, S6**) and compared the number of paralogs in the epigenetic gene families versus the housekeeping gene families (**Table S8**), using methods described in the supplementary text.

Acknowledgements

This study was financially supported by grants from the National Institute of Health (grant number R15HG010409) and the National Science Foundation (grant numbers OCE-1924570, DEB-1651908) to LAK. We thank current and previous members of the Katzlab for help searching databases and for helpful comments on the manuscript. Further, we thank Jan Pawlowski, Roberto Sierra, Florian Mauffrey and Joana Cruz from the University of Geneva for contributing Foraminifera transcriptome data. Their work was supported by grant 31003A_179125 from the Swiss National Foundation. We also extend our thanks to the sequencing center at the Institute for Genome Sciences at the University of Maryland.

Data availability:

All sequenced transcriptomes are available on GenBank under the SRA BioProject PRJNA637648. The scripts used in the analyses of the data are available under github.com/Katzlab/Epigenetics.

References

Anantharaman V, Iyer LM, Aravind L. 2007. Comparative genomics of protists: New insights into the evolution of eukaryotic signal transduction and gene regulation. Annual Review of Microbiology 61:453-475.

Aramayo R, Selker EU. 2013. Neurospora crassa, a Model System for Epigenetics Research. Cold Spring Harbor Perspectives in Biology 5.

Aravind L, Burroughs AM, Zhang DP, Iyer LM. 2014. Protein and DNA Modifications: Evolutionary Imprints of Bacterial Biochemical Diversification and Geochemistry on the Provenance of Eukaryotic Epigenetics. Cold Spring Harbor Perspectives in Biology 6.

Belli SI. 2000. Chromatin remodelling during the life cycle of trypanosomatids. International Journal for Parasitology 30:679-687.

Bernstein E, Allis CD. 2005. RNA meets chromatin. Genes Dev 19:1635-1655.

Bhandari PN, Cui Y, Elzey BD, Goergen CJ, Long CM, Irudayaraj J. 2017. Oxygen nanobubbles revert hypoxia by methylation programming. Scientific Reports 7.

Bird A. 2007. Perceptions of epigenetics. Nature 447:396-398.

Blevins T, Wang J, Pflieger D, Pontvianne F, Pikaard CS. 2017. Hybrid incompatibility caused by an epiallele. Proceedings of the National Academy of Sciences of the United States of America 114:3702-3707.

Bond DM, Baulcombe DC. 2014. Small RNAs and heritable epigenetic variation in plants. Trends in Cell Biology 24:100-107.

Boskovic A, Rando OJ. 2018. Transgenerational Epigenetic Inheritance. Annual Review of Genetics, Vol 52 52:21-41.

Burki F, Roger AJ, Brown MW, Simpson AGB. 2020. The New Tree of Eukaryotes. Trends in Ecology & Evolution 35:43-55.

Camuzi D, de Amorim ISS, Pinto LFR, Trivilin LO, Mencalha AL, Lima SCS. 2019. Regulation Is in the Air: The Relationship between Hypoxia and Epigenetics in Cancer. Cells 8.

Carranza PG, Gargantini PR, Prucca CG, Torri A, Saura A, Svard S, Lujan HD. 2016. Specific histone modifications play critical roles in the control of encystation and antigenic variation in the early-branching eukaryote Giardia lamblia. International Journal of Biochemistry & Cell Biology 81:32-43.

Ceron-Romero M, Maurer-Alcala X, Grattepanche J, Yan Y, Fonseca M, Katz L. 2019. PhyloToL: A Taxon/Gene-Rich Phylogenomic Pipeline to Explore Genome Evolution of Diverse Eukaryotes. MBE 36:1831-1842.

Cerutti H, Casas-Mollano JA. 2006. On the origin and functions of RNA-mediated silencing: from protists to man. Current Genetics 50:81-99.

Chalker DL, Meyer E, Mochizuki K. 2013. Epigenetics of Ciliates. Cold Spring Harbor Perspectives in Biology 5.

Christensen DG, Baumgartner JT, Xie X, Jew KM, Basisty N, Schilling B, Kuhn ML, Wolfe AJ. 2019. Mechanisms, Detection, and Relevance of Protein Acetylation in Prokaryotes. mBio 10.

Clayton C. 2019. Regulation of gene expression in trypanosomatids: living with polycistronic transcription. Open Biology 9.

Cortes A, Crowley VM, Vaquero A, Voss TS. 2012. A View on the Role of Epigenetics in the Biology of Malaria Parasites. Plos Pathogens 8.

Croken MM, Nardelli SC, Kim K. 2012. Chromatin modifications, epigenetics, and how protozoan parasites regulate their lives. Trends in Parasitology 28:202-213.

Derelle R, Torruella G, Klimes V, Brinkmann H, Kim E, Vlcek C, Lang BF, Elias M. 2015. Bacterial proteins pinpoint a single eukaryotic root. Proceedings of the National Academy of Sciences of the United States of America 112:E693-E699.

Di Croce L, Helin K. 2013. Transcriptional regulation by Polycomb group proteins. Nature Structural & Molecular Biology 20:1147-1155.

El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, et al. 2005. Comparative genomics of trypanosomatid parasitic protozoa. Science 309:404-409.

Elias MC, Faria M. 2009. Are There Epigenetic Controls in Trypanosoma cruzi? Natural Genetic Engineering and Natural Genome Editing 1178:285-290.

Fazzari MJ, Greally JM. 2004. Epigenomics: beyond CpG islands. Nat Rev Genet 5:446-455.

Fedoroff NV. 2012. Presidential address. Transposable elements, epigenetics, and genome evolution. Science 338:758-767.

Finnegan EJ, Genger RK, Peacock WJ, Dennis ES. 1998. DNA Methylation in Plants. Annu Rev Plant Physiol Plant Mol Biol 49:223-247.

Fuks F. 2005. DNA methylation and histone modifications: teaming up to silence genes. Current Opinion in Genetics & Development 15:490-495.

Glastad KM, Hunt BG, Yi SV, Goodisman MA. 2011. DNA methylation in insects: on the brink of the epigenomic era. Insect Mol Biol 20:553-565.

Gluckman PD, Hanson MA, Buklijas T, Low FM, Beedle AS. 2009. Epigenetic mechanisms that underpin metabolic and cardiovascular diseases. Nature Reviews Endocrinology 5:401-408.

Goldberg AD, Allis CD, Bernstein E. 2007. Epigenetics: A landscape takes shape. Cell 128:635-638.

Grattepanche JD, Walker LM, Ott BM, Pinto DLP, Delwiche CF, Lane CE, Katz LA. 2018. Microbial Diversity in the Eukaryotic SAR Clade: Illuminating the Darkness Between Morphology and Molecular Data. Bioessays 40.

Grewal SIS. 2000. Transcriptional silencing in fission yeast. Journal of Cellular Physiology 184:311-318.

Haig D. 2004. The (dual) origin of epigenetics. Cold Spring Harbor Symposia on Quantitative Biology 69:67-70.

Handel AE, Ebers GC, Ramagopalan SV. 2010. Epigenetics: molecular mechanisms and implications for disease. Trends in Molecular Medicine 16:7-16.

He D, Fiz-Palacios O, Fu CJ, Fehling J, Tsai CC, Baldauf SL. 2014. An Alternative Root for the Eukaryote Tree of Life. Current Biology 24:465-470.

Heard E, Martienssen RA. 2014. Transgenerational epigenetic inheritance: myths and mechanisms. Cell 157:95-109.

Hollick JB. 2008. Sensing the epigenome. Trends in Plant Science 13:398-404.

Iyer LM, Anantharaman V, Wolf MY, Aravind L. 2008. Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. Int J Parasitol 38:1-31.

Jiang YH, Bressler J, Beaudet AL. 2004. Epigenetics and human disease. Annual Review of Genomics and Human Genetics 5:479-510.

Johnson KM, Kelly MW. 2020. Population epigenetic divergence exceeds genetic divergence in the Eastern oyster Crassostrea virginica in the Northern Gulf of Mexico. Evolutionary Applications.

Katz LA. 2001. Evolution of nuclear dualism in ciliates: a reanalysis in light of recent molecular data. Int. J. Syst. Evol. Microbiol. 51:1587-1592.

Katz LA. 2006. Genomes: Epigenomics and the future of genome sciences. Current Biology 16:R996-R997.

Katz LA, Grant JR, Parfrey LW, Burleigh JG. 2012. Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. Syst. Biol.

Klobutcher LA, Herrick G. 1997. Developmental genome reorganization in ciliated protozoa: The transposon link. Progress in Nucleic Acid Research and Molecular Biology, Vol. 56 56:1-62.

Kohler C, Villar CBR. 2008. Programming of gene expression by Polycomb group proteins. Trends in Cell Biology 18:236-243.

Kolisko M, Boscaro V, Burki F, Lynn DH, Keeling PJ. 2014. Single-cell transcriptomics for microbial eukaryotes. Current Biology 24:R1081-R1082.

Lagunas-Rangel FA, Bermudez-Cruz RM. 2019. Epigenetics in the early divergent eukaryotic Giardia duodenalis: An update. Biochimie 156:123-128.

Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178-2189.

Li SS, Patel DJ. 2016. Drosha and Dicer: Slicers cut from the same cloth. Cell Research 26:511-512.

Lisch D. 2009. Epigenetic Regulation of Transposable Elements in Plants. Annual Review of Plant Biology 60:43-66.

Liu B, Wendel JF. 2003. Epigenetic phenomena and the evolution of plant allopolyploids. Mol. Phyl. Evol. 29:365-379.

Liu Y, Taverna SD, Muratore TL, Shabanowitz J, Hunt DF, Allis CD. 2007. RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in Tetrahymena. Genes & Development 21:1530-1545.

Marsh AG, Pasqualone AA. 2014. DNA methylation and temperature stress in an Antarctic polychaete, Spiophanes tcherniai. Frontiers in Physiology 5.

Massana R, Gobet A, Audic S, Bass D, Bittner L, Boutte C, Chambouvet A, Christen R, Claverie JM, Decelle J, et al. 2015. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. Environ Microbiol 17:4035-4049.

Maumus F, Rabinowicz P, Bowler C, Rivarola M. 2011. Stemming Epigenetics in Marine Stramenopiles. Current Genomics 12:357-370.

Maurer-Alcalá XX, Katz LA. 2015. An epigenetic toolkit allows for diverse genome architectures in eukaryotes. Current Opinion in Genetics & Development 35:93-99.

Maurer-Alcalá XX, Yan Y, Pilling OA, Knight R, Katz LA. 2018. Twisted Tales: Insights into Genome Diversity of Ciliates Using Single-Cell ‘Omics. Genome biology evolution 10:1927-1939.

What is Epigenetics [Internet]. © Epigenome NoE; 2006 [cited 2013. Available from: http://epigenome.eu/en/1,1,0

Narsai R, Secco D, Schultz MD, Ecker JR, Lister R, Whelan J. 2017. Dynamic and rapid changes in the transcriptome and epigenome during germination and in developing rice (Oryza sativa) coleoptiles under anoxia and re-oxygenation. Plant J 89:805-824.

Ng HH, Bird A. 1999. DNA methylation and chromatin modification. Current Opinion in Genetics & Development 9:158-163.

Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM. 2009. The evolution of RNAi as a defence against viruses and transposable elements. Philosophical Transactions of the Royal Society B-Biological Sciences 364:99-115.

Ortega-Pierres MG, Jex AR, Ansell BRE, Svard SG. 2018. Recent advances in the genomic and molecular biology of Giardia. Acta Tropica 184:67-72.

Parfrey LW, Lahr DJG, Katz LA. 2008. The dynamic nature of eukaryotic genomes. MBE 25:787-794.

Peng JC, Lin HF. 2013. Beyond transposons: the epigenetic and somatic functions of the Piwi-piRNA mechanism. Current Opinion in Cell Biology 25:190-194.

Perez MF, Lehner B. 2019. Intergenerational and transgenerational epigenetic inheritance in animals. Nature Cell Biology 21:143-151.

Peters L, Meister G. 2007. Argonaute proteins: Mediators of RNA silencing. Molecular Cell 26:611-623.

Pilling OA, Rogers AJ, Gulla-Devaney B, Katz LA. 2017. Insights into transgenerational epigenetics from studies of ciliates. Eur J Protistol 61:366-375.

Ponger L, Li WH. 2005. Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. Mol Biol Evol 22:1119-1128.

Prescott DM. (Prescott, D.M. co-authors). 1994. The DNA of ciliated protozoa. Microbiol. Rev. 58:233-267.

Prucca CG, Slavin I, Quiroga R, Elias EV, Rivero FD, Saura A, Carranza PG, Lujan HD. 2008. Antigenic variation in Giardia lamblia is regulated by RNA interference. Nature 456:750-754.

Rapp RA, Wendel JF. 2005. Epigenetics and plant evolution. New Phytol 168:81-91.

Rastogi A, Lin X, Lombard B, Loew D, Tirichine L. 2015. Probing the evolutionary history of epigenetic mechanisms: what can we learn from marine diatoms. Aims Genetics 2:173-191.

Razin A, Riggs AD. 1980. DNA Methylation and Gene-Function. Science 210:604-610.

Rey O, Danchin E, Mirouze M, Loot C, Blanchet S. 2016. Adaptation to Global Change: A Transposable Element-Epigenetics Perspective. Trends in Ecology & Evolution 31:514-526.

Saliba AE, Westermann AJ, Gorski SA, Vogel J. 2014. Single-cell RNA-seq: advances and future challenges. NAR 42:8845-8860.

Shabalina SA, Koonin EV. 2008. Origins and evolution of eukaryotic RNA interference. Trends in Ecology & Evolution 23:578-587.

Sharp PA. 2001. RNA interference - 2001. Genes & Development 15:485-490.

Shaver S, Casas-Mollano JA, Cerny RL, Cerutti H. 2010. Origin of the polycomb repressive complex 2 and gene silencing by an E(z) homolog in the unicellular alga Chlamydomonas. Epigenetics 5:301-312.

Smith G, Ritchie MG. 2013. How might epigenetics contribute to ecological speciation? Current Zoology 59:686-696.

Smith TA, Martin MD, Nguyen M, Mendelson TC. 2016. Epigenetic divergence as a potential first step in darter speciation. Molecular Ecology 25:1883-1894.

Stechmann A, Cavalier-Smith T. 2003. Phylogenetic analysis of eukaryotes using heat-shock protein Hsp90. J. Mol. Evol. 57:408-419.

Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science 297:89-91.

Stotz K, Griffiths P. 2016. Epigenetics: ambiguities and implications. History and Philosophy of the Life Sciences 38.

Sullivan WJ, Naguleswaran A, Angel SO. 2006. Histones and histone modifications in protozoan parasites. Cellular Microbiology 8:1850-1861.

Tollefsbol TO. 2017. An Overview of Epigenetics. Handbook of Epigenetics: The New Molecular and Medical Genetics, 2nd Edition:3-8.

Vanacova S, Liston DR, Tachezy J, Johnson PJ. 2003. Molecular biology of the amitochondriate parasites, Giardia intestinalis, Entamoeba histolytica and Trichomonas vaginalis. Int J Parasitol 33:235-255.

Vogt G. 2017. Evolution of Epigenetic Mechanisms in Animals and Their Role in Speciation. Handbook of Epigenetics: The New Molecular and Medical Genetics, 2nd Edition:409-426.

Waddington CH. 1942. The Epigenotype. Endeavour 1:18-20.

Wogan GOU, Yuan ML, Mahler DL, Wang IJ. 2020. Genome-wide epigenetic isolation by environment in a widespread Anolis lizard. Molecular Ecology 29:40-55.

Wolffe AP, Matzke MA. 1999. Epigenetics: Regulation through repression. Science 286:481-486.

Zemach A, Zilberman D. 2010. Evolution of eukaryotic DNA methylation and the pursuit of safer sex. Curr Biol 20:R780-785.

Table 1: Summary of epigenetic gene families and their functional categories
Shown are the two main categories of epigenetic processes, chromatin modification and non-protein-coding RNAs (npc-RNAs), which are split into subcategories and associated pathways/processes. The number of gene families for each representative pathway is indicated. In total we analyzed 118 gene families. Details on individual genes and their functions are shown in **Table S1**.

Category	Subcategory	Pathway/Process	# gene families
Chromatin modification	DNA methylation	DNA methyltransferases	12
		methyl-DNA binding	3
	Histone modification	Lysine Acetyltransferase	10
		Lysine Deacetylase	10
		Lysine Methyltransferase	20
		Lysine Demethylase	12
		Arginine Methyltransferase	9
		Polycomb-group proteins	12
		Others	6
npc-RNAs	NA	non-protein-coding RNAs	24

Table 2: Eukaryotic and prokaryotic lineages included in the analysis

The names and abbreviations used throughout the manuscript for the major eukaryotic clades, bacteria and archaea. Shown are exemplary nested clades for each major clade and the number of species included in the different taxon sub-selections. Numbers in parenthesis indicate genomes and transcriptomes, respectively. For details on chosen species, their taxonomy and accession numbers see **Table S2**.

Major clade	Nested clades	ALL	INFORMED	RANDOM
Stramenopila (St)	Diatoms, Bikosea, Blastocystida, Chrysophytes, Eustigmatophytes, Labyrinthulomycetes, Oomycetes, Brown Algae, Pinguiophyceae, Raphidophytes, Synchromophytes, Synurophytes	77 (13/64)	25 (6/19)	25 (4/21)
Alveolata (Al)	Apicomplexa, Chromerida, Ciliates, Dinoflagellates, Perkinsozoa	87 (28/59)	25 (11/14)	25 (13/12)
Rhizaria (Rh)	Cercozoa, Foraminifera, Sticholonchida	31 (2/29)	25 (1/24)	25 (2/23)
Archaeplastida (Pl)	Green Algae and plants, Glaucophytes, Red Algae	59 (20/39)	25 (12/13)	25 (12/13)
Orphan lineages (EE)	Apusozoa, Breviatea, Centroheliozoa, Cryptomonads, Haptophytes, Katablepharids	42 (3/39)	25 (2/23)	25 (3/22)
Excavata (Ex)	Euglenozoa, Fornicata, Heterolobosea, Jakobida, Malawimonadidae, Oxymonadida, Parabasalia	31 (20/11)	25 (14/11)	25 (15/10)
Amoebozoa (Am)	Archamoeba, Discosea, Mycetozoa, Stereomyxa, Tubulinea	36 (8/28)	25 (5/20)	25 (5/20)
Opisthokonta (Op)	Choanoflagellates, Fungi, Ichthyosporea, Metazoa	97 (88/9)	25 (22/3)	25 (24/1)
Archaea (Za)	Archaeoglobi, Asgard group, Bathyarchaeota, Crenarchaeota, Halobacteria, Korarchaeota, Methanobacteria, Methanococci, Methanomicrobia, Methanopyri, Nanoarchaeota, Thaumarchaeota, Thermococci, Thermoplasmata	25 (25/0)	25 (25/0)	25 (25/0)
Bacteria (Ba)	Actinobacteria,Proteobacteria, Aquificae, Bacilli, Bacteroidia, Chlamydiales, Chlorobi, Chloroflexia, Clostridia, Cyanobacteria, Cytophagia, Deinococcus-Thermus, Dictyoglomi, Fusobacteriia, Nitrospira, Planctomycetes, Spirochaetia, Tenericutes, Thermotogae, Verrucomicrobia	89 (89/0)	25 (25/0)	25 (25/0)

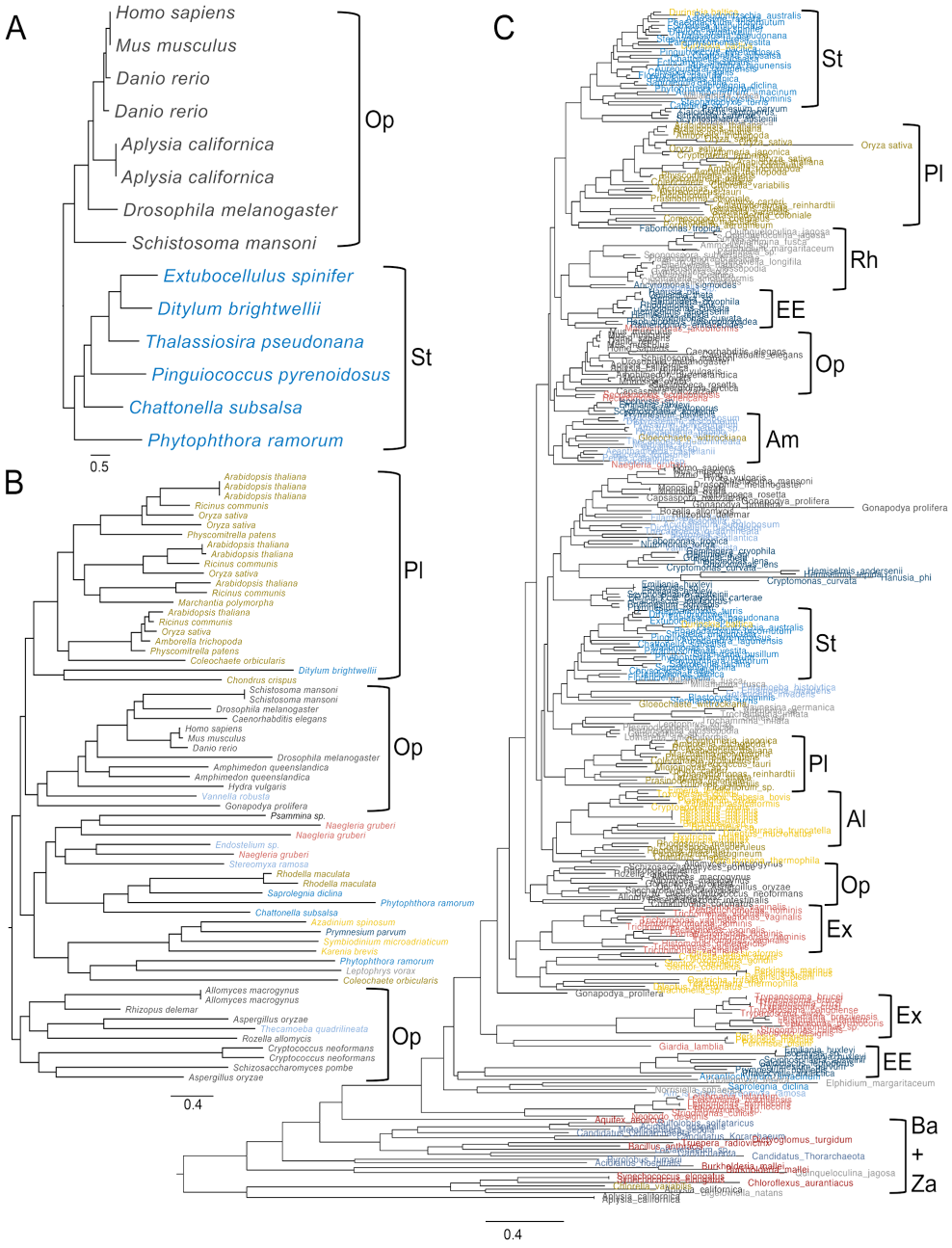

Figure 1: Exemplary epigenetic gene families showing variability in size
A) The phylogenetically-restricted Polycomb-related gene *SUZ12*, B) the punctate npc-RNA related gene family *DICER* and C) the complex Lysine deacetylase *HDAC1* gene family. Single gene trees do not generate well-resolved phylogenetic relationships across the ~1.8 billion years of eukaryotic evolution, and these trees are included to show the variability in conservation of the epigenetic genes across eukaryotes. Taxon selection is the INFORMED dataset and taxa are colored by major clades: Stramenopila (St) = blue, Alveolata (Al) = yellow, Rhizaria (Rh) = grey, Archaeplastida (Pl) = olive, orphans (EE) = dark blue, Excavata (Ex) = red, Amoebozoa (Am) = light blue, Opisthokonta (Op) = dark grey, Archaea (Za) = blue grey,

Bacteria (Ba) = dark red. The trees were manually rooted on bacteria, fungi or metazoa depending on which lineages were present.

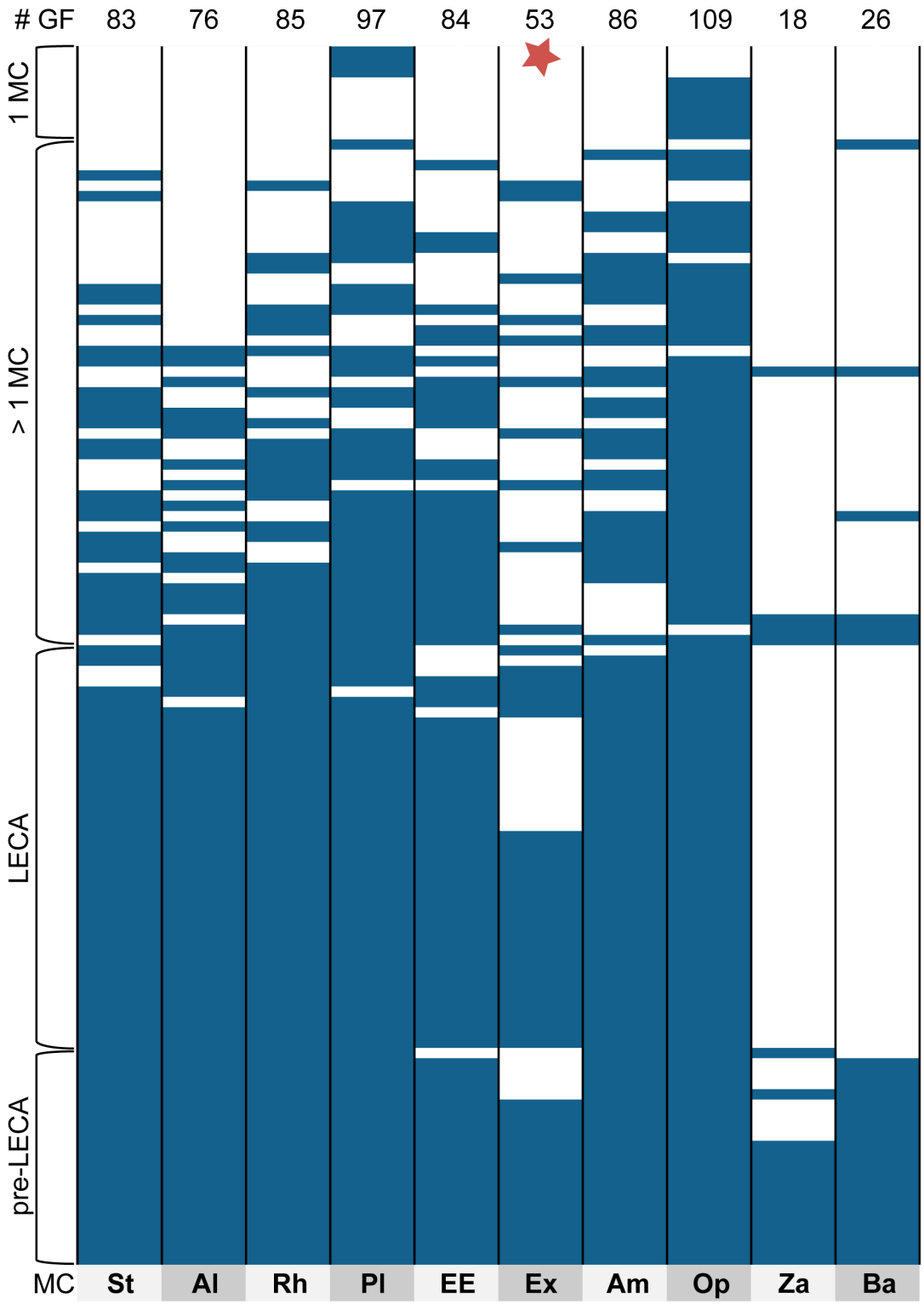

Figure 2: Differential conservation of the epigenetic toolkit across major eukaryotic clades with relative paucity of gene families in Excavata

Each column represents the presence/absence pattern per major clade (MC, abbreviations of the major clades as in **Table 2**), and the rows represent the 118 epigenetic gene families sorted

by degree of conservation across the tree of life (for the exact order of gene families see **Table S5)**. The numbers on top indicate the number of gene families (GF) present in each major clade. Shown are the results for the INFORMED taxon selection (250 species) and the presence (blue) and absence (white) of the epigenetic gene families in the major eukaryotic clades, bacteria and archaea. There is a striking difference in degree of conservation among epigenetic gene families: about half of them seem to have been present already before the LECA (pre-LECA) or in the LECA, whereas the other half are more restricted. Another strong signal is the absence of the majority of gene families in the Excavata (highlighted by a red star).

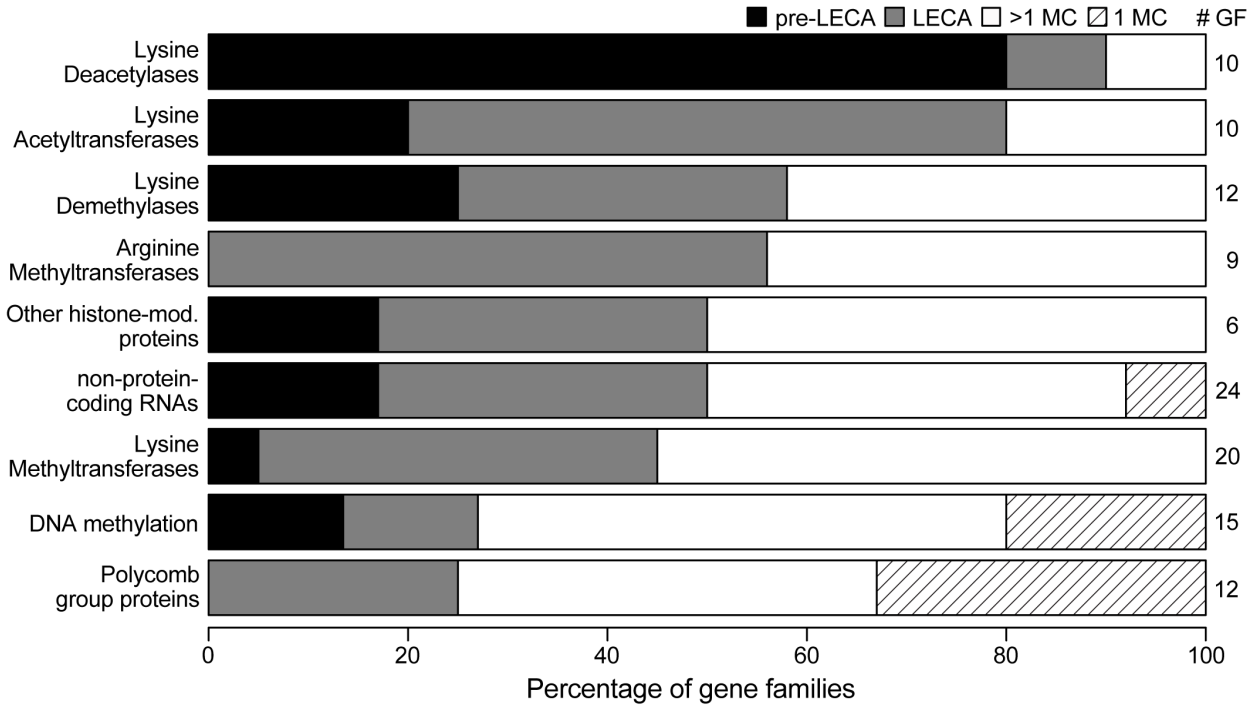

Figure 3: Conservation of epigenetic gene families shows differences across functional categories

Epigenetic gene families classified by functional category, as shown in **Table 1**, show variable numbers of conserved genes. “DNA methylation” and “non-protein-coding RNAs” represent higher level categories, comprising a variety of genes with different functions. To the right, the number of gene families (GF) is listed for each category. For each category, the percentage of pre-LECA (black) and LECA (grey) gene families is indicated, as well as the gene families that are present in fewer major clades (white) or are even restricted to one major clade (stippled). Data are based on the results of the INFORMED taxon selection.

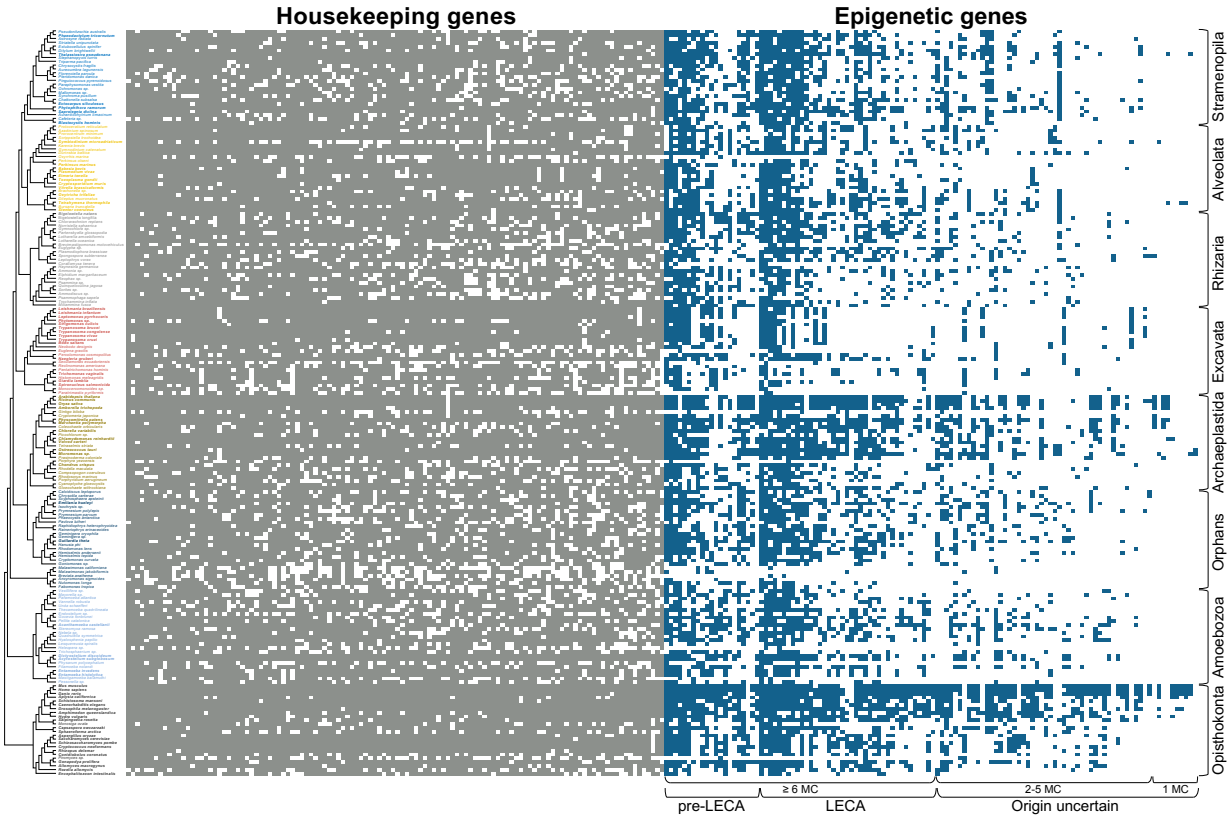

Figure 4: Comparison of the presence/absence between epigenetic and housekeeping gene families in eukaryotic species shows punctate retention of the epigenetic toolkit
The phylogenetic tree contains the eukaryotic species of the INFORMED dataset (196 species, four species were removed due to low data quality) and is a concatenated tree based on the 391 housekeeping gene families (see methods for details). The color coding of the major clades follows the colors in **Figure 1**, genome taxa are in bold. Two *Malawimonas* species were originally classified as Excavata, but fell among the orphan lineages in the tree. The panel on the left shows the presence (grey) or absence (white) for 118 of the 391 housekeeping gene families (columns) in each of the eukaryotic taxa (rows). The panel on the right shows the presence (blue) or absence (white) of the 118 epigenetic gene families. The orphan lineages are disregarded in counting the number of major clades.

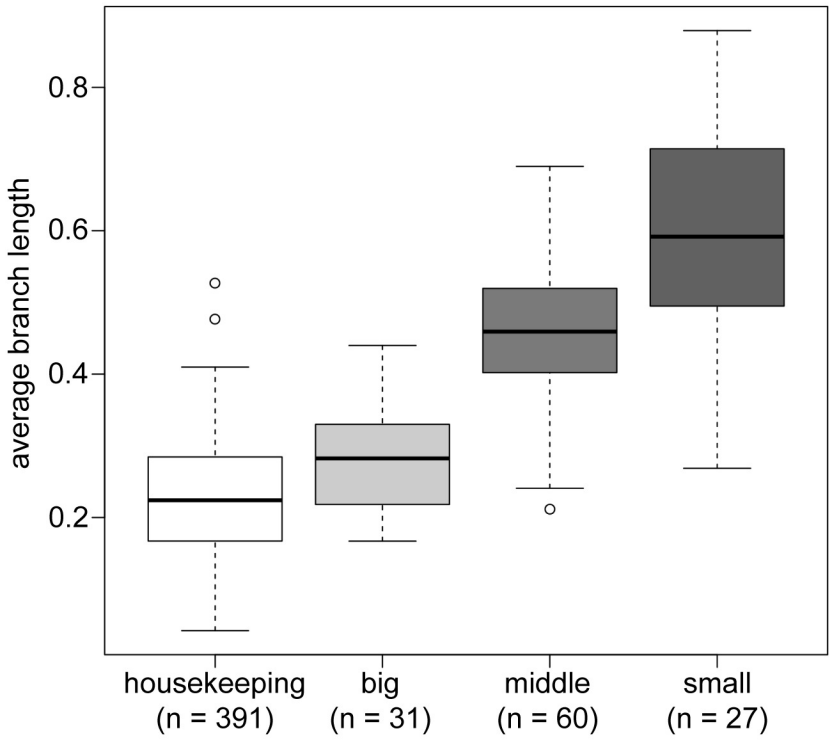

Figure 5: Differences between the average branch lengths of the housekeeping and epigenetic gene trees

We calculated average branch lengths for every tree of the housekeeping and epigenetic gene families based on the INFORMED taxon selection. The epigenetic trees are clustered into three groups (big, middle, small) based on the number of branches they contain. Statistical analysis (Shappiro Wilk, big: $p > 0.5102$, middle: $p > 0.8219$, small: $p > 0.496$, housekeeping: $p < 0.002036$) and analyses of QQ plots (**Figure S1**) of the four datasets suggest that the data likely are normally distributed. The means of all four datasets are significantly different from each other (Welch's t-test, **Table S6**), even though the difference between the housekeeping and big trees is smaller than between all other combinations. The box-whisker plots include medians for each dataset.

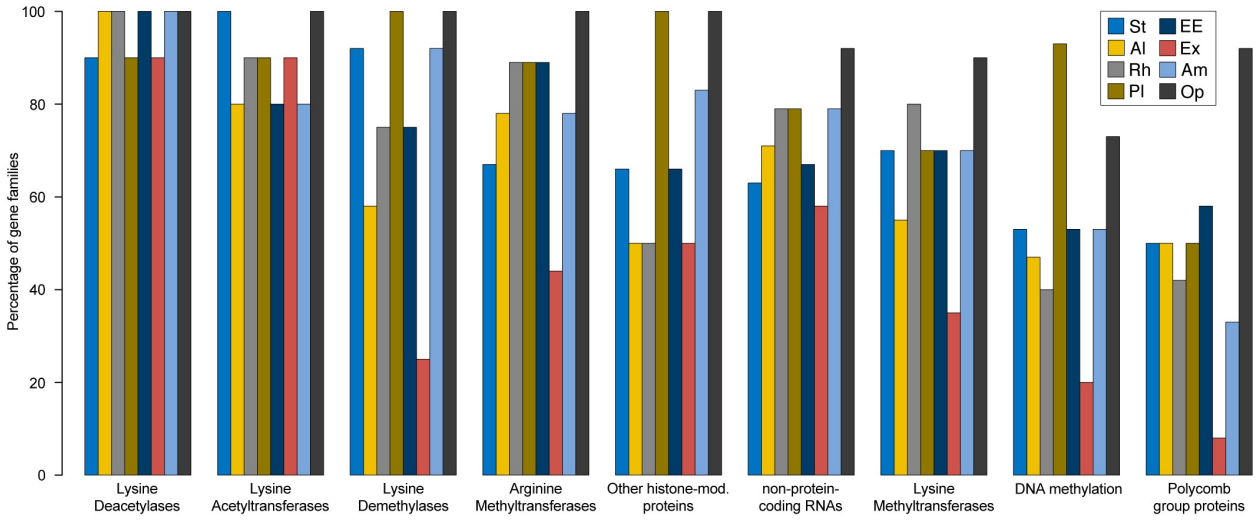

Figure 6: Distribution of the epigenetic gene families by functional categories in the major eukaryotic clades

Shown is the percentage of gene families per functional category that each major eukaryotic clade contains, based on the INFORMED taxon selection. Color coding of the major eukaryotic clades follows the colors in **Figure 1**. Noteworthy is the limited number of gene families related to methylation processes and Polycomb-group proteins in the Excavata.

Supplementary Material

Text S1: Detailed Material and Methods section

Table S1: Detailed list of the genes selected as part of the epigenetic toolkit and the housekeeping genes

Table S2: Detailed list of taxa of all major eukaryotic clades, bacteria and archaea

Table S3: Number of gene trees in which each major clade was present for each of the different taxon sub-selections

Table S4: Presence/absence of taxa in epigenetic gene trees and number of paralogs – INFORMED taxon selection, ALL taxon selection, RANDOM taxon selection, GENOME taxon selection, housekeeping gene trees INFORMED taxon selection

Table S5: Datafiles underlying the figures

Table S6: p-values of comparisons of means of branch lengths

Table S7: Results of guidance runs for potential homologs

Table S8: Comparison between number of paralogs of epigenetic and housekeeping gene families

Figure S1: QQplots for branch lengths analysis

Figure S2: Gene trees from the analysis of potential homologs