scientific reports

OPEN

Detection of horizontal gene transfer in the genome of the choanoflagellate *Salpingoeca rosetta*

Danielle M. Matriano[1], Rosanna A. Alegado[2] & Cecilia Conaco[1]✉

Horizontal gene transfer (HGT), the movement of heritable materials between distantly related organisms, is crucial in eukaryotic evolution. However, the scale of HGT in choanoflagellates, the closest unicellular relatives of metazoans, and its possible roles in the evolution of animal multicellularity remains unexplored. We identified at least 175 candidate HGTs in the genome of the colonial choanoflagellate *Salpingoeca rosetta* using sequence-based tests. The majority of these were orthologous to genes in bacterial and microalgal lineages, yet displayed genomic features consistent with the rest of the *S. rosetta* genome—evidence of ancient acquisition events. Putative functions include enzymes involved in amino acid and carbohydrate metabolism, cell signaling, and the synthesis of extracellular matrix components. Functions of candidate HGTs may have contributed to the ability of choanoflagellates to assimilate novel metabolites, thereby supporting adaptation, survival in diverse ecological niches, and response to external cues that are possibly critical in the evolution of multicellularity in choanoflagellates.

In Bacteria and Archaea, several mechanisms, such as transformation, conjugation, and transduction[1–8], facilitate the introduction of novel genes between neighboring strains and species, a process known as horizontal (or lateral) gene transfer (HGT)[9,10]. Horizontal transfers into eukaryote genomes, however, are thought to be low frequency events, as genetic material must enter the recipient cell's nucleus to be incorporated into the genome and vertically transmitted[11]. While HGT has been difficult to demonstrate in eukaryotic lineages[12], conflicting branching patterns between individual gene histories and species phylogenies[4] has garnered a number of explanations: a gene transfer ratchet to fix prey-derived genes ("you are what you eat")[13], movement of DNA from organelles to the nucleus through endosymbiosis (e.g. endosymbiotic gene transfer)[14,15], the presence of "weak-links" or unprotected windows during unicellular or early developmental stages that may enable integration of foreign DNA into eukaryotic genomes[11], and horizontal transposon transfers[16–18]. Indeed, HGT may have accelerated genome evolution and innovation in microbial eukaryotes by contributing to species divergence[19–21], metabolic diversity and versatility[22,23], and the establishment of interkingdom genetic exchange[10,19,24,25].

Recently, Choanoflagellatea, a clade of aquatic, free-living, heterotrophic nanoflagellates[26,27], has emerged as a model for understanding the unicellular origins of animals. Choanoflagellates are globally distributed in marine, brackish, and freshwater environments[28]. All choanoflagellates have a life history stage characterized by an ovoid unicell capped by a single posterior flagellum that functions to propel the cell in the water column and to generate currents that sweep food items toward an actin microvilli collar[26,27,29]. Captured food items, such as detritus and a diverse array of microorganisms, are phagocytosed by the organism. The choanoflagellate cellular architecture bears striking similarity to sponge choanocytes (or "collar cells"), leading to speculation on the evolutionary relationship between choanoflagellates and animals[27]. Abundant molecular phylogenetic evidence support choanoflagellates as the closest extant unicellular relatives of metazoans[30–35].

Choanoflagellates were historically grouped into two distinct orders based on taxonomy[28,32,33,36]: Acanthoecida (families Stephanoecidae and Acanthoecidae), which produce a siliceous cage-like basket exoskeleton or lorica; and non-loricate Craspedida (families Codonosigidae and Salpingoecidae), which produce an organic extracellular sheath or theca[32,33]. Previous studies suggest that silicate transport genes and transposable elements

[1]Marine Science Institute, University of the Philippines, Diliman, Quezon City, Philippines. [2]Department of Oceanography, Hawai'i Sea Grant, Daniel K. Inouye Center for Microbial Oceanography: Research and Education, University of Hawai'i at Manoa, Honolulu, USA. ✉email: cconaco@msi.upd.edu.ph

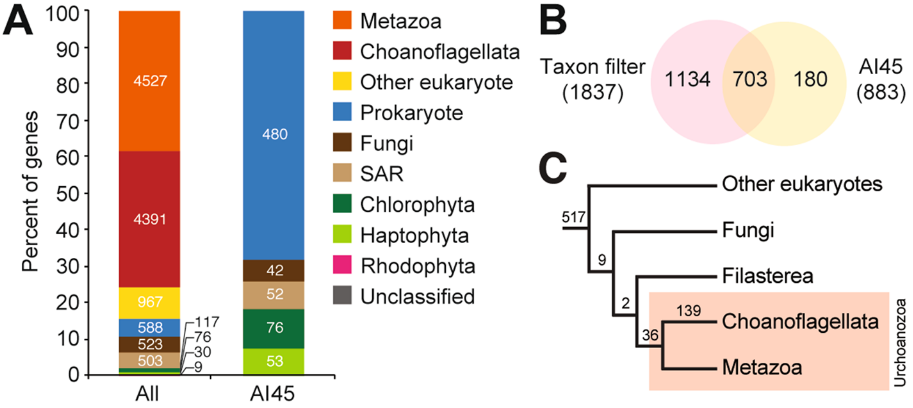

Figure 1. HGT candidates in the *S. rosetta* genome. (**A**) Taxonomic affiliation of all *S. rosetta* genes and genes that passed the Alien Index analysis (AI45). (**B**) Number of putative HGTs detected based on taxonomic affiliation (taxon filter) and Alien Index (AI45) analysis. (**C**) Number of HGT candidates with orthologs in other eukaryotic lineages.

in loricate choanoflagellates may have been acquired from microalgal prey[16,37]. At least 1000 genes, including entire enzymatic pathways in extant choanoflagellates, may have resulted from HGT events[38–46].

The colonial craspedid *Salpingoeca rosetta* has been established as a model organism for investigating the origins of animal multicellularity[27,47]. In this study, we used sequence-based methods previously employed to identify HGT events in other eukaryotes[48,49] to detect putative horizontally acquired genes in the *S. rosetta* genome. *S. rosetta* has a rich life history entwined with affiliated environmental bacteria[26,27,50–52]. Transient differentiation from the solitary morphotype into chain or rosette colonies, as well as sexual mating in choanoflagellates, are triggered by distinct prey bacteria[50]. We hypothesized that sources of novel horizontal transfers in *S. rosetta* are primarily from food sources, as has been demonstrated in other phagocytic and parasitic organisms[13,53]. We compared the unique gene signatures (i.e. GC content, codon usage bias, intron number) of potential HGTs against the rest of the *S. rosetta* genome to determine whether the genes were ancient or recent transfers. We also assessed the extent of taxonomic representation of the candidate HGTs in other choanoflagellates. Characterizing the magnitude and functions of HGTs in choanoflagellates may provide insight into their role in the evolution and adaptation of the lineage to new ecological niches and lifestyles. These HGT events may also shed light on the evolutionary history of genes involved in the origin of multicellularity in choanoflagellates and animals.

Results

Identification and genic architecture of candidate HGTs in *S. rosetta*.

Of the 11,731 reference *S. rosetta* protein coding genes, 4527 (38.59%) were related to metazoan sequences, while 4391 (37.43%) were specific to *S. rosetta* (Fig. 1A). Our taxon filter flagged 2804 (23.90%) genes with highest orthology to sequences from prokaryotes and unicellular non-metazoan taxa. Of these, 1837 (15.66%) genes had best hits to sequences from prokaryotes, fungi, and unicellular eukaryotic taxa with marine representatives, such as chlorophytes, rhodophytes, haptophytes, and the stramenopile, alveolate, and Rhizaria supergroup (SAR) (Fig. 1B). Best hits from prokaryotes (i.e. bacteria and archaea) and unicellular eukaryotic taxa (i.e. unicellular algae and fungi) were scored as potential HGT candidates and subjected to further analysis.

Identification of candidate HGTs by Alien Index analysis.

Using metrics described by Gladyshev et al.[54], Alien Index (AI) analysis of the 1,837 potential HGT candidate genes identified 703 candidate HGTs from potential bacteria, archaea, and common unicellular marine fungi (i.e. Basidiomycota and Ascomycota) and eukaryotic (i.e. Chlorophyta, Haptophyta, Pelagophyta, and Bacillariophyta) donors (Fig. 1A, Table S1), which included only 38% of the genes flagged by sequence alignment to the NCBI nr database (Fig. 1B). A total of 111 (16%) of the potential HGTs in *S. rosetta* have orthologs to HGT candidates that were previously identified in the choanoflagellate, *Monosiga brevicollis*[40].

Orthologs of candidate HGTs in other taxa.

To estimate when horizontally transferred genes in the *S. rosetta* genome were acquired, we assessed the number of candidate HGTs with orthologs in 20 other choanoflagellates and in representative eukaryotes, opisthokonts, filasterean, and metazoans, based on OrthoMCL groups or ortholog families identified by the study of Richter et al.[55] (Table S2). Majority of candidate HGTs flagged in the AI analysis (517 genes) had orthologs in other eukaryotes (Excavata, Diaphoretickes, and Amoebozoa), fungi (9 genes), and filasterean (2 genes). The detection of orthologs of candidate HGTs in multiple eukaryotic lineages likely reflect genes that are present in eukaryotic donor taxa, genes that were transferred into older lineages, or ancestral genes that were lost in multiple lineages but retained in choanoflagellates. Thirty-six genes had orthologs in animals while 139 were only found in choanoflagellates (Fig. 1C). These 175 genes were potentially acquired in the last common ancestor of choanoflagellates and animals (Urchoanozoan) or in the choanoflagellate lineage. Given the difficulty of ascertaining the evolutionary origins of genes with orthologs

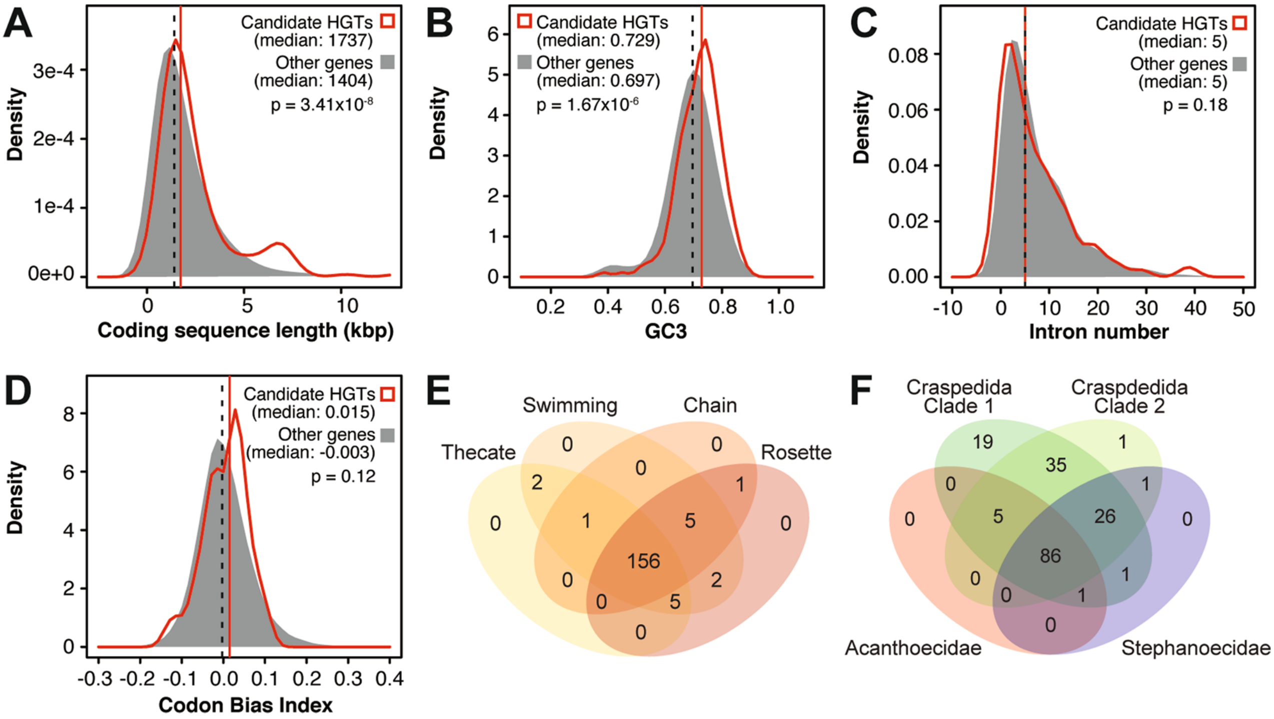

Figure 2. Gene architecture of candidate HGTs. Density plots showing the distribution of (**A**) coding sequence length, (**B**) GC content at the third codon position (GC3), (**C**) intron number, and (**D**) codon bias index in candidate horizontally transferred genes (red) in comparison to the bulk of *S. rosetta* genes (gray). (**E**) Number of HGT candidates that are expressed in the indicated life stages of *S. rosetta*. (**F**) Number of HGT candidates conserved in other choanoflagellate groups.

spanning multiple eukaryotic lineages, succeeding analyses focused only on genes gained in the Urchoanozoan and in the choanoflagellate lineage.

Genic architecture and expression of candidate HGTs. In *S. rosetta*, the difference in median protein coding sequence length of candidate HGTs (1737 bp) from the overall median CDS length (1404 bp) was small but significant (Kruskal–Wallis test, $p < 0.001$; Fig. 2A). Candidate HGTs also had significantly higher GC content at the third codon position (GC3) relative to other *S. rosetta* genes (Fig. 2B; Kruskal–Wallis test, $p < 0.001$). However, candidate HGTs were observed to have a similar median number of introns, specifically five, as compared to other *S. rosetta* genes (Fig. 2C). Only 1103 (8.79%) of *S. rosetta* genes lacked introns, of which only 23 passed both the AI and OrthoMCL filters. Codon bias index (CBI), which is a measure of codon usage frequency, was also similar between the putative HGTs and other genes (Fig. 2D). In addition, 156 of the candidate HGTs were expressed in the four life stages of *S. rosetta* (i.e. thecate cells, swimming cells, chain colonies, and rosette colonies), 16 genes were expressed in at least one stage, and only 3 were not detected in any of the stages (Fig. 2E, Table S3), based on data from the study of Fairclough et al.[34]. Candidate HGTs were conserved in other choanoflagellates, with 86 genes (49%) found in all choanoflagellate families, 66 genes (38%) found in both Craspedida clades, and 19 genes (11%) found only in Craspedida (clade 1), which includes *S. rosetta* (Fig. 2F).

Potential sources of *S. rosetta* HGTs. Majority of the candidate HGTs exhibited highest similarity to bacteria (65%), unicellular algae (27%), and unicellular fungal (Ascomycota and Basidiomycota) (8%) sequences (Fig. 3A). Phylogenetic analysis of 130 of 175 putative HGTs revealed conflicting phylogenetic signals relative to the consensus eukaryote reference phylogeny of Richter et al.[55] (Fig. 3B–F, Fig. S1), confirming their possible origin from donor taxa. The most common potential bacterial donors belonged to the following phyla: Proteobacteria, Terrabacteria, Planctomycetes, Verrucomicrobia, and Chlamydiae (PVC) group, and Fibrobacteres, Chlorobi, and Bacteroidetes (FCB) group (Fig. 3A). At least 4 sequences may have been derived from marine microorganisms that *S. rosetta* potentially interacts with, including known food sources, such as *Vibrio* spp. (Proteobacteria) (3 genes) and *Cytophaga* spp. (Bacteroidetes) (1 gene)[50].

Functions of candidate HGTs in *S. rosetta*. Of the 175 candidate HGTs in *S. rosetta*, 125 (71%) had identifiable PFAM domains and 114 (65%) were assigned gene ontology (GO) annotations (Table S1). 41 (23%) candidate HGTs contained two or more annotated PFAM domains. The most notable protein domains represented in the set of putative HGTs included enzymes (i.e. sulfatase, dehydrogenase, ubiquitin-activating enzyme active site, pyrroloquinolone quinolone-like domain, hydrolases, oxidoreductases, transposases), ECM-associ-

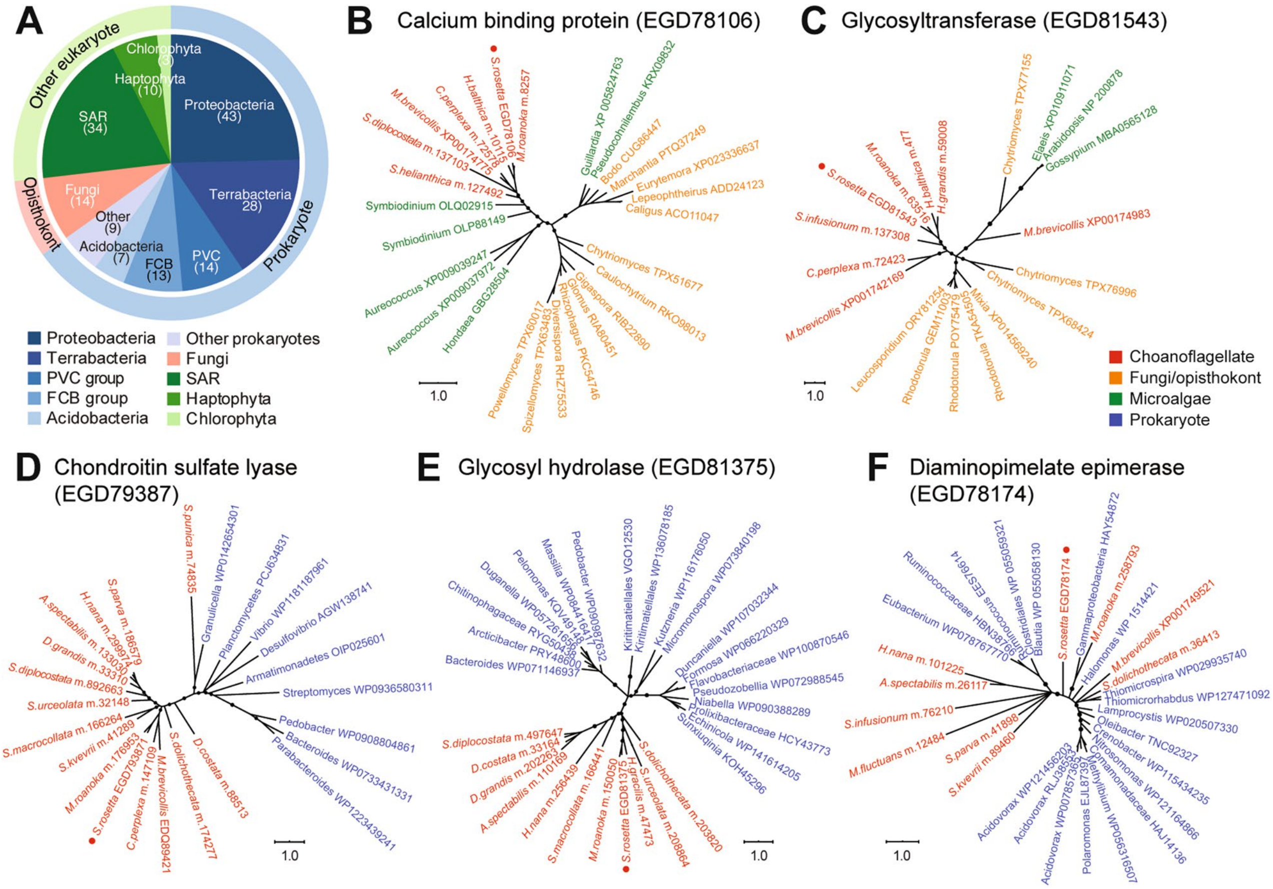

Figure 3. Potential donor phyla of candidate HGTs in *S. rosetta*. (**A**) Taxon affiliation of candidate HGTs based on the best sequence match for each gene. Phylogenetic analysis of selected candidate HGTs, including (**B**) calcium binding protein of microalgal origin, (**C**) glycosyltransferase of fungal origin, and (**D**) chondroitin sulfate lyase, (**E**) glycosyl hydrolase, and (**F**) diaminopimelate epimerase of prokaryotic origin. Genes from choanoflagellates are shown in red, fungi or opisthokonts in orange, microalgae in green, and prokaryotes in blue. *S. rosetta* genes are indicated by a red dot. Trees were generated using MrBayes 3.2.6. Circles at the branches indicate posterior probabilities of 0.70–1.00.

ated domains (i.e. dermatopontin, calcium-binding EGF domain, IPT/TIG, cadherin, galactose-binding lectin), and signaling domains (i.e. protein kinase) (Fig. 4A, Table S4).

Gene ontology analysis further supported the finding that majority of HGT candidates were enzymes with a variety of catalytic activities and were associated with cellular membranes where most biosynthetic and energy transduction processes of the cell occur[56] (Fig. 4B, Table S5). More specifically, potential HGTs in *S. rosetta* were associated with functions related to carbohydrate and protein metabolic processes (i.e. carbohydrate metabolism, peptidyl-aspartic acid hydroxylation, diaminopimelate epimerase activity), and adhesion and signaling (i.e. homophilic cell adhesion, receptor-mediated endocytosis, calcium ion binding, scavenger receptor activity).

Selected candidate HGTs. To further examine the potential functions of candidate HGTs, the genes were manually curated and assigned to a general cellular function based on GO affiliations and PFAM domains (Fig. 4C, Table S6). The most common functions within the set of putative HGTs were extracellular matrix (ECM) components. ECM-associated genes identified as potential HGTs included 10 dermatopontin/calcium-binding EGF (DPT/Ca^{2+} EGF) domain-containing genes. These genes had orthologs in craspedids but not in loricate choanoflagellates. Genes containing DPT/Ca^{2+} EGF domains had no orthologs in representative filastereans but were present in higher animals, including cnidarians and bilaterians. Two chondroitin sulfate lyase genes were also detected as potential HGTs in the AI analysis, although only one gene (EGD79387) passed the OrthoMCL filter and was unique to the choanoflagellate lineage.

Functions related to the metabolism of carbohydrates and proteins were also common within the set of candidate HGTs. Carbohydrate metabolism-related HGTs included 6 genes with glycosyl hydrolase domains and 2 with glycosyltransferase domains. Protein metabolism-related genes included various peptidases and ubiquitin-activating enzymes. Several genes involved in amino acid metabolism were also identified as HGTs, including diaminopimelate epimerase (*dapF*), a gene involved in lysine biosynthesis through the diaminopimelic acid (DAP) pathway. It should be noted that other DAP-associated genes, specifically, aspartate-semialdehyde dehydrogenase (*asd*) and diaminopimelate decarboxylase and aspartate kinase (*lysAC*), as well as other enzymes

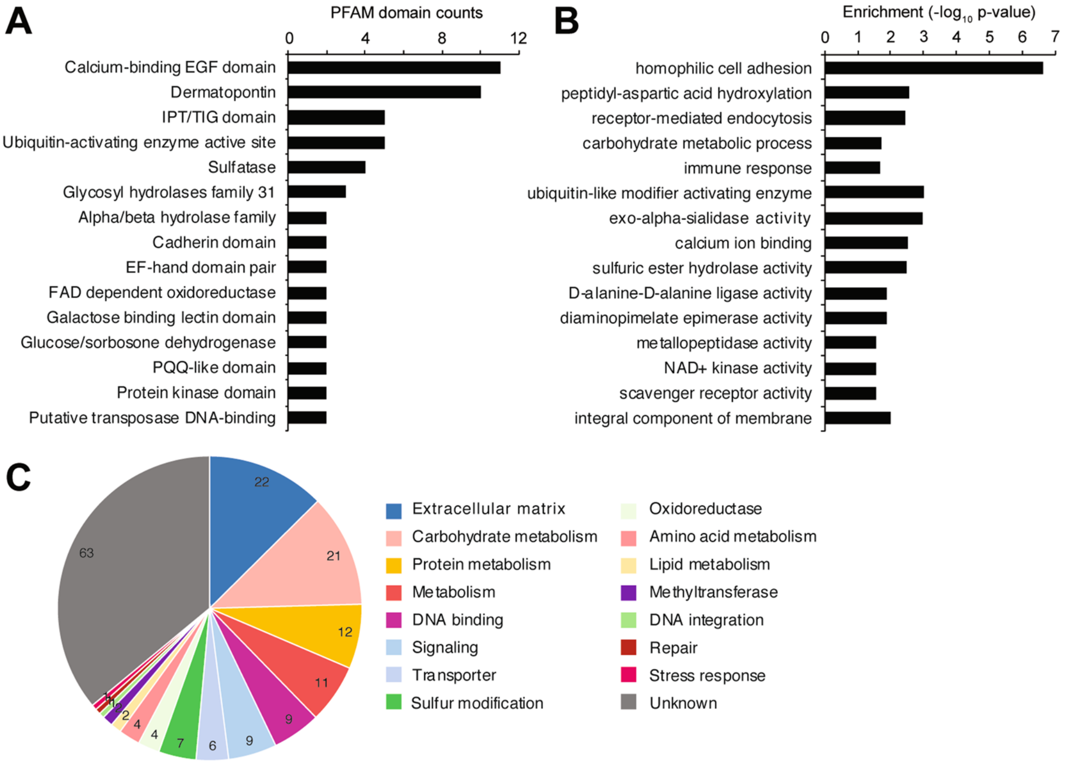

Figure 4. Functional analysis of candidate HGTs. (**A**) Most common PFAM protein domains in the set of candidate HGTs. (**B**) Gene ontology functions enriched in the set of putative HGTs. Enrichment p-values ($p \leq 0.05$) for selected functions are shown. (**C**) Number of candidate HGTs with associated functions based on manual curation.

involved in the biosynthesis of histidine, threonine, methionine, cysteine, tryptophan, and arginine[57], were also flagged as candidate HGTs by AI analysis but did not pass the OrthoMCL filter (Fig. S2). Candidate HGTs related to the metabolism of amino acids have orthologs in other choanoflagellates, sponges, and cnidarians.

Discussion

Our analysis provides evidence of a rich repertoire of *S. rosetta* genes that may have been acquired through horizontal gene transfer. Recent gene acquisitions are usually distinguished by divergent genetic characteristics, such as GC content, codon usage bias, and genetic architecture (e.g. intron content, coding sequence length)[58,59]. Over time, transferred genes undergo sequence changes to adapt to host genome characteristics, enabling improved transcription and translation[60]. A majority of *S. rosetta* HGT candidates had gene features similar to the rest of the genome and were expressed, based on the study by Fairclough et al.[34], indicating that these acquisitions were not recent[34]. We noted a higher GC content at the third codon position of the HGT candidate genes, which is a typical marker for genes derived from bacterial sources[61,62], yet the genes did not exhibit a divergent intron count. It is possible that foreign genes with translationally optimal codons and high GC3 content that resembles the GC-rich genome of *S. rosetta* are more likely to be positively selected, as this allows for more efficient translation and greater expression. Alternatively, selection for genes with metabolic functions, which tend to have a higher GC bias[63], may also explain the high GC3 content of the horizontally acquired genes. The prevalence of introns in *S. rosetta* candidate HGTs further indicate adaptation of acquired genes to the intron-rich genome of *S. rosetta*. Moreover, most HGT candidates had orthologs in other choanoflagellate taxa, suggesting that the gene transfer events occurred before divergence of the various choanoflagellate groups. Nevertheless, phylogenetic analysis of the candidate HGTs revealed incongruent gene trees, with candidate HGTs clustering with genes from potential donors, including bacteria, unicellular algae, and fungi. This suggests that candidate horizontally transferred genes in *S. rosetta* were acquired from multiple prokaryotic and unicellular eukaryotic donors.

Most of the potential donors of HGTs in *S. rosetta* clustered with possible bacterial donors, in contrast to *M. brevicollis* where HGTs were identified as coming mostly from algal donors[40]. *S. rosetta* is an active phagotroph of bacteria, such as *A. machipongonensis* and *Vibrio* spp., which have also been shown to influence its development and metabolic processes[26,27,50,52]. Thus, the major mechanism of HGT in *S. rosetta* may be through the engulfing of food or associated microbes. HGT events may also be facilitated by the activity of TEs, which are abundant in the genome of *S. rosetta*[16]. However, it should be noted that because some taxonomic lineages are underrepresented in sequence databases and since some HGTs may have been integrated into the host genome for a long time, the identification of the donor species for HGTs is not straightforward.

Most candidate HGTs with known functions were orthologous to operational genes, such as enzymes that function in amino acid and carbohydrate biosynthesis, as well as genes that function in intercellular signaling and in the establishment and modification of ECM components. Based on the complexity hypothesis proposed by Jain et al.[22], gene transferability is dependent on two factors: gene function and protein–protein interaction/network interaction[22,23,64–70]. Operational genes are more likely to be passed horizontally because they can function independently of other genes[40,48,54,71–74]. On the other hand, informational genes or genes involved in transcription and translational processes, physically interact with more gene products, limiting their functionality when transferred individually and reducing the possibility that they will be successfully retained as HGTs[65].

The acquisition of genes through ingestion of prey corroborates the "you are what you eat" gene transfer ratchet theory, which suggests that the evolution of the nuclear genome of most protists was driven by acquisition of exogenous genes by phagotrophy or engulfment and acquisition of gene fragments from their food sources[13]. This mode of acquisition allows for the transfer of more diverse genetic functions, as opposed to a more selective one via endosymbiosis. Interestingly, horizontal gene transfer in diverse eukaryotic lineages often involves enzymes that function in common metabolic pathways[40,41,46,48,53,54,72,75]. As mentioned above, the selective retention of these types of genes may be due to their ability to function independently and to be incorporated into pre-existing metabolic processes[22,76]. The acquisition of novel functions through HGT can extend the metabolic capability of the host, allowing it to explore and establish new niches or adapt to various environmental conditions. Horizontally acquired genes may also mediate interactions with other organisms in the environment and facilitate life stage transitions. The contribution of novel functions acquired via HGT may be greater in choanoflagellates like *M. brevicollis* and *S. rosetta* that have retained fewer ancestral gene families compared to other choanoflagellates[55].

Novel combinations of protein domains could potentially enhance catalytic efficiency and functional novelty of enzymes. The genome of *S. rosetta* contains a rich complement of multidomain genes that may have contributed to its unique biology and morphology[31,34]. It is possible that some of these molecular innovations emerged through gene fusion or domain shuffling events that incorporated pre-existing domains with domains acquired from horizontally transferred genes, thereby expanding adaptive functional novelty of genes in choanoflagellates and other eukaryotes[77].

Many of the putative HGTs in *S. rosetta* contribute to nutrient acquisition and metabolic processes by enabling efficient use of available organic substrates. Conservation of horizontally acquired proteases and glycosyl hydrolases in phagotrophic choanoflagellates suggests their importance in digesting various food sources and adapting to an environment high in plant biomass, such as the mud core samples from where *S. rosetta* was isolated[27,40,78]. Proteases and glycosyl hydrolases have also been flagged as candidate HGTs in *M. brevicollis,* bdelloid rotifers, sponge, rumen ciliates, and fungi[40,54,72,78,79].

HGTs with functions related to amino acid biosynthesis contribute to the metabolic flexibility of *S. rosetta*. Some of these enzymes, particularly those involved in the DAP pathway of lysine biosynthesis, as well as those involved in the biosynthesis of arginine, threonine and methionine, have previously been identified as potential HGTs in *M. brevicollis*[40,41]. Conservation of these HGTs in the choanoflagellate lineage and their absence in some metazoans, filastereans, and fungi, suggest that they were likely transferred into older lineages and retained in choanoflagellates. Retention of genes involved in amino acid metabolism in choanoflagellates may contribute to unique metabolic competencies. On the other hand, these genes have been lost in the animal stem lineage[80,81] as multicellular animals rely on direct acquisition of essential amino acids from their diet[82]. Cnidarians, however, regained the ability to synthesize aromatic amino acids tryptophan, phenylalanine, and other aromatic compounds through HGT events[80].

Salpingoeca rosetta has several life stages regulated by extrinsic factors. Its genome harbors multiple adhesion receptors and cell membrane enzymes for substrate attachment, cell–cell communication, and colony formation that may regulate these life stage transitions[26,34,50–52,55]. Among these are candidate HGTs, including several genes with dermatopontin (DPT/Ca^{2+} EGF) domains, which may function similar to dermatopontin, an acidic multifunctional matrix protein that promotes cell adhesion, ECM collagen fibrillogenesis, and cell assembly mediated by cell surface integrin binding[83–87]. It is also a major component of the organic matrix of biomineralized tissues (e.g. mussels)[88]. Absence of orthologs of DPT/Ca^{2+} EGF domain-containing genes in some loricate choanoflagellates suggests the possible importance of these genes in the production of the organic theca in Craspedida.

Salpingoeca rosetta also has multiple glycosyl hydrolases and glycosyltransferases that were potentially acquired through horizontal transfer. Some glycosyl hydrolases and glycosyltransferases were found exclusively in the choanoflagellate lineage. These enzymes are known to sculpt the ECM of animals by changing the structure of the extracellular matrix components and by modifying functional groups on molecules to regulate their interactions[89,90]. Glycosyl hydrolases may degrade proteoglycans to control the mating process of *S. rosetta*[52,91]. Glycosyltransferases, on the other hand, regulate key signaling and adhesion proteins like cadherins and integrins[92–94].

The *S. rosetta* genome contains two chondroitin sulfate lyase genes, both of which are candidate HGTs from potential bacterial donors, although only one gene is unique to the choanoflagellate lineage. It was suggested that these proteins may be involved in endogenous processes for regulating mating in *S. rosetta*[52]. It is also possible that chondroitin sulfate lyases contribute to the ability of choanoflagellates to modify cell walls and extracellular matrix proteins, which may facilitate transitions from one life stage to the next. Chondroitin sulfate lyase cleaves chondroitin sulfate glycosaminoglycans (GAGs) via an elimination mechanism resulting in disaccharides or oligosaccharides[95]. GAGs are typically found as side chains on proteoglycans on cell membranes and the ECM of animal tissues where they regulate processes such as adhesion, differentiation, migration, proliferation, and cell–cell communication[95]. The proteoglycan, chondroitin sulfate, was found to partially suppress the adhesive properties of dermatopontin[83–87], suggesting that breakdown of chondroitin sulfate through lyase activity may promote stronger cell adhesion. The bacterium, *Vibrio fischerii*, produces a chondroitin sulfate lyase called *EroS* (or "extracellular regulator of sex") that induces mating in the choanoflagellate, *S. rosetta*[52]. The conservation of chondroitin sulfate lyase genes in *S. rosetta* and most choanoflagellate representatives suggest that these genes may constitute unique adaptive mechanisms in choanoflagellates. Further studies on the expression of these genes in different environmental conditions are needed in order to better understand their potential functions in the host.

As with other studies on the detection of HGTs in eukaryotes, it is important to note that the current work is limited by the lack of broader taxon representation in publicly available databases. In addition, parametric tests and Alien Index analysis may not accurately estimate the number of HGT events, particularly for ancient gene transfers that have been retained in metazoans. Moreover, it can be difficult to distinguish HGT events from gene gain or gene loss events in the last universal common ancestor, as both result in patchy distribution of genes in the species tree. For many of the candidate HGTs in *S. rosetta*, presence of homologs in only a few lineages also constrained phylogenetic analyses to just a few representative taxa. These limitations may have resulted in the underestimation of the number of horizontally acquired genes in *S. rosetta*. Further development of methods to effectively filter out false positives and fine tune the results of HGT analysis is needed to reveal the true extent of horizontally acquired genes in the last common ancestor of choanoflagellates and animals.

In conclusion, we have shown that the genome of *S. rosetta* contains a rich repertoire of genes potentially acquired through horizontal transfer. Most genes were possibly ancient horizontal transfers gained prior to the divergence of choanoflagellates, as evidenced by similar genomic signatures and expression motifs to the host, as well as conservation in other choanoflagellate taxa. Horizontal acquisition of genes may have played a key role in the diversification of cellular metabolic processes and contributed novel functions that enhanced the catalytic ability of enzymes, thereby allowing *S. rosetta* to colonize diverse ecological niches. In addition, the acquisition of genes involved in the extra cellular senses and sensory responses to the environment, through the induction of reproduction and multicellular development, may have helped shape choanoflagellate evolution and multicellular life stages. We anticipate that future expression analysis of these promising candidate HGTs will provide a more in-depth understanding on their potential roles in choanoflagellates and animal multicellularity.

Materials and methods

Identification of HGTs by sequence alignment. The 11,731 predicted *S. rosetta* protein sequences downloaded from the Ensembl protist database[96] (http://www.ensembl.org/; last accessed October, 2018) were aligned to the NCBI non-redundant (nr) protein database (http://www.ncbi.nlm.nih.gov/; last accessed October, 2018), comprised of protein sequences from GenBank, EMBL, DDBJ, PDB, and RefSeq, using Diamond[97] local sequence alignment with a threshold E-value of 1×10^{-5}. The taxonomic affiliation of hits were retrieved from NCBI taxonomy (http://www.ncbi.nlm.nih.gov/taxonomy), and a Python[98] script was used to collect hits with specific taxon IDs. Only sequences with the lowest E-value corresponding to bacteria, archaea, unicellular algae (i.e. Chlorophyta, Stramenopiles, Bacillariophyta, Pelagophyta, and Haptophta), or unicellular fungi (i.e. Ascomycota and Basidiomycota) were considered as potential HGTs; peptides with no hits to the indicated taxa were disregarded from further analysis.

Determining the Alien Index scores of candidate HGTs. Alien Index (AI) analysis quantitatively measures how well the *S. rosetta* protein sequences align to non-metazoan versus metazoan protein sequences[54]. The E-value of the best sequence alignment match of *S. rosetta* peptides against all metazoan or non-metazoan gene sequences from the NCBI nr database were used to compute the AI score of each gene using the formula[54]:

$$AI = \log((best\ metazoan\ E\ value) + 1E - 200) - \log((best\ non{-}metazoan\ E\ value) + 1E - 200).$$

In cases where *S. rosetta* sequences had no hits to non-metazoans (excluding choanoflagellates) or metazoans, the E-value was set to 1. Genes that scored ≥ 45 were classified as foreign, $0 \leq AI \leq 45$ as indeterminate, and less than 0 as metazoan genes[54]. Orthologs of the foreign genes or candidate HGTs were determined using OrthoMCL group data from the study of Richter et al.[55]. Only genes gained in the Urchoanozoan stem were included in subsequent analyses.

Genomic feature analysis of candidate HGTs. Genomic features of putative HGTs, specifically GC content at the third codon position (GC3) and codon usage, as represented by the codon bias index (CBI), were determined using correspondence analysis on CodonW[99]. Coding sequence (CDS) lengths and intron numbers were determined from published information on the *S. rosetta* genome on Ensembl[96]. To determine if there were mean differences between the genomic features of candidate HGTs compared to all *S. rosetta* genes, we performed Kruskal–Wallis test in RStudio version 1.2.1335[100] with R version 3.6.1[100]. Density plots were generated using the sm package[101] in R and edited using Adobe Illustrator version 24.2.1.

Gene expression analysis. To examine expression of candidate HGTs, we obtained transcriptome data from the study of Fairclough et al.[34]. Sequence libraries representing various life stages of *S. rosetta* were downloaded from the NCBI Short Read Archive (rosette cells, SRX042054; chain cells, SRX042047; thecate cells, SRX042052; swimming cells, SRX042053). Reads were mapped against the predicted cDNA sequences of *S. rosetta* using kallisto[102] with default settings to obtain gene expression values in transcripts per million reads (TPM).

Protein domain and gene ontology analysis. Identification of protein domains and gene ontology associations was conducted on Blast2GO[103] to determine possible functions of candidate HGTs in *S. rosetta*.

Gene ontology terms associated with each predicted peptide were determined from its best Diamond hit to the UniProt database[104] at an E-value ≤ 1×10^{-5}. Enriched functions in the set of putative HGTs versus non-HGTs were identified using the Fisher's exact test in topGO[105] in R package version 3.6.1[100]. Only functions with an FDR-corrected p-value ≤ 0.05 were considered statistically significant.

Phylogenetic analysis of candidate HGTs. Peptide sequences of candidate HGTs were aligned with representative sequences from selected metazoa, eukaryotic, and prokaryotic taxa using MAFFT version 7[106]. Alignment confidence scores were calculated using GUIDANCE2[107] and alignments were trimmed using Gblocks[108] with default settings to remove ambiguous and divergent protein alignments. For each protein sequence, Akaike information criterion (AIC) and Bayesian information criterion (BIC) scores were calculated in MEGAX[109]. The best evolutionary model for phylogeny between different models tested for the protein sequences was determined by using the calculated AIC and BIC scores that had the smallest score difference, in this case, mtrev substitution model was consistently used. Markov Chain Monte Carlo (MCMC) parameters of each analysis were set to 100,000 generations sampled every 100 trees. By default, the first 25% of the trees were discarded as burn-in. Bayesian phylogenetic trees were constructed using MrBayes 3.2.6[110]. Trees were edited online using Interactive Tree Of Life (iTOL)[111] and Adobe Illustrator version 24.2.1. Genes with an insufficient number of homologs, an alignment confidence score ≤ 70%, or ambiguous alignment regions were not included in this analysis.

Data availability

The datasets generated during and/or analyzed in the current study are available in Figshare: https://figshare.com/projects/Detection_of_Horizontal_Gene_Transfer_in_the_Genome_of_the_Choanoflagellate_Salpingoeca_rosetta/79619.

Received: 3 July 2020; Accepted: 28 February 2021
Published online: 16 March 2021

References

1. Lorenz, M. G. & Wackernagel, W. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Rev.* **58**, 563–602 (1994).
2. Dubnau, D. DNA uptake in bacteria. *Annu. Rev. Microbiol.* **53**, 217–244 (1999).
3. Chen, I. & Dubnau, D. DNA uptake during bacterial transformation. *Nat. Rev. Microbiol.* **2**, 241–249 (2004).
4. Heinemann, J. A. & Sprague, G. F. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature* **340**, 205–209 (1989).
5. Llosa, M., Gomis-Rüth, F. X., Coll, M. & De la Cruz, F. Bacterial conjugation: A two-step mechanism for DNA transport. *Mol. Microbiol.* **45**, 1–8 (2002).
6. Haas, A. L., Baboshina, O., Williams, B. & Schwartz, L. M. Coordinated induction of the ubiquitin conjugation pathway accompanies the developmentally programmed death of insect skeletal muscle. *J. Biol. Chem.* **270**, 9407–9412 (1995).
7. Norman, A., Hansen, L. H. & Sørensen, S. J. Conjugative plasmids: Vessels of the communal gene pool. *Philos. Trans. R. Soc. B Biol. Sci.* **364**, 2275–2289 (2009).
8. Kyndt, T. *et al.* The genome of cultivated sweet potato contains Agrobacterium T-DNAs with expressed genes: An example of a naturally transgenic food crop. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 5844–5849 (2015).
9. Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* **9**, 605–618 (2008).
10. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: Building the web of life. *Nat. Rev. Genet.* **16**, 472–482 (2015).
11. Huang, J. Horizontal gene transfer in eukaryotes: The weak-link model. *BioEssays* **35**, 868–875 (2013).
12. Takeuchi, N., Kaneko, K. & Koonin, E. V. Horizontal gene transfer can rescue prokaryotes from Muller's ratchet: Benefit of DNA from dead cells and population subdivision. *G3 Genes Genomes Genet.* **4**, 325–339 (2014).
13. Doolittle, W. F. You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* **14**, 307–311 (1998).
14. O'Malley, M. A. Endosymbiosis and its implications for evolutionary theory. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 10270–10277 (2015).
15. Margulis, L., Chapman, M., Guerrero, R. & Hall, J. The last eukaryotic common ancestor (LECA): Acquisition of cytoskeletal motility from aerotolerant spirochetes in the Proterozoic Eon. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 13080–13085 (2006).
16. Southworth, J., Grace, C. A., Marron, A. O., Fatima, N. & Carr, M. A genomic survey of transposable elements in the choanoflagellate *Salpingoeca rosetta* reveals selection on codon usage. *Mobile DNA.* https://doi.org/10.1186/s13100-019-0189-9 (2019).
17. Pace, J. K., Gilbert, C., Clark, M. S. & Feschotte, C. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 17023–17028 (2008).
18. Mizrokhi, L. J. & Mazo, A. M. Evidence for horizontal transmission of the mobile element jockey between distant Drosophila species. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 9216–9220 (1990).
19. Williams, D. *et al.* A rooted net of life. *Biol. Direct* **6**, 45 (2011).
20. Andam, C. P. & Gogarten, J. P. Biased gene transfer and its implications for the concept of lineage. *Biol. Direct* **6**, 47 (2011).
21. Huang, J. & Gogarten, J. P. Ancient horizontal gene transfer can benefit phylogenetic reconstruction. *Trends Genet.* **22**, 361–366 (2006).
22. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3801–3806 (1999).
23. Jain, R., Rivera, M. C., Moore, J. E. & Lake, J. A. Horizontal gene transfer accelerates genome innovation and evolution. *Mol. Biol. Evol.* **20**, 1598–1602 (2003).
24. Swithers, K. S., Gogarten, J. P. & Fournier, G. P. Trees in the web of life. *J. Biol.* **8**, 54 (2009).
25. Redrejo-Rodriǵuez, M., Munõz-Espín, D., Holguera, I., Menciá, M. & Salas, M. Functional eukaryotic nuclear localization signals are widespread in terminal proteins of bacteriophages. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 18482–18487 (2012).
26. Dayel, M. J. & King, N. Prey capture and phagocytosis in the choanoflagellate *Salpingoeca rosetta*. *PLoS ONE* **9**, e95577 (2014).
27. Dayel, M. J. *et al.* Cell differentiation and morphogenesis in the colony-forming choanoflagellate *Salpingoeca rosetta*. *Dev. Biol.* **357**, 73–82 (2011).
28. Richter, D. J. & Nitsche, F. Choanoflagellatea. In *Handbook of the Protists* 2nd edn (eds Archibald, J. M. *et al.*) 1479–1496 (Springer, 2017).

29. Hoffmeyer, T. T. & Burkhardt, P. Choanoflagellate models—*Monosiga brevicollis* and *Salpingoeca rosetta*. *Curr. Opin. Genet. Dev.* **39**, 42–47 (2016).
30. Lang, B. F., O'Kelly, C., Nerad, T., Gray, M. W. & Burger, G. The closest unicellular relatives of animals. *Curr. Biol.* **12**, 1773–1778 (2002).
31. Ruiz-Trillo, I., Roger, A. J., Burger, G., Gray, M. W. & Lang, B. F. A phylogenomic investigation into the origin of Metazoa. *Mol. Biol. Evol.* **25**, 664–672 (2008).
32. Carr, M. *et al.* A six-gene phylogeny provides new insights into choanoflagellate evolution. *Mol. Phylogenet. Evol.* **107**, 166–178 (2017).
33. Nitsche, F., Carr, M., Arndt, H. & Leadbeater, B. S. C. Higher level taxonomy and molecular phylogenetics of the Choanoflagellatea. *J. Eukaryot. Microbiol.* **58**, 452–462 (2011).
34. Fairclough, S. R. *et al.* Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. *Genome Biol.* **14**, 1–15 (2013).
35. King, N. *et al.* The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* **451**, 783–788 (2008).
36. Jeuck, A., Arndt, H. & Nitsche, F. Extended phylogeny of the Craspedida (Choanomonada). *Eur. J. Protistol.* **50**, 430–443 (2014).
37. Marron, A. O. *et al.* A family of diatom-like silicon transporters in the siliceous loricate choanoflagellates. *Proc. R. Soc. B Biol. Sci.* **280**, 20122543 (2013).
38. Bapteste, E., Moreira, D. & Philippe, H. Rampant horizontal gene transfer and phospho-donor change in the evolution of the phosphofructokinase. *Gene* **318**, 185–191 (2003).
39. Malik, S. B., Ramesh, M. A., Hulstrand, A. M. & Logsdon, J. M. Protist homologs of the meiotic Spo11 gene and topoisomerase VI reveal an evolutionary history of gene duplication and lineage-specific loss. *Mol. Biol. Evol.* **24**, 2827–2841 (2007).
40. Yue, J., Sun, G., Hu, X. & Huang, J. The scale and evolutionary significance of horizontal gene transfer in the choanoflagellate *Monosiga brevicollis*. *BMC Genomics* **14**, 729 (2013).
41. Sun, G. & Huang, J. Horizontally acquired DAP pathway as a unit of self-regulation. *J. Evol. Biol.* **24**, 587–595 (2011).
42. Maruyama, S., Matsuzaki, M., Misawa, K. & Nozaki, H. Cyanobacterial contribution to the genomes of the plastid-lacking protists. *BMC Evol. Biol.* **9**, 197 (2009).
43. Nedelcu, A. M., Miles, I. H., Fagir, A. M. & Karol, K. Adaptive eukaryote-to-eukaryote lateral gene transfer: Stress-related genes of algal origin in the closest unicellular relatives of animals. *J. Evol. Biol.* **21**, 1852–1860 (2008).
44. Nedelcu, A. M., Blakney, A. J. C. & Logue, K. D. Functional replacement of a primary metabolic pathway via multiple independent eukaryote-to-eukaryote gene transfers and selective retention. *J. Evol. Biol.* **22**, 1882–1894 (2009).
45. Tucker, R. P., Beckmann, J., Leachman, N. T., Schöler, J. & Chiquet-Ehrismann, R. Phylogenetic analysis of the teneurins: Conserved features and premetazoan ancestry. *Mol. Biol. Evol.* **29**, 1019–1029 (2012).
46. Torruella, G., Suga, H., Riutort, M., Peretó, J. & Ruiz-Trillo, I. The evolutionary history of lysine biosynthesis pathways within eukaryotes. *J. Mol. Evol.* **69**, 240–248 (2009).
47. Fairclough, S. R., Dayel, M. J. & King, N. Multicellular development in a choanoflagellate. *Curr. Biol.* **20**, R875 (2010).
48. Eyres, I. *et al.* Horizontal gene transfer in bdelloid rotifers is ancient, ongoing and more frequent in species from desiccating habitats. *BMC Biol.* https://doi.org/10.1186/s12915-015-0202-9 (2015).
49. Fan, X. *et al.* Phytoplankton pangenome reveals extensive prokaryotic horizontal gene transfer of diverse functions. *Sci. Adv.* https://doi.org/10.1126/sciadv.aba0111 (2020).
50. Alegado, R. A. *et al.* A bacterial sulfonolipid triggers multicellular development in the closest living relatives of animals. *Elife* **1**, e00013 (2012).
51. Woznica, A. *et al.* Bacterial lipids activate, synergize, and inhibit a developmental switch in choanoflagellates. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7894–7899 (2016).
52. Woznica, A., Gerdt, J. P., Hulett, R. E., Clardy, J. & King, N. Mating in the closest living relatives of animals is induced by a bacterial chondroitinase. *Cell* **170**, 1175–1183 (2017).
53. Loftus, B. *et al.* The genome of the protist parasite *Entamoeba histolytica*. *Nature* **433**, 865–868 (2005).
54. Gladyshev, E. A., Meselson, M. & Arkhipova, I. R. Massive horizontal gene transfer in bdelloid rotifers. *Science* **320**, 1210–1213 (2008).
55. Richter, D. J., Fozouni, P., Eisen, M. B. & King, N. Gene family innovation, conservation and loss on the animal stem lineage. *Elife*. https://doi.org/10.7554/eLife.34226 (2018).
56. Ray, S., Kassan, A., Busija, A. R., Rangamani, P. & Patel, H. H. The plasma membrane as a capacitor for energy and metabolism. *Am. J. Physiol. Cell Physiol.* **310**, C181–C192 (2016).
57. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* https://doi.org/10.1093/nar/28.1.27 (2000).
58. Daubin, V., Lerat, E. & Perrière, G. The source of laterally transferred genes in bacterial genomes. *Genome Biol.* **4**, R57 (2003).
59. Lawrence, J. G. & Ochman, H. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9413–9417 (1998).
60. Langille, M. G. I., Hsiao, W. W. L. & Brinkman, F. S. L. Detecting genomic islands using bioinformatics approaches. *Nat. Rev. Microbiol.* **8**, 373–382 (2010).
61. Hildebrand, F., Meyer, A. & Eyre-Walker, A. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* **6**, e1001107 (2010).
62. Hershberg, R. & Petrov, D. A. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* **6**, e1001115 (2010).
63. Southworth, J. *et al.* Patterns of ancestral animal codon usage bias revealed through holozoan protists. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msy157 (2018).
64. Doolittle, W. F. Lateral genomics. *Trends Biochem. Sci.* **24**, M5 (1999).
65. Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6239–6244 (1998).
66. Sicheritz-Ponten, T. A phylogenomic approach to microbial evolution. *Nucleic Acids Res.* **29**, 545–552 (2001).
67. Gogarten, J. P., Senejani, A. G., Zhaxybayeva, O., Olendzenski, L. & Hilario, E. Inteins: Structure, function, and evolution. *Annu. Rev. Microbiol.* **56**, 263–287 (2002).
68. Brown, J. R. Ancient horizontal gene transfer. *Nat. Rev. Genet.* **4**, 121–132 (2003).
69. Wellner, A., Lurie, M. N. & Gophna, U. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. *Genome Biol.* **8**, R156 (2007).
70. Lercher, M. J. & Pál, C. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol. Biol. Evol.* **25**, 559–567 (2008).
71. Yue, J., Hu, X. & Huang, J. Origin of plant auxin biosynthesis. *Trends Plant Sci.* **19**, 764–770 (2014).
72. Conaco, C. *et al.* Detection of prokaryotic genes in the *Amphimedon queenslandica* genome. *PLoS ONE* **11**, e0151092 (2016).
73. Boschetti, C. *et al.* Biochemical diversification through foreign gene expression in bdelloid rotifers. *PLoS Genet.* **8**, e1003035 (2012).
74. Lal, D. & Lal, R. Evolution of mercuric reductase (merA) gene: A case of horizontal gene transfer. *Microbiology* **79**, 500–508 (2010).

75. Tucker, R. P. Horizontal gene transfer in Choanoflagellates. *J. Exp. Zool. B Mol. Dev. Evol.* **320**, 1–9 (2013).
76. Ochman, H., Lawrence, J. G. & Grolsman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature*. https://doi.org/10.1038/35012500 (2000).
77. Leadbeater, B. S. C. *The choanoflagellates: Evolution, Biology and Ecology* (Taylor and Francis Group, 2014).
78. Ricard, G. *et al.* Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics*. https://doi.org/10.1186/1471-2164-7-22 (2006).
79. Garcia-Vallvé, S., Romeu, A. & Palau, J. Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Mol. Biol. Evol.* **17**, 352–361 (2000).
80. Starcevic, A. *et al.* Enzymes of the shikimic acid pathway encoded in the genome of a basal metazoan, *Nematostella vectensis*, have microbial origins. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 2533–2537 (2008).
81. López-Escardó, D. *et al.* Reconstruction of protein domain evolution using single-cell amplified genomes of uncultured choanoflagellates sheds light on the origin of animals. *Philos. Trans. R. Soc. B Biol. Sci.* **374**, 20190088 (2019).
82. Payne, S. H. & Loomis, W. F. Retention and loss of amino acid biosynthetic pathways based on analysis of whole-genome sequences. *Eukaryot. Cell* **5**, 272–276 (2006).
83. Kuroda, K., Okamoto, O. & Shinkai, H. Dermatopontin expression is decreased in hypertrophic scar and systemic sclerosis skin fibroblasts and is regulated by transforming growth factor- β1, interleukin-4, and matrix collagen. *J. Investig. Dermatol.* **112**, 706–710 (1999).
84. Okamoto, O. & Fujiwara, S. Dermatopontin, a novel player in the biology of the extracellular matrix. *Connect. Tissue Res.* **47**, 177–189 (2006).
85. Lewandowska, K. *et al.* Extracellular matrix adhesion-promoting activities of a dermatan sulfate proteoglycan-associated protein (22K) from bovine fetal skin. *J. Cell Sci.* **99**, 657–668 (1991).
86. Lewandowska, K., Choi, H. U., Rosenberg, L. C., Zardi, L. & Culp, L. A. Fibronectin-mediated adhesion of fibroblasts: Inhibition by dermatan sulfate proteoglycan and evidence for a cryptic glycosaminoglycan-binding domain. *J. Cell Biol.* **105**, 1443–1454 (1987).
87. Winnemoller, M., Schon, P., Vischer, P. & Kresse, H. Interactions between thrombospondin and the small proteoglycan decorin: Interference with cell attachment. *Eur. J. Cell Biol.* **59**, 47–55 (1992).
88. Marxen, J. C., Nimtz, M., Becker, W. & Mann, K. The major soluble 196 kDa protein of the organic shell matrix of the freshwater snail *Biomphalaria glabrata* is an N-glycosylated dermatopontin. *Biochim. Biophys. Acta Proteins Proteomics* **1650**, 92–98 (2003).
89. Wetzel, L. A. *et al.* Predicted glycosyltransferases promote development and prevent spurious cell clumping in the choanoflagellate *S. rosetta*. *Elife*. https://doi.org/10.7554/eLife.41482 (2018).
90. Larson, B. T. *et al.* Biophysical principles of choanoflagellate self-organization. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 1303–1311 (2020).
91. Trincone, A., Tramice, A., Giordano, A. & Andreotti, G. Glycoside hydrolases in *Aplysia fasciata*: Analysis and applications. *Biotechnol. Genet. Eng. Rev.* **25**, 129–148 (2008).
92. Sawaguchi, S. *et al.* O-GlcNAc on NOTCH1 EGF repeats regulates ligand-induced Notch signaling and vascular development in mammals. *Elife*. https://doi.org/10.7554/eLife.24419 (2017).
93. Stratford, M. Yeast flocculation: Receptor definition by mnn mutants and concanavalin A. *Yeast* **8**, 635–645 (1992).
94. Naderer, T., Vince, J. & McConville, M. Surface determinants of leishmania parasites and their role in infectivity in the mammalian host. *Curr. Mol. Med.* **4**, 649–665 (2005).
95. Zhang, F., Zhang, Z. & Linhardt, R. J. Glycosaminoglycans. In *Handbook of Glycomics* (eds Zhang, F. *et al.*) 59–80 (Elsevier, 2020).
96. Kersey, P. J. *et al.* Ensembl Genomes 2018: An integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* **46**, D802–D808 (2018).
97. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2014).
98. Python Software Foundation. Python Language Reference, version 3.5. *Python Software Foundation* (2016).
99. Wang, M. *et al.* Analysis of codon usage in type 1 and the new genotypes of duck hepatitis virus. *BioSystems* **106**, 45–50 (2011).
100. R Development Core Team. R: A language and environment for statistical computing. *R Found. Stat. Comput.* Vol. 1, 409 (2011).
101. Bowman, A. W. & Azzalini, A. Computational aspects of nonparametric smoothing with illustrations from the sm library. *Comput. Stat. Data Anal.* https://doi.org/10.1016/S0167-9473(02)00118-4 (2003).
102. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
103. Conesa, A. *et al.* Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
104. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
105. Alexa, A. & Rahnenfuhrer, J. topGO: Enrichment analysis for gene on-tology. In *R package version 2.26.0.* R package version 2.22.0 (2009).
106. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
107. Sela, I., Ashkenazy, H., Katoh, K. & Pupko, T. GUIDANCE2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkv318 (2015).
108. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
109. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
110. Ronquist, F. *et al.* Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
111. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**, W256 (2019).

Acknowledgements

We thank Joshua Dizon and Francis Tablizo of the Philippine Genome Center Core Facility for Bioinformatics for assistance with scripts and database construction. We thank Becca Lensing (University of Hawai'i), Cheryl Andam (University of New Hampshire), Deo Onda and Ron Leonard Dy (University of the Philippines) for insightful comments and suggestions on the analysis and interpretation of the data. This work was supported by thesis grants from the Department of Science and Technology Accelerated Science and Technology Human Resource Development Program-National Science Consortium (DOST-ASTHRDP-NSC) and the University of the Philippines Marine Science Institute to DM.

Author contributions

D.M. and C.C. together conceived and designed the study. D.M. and C.C. conducted the analyses and drafted the manuscript. C.C. and R.A. advised on data analyses and helped to draft and edit the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-85259-6.

Correspondence and requests for materials should be addressed to C.C.