Archaeal origins of gamete fusion

David Moi[1,2,3,*], Shunsuke Nishio[4,*], Xiaohui Li[5,*], Clari Valansi[5], Mauricio Langleib[6,7], Nicolas G. Brukman[5], Kateryna Flyak[5], Christophe Dessimoz[2,3,8,9], Daniele de Sanctis[10], Kathryn Tunyasuvunakool[11], John Jumper[11], Martín Graña[7,#], Héctor Romero[6,12,#], Pablo S. Aguilar[1,13,#], Luca Jovine[4,#] and Benjamin Podbilewicz[5,#]

[1]Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE-CONICET), Buenos Aires, Argentina

[2]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

[3]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[4]Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden

[5]Department of Biology, Technion- Israel Institute of Technology, Haifa, Israel

[6]Unidad de Genómica Evolutiva, Facultad de Ciencias, Universidad de la República, Uruguay

[7]Unidad de Bioinformática, Institut Pasteur de Montevideo, Uruguay

[8]Centre for Life's Origins and Evolution, Dept. of Genetics, Evolution and Environment, University College London, United Kingdom

[9]Department of Computer Science, University College London, United Kingdom

[10]Structural Biology Group, ESRF - The European Synchrotron, Grenoble, France

[11]DeepMind, London, UK

[12]Centro Universitario Regional Este - CURE, Centro Interdisciplinario de Ciencia de Datos y Aprendizaje Automático - CICADA, Universidad de la República, Uruguay

[13]Instituto de Investigaciones Biotecnológicas Dr. Rodolfo A. Ugalde, Universidad Nacional de San Martín (IIB-CONICET), San Martín, Buenos Aires, Argentina

*These authors contributed equally: David Moi, Shunsuke Nishio, Xiaohui Li.

#Correspondence and requests for materials should be addressed to Martin Graña (mgrana@pasteur.edu.uy), Héctor Romero (eletor@fcien.edu.uy), Pablo S. Aguilar (paguilar@iib.unsam.edu.ar), Luca Jovine (luca.jovine@ki.se), Benjamin Podbilewicz (podbilew@technion.ac.il).

Abstract

Sexual reproduction in Eukarya consists of genome reduction by meiosis and subsequent gamete fusion. The presence of meiotic genes in Archaea and Bacteria suggests that prokaryotic DNA repair mechanisms evolved towards meiotic recombination[1,2]. However, the evolutionary origin of gamete fusion is less clear because fusogenic proteins resembling those found in Eukarya have so far not been identified in prokaryotes[3–5]. Here, using bioinformatics, we identified archaeal genes encoding candidates of fusexins, a superfamily of fusogens mediating somatic and gamete fusion in multiple eukaryotic lineages. Crystallographic structure determination of a candidate archaeal FusexinA reveals an archetypical trimeric fusexin architecture with novel features such as a six-helix bundle and an additional globular domain. We demonstrate that ectopically expressed FusexinA can fuse mammalian cells, and that this process involves the additional domain and a more broadly conserved fusion loop. Genome content analyses reveal that archaeal fusexins genes are within integrated mobile elements. Finally, evolutionary analyses place these archaeal fusogens as the founders of the fusexin superfamily. Based on these findings, we propose a new hypothesis on the origins of eukaryotic sex where an archaeal fusexin, originally used by selfish elements for horizontal transmission, was repurposed to enable gamete fusion.

Introduction

How the earliest eukaryotes developed the capacity for gamete fusion is a central question that is entangled with the origins of the eukaryotic cell itself. The widespread presence of a conserved set of meiotic, gamete and nuclear fusion proteins (fusogens) among extant eukaryotes suggests that meiotic sex emerged once, predating the last eukaryotic common ancestor (LECA)[1,6]. The conserved gamete fusogen HAP2/GCS1 belongs to a superfamily of fusion proteins called fusexins[3–5]. This superfamily encompasses class II viral fusogens (viral fusexins) that fuse the envelope of some animal viruses with the membranes of host cells during infection[7–9]; EFF-1 and AFF-1 (somatic fusexins) that promote cell fusion during syncytial organ development[10–14]; and HAP2/GCS1 (sexual fusexins) that mediate gamete fusion[15–17]. Although it is assumed that sexual fusexins were already present in the LECA[8], their shared ancestry with viral fusexins posed a "the virus or the egg" evolutionary dilemma[18]. In one scenario, fusexins are proper eukaryal innovations that were transferred to some viruses and used for host invasion. Alternatively, a viral fusexin gene was captured by an early eukaryotic cell and then repurposed for gamete fusion.

Here, we identify a fourth family of fusexins in the genomes of Archaea and in the prokaryotic fractions of metagenomes from diverse environments. We provide structural and functional evidence indicating that these proteins are cellular fusogens. Genomic and evolutionary analyses reveal the ancient origins of these archaeal fusexins and their lateral mobility within Archaea, leading us to provide a working model for the emergence of meiotic sex during eukaryogenesis.

Results

Fusexin genes in Archaea

To search for fusexins we used the crystallographic structures of HAP2/GCS1 of *C. reinhardtii*, *A. thaliana,* and *T. cruzi* (Cr/At/TcHAP2)[4,19,20] to build dedicated Hidden Markov Models (HMMs). These were used to scan the Uniclust30 database with HHblits (see Methods, Supplementary Information). We detected 24 high confidence candidates in prokaryotes: eight belong to isolated and cultivated archaea, and the remaining sixteen come from metagenomics-assembled genomes (MAGs, **Extended Data Table 1**). We then built HMMs of the candidate ectodomains and compared them to HMMs of sexual, somatic and viral fusexins. **Figure 1a** shows that the prokaryotic candidates are closely related to HAP2/GCS1 (hereafter referred to as HAP2), with E-values below 0.001 and HHblits derived probabilities higher than 90% (**Supplementary Fig. 1**). Since all candidate sequences from pure culture genomes (PCGs) are from Archaea, we decided to name these proteins FusexinA (FsxA). All *fsxA* genes found in cultivated and isolated prokaryotes are restricted to the Halobacteria class (Euryarchaeota superphylum) whereas MAGs containing FsxAs include all major Archaea superphyla (**Extended Data Table 1**). Next, we used this set of FsxA sequences to search the Metaclust database, which comprises 1.59 billion clustered proteins from over 2200 metagenomic and metatranscriptomic datasets. Performing a scan pipeline using PSI-BLAST, HMM-HMM comparisons and topology filtering (see Methods) we found 96 high-confidence *fsxA* genes. The identified *fsxA* genes come from different environments (with preeminence of saline samples), and from a wide range of temperatures (-35 to 80ºC, **Source Data Table 1**).

FsxA is a structural homologue of HAP2/GCS1

To obtain experimental evidence for the presence of fusexin-like proteins in Archaea, a selection of the candidate genes was screened for expression in mammalian cells. High-level expression was observed for a metagenomic FsxA sequence from a hypersaline environment, predicted to encode a large ectodomain region followed by three transmembrane helices (**Supplementary Fig. 2a, b; Source Data Table 1**). Although the protein was prone to denaturation on cryo-EM grids, we could grow crystals of its ~55 kDa ectodomain (FsxA$_E$) in the presence of 2.5 M NaCl, 0.2 M CaCl$_2$ (**Extended Data Fig. 1**). These yielded data to 2.3 Å resolution (**Extended Data Table 2**), which however could not be phased experimentally, probably because the high-salt mother liquor composition hindered heavy atom binding. Molecular replacement with HAP2-derived homology models also failed, but we succeeded in solving the structure using a combination of fragments from models generated by AlphaFold2[21] (**Extended Data Fig. 2**).

Despite being a monomer in solution (**Extended Data Fig. 1c,d**), FsxA$_E$ crystallized as a homotrimer of 119x78x75 Å (**Fig. 1b** and **Extended Data Figs. 1f** and **3**). Each protomer consists of four domains, the first three of which match the approximate dimensions and relative arrangement of domains I-III of fusexins in their post-fusion conformation[22] (**Fig. 1b** and **Extended Data Fig. 4a,b**); accordingly, fold and interface similarity searches identify HAP2 as the closest structural homologue of FsxA$_E$, followed by viral fusexins and *C. elegans* EFF-1 (**Extended Data Fig. 4c**). FsxA domains I and III are relatively sequence-conserved among archaeal homologues (**Extended Data Fig. 5a; Supplementary Fig. 3**) and closely resemble the corresponding domains of HAP2 (RMSD 2.1 Å over 218 Cα), including the invariant disulfide bond between domain III strands βC and βF[4] (C$_3$389-C$_4$432; **Extended Data Figs. 3d** and **4**). On the other hand, FsxA domain II shares the same topology as that of HAP2 but differs significantly in terms of secondary structure elements and their relative orientation, as well as disulfide bonds (**Extended Data Fig. 4d**). In particular, FsxA domain II is characterized by a four-helix hairpin, the N-terminal half of which interacts with the same region of the other two subunits to generate a six-helix bundle around the molecule's three-fold axis (**Figs. 1b** and **2a-c; Extended Data Figs. 3a** and **5b**).

Notably, unlike previously characterized viral and eukaryotic fusexins, FsxA also contains a fourth globular domain conserved among archaeal homologues (**Fig. 1b** and **Extended Data Figs. 3, 4; Supplementary Fig. 3**), whose antiparallel β-sandwich, including the two C-terminal disulfides of the protein, resembles the carbohydrate-binding fold of dust mite allergen Der p 23 and related chitin-binding proteins[23] (**Fig. 2d**); accordingly, it is also structurally similar to a high-confidence AlphaFold2 model of the C-terminal domain of acidic mammalian chitinase[24]. In addition to being coaxially stacked with domain III as a result of a loop/loop interaction stabilized by the $C_5$457-$C_6$477 disulfide, domain IV contributes to the quaternary structure of the protein by interacting with domain II of the adjacent subunit to which domain III also binds (**Figs. 1b** and **2c; Extended Data Fig. 4a, b**).

The $FsxA_E$ monomer has a net charge of -67 (**Fig. 2a**), and another important feature stabilizing its homotrimeric assembly is a set of Ca^{2+} and Na^{+} ions that interacts with negatively charged residues at the interface between subunits (**Fig. 2b**; **Extended Data Figs. 3a** and **5b**). Additional metal ions bind to sites located within individual subunits; in particular, a Ca^{2+} ion shapes the conformation of the domain II c-d loop (S143-V148) so that its uncharged surface protrudes from the rest of the molecule (**Fig. 2b, c, e; Extended Data Figs. 3b** and **5c**). Strikingly, the position of this element matches that of the fusion loops of other fusexins, including the Ca^{2+}-binding fusion surface of rubella virus E1 protein[25,26] (**Fig. 2e**). Moreover, as previously observed in the case of CrHAP2[20], the loops of each trimer interact with those of another trimer within the FsxA crystal lattice.

In summary, despite significant differences in the fold of domain II, the unprecedented presence of a domain IV, and extreme electrostatic properties, the overall structural similarity between FsxA and viral or eukaryotic fusexins strongly suggests that this prokaryotic molecule also functions to fuse membranes.

FsxA can fuse eukaryotic cells

To test the fusogenic activities of the candidate archaeal fusexins we studied their fusion activity upon transfection in eukaryotic cells[3,12,14]. For this, we co-cultured two batches of BHK cells independently transfected with FsxA and coexpressing either nuclear H2B-RFP or H2B-GFP [3]. Following co-culture of the two batches, we fixed, permeabilized and performed immunofluorescence against a V5 tag fused to the cytoplasmic tail of FsxA (**Fig. 3a, b**). We observed a five-fold increase in the mixing of the nuclear H2B-GFP and H2B-RFP compared to vector control, showing that FsxA is a *bona fide* fusogen, comparable in efficiency to the eukaryotic gamete fusexin AtHAP2 (**Fig. 3c; Extended Data Fig. 6**). To determine whether FsxA expression is required in both fusing cells or, alternatively, it suffices in one of the fusing partners, we mixed BHK-FsxA coexpressing cytoplasmic GFP with BHK cells expressing only nuclear RFP. We found increased multinucleation of GFP+ cells but very low mixing with RFP+ cells not expressing FsxA. In contrast, the vesicular stomatitis virus G-glycoprotein (VSVG) fusogen induced efficient unilateral fusion[14] (**Fig. 3d-f; Extended Data Fig. 6**). Thus, FsxA acts in a bilateral way, similarly to the *C. elegans* EFF-1 and AFF-1 fusexins[14,27–29]. We then performed live-imaging using spinning disk confocal microscopy and observed cell-cell fusion of BHK-FsxA cells (**Fig. 3g, h; Supplementary Videos 1** and **2**).

Structure-function analysis of FsxA

To compare archaeal FsxA activity with fusexins from eukaryotes and viruses, we introduced mutations into two specific structural domains of FsxA and tested their surface expression and fusogenic activities in mammalian cells.

First, to test whether the putative fusion loop (FL) of FsxA (143-SVTSPV-148) is involved in fusion, we replaced it with a linker of 4G between Y142A and Y149A (**Figs. 2e** and **3i; Supplementary Figs. 3** and **4;** ΔFL→AG$_4$A). This FL replacement did not affect surface expression yet resulted in a reduction in content mixing to levels similar to those of the negative control (**Fig. 3j; Extended Data Fig. 7**).

Second, we asked whether domain IV, which is present in archaeal fusexins but absent in known eukaryotic and viral fusexins, has a function in the fusion process. For this, we replaced the entire domain with the stem region of *C. elegans'* EFF-1 (**Fig. 3i;** ΔDIV→EFF-1 stem). While this mutant FsxA reaches the cell surface, suggesting that it folds normally, it showed a significantly reduced activity compared to wildtype FsxA (**Fig. 3j; Extended Data Fig. 7; Supplementary Fig. 4**).

FsxAs are ancestral fusogens associated with integrated mobile elements

The sparse pattern of FsxA presence in Archaea led us to perform genomic comparisons of related species with and without the *fsxA* gene. These comparisons revealed large DNA insertions (> 50 kbp) in the genomes of species with *fsxA* genes (**Supplementary Fig. 5**), which we analysed in more detail. We first performed k-mer spectrum analysis on *fsx*A-containing genomes of pure cultured species and found divergent regions containing the *fsxA* ORF (**Figure 4a; Supplementary Fig. 6**). Then, performing homology searches (**Supplementary Fig. 7**), we studied the gene content of *fsxA*-containing regions. We found that they share a portion of their genes (**Supplementary Fig. 8**) and display conserved synteny (**Fig. 4b**; **Supplementary Fig. 9**), suggesting common ancestry. These regions are enriched in ORFs that show homology with proteins involved in DNA mobilization and integration such as the type-IV secretion system VirB4/TrbE and TraG/VirD4 ATPases, the HerA helicase and the XerC/D tyrosine recombinase (**Figure 4b**; **Supplementary Table 1**). Thus, our results suggest that *fsxA* genes are contained in integrated mobile elements (IMEs) that can be mobilized by a conjugative-like, cell fusion-dependent mechanism.

To describe FsxA's tempo and mode of evolution we built maximum likelihood phylogenies for a set of FsxA sequences derived from isolated species, metagenomic samples and MAGs, and a subset of HAP2s. We found that the branching pattern of FsxA sequences is incompatible with their species tree (**Extended Data Fig. 8**). This incongruence supports a history of horizontal gene transfer (HGT) events within Archaea, in line with *fsxAs* presence in IMEs. Moreover, HAP2 monophyly indicates an ancient split before the eukaryal radiation (**Extended Data Fig. 8**).

To analyze deep homologies with no sequence signal, we built a structural comparison tree between archaeal, eukaryotic and viral fusexins (**Fig. 4c**). In this minimum evolution tree, both Minimal Ancestor Deviation[30] (MAD) and midpoint rooting reveal the position of the root within the archaeal branch. These results are in line with current eukaryogenesis models that point to a history of massive acquisition of prokaryotic genes by HGT during the transition to LECA, suggesting FsxAs are basal, predating the divergence between eukaryotic and viral fusexins. Furthermore, relative acquisition time analysis (**Supplementary Fig. 10**) suggests that *fsxA* was an intermediate-to-late acquisition in the transition from First Eukaryotic Common Ancestor (FECA) to LECA .

Discussion

The archaeal fusexins herein identified reveal a broader presence of these fusogens in yet another domain of life and with different types of membranes. We also unveil a wider physicochemical landscape for this protein superfamily, from cold hypersaline lakes to hot springs and hydrothermal vents (**Source Data Table 1** and **Extended Data Fig. 8**).

Our structural and functional analyses show that FsxA has both conserved and diverged properties when compared to eukaryotic and viral fusexins (**Extended Data Fig. 4; Fig. 3**). Like its viral counterparts, FsxA has an uncharged loop that is essential for fusion. Unlike any other previously known fusexin, FsxA possesses an additional domain (domain IV), that is important for FsxA activity and may bind sugars (**Figs. 2d** and **3j**). Considering that cell surface glycosylation was found to be important for fusion-based mating of halophilic archaea[31], this domain may actively promote fusion by interacting with carbohydrates attached to lipids or proteins such as S-layer glycoproteins[32]. Like somatic and sexual fusexins, FsxA mediates BHK cell fusion in a bilateral fashion (**Fig. 3f**). Future studies will aim at understanding the importance of the six-helix bundle formed by FsxA domain II, which is unprecedented among fusexins and raises an unexpected structural connection with class I viral fusogens[8,9].

The presence of FsxAs in Halobacteria IMEs is consistent with the evolutionary history and genetic structure of members of this class. Halophilic archaea are notorious for being polyploid[33] and undergoing HGT events that overcome species and genera barriers[34,35]. The best evidence of archaeal cell fusion comes from studies showing bilateral DNA exchange that correlates with cytoplasmic bridges made up of fused lipid bilayers connecting haloarchaeal cells[32,36,37]. Thus, it is plausible that Halobacteria evolved HGT mechanisms based on conjugative-like DNA mobilization and cell-cell fusion[38]. FsxAs seem to be absent in some genomes of archaeal species known to undergo fusion-based mating for gene transfer[37]. Their relative confinement to few archaeal lineages suggests limited fitness advantages to their present bearers indicating they are molecular relics, playing a marginal role in

Archaea. Current evidence suggests that major eukaryotic lineages such as chordata and fungi replaced HAP2 with other fusogens during evolution[39]. Therefore, other unidentified archaeal fusogens may be at play. More broadly, cell fusion-based HGT might have declined during archaeal evolution in favour of conjugation, transduction and natural transformation.

The lateral mobility of *fsxA* genes and their likely ancestral position (**Extended Data Fig. 8**, **Fig. 4c**) prompts us to abandon the “virus or the egg” dilemma of the origin of fusexins[18] in favour of a hypothesis where it was archaeal fusexins who were repurposed (exapted) for gamete fusion. In the hypothesis that we are calling "eukaryotic sexaptation", fusexins paved the way for sexual reproduction and other processes relying on membrane fusion during the FECA to LECA transition (**Fig. 4d**).

Discovery of the Asgard superphylum[40] and the successful recent cultivation of one of its members[41] have lent weight to eukaryogenesis models where heterogeneous populations of bacteria and archaea lived in syntrophy transferring metabolites and genetic information[42]. Lateral transfer of a *fsxA* gene, presumably at a mid-stage of eukaryogenesis (**Supplementary Fig. 10, Fig. 4d**), could thus have enabled pre-LECA cells to undergo genome expansion, explore syncytial forms[43], and evolve into mononucleated cells fully equipped for meiosis and gamete fusion[44,45]. Our findings suggest that today's eukaryotic sexual reproduction is the result of over two billion years of evolution of this ancient archaeal cell fusion machine.

References

1. Ramesh, M. A., Malik, S.-B. & Logsdon, J. M., Jr. A phylogenomic inventory of meiotic genes; evidence for sex in Giardia and an early eukaryotic origin of meiosis. *Curr. Biol.* **15**, 185–191 (2005).
2. Bergerat, A. *et al.* An atypical topoisomerase II from Archaea with implications for meiotic recombination. *Nature* **386**, 414–417 (1997).
3. Valansi, C. *et al.* Arabidopsis HAP2/GCS1 is a gamete fusion protein homologous to somatic and viral fusogens. *J. Cell Biol.* **216**, 571–581 (2017).
4. Fédry, J. *et al.* The Ancient Gamete Fusogen HAP2 Is a Eukaryotic Class II Fusion Protein. *Cell* **168**, 904–915.e10 (2017).

5. Pinello, J. F. *et al.* Structure-Function Studies Link Class II Viral Fusogens with the Ancestral Gamete Fusion Protein HAP2. *Curr. Biol.* **27**, 651–660 (2017).
6. Speijer, D., Lukeš, J. & Eliáš, M. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8827–8834 (2015).
7. Rey, F. A., Heinz, F. X., Mandl, C., Kunz, C. & Harrison, S. C. The envelope glycoprotein from tick-borne encephalitis virus at 2 Å resolution. *Nature* **375**, 291–298 (1995).
8. Harrison, S. C. Viral membrane fusion. *Virology* **479-480**, 498–507 (2015).
9. Kielian, M. Mechanisms of Virus Membrane Fusion Proteins. *Annu. Rev. Virol.* **1**, 171–189 (2014).
10. Mohler, W. A. *et al.* The type I membrane protein EFF-1 is essential for developmental cell fusion. *Dev. Cell* **2**, 355–362 (2002).
11. Podbilewicz, B. *et al.* The *C. elegans* developmental fusogen EFF-1 mediates homotypic fusion in heterologous cells and in vivo. *Dev. Cell* **11**, 471–481 (2006).
12. Pérez-Vargas, J. *et al.* Structural basis of eukaryotic cell-cell fusion. *Cell* **157**, 407–419 (2014).
13. White, J. M. The first family of cell-cell fusion. *Dev. Cell* **12**, 667–668 (2007).
14. Avinoam, O. *et al.* Conserved eukaryotic fusogens can fuse viral envelopes to cells. *Science* **332**, 589–592 (2011).
15. Johnson, M. A. *et al.* Arabidopsis hapless mutations define essential gametophytic functions. *Genetics* **168**, 971–982 (2004).
16. Mori, T., Kuroiwa, H., Higashiyama, T. & Kuroiwa, T. GENERATIVE CELL SPECIFIC 1 is essential for angiosperm fertilization. *Nat. Cell Biol.* **8**, 64–71 (2006).
17. Liu, Y. *et al.* The conserved plant sterility gene *HAP2* functions after attachment of fusogenic membranes in *Chlamydomonas* and *Plasmodium* gametes. *Genes Dev.* **22**, 1051–1068 (2008).
18. Doms, R. W. What Came First-the Virus or the Egg? *Cell* **168**, 755–757 (2017).
19. Fedry, J. *et al.* Evolutionary diversification of the HAP2 membrane insertion motifs to drive gamete fusion across eukaryotes. *PLoS Biol.* **16**, e2006357 (2018).
20. Feng, J. *et al.* Fusion surface structure, function, and dynamics of gamete fusogen HAP2. *Elife* **7**, e39772 (2018).

21. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
22. Modis, Y., Ogata, S., Clements, D. & Harrison, S. C. Structure of the dengue virus envelope protein after membrane fusion. *Nature* **427**, 313–319 (2004).
23. Mueller, G. A. *et al.* Serological, genomic and structural analyses of the major mite allergen Der p 23. *Clin. Exp. Allergy* **46**, 365–376 (2016).
24. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
25. DuBois, R. M. *et al.* Functional and evolutionary insight from the crystal structure of rubella virus protein E1. *Nature* **493**, 552–556 (2013).
26. Dubé, M., Etienne, L., Fels, M. & Kielian, M. Calcium-Dependent Rubella Virus Fusion Occurs in Early Endosomes. *J. Virol.* **90**, 6303–6313 (2016).
27. Shemer, G. *et al.* EFF-1 is sufficient to initiate and execute tissue-specific cell fusion in *C. elegans*. *Curr. Biol.* **14**, 1587–1591 (2004).
28. Sapir, A. *et al.* AFF-1, a FOS-1-regulated fusogen, mediates fusion of the anchor cell in *C. elegans*. *Dev. Cell* **12**, 683–698 (2007).
29. Gattegno, T. *et al.* Genetic control of fusion pore expansion in the epidermis of *Caenorhabditis elegans*. *Mol. Biol. Cell* **18**, 1153–1166 (2007).
30. Tria, F. D. K., Landan, G. & Dagan, T. Phylogenetic rooting using minimal ancestor deviation. *Nat. Ecol. Evol.* **1**, 193 (2017).
31. Shalev, Y., Turgeman-Grott, I., Tamir, A., Eichler, J. & Gophna, U. Cell Surface Glycosylation Is Required for Efficient Mating of *Haloferax volcanii*. *Front. Microbiol.* **8**, 1253 (2017).
32. Sivabalasarma, S. *et al.* Analysis of Cell-Cell Bridges in *Haloferax volcanii* Using Electron Cryo-Tomography Reveal a Continuous Cytoplasm and S-Layer. *Front. Microbiol.* **11**, 612239 (2020).
33. Ludt, K. & Soppa, J. Polyploidy in halophilic archaea: regulation, evolutionary advantages, and gene conversion. *Biochem. Soc. Trans.* **47**, 933–944 (2019).
34. Turgeman-Grott, I. *et al.* Pervasive acquisition of CRISPR memory driven by inter-species mating of archaea can limit gene transfer and influence speciation. *Nat. Microbiol.* **4**, 177–186 (2019).
35. DeMaere, M. Z. *et al.* High level of intergenera gene exchange shapes the evolution of haloarchaea in an isolated Antarctic lake. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 16939–16944 (2013).

36. Rosenshine, I., Tchelet, R. & Mevarech, M. The mechanism of DNA transfer in the mating system of an archaebacterium. *Science* **245**, 1387–1389 (1989).
37. Naor, A., Lapierre, P., Mevarech, M., Papke, R. T. & Gophna, U. Low Species Barriers in Halophilic Archaea and the Formation of Recombinant Hybrids. *Curr. Biol.* **22**, 1444–1448 (2012).
38. Wagner, A. *et al.* Mechanisms of gene flow in archaea. *Nat. Rev. Microbiol.* **15**, 492–501 (2017).
39. Brukman, N. G., Uygur, B., Podbilewicz, B. & Chernomordik, L. V. How cells fuse. *J. Cell Biol.* **218**, 1436–1451 (2019).
40. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
41. Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).
42. O'Malley, M. A., Leger, M. M., Wideman, J. G. & Ruiz-Trillo, I. Concepts of the last eukaryotic common ancestor. *Nat. Ecol. Evol.* **3**, 338–344 (2019).
43. Skejo, J. *et al.* Evidence for a Syncytial Origin of Eukaryotes from Ancestral State Reconstruction. *Genome Biol. Evol.* **13**, (2021).
44. Baum, D. A. & Baum, B. An inside-out origin for the eukaryotic cell. *BMC Biol.* **12**, 76 (2014).
45. Koonin, E. V. Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140333 (2015).
46. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
47. Baquero, E., Fedry, J., Legrand, P., Krey, T. & Rey, F. A. Species-Specific Functional Regions of the Green Alga Gamete Fusion Protein HAP2 Revealed by Structural Studies. *Structure* **27**, 113–124.e4 (2019).
48. Mirdita, M. *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176 (2017).
49. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).
50. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).

51. Sievers, F. & Higgins, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* **1079**, 105–116 (2014).
52. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
53. Tsirigos, K. D., Peters, C., Shu, N., Kall, L. & Elofsson, A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.* **43(W1)**, W401-W407 (2015).
54. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960 (2005).
55. Raj, I. *et al.* Structural Basis of Egg Coat-Sperm Recognition at Fertilization. *Cell* **169**, 1315–1326.e17 (2017).
56. Dunsing, V. *et al.* Optimal fluorescent protein tags for quantifying protein oligomerization in living cells. *Sci. Rep.* **8**, 1–12 (2018).
57. DuBridge, R. B. *et al.* Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. *Mol. Cell. Biol.* **7**, 379–387 (1987).
58. Pernot, P. *et al.* Upgraded ESRF BM29 beamline for SAXS on macromolecules in solution. *J. Synchrotron Radiat.* **20**, 660–664 (2013).
59. Round, A. *et al.* BioSAXS Sample Changer: a robotic sample changer for rapid and reliable high-throughput X-ray solution scattering experiments. *Acta Crystallogr. D Biol. Crystallogr.* **71**, 67–75 (2015).
60. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J. & Svergun, D. I. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *J. Appl. Crystallogr.* **36**, 1277–1282 (2003).
61. Manalastas-Cantos, K. *et al.* ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *J. Appl. Crystallogr.* **54**, 343–355 (2021).
62. Svergun, D., Barberato, C. & Koch, M. H. J. CRYSOL– a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Crystallogr.* **28**, 768–773 (1995).
63. Franke, D. & Svergun, D. I. DAMMIF, a program for rapid *ab-initio* shape determination in small-angle scattering. *J. Appl. Crystallogr.* **42**, 342–346 (2009).
64. Volkov, V. V., Svergun, D. I. & IUCr. Uniqueness of *ab initio* shape determination in small-angle scattering. *J. Appl. Crystallogr.* **36**, 860–864 (2003).
65. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

66. Nurizzo, D. *et al.* The ID23-1 structural biology beamline at the ESRF. *J. Synchrotron Radiat.* **13**, 227–238 (2006).
67. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
68. Vagin, A. & Teplyakov, A. MOLREP: an Automated Program for Molecular Replacement. *J. Appl. Crystallogr.* **30**, 1022–1025 (1997).
69. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).
70. Matthews, B. W. Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497 (1968).
71. Kantardjieff, K. A. & Rupp, B. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci.* **12**, 1865–1871 (2003).
72. AlQuraishi, M. Machine learning in protein structure prediction. *Curr. Opin. Chem. Biol.* **65**, 1–8 (2021).
73. McCoy, A. J., Sammito, M. D. & Read, R. J. Possible Implications of AlphaFold2 for Crystallographic Phasing by Molecular Replacement. *bioRxiv* 2021.05.18.444614 (2021). doi:10.1101/2021.05.18.444614.
74. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
75. Terwilliger, T. C. *et al.* Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (2008).
76. Chojnowski, G., Pereira, J. & Lamzin, V. S. Sequence assignment for low-resolution modelling of protein crystal structures. *Acta Crystallogr. D Struct. Biol.* **75**, 753–763 (2019).
77. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355–367 (2011).
78. Casañal, A., Lohkamp, B. & Emsley, P. Current developments in Coot for macromolecular model building of Electron Cryo-microscopy and Crystallographic Data. *Protein Sci.* **29**, 1069–1078 (2020).
79. Croll, T. I. *ISOLDE*: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. D Struct. Biol.* **74**, 519–530 (2018).

80. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).
81. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. *Acta Crystallogr. D Struct. Biol.* **75**, 861–877 (2019).
82. Thorn, A. & Sheldrick, G. M. ANODE: anomalous and heavy-atom density calculation. *J. Appl. Crystallogr.* **44**, 1285–1287 (2011).
83. Zheng, H. *et al. CheckMyMetal*: a macromolecular metal-binding validation tool. *Acta Crystallogr. D Biol. Crystallogr.* **73**, 223–233 (2017).
84. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).
85. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182 (1998).
86. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).
87. Holm, L. Using Dali for Protein Structure Comparison. *Methods Mol. Biol.* **2112**, 29–42 (2020).
88. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2256–2268 (2004).
89. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
90. de Beer, T. A. P., Berka, K., Thornton, J. M. & Laskowski, R. A. PDBsum additions. *Nucleic Acids Res.* **42**, D292–6 (2014).
91. Tina, K. G., Bhadra, R. & Srinivasan, N. PIC: Protein Interactions Calculator. *Nucleic Acids Res.* **35**, W473–W476 (2007).
92. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
93. Krieger, E. *et al.* Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins* **77**, 114–122 (2009).

94. Dolinsky, T. J. *et al.* PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35**, W522–W525 (2007).
95. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10037–10041 (2001).
96. Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350 (2016).
97. Sievers, A. *et al.* K-mer Content, Correlation, and Position Analysis of Genome DNA Sequences for the Identification of Function and Evolutionary Features. *Genes* **8**, 122 (2017).
98. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–41 (2004).
99. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life* **5**, 818–840 (2015).
100. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
101. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
102. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
103. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
104. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
105. Vosseberg, J. *et al.* Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat. Ecol. Evol.* **5**, 92–100 (2021).

106. Vosseberg, J. *et al.* Data for: Timing the origin of eukaryotic cellular complexity with ancient duplications. (2019) doi:10.6084/m9.figshare.10069985.v1.

107. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826 (1986).

108. Holm, L. & Sander, C. Mapping the protein universe. *Science* **273**, 595–603 (1996).

109. Ye, Y. & Godzik, A. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.* **32**, W582–W585 (2004).

110. Kanai R. *et al.* Crystal Structure of West Nile Virus Envelope Glycoprotein Reveals Viral Surface Epitopes. *J. Virol.* **80**, 11000–11008 (2006).

111.Klein D. E., Choi J. L. & Harrison S. C. Structure of a Dengue Virus Envelope Protein Late-Stage Fusion Intermediate. *J. Virol.* **87**, 2287–2293 (2013).

112. Gibbons, D. L. *et al.* Conformational change and protein-protein interactions of the fusion protein of Semliki Forest virus. *Nature* **427**, 320–325 (2004).

113. Voss, J. E. *et al.* Glycoprotein organization of Chikungunya virus particles revealed by X-ray crystallography. *Nature* **468**, 709–712 (2010).

114. Guardado-Calvo, P. *et al.* A glycerophospholipid-specific pocket in the RVFV class II fusion protein drives target membrane insertion. *Science* **358**, 663–667 (2017).

115. Li, Z., Natarajan, P., Ye, Y., Hrabe, T. & Godzik, A. POSA: a user-driven, interactive multiple protein structure alignment server. *Nucleic Acids Res.* **42**, W240–5 (2014).

116. Dong, R., Peng, Z., Zhang, Y. & Yang, J. mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics* **34**, 1719–1725 (2018).

117. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

118. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).

119. Lefort, V., Desper, R. & Gascuel, O. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Mol. Biol. Evol.* **32**, 2798–2800 (2015).

120. Edelstein, A. D. *et al.* Advanced methods of microscope control using μManager software. *J. Biol. Methods* **1**, e10 (2014).

121. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).

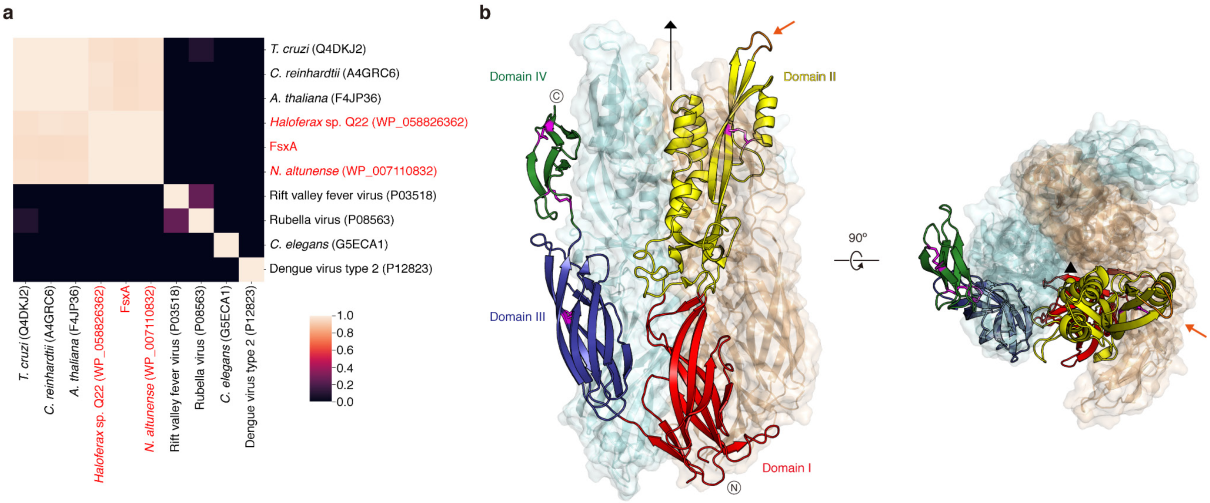


Fig. 1 | FsxA is a member of the fusexin protein superfamily.

a, HMM vs HMM homology probabilities (colour scale) of eukaryotic, viral and archaeal fusexin ectodomains. HMMs were constructed for known fusexin ectodomain sequences with corresponding crystal structures (UniProt identifiers shown in black) and FsxA sequences (NCBI identifiers shown in red except for FsxA; see Methods). All vs. all probabilities of homology as determined by HHblits were clustered using UPGMA.

b, Crystal structure of the trimeric ectodomain of FsxA. Subunit A is shown as a cartoon, with domains I, II, III and IV coloured red, yellow, blue and green, respectively; disulfide bonds are magenta; the putative fusion loop of domain II is coloured orange and indicated by an orange arrow. Subunits B, C are in mixed cartoon/surface representation, and the position of the three-fold non-crystallographic axis is indicated.

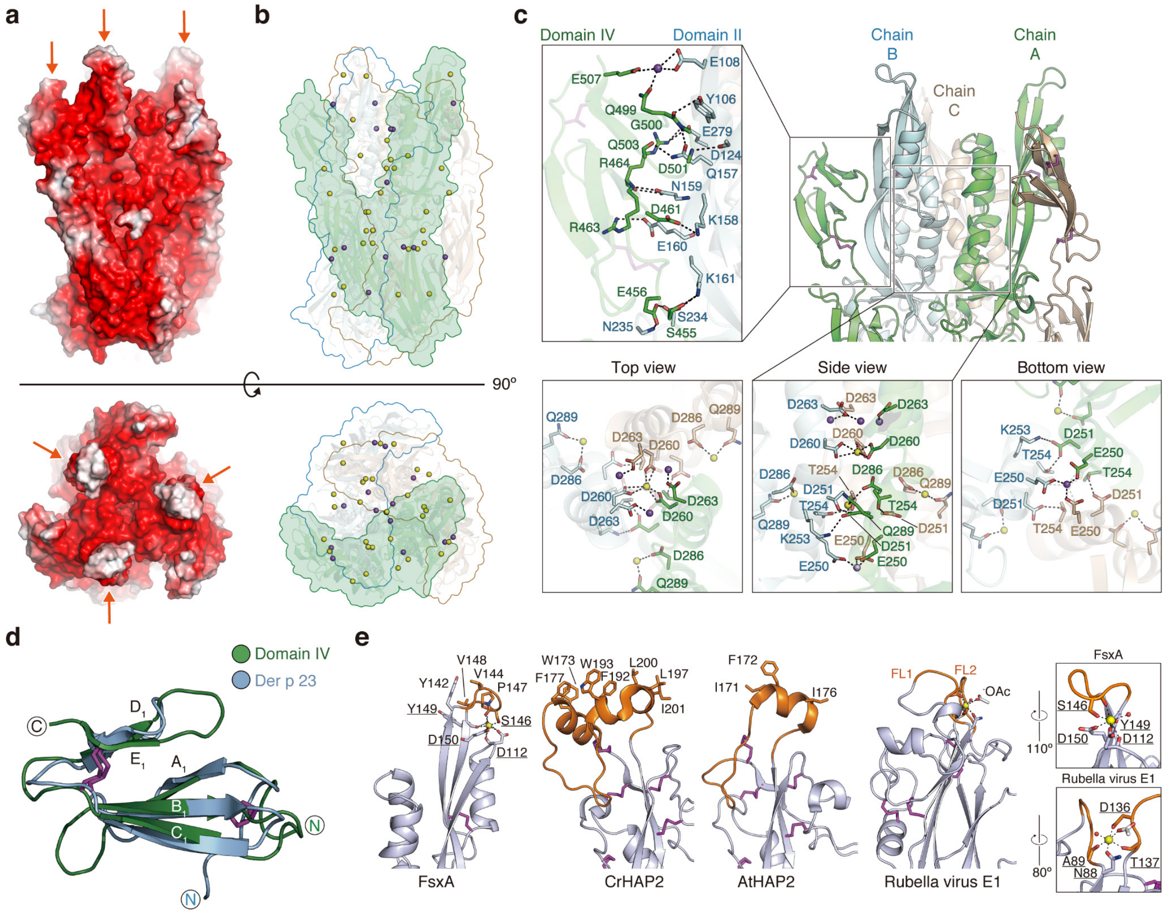


Fig. 2 | Distinct structural features of FsxA.

a, FsxA$_E$ surface coloured by electrostatic potential from red (-5 kT/e) to blue (+5 kT/e) through white (0 kT/e). Orange arrows indicate the putative fusion loops.

b, Location of the ions stabilizing the FsxA$_E$ trimer. Ca^{2+} and Na^{+} ions are depicted as yellow and purple spheres, respectively, with the molecular surface of a protomer shaded green and the outline of the other two subunits coloured cyan and wheat.

c, Details of interactions at the level of the six-helical bundle made by domain II of the FsxA subunits (bottom subpanels) and the domain IV/domain II interface (left subpanel). Selected side chains are shown in stick representation, with hydrogen bonds indicated by dashed lines.

d, Superposition of FsxA domain IV and Der p 23 (PDB 4CZE; Dali Z-score 3.6, RMSD 2.2 Å).

e, Comparison of the FsxA region that includes the putative fusion loop and the corresponding parts of other fusexins (PDB 6E18, 5OW3, 4B3V). Residues coordinating the Ca^{2+} ion (yellow sphere) that stabilizes the FsxA fusion loop are underlined, and compared to the Ca^{2+}-binding region of Rubella virus E1 protein in the boxed panels on the far right.

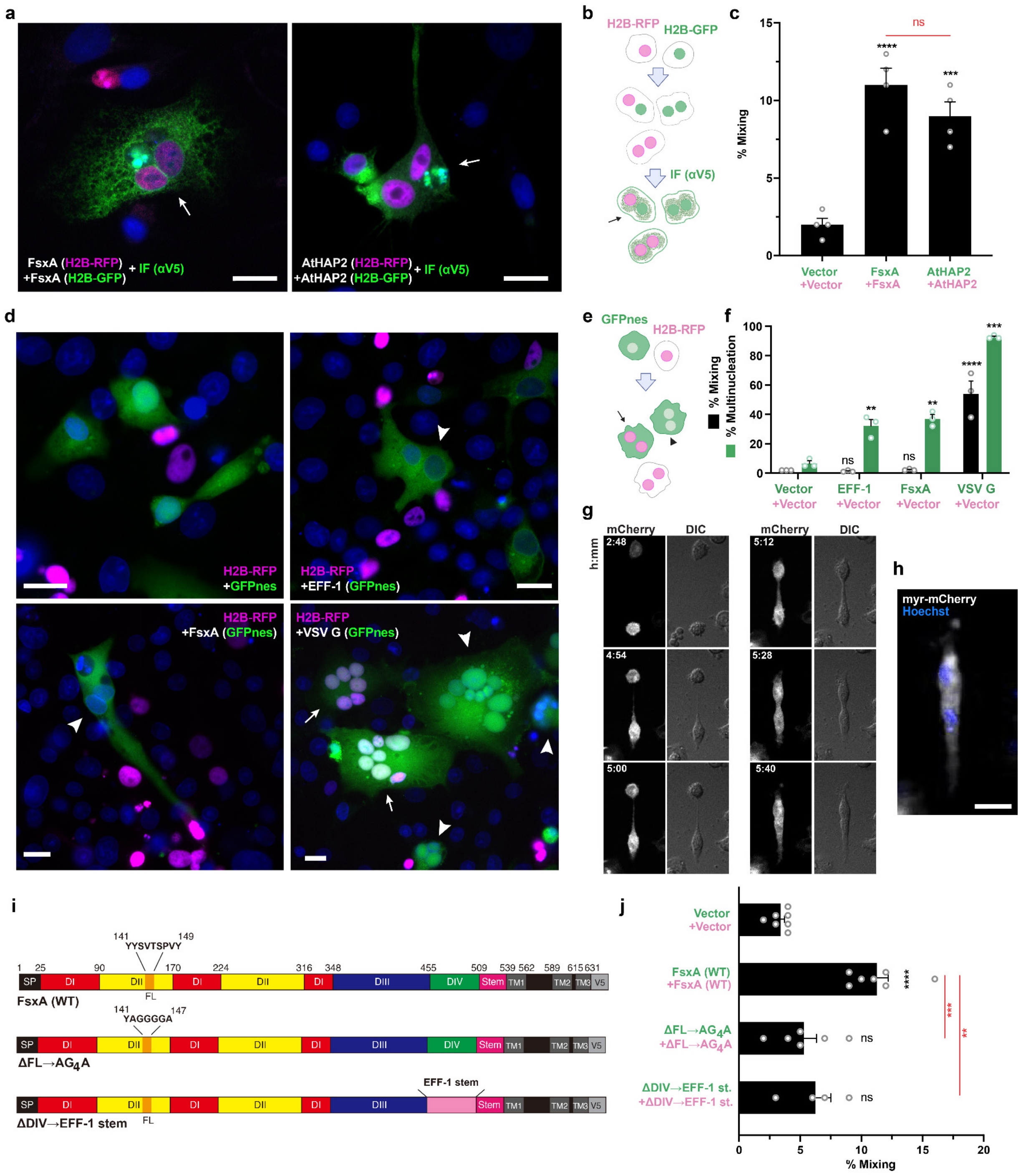


Fig. 3 | FsxA mediates bilateral cell-cell fusion.

a-c, Cell-cell fusion was measured by content-mixing, indicated by the appearance of multinucleated cells containing green nuclei (H2B-GFP) and magenta nuclei (H2B-RFP). Immunofluorescence against the V5 tag was performed (green).

a, Representative images of mixed cells. DAPI, blue.

b, Scheme of experimental design.

c, Quantification of content-mixing. The mixing indices presented as means ± SEM of four independent experiments. Comparisons by one-way ANOVA followed by Bonferroni's test. ns = non-significant, *** $p < 0.001$, **** $p < 0.0001$.

d-f, Unilateral fusion was evaluated by mixing control cells expressing nuclear H2B-RFP (magenta) with cells expressing GFP with a nuclear export signal (GFPnes, green cytoplasm) only or together with FsxA, EFF-1 or VSV G.

d, Images of cells transfected with empty::GFPnes vector, FsxA::GFPnes, EFF-1::GFPnes or VSV G::GFPnes. FsxA and EFF-1 show multinucleated GFPnes positive cells (arrowheads). VSV G multinucleated cells are found with GFP only (arrowheads) or with both markers (arrows). Scale Bars, 20 μm.

e, Scheme of experimental design.

f, Quantification of content-mixing experiments in which only the GFP population of cells express FsxA, EFF-1, VSV G or none of them (vector). Bar chart showing means ± SEM of three independent experiments. Comparisons by one-way ANOVA followed by Dunett's test against the vector. ns = non-significant, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

g, Time-lapse images taken by spinning disk microscopy experiments indicating the merging of two cells expressing myr-mCherry and FsxA. Time in hours:minutes. The red channel (mCherry, white) and the DIC are shown. Refer to **Supplementary Video 1**.

h, For the last point a Z-projection showing the myr-mCherry fluorescence (white) and the nuclei Hoechst (blue; **Supplementary Video 2**). Scale Bar, 20 μm.

i-j Structure function analysis of FsxA. **i,** Schematic diagram of wild type FsxA and two mutants. SP, signal peptide; FL, putative fusion loop (143-SVTSPV-148; **Fig. 2e**); TM, predicted transmembrane helices.

j, Quantification of content-mixing experiments in which both populations of cells express FsxA (wt) (n=7), ΔFL→AG_4A (n=6), ΔDIV→EFF-1 stem (n=4) or none of them (vector) (n=7). Bar chart showing means ± SEM. Comparisons by one-way ANOVA followed by Bonferroni's test against the vector (black) and against FsxA (red). ns = non-significant, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

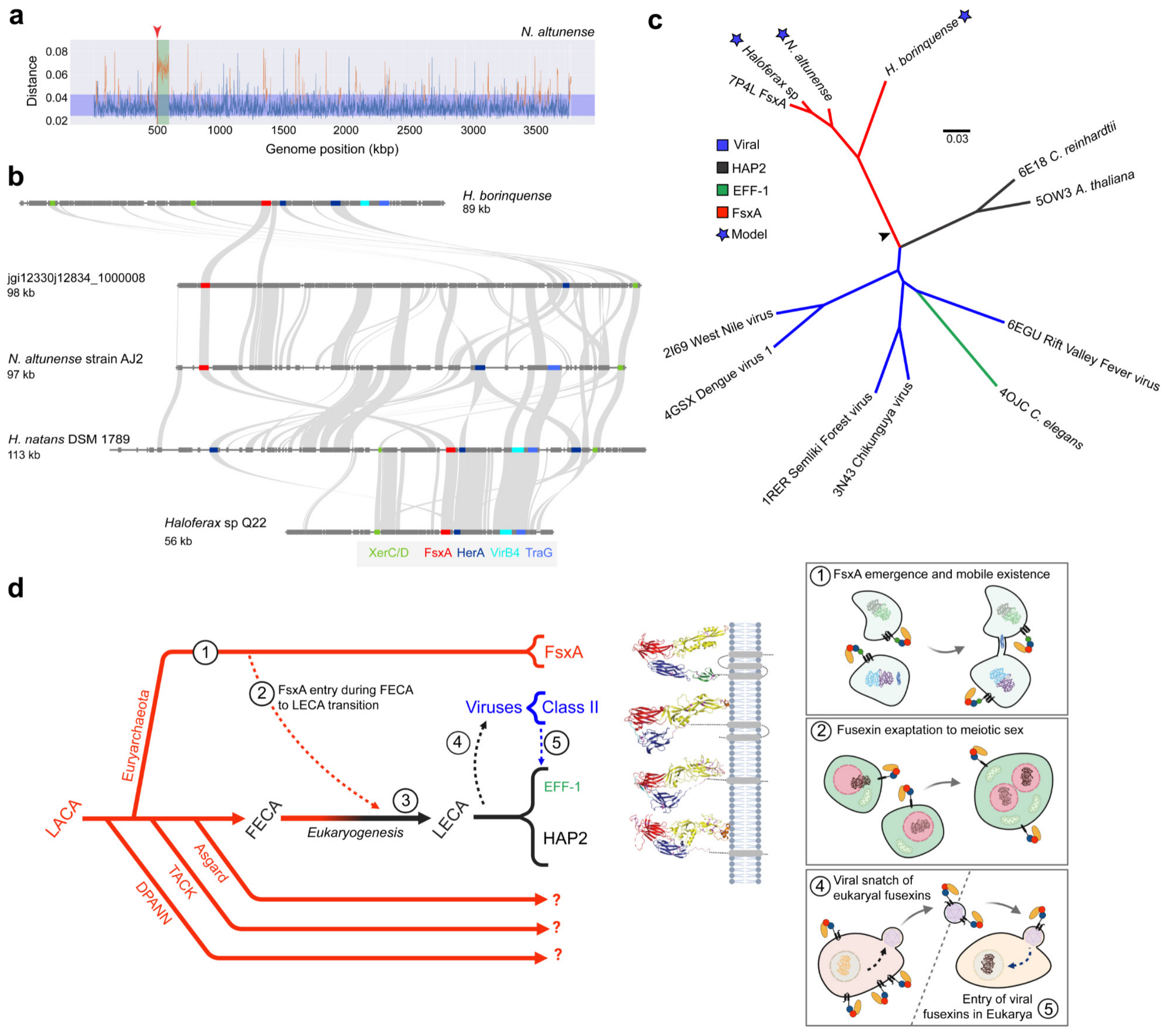


Fig. 4 | Genomic features of FsxAs and evolutionary history of fusexins.

a, FsxAs are embedded in integrated mobile elements (IMEs). *Natrinema altunense* AJ2 complete genome k-mer spectrum deviation from centroid. Blue region shows the standard deviation. Locus of *fsxA* is indicated by the red arrowhead; the mobile element containing *fsxA* is in green.

b, Four IMEs from Pure Culture Genome sequences and the metagenomic sequence from which the crystallized FsxA was obtained (second from top) are represented with their annotated genes (thick segments). Grey links are drawn between homologues. The *fsxA* gene is marked in red and selected ORFs homologous to IME signature genes are labelled and colour-coded. XerC/XerD recombinases from family integrase (green); HerA helicase (dark blue); VirB4, Type IV secretory (T4SS) pathway (cyan); TraG/TraD/VirD4 family enzyme, ATPase T4SS (light blue); (see also **Supplementary Tables 1** and **2**; **Supplementary Fig. 9**).

c, Models based on the FsxA$_E$ structure (PDB 7P4L), and crystal structures of representative fusexins were compared using flexible structural alignment to build a minimum evolutionary tree. Scale bar represents distance as 1-TMscore (see Methods). These TM distances indicate that all structures at the leaves are homologous. FsxA, HAP2 and viral lineages appear monophyletic, with eukaryotic and archaeal structures conforming a distinct clade. Black arrowhead indicates the position of the root estimated by Minimal Ancestral Deviation method[30]. Flavivirus E: West Nile virus (PDB 2I69), Dengue virus serotype 1 (PDB 4GSX); Alphavirus E1: Semliki Forest virus (PDB 1RER); Chikungunya virus (PDB 3N43); *C. elegans* EFF-1 (PDB 4OJC); Bunyavirus Gc Rift Valley fever virus (PDB 6EGU); Eukaryotic HAP2 from *A. thaliana* (PDB 5OW3) and *C. reinhardtii* (PDB 6E18).

d, Fusexins evolutionary model. Likely originated in Euryarchaeota, fusexins form part of integrated mobile elements that promote cell-cell fusion and horizontal gene transfer (HGT) (1). pre-LECA cells received a *fsxA* gene that became fixed during eukaryogenesis and the emergence of meiotic sex (2,3). Viral capture of fusexin genes from early eukaryotic cells and further evolution within the virosphere led to extant viral (class II) fusexins (4). Viral fusexin genes were captured by different eukaryotic lineages, presumably leading to somatic fusogens like EFF-1 (5). Solid lines represent evolutionary trajectories of Archaea (red) and eukaryogenesis (red to black gradient). Dashed arrows represent HGT events. Question marks denote uncertainty regarding the presence of *fsxA*-related genes in the respective evolutionary branches. LACA, Last Archaeal Common Ancestor.

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized.

Initial fusexin search using structurally-guided MSAs

HMMs were prepared using structurally guided multiple sequence alignments (MSAs) of known eukaryotic HAP2 sequences. Structural MSAs were derived using I-TASSER[46] generated models of HAP2 homologues for *Erythranthe guttata* (A0A022QRC8), *Phytomonas* sp. isolate EM1 (W6KUI1), *Plasmodium falciparum* (A0A1C3KGX6), *Chlorella variabilis* (E1Z455) and the HAP2 crystal structures for *Chlamydomonas reinhardtii* (PDB 6E18[47] and 6DBS[20]) and *Arabidopsis thaliana* (PDB 5OW3[19]).

Searches for fusexin homologues using structurally guided MSAs were performed for 3 iterations on the Uniclust database[48] using default HHBlits parameters[49].

HMM-based distance matrices

A taxonomically representative list of known viral and eukaryotic fusexin homologues, covering major lineages, was manually curated. A MSA was built for each homologue by using the sequence as a query on the Uniclust database with HHBlits[49] for 3 iterations (see Supplementary Information). This set of MSAs was compiled into an HHSuite database and each MSA was used as a query against this database to establish a profile-based distance matrix using the probability of homology (**Supplementary Fig. 1**, **Fig. 1a**).

Metaclust database search

We searched the Metaclust[50] dataset using an HMM made of FsxA sequences found in PCGs and MAGs (**Source Data Table 1**; Supplementary Information). FsxA sequences were aligned using ClustalO[51] on default settings for 3 iterations and the resulting MSA was used as a query with hmmsearch[52] against the Metaclust50[50] dataset. All returned sequences with an E-value < 0.0001 with a match length greater than 100 residues were selected for further analysis. Manual curation was performed using membrane protein topology predictor TOPCONS[53] and distant homology searches using HHPRED[54] against the PDB70.

DNA constructs

Ten archaeal genes were synthesized (GenScript) and cloned into pGene/V5-His vectors (**Supplementary Table 3**). Details of nucleotides used for synthesis and protein sequences are described in **Source Data 1**.

For structural studies, a synthetic gene fragment encoding the extracellular region of a metagenomic FsxA ORF (IMG genome 3300000868, scaffold JGI12330J12834_ 1000008, ORF 8; **Source Data Table 1**) (GenScript) was subcloned by PCR in frame with the 5' chicken Crypα signal peptide- and 3' 8xHis-tag-encoding sequences of pLJ6, a mammalian expression vector derived from pHLsec3[55]. The protein construct that yielded the final high-resolution dataset included residues D25-S535 and contained a T369C substitution, introduced by PCR mutagenesis with the aim of facilitating heavy atom derivatization for experimental phasing. Oligonucleotides were from Sigma-Aldrich or IDT and all constructs were verified by DNA sequencing (Eurofins Genomics or Macrogen).

To generate pCI::GFPnes plasmid (see list of plasmids in **Supplementary Table 4**), an oligo DNA encoding for the nuclear export signal (LQKKLEELELD) was cloned into the C-terminal end of EGFP of the pCAGIG plasmid using the enzyme BsrGI. Then, the GFPnes coding sequence was amplified using the pCAGGS FW and pCAGGS RV primers, cut with BmgBI and BglII and used to replace the H2B-GFP coding sequence of the pCI::H2B-GFP plasmid (see list of primers in **Supplementary Table 5**). FsxA-V5, AtHAP2-V5[3], EFF-1-V5, VSV-G[14] and other archaeal fusexins (NaFsxA, HQ22FsxA, HnFsxA) were subcloned into corresponding pCI::H2B-RFP/H2B-GFP/GFPnes vectors separately. For mutagenesis of FsxA, i) FsxA-ΔFL-AG$_4$A: The mutation of Y142A, Y149A and four glycines inserted between them were achieved using PCR with overlapping primers. ii) FsxA-ΔDIV-EFF-1 stem: the stem region of EFF-1(E510-D561) was amplified from pGene::EFF-1-V5 and fused to the upstream and downstream of FsxA-DIV region with overlapping primers. All mutants were ligated into pCI::H2B-RFP and pCI::GFPnes vectors for mixing assay. Additional details are found in **Supplementary Tables 4 and 5**.

Protein expression and purification

HEK293T cells[57] were transiently transfected using 25 kDa branched polyethyleneimine and cultured in DMEM media (Invitrogen) supplemented with 2% (v/v) foetal bovine serum (Biological Industries). 90-96 hours after transfection, the conditioned media from HEK293T cells was harvested, 0.2 µm-filtered (Pall) and adjusted to 20 mM Na-HEPES pH 7.8, 2.5 M NaCl, 5 mM imidazole. 10 ml Ni Sepharose excel beads (GE Healthcare) pre-equilibrated with immobilized metal affinity chromatography (IMAC) buffer (20 mM Na-HEPES pH 7.8, 2.5 M NaCl, 10 mM imidazole) were added to 1 L adjusted conditioned media and incubated overnight at 4ºC. After washing the beads with 100 column volumes IMAC buffer, captured FsxA$_E$ was batch-eluted with 30 mL 20 mM Na-HEPES pH 7.8, 2.5 M NaCl, 500 mM imidazole and concentrated with 30 kDa-cutoff centrifugal filtration devices (Amicon). The material was then further purified by SEC at 4ºC, using an ÄKTAfplc chromatography system (GE Healthcare) equipped with a Superdex 200 Increase 10/300 GL column (GE Healthcare) pre-equilibrated with 20 mM Na-HEPES pH 7.8, 2.5 M NaCl. Peak fractions were pooled and concentrated to 5 mg mL^{-1} (**Extended Data Fig. 1a, b**).

Size exclusion chromatography-multiangle light scattering (SEC-MALS)

Purified FsxA$_E$ (200 µg) were measured using an Ettan LC high-performance liquid chromatography system with a UV-900 detector (Amersham Pharmacia Biotech; λ = 280 nm), coupled with a miniDawn Treos MALS detector (Wyatt Technology; λ = 658 nm) and an Optilab T-rEX dRI detector (Wyatt Technology; λ = 660 nm). Separation was performed at 20ºC using a Superdex 200 Increase 10/300 GL column (GE Healthcare) with a flow rate of 0.5 mL min^{-1} and a mobile phase consisting of 20 mM Na-HEPES pH 7.8, 150 mM NaCl was used (**Extended Data Fig. 1c**). The data processing and weight-averaged molecular mass calculation were performed using the ASTRA 7.1.3 software (Wyatt Technology). BSA (150 µg) was used as a control.

Small-angle X-ray scattering (SAXS)

SAXS experiments were performed at beamline BM29 of the European Synchrotron Radiation Facility (ESRF)[58], using FsxA$_E$ (4.5 mg mL-1) in 20 mM Na-HEPES pH 7.8, 150 mM NaCl. Sample delivery and measurements were performed using a 1 mm thick quartz capillary, which is part of the BM29 BioSAXS automated sample changer unit[59]. Data were collected at 1 Å wavelength in 10 frames of 1 s at 20ºC, using an estimated beam size of 1 mm x 100 µm; buffer blank measurements were carried out under the same conditions, both before and after sample measurement. Data were averaged and subtracted using PRIMUS[60] from the ATSAS package[61], which was also used to calculate the pair-distance distribution function, as well as the radius of gyration and the Porod volume. Theoretical scattering curves for monomeric and trimeric FsxA$_E$ were calculated and compared with the experimental data using CRYSOL[62]. *Ab initio* envelope reconstruction was performed using the program DAMMIF[63], resulting in twenty models that were superimposed and averaged with DAMAVER[64]. Chain A of the refined FsxA$_E$ model was fitted into the SAXS reconstruction using UCSF ChimeraX[65] (**Extended Data Fig. 1d**).

Crystallization and X-ray diffraction data collection

Two similar initial hits obtained from extensive screening using a mosquito crystallization robot (TTP Labtech) were manually optimized by setting up vapour diffusion experiments at 20ºC in 24-well plates. To grow diffraction-quality crystals, 1 µl purified FsxA$_E$ was mixed with 1 µL 23% (w/v) PEG 4000, 0.1 M Tris-HCl pH 8.5, 0.2 M $CaCl_2$ and equilibrated against 1 mL of the same solution. Rhomboidal plates of FsxA$_E$ grew in 1-3 months from protein precipitate that appeared after overnight equilibration of the crystallization drops (**Extended Data Fig. 1e**). For data collection, specimens were freed from the precipitate by micromanipulation with MicroMounts (MiTeGen) and flash frozen in liquid nitrogen. More than a hundred crystals were screened at beamlines ID23-1 of the ESRF[66] and I04 of Diamond Light Source, yielding datasets of highly variable quality. The final X-ray diffraction dataset at 2.3 Å resolution was collected at ESRF ID23-1.

Data reduction and non-crystallographic symmetry analysis

Datasets were processed in space group *C*2 with XDS[67] (**Extended Data Table 2**). By revealing a strong non-origin peak at chi=120 (**Extended Data Fig. 1f**), self rotation functions calculated with MOLREP[68] or POLARRFN[69] clearly indicated the presence of three-fold non-crystallographic symmetry (NCS) within the asymmetric unit of the centred monoclinic crystals. Combined with Matthews coefficient calculations[70,71], this strongly suggested that FsxA$_E$ crystallized as a homotrimer.

Structure determination by molecular replacement with AlphaFold2 models

Because multiple attempts to experimentally determine the structure of FsxA$_E$ using a variety of heavy atoms failed, we took advantage of the recent significant advances in protein 3D structure prediction using machine learning[72] to phase the data by molecular replacement (MR)[73] (**Extended Data Fig. 2**). To do so, we used AlphaFold2[21] to generate five independent models of FsxA ectodomain residues D25-S535, with per-residue pseudo-B factors corresponding to 100-(per-residue confidence (pLDDT[21])). These models had relative root-mean-square deviations (RMSD) of 1.4-3.3 Å, or 0.7-1.9 Å after excluding 26 C-terminal residues predicted with low-confidence. Initial attempts to solve the structure with Phaser[74], using an ensemble including these models (further truncated to Q453, the predicted C-terminal end of domain III), yielded 4 solutions (with top Log Likelihood Gain (LLG) 188, final Translation Function Z score (TFZ) 9.6) that were retrospectively correct in terms of domain I/II placement, but completely wrong in the positioning of domain III. Because of the latter, automatic refinement of these solutions did not progress beyond R_{free} ~0.53. On the other hand, a parallel consecutive search for three copies of a domain I/II ensemble (D25-A335; RMSD 0.3-0.9 Å) followed by three copies of domain III (P350-Q453; RMSD 0.1-0.3 Å), using a model RMSD variance of 1 Å, yielded a clear single solution (LLG 876, TFZ 23.1) that could be automatically refined to initial R 0.45, R_{free} 0.46.

Remarkably, although a single copy of domain 3 corresponds to only 7% of the total scattering mass in the asymmetric unit of the FsxA$_E$ crystal, the very high accuracy of its AlphaFold2 model (reflected by *a posteriori*-calculated global RMSD and Distance Test Total Score (GDT_TS) of 0.7 Å and 97.6, respectively) allowed Phaser to also find a correct MR solution using just this part of the structure. Specifically, a consecutive search for three copies of the domain resulted in a trimeric model with LLG 275 and TFZ 15.1, which could be refined to starting R 0.51, R_{free} 0.51.

Also worth mentioning is the observation that the same domain I/II + domain III MR strategy used to phase the 2.3 Å resolution data could also be successfully applied to an initial dataset at much lower resolution (3.5 Å, with outer shell mean I/σI 0.6 and $CC_{1/2}$ 0.31); in this case, the Phaser LLG and TFZ values for the solution were 361 and 13.5, respectively, and initial automatic refinement of the corresponding model yielded R 0.44, R_{free} 0.48.

Model building, refinement and validation

The initial model of $FsxA_E$ was first automatically rebuilt using PHENIX AutoBuild[75] (1083 residues; R 0.34, R_{free} 0.38) and then significantly improved with the machine-learning-based sequence-docking method of ARP/wARP[76], as implemented in CCP4[69] (1390 residues; REFMAC[77] R 0.23). The resulting set of coordinates was subsequently subjected to alternating cycles of manual rebuilding with Coot[78]/ISOLDE[79] and refinement with phenix.refine[80], using torsion-based NCS restraints and three Translation-Libration-Screw-rotation groups per chain. Metal ions were assigned based on electron density level, difference Fourier maps generated using alternative atom types, correspondence with peaks in phased anomalous difference maps generated with PHENIX[81] or ANODE[82] and coordination properties[83]. Protein geometry was validated using MolProbity[84] (**Extended Data Table 2**).

Sequence-structure analysis

Transmembrane helices were predicted using TMHMM[85]. GDT_TS scores were calculated using LGA[86] and structural similarities were assessed with Dali[87] and PDBeFold[88]. Secondary structure was assigned using DSSP[89]. Subunit interfaces were analyzed using PDBsum[90], PIC[91] and PISA[92]. Molecular charge was calculated using the YASARA2 force field[93] and electrostatic surface potential calculations were performed with PDB2PQR[94] and APBS[95], via the APBS Tools plugin of PyMOL. Mapping of amino acid conservation onto the 3D structure of $FsxA_E$ was carried out by analyzing a sequence alignment of archaeal homologues with ConSurf[96]. Structural figures were generated with PyMOL (Schrödinger, LLC).

IME identification and analyses

K-mer (K = 4) spectrum for each genome was calculated for a sliding window of 1 kb using 500 bp steps and subtracted from the genomic average at each window position[97]. The absolute value of the difference between the genomic average and window spectra was represented graphically over the entire genome **(Supplementary Fig. 6, Fig. 4a)**. Gaussian mixture models using two distributions were fitted to the K-mer content of all windows to classify windows as belonging to either the core genome or transferred elements[97]. HMMER[52] and Pfam[98] were used to assign domains and their associated arCOG[99] identifiers to ORFs using default parameters (**Source Data Table 3**). Synteny conservation plots were made with MCscan tool[100] from the JCVI pipeline, creating relevant files by formatting data regarding the inferred homology relationship with homemade Python scripts. For details, see Supplementary Information.

Phylogenetic analyses

Maximum likelihood phylogenetic trees were generated with sequences aligned with MAFFT[101] (L-INS-i option) as input for IQ-TREE[102] and selecting the best evolutionary model with ModelFinder[103]. Homology trimeric models for FsxA$_E$ archaeal homologues (**Extended Data Fig. 8**) were built with MODELLER[104] using our crystal structure as template. Stem length and timing of acquisition of FsxA-HAP2 was done as described[105] in order to compare it with their data, generously made available[106]. The root for the structural tree was inferred using both midpoint-rooting and the Minimal Ancestor Deviation method[30].

Structural alignment and phylogeny

The overall assumption here is common knowledge, namely, that protein folds and their decorations evolve more slowly than sequences, hence preserving deep evolutionary signals[107,108]. FsxA models and crystal structures of FsxA$_E$ and eukaryotic and viral fusexins were used in all-vs-all comparisons with FATCAT[109] to establish structural distances between them and write PDB files for each superimposed pair. The following experimental crystal structures from other works were used: Flavivirus E: West Nile virus (2I69)[110]; Dengue virus serotype 1 (4GSX)[111]; Alphavirus E1: Semliki Forest virus (1RER)[112]; Chikungunya virus (3N43)[113]*; C. elegans* EFF-1 (4OJC)[12]; Bunyavirus Gc Rift Valley fever virus (6EGU)[114]; eukaryotic HAP2/GCS1 from *A. thaliana* (5OW3)[19] and *C. reinhardtii* (6E18)[47]. The text output of FATCAT was parsed and compiled into pairwise alignments, which were in turn merged iteratively with ClustalO[51] to generate a structure-based MSA. Essentially equivalent alignments can be obtained with online servers such as POSA[115] or mTM-align[116], that will derive TM-score matrices and multiple structure alignments. The PDB files produced by flexible alignment with FATCAT were compared with TMalign[117] to build a distance matrix filled with TM scores[118] between all structures. This distance matrix was the basis to compute a minimum evolution tree with FastME[119] on default parameters.

Cells and reagents

BHK-21 cells (kindly obtained from Judith White, University of Virginia) were maintained in DMEM supplemented with 10% FBS (Biological Industries), 100 U/ml penicillin, 100 μg/ml streptomycin (Biological Industries), 2 mM L-glutamine (Biological Industries), 1 mM sodium pyruvate (Gibco), and 30 mM HEPES buffer, pH 7.3, at 37°C with 5% CO_2. Transfections were performed using Fugene HD (Promega) or jetPRIME (Polyplus) according to the manufacturer's instructions.

Immunofluorescence

BHK cells were grown on 24-well tissue-culture plates with glass coverslips. Permeabilized cells were fixed with 4% paraformaldehyde (EM grade, Bar Naor, Israel) in PBS, followed by incubation in 40 mM NH_4Cl to block free aldehydes, permeabilized in 0.1% Triton X-100 in PBS and blocked in 1% FBS in PBS. After fixation, the coverslips were incubated 1 h with mouse anti–V5 antibody (Invitrogen, 1:500) and 1 h with the secondary antibody which was donkey anti–mouse coupled to Alexa Fluor 488 (Invitrogen, 1:500). Alternatively, for immunofluorescence without permeabilization, cells were blocked on ice in PBS with 1% FBS for 20 minutes, and then stained with Monoclonal ANTI-FLAG M2 antibody (Sigma, 1:1000) on ice for 1h. After anti-FLAG staining, cells were washed and fixed with 4% PFA in PBS. Cells were blocked again and stained with the secondary antibody (donkey anti–mouse coupled to Alexa Fluor 488; Invitrogen) diluted 1:500 in PBS for 1 h. In all cases, nuclei were stained with 1 μg/ml DAPI. Images were captured using a Nikon Eclipse E800 with a 60X/1.40 Plan Apochromat objective and an optical zoom lens (Nikon) using a Hamamatsu ORCA-ER camera controlled by Micro-Manager software[120] (**Extended Data Fig. 7d**).

Western blots

24 h post-transfection, cells were treated with Lysis Buffer (50 mM Tris-HCl pH 8.0, 100 mM NaCl, 5 mM EDTA, 1% Triton X-100 supplemented with chymostatin, leupeptin, antipain and pepstatin) on ice for 10 min. After 10 min centrifugation at 4 °C,14,000 rpm, supernatants of lysates were mixed with reducing sample buffer (+ DTT) and incubated 5 min at 95°C. Samples were loaded on a 10% SDS-PAGE gel and transferred to PVDF membrane. After blocking, membranes were incubated with primary antibody anti–V5 mouse monoclonal antibody (1:5,000; Invitrogen) or

anti-actin (1:2,000; MP Biomedicals) at 4 °C overnight and HRP-conjugated goat anti-mouse secondary antibody 1 h at room temperature. Membranes were imaged by the ECL detection system using FUSION-PULSE.6 (VILBER).

Content mixing assays with immunofluorescence

BHK-21 cells at 70% confluence were transfected (using JetPrime; Polyplus at a ratio of 1:2 DNA:transfection reagent) with 1 µg pCI::FsxA-V5::H2B-eGFP, pCI::FsxA-V5::H2B-RFP, pCI::AtHAP2-V5::H2B-eGFP, pCI::AtHAP2-V5::H2B-RFP, respectively. Control cells were co-transfected with pCI::H2B-eGFP and pRFPnes or pCI::H2B-RFP and pRFPnes. 4 h after transfection, the cells were washed 4 times with DMEM with 10% serum (Invitrogen), 4 times with PBS and detached using Trypsin (Biological Industries). The transfected cells were collected in Eppendorf tubes, resuspended in DMEM with 10% serum, and counted. Equal amounts of H2B-RFP and H2B-eGFP cells were mixed and seeded on glass-bottom plates (12-well black, glass-bottom #1.5H; Cellvis) and incubated at 37°C and 5% CO_2. 18 h after mixing, 20 µM 5-fluoro-2'-deoxyuridine (FdUrd) was added to the plates to arrest the cell cycle and 24 h later, the cells were fixed with 4% PFA in PBS and processed for immunofluorescence. To assay mixed cells and detect the transfected proteins (FsxA-V5 or AtHAP2-V5), we stained cells with anti-V5 mAb (Life Science). The secondary antibody was Alexa Fluor 488 goat anti-mouse, with 1 µg/ml DAPI[3]. Micrographs were obtained using wide-field laser illumination using an ELYRA system S.1 microscope (Plan-Apochromat 20X NA 0.8; Zeiss). The GFP + RFP mixing index was calculated as the number of Red and Green nuclei in mixed cells out of the total number of nuclei of fluorescent (green cytoplasm) cells in contact (**Fig. 3b**).

Cell fusion assay by content mixing with nuclear and cytoplasmic markers.

For the unilateral setup, BHK-21 cells were transfected (as explained above) with 1 µg pCI::H2B-RFP; pCI::GFPnes; pCI::FsxA-V5::GFPnes; 0.25 µg pCI::EFF-1-V5::GFPnes; 1 µg pCI::VSV-G::GFPnes in respective 35mm plates. The cells were incubated, washed, and mixed with pCI::H2B-RFP (empty vector) transfected cells. For evaluating the mutants, BHK-21 cells were transfected with 1 µg pCI::FsxA-V5::GFPnes or pCI::FsxA-V5::H2B-RFP or the plasmids encoding for each mutant: ΔFL→AG$_4$A or ΔDIV→EFF-1 stem (**Fig. 3i**). Empty pCI::GFPnes or pCI::H2B-RFP were used as negative controls. 4 h after transfection, the cells were washed, counted, mixed, and incubated as previously described. In all cases, 18 h after mixing, 20 µM FdUrd was added to the plates, and 24 h later, the cells were fixed with 4% paraformaldehyde diluted in PBS. Nuclei were stained with 1 µg/ml DAPI. Micrographs were obtained using wide-field laser illumination using an ELYRA system S.1 microscope as described above. The GFP + RFP mixing index was calculated as the number of nuclei in mixed cells, green cytoplasm (GFPnes) with red (H2B-RFP) and blue (DAPI) nuclei out of the total number of nuclei in fluorescent cells in contact (**Fig. 3e** and **3j**). For the unilateral assay, multinucleation was determined as the ratio between the number of nuclei in multinucleated green cells and the total number of nuclei in green multinucleated cells and GFPnes expressing cells that were in contact but did not fuse.

Live imaging of fusing cells

BHK cells were plated on 15 mm glass bottom plates (Wuxi NEST Biotechnology Co., Ltd.) and transfected with 1 µg pCI::FsxA-V5::H2B-GFP together with 0.5 µg myristoylated-mCherry (myr-palm-mCherry; kindly provided by Valentin Dunsing and Salvatore Chiantia [56]). 18 h after transfection, the cells were incubated with 2 µg/ml Hoechst dye for 10 min at 37°C and washed once with fresh medium. Time-lapse microscopy to identify fusing cells was performed using a spinning disc confocal microscope (CSU-X; Yokogawa Electric Corporation) with an Eclipse Ti and a Plan-Apochromat 20X (NA, 0.75; Nikon) objective. Images in differential interference contrast and red channels were recorded every 4 min in different positions of the plate using high gain and minimum laser exposure. Time lapse images were captured with an iXon 3 EMCCD camera (Andor Technology). After 5 h, confocal z-series, including detection of the DAPI channel, were obtained to confirm the formation of multinucleated cells. Image analyses were performed in MetaMorph (Molecular Devices) and ImageJ[121] (National Institutes of Health).

Surface biotinylation

Proteins localizing on the surface were detected as previously described[3]. Briefly, BHK cells were transfected with 1 µg pCAGGS, pCAGGS::EFF-1-V5, pCAGGS::FsxA-V5, pCAGGS::ΔFL→AG$_4$A-V5 or pCAGGS::ΔDIV→EFF-1 stem-V5. 24 h later, cells were washed twice with ice-cold PBS^{2+} (with Ca^{2+} and Mg^{2+}) and incubated with 0.5 mg/ml EZ-Link Sulfo NHS-Biotin (Thermo Fisher Scientific) for 30 min on ice. The cells were washed four times with ice-cold PBS^{2+}, once with DMEM with 10% FBS (to quench residual biotin), followed by two more washes with PBS^{2+}. To each plate 300 µl of Lysis Buffer supplemented with 10 mM iodoacetamide were added and the cells detached using a scrapper. The insoluble debris was separated by centrifugation (10 min at 21,000 *g*), and the lysate was mixed with NeutrAvidin Agarose Resin (Thermo Fisher Scientific) and 0.3% SDS. After an incubation of 12 h at 4°C the resin was separated by centrifugation (2 min at 21,000 *g*), washed three times with lysis buffer and then mixed with SDS-PAGE loading solution with freshly added 5% b-mercaptoethanol and incubated 5 min at 100°C. After pelleting by centrifugation, the samples were separated by SDS-PAGE gel and analyzed by Western blotting as described above using anti–V5 mouse monoclonal antibody. Loading was controlled using anti-actin C4 monoclonal (1:2,000; MP Biomedicals).

Data analysis

Counting of content mixing and multinucleation was made blind for the experiments included in **Fig. 3f** and **3j**. Interobserver error was estimated for counting of multinucleated cells, cells in contact, and content-mixing experiments: the differences in percentages of multinucleation and content mixing obtained by two observers was <10%. Figures were prepared with Photoshop and Illustrator CS (Adobe), BioRender and ImageJ[121].

Statistical tests

Results are presented as means ± SEM. For each experiment we performed at least three independent biological repetitions. To evaluate the significance of differences between the averages we used one-way ANOVA as described in the legends (GraphPad Prism).

Data availability

All relevant data are included as Supplementary Information files (see Suppl. Inf. Guide). Crystallographic structure factors and atomic coordinates have been deposited in the Protein Data Bank under accession code 7P4L.

Code availability

All relevant codes, notebooks and datasets necessary for: HHblits and Hmmer searches and comparisons (Fig1a, Suppl. Fig1, Source Data Table 1); Kmer spectra analyses (Fig. 4a, Suppl Fig. 6); IMEs clustering, content and synteny analyses (Source Data Tables 2 and 3, Fig. 4b, Suppl. Figs.8 and 9), protein sequence and structure-based comparisons (Extended Data Fig. 8, Fig. 4c) and timing of gene acquisition analysis (Suppl. Fig. 10) are available upon request. GitHub repository URL will be provided upon publication.

Acknowledgements

We thank Sonja-Verena Albers, Dan Cassel, Peter Walter, Alejandro Colman-Lerner, Uri Gophna, Yael Iosilevskii, Shahar Lavid, and members of our laboratories for discussion and comments on the manuscript; Kira Makarova for discussions on archaea; Jose Flores for advice on searches of fusexins in metagenomes. This work was supported by grants from Comisión Sectorial de Investigación Científica, Uruguay (CSIC I+D-2020-682 to H.R. and M.G.); Fondo para la Investigación Científica y Tecnológica, Argentina (PICT-2017-0854 to P.S.A.); Beca de Doctorado Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina (to D.M.); the Knut and Alice Wallenberg Foundation (project grant 2018.0042 to L.J.); the Novartis Foundation for Medical-Biological Research (grant #17B111 to C.D.); the Swedish Research Council (project grants 2016-03999 and 2020-04936 to L.J.); the Swiss Leading House for the Latin American Region (C.D. and P.S.A.); the Swiss National Science Foundation (grant 183723 to C.D.); the Israel Science Foundation (grants 257/17, 2462/18, 2327/19, and 178/20 to B.P.); FOCEM, Fondo para la Convergencia Estructural del Mercosur (COF 03/11, M.G.). H.R., M.G. and M.L. also thank Programa de Desarrollo de las Ciencias Básicas and ANII, Uruguay. The computations for this work were performed at the Vital-IT Center for high-performance computing of the SIB Swiss Institute of Bioinformatics.

Author contributions

B.P. conceived the experiments; performed some imaging work; designed, supervised and analyzed cell fusion experiments. C.D. helped devise analysis strategies for k-mer and phylogenetic analysis. C.V. designed, performed and analyzed cell fusion assays. D.deS. collected X-ray data, took part in structural analysis and validated metal substructure. D.M. carried out deep homology detection of FsxA; designed gene content analysis strategies and phylogenetic analysis, collected sequence data; performed k-mer, functional, structural and phylogenomics analyses, built homology models, coded analysis routine pipelines. H.R. supervised the bioinformatics part of the work, estimated relative acquisition times, designed and performed phylogenetic and phylogenomic surveys. K.F. made constructs of FsxA mutants. K.T. and J.J. generated AlphaFold2 models of FsxA. L.J. supervised the biochemical and structural part of the work; collected X-ray data; solved the FsxA structure, refined and analyzed it. M.G. designed bioinformatic strategies, supervised bioinformatic aspects of the work, analyzed sequence and structural data. M.L. performed IME synteny analyses, phylogenetic and phylogenomic surveys. N.G.B. carried out live imaging and surface biotinylation experiments, assisted with the preparation and design of plasmids; analyzed data. P.S.A. supervised the bioinformatics part of the work, analyzed data. S.N. expressed, purified and crystallized FsxA; performed SEC-MALS experiments; analyzed SAXS data; collected X-ray data and took part in structure determination, model building and structure analysis. X.L. carried out immunofluorescence and western blots for archaeal fusexins in mammalian cells, designed and constructed plasmids, performed imaging work. D.M., S.N., X.L., N.G.B, M.G., H.R., P.S.A., L.J. and B.P. made figures and tables. D.M., S.N., X.L., M.G., H.R., P.S.A., L.J. and B.P. wrote the manuscript. All authors reviewed the manuscript.

Correspondence and requests for materials should be addressed to M.G. or H.R. or P.S.A. or L.J. or B.P.

Competing interests

J.J. has filed provisional patent applications relating to machine learning for predicting protein structures. The other authors declare no competing interests.

Extended data

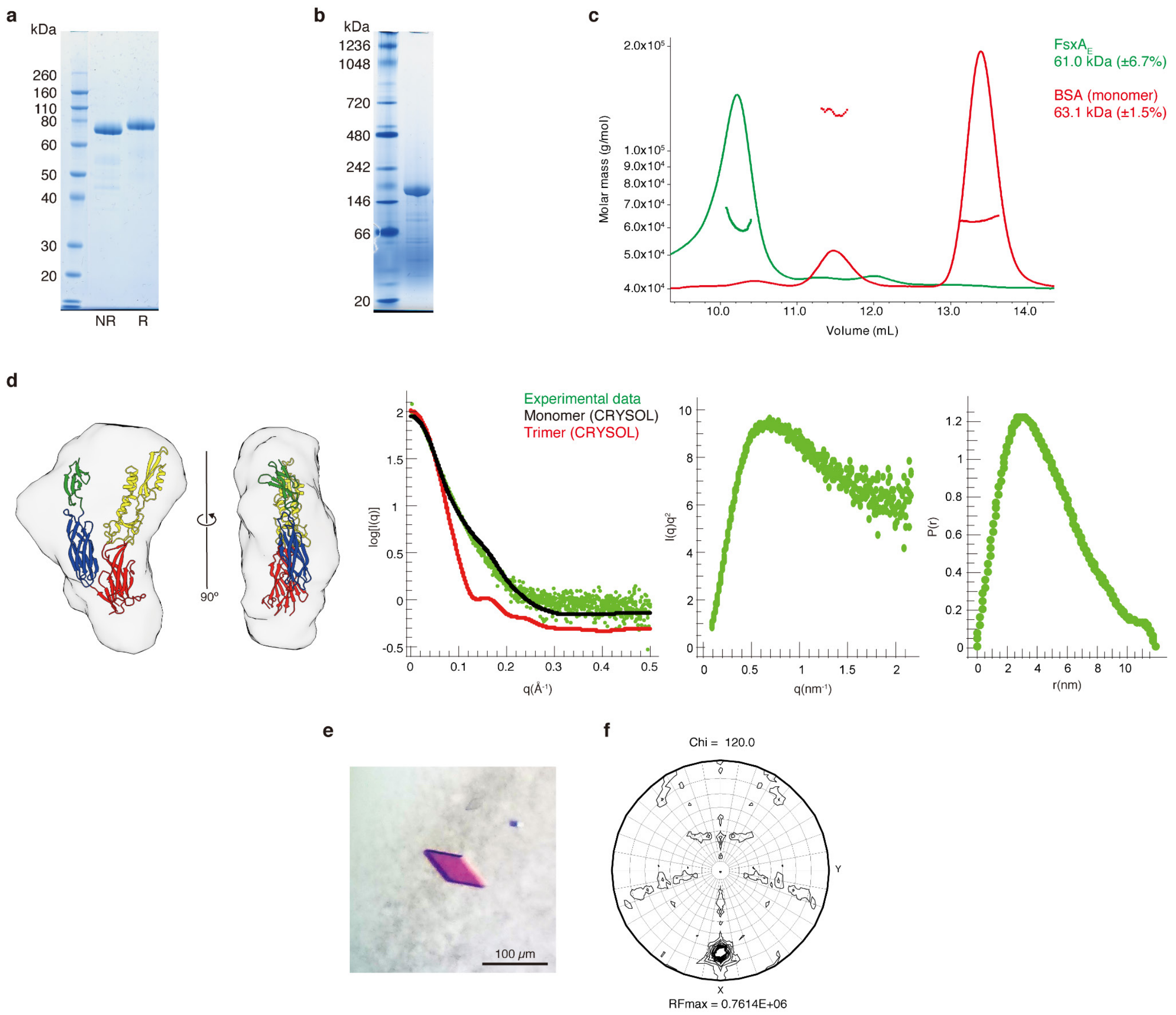
Extended Data Table 1 | FsxAs in genomes with assigned taxonomy

FsxAs in Pure Culture Genomes (PCGs)

Sequence ID	Species	Taxonomy
WP_058826362.1	*Haloferax* sp. Q22	Archaea › Euryarchaeota › Stenosarchaea group › Halobacteria › Haloferacales › Haloferacaceae
WP_144240185.1	*Natrinema altunense* (AJ2)	Archaea › Euryarchaeota › Stenosarchaea group › Halobacteria › Natrialbales › Natrialbaceae
ELY83688.1	*Natrinema altunense* JCM12890)	Archaea › Euryarchaeota › Stenosarchaea group › Halobacteria › Natrialbales › Natrialbaceae
WP_157573584.1	*Haloplanus natans* DSM 17983	Archaea › Euryarchaeota › Stenosarchaea group › Halobacteria › Haloferacales › Haloferacaceae
WP_174701778.1	*Haloterrigena* sp. SYSU A121-1	Archaea › Euryarchaeota › Stenosarchaea group › Halobacteria › Natrialbales › Natrialbaceae
WP_179268568.1	*Halobonum* sp. NJ-3-1	Archaea › Euryarchaeota › Stenosarchaea group › Halobacteria › Haloferacales › Halorubraceae
WP_163487151.1	*Halogeometricum borinquense* strain wsp4	Archaea › Euryarchaeota › Stenosarchaea group › Halobacteria › Haloferacales › Haloferacaceae
WP_207587115.1	*Halovivax* sp. KZCA124	Archaea › Euryarchaeota › Stenosarchaea group › Halobacteria › Natrialbales › Natrialbaceae

FsxAs in Metagenomics-Assembled Genomes (MAGs)

Sequence ID	Assigned taxon	Taxonomy
MGYP000598426430	Halobacteriales	Archaea › Euryarchaeota › Stenosarchaea group › Halobacteria › Halobacteriales
LKMP01000007_1	Nanohaloarchaea archaeon B1-Br10_U2g21	Archaea › Euryarchaeota › Stenosarchaea group › Candidatus Nanohaloarchaeota
RLG58774.1	Candidatus Geothermarchaeota B85_G16	Archaea ›TACK group › Candidatus Geothermarchaeota
RLI53188.1	Candidatus Thorarchaeota archaeon	Archaea › Asgard group › Candidatus Thorarchaeota
RKX41251.1	Thermotogae bacterium	Bacteria › Thermotogae
RKZ11204.1	Candidatus Fermentibacteria bacterium	Bacteria › Candidatus Fermentibacteria
RLG94066.1	Candidatus Bathyarchaeota archaeon	Archaea ›TACK group › Candidatus Bathyarchaeota
AJF63093.1	archaeon GW2011_AR20	Archaea ›unclassified
HEX32987.1	Candidatus Aenigmarchaeota archaeon	Archaea › DPANN group › Candidatus Aenigmarchaeota
HDD44259.1	Candidatus Desulfofervidus auxilii	Bacteria › Proteobacteria › Deltaproteobacteria › Candidatus Desulfofervidaceae
HDI72891.1	Candidatus Altiarchaeales archaeon	Archaea › DPANN group › Candidatus Altiarchaeota › Candidatus Altiarchaeales
HHR27186.1	Candidatus Bathyarchaeota archaeon	Archaea ›TACK group › Candidatus Bathyarchaeota
NJD53946.1	Candidatus Methanoperedens sp.	Archaea › Euryarchaeota › Stenosarchaea group › Methanomicrobia › Methanosarcinales › Cand. Methanoperedenaceae
NOZ47386.1	Chlorobi bacterium	Bacteria › Chlorobi
HGF63239.1	Candidatus Micrarchaeota archaeon	Archaea › DPANN group › Candidatus Micrarchaeota
HID09282.1	Candidatus Micrarchaeota archaeon	Archaea › DPANN group › Candidatus Micrarchaeota



Extended Data Fig. 1 | The FsxA ectodomain is a monomer in solution but crystallizes as a trimer.

a, b, SDS-PAGE (a) and blue native PAGE (b) gels of purified FsxA ectodomain (FsxA$_E$). NR; non-reducing conditions. R; reducing conditions (see **Supplementary Fig. 11**).

c, Size exclusion chromatography-multiangle light scattering (SEC-MALS) shows that FsxA$_E$ is a monomer in solution. BSA is used as a control.

d, SAXS analysis of FsxA$_E$. Left panel, The SAXS envelope of FsxA$_E$, obtained by averaging of *ab initio* shape reconstructions, is consistent with the crystallographic model of FsxA$_E$ chain A. Centre-left panel, Comparison of the experimental SAXS profile of FsxA$_E$ (green dots) and theoretical scattering curves calculated from the refined coordinates of FsxA$_E$ chain A (black dots) or the whole FsxA$_E$ trimer (red dots). Centre-right panel, the Kratky plot of FsxA$_E$ suggests the presence of significant flexibility between the domains of the monomeric protein. Right panel, Pairwise interatomic distance distribution of FsxA$_E$.

e, Representative rhomboidal plate crystal of FsxA$_E$.

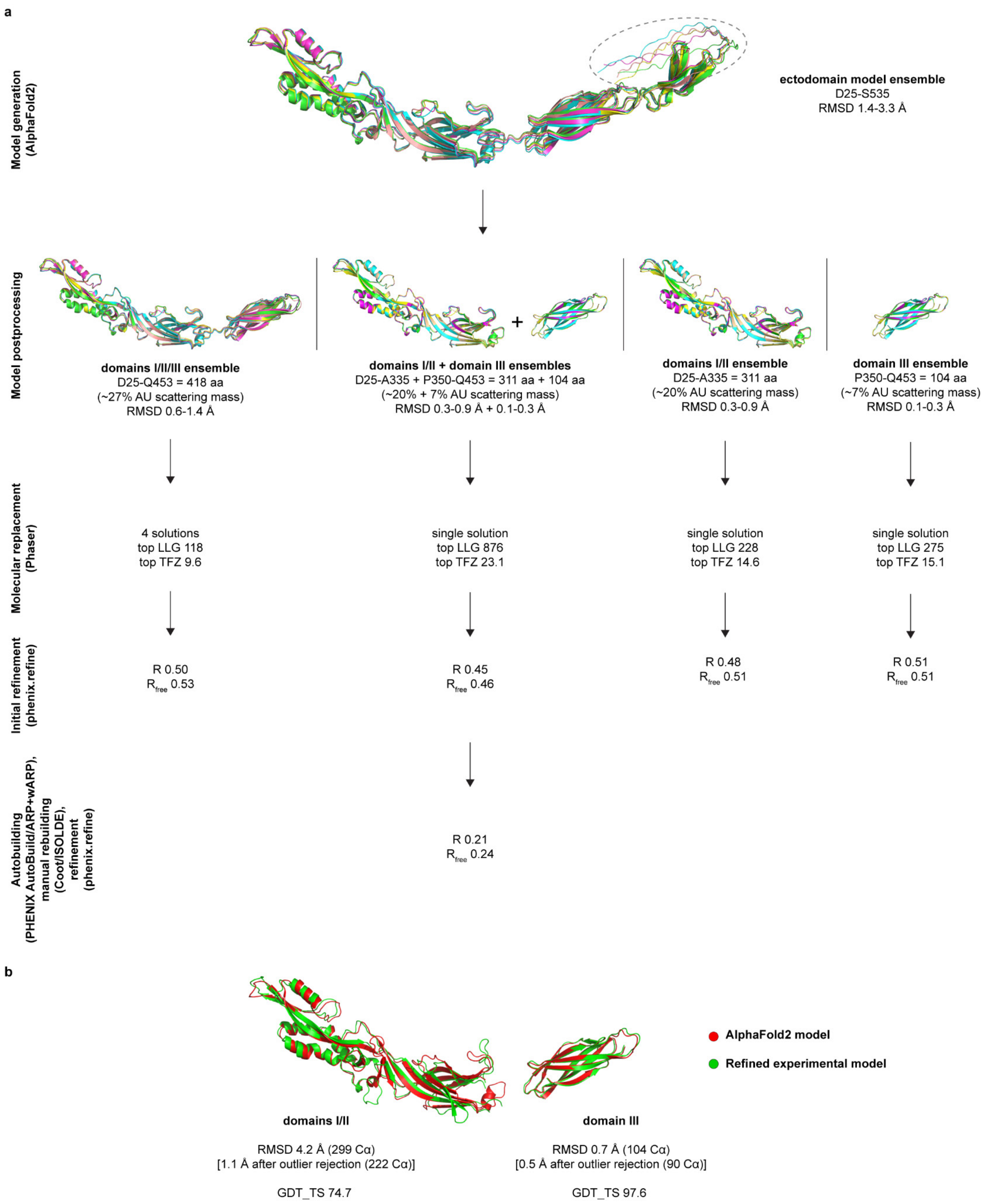
f, The Chi=120 section of the self-rotation function of FsxA$_E$ (calculated using a 67.3-2.6 Å resolution range) shows a prominent peak with a height of 72% of the origin peak.

Extended Data Table 2 | Data collection and refinement statistics

Data collection	
Space group	*C*2 (5)
Cell dimensions	
a, *b*, *c* (Å)	262.51, 111.33, 68.51
α, *β*, *γ* (°)	90, 100.709, 90
Wavelength (Å)	1.005
Resolution range (Å)	67.3-2.3 (2.38-2.30)*
Unique reflections	85618 (8536)
Multiplicity	4.2 (4.4)
Completeness (%)	99.5 (99.9)
Mean I / σI	11.4 (1.4)
Wilson B-factor	42.8
R_{merge}	0.100 (1.218)
R_{meas}	0.115 (1.385)
R_{pim}	0.055 (0.651)
$CC_{1/2}$	1.0 (0.58)
CC*	1.0 (0.86)
Refinement	
Reflections used in refinement	85574 (6092)**
Reflections used for R_{free}	2012 (149)
R_{work}	0.199 (0.303)
R_{free}	0.243 (0.377)
Number of non-H atoms	11733
macromolecules / ligand / solvent	11051 / 66 / 616
Protein residues	1432
RMS	
bonds (Å)	0.004
angles (°)	0.58
Ramachandran favoured / allowed / outliers (%)	98.9 / 1.1 / 0.0
Rotamer outliers (%)	0.2
Clashscore	2.5
Average B-factor	53.4
macromolecules / ligand / solvent	53.5 / 58.8 / 50.0

* Values in parenthesis are for the highest resolution shell

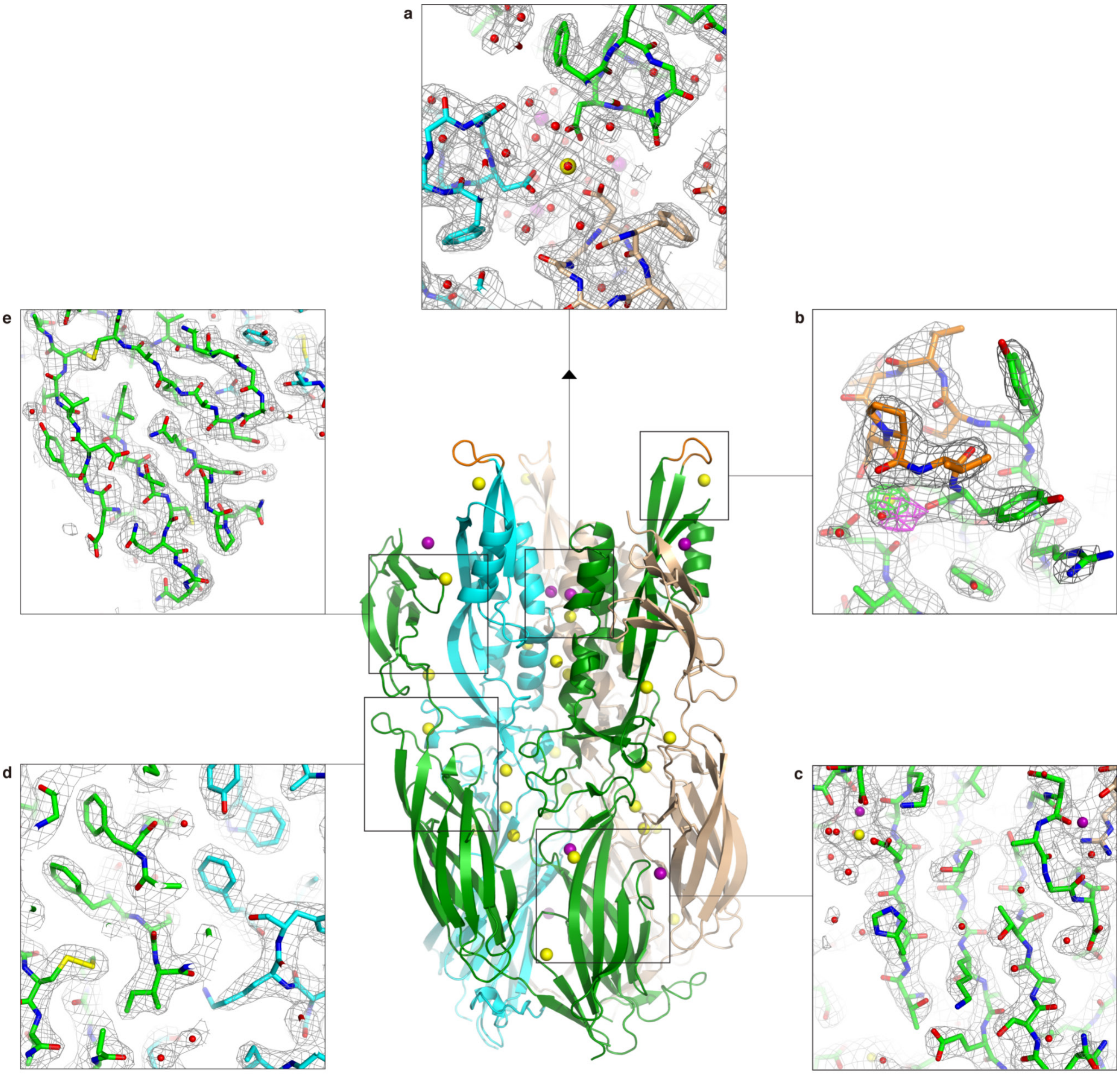
** The highest resolution shell used in refinement included reflections between 2.36 and 2.30 Å



Extended Data Fig. 2 | AlphaFold2-aided MR phasing of FsxA.

a, Flowchart of FsxA structure determination using different AlphaFold2 model fragments or a combination thereof. The dashed oval in the top panel indicates C-terminal residues D510-S535, predicted with low-confidence by AlphaFold2. aa, amino acid; AU, asymmetric unit.

b, Comparison of the AlphaFold2 and final crystallographic models of FsxA domains I/II and III.



Extended Data Fig. 3 | Details of the electron density map of FsxA.

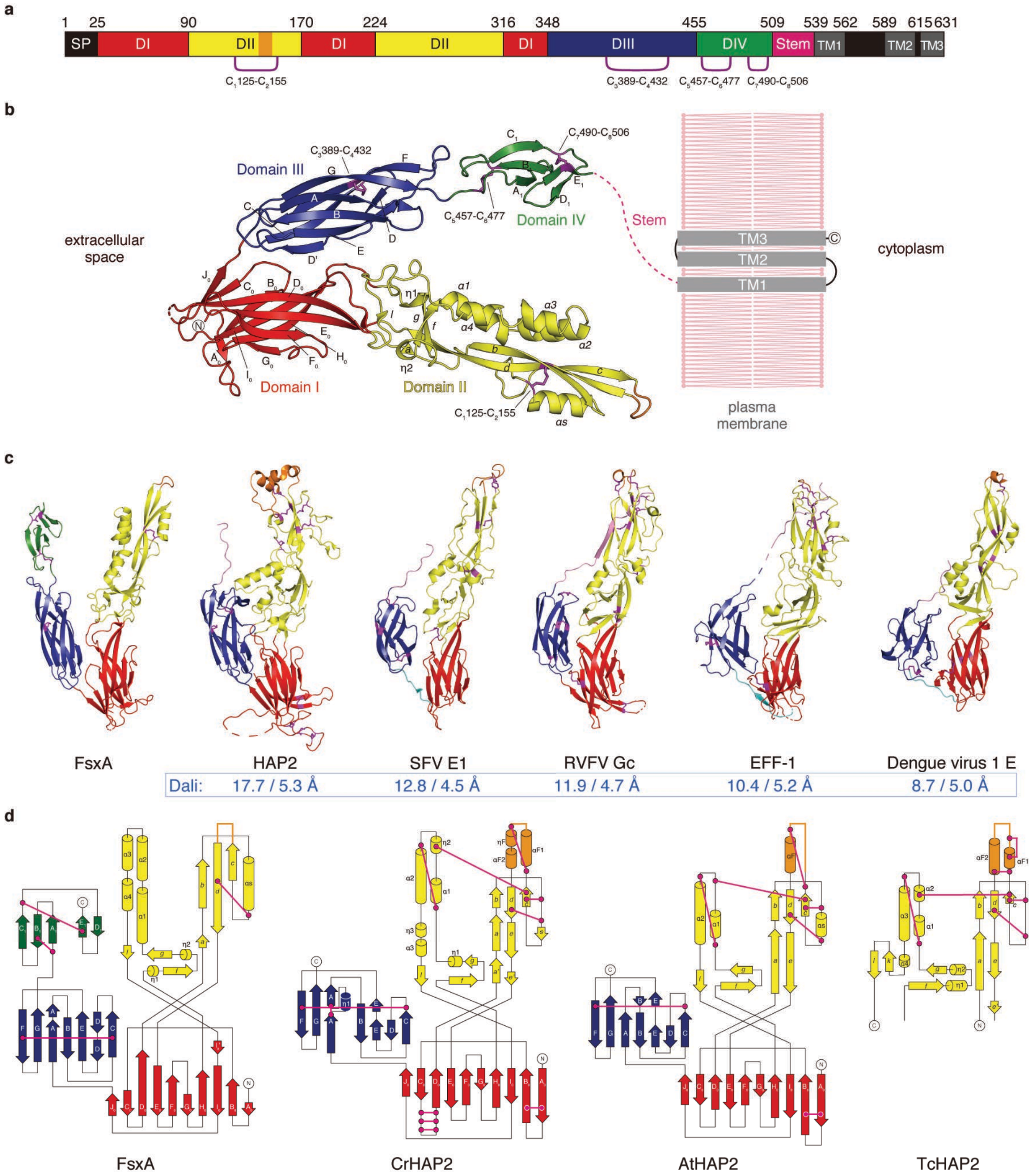
a, View of the FsxA domain II helical bundle, looking down the molecular three-fold axis from the centre of the structure towards the putative fusion loop end. Domain II α1 helix residues D260 and D263 coordinate a Ca^{2+} ion sitting on the NCS axis and three symmetrically positioned Na^{+} ions, respectively. The refined *2mFo-DFc* electron density map, contoured at 1.0 σ, is shown as a grey mesh superimposed onto the protein model in stick representation. FsxA subunits and metal ions are coloured as in Fig. 2c.

b, The putative fusion loop of FsxA adopts a highly ordered conformation stabilized by a Ca^{2+} ion. Presence and identity of the latter are supported by two other maps shown in addition to the *2mFo-DFc* map: a difference map calculated upon omitting all metal ions from the model (thick green(+)/red(-) mesh, contoured at 6.0 σ) and a phased anomalous difference map calculated from a 2.9 Å-resolution dataset collected at 7.1 KeV (thick magenta mesh, contoured at 3.2 σ).

c, Section of the map centred around β-strands D_0, E_0 and F_0 of domain I.

d, Closeup of the map region where domain III of one FsxA subunit interacts with domains I and II of another. Clear density for the $C_3$389-$C_4$432 disulfide is visible near the bottom left corner.

e, Map of FsxA domain IV. The conserved $C_7$490-$C_8$506 disulfide is at the top left corner, whereas $C_6$477 of the $C_5$457-$C_6$477 disulfide is visible at the bottom and the domain II $C_1$125-$C_2$155 disulfide of the adjacent subunit can be seen on the top right corner.



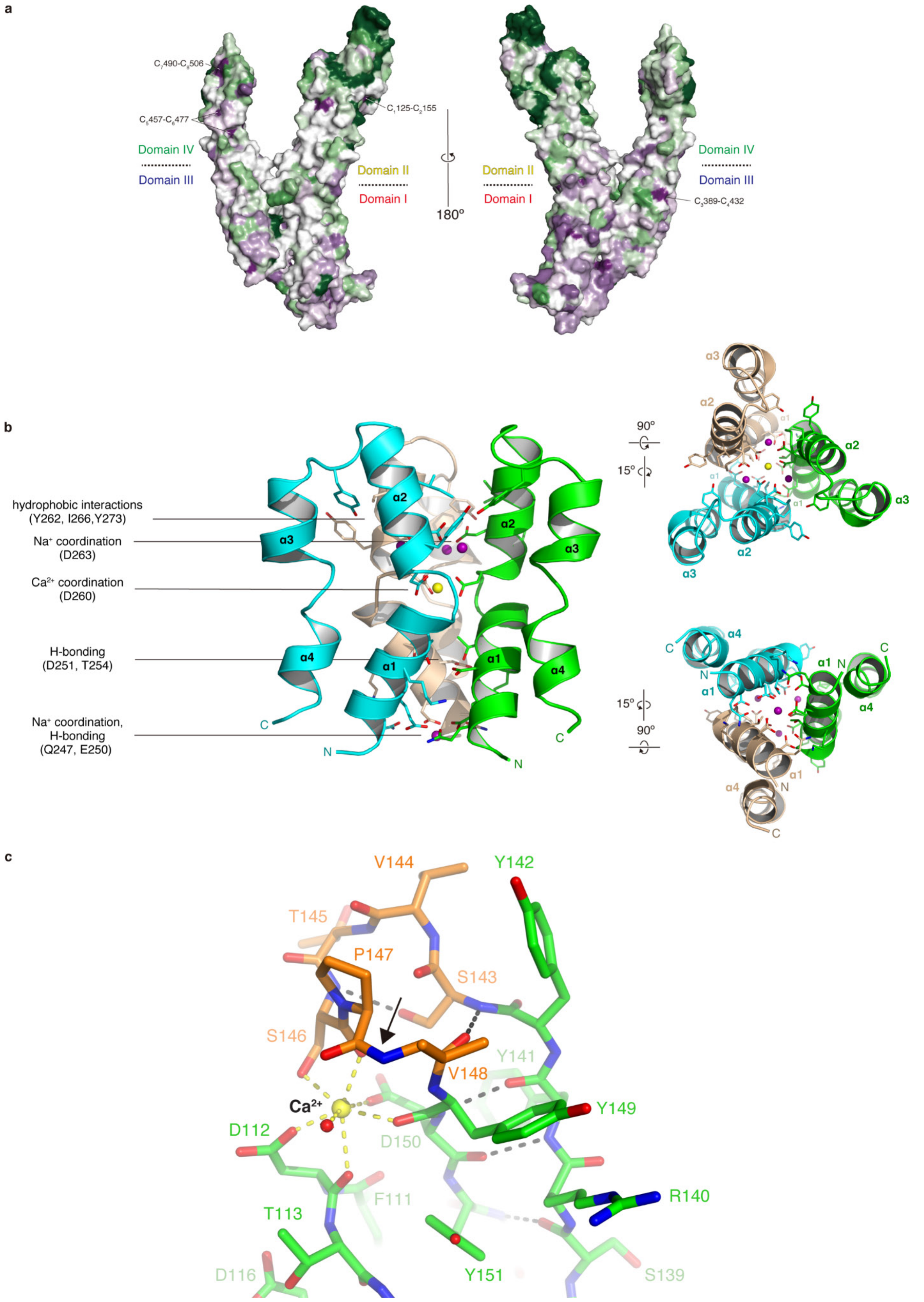
Extended Data Fig. 4 | Structural comparison of FsxA with class II fusogens.

a, Schematic diagram of the domains of FsxA. SP, signal peptide; TM, predicted transmembrane helices.

b, Crystal structure of the ectodomain of FsxA and predicted topology of the full-length protein relative to the plasma membrane. Domains I, II, III and IV are shown in red, yellow, blue and green, respectively; disulfide bonds are indicated and coloured magenta.

c, Side by side comparison of FsxA and known type II fusogens. Domains and disulfide bonds are coloured as in panel b; the stem region and the linker between domains I and III are shown in pink and cyan, respectively. *C. reinhardtii* HAP2 (CrHAP2; PDB 6E18[47]), Semliki Forest Virus glycoprotein E1 (SFV E1; PDB 1RER[112]), Rift Valley Fever Virus Glycoprotein c (RVFV Gc; PDB 6EGU[114]), *C. elegans* EFF-1 (CeEFF-1; PDB 4OJC[12]) and Dengue virus 1 E (PDB 4GSX[111]) are shown, with Dali Z-scores/RMSD values (blue) indicating the respective structural similarity to FsxA.

d, Topology diagrams of FsxA, *C. reinhardtii* HAP2 (PDB 6E18[47]), *A. thaliana* HAP2 (PDB 5OW3[19]) and *T. cruzi* HAP2 (PDB 5OW4[19]). Domains and disulfide bonds are coloured as in panel b. Note how, although domain II of FsxA has the same topology as the corresponding domain of HAP2, it contains only one of its conserved disulfide bonds ($C_1$125-$C_2$155, corresponding to disulfide bond 3 of CrHAP2[4] and disulfide bond 5 of CeEFF-1[12]).

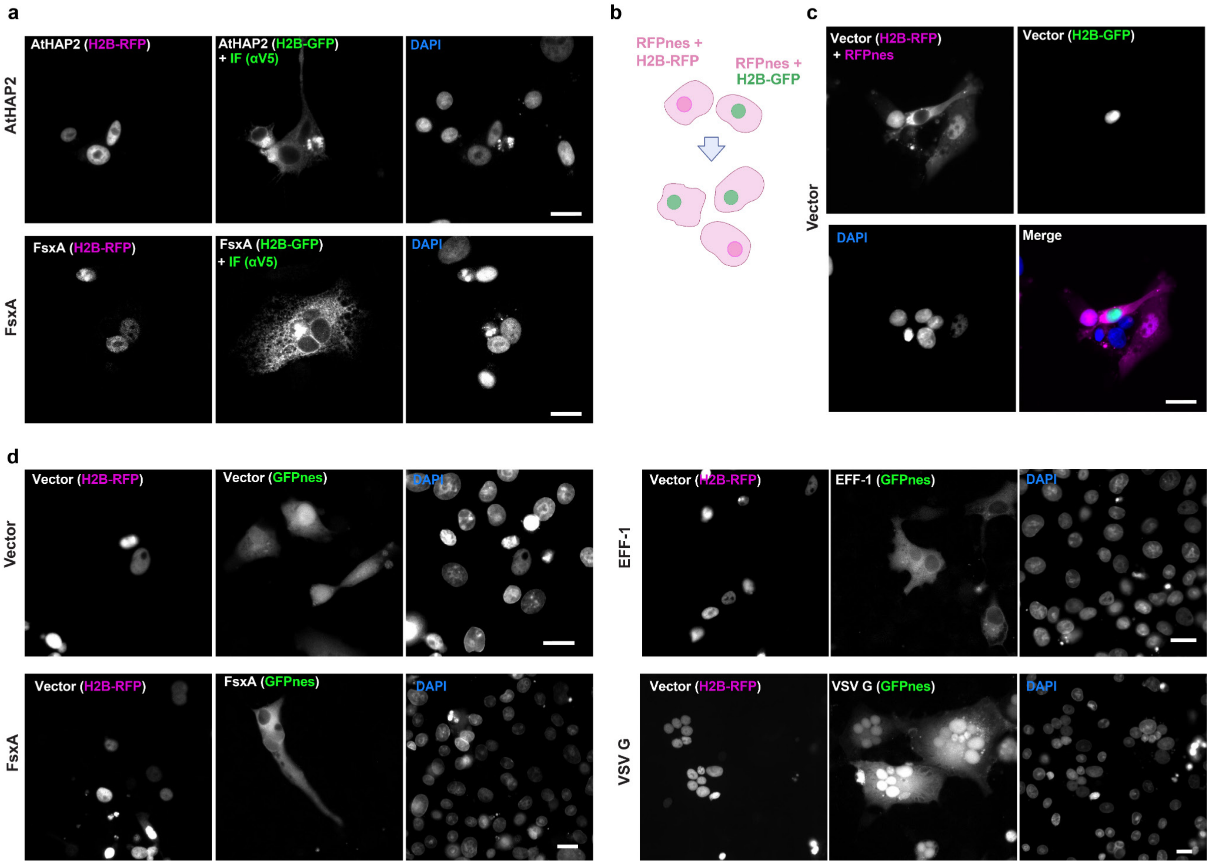

a
$C_7$490-$C_8$506
$C_5$457-$C_6$477
$C_1$125-$C_2$155
Domain IV
Domain III
Domain II
Domain I
180°
Domain II
Domain I
Domain IV
Domain III
$C_3$389-$C_4$432
b
hydrophobic interactions (Y262, I266,Y273)
Na⁺ coordination (D263)
Ca²⁺ coordination (D260)
H-bonding (D251, T254)
Na⁺ coordination, H-bonding (Q247, E250)
90°
15°
15°
90°
c
V144
Y142
T145
P147
S143
S146
Y141
V148
Ca²⁺
Y149
D112
D150
F111
R140
T113
D116
Y151
S139

Extended Data Fig. 5 | 3D mapping of the evolutionary conservation of FsxA residues and key structural features of its domain II.

a, FsxA domains I and III are more evolutionarily conserved than domains II and IV. Surface representation of the FsxA monomer, with residues coloured from green to violet by increasing conservation among archaeal homologues. Approximate domain boundaries are marked, and the position of the four highly conserved disulfides of FsxA is indicated.

b, Helices α1 and α2 of FsxA domain II form a six-helix bundle around the molecular three-fold axis. Top, side and bottom view of the helical region of domain II. Subunits are shown in cartoon representation, with residues mediating direct or ion-mediated interactions between chains depicted in stick representation (for clarity, water-mediated interactions are not shown) and elements coloured as in **Fig. 2c**.

c, Structure of the fusion loop of FsxA. The domain II region encompassing the loop is shown in the same orientation as in **Extended Data Fig. 3b**, with black and yellow dashes indicating protein hydrogen bonds and the coordination of the Ca^{2+} ion, respectively. Note how binding of the ion locally twists the protein main chain, with the peptide bond between P147 and V148 adopting a *cis* configuration (black arrow).



Extended Data Fig. 6 | FsxA mediates bilateral cell-cell fusion.

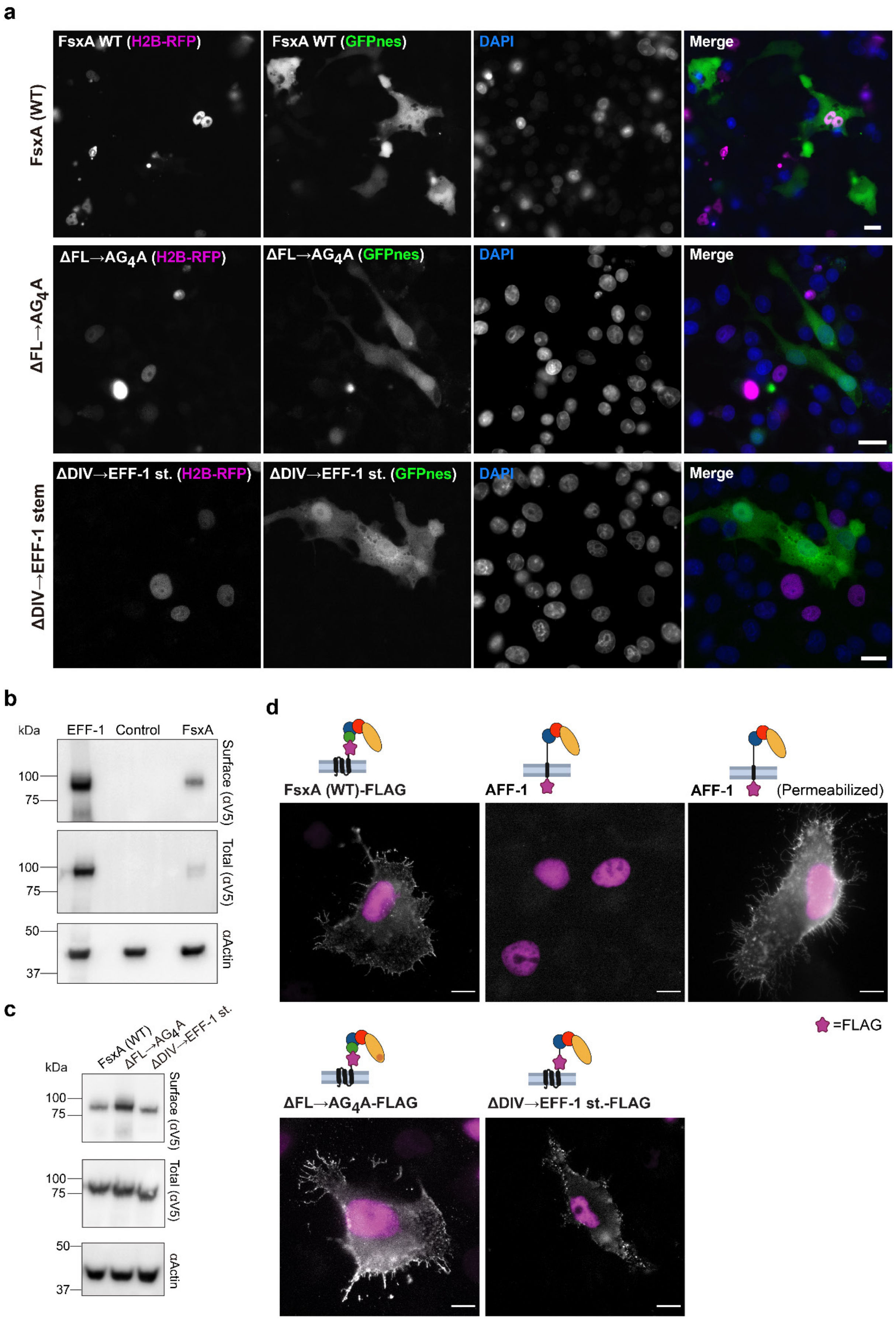
a, Images from **Fig. 3a** in each separate channel (red, green and DAPI).

b-c, In the negative control, cell-cell fusion was measured by content-mixing, indicated by the appearance of multinucleated cells containing green nuclei (H2B-GFP) and magenta nuclei (H2B-RFP). To reveal the cytoplasm of the transfected cells, a plasmid encoding for cytoplasmic RFP (RFPnes) was co-transfected.

b, Cartoon showing the experimental design for negative control.

c, Representative images of mononucleated cells with a green or red nucleus. DAPI staining is shown in blue.

d, Images from **Fig. 3d** in each separate channel (red, green and DAPI). Scale bars, 20 µm.


a
FsxA (WT)
FsxA WT (H2B-RFP)
FsxA WT (GFPnes)
DAPI
Merge
ΔFL→AG4A
ΔFL→AG4A (H2B-RFP)
ΔFL→AG4A (GFPnes)
DAPI
Merge
ΔDIV→EFF-1 stem
ΔDIV→EFF-1 st. (H2B-RFP)
ΔDIV→EFF-1 st. (GFPnes)
DAPI
Merge
b
kDa
EFF-1
Control
FsxA
100
75
Surface (αV5)
100
75
Total (αV5)
50
37
αActin
c
kDa
FsxA (WT)
ΔFL→AG4A
ΔDIV→EFF-1 st.
100
75
Surface (αV5)
100
75
Total (αV5)
50
37
αActin
d
FsxA (WT)-FLAG
AFF-1
AFF-1 (Permeabilized)
=FLAG
ΔFL→AG4A-FLAG
ΔDIV→EFF-1 st.-FLAG

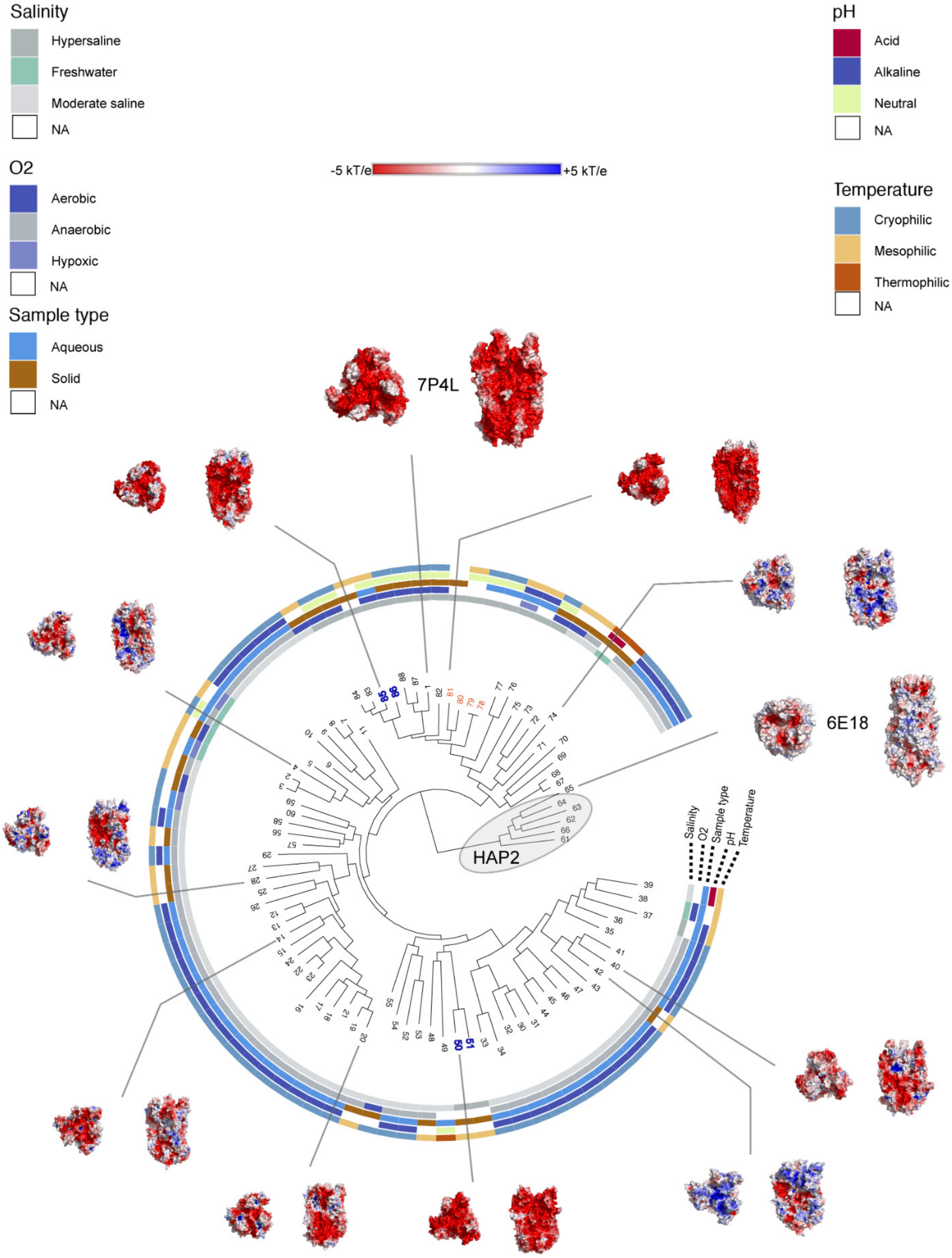
Extended Data Fig. 7 | Structure-function analysis of FsxA.

a, Images from **Fig. 3j** in each separate channel: red (RFP); green (GFP) and blue (DAPI) and the merged images. Scale bars, 20 μm.

b, Immunoblot of EFF-1-V5, control (untransfected cells) and FsxA-V5 expressing cells. “Surface” indicates surface biotinylation followed by affinity purification using neutravidin agarose beads; “Total” indicates the expression in whole cell extracts. Actin is used as a loading control. The amount of initial cells for FsxA is 4 times higher than EFF-1 (**Supplementary Fig. 12b**).

c, Surface biotinylation as explained in “**b”** for cells expressing FsxA-V5 (WT), ΔFL→AG$_4$A-V5 or ΔDIV→EFF-1 stem-V5 (**Fig. 3i; Supplementary Fig. 12c**).

d, Images from immunofluorescence on non-permeabilized cells expressing FsxA-FLAG (WT), AFF-1-FLAG (negative control, cytotail), FsxA-ΔFL→AG$_4$A-FLAG, FsxA-ΔDIV→EFF-1 stem-FLAG and AFF-1-FLAG (permeabilized). The FLAG tag was inserted before the first transmembrane segment of each construct except for *C. elegans* AFF-1 in which the FLAG is at C-terminal after the cytoplasmic tail. Transfected BHK cells were incubated with anti-FLAG antibody on ice before fixation. Non-permeabilized staining of FLAG antibody showed the surface expression of FsxA and the mutants. *C. elegans* AFF-1 tagged with a cytoplasmic FLAG is a negative control for non-permeabilized staining. Permeabilized staining of *Ce*AFF-1-FLAG shows the localization on plasma membrane and internal compartments (Golgi region). Scale bars, 10 μm.


Salinity
Hypersaline
Freshwater
Moderate saline
NA
O2
Aerobic
Anaerobic
Hypoxic
NA
Sample type
Aqueous
Solid
NA
-5 kT/e
+5 kT/e
pH
Acid
Alkaline
Neutral
NA
Temperature
Cryophilic
Mesophilic
Thermophilic
NA
7P4L
6E18
HAP2
Salinity
O2
Sample type
pH
Temperature

Extended Data Fig. 8 | Tree of FsxAs, surface charge and environments.
Rooted phylogenetic tree with a selection of trimers generated by homology modelling[104]. A maximum likelihood tree was rooted with eukaryotic HAP2 sequences. The core trimer was modelled based on the crystal structure of FsxA (7P4L), ignoring N- or C-terminal extensions as well as some specific insertions. Despite the great diversity of environments where the sequences come from, all of them display uncharged (hydrophobic) tips, even when differing in surface electrostatics for the rest of the molecule. Indeed, for many closely related sequences coming from different environments, surface charge differences are readily apparent. Molecular surfaces coloured by electrostatic potential from negative red (-5 kT/e) to positive blue (+5 kT/e), through neutral/hydrophobic white (0 kT/e). For each modelled structure, top and side views are shown oriented as in **Fig. 2a**. Protein IDs at tree tips were number-coded for clarity (see **Source Data Table 1**), with colours for proteins from genomes of the Halobacteria class: Natrialbales order shown in red; Haloferacales order in bold blue show how FsxA proteins are incongruent with the species tree, evidencing HGT within Archaea. Sequences with less than four environmental descriptors were removed from the phylogenetic computation (except if coming from complete genomes). Both crystal structures (7P4L and *Chlamydomonas reinhardtii* 6E18[47]) are enlarged with respect to the models. Eukaryotic HAP2 tip labels: 61, *Vitrella brassicaformis*; 62, *Tetrahymena thermophila*; 63, *Trypanosoma cruzi*; 64, *Acanthamoeba castellanii*; 66, *Dictyostelium purpureum*. For details on archaeal protein IDs, metagenomics data sources and environmental descriptors refer to **Source Data Table 1**.

Supplementary Information for:

Archaeal origins of gamete fusion

David Moi[1,2,3,*], Shunsuke Nishio[4,*], Xiaohui Li[5,*], Clari Valansi[5], Mauricio Langleib[6,7], Nicolas G. Brukman[5], Kateryna Flyak[5], Christophe Dessimoz[2,3,8,9], Daniele de Sanctis[10], Kathryn Tunyasuvunakool[11], John Jumper[11], Martín Graña[7,#], Héctor Romero[6,12,#], Pablo S. Aguilar[1,13,#], Luca Jovine[4,#] and Benjamin Podbilewicz[5,#]

[1]Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE-CONICET), Buenos Aires, Argentina

[2]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

[3]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[4]Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden

[5]Department of Biology, Technion- Israel Institute of Technology, Haifa, Israel

[6]Unidad de Genómica Evolutiva, Facultad de Ciencias, Universidad de la República, Uruguay

[7]Unidad de Bioinformática, Institut Pasteur de Montevideo, Uruguay

[8]Centre for Life's Origins and Evolution, Dept. of Genetics, Evolution and Environment, University College London, United Kingdom

[9]Department of Computer Science, University College London, United Kingdom

[10]Structural Biology Group, ESRF - The European Synchrotron, Grenoble, France

[11]DeepMind, London, UK

[12]Centro Universitario Regional Este - CURE, Centro Interdisciplinario de Ciencia de Datos y Aprendizaje Automático - CICADA, Universidad de la República, Uruguay

[13]Instituto de Investigaciones Biotecnológicas Dr. Rodolfo A. Ugalde, Universidad Nacional de San Martín (IIB-CONICET), San Martín, Buenos Aires, Argentina

This file contains Supplementary Methods, Supplementary Figs. 1–12, Supplementary Tables 1-5 and Supplementary References.

Table of Contents

Supplementary Methods

Integrated Mobile Element (IME) identification by k-mer spectra analysis and comparative genomics

Among different methodologies that rely on DNA composition to identify horizontally transferred genomic regions[1], k-mer spectrum analysis is a standard tool for this purpose[2,3]. Normalized k-mer spectra for DNA sequences of arbitrary length were generated by counting occurrences of all k-mers and normalizing by the total amount of words counted. k-mer sizes from 3 bp to 8 bp were tested with no effect on results. A length of 4 bp was selected. To detect possible horizontally transferred regions, an average spectrum for each genome was calculated. A spectrum was calculated for a sliding window of 1 kb using 500 bp steps and subtracted from the genomic average at each window position **(Supplementary Fig. 6)**. The absolute value of the difference between the genomic average and window spectra is represented over the entire genome. Gaussian mixture models using two distributions were fit[4] to the k-mer content of all windows, to classify these as belonging to either the core genome or transferred elements. This deviation in k-mer spectra has been explored in the context of the archaeal mobilome and contains information on the ecological niche and evolutionary history of DNA sequences[5].

Comparison between close species with presence (*fsxA+)* or absence (*fsxA-)* of archaeal fusexins to detect insertion sites was done performing sequence similarity searches in complete genomes from the closest relatives available in the PATRIC database[6] (**Extended Data Table 1** and **Supplementary Fig. 5**). Coordinates of *fsxA*-containing IMEs present in pure culture genomes (PCGs) are annotated in **Supplementary Table 2.**

Homology groups for gene sequences encoded in IMEs from PCGs, metagenomics-assembled genomes (MAGs) and metagenome assembled scaffolds (see **Source Data Table 1**) were created by means of the pipeline represented in **Supplementary Fig. 7**. Briefly, after in-house gene re-annotation in each mobile element (ME), successive rounds of similarity searches using Hidden Markov Model (HMM)-protein and HMM-HMM comparisons were used to establish group belonging. Finally, we clustered MEs using a Jaccard index based distance matrix from these homology groups to assess their similarity.

Synteny conservation plots were made with MCscan tool from the JCVI pipeline[7], creating relevant files by formatting data regarding the inferred homology relationship with homemade Python scripts.

HMMER[8] and Pfam[9] were used on default parameters to assign domains and their associated arCOG[10,11] identifiers to ORFs (**Source Data Table 2**).

IME homology analyses

We followed the pipeline depicted in **Supplementary Fig. 7**. Briefly, PCGs' IMEs were determined by a combination of k-mer spectra and genomic alignments (see **Supplementary Table 2**). We initially inspected *fsxA*-containing scaffolds and kept only sequences that were 20 kb or longer for downstream analyses. We generated an enriched annotation for each IME. Then, we obtained an initial set of groups of homologous sequences, and each of these groups was enriched by means of HMM searches. Subsequently, the enriched homology groups showing similarity between them, as judged by HMM-HMM comparisons, were collapsed into unique groups.

In detail, first, we re-annotated the identified mobile elements (see Methods), combining the corresponding segment of the PATRIC[6] GFF annotation file with in-house ORF predictions (minimum ORF length of 30 nucleotides, option by default). ORF inference was done by means of getorf of the EMBOSS package v6.6.0.0[12], specifying genetic code by Table 11 (Bacteria and Archaea) and other parameters running by default. The similarity of inferred ORFs and annotated features in these mobile elements (i.e. features in their GFF annotation file) was established by means of BLASTP reciprocal searches[13]. We kept all the predicted ORFs and homologues that were at least annotated in one genome, in this way we tried to recover missanotated conserved ORFs.

Initial sets of homologues were generated with get_homologues v20210305[14]. Sequence identity and query coverage thresholds were set to 35% and 70%, respectively. In-paralogues were not allowed within these groups (option '-e'), and remaining parameters were run by default.

HMM profiles were constructed for each homologue group. To this aim, homologous sequences were retrieved for members of each group from the UniRef50 database[15] with jackhmmer (HMMER package v3.1b2; http://hmmer.org)[8] running with one iteration ('-N 1' parameter). MSAs were then generated for each group and its relevant hits with MAFFT[16] (v7.310) running under '--auto' parameter, and HMMs were created with hmmbuild (HMMER[8]). Homologue groups were enriched by means of HMM searches with hmmsearch (HMMER[8]), using each HMM as a query against a database comprising all predicted ORFs described above. Hits showing an e-value < 1e-10 and covering at least 50% of the HMM were added to the groups.

Enriched homology groups showing homology were collapsed. For this purpose, HMM-vs-HMM comparisons were performed with HHalign v3.3.0 from the HHsuite[17]. A graph was created with the Python library networkx v2.5.1, each node being an enriched group of homologues. An edge was established between nodes if their HMM-HMM alignment was significant (i.e. e-value < 1e-10, HMM coverage of longest HMM >= 50%). Groups of interconnected nodes were established with the 'connected_components()' routine, creating a collapsed homology group in each case.

Finally, we assessed the gene content similarity between mobile elements using a Jaccard Index based on the homology groups defined above. Usual Jaccard index of two sets is defined as (# of the intersection)/(# of the union). In this case:

$$J\,(MEA, MEB) \;=\; \frac{Nhomology\ groups\ shared\ between\ ME\ A\ \&\ ME\ B}{N\ homol.\ groups\ MEA + N\ homol.\ groups\ MEB - Nhomol.\ groups\ shared\ between\ ME\ A\ \&\ ME\ B}$$

We performed a hierarchical clustering of the MEs based on a distance matrix obtained from the pairwise Jaccard Indexes (distance(A,B) = 1 - $J_{A,B}$). This was done in Python with seaborn v0.11.1[18], employing the clustermap function. A subset of 11 mobile elements (**Supplementary Fig. 9**, in red), which included ME from PCGs and

FsxA.11, was selected for synteny conservation analysis. Plots depicting synteny in gene content between homolog groups were generated employing the MCscan tool[7]. Collapsed clusters can be found in Supplementary Information.

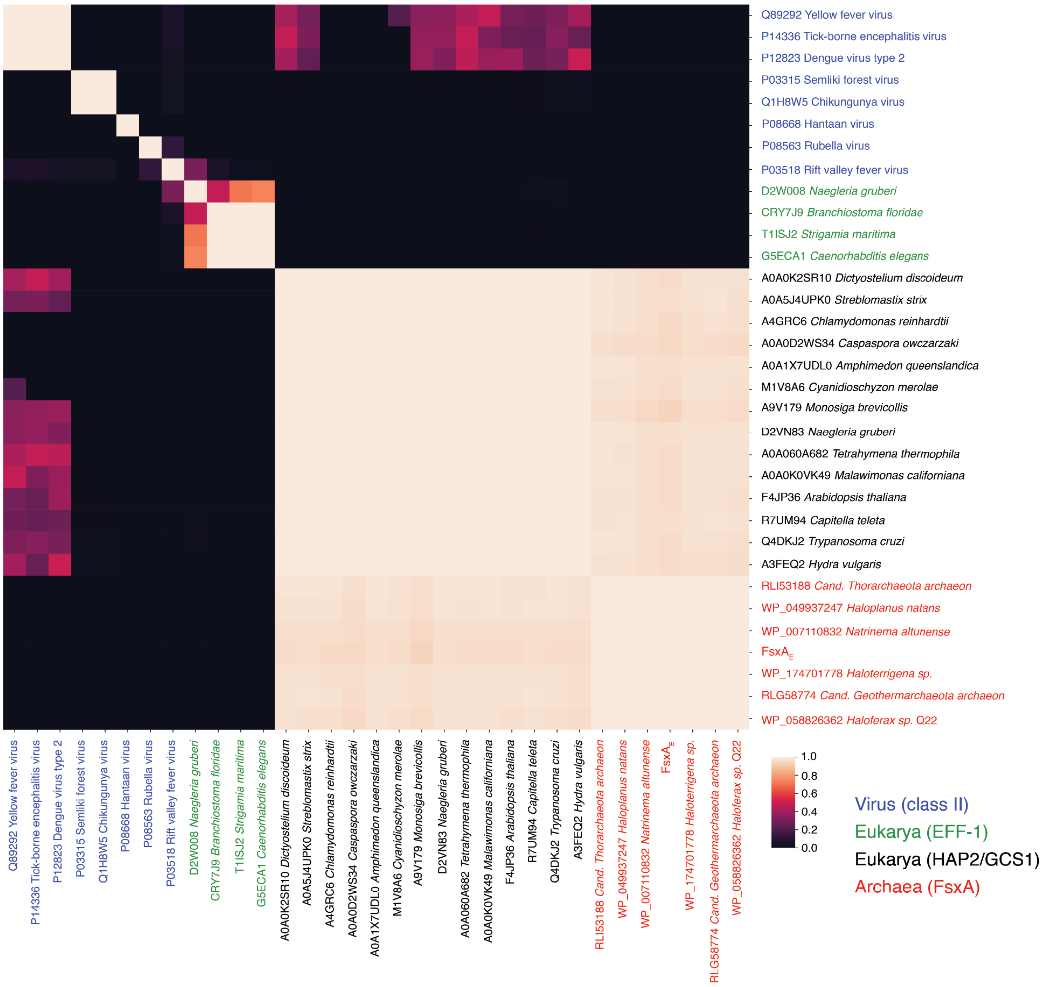
Relative time of acquisition analysis

Previous work[19,20] has closely examined relative acquisition times of thousands of eukaryotic genes. Briefly, given a set of homologous genes present in Prokaryotes and Eukaryotes, analysis of the corresponding phylogeny may shed light on the relative time of acquisition of this gene. First, we have to determine if this gene was in the LECA. Then, with its phylogeny (**Supplementary Fig. 10a**) we can define: i) a LECA node indicating the node where this gene started to diversify within the Eukaryotic clade, ii) an acquisition node connecting eukaryotic to prokaryotic lineages, and iii) the ‘stem’, which is the branch connecting the acquisition node with the LECA node. At this point a sister lineage can be identified as the potential donor of the gene and the different stem lengths will inform on how early (or late) this gene was acquired. Long stems are indicative of older genes, thus probably present in the FECA, while short stems suggest a late acquisition during the FECA-to-LECA transition. Given that genes have different evolutionary rates, it is important to normalize the raw stem length (Rsl) dividing it by the median of the eukaryotic branch length (from the LECA node to each tip of the tree) obtaining a normalized stem length (Sl). For a complete description of this approach, see previous reports[19,20]. Based on this idea, we decided to perform a similar analysis of the *fsxA* gene. Some singularities of this family, such as its existence within IMEs in a limited set of completely sequenced genomes, may influence the results. Yet, it meets the general criteria for the analysis and its comparison with the results obtained by these authors should be informative.

Thus, we closely followed the approach described in references 19 and 20 in order to compare our results with these works. We downloaded the same 209 complete eukaryotic proteomes and scanned them with hmmsearch[8] with a HAP2 HMM profile built from a curated alignment from our previous work[21], obtaining 86 sequences. Then, following the scrollsaw[22] method to pick slowly evolving sequences, we generated automatic and manual datasets, discarding fast evolving sequences, clade-specific duplications and possible phylogenetic artifacts. The scrollsaw method has proved useful in certain inferences, as it increases well-supported deep nodes in phylogenetic trees. We next proceeded to reduce the PCGs + MAGs + Metagenomic FsxA dataset with kClust v1.037[23] using a clustering threshold of 2.93, which corresponds to a 60% sequence identity. Finally, we combined eukaryotic and prokaryotic results into several datasets in order to perform sensitivity analyses. In the most stringent dataset, we excluded metagenomic sequences outside the PCG clade (n=2), addressing the possibility of tree artifacts or even the accidental inclusion of a eukaryotic sequence from metagenomic data.

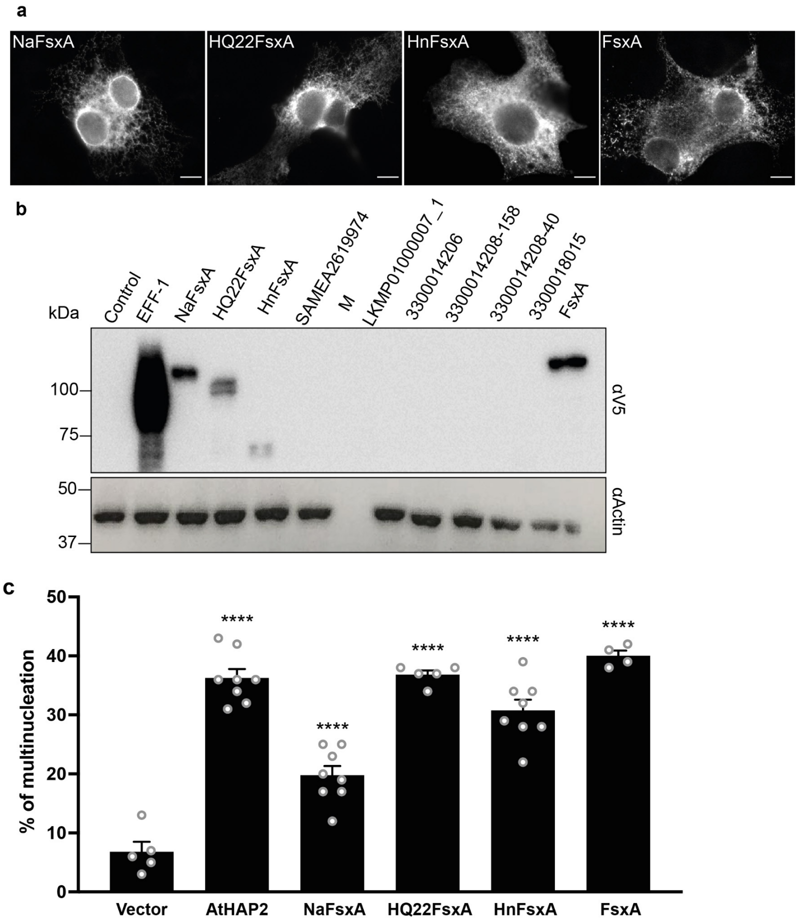
Multiple sequence alignments were computed with MAFFT[16] v7.310 ('--auto' option) and trimmed with trimAl v1.4.rev1544[24] with a 10% gap threshold ('-gt 0.1'), i.e. only removing columns with more than 90% gaps. Phylogenetic trees were inferred with IQ-TREE v1.6.12 (LG4X evolutionary model[25], 1,000 ultrafast bootstraps[26]). Trees were analyzed with ETE Toolkit[27]. All data and scripts are described in the Supplementary Information Guide and available upon request. The stem length distributions and sister clade identification for the eukaryotic genes of references 19 and 20 were obtained from data made available by the authors[28].

Supplementary Figures



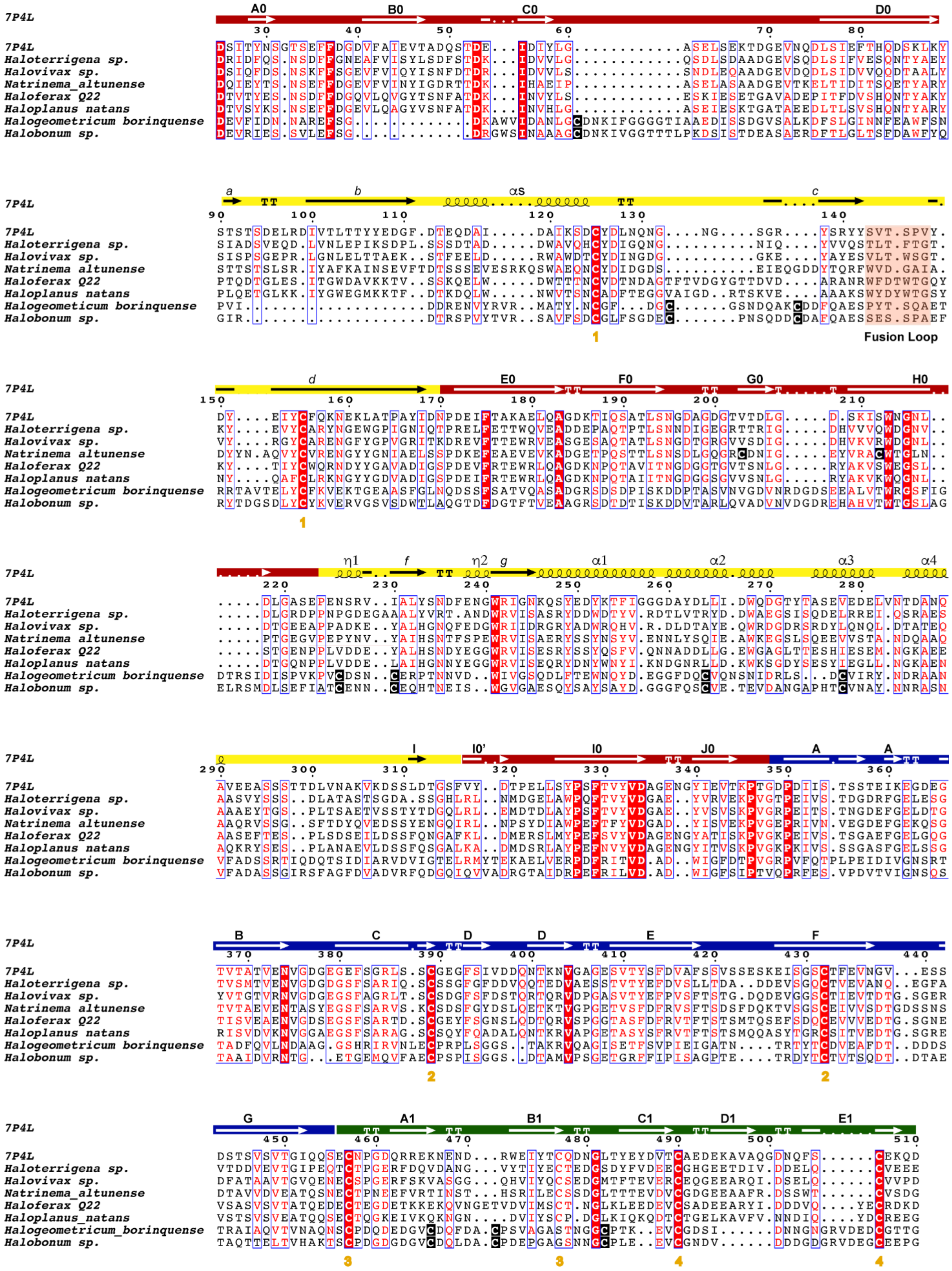
Supplementary Fig. 1 | Full version of the HMM homology probabilities of eukaryotic, viral and archaeal fusexin ectodomains.

HMMs were constructed for each ectodomain sequence (UniProt and NCBI identifiers shown)[29]. HAP2/GCS1 and EFF-1 sequences were chosen from representative species of the major eukaryotic lineages where these fusexins are present. Flavi-, alpha-, rubi- and bunyaviruses encompass all currently known viral fusexins. All vs all probabilities of homology as determined by HHblits[29] were clustered along rows and columns using UPGMA with Hamming distance. Several sequences selected for this analysis have corresponding crystal structures: Yellow fever virus (UniProt: Q89292, PDB: 6IW5[30]), Chikingunya virus (UniProt: Q1H8W5, PDB: 3N43[31]), Dengue virus (UniProt: P12823, PDB: 1OAN[32]), Semliki forest virus (UniProt: P03315, PDB: 1RER[33]), Tick-borne encephalitis virus (UniProt: P14336, PDB: 1SVB[34]), *Arabidopsis thaliana* (UniProt: F4JP36, PDB: 5OW3[35]), Rubella virus (UniProt: P08563, PDB: 4ADG[36]), *Chlamydomonas reinhardtii* (UniProt: A4GRC6, PDB: 5MF1[35]), Hantavirus (UniProt: P08668, PDB: 5LK1[37]) and Rift valley fever virus (UniProt: P03518, PDB: 4HJC[38]). Although all of the sequences used as input belong to the fusexin structural superfamily, HMM vs HMM comparisons can only detect homology within subsets of the superfamily.



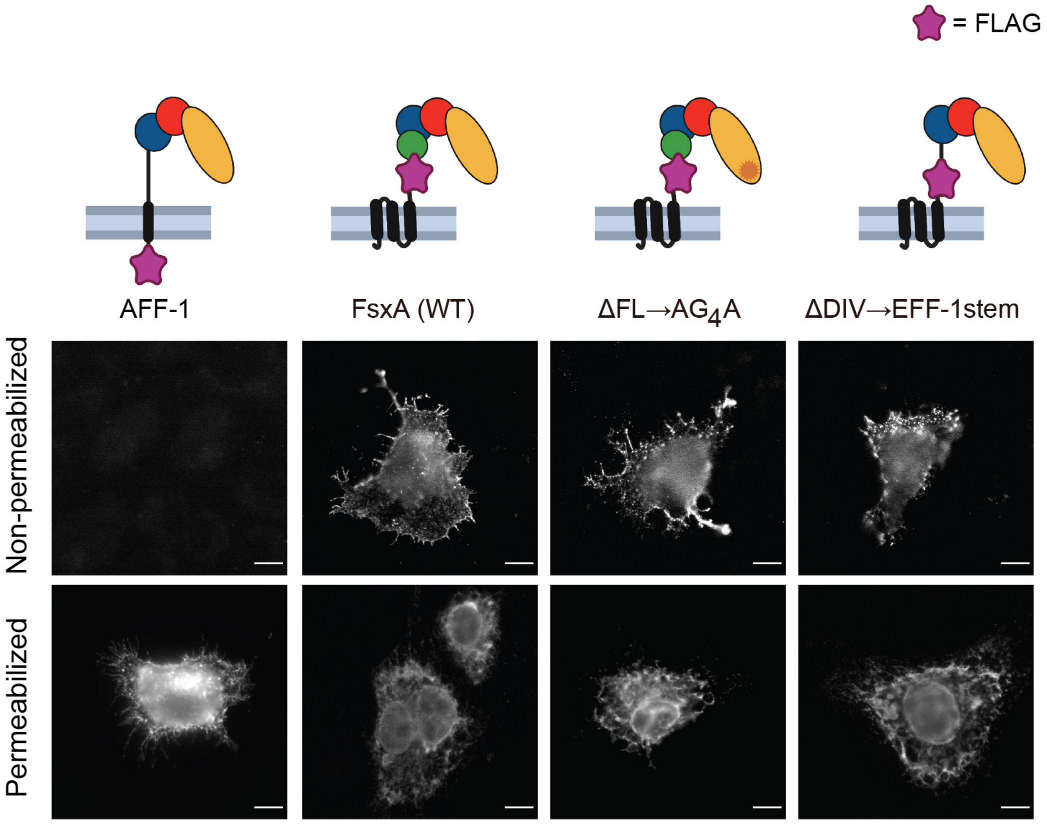
Supplementary Fig. 2 | Ectopic expression of archaeal fusexins in BHK cells.
a-b, Ten archaeal genes were synthesized (**Supplementary Table 3**) and independently expressed in BHK cells using an inducible promoter. Immunofluorescence (a) and Western blot (b) showing ectopic expression detected with anti-V5 antibody. EFF-1 from *C. elegans* was used as a positive control. NaFsxA, *Natrinema altunense* FsxA. HQ22FsxA, *Haloferax* sp. Q22 FsxA. HnFsxA, *Haloplanus natans* FsxA. LKMP01000007_1 was obtained from Nanohaloarchaea B1-Br10_U2g21 LB-BRINE-C121. FsxA (the protein subsequently characterized), SAMEA2619974 and sequences starting with "330" were obtained from metagenomic databases (see **Supplementary Table 3** for complete accession numbers). Scale Bars, 10 µm.

c, Quantification of multinucleation in cells expressing archaeal fusexins. Cells were transfected with archaeal fusexins cloned into pCI::H2B-RFP/GFP vectors separately. 48 h post-transfection, immunofluorescence was performed with anti-V5 antibody. Empty vector pCI::H2B-RFP or pCI::H2B-GFP co-transfected with myr-EGFP were the negative controls. AtHAP2 was used as a positive control. Multinucleation was determined as the ratio between the number of nuclei in multinucleated cells and the total number of nuclei in multinucleated cells and expressing cells that were in contact but did not fuse. The percentage of multinucleation is presented as means ± SEM of independent experiments ($n \geq 4$). Total number of nuclei counted in multinucleated cells and in cells in contact $n \geq 1{,}000$ for each experimental condition. Comparisons were made with one-way ANOVA followed by Dunett's test against the empty vector. **** $p < 0.0001$.


7P4L
A0 B0 C0 D0
30 40 50 60 70 80
7P4L DSITYNSGTSEFFDGDVFAIEVTADQSTDE...IDIYLG............ASELSEKTDGEVNQDLSIEFTHQDSKLKY
Haloterrigena sp. DRIDFQS.NSDFFGNEAFVISYLSDFSTDK...IDVVLG............QSDLSDAADGEVSQDLSIFVESQNTYAEY
Halovivax sp. DSIQFDS.NSKFFSGEVFVIQYISNFDTDR...IDVVLS............SNDLEQAADGEVDQDLSIDVVQQDTAALY
Natrinema_altunense DQIEYTS.NSEFFDGEVFVINYIGDRTTDK...IHAEIP............ASELASAADGEVTKELTIDITSQETYARY
Haloferax Q22 DTVTYES.NSDFFDGQVLQVGYTSNFATDK...INVYLS............ESKIESETGAVADEPITFDVSHQNTYAKY
Haloplanus natans DTVSYKS.NSEFFDGEVLQAGYVSNFATDK...INVHLG............ASEIESKTGATAEEDLTLSVSHQNTYARY
Halogeometricum borinquense DEVFIDN.NAREFSG.............DKAWVIDANLGCDNKIFGGGGTIAAEDISSDGVSALKDFSLGINNFEAWFSN
Halobonum sp. DEVRIES.SVLEFSG.............DRGWSINAAAGCDNKIVGGTTTLPKDSISTDEASAERDFTLGLTSFDAWFYQ
7P4L
a b αS c
TT TT
90 100 110 120 130 140
7P4L STSTSDELRDIVTLTTYYEDGF.DTEQDAI.....DAIKSDCYDLNQNG....NG....SGR....YSRYYSVT.SPVY.
Haloterrigena sp. SIADSVEQD.LVNLEPIKSDPL.SSSDTAD.....DWAVQHCYDIGQNG.........NIQ....YVVQSTLT.FTGT.
Halovivax sp. SISPSGEPR.LGNLELTTAEK..STFEELD.....RWAWDTCYDINGDG.........GKE....YAYESVLT.WSGT.
Natrinema altunense STTSTSLSR.IYAFKAINSEVFTDTSSSEVESRKQSWAEQNCYDIDGDS.........EIEQGDDYTQRFWVD.GAIA.
Haloferax Q22 PTQDTGLES.ITGWDAVKKTV..SSKQELW.....DWTTTNCVDTNDAGTFTVDGYGTTDVD....ARANRWFDTWTGQY
Haloplanus natans PLQETGLKK.IYGWEGMKKTF..DTKDQLW.....NWVTSNCADFTEGGVAIGD..RTSKVE....AAAKSWYDYWTGSY
Halogeometicum borinquense PVI..................DDRENVYRVR..MATY.NCGF...DGC......GSNDQAKCDDFQAESPYT.SQAET
Halobonum sp. GIR..................DTRSPVYTVR..SAVFSDCGLFSGDEC.......PNSQDDCDAFQAESSES.SPAEF
1
Fusion Loop
7P4L
d E0 F0 G0 H0
TT TT T T
150 160 170 180 190 200 210
7P4L DY....EIYCFQKNEKLATPAYIDNPDEIFTAKAELQAGDKTIQSATLSNGDAGDGTVTDLG.....D.SKISWNGNL..
Haloterrigena sp. KY....EVYCARYNGEWGPIGNIQTPRELFETTWQVEADDEPAQTPTLSNNDIGEGRTTRIG.....DHVVVQWDGNV..
Halovivax sp. VY....RGYCARENGFYGPVGRITKDREVFTTEWRVEASGESAQTATLSNGDTGRGVVSDIG.....DHVKVRWDGNL..
Natrinema altunense DYYN.AQVYCVRENGYYGNIAELSSPDKEFEAEVEVKADGETPQSTTLSNSDLGQGRCDNIG.....EYVRACWTGLN..
Haloferax Q22 KY....TIYCWQRNDYYGAVADIGSPDEVFRTEWRLQAGDKNPQTAVITNGDGGTGVTSNLG.....RYAKVSWEGSL..
Haloplanus natans NY....QAFCLRKNGYYGDVADIGSPDEIFRTEWRLQAGDKNPQTAIITNGDGGSGVVSNLG.....RYAKVKWQGNL..
Halogeometricum borinquense RRTAVTELYCFKVEKTGEAASFGLNQDSSFSATVQASADGRSDSDPISKDDPTASVNVGDVNRDGDSEEALVTWRGSFIG
Halobonum sp. RYTDGSDLYCYKVERVGSVSDWTLAQGTDFDGTFTVEAAGRSDTDTISKDDVTARLQVADVNVDGDREHAHVTWTGSLAG
1
7P4L
η1 f η2 g α1 α2 α3 α4
TT
220 230 240 250 260 270 280
7P4L .....DLGASEPENSRV..IALYSNDFENGWRIGNKQSYEDYKTFIGGGDAYDLLI.DWQDGTYTASEVEDELVNTDANQ
Haloterrigena sp. .....DLGRDPPNPGDEGAAALYVRT.ANDWRVISASRYDDWDTYV.RDTLVTRYD.DWAEGSISQDELRREI.QSRAES
Halovivax sp. .....DTGEEAPPADKE..YALHGNQFEDGWRIIDRGRYADWRQHV.R.DLDTAYE.QWRDGDRSRDYLQNQL.DTATEQ
Natrinema altunense .....PTGEGVPEPYNV..YAIHSNTFSPEWRVISAERYSSYNSYV.ENNLYSQIE.AWKEGSLSQEEVVSTA.NDQAAQ
Haloferax Q22 .....STGENPPLVDDE..YALHSNDYEGGWRVISESRYSSYQSFV.QNNADDLLG.EWGAGLTTESHIESEM.NGKAEE
Haloplanus natans .....DTGQNPPLVDDE..LAIHGNNYEGGWRVISEQRYDNYWNYI.KNDGNRLLD.KWKSGDYSESYIEGLL.NGKAEN
Halogeometricum borinquense DTRSIDISPVKPVCDSN..CERPTNNVD..WIVGSQDLFTEWNQYD.EGGFDQCVQNSNIDRLS...DCVIRY.NDRAAN
Halobonum sp. ELRSMDLSEFIATCENN..CEQHTNEIS..WGVGAESQYSAYSAYD.GGGFQSCVE.TEVDANGAPHTCVNAY.NNRASN
7P4L
I I0' I0 J0 A A
TT TT
290 300 310 320 330 340 350 360
7P4L AVEEASSSTTDLVNAKVKDSSLDTGSFVY..DTPELLSYPSFTVYVDAGENGYIEVTKPTGDPDIIS.TSSTEIKEGDEG
Haloterrigena sp. AASVYSSS..DLATASTSGDA.SSGHLRL..NMDGELAWPQFTVYVDGAE..YVRVEKPVGTPEIVS.TDGDRFGELESG
Halovivax sp. AAAEYTGS..PLTSAETVSSTYTDGQLRL..EMDTDLAYPSFTVYVDGAE..YVSVSKPVGRPEITS.TNGDEFGELDTG
Natrinema altunense AAQRVSSG..SFTDYQVEDSSYENGQIRL..NPSYDIAWPEFTFYVDGAD..YISVEKPVGEPRIVN.VEGDEFGEKQSG
Haloferax Q22 AASEFTES..PLSDSEILDSSFQNGAFKL..DMERSLMYPEFSVYVDAGENGYATISKPVGKPEIVS.TSGAEFGELGQG
Haloplanus natans AQKRYSES..PLANAEVLDSSFQSGALKA..DMDSRLAYPEFNVYVDAGENGYITVSKPVGKPKIVS.SSGASFGELSSG
Halogeometricum borinquense VFADSSRTIQDQTSIDIARVDVIGTELRMYTEKAELVERPDFRITVD.AD..WIGFDTPVGRPVFQTPLPEIDIVGNSRT
Halobonum sp. VFADASSGIRSFAGFDVADVRFQDGQIQVVADRGTAIDRPEFRILVD.AD..WIGFSIPTVQPRFES.VPDVTVIGNSQS
7P4L
B C D D E F
TT TT
370 380 390 400 410 420 430 440
7P4L TVTATVENVGDGEGEFSGRLS.SCGEGFSIVDDQNTKNVGAGESVTYSFDVAFSSVSSESKEISGSCTFEVNGV...ESS
Haloterrigena sp. TVSMTVENVGDGDGSFSARIQ.SCSSGFGFDDVQQTEDVAESSTVTYEFDVSLLTDA..DDEVSGQCTVEVANQ..EGFA
Halovivax sp. YVTGTVRNVGDGEGSFAGRLT.SCSDGFSFDSTQRTQRVDPGASVTYEFPVSFTSTG.DQDEVGGSCTIEVTDTG.SGER
Natrinema altunense TVTAEVENTASYEGSFSARVS.KCSDSFGYDSLQETKTVGPGETVSFDFRVSFTSDSFDQKTVSGSCEIVVSDTGDSSNS
Haloferax Q22 TISVEAENVGDSEGSFSARTD.SCGEYFSGNSLQDTQRVSPGGTASFDFRVTFTSTSMTQSEFSDQCEVVVEDTG.SGNE
Haloplanus natans RISVDVKNVGGAEGSFSARAG.SCSQYFQADALQNTKRVAPGETASYSFRVTFTSTSMQQASYTGRCSITVEDTG.SGRE
Halogeometricum borinquense TADFQVLNDAAG.GSHRIRVNLECPRPLSGGS..TAKRVQAGISETFSVPIEIGATN....TRTYTCDVEAFDT..DDDS
Halobonum sp. TAAIDVRNTG...ETGEMQVFAECPSPISGGS..DTAMVPSGETGRFFIPISAGPTE....TRDYTCTVTSQDT..DTAE
2
2
7P4L
G A1 B1 C1 D1 E1
TT TT TT TT TT TT
450 460 470 480 490 500 510
7P4L DSTSVSVTGIQQSECNPGDQRREKNEND...RWEIYTCQDNGLTYEYDVTCAEDEKAVAQGDNQFS......CEKQD
Haloterrigena sp. VTDDVEVTGIPEQTCTPGERFDQVDANG...VYTIYECTEDGSDYFVDEECGHGEETDIV.DDELQ......CVEEE
Halovivax sp. DFATAAVTGVQENECSPGERFSKVASGG...QHVIYQCSEDGMTFTEVERCEQGEEARQI.DSELQ......CVVPD
Natrinema_altunense DTAVVDVEATQSNECTPNEEFVRTINST...HSRILECSSDGLTTTEVDVCGDGEEAAFR.DSSWT......CVSDG
Haloferax Q22 VSASVSVTATQEDECTEGDETKKEKQVNGETVDVIMSCT.NGLKLEEDEVCSADEEARYI.DDDVQ....YECRDKD
Haloplanus_natans VSTSVSVEATQQSECTQGKEIVKQKNGN....DVIYSCP.DGLKIQKQDTCTGELKAVFV.NNDIQ....YDCREEG
Halogeometricum_borinquense TRAIAQVTVNAQNSCPDQDEDGVCDQFDA.CPSVAGASTNGGCPTK..EVCGDSI......DNNGNGRVDEDCGTTG
Halobonum sp. TAQTTELTVHAKTSCPDGDGDGVCDQLDA.CPDEPGAGSNNGCPLE..EVCGNDV......DDDGDGRVDEGCEEPG
3
3
4
4

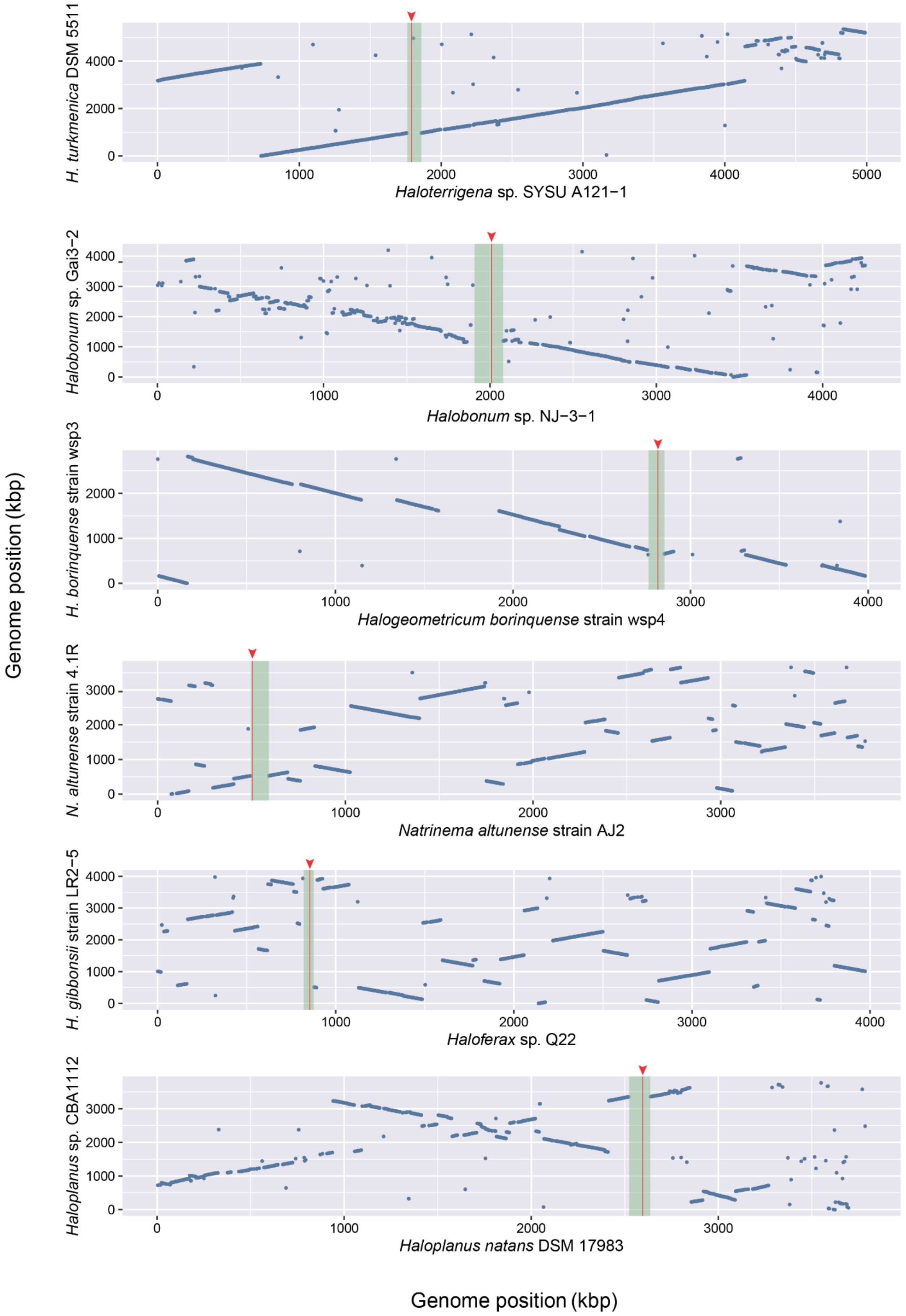
Supplementary Fig. 3 | Sequence alignment of archaeal fusexins.

Archaeal sequences from PCGs, with N- and C-terminal regions cropped to match FsxA$_E$ (PDB: 7P4L). Secondary structure elements are shown within boxed domains, coloured and labelled following the previous nomenclature (domain I, red; domain II, yellow; domain III, blue). Disulfide bonds are indicated by orange numbers below the alignment. Additional, lineage-specific cysteines are black-boxed and depicted in bold white. The fusion (cd) loop is highlighted in light orange within the alignment; as in eukaryotic HAP2, it has poor sequence conservation (see also **Extended Data Fig. 5a**) but shows a high prevalence of hydrophobic residues (see also **Extended Data Fig. 8**). Domain IV (green) has no eukaryotic counterpart and has relatively poor sequence conservation within archaea (**Extended Data Fig. 5a**), yet preserves its disulfides. For reproducibility, no gap or block was altered from the alignment. Identical column residues are depicted in bold white on a red background; conserved positions are boxed and labelled red. The figure was made with ESPript[39] and manually edited.



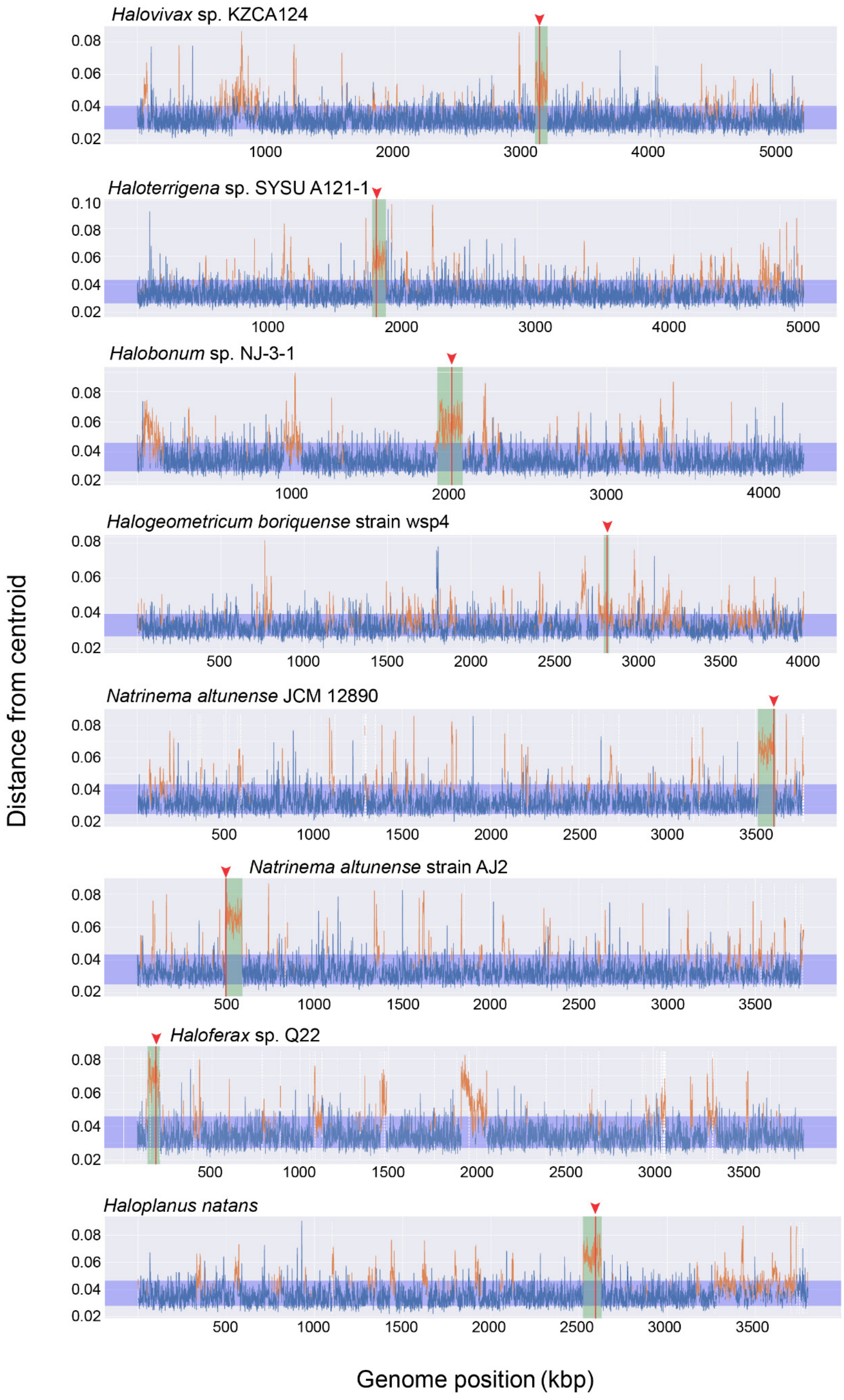
Supplementary Fig. 4 | Surface expression of FsxA and mutants.

BHK cells were transfected with FLAG-tagged FsxA (WT) and the indicated mutants; the FLAG tag was inserted before the membrane anchor (see **Extended Data Fig. 7d**). Non-permeabilized staining using anti-FLAG antibody showed surface expression of FsxA and the various mutants. The proportion of non-permeabilized cells showing surface expression was: AFF-1-FLAG (negative control; 0%, n~1000), FsxA-FLAG (3.9%, n=1242), FsxA-ΔFL→AG$_4$A-FLAG (4.4%, n=1176), FsxA-ΔDIV→EFF-1stem-FLAG (2.6%, n=1263). Another group of transfected BHK cells in parallel were fixed, permeabilized and stained with anti-FLAG antibody. Permeabilized staining showed main distribution in the cytoplasm (endoplasmic reticulum) of FsxA and its mutants. *C. elegans* AFF-1 tagged with FLAG at the C terminus (cytoplasmic tail) worked as a negative control for non-permeabilized staining. Scale Bars, 10 μm.



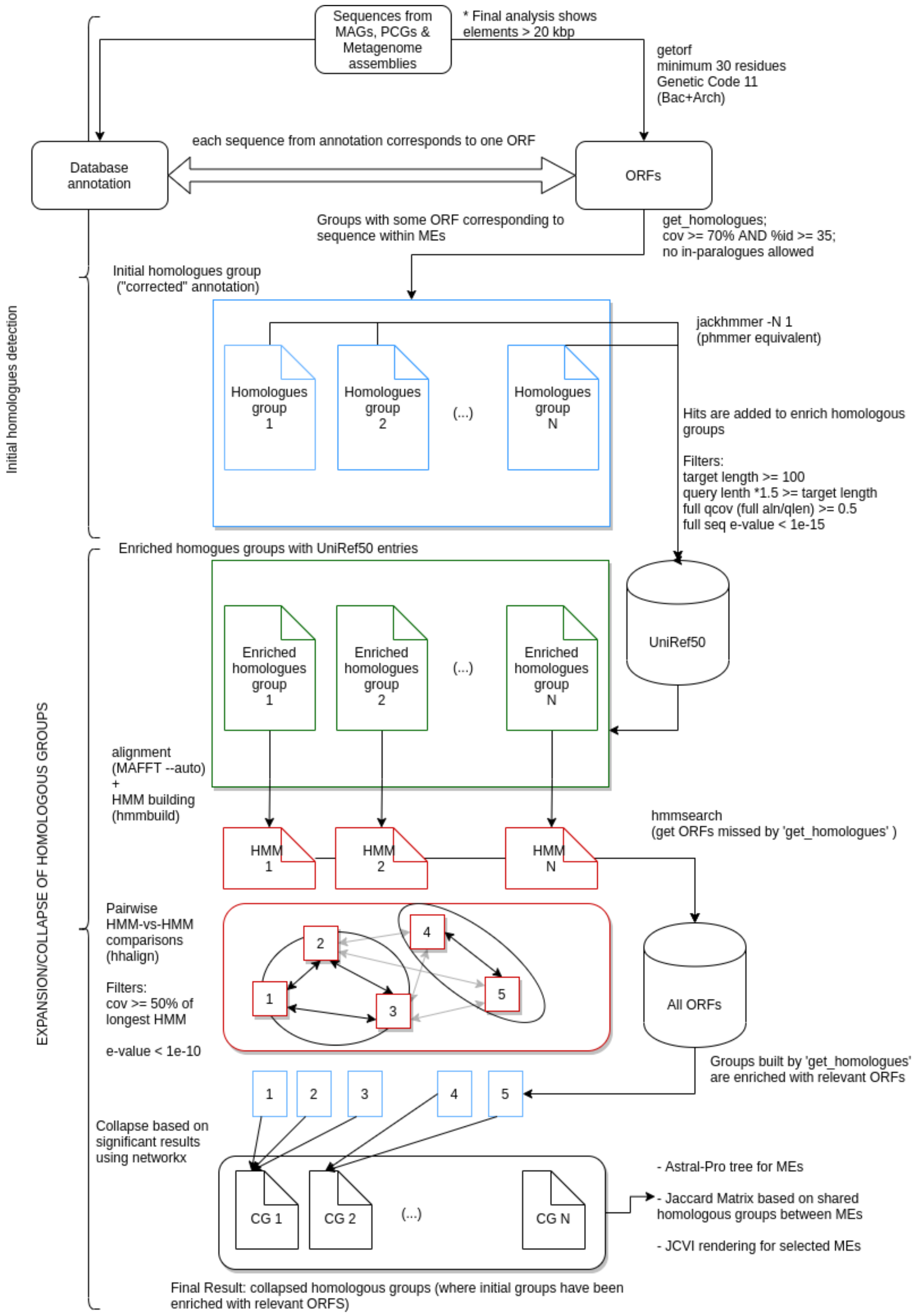
Supplementary Fig. 5 | Whole genome comparison of species with and without *fsxA*.

Each blue dot represents a segment of 500 bp with more than 80% identity between the species harbouring *fsxA* (*e.g. Haloplanus natans* DSM 17083) and the species with no *fsxA* (e.g. *Haloplanus* sp. CBA1112). Species with *fsxA* are in the x axis, the base of the green rectangles represent the detected IME carrying the *fsxA* gene, locus of *fsxA* is in red vertical line and pointed with a red arrowhead.



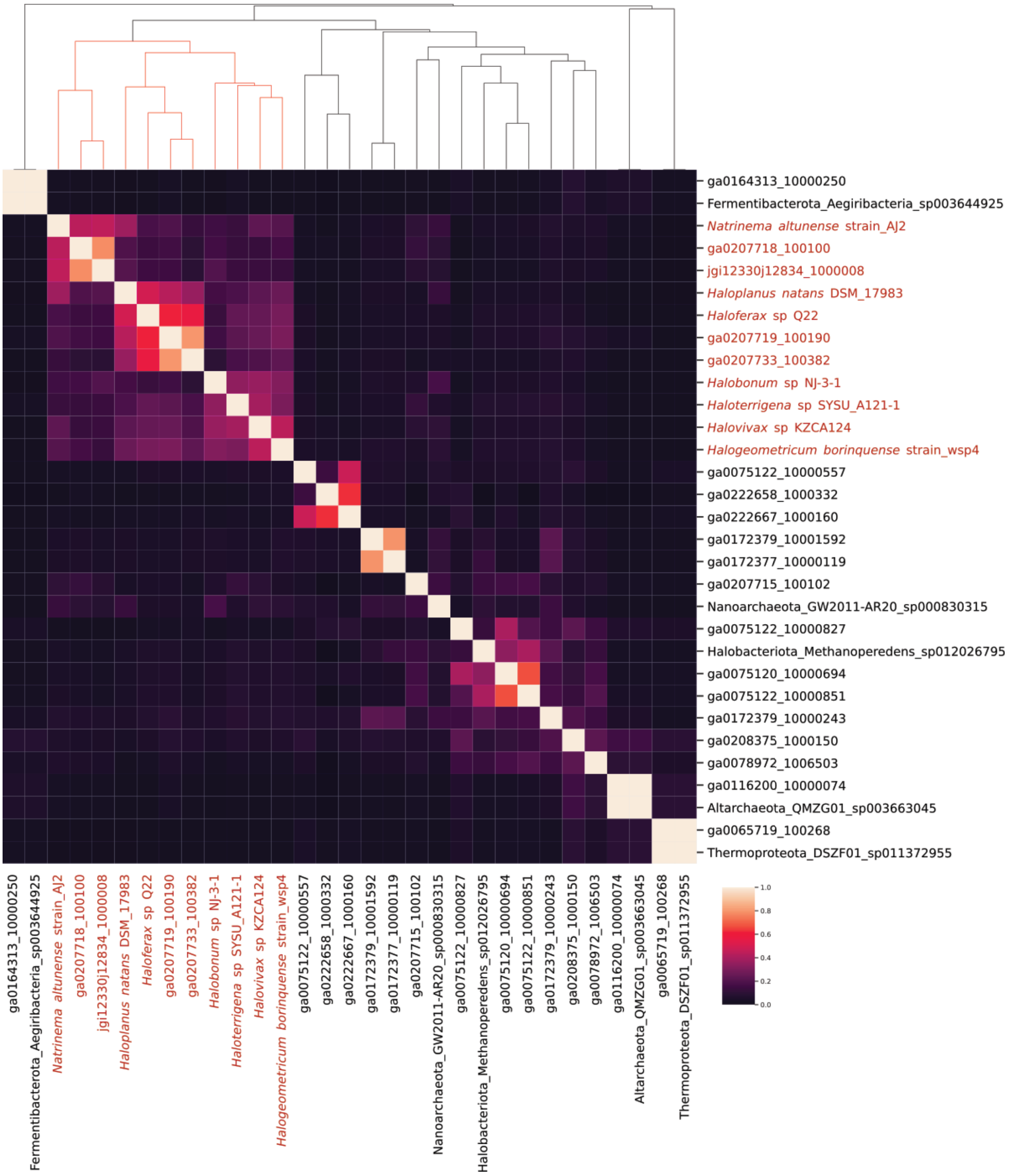
Supplementary Fig. 6 | K-mer spectra deviation of *fsxA*-containing IMEs.

K-mer spectrum deviation from centroid is shown for each of the PCGs where *fsxA* was detected (see Methods). Blue region shows the standard deviation. Locus of *fsxA* is in red vertical line and pointed with a red arrowhead, the mobile element containing *fsxA* is in green. Dashed vertical white lines indicate the end of a contig in the genome assembly. *fsxA* is consistently found within regions that deviate from the core genome's spectrum, indicating they belong to a mobile element.


Sequences from MAGs, PCGs & Metagenome assemblies
* Final analysis shows elements > 20 kbp
getorf
minimum 30 residues
Genetic Code 11
(Bac+Arch)
each sequence from annotation corresponds to one ORF
Database annotation
ORFs
Groups with some ORF corresponding to sequence within MEs
get_homologues;
cov >= 70% AND %id >= 35;
no in-paralogues allowed
Initial homologues group ("corrected" annotation)
Initial homologues detection
jackhmmer -N 1
(phmmer equivalent)
Homologues group 1
Homologues group 2
(...)
Homologues group N
Hits are added to enrich homologous groups
Filters:
target length >= 100
query lenth *1.5 >= target length
full qcov (full aln/qlen) >= 0.5
full seq e-value < 1e-15
Enriched homogues groups with UniRef50 entries
Enriched homologues group 1
Enriched homologues group 2
(...)
Enriched homologues group N
UniRef50
EXPANSION/COLLAPSE OF HOMOLOGOUS GROUPS
alignment
(MAFFT --auto)
+
HMM building
(hmmbuild)
HMM 1
HMM 2
HMM N
hmmsearch
(get ORFs missed by 'get_homologues')
Pairwise HMM-vs-HMM comparisons (hhalign)
Filters:
cov >= 50% of longest HMM
e-value < 1e-10
2
4
1
3
5
All ORFs
Groups built by 'get_homologues' are enriched with relevant ORFs
1
2
3
4
5
Collapse based on significant results using networkx
CG 1
CG 2
(...)
CG N
- Astral-Pro tree for MEs
- Jaccard Matrix based on shared homologous groups between MEs
- JCVI rendering for selected MEs
Final Result: collapsed homologous groups (where initial groups have been enriched with relevant ORFS)

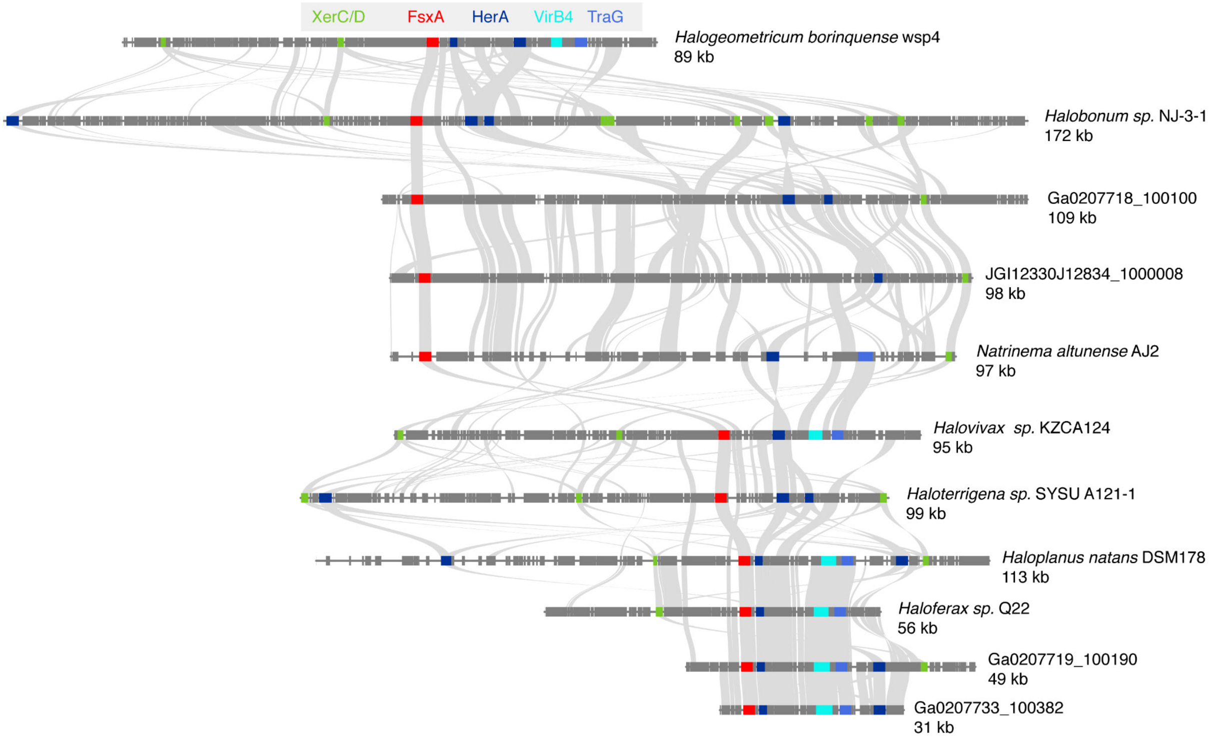
Supplementary Fig. 7 | Bioinformatics workflow for IMEs.

The general workflow is divided into two parts. First, we re-annotated the ORFs of each IME and searched for potential homologues between them. Then we enriched these initial groups by searching the UNIREF50 database and generated new HMM profiles. With these new HMMs we searched again within each IME to capture any potentially missing homologue. Finally, we performed HMM vs HMM[40] search to collapse similar groups into one. CG1, CG2...CGN are final collapsed homologous groups which are the basis for IME clustering, synteny conservation and gene content analyses shown in **Supplementary Figs. 8, 9** and **Supplementary Table 1**, respectively.



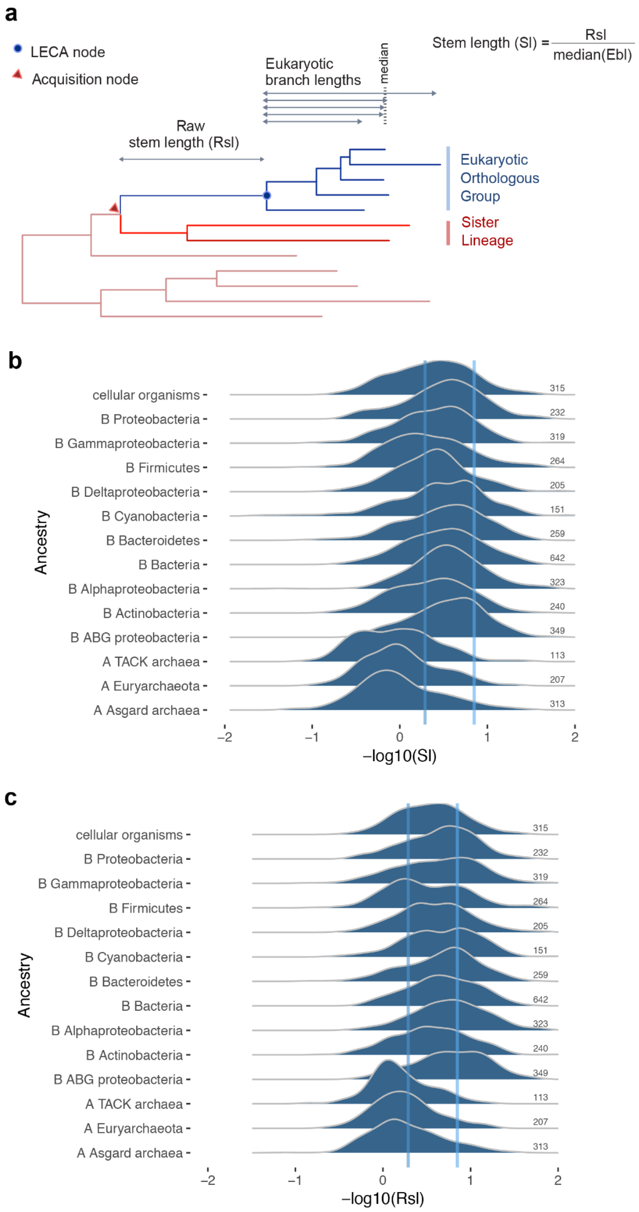
Supplementary Fig. 8 | Clustering of IMEs from complete genomes, MAGs and metagenomic contigs based on gene content.

A distance metric based on the sharing of homologous genes between all IMEs was computed (see Methods). Then a pairwise distance matrix was built to perform hierarchical clustering. There is a clear cluster, marked in red, which contains all PCG IMEs and the DNA contig containing the crystallized FsxA (jgi12330j12834_1000008). This cluster of 11 IMEs was used for the synteny conservation analysis.



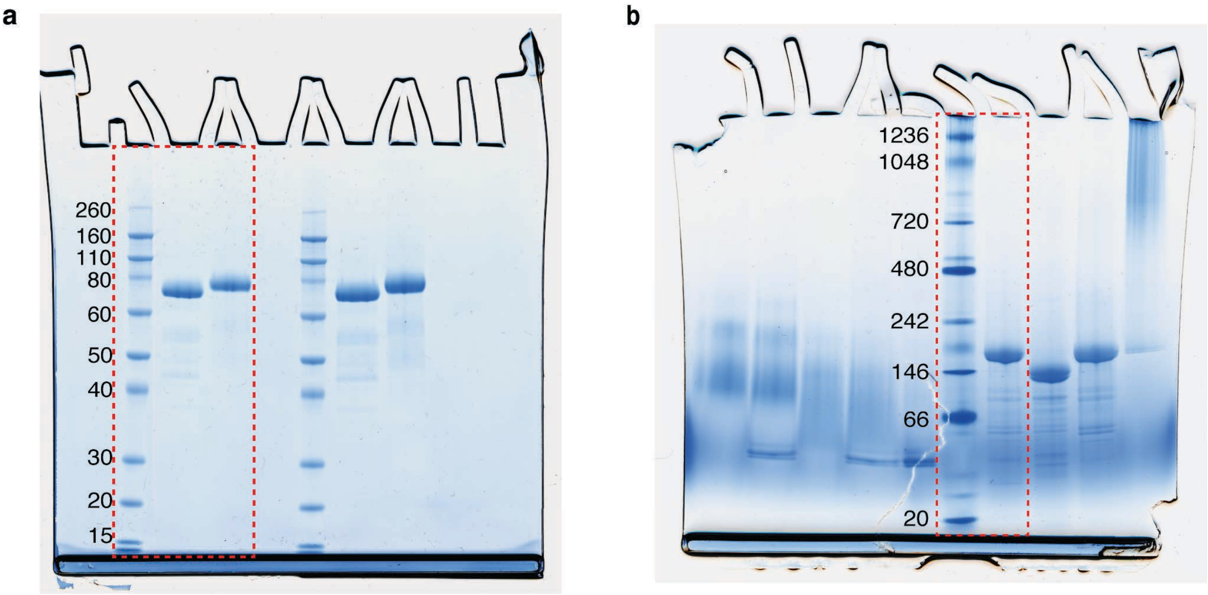
Supplementary Fig. 9 | Synteny plots for IMEs from complete genomes and metagenomic data.

Annotated regions plus inferred ORFs belonging to homologous clusters identified by the workflow depicted in **Supplementary Fig. 7**. Homology relationships are represented by grey links. *fsxA* genes are marked in red and selected ORFs homologous to IME signature genes are labelled and colour-coded. XerC/XerD recombinases (green); HerA helicase (dark blue); VirB4, Type IV secretory (T4SS) pathway (cyan); TraG/TraD/VirD4 family enzyme, ATPase, T4SS (see **Supplementary Table 1** and **Source Data Table 3** for details). The eleven segments analyzed correspond to the cluster marked in **Supplementary Fig. 8.**



Supplementary Fig. 10 | Timing of gene acquisitions from different phylogenetic origins during eukaryogenesis.

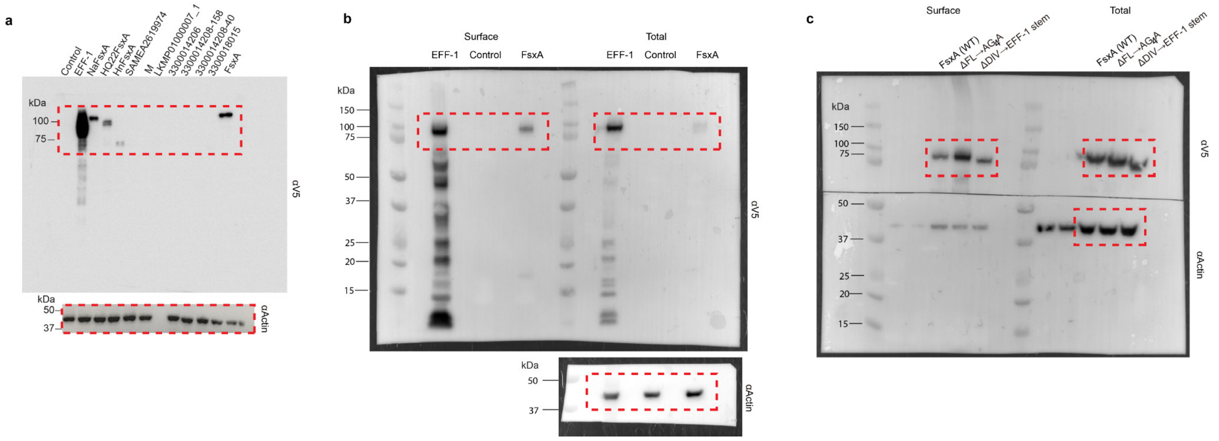
a, Stem length analysis rationale. The stem length of a given set of homologues is calculated as the Raw length of the branch that connects the LECA node with the acquisition node divided by the median of all the branch lengths from the LECA node to the different tips of the tree. Shorter stems indicate more recent acquisitions of the gene. The sister lineage is the immediate divergent lineage before the LECA node. **b** and **c**, Ridgeline plots showing the distribution of the normalized (**b**) and raw (**c**) stem lengths, depicted as the additive inverse of the log-transformed values. Consequently, longer branches have a smaller value and vice versa. The total number of genes for each identified sister lineage is on the right; “cellular organisms” means that the sister lineage is composed of archaea and bacteria. Steel-blue vertical lines indicate the maximum and minimum estimates of the FsxA-HAP2 stem length (see Supplementary Methods). Data obtained from elsewhere[20,28], except for FsxA-HAP2.



Supplementary Fig. 11 | Uncropped gel images from Extended Data Fig. 1.

a, SDS-PAGE gel stained with Coomassie G-250. The section indicated by a red dashed square is used in **Extended Data Fig. 1a**.

b, BN-PAGE gel stained with Coomassie G-250. The section indicated by a red dashed square is used in **Extended Data Fig. 1b**.



Supplementary Fig. 12 | Uncropped Western blot images from Supplementary Fig. 2b and Extended Data Fig. 7b, c.

a, Western blot probed with anti-V5 (upper) and anti-actin (lower) antibodies. The sections indicated by red dashed squares are used in **Supplementary Fig. 2b**.

b, Western blot probed with anti-V5 (upper) and anti-actin (lower) antibodies. The sections indicated by red dashed squares are used in **Extended Data Fig. 7b**.

c, Western blot probed with anti-V5 (upper) and anti-actin (lower) antibodies. The sections indicated by red dashed squares are used in **Extended Data Fig. 7c**.

Supplementary Tables

Supplementary Table 1 | Most common arCOGs from 11 IMEs[a]

Collapsed Group Name	IMEs count[b]	arCOG count[c]	arCOG	Category	Annotation
CG_2	11	22	arCOG00280, arCOG00285, arCOG06224	L	HerA helicase
CG_17	10	21	arCOG01241, arCOG01248, arCOG01250	X	XerC XerD/XerC family integrase
41684	10	3	arCOG01680, arCOG02808, arCOG04362	K	Transcriptional regulator containing HTH domain
43214	9	9	arCOG08903	S	Uncharacterized protein
CG_1	8	8	arCOG12186	S	Uncharacterized protein
CG_2	7	7	arCOG04816	U	TraG/TraD/VirD4 family enzyme, ATPase
43797	7	7	arCOG12187	S	Uncharacterized membrane protein
43810	7	7	arCOG10296	S	Uncharacterized membrane protein
43833	7	7	arCOG08907	S	Uncharacterized protein
43868	7	7	arCOG10381	S	Uncharacterized protein
CG_2	6	7	arCOG00467	L	Cdc6-related protein, AAA superfamily ATPase
CG_2	6	6	arCOG07496	U	VirB4, Type IV secretory pathway
42763	6	6	arCOG06216	K	Transcriptional regulator
CG_2	5	5	arCOG01308	O	ATPase of the AAA+ class , CDC48 family
CG_21	3	3	arCOG07871	K	Helicase
CG_2	2	2	arCOG00439	L	ATPase involved in replication control
CG_2	2	2	arCOG03779	V	GTPase subunit of restriction endonuclease
CG_2	2	2	arCOG05935	R	Helicase of FtsK superfamily
CG_21	2	2	arCOG00878	V	restriction-modification related helicase
CG_9	1	1	arCOG03600	E	Transglutaminase-like cysteine protease
CG_21	1	1	arCOG04818	K	Superfamily II DNA/RNA helicase, SNF2 family

a. Collapsed Homology Groups (**Supplementary Fig. 7**) corresponding to the 11 IME cluster (**Supplementary Fig. 8**) were analyzed using HMMER against the arCOGs database. Collapsed Homology Groups with zero identified arCOGs are not shown. Full dataset with results is in **Source Data Table 3.**

b. Number of IMEs where the indicated arCOGs are present.

c. Total number of ORFs belonging to arCOGs in the next column identified in the set of 11 IMEs.

Supplementary Table 2 | IMEs carrying the *fsxA* gene in completely sequenced genomes

Species with *fsxA*	NCBI TaxID	PATRIC ID	sequence ID[a]	ME start k-mer[b]	ME end k-mer[b]	Length ME k-mer	ME start CG[c]	ME end CG[c]	Length ME CG	Species w/o *fsxA*[d]	NCBI TaxID	PATRIC ID	ANI[e]
Haloplanus natans DSM 17983	926690	926690.3	ATYM01000002	1422500	1526500	104001	1422548	1535558	113011	*Haloplanus* sp. CBA1112	1547898	1547898.3	88.0
Natrinema altunense strain AJ2	222984	222984.5	JNCS01000001	496500	593500	97001	497005	591892	94888	*Natrinema altunense* strain 4.1R	222984	222984.10	98.0
Halobonum sp. NJ-3-1	2743089	2743089.3	CP058579	1918000	2078500	160501	1906722	2078630	171909	*Halobonum* sp. Gai3-2	2743090	2743090.3	85.0
Haloferax sp. Q22	1526048	1526048.3	LOEP01000012	1	56511	56511	1	56511	56511	*Haloferax gibbonsii* strain LR2-5	35746	35746.12	94.8
Haloterrigena sp. SYSU A121-1	2496101	2496101.3	JABURA010000001	1761500	1860500	99001	1761690	1860499	98810	*Haloterrigena turkmenica* DSM 5511	543526	543526.13	92.4
Halogeometricum borinquense strain wsp4	60847	60847.21	CP048739	2797000	2827000	30001	2763703	2853374	89672	*Halogeometricum borinquense* strain wsp3	60847	60847.22	99.6
Halovivax sp. KZCA124	2817025	2817025.3	NZ_CP071597	3085000	3179500	94501	--[f]	--	--	--	--	--	--

a. Genomic contig carrying the IME.

b. Sequence coordinates of the start and end of the IME identified by k-mer spectrum.

c. Sequence coordinates of the start and end of the IME identified by comparative genomics (CG).

d. Closest species not carrying the *fsxA* gene with a completely sequenced genome (used for CG analysis).

e. Average nucleotide identity between the complete genomes of compared species.

f. No species with a completely sequenced genome was similar enough to *Halovivax* sp. KZCA124 to compute an accurate estimate of IME's insertion sites with CG.

Supplementary Table 3 | Synthesized archaeal fusexin genes for fusogenic tests in mammalian cells

Synthesized plasmids	Accession number/ sequence	Species/Assembly source
pBPT01	WP_007110832	*Natrinema altunense*
pBPT02	WP_058826362	*Haloferax* sp.Q22
pBPT03	WP_049937247	*Haloplanus natans*
pBPT04	SAMEA2619974_10776_4	MAG/assembled metagenome
pBPT05	LKMP01000007_1	Nanohaloarchaea archaeon B1-Br10_U2g21 LB-BRINE-C121
pBPT06	3300014206-Ga0172377-10000119-870930-129	MAG/assembled metagenome
pBPT07	3300014208-Ga0172379-10000243-871512-158	MAG/assembled metagenome
pBPT08	3300014208-Ga0172379-10001592-871560-40	MAG/assembled metagenome
pBPT10	3300018015-Ga0187866_1000629|915963_9	MAG/assembled metagenome
pBPT11	3300000868-JGI12330J12834-1000008-299010-8 , **FsxA**	MAG/assembled metagenome

Supplementary Table 4 | Plasmids used in this study

Plasmid name	Description	Use	Source
pLJFX11B	pLJ6-FsxA$_E$	For SAXS and SEC-MALS (Extended Data Fig. 1)	This study, amplified from pBPT11
pLJFX11B_T369C	pLJ6-FsxA$_E$-T369C mutant	For crystallographic study (Figs. 1-2; Extended Data Fig. 3-5)	This study, amplified from pBPT11
pBPT01	*Natrinema altunense fsxA* synthesized into pGene/V5-His	Inducible expression in mammalian cells (Supplementary Fig. 2)	This study
pBPT02	*Haloferax* sp. Q22 *fsxA* synthesized into pGene/V5-His	Inducible expression in mammalian cells (Supplementary Fig. 2)	This study
pBPT03	*Haloplanus natans fsxA* synthesized into pGene/V5-His	Inducible expression in mammalian cells (Supplementary Fig. 2)	This study
pBPT04	SAMEA2619974 synthesized into pGene/V5-His	Inducible expression in mammalian cells (Supplementary Fig. 2)	This study
pBPT05	LKMP01000007_1 synthesized into pGene/V5-His	Inducible expression in mammalian cells (Supplementary Fig. 2)	This study
pBPT06	3300014206 synthesized into pGene/V5-His	Inducible expression in mammalian cells (Supplementary Fig. 2)	This study
pBPT07	3300014208-158 synthesized into pGene/V5-His	Inducible expression in mammalian cells (Supplementary Fig. 2)	This study
pBPT08	3300014208-40 synthesized into pGene/V5-His	Inducible expression in mammalian cells (Supplementary Fig. 2)	This study
pBPT10	3300018015 synthesized into pGene/V5-His	Inducible expression in mammalian cells (Supplementary Fig. 2)	This study
pBPT11	*fsxA* synthesized into pGene/V5-His	Inducible expression in mammalian cells (Supplementary Fig. 2)	This study
pGene/V5-His	pGene/V5-His	GeneSwitch™ inducible Mammalian Expression	INVITROGEN
pSwitch	pSwitch	Regulatory vector for Mifepristone induction	INVITROGEN
pOA34	pGene::EFF-1-V5	*C. elegans eff-1* fused to a C-terminal V5 tag (EFF-1-V5) in pGene	Avinoam et al., 2011[41]
pCI H2B-RFP	pCI::H2B-RFP	A CAG promoter (CMV immediate early enhancer and chicken beta actin promoter) and IRES controlled Histone2B-mRFP1 reporter.	Addgene plasmid # 92398[42]
pCI H2B-GFP	pCI::H2B-GFP	A CAG promoter (CMV immediate early enhancer and chicken beta actin promoter) and IRES controlled Histone2B-EGFP reporter.	Addgene plasmid # 92399[42]
pCAGIG	pCAGIG	A CAG promoter (CMV immediate early enhancer and chicken beta actin promoter) and IRES controlled EGFP reporter.	Addgene plasmid # 11159[43]
pNB25	pCAGIGnes	Intermediate construct to create pNB32	This study
pNB32	pCI::GFPnes	Content-mixing, Fig. 3a-c	This study
pRFPnes	DsRed2 with a nuclear export signal	Content-mixing, Fig. 3a-c	Avinoam et al., 2011[41]
pXL27	pCI::FsxA-V5::H2B-RFP	Content-mixing, Fig. 3a-c	This study
pXL28	pCI::FsxA-V5::H2B-GFP	Content-mixing, Fig. 3a-c; live imaging of fusion, Fig. 3g	This study
pXL29	pCI::AtHAP2-V5::H2B-RFP	Content-mixing, Fig. 3a-c; Multinucleation assay (Supplementary Fig. 2)	This study
pXL30	pCI::AtHAP2-V5::H2B-GFP	Content-mixing, Fig. 3a-c; Multinucleation assay (Supplementary Fig. 2)	This study
pXL49	pCI::FsxA-V5::GFPnes	Content-mixing, Fig. 3d-e	This study

Plasmid name	Description	Use	Source
pNB34	pCI::EFF-1-V5::GFPnes	Content-mixing, Fig. 3d-e	This study
pXL68	pCI::VSV-G::GFPnes	Content-mixing, Fig. 3d-e	This study
pOA19	pCAGGS::EFF-1-V5	Surface biotinylation of EFF-1 (Extended Data Fig. 7)	Avinoam et al., 2011[41]
pXL50	pCAGGS::FsxA-V5	Surface biotinylation of FsxA (Extended Data Fig. 7)	This study
myr-mCherry	myr-mCherry	mCherry linked to a myristoylated and palmitoylated peptide, live imaging of fusion, Fig. 3g	Dunsing et al., Sci. Rep. 2018 [44]
myr-EGFP	myr-EGFP	EGFP linked to a myristoylated and palmitoylated peptide	Dunsing et al., Sci. Rep. 2018 [44]
pXL21	pCI::NaFsxA-V5::H2B-RFP	Multinucleation assay (Supplementary Fig. 2)	This study
pXL22	pCI::NaFsxA-V5::H2B-GFP	Multinucleation assay (Supplementary Fig. 2)	This study
pXL23	pCI::HQ22FsxA-V5::H2B-RFP	Multinucleation assay (Supplementary Fig. 2)	This study
pXL24	pCI::HQ22FsxA-V5::H2B-GFP	Multinucleation assay (Supplementary Fig. 2)	This study
pXL25	pCI::HnFsxA-V5::H2B-RFP	Multinucleation assay (Supplementary Fig. 2)	This study, subcloned from pBPT03 with modification of complete signal peptide
pXL26	pCI::HnFsxA-V5::H2B-GFP	Multinucleation assay (Supplementary Fig. 2)	This study, subcloned from pBPT03 with modification of complete signal peptide
pXL57	pCI::FsxA-ΔFL-AG$_4$A::GFPnes	Content-mixing, Fig. 3j	This study
pXL58	pCI::FsxA-ΔFL-AG$_4$A::H2B-RFP	Content-mixing, Fig. 3j	This study
pXL63	pCI::FsxA-ΔDIV-EFF-1-stem::H2B-RFP	Content-mixing, Fig. 3j	This study
pXL64	pCI::FsxA-ΔDIV-EFF-1-stem::GFPnes	Content-mixing, Fig. 3j	This study
pXL108	pCAGGS::FsxA-ΔDIV-EFF-1-stem	Surface biotinylation of FsxA-ΔDIV-EFF-1-stem mutant (Extended Data Fig. 7)	This study
pXL82	pCI::FsxA-WT-FLAG-3TMs::H2B-RFP	Surface expression tests, FLAG tag inserted before the first TM segment of FsxA (Supplementary Fig. 4)	This study
pXL86	pCI::FsxA-ΔFL-AG$_4$A-FLAG-3TMs::H2B-RFP	Surface expression tests, FLAG tag inserted before the first TM segment of FsxA-ΔFL-AG$_4$A mutant (Supplementary Fig. 4)	This study
pXL92	pCI::FsxA-ΔDIV-EFF-1-stem-FLAG-3TMs::H2B-RFP	Surface expression tests, FLAG tag inserted before the first TM segment of FsxA-ΔDIV-EFF-1-stem mutant (Supplementary Fig. 4)	This study
pOA20	pCAGGS::AFF-1-FLAG	*C. elegans aff-1* fused to a C-terminal FLAG tag (AFF-1-FLAG) in pCAGGS	Avinoam et al., 2011[41]
pXL100	pCI::AFF-1-FLAG::H2B-RFP	Surface expression tests of AFF-1 (Supplementary Fig. 4)	This study
pXL106	pCAGGS::FsxA-ΔFL-AG$_4$A	Surface biotinylation of FsxA-ΔFL-AG$_4$A mutant (Extended Data Fig. 7)	This study

Supplementary Table 5 | Primers used in this study

Primer name	Sequence	Description
SNFX11_F	CGTAGCTGAAACCGGTGATTCAATCACGTATAACTCTGG	Forward primer for cloning FsxA$_E$ and T369C mutant into pLJ6 with AgeI
SNFX11_R	GGTGATGGTGCTCGAGGGAACCAGAACCTCCGAA	Reverse primer for cloning FsxA$_E$ and T369C mutant into pLJ6 with XhoI
SNFX11_T369C_F	ACTGTAtgcGCCACCGTTGAGAATGTC	Forward primer for T369C mutant
SNFX11_T369C_R	GGTGGCgcaTACAGTTCCCTCATCTCCCTC	Reverse primer for T369C mutant
seq_up	GCTGGTTGTTGTGCTGTCTCATC	Sequencing primer for pLJ6
seq_down	CACCAGCCACCACCTTCTGATAG	Sequencing primer for pLJ6
LXH1	GATGGTGCGATTGCGGAT	Sequencing primer for pBPT01
LXH2	CCTACGAGAATGGGCAGA	Sequencing primer for pBPT01
LXH3	TTGCTGGCAGAGAAATGA	Sequencing primer for pBPT02
LXH4	TGATGTACCCCGAGTTCA	Sequencing primer for pBPT02
LXH5	GGATGAAATCTTCAGAAC	Sequencing primer for pBPT03
LXH6	ACTGTCTCGAAGCCGGTT	Sequencing primer for pBPT03
LXH7	CAAAATCACCCTCACATC	Sequencing primer for pBPT04
LXH8	CCTACAATATTAAGTTGTG	Sequencing primer for pBPT04
LXH9	TGAGTCTGAATGGATTAT	Sequencing primer for pBPT05
LXH10	TAGGACTACAGCGAAAAT	Sequencing primer for pBPT05
LXH11	TTGGGGAGGAAATGTAAA	Sequencing primer for pBPT06
LXH12	TAGAAGAATAAATATTCC	Sequencing primer for pBPT06
LXH13	TCCTCTTCCCTCGGAGAA	Sequencing primer for pBPT07
LXH14	CCTACTCAGGTAACGTAA	Sequencing primer for pBPT07
LXH15	CAGTAACAATAAATGGTG	Sequencing primer for pBPT08
LXH16	CAGAAGAATAAACATTCC	Sequencing primer for pBPT08
LXH17	AACAATAGGACAAGCAAA	Sequencing primer for pBPT10
LXH18	ACCAAAAATATTGTCTGC	Sequencing primer for pBPT10
LXH19	AGCATACATAGACAACCC	Sequencing primer for pBPT11, FsxA
LXH20	ACGTCGATGCCGGAGAAA	Sequencing primer for pBPT11, FsxA
pGene FW	CTGCTCAACCTTCCTATC	pGene backbone sequencing forward primer
pGene REV	TTAGGAAAGGACAGTGGGAGTG	pGene backbone sequencing reverse primer
PCA-5	GGTTCGGCTTCTGGCGTGTGACC	pCI::H2B-RFP/H2B-GFP/GFPnes backbones sequencing forward primer

IRES-REV	GCATTCCTTTGGCGAGAG	pCI::H2B-RFP/H2B-GFP/GFPnes backbones sequencing reverse primer
pCAGGS FW	GCAACGTGCTGGTTGTTGTGCTGTC	pCAGGS backbone sequencing forward primer
LXH24	CGGGGTACCATGAGACGTGCAGCATTG	Forward primer for cloning FsxA and its mutants into pCAGGS vector with KpnI
LXH42	CTAGCTAGCGGTACCATGAGACGTGCAGCATTGATT	Forward primer for cloning FsxA into pCI::H2B-RFP/H2B-GFP/GFPnes vectors with NheI and KpnI
LXH44	CTAGCTAGCGGTACCATGGAACCGCCGTTTGAGTGG	Forward primer for cloning EFF-1 into pCI::GFPnes vector with NheI and KpnI
LXH45	CTAGCTAGCGGTACCATGGTGAACGCGATTTTAATG	Forward primer for cloning AtHAP2 into pCI::H2B-RFP/GFP vectors with NheI and KpnI
LXH79	TCCCCCGGGCTAATGGTGATGGTGATGATGACC	Reverse primer for cloning fusexins into pCI vectors with SmaI which binds to 6xHis tag
LXH81	CTAGCTAGCTCAATGGTGATGGTGATGATGACC	Reverse primer for cloning FsxA and its mutants into pCAGGS vector with NheI
LXH111	CTAGCTAGCATGAAGTGCCTTTTGTACTTAG	Forward primer for cloning VSV-G into pCI::GFPnes vector with NheI
LXH112	TCCCCCGGGTTACTTTCCAAGTCGGTTCATC	Reverse primer for cloning VSV-G into pCI::GFPnes vector with SmaI
LXH39	CTAGCTAGCGGTACCATGCGGGCGGTGTCTGATTTC	Forward primer for cloning NaFsxA into pCI::H2B-RFP/GFP vectors with NheI and KpnI
LXH40	CTAGCTAGCGGTACCATGAAAAACGGGTTGAAGGCC	Forward primer for cloning HQ22FsxA into pCI::H2B-RFP/GFP vectors with NheI and KpnI
LXH134	CTAGCTAGCATGGTGAAACGAGTGGGTAATTGTTGG AAGGCCTCAGTAGCGGCATTCTTCCTTCTCATGTTCACTGCATTT	Forward primer for cloning HnFsxA into pCI::H2B-RFP/GFP vectors with NheI; containing modification of complete signal peptide
LXH107	GCGGAAGGTACAGCAGGTACGCCGGTGGAGGTGGAGCTGATTACGAGATCTATTGTTT	Forward primer for cloning downstream of FsxA-ΔFL-AG$_{4}$A mutant by overlap PCR
LXH108	AAACAATAGATCTCGTAATCAGCTCCACCTCCACCGGCGTACCTGCTGTACCTTCCGC	Reverse primer for cloning upstream of FsxA-ΔFL-AG$_{4}$A mutant by overlap PCR
LXH101	ACCGGTATCCAGCAGGAAATCGATCTTGTT	Forward primer 1 for cloning FsxA-ΔDIV-EFF-1-stem by overlap PCR
LXH102	AACAAGATCGATTTCCTGCTGGATACCGGT	Reverse primer 1 for cloning FsxA-ΔDIV-EFF-1-stem by overlap PCR
LXH103	ATGATTGCTACGGATCAGGACGATGATTCA	Forward primer 2 for cloning FsxA-ΔDIV-EFF-1-stem by overlap PCR
LXH104	TGAATCATCGTCCTGATCCGTAGCAATCAT	Reverse primer 2 for cloning FsxA-ΔDIV-EFF-1-stem by overlap PCR
LXH128	TGTTCGGAGGTTCTGGTTCCGACTACAAGGACGACGATGACAAAGGAGATCTGCTTAC	Forward primer for cloning downstream of FsxA-WT/mutants-FLAG by overlap PCR
LXH129	GGAACCAGAACCTCCGAACA	Reverse primer for cloning upstream of FsxA-WT/mutants-FLAG by overlap PCR
LXH135	CTAGCTAGCATGGTACTGTGGCAATGGTCAATAG	Forward primer for cloning *C. elegans* AFF-1-FLAG into pCI::H2B-RFP vector with NheI
LXH136	TCCCCCGGGTTATTTGTCATCGTCGTCCTTGTAGTC	Reverse primer for cloning *C. elegans* AFF-1-FLAG into pCI::H2B-RFP vector with SmaI

Supplementary Video Legends

Supplementary Video 1 | Time-lapse experiment using spinning disk confocal microscopy reveals merging of two cells expressing myr-mCherry and FsxA. Time in hours:minutes. Merge of the red and DIC channels is shown.

Supplementary Video 2 | Z-series of the binucleated BHK cell from Fig. 3h. Labeled nuclei (blue) and myr-mCherry (white). Each optical section obtained with spinning disc confocal microscopy is 1 μm apart.

Supplementary References

1. Lu, B. & Leong, H. W. Computational methods for predicting genomic islands in microbial genomes. *Comput. Struct. Biotechnol. J.* **14**, 200–206 (2016).
2. Sievers, A. *et al.* K-mer Content, Correlation, and Position Analysis of Genome DNA Sequences for the Identification of Function and Evolutionary Features. *Genes* **8**, 122 (2017).
3. Zhou, F., Olman, V. & Xu, Y. Barcodes for genomes and applications. *BMC Bioinformatics* **9**, 546 (2008).
4. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* **39**, 1–38 (1977).
5. Bize, A. *et al.* Exploring short k-mer profiles in cells and mobile elements from Archaea highlights the major influence of both the ecological niche and evolutionary history. *BMC Genomics* **22**, 186 (2021).
6. Davis, J. J. *et al.* The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res.* **48**, D606–D612 (2020).
7. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
8. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
9. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–41 (2004).
10. Makarova, K. S., Sorokin, A. V., Novichkov, P. S., Wolf, Y. I. & Koonin, E. V. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct* **2**, 1–20 (2007).
11. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life* **5**, 818–840 (2015).
12. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

13. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
14. Contreras-Moreira, B. & Vinuesa, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* **79**, 7696–7701 (2013).
15. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
16. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
17. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
18. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
19. Pittis, A. A. & Gabaldón, T. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* **531**, 101–104 (2016).
20. Vosseberg, J. *et al.* Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat. Ecol. Evol.* **5**, 92–100 (2021).
21. Valansi, C. *et al.* Arabidopsis HAP2/GCS1 is a gamete fusion protein homologous to somatic and viral fusogens. *J. Cell Biol.* **216**, 571–581 (2017).
22. Elias, M., Brighouse, A., Gabernet-Castello, C., Field, M. C. & Dacks, J. B. Sculpting the endomembrane system in deep time: high resolution phylogenetics of Rab GTPases. *J. Cell Sci.* **125**, 2500–2508 (2012).
23. Hauser, M., Mayer, C. E. & Söding, J. kClust: fast and sensitive clustering of large protein sequence databases. *BMC Bioinformatics* **14**, 248 (2013).
24. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
25. Le, S. Q., Dang, C. C. & Gascuel, O. Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. *Mol. Bio. Evol.* **29**, 2921–2936 (2012).

26. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
27. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
28. Vosseberg, J. *et al.* Data for: Timing the origin of eukaryotic cellular complexity with ancient duplications. (2019) doi:10.6084/m9.figshare.10069985.v1.
29. Remmert, M., Biegert, A., Hauser, A. & Soding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).
30. Lu, X. *et al.* Double Lock of a Human Neutralizing and Protective Monoclonal Antibody Targeting the Yellow Fever Virus Envelope. *Cell Rep.* **26**, 438–446.e5 (2019).
31. Voss, J. E. *et al.* Glycoprotein organization of Chikungunya virus particles revealed by X-ray crystallography. *Nature* **468**, 709–712 (2010).
32. Modis, Y., Ogata, S., Clements, D. & Harrison, S. C. A ligand-binding pocket in the dengue virus envelope glycoprotein. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 6986–6991 (2003).
33. Gibbons, D. L. *et al.* Conformational change and protein-protein interactions of the fusion protein of Semliki Forest virus. *Nature* **427**, 320–325 (2004).
34. Rey, F. A., Heinz, F. X., Mandl, C., Kunz, C. & Harrison, S. C. The envelope glycoprotein from tick-borne encephalitis virus at 2 Å resolution. *Nature* **375**, 291–298 (1995).
35. Fédry, J. *et al.* The Ancient Gamete Fusogen HAP2 Is a Eukaryotic Class II Fusion Protein. *Cell* **168**, 904–915.e10 (2017).
36. DuBois, R. M. *et al.* Functional and evolutionary insight from the crystal structure of rubella virus protein E1. *Nature* **493**, 552–556 (2013).
37. Guardado-Calvo, P. *et al.* Mechanistic Insight into Bunyavirus-Induced Membrane Fusion from Structure-Function Analyses of the Hantavirus Envelope Glycoprotein Gc. *PLoS Pathog.* **12**, e1005813 (2016).
38. Dessau, M. & Modis, Y. Crystal structure of glycoprotein C from Rift Valley fever virus. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 1696–1701 (2013).

39. Gouet, P., Robert, X. & Courcelle, E. ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res.* **31**, 3320–3323 (2003).
40. Söding, J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**, 951–960 (2004).
41. Avinoam, O. *et al.* Conserved eukaryotic fusogens can fuse viral envelopes to cells. *Science* **332**, 589–592 (2011).
42. Williams, R. M. *et al.* Genome and epigenome engineering CRISPR toolkit for in vivo modulation of cis-regulatory interactions and gene expression in the chicken embryo. *Development* **145**, (2018).
43. Matsuda, T. & Cepko, C. L. Electroporation and RNA interference in the rodent retina in vivo and in vitro. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 16–22 (2004).
44. Dunsing, V. *et al.* Optimal fluorescent protein tags for quantifying protein oligomerization in living cells. *Sci. Rep.* **8**, 1–12 (2018).