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

Danielle M. Matriano1, Rosanna A. Alegado2 & Cecilia Conaco1*

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 transduction1–8, 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 transmitted11. While HGT has been difficult to demonstrate in eukaryotic lineages12, conflicting branching patterns between individual gene histories and species phylogenies4 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 genomes11, and horizontal transposon transfers16–18. Indeed, HGT may have accelerated genome evolution and innovation in microbial eukaryotes by contributing to species divergence19–21, metabolic diversity and versatility22,23, and the establishment of interkingdom genetic exchange10,19,24,25.

Recently, Choanoflagellatea, a clade of aquatic, free-living, heterotrophic nanoflagellates26,27, has emerged as a model for understanding the unicellular origins of animals. Choanoflagellates are globally distributed in marine, brackish, and freshwater environments28. 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 collar26,27,29. 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 animals27. Abundant molecular phylogenetic evidence support choanoflagellates as the closest extant unicellular relatives of metazoans30–33.

Choanoflagellates were historically grouped into two distinct orders based on taxonomy28,32,33,36: Acanthoe-cida (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 theca32,33. Previous studies suggest that silicate transport genes and transposable elements...
in loricate choanoflagellates may have been acquired from microalgal prey. At least 1000 genes, including entire enzymatic pathways in extant choanoflagellates, may have resulted from HGT events.

The colonial craspedid Salpingoeca rosetta has been established as a model organism for investigating the origins of animal multicellularity. In this study, we used sequence-based methods previously employed to identify HGT events in other eukaryotes to detect putative horizontally acquired genes in the S. rosetta genome.

S. rosetta has a rich life history entwined with affiliated environmental bacteria. Transient differentiation from the solitary morphotype into chain or rosette colonies, as well as sexual mating in choanoflagellates, are triggered by distinct prey bacteria. 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. 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., 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.

**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. (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...
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.35 (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)36.

**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-
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 occur56 (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/Ca²⁺ EGF) domain-containing genes. These genes had orthologs in craspedids but not in loricate choanoflagellates. Genes containing DPT/Ca²⁺ 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...
involved in the biosynthesis of histidine, threonine, methionine, cysteine, tryptophan, and arginine57, 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 translation60. 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 recent34. 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 sources61,62, 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 bias63, 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 donors40. *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 processes26,27,50,52. 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.

**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 ≤ 0.05) for selected functions are shown. (C) Number of candidate HGTs with associated functions based on manual curation.
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 complex hypothesis proposed by Jain et al.22, gene transferability is dependent on two factors: gene function and protein–protein interaction/network interaction22,23,64–70. Operational genes are non-likely to be passed horizontally because they can function independently of other genes60,48,54,71–74. 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 HGTs49.

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 sources13. 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 pathways40,41,46,48,53,54,72,75. 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 processes22,76. 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 fewest ancestral gene families compared to other choanoflagellates65.

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 morphology11,14. 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 eukaryotes77.

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 isolated27,40,78. Proteases and glycosyl hydrolases have also been flagged as candidate HGTs in M. brevicollis, bdelloid rotifers, sponge, rumen ciliates, and fungi40,54,78,79.

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. brevicollis60,61. 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 lineage80,81 as multicellular animals rely on direct acquisition of essential amino acids from their diet82. Cnidarians, however, regained the ability to synthesize aromatic amino acids tryptophan, phenylalanine, and other aromatic compounds through HGT events83.

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 transitions26,34,39–52,55. Among these are candidate HGTs, including several genes with dermatoponint (DPT/Ca2+ EGF) domains, which may function similar to dermatoponint, an acidic multifunctional matrix protein that promotes cell adhesion, ECM collagen fibrillogenesis, and cell assembly mediated by cell surface integrin binding43–45. It is also a major component of the organic matrix of bimineralized tissues (e.g. mussels)84. Absence of orthologs of DPT/Ca2+ 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 interactions89,90. Glycosyl hydrolases may degrade proteoglycans to control the mating process of S. rosetta12,21. Glycosyltransferases, on the other hand, regulate key signaling and adhesion proteins like cadherins and integrins25,92.

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. rosetta22. 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 chondroitin sulfate glycosaminoglycans (GAGs) via an elimination mechanism resulting in disaccharides or
(or “extracellular regulator of sex”) that induces mating in the choanoflagellate, S. rosetta\textsuperscript{32}. 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 metazoa. 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\textsuperscript{40} (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\textsuperscript{41} local sequence alignment with a threshold E-value of $1 \times 10^{-5}$. The taxonomic affiliation of hits were retrieved from NCBI taxonomy (http://www.ncbi.nlm.nih.gov/taxonomy), and a Python\textsuperscript{42} 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 Haptophyta), 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\textsuperscript{48}. 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\textsuperscript{54}:

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

In cases where S. rosetta sequences had no hits to non-metazoa (excluding choanoflagellates) or metazoa, the E-value was set to 1. Genes that scored $\geq 45$ were classified as foreign, $0 \leq \text{AI} \leq 45$ as indeterminate, and less than 0 as metazoan genes\textsuperscript{54}. Orthologs of the foreign genes or candidate HGTs were determined using OrthoMCL group data from the study of Richter et al.\textsuperscript{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\textsuperscript{96}. Coding sequence (CDS) lengths and intron numbers were determined from published information on the S. rosetta genome on Ensembl\textsuperscript{10}. 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 1.2.1335100 with R version 3.6.1100. Density plots were generated through the use of the ggplot2\textsuperscript{101} R package.

Gene expression analysis. To examine expression of candidate HGTs, we obtained transcriptome data from the study of Fairclough et al.\textsuperscript{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\textsuperscript{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\textsuperscript{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 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 in R package version 3.6.1. 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 metazoans, eukaryotes, and prokaryotes taxa using Mafft version 7. Alignment confidence scores were calculated using GUIDANCE and alignments were trimmed using Gblocks 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. 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. Trees were edited online using Interactive Tree Of Life (iTOL) 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_rossetta/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 accompanied 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. Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. 9, 605–618 (2008).

10. Southworth, J., Grace, C. A., Marron, A. O., Fatima, N. & Carr, M. A genomic survey of transposable elements in the choanoflagellate Salpingoeca rosetta. Trends Genet. 4, 325–339 (2014).

11. 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).

12. O’Malley, M. A. Endosymbiosis and its implications for evolutionary theory. Proc. Natl. Acad. Sci. U.S.A. 112, 10270–10277 (2015).

13. Margulis, L., Chapman, M., Guererro, 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).

14. 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).

15. Pace, J. K., Gilbert, C., Clark, M. S. & Feschotte, C. Repeate d horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc. Natl. Acad. Sci. U.S.A. 105, 17023–17028 (2008).

16. Mizroki, 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).

17. Williams, D. A. A rooted net of life. Biol. Direct 6, 45 (2011).

18. Andam, C. P. & Gogarten, J. P. Biased gene transfer and its implications for the concept of lineage. Biol. Direct 6, 47 (2011).

19. Huang, J. & Gogarten, J. P. Ancient horizontal gene transfer can benefit phylogenetic reconstruction. Trends Genet. 22, 361–366 (2006).

20. 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).

21. 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).

22. Swithers, K. S., Gogarten, J. P. & Fournier, G. P. Trees in the web of life. J. Biol. 8, 54 (2009).

23. Redondo-Rodriguez, M., Munoz-Espin, D., Holguera, I., Mencía, 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).

24. Dayel, M. J. & King, N. Prey capture and phagocytosis in the choanoflagellate Salpingoeca rosetta. PLoS ONE 9, e95577 (2014).

25. Dayel, M. J. et al. Cell differentiation and morphogenesis in the colony-forming choanoflagellate Salpingoeca rosetta. Dev. Biol. 357, 73–82 (2011).

26. Richter, D. J. & Nitsche, F. Choanoflagellatea. In Mobile DNA. https://doi.org/10.1186/s13100-019-0189-9 (2019).

27. Redrejo-Rodríguez, 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).

28. Richter, D. J. & Nitsche, F. Choanoflagellatea. In Mobile DNA. https://doi.org/10.1186/s13100-019-0189-9 (2019).

29. Redrejo-Rodríguez, 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).

30. Dayel, M. J. & King, N. Prey capture and phagocytosis in the choanoflagellate Salpingoeca rosetta. PLoS ONE 9, e95577 (2014).

31. Dayel, M. J. et al. Cell differentiation and morphogenesis in the colony-forming choanoflagellate Salpingoeca rosetta. Dev. Biol. 357, 73–82 (2011).

32. Richter, D. J. & Nitsche, F. Choanoflagellatea. In Mobile DNA. https://doi.org/10.1186/s13100-019-0189-9 (2019).
29. Hoffmeyer, T. T. & Burkhardt, P. Chaoanoflagellate 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 chaoanoflagellate 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 Choanoflagellate. J. Eukaryot. Microbiol. 58, 452–462 (2011).
34. Fairclough, S. R.
35. Nitsche, F., Carr, M., Arndt, H. & Leadbeater, B. S. C. Higher level taxonomy and molecular phylogenetics of the Choanoflagellate. J. Eukaryot. Microbiol. 58, 452–462 (2011).
36. King, N. et al. The genome of the chaoanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 451, 783–788 (2008).
37. Jeuck, A., Arndt, H. & Nitsche, F. Extended phylogeny of the Craspedida (Choanomona). Eur. J. Protistol. 50, 430–443 (2014).
38. Marron, A. O. et al. A family of diatom-like silicon transporters in the siliceous loricacean chaoanoflagellates. Proc. R. Soc. B Biol. Sci. 280, 20122543 (2013).
39. Baptiste, E., Moreira, D. & Philippe, H. Rampant horizontal gene transfer and phospo-donor change in the evolution of the phosphofructokinase. Gene 318, 185–191 (2003).
40. Malik, S. B., Ramesh, M. A., Hulstrand, A. M. & Logsdon, J. M. Protist homologs of the meiotic Spo11 gene and topoisomerase.
41. Sun, G. & Huang, J. Horizontally acquired DAP pathway as a unit of self-regulation.
42. Maruyama, S., Matsuzaki, M., Misawa, K. & Nozaki, H. Cyanobacterial contribution to the genomes of the plastid-lacking.
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. Fairclough, S. R., Dayel, M. J. & King, N. Multicellular development in a chaoanoflagellate. Curr. Biol. 20, R875 (2010).
45. 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/s12915-015-0209-7 (2015).
46. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 277–280 (2000).
47. Daubin, V., Lerat, E. & Perrière, G. The source of laterally transferred genes in bacterial genomes. Proc. Natl. Acad. Sci. U.S.A. 112, 18695–18700 (2015).
48. Southworth, J. et al. A bacterial sulfonolipid triggers multicellular development in the closest living relatives of animals. Elife 1, e00013 (2012).
49. Woznica, A. et al. Bacterial lipids activate, synergize, and inhibit a developmental switch in chaoanoflagellates. Proc. Natl. Acad. Sci. U.S.A. 113, 7894–7899 (2016).
50. 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).
51. Gladyshnev, E. A., Meselson, M. & Arkhipova, I. R. Massive horizontal gene transfer in bdelloid rotifers. Science 320, 1210–1213 (2008).
52. Richter, D. J., Fozoumi, 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).
53. 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).
54. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz157 (2018).
55. Rivero, 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, 5629–5634 (1998).
56. Langille, M. G. I., Hsiaw, W. W. L. & Brinkman, F. S. L. Detecting genomic islands using bioinformatics approaches. Nat. Rev. Microbiol. 8, 373–382 (2010).
57. Hildebrand, F., Meyer, A. & Eyre-Walker, A. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 6, e1001107 (2010).
58. Hershberg, R. & Petrov, D. A. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 6, e1001115 (2010).
59. 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).
60. Doolittle, W. F. Lateral genomics. Trends Biochem. Sci. 24, M5 (1999).
61. 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).
62. Carr, M. et al. A phylogenomic approach to microbial evolution. Nucleic Acids Res. 29, 545–552 (2001).
63. Rogan, J. P., Senejani, A. G., Zhaohuyaye, O., Olenzulence, H. & Hilario, E. Inteins: Structure, function, and evolution. Annu. Rev. Microbiol. 56, 263–287 (2002).
64. Brown, J. R. Ancient horizontal gene transfer. Nat. Rev. Genet. 4, 121–132 (2003).
65. Wellner, A., Lurie, M. N. & Gophna, U. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. Genome Biol. 8, R156 (2007).
66. 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).
67. Yue, J., Hu, X. & Huang, J. Origin of plant auxin biosynthesis. Trends Plant Sci. 19, 764–770 (2014).
68. Conaco, C. et al. Detection of prokaryotic genes in the Amphilodomon queenslandica genome. PLoS ONE 11, e0151092 (2016).
69. Boschetti, C. et al. Biochemical diversification through foreign gene expression in bdelloid rotifers. PLoS Genet. 8, e1003035 (2012).
70. 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. & Groisman, 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-Valvé, S., Romeu, A. & Palau, J. Horizontal gene transfer of glycoyl hydroxyls 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, Nemastrella 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. Photos. 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. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v4: Recent updates and new developments. Mol. Biol. Evol. 35, 539–542 (2018).
85. Starcevic, A. et al. Enzymes of the shikimic acid pathway encoded in the genome of a basal metazoan, Nemastrella vectensis, have microbial origins. Proc. Natl. Acad. Sci. U.S.A. 105, 2533–2537 (2008).
86. 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).
87. Winnemoller, M., Schon, P. & Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 259–261 (2015).
88. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 259–261 (2015).
89. Conesa, A. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676 (2005).
90. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527 (2016).
91. Trincone, A., Tramice, A., Giordano, A. & Andreotti, G. Glycoside hydrolases in the Philippines Marine Science Institute to DM.
92. Sawaguchi, S. et al. Yeast flocculation: Receptor definition by mnn mutants and concanavalin A. J. Cell Biol. 177, 171–177 (2006).
93. 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).
94. Zhang, F. et al. Yeast flocculation: Receptor definition by mnn mutants and concanavalin A. J. Cell Biol. 177, 171–177 (2006).
95. Stratford, M. Yeast flocculation: Receptor definition by mnn mutants and concanavalin A. J. Cell Biol. 177, 171–177 (2006).
96. Wang, M., Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 259–261 (2015).
97. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552 (2000).
98. Conesa, A. et al. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676 (2005).
99. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60 (2014).
100. R Development Core Team. R: A language and environment for statistical computing.
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.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021