Title:

Discovery of archaeal Fusexins homologous to eukaryotic HAP2/GCS1 gamete fusion proteins

Authors:

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], Martin Graña[7,*], Héctor Romero[6,12,*], Pablo S. Aguilar[1,13,*], Luca Jovine[4,*], Benjamin Podbilewicz[5,*]

Affiliations:

[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.

***Corresponding authors. Email: mgrana@pasteur.edu.uy, eletor@fcien.edu.uy, paguilar@iib.unsam.edu.ar, luca.jovine@ki.se, podbilew@technion.ac.il**

†These authors contributed equally to this work.

Abstract:

Sexual reproduction consists of genome reduction by meiosis and subsequent gamete fusion. Presence of meiotic genes in prokaryotes suggests that DNA repair mechanisms evolved toward meiotic recombination; however, fusogenic proteins resembling those found in eukaryotes were not identified in prokaryotes. Here, we identify archaeal proteins that are homologs of fusexins, a superfamily of fusogens that mediate eukaryotic gamete and somatic cell fusion, as well as virus entry. The crystal structure of a trimeric archaeal Fusexin1 reveals novel features such as a six-helix bundle and an additional globular domain. Ectopically expressed Fusexin1 can fuse mammalian cells, and this process involves the additional domain and a conserved fusion loop. Archaeal fusexin genes exist within integrated mobile elements, potentially linking ancient archaeal gene exchanges and eukaryotic sex.

One-Sentence Summary:

Cell membrane fusion proteins of viruses and eukaryotes are also present in archaea.

Main text:

How early eukaryotes developed the capacity for gamete fusion is a central question entangled with the origin of the eukaryotic cell itself. The widespread presence of a conserved set of meiosis, gamete and nuclear fusion proteins among extant eukaryotes indicates that meiotic sex emerged once, predating the last eukaryotic common ancestor (LECA) (*1*, *2*). The conserved gamete fusion protein HAP2/GCS1 belongs to a superfamily of fusion proteins (fusogens) 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 (*6*, *7*); EFF-1 and AFF-1

(somatic fusexins) that promote cell fusion during syncytial organ development (*8–11*); and HAP2/GCS1 (sexual fusexins; hereafter referred to as HAP2) that mediate gamete fusion (*12–14*). Although it is assumed that sexual fusexins were already present in the LECA (*1*, *15*), their shared ancestry with viral fusexins posed a "the virus or the egg" evolutionary dilemma (*16*). In one scenario, fusexins are proper eukaryal innovations that were captured by some viruses and used for host invasion. Alternatively, a viral fusexin gene was transferred to an early eukaryotic cell and then repurposed for gamete fusion.

Here, we identify a fourth family of fusexins in genomes of Archaea and prokaryotic fractions of metagenomes from very diverse environments. We provide structural and functional evidence indicating that these proteins are cellular fusogens. Genomic analyses indicate that archaeal fusexins are carried by integrated mobile genetic elements. Evolutionary analyses of the whole fusexin superfamily reveal new working models for the emergence of meiotic sex during eukaryogenesis.

Results

Fusexin genes in Archaea

To search for fusexins we used the crystallographic structures of *C. reinhardtii*, *A. thaliana,* and *T. cruzi* HAP2 (Cr/At/TcHAP2) (*4*, *17*, *18*) to build dedicated Hidden Markov Models (HMMs). These were used to scan the Uniclust30 database with HHblits (see Methods in Supplementary Materials). We detected 24 high-confidence candidates in prokaryotes: eight belong to isolated and cultivated archaea, and the remaining sixteen to metagenome-assembled genomes (MAGs, table S1). We then built HMMs of the candidate ectodomains and compared them to HMMs of sexual, somatic and viral fusexins. Fig. 1A shows that the prokaryotic candidates have detectable sequence similarities with HAP2, with E-values below 0.001 and HHblits-derived probabilities

higher than 0.95 (fig. S1). We named these proteins Fusexin1 (Fsx1). *fsx1* genes found in cultivated and isolated prokaryotes are restricted to the Halobacteria class (Euryarchaeota superphylum) whereas MAGs containing Fsx1s include all major Archaea superphyla (table S1). Next, we used this Fsx1 sequence set to search the Metaclust database, which comprises 1.59 billion clustered proteins from over 2200 metagenomic/metatranscriptomic datasets. Performing a scan pipeline using PHMMER, PSI-BLAST, HMM-HMM comparisons and topology filtering we found 96 high-confidence *fsx1* genes. The identified *fsx1*s come from different environments (with preeminence of saline samples) and a wide temperature range (-35 to 80ºC, data S1).

Fsx1 is a structural homolog of HAP2/GCS1

To experimentally investigate the presence of fusexin-like proteins in Archaea, a selection of the candidate genes was expressed in mammalian cells (fig. S2A, B). High-level expression was observed for a metagenomic Fsx1 sequence from a hypersaline environment, predicted to encode a ~55 kDa ectodomain region ($Fsx1_E$) followed by three transmembrane domains (TMs) (data S1). Crystals of $Fsx1_E$, obtained in the presence of 2.5 M NaCl, 0.2 M $CaCl_2$, yielded data to 2.3 Å resolution that was phased by molecular replacement using a combination of fragments from models generated by AlphaFold2 (*19*) (figs. S3 and S4; table S2).

$Fsx1_E$ is a monomer in solution but crystallized as a homotrimer of 119x78x75 Å (Fig. 1B and figs. S3 and S5). Each protomer consists of four domains, the first three of which match the approximate dimensions and relative arrangement of domains I-III of known fusexins in their post-fusion conformation (*20*); accordingly, fold and interface similarity searches identify HAP2 as the closest structural homolog of $Fsx1_E$, followed by viral fusexins and *C. elegans* EFF-1 (Fig. 1B, C and fig. S6). Fsx1 domains I and III are relatively sequence-conserved among archaeal homologs (figs. S7A and S8A) 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; figs. S5C and S6). On the other hand, Fsx1 domain II shares the same topology as that of HAP2 but differs significantly in its secondary structure elements and their relative orientation, as well as disulfide bonds (fig. S6C). In particular, Fsx1 domain II is characterized by a four-helix hairpin, whose N-terminal half 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 2, A to C, and figs. S5A and S7, B and C).

Notably, unlike previously characterized viral and eukaryotic fusexins, Fsx1 also contains a fourth globular domain conserved among archaeal homologs (Fig. 1, B and C, and figs. S5D, S6 and S8). Its antiparallel β-sandwich, which includes the two C-terminal disulfides of Fsx1, resembles the carbohydrate-binding fold of dust mite allergen Der p 23 and related chitin-binding proteins (*21*) (Fig. 2D); accordingly, it is also structurally similar to a high-confidence AlphaFold2 model of the C-terminal domain of acidic mammalian chitinase (*22*). 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).

The $Fsx1_E$ monomer has a net charge of -67, and another 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. 2A, B). 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. 2 and fig. S7D). Strikingly, the position of this element matches that of the fusion loops (FLs) of other fusexins, including the Ca^{2+}-binding fusion surface of rubella virus E1 protein (*23*, *24*) (Fig. 2E).

Moreover, as previously observed in the case of CrHAP2 (*18*), the loops of each trimer interact with those of another trimer within the Fsx1 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 Fsx1 and viral or eukaryotic fusexins suggests that this prokaryotic molecule also functions to fuse membranes.

Fsx1 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*, *10*, *11*). Cells with either red or green nuclei are mixed with each other and fusion is measured by the formation of hybrid cells with both red and green nuclei revealing merger of their cytoplasms. For this, we co-cultured two batches of Baby Hamster Kidney (BHK) cells independently transfected with Fsx1 and co-expressing either nuclear H2B-RFP or H2B-GFP (*3*). We then performed immunofluorescence against a V5 tag fused to the cytoplasmic tail of Fsx1 (Fig. 3A and B, and fig. S9). We observed a five-fold increase in the mixing of the nuclear H2B-GFP and H2B-RFP compared to vector control, showing that Fsx1 is a *bona fide* fusogen, as efficient as the fusexin AtHAP2 (Fig. 3C). To determine whether Fsx1 expression is required in both fusing cells or, alternatively, it suffices in one of the fusing partners, we mixed BHK-Fsx1 coexpressing cytoplasmic GFP with BHK cells expressing only nuclear RFP. We found increased multinucleation of GFP+ cells (revealing cell-cell fusion) but very low mixing with RFP+ cells not expressing Fsx1. In contrast, the vesicular stomatitis virus G-glycoprotein (VSVG) fusogen induced efficient unilateral fusion (*11*) (Fig. 3D-F). Thus, Fsx1 acts in a bilateral way, similarly to EFF-1 and AFF-1 fusexins (*11*, *25*, *26*). We then performed live imaging using

spinning disk confocal microscopy and observed cell-cell fusion of BHK-Fsx1 cells (Fig. 3, G and H).

Structure-function analysis of Fsx1

To compare archaeal Fsx1 activity with fusexins from eukaryotes and viruses, we introduced mutations into three different structural domains of Fsx1 and tested surface expression and fusogenic activity in mammalian cells.

First, to test whether the putative FL of Fsx1 (143-SVTSPV-148) is involved in fusion, we replaced it with a linker of 4G between Y142A and Y149A (Figs. 2E and 3I, and fig. S8; ΔFL→AG$_4$A). This replacement does not affect surface expression yet reduces content mixing to levels similar to those of the negative control (Fig. 3J and fig. S10).

Second, we asked whether domain IV, which is only present in archaeal fusexins, has a function in the fusion process. For this, we replaced the entire domain with the stem region of EFF-1 (Fig. 3I; ΔDIV→EFF-1 stem). While this mutant Fsx1 reaches the cell surface, suggesting that it folds normally, it shows a significantly reduced activity compared to wildtype Fsx1 (Fig. 3J and fig. S10).

Third, to test whether the three TMs of Fsx1 are required for fusion, we replaced them with the TM and cytoplasmic domains of EFF-1 (Fig. 3I; ΔTMs→EFF-1 TM) or a glycosylphosphatidylinositol (GPI) anchor signal (Fig. 3I; ΔTMs→GPI). We found that both Fsx1 mutants remained active (Fig. 3J), indicating that the Fsx1 TMs are not essential for fusion. Finally, we also replaced the TM and cytoplasmic domains of AtHAP2 with a signal for GPI and found that the protein also maintained its fusogenic activity (Fig. 3, C, I, and J). Thus, contrary to some viral fusogens in which the GPI-anchored glycoproteins fail to drive complete fusion (*27–29*),

lipid-anchored Fsx1 or eukaryotic HAP2 promote syncytia formation when expressed on the surface of BHK cells.

Fsx1s are ancient fusogens associated with integrated mobile elements

The sparse pattern of Fsx1 presence in Archaea led us to perform genomic comparisons of related species with and without the *fsx1* gene. These revealed >50 kbp DNA insertions in the genomes of species with *fsx1* genes (fig. S12). To investigate them, we performed k-mer spectrum analysis on *fsx1*-containing Pure Culture Genomes (PCGs) and found divergent regions containing the *fsx1* ORF (Fig. 4A and fig. S13). Gene content analyses of *fsx1*-containing regions (fig. S14) show that they share a portion of their genes (fig. S15) and display conserved synteny (Fig. 4B and fig. S16), suggesting common ancestry. These regions are enriched in ORFs homologous to proteins involved in DNA mobilization and integration (Fig. 4B and table S3). Thus, our results indicate that *fsx1* genes are contained in integrated mobile elements (IMEs) that can be mobilized by a conjugative-like, cell fusion-dependent mechanism.

To describe Fsx1's tempo and mode of evolution we first compared archaeal and sexual fusexins, which share enough sequence conservation to apply standard phylogenetic analyses, not possible for somatic and viral fusexins (Fig. 1A and fig. S1). We built maximum likelihood (ML) phylogenies for a set of Fsx1 sequences derived from isolated species and metagenomes, and a subset of HAP2s encompassing a wide breadth of eukaryotic lineages. Two major findings come from these phylogenies. First, the branching pattern of Fsx1 sequences from complete genomes is incompatible with their species tree (fig. S17), supporting a history of Horizontal Gene Transfer (HGT) events within Archaea, in line with *fsx1s* presence in IMEs. Second, eukaryal fusexins still cluster into a strongly supported clade suggesting they diverged before LECA.

To place *fsx1* in the broader fusexin superfamily context, we performed structural phylogenetic analysis comparing crystal structures from viral, somatic and eukaryotic gamete fusogens (Fig. 4C). This structure-based tree supports a viral origin of somatic fusexins (EFF-1) (*10*) and is also compatible with archaeal fusexins appearing before the radiation of eukaryotes.

Discussion

The archaeal proteins herein identified place fusexins in yet another domain of life, with different membrane chemistry and along a wide niche landscape, from cold hypersaline lakes to hot springs and hydrothermal vents (fig. S17).

Our structural and functional analyses show that Fsx1 has both conserved and diverged properties when compared to eukaryotic and viral fusexins (Fig. 3 and fig. S6). Like its viral counterparts, Fsx1 has an uncharged loop required for fusion. However, unlike previously known fusexins, Fsx1 harbors an additional domain (IV) involved in fusogenic activity that may bind sugars (Figs. 2D and 3J). Considering that cell surface glycosylation was found to be important for fusion-based mating of halophilic archaea (*30*), this domain may actively promote fusion by interacting with carbohydrates attached to lipids or proteins such as S-layer glycoproteins (*31*). Unlike HAP2s, Fsx1 homologs have 1-4 TMs and a variable Cys number (5 to 30, see data S1). Like eukaryotic fusexins, Fsx1 mediates BHK cell fusion in a bilateral fashion (Fig. 3F). However, in contrast to viral fusogens (*27–29*, *32*, *33*), the fusion activity of Fsx1 is maintained following substitution of its three TMs with a single TM or a GPI anchor. Since GPI-anchored AtHAP2 is also fusogenic, other fusexins may also drive complete cell fusion without a specific involvement of TMs. Future studies will address the function of the six-helix bundle formed by Fsx1 domain II, which is unprecedented among fusexins and raises an unexpected structural connection with class I viral fusogens (*6*, *7*).

The "virus or the egg" dilemma (*16*) posits that fusexins may have been viral innovations (class II fusogens), later acquired by eukaryotes, or vice versa. Archaeal fusexins expand this dilemma: the origins of gamete fusion may have had a prokaryotic player. Both structure- and sequence-based trees (Fig. 4C and fig. S17) do not solve but provide insights to address this predicament, in which we distinguish three main alternative origins: viral, eukaryotic and archaeal.

Although exaptation of viral genes by eukaryotic organisms is well documented (*34*), known viral fusexins are restricted to very few divergent lineages infecting multicellular eukaryotes, suggesting that different viruses captured fusexin genes in different occasions. The Virus-first hypothesis requires more than one HGT event from an ancient viral lineage into different domains. Support for the existence of such fusexin-containing viral lineage infecting both archaeal and eukaryotic hosts is currently lacking (*35*).

The ubiquity of sexual fusexins in Eukarya strongly suggests evolutionary success, in line with the Eukarya-first hypothesis. However, examples of interdomain gene transfers from eukaryotes to archaeas are rare (*36*). The presence of Fsx1s in Halobacteria IMEs is consistent with HGT in the opposite direction. In addition to the Asgard- and alphaproteobacterial endosymbiont-related inherited genes, the pre-LECA may have received hundreds of archaeal genes from other lineages, including Euryarchaeota (*37*). Halophilic archaea are polyploid (*38*) and undergo HGT events that overcome species and genera barriers (*39*, *40*). Fsx1s seem absent in some archaeal species known to undergo fusion-based gene exchange events (*41*). As genetic exchange in extant prokaryotes is dominated by conjugation, transduction and transformation, we hypothesize that Fsx1s are molecular fossils still playing a role in archaeal 'sex-like' events.

Overall, the pre-LECA divergence between archaeal and eukaryal fusexins, their presence in mobile elements and our current understanding of the role played by HGT during eukaryogenesis (*42–44*) prompt us to consider the Archaea-first as a plausible hypothesis.

Discovery of the Asgard superphylum (*45*) and the recent cultivation of one of its members (*46*) support eukaryogenesis models where populations of bacteria and archaea lived in syntrophy, transferring metabolites and genes (*47*). Acquisition of a *fsx1* gene during the FECA to LECA transition could have enabled pre-LECA cells to undergo genome expansion, explore syncytial forms (*48*) and evolve into mononucleated cells fully equipped for meiosis and gamete fusion (*49*). Our findings raise the possibility that gamete fusion is the product of over two billion years of evolution of this ancient archaeal cell fusion machine.

References

1. D. Speijer, J. Lukeš, M. Eliáš, Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8827–8834 (2015).

2. M. A. Ramesh, S.-B. Malik, J. M. Logsdon 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).

3. C. Valansi, D. Moi, E. Leikina, E. Matveev, M. Graña, L. V. Chernomordik, H. Romero, P. S. Aguilar, B. Podbilewicz, Arabidopsis HAP2/GCS1 is a gamete fusion protein homologous to somatic and viral fusogens. *J. Cell Biol.* **216** (2017), pp. 571–581.

4. J. Fédry, Y. Liu, G. Péhau-Arnaudet, J. Pei, W. Li, M. A. Tortorici, F. Traincard, A. Meola, G. Bricogne, N. V. Grishin, W. J. Snell, F. A. Rey, T. Krey, The Ancient Gamete Fusogen HAP2 Is a Eukaryotic Class II Fusion Protein. *Cell*. **168** (2017), pp. 904–915.e10.

5. J. F. Pinello, A. L. Lai, J. K. Millet, D. Cassidy-Hanley, J. H. Freed, T. G. Clark, Structure-Function Studies Link Class II Viral Fusogens with the Ancestral Gamete Fusion Protein

HAP2. *Curr. Biol.* **27**, 651–660 (2017).

6. S. C. Harrison, Viral membrane fusion. *Virology*. **479-480**, 498–507 (2015).

7. M. Kielian, Mechanisms of Virus Membrane Fusion Proteins. *Annu Rev Virol*. **1**, 171–189 (2014).

8. W. A. Mohler, G. Shemer, J. J. del Campo, C. Valansi, E. Opoku-Serebuoh, V. Scranton, N. Assaf, J. G. White, B. Podbilewicz, The type I membrane protein EFF-1 is essential for developmental cell fusion. *Dev. Cell*. **2**, 355–362 (2002).

9. B. Podbilewicz, E. Leikina, A. Sapir, C. Valansi, M. Suissa, G. Shemer, L. V. Chernomordik, The *C. elegans* developmental fusogen EFF-1 mediates homotypic fusion in heterologous cells and in vivo. *Dev. Cell*. **11**, 471–481 (2006).

10. J. Pérez-Vargas, T. Krey, C. Valansi, O. Avinoam, A. Haouz, M. Jamin, H. Raveh-Barak, B. Podbilewicz, F. A. Rey, Structural basis of eukaryotic cell-cell fusion. *Cell*. **157**, 407–419 (2014).

11. O. Avinoam, K. Fridman, C. Valansi, I. Abutbul, T. Zeev-Ben-Mordehai, U. E. Maurer, A. Sapir, D. Danino, K. Grünewald, J. M. White, B. Podbilewicz, Conserved eukaryotic fusogens can fuse viral envelopes to cells. *Science*. **332**, 589–592 (2011).

12. M. A. Johnson, K. von Besser, Q. Zhou, E. Smith, G. Aux, D. Patton, J. Z. Levin, D. Preuss, Arabidopsis hapless mutations define essential gametophytic functions. *Genetics*. **168**, 971–982 (2004).

13. T. Mori, H. Kuroiwa, T. Higashiyama, T. Kuroiwa, GENERATIVE CELL SPECIFIC 1 is essential for angiosperm fertilization. *Nat. Cell Biol.* **8**, 64–71 (2006).

14. Y. Liu, R. Tewari, J. Ning, A. M. Blagborough, S. Garbom, J. Pei, N. V. Grishin, R. E. Steele, R. E. Sinden, W. J. Snell, O. Billker, The conserved plant sterility gene HAP2 functions after attachment of fusogenic membranes in Chlamydomonas and Plasmodium gametes. *Genes Dev.* **22**, 1051–1068 (2008).

15. J. L. Wong, M. A. Johnson, Is HAP2-GCS1 an ancestral gamete fusogen? *Trends Cell Biol.* **20**, 134–141 (2010).

16. R. W. Doms, What Came First-the Virus or the Egg? *Cell*. **168** (2017), pp. 755–757.

17. J. Fedry, J. Forcina, P. Legrand, G. Péhau-Arnaudet, A. Haouz, M. Johnson, F. A. Rey, T. Krey, Evolutionary diversification of the HAP2 membrane insertion motifs to drive gamete fusion across eukaryotes. *PLoS Biol.* **16**, e2006357 (2018).

18. J. Feng, X. Dong, J. Pinello, J. Zhang, C. Lu, R. E. Iacob, J. R. Engen, W. J. Snell, T. A. Springer, Fusion surface structure, function, and dynamics of gamete fusogen HAP2. *Elife*. **7**, e39772 (2018).

19. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**, 583–589 (2021).

20. Y. Modis, S. Ogata, D. Clements, S. C. Harrison, Structure of the dengue virus envelope protein after membrane fusion. *Nature*. **427**, 313–319 (2004).

21. G. A. Mueller, T. A. Randall, J. Glesner, L. C. Pedersen, L. Perera, L. L. Edwards, E. F. DeRose, M. D. Chapman, R. E. London, A. Pomés, Serological, genomic and structural analyses of the major mite allergen Der p 23. *Clin. Exp. Allergy*. **46**, 365–376 (2016).

22. K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, D. Hassabis, Highly accurate protein structure prediction for the human proteome. *Nature*. **596**, 590–596 (2021).

23. R. M. DuBois, M.-C. Vaney, M. A. Tortorici, R. A. Kurdi, G. Barba-Spaeth, T. Krey, F. A. Rey, Functional and evolutionary insight from the crystal structure of rubella virus protein E1. *Nature*. **493**, 552–556 (2013).

24. M. Dubé, L. Etienne, M. Fels, M. Kielian, Calcium-Dependent Rubella Virus Fusion Occurs in Early Endosomes. *J. Virol.* **90**, 6303–6313 (2016).

25. A. Sapir, J. Choi, E. Leikina, O. Avinoam, C. Valansi, L. V. Chernomordik, A. P. Newman, B. Podbilewicz, AFF-1, a FOS-1-regulated fusogen, mediates fusion of the anchor cell in *C. elegans*. *Dev. Cell*. **12**, 683–698 (2007).

26. T. Gattegno, A. Mittal, C. Valansi, K. C. Q. Nguyen, D. H. Hall, L. V. Chernomordik, B. Podbilewicz, Genetic control of fusion pore expansion in the epidermis of *Caenorhabditis elegans*. *Mol. Biol. Cell*. **18**, 1153–1166 (2007).

27. G. W. Kemble, T. Danieli, J. M. White, Lipid-anchored influenza hemagglutinin promotes hemifusion, not complete fusion. *Cell*. **76**, 383–391 (1994).

28. R. M. Markosyan, F. S. Cohen, G. B. Melikyan, The lipid-anchored ectodomain of influenza virus hemagglutinin (GPI-HA) is capable of inducing nonenlarging fusion pores. *Mol. Biol. Cell*. **11**, 1143–1152 (2000).

29. N. A. Jones, R. J. Geraghty, Fusion activity of lipid-anchored envelope glycoproteins of herpes simplex virus type 1. *Virology*. **324**, 213–228 (2004).

30. Y. Shalev, I. Turgeman-Grott, A. Tamir, J. Eichler, U. Gophna, Cell Surface Glycosylation Is Required for Efficient Mating of *Haloferax volcanii*. *Front. Microbiol.* **8**, 1253 (2017).

31. S. Sivabalasarma, H. Wetzel, P. Nußbaum, C. van der Does, M. Beeby, S.-V. Albers, Analysis of Cell-Cell Bridges in *Haloferax volcanii* Using Electron Cryo-Tomography Reveal a Continuous Cytoplasm and S-Layer. *Front. Microbiol.* **11**, 612239 (2020).

32. D. Odell, E. Wanas, J. Yan, H. P. Ghosh, Influence of membrane anchoring and cytoplasmic domains on the fusogenic activity of vesicular stomatitis virus glycoprotein G. *J. Virol.* **71**, 7996–8000 (1997).

33. S. Tong, R. W. Compans, Oligomerization, secretion, and biological function of an anchor-free parainfluenza virus type 2 (PI2) fusion protein. *Virology*. **270**, 368–376 (2000).

34. E. V. Koonin, M. Krupovic, The depths of virus exaptation. *Curr. Opin. Virol.* **31**, 1–8 (2018).

35. K. El Omari, S. Li, A. Kotecha, T. S. Walter, E. A. Bignon, K. Harlos, P. Somerharju, F. De Haas, D. K. Clare, M. Molin, F. Hurtado, M. Li, J. M. Grimes, D. H. Bamford, N. D. Tischler, J. T. Huiskonen, D. I. Stuart, E. Roine, The structure of a prokaryotic viral envelope protein expands the landscape of membrane fusion proteins. *Nat. Commun.* **10**, 846 (2019).

36. A. Wagner, R. J. Whitaker, D. J. Krause, J.-H. Heilers, M. van Wolferen, C. van der Does, S.-

V. Albers, Mechanisms of gene flow in archaea. *Nat. Rev. Microbiol.* **15**, 492–501 (2017).

37. J. Vosseberg, J. J. E. van Hooff, M. Marcet-Houben, A. van Vlimmeren, L. M. van Wijk, T. Gabaldón, B. Snel, Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat Ecol Evol*. **5**, 92–100 (2021).

38. K. Ludt, J. Soppa, Polyploidy in halophilic archaea: regulation, evolutionary advantages, and gene conversion. *Biochem. Soc. Trans.* **47**, 933–944 (2019).

39. I. Turgeman-Grott, S. Joseph, S. Marton, K. Eizenshtein, A. Naor, S. M. Soucy, A.-E. Stachler, Y. Shalev, M. Zarkor, L. Reshef, N. Altman-Price, A. Marchfelder, U. Gophna, 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).

40. M. Z. DeMaere, T. J. Williams, M. A. Allen, M. V. Brown, J. A. E. Gibson, J. Rich, F. M. Lauro, M. Dyall-Smith, K. W. Davenport, T. Woyke, N. C. Kyrpides, S. G. Tringe, R. Cavicchioli, 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).

41. A. Naor, P. Lapierre, M. Mevarech, R. T. Papke, U. Gophna, Low Species Barriers in Halophilic Archaea and the Formation of Recombinant Hybrids. *Curr. Biol.* **22**, 1444–1448 (2012).

42. E. V. Koonin, Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Res.* **5** (2016), doi:10.12688/f1000research.8737.1.

43. L. Eme, A. Spang, J. Lombard, C. W. Stairs, T. J. G. Ettema, Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* **16**, 120 (2018).

44. T. Gabaldón, Origin and Early Evolution of the Eukaryotic Cell. *Annu. Rev. Microbiol.* **75**, 631–647 (2021).

45. K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, D. Bäckström, L. Juzokaite, E. Vancaester, K. W. Seitz, K. Anantharaman, P. Starnawski, K. U. Kjeldsen, M. B. Stott, T. Nunoura, J. F. Banfield, A. Schramm, B. J. Baker, A. Spang, T. J. G. Ettema, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*. **541**, 353–358 (2017).

46. H. Imachi, M. K. Nobu, N. Nakahara, Y. Morono, M. Ogawara, Y. Takaki, Y. Takano, K. Uematsu, T. Ikuta, M. Ito, Y. Matsui, M. Miyazaki, K. Murata, Y. Saito, S. Sakai, C. Song, E. Tasumi, Y. Yamanaka, T. Yamaguchi, Y. Kamagata, H. Tamaki, K. Takai, Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature*. **577**, 519–525 (2020).

47. M. A. O'Malley, M. M. Leger, J. G. Wideman, I. Ruiz-Trillo, Concepts of the last eukaryotic common ancestor. *Nat Ecol Evol*. **3**, 338–344 (2019).

48. J. Skejo, S. G. Garg, S. B. Gould, M. Hendriksen, F. D. K. Tria, N. Bremer, D. Franjević, N. W. Blackstone, W. F. Martin, Evidence for a Syncytial Origin of Eukaryotes from Ancestral State Reconstruction. *Genome Biol. Evol.* **13** (2021), doi:10.1093/gbe/evab096.

49. D. A. Baum, B. Baum, An inside-out origin for the eukaryotic cell. *BMC Biol.* **12**, 76 (2014).

50. J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, The I-TASSER Suite: protein structure and function prediction. *Nat. Methods*. **12**, 7–8 (2015).

51. E. Baquero, J. Fedry, P. Legrand, T. Krey, F. A. Rey, Species-Specific Functional Regions of the Green Alga Gamete Fusion Protein HAP2 Revealed by Structural Studies. *Structure*. **27**, 113–124.e4 (2019).

52. M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, M. Steinegger, Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176 (2017).

53. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*. **9**, 173–175 (2011).

54. M. Steinegger, J. Söding, Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).

55. F. Sievers, D. G. Higgins, Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* **1079**, 105–116 (2014).

56. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

57. K. D. Tsirigos, C. Peters, N. Shu, L. Kall, A. Elofsson, The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.* (2015), doi:10.1093/nar/gkv485.

58. J. Söding, Protein homology detection by HMM-HMM comparison. *Bioinformatics*. **21**, 951–960 (2005).

59. I. Raj, H. Sadat Al Hosseini, E. Dioguardi, K. Nishimura, L. Han, A. Villa, D. de Sanctis, L. Jovine, Structural Basis of Egg Coat-Sperm Recognition at Fertilization. *Cell*. **169**, 1315–1326.e17 (2017).

60. G. W. Kemble, Y. I. Henis, J. M. White, GPI- and transmembrane-anchored influenza hemagglutinin differ in structure and receptor binding activity. *J. Cell Biol.* **122**, 1253–1265 (1993).

61. R. B. DuBridge, P. Tang, H. C. Hsia, P. M. Leong, J. H. Miller, M. P. Calos, Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. *Mol. Cell. Biol.* **7**, 379–387 (1987).

62. P. Pernot, A. Round, R. Barrett, A. De Maria Antolinos, A. Gobbo, E. Gordon, J. Huet, J. Kieffer, M. Lentini, M. Mattenet, C. Morawe, C. Mueller-Dieckmann, S. Ohlsson, W. Schmid, J. Surr, P. Theveneau, L. Zerrad, S. McSweeney, Upgraded ESRF BM29 beamline for SAXS on macromolecules in solution. *J. Synchrotron Radiat.* **20**, 660–664 (2013).

63. A. Round, F. Felisaz, L. Fodinger, A. Gobbo, J. Huet, C. Villard, C. E. Blanchet, P. Pernot, S. McSweeney, M. Roessle, D. I. Svergun, F. Cipriani, 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).

64. P. V. Konarev, V. V. Volkov, A. V. Sokolova, M. H. J. Koch, D. I. Svergun, PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *J. Appl. Crystallogr.* **36**, 1277–1282 (2003).

65. K. Manalastas-Cantos, P. V. Konarev, N. R. Hajizadeh, A. G. Kikhney, M. V. Petoukhov, D. S. Molodenskiy, A. Panjkovich, H. D. T. Mertens, A. Gruzinov, C. Borges, C. M. Jeffries, D. I. Svergun, D. Franke, ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *J. Appl. Crystallogr.* **54**, 343–355 (2021).

66. D. Svergun, C. Barberato, M. H. J. Koch, CRYSOL– a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J Appl Crystallogr*. **28** (1995), pp. 768–773.

67. D. Franke, D. I. Svergun, DAMMIF, a program for rapid *ab-initio* shape determination in

small-angle scattering. *J. Appl. Crystallogr.* **42**, 342–346 (2009).

68. V. V. Volkov, D. I. Svergun, IUCr, Uniqueness of *ab initio* shape determination in small-angle scattering. *J. Appl. Crystallogr.* **36**, 860–864 (2003).

69. E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, T. E. Ferrin, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

70. D. Nurizzo, T. Mairs, M. Guijarro, V. Rey, J. Meyer, P. Fajardo, J. Chavanne, J. C. Biasci, S. McSweeney, E. Mitchell, The ID23-1 structural biology beamline at the ESRF. *J. Synchrotron Radiat.* **13**, 227–238 (2006).

71. W. Kabsch, XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).

72. A. Vagin, A. Teplyakov, MOLREP: an Automated Program for Molecular Replacement. *J. Appl. Crystallogr.* **30**, 1022–1025 (1997).

73. M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).

74. B. W. Matthews, Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497 (1968).

75. K. A. Kantardjieff, B. Rupp, Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci.* **12**, 1865–1871 (2003).

76. M. AlQuraishi, Machine learning in protein structure prediction. *Curr. Opin. Chem. Biol.* **65**, 1–8 (2021).

77. C. Millán, R. M. Keegan, J. Pereira, M. D. Sammito, A. J. Simpkin, A. J. McCoy, A. N. Lupas, M. D. Hartmann, D. J. Rigden, R. J. Read, Assessing the utility of CASP14 models for molecular replacement. *Proteins*. **89**, 1752–1769 (2021).

78. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

79. T. C. Terwilliger, R. W. Grosse-Kunstleve, P. V. Afonine, N. W. Moriarty, P. H. Zwart, L. W. Hung, R. J. Read, P. D. Adams, Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (2008).

80. G. Chojnowski, J. Pereira, V. S. Lamzin, Sequence assignment for low-resolution modelling of protein crystal structures. *Acta Crystallogr D Struct Biol*. **75**, 753–763 (2019).

81. G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, A. A. Vagin, REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355–367 (2011).

82. A. Casañal, B. Lohkamp, P. Emsley, Current developments in Coot for macromolecular model building of Electron Cryo-microscopy and Crystallographic Data. *Protein Sci.* **29**, 1069–1078 (2020).

83. T. I. Croll, *ISOLDE*: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr D Struct Biol*. **74**, 519–530 (2018).

84. P. V. Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, P. D. Adams, Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).

85. D. Liebschner, P. V. Afonine, M. L. Baker, G. Bunkóczi, V. B. Chen, T. I. Croll, B. Hintze, L. W. Hung, S. Jain, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams, P. D. Adams, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. *Acta Crystallogr D Struct Biol*. **75**, 861–877 (2019).

86. A. Thorn, G. M. Sheldrick, ANODE: anomalous and heavy-atom density calculation. *J. Appl. Crystallogr.* **44**, 1285–1287 (2011).

87. H. Zheng, D. R. Cooper, P. J. Porebski, I. G. Shabalin, K. B. Handing, W. Minor, *CheckMyMetal*: a macromolecular metal-binding validation tool. *Acta Crystallogr. D Biol. Crystallogr.* **73**, 223–233 (2017).

88. C. J. Williams, J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall 3rd, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson, D. C. Richardson, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).

89. E. L. Sonnhammer, G. von Heijne, A. Krogh, A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182 (1998).

90. A. Zemla, LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).

91. L. Holm, Using Dali for Protein Structure Comparison. *Methods Mol. Biol.* **2112**, 29–42 (2020).

92. E. Krissinel, K. Henrick, Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2256–2268 (2004).

93. W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. **22**, 2577–2637 (1983).

94. T. A. P. de Beer, K. Berka, J. M. Thornton, R. A. Laskowski, PDBsum additions. *Nucleic Acids Res.* **42**, D292–6 (2014).

95. K. G. Tina, R. Bhadra, N. Srinivasan, PIC: Protein Interactions Calculator. *Nucleic Acids Res.* **35**, W473–6 (2007).

96. E. Krissinel, K. Henrick, Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).

97. E. Krieger, K. Joo, J. Lee, J. Lee, S. Raman, J. Thompson, M. Tyka, D. Baker, K. Karplus, Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins: Structure, Function, and Bioinformatics*. **77** (2009), pp. 114–122.

98. T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe, N. A. Baker, PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for

molecular simulations. *Nucleic Acids Res.* **35**, W522–5 (2007).

99. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10037–10041 (2001).

100. H. Ashkenazy, S. Abadi, E. Martz, O. Chay, I. Mayrose, T. Pupko, N. Ben-Tal, ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–50 (2016).

101. R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer. *bioRxiv* (2021), p. 2021.10.04.463034.

102. A. D. Edelstein, M. A. Tsuchida, N. Amodaj, H. Pinkard, R. D. Vale, N. Stuurman, Advanced methods of microscope control using μManager software. *J Biol Methods*. **1** (2014), doi:10.14440/jbm.2014.36.

103. V. Dunsing, M. Luckner, B. Zühlke, R. A. Petazzi, A. Herrmann, S. Chiantia, Optimal fluorescent protein tags for quantifying protein oligomerization in living cells. *Sci. Rep.* **8**, 1–12 (2018).

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

105. J. J. Davis, A. R. Wattam, R. K. Aziz, T. Brettin, R. Butler, R. M. Butler, P. Chlenski, N.

Conrad, A. Dickerman, E. M. Dietrich, J. L. Gabbard, S. Gerdes, A. Guard, R. W. Kenyon, D. Machi, C. Mao, D. Murphy-Olson, M. Nguyen, E. K. Nordberg, G. J. Olsen, R. D. Olson, J. C. Overbeek, R. Overbeek, B. Parrello, G. D. Pusch, M. Shukla, C. Thomas, M. VanOeffelen, V. Vonstein, A. S. Warren, F. Xia, D. Xie, H. Yoo, R. Stevens, The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res.* **48**, D606–D612 (2020).

106. B. Lu, H. W. Leong, Computational methods for predicting genomic islands in microbial genomes. *Comput. Struct. Biotechnol. J.* **14**, 200–206 (2016).

107. A. Sievers, K. Bosiek, M. Bisch, C. Dreessen, J. Riedel, P. Froß, M. Hausmann, G. Hildenbrand, K-mer Content, Correlation, and Position Analysis of Genome DNA Sequences for the Identification of Function and Evolutionary Features. *Genes* . **8** (2017), doi:10.3390/genes8040122.

108. F. Zhou, V. Olman, Y. Xu, Barcodes for genomes and applications. *BMC Bioinformatics*. **9**, 546 (2008).

109. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* **39**, 1–38 (1977).

110. A. Bize, C. Midoux, M. Mariadassou, S. Schbath, P. Forterre, V. Da Cunha, 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).

111. P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

112. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

113. B. Contreras-Moreira, P. Vinuesa, GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* **79**, 7696–7701 (2013).

114. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, UniProt Consortium, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*. **31**, 926–932 (2015).

115. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

116. M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, J. Söding, HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*. **20**, 473 (2019).

117. M. Waskom, seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

118. H. Tang, J. E. Bowers, X. Wang, R. Ming, M. Alam, A. H. Paterson, Synteny and collinearity in plant genomes. *Science*. **320**, 486–488 (2008).

119. A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, S. R. Eddy, The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–41 (2004).

120. K. S. Makarova, A. V. Sorokin, P. S. Novichkov, Y. I. Wolf, E. V. Koonin, Clusters of

orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct*. **2**, 1–20 (2007).

121. K. S. Makarova, Y. I. Wolf, E. V. Koonin, 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).

122. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

123. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*. **14**, 587–589 (2017).

124. A. Sali, T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).

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

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

127. D. Lundin, A. M. Poole, B.-M. Sjöberg, M. Högbom, Use of structural phylogenetic networks for classification of the ferritin-like superfamily. *J. Biol. Chem.* **287**, 20565–20575 (2012).

128. Y. Ye, A. Godzik, FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.* **32**, W582–5 (2004).

129. Kanai R., Kar Kalipada, Anthony Karen, Gould L. Hannah, Ledizet Michel, Fikrig Erol, Marasco Wayne A., Koski Raymond A., Modis Yorgo, Crystal Structure of West Nile Virus Envelope Glycoprotein Reveals Viral Surface Epitopes. *J. Virol.* **80**, 11000–11008 (2006).

130. 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).

131. D. L. Gibbons, M.-C. Vaney, A. Roussel, A. Vigouroux, B. Reilly, J. Lepault, M. Kielian, F. A. Rey, Conformational change and protein-protein interactions of the fusion protein of Semliki Forest virus. *Nature*. **427**, 320–325 (2004).

132. J. E. Voss, M.-C. Vaney, S. Duquerroy, C. Vonrhein, C. Girard-Blanc, E. Crublet, A. Thompson, G. Bricogne, F. A. Rey, Glycoprotein organization of Chikungunya virus particles revealed by X-ray crystallography. *Nature*. **468**, 709–712 (2010).

133. P. Guardado-Calvo, K. Atkovska, S. A. Jeffers, N. Grau, M. Backovic, J. Pérez-Vargas, S. M. de Boer, M. A. Tortorici, G. Pehau-Arnaudet, J. Lepault, P. England, P. J. Rottier, B. J. Bosch, J. S. Hub, F. A. Rey, A glycerophospholipid-specific pocket in the RVFV class II fusion protein drives target membrane insertion. *Science*. **358**, 663–667 (2017).

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

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

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

137. X. Lu, H. Xiao, S. Li, X. Pang, J. Song, S. Liu, H. Cheng, Y. Li, X. Wang, C. Huang, T. Guo, J. Ter Meulen, S. Daffis, J. Yan, L. Dai, Z. Rao, H.-D. Klenk, J. Qi, Y. Shi, G. F. Gao, Double Lock of a Human Neutralizing and Protective Monoclonal Antibody Targeting the Yellow Fever Virus Envelope. *Cell Rep.* **26**, 438–446.e5 (2019).

138. Y. Modis, S. Ogata, D. Clements, S. C. Harrison, A ligand-binding pocket in the dengue virus envelope glycoprotein. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 6986–6991 (2003).

139. F. A. Rey, F. X. Heinz, C. Mandl, C. Kunz, S. C. Harrison, The envelope glycoprotein from tick-borne encephalitis virus at 2 A resolution. *Nature*. **375**, 291–298 (1995).

140. P. Guardado-Calvo, E. A. Bignon, E. Stettner, S. A. Jeffers, J. Pérez-Vargas, G. Pehau-Arnaudet, M. A. Tortorici, J.-L. Jestin, P. England, N. D. Tischler, F. A. Rey, Mechanistic Insight into Bunyavirus-Induced Membrane Fusion from Structure-Function Analyses of the Hantavirus Envelope Glycoprotein Gc. *PLoS Pathog.* **12**, e1005813 (2016).

141. M. Dessau, Y. Modis, Crystal structure of glycoprotein C from Rift Valley fever virus. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 1696–1701 (2013).

142. Y. Liu, J. Pei, N. Grishin, W. J. Snell, The cytoplasmic domain of the gamete membrane fusion protein HAP2 targets the protein to the fusion site in *Chlamydomonas* and regulates the fusion reaction. *Development*. **142**, 962–971 (2015).

143. P. Gouet, X. Robert, E. Courcelle, ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res.* **31**, 3320–3323 (2003).

144. R. M. Williams, U. Senanayake, M. Artibani, G. Taylor, D. Wells, A. A. Ahmed, T. Sauka-

Spengler, Genome and epigenome engineering CRISPR toolkit for in vivo modulation of cis-regulatory interactions and gene expression in the chicken embryo. *Development*. **145** (2018), doi:10.1242/dev.160333.

145. T. Matsuda, C. L. Cepko, Electroporation and RNA interference in the rodent retina in vivo and in vitro. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 16–22 (2004).

Acknowledgments: 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; Jose Flores for advice on searches of fusexins in metagenomes; Olivier Gascuel for discussions on phylogenetics; Pablo Dans for advice on molecular dynamics; Yoav Henis for the GPI-BHA plasmid. Part of the computations for this work was performed at the Vital-IT Center for high-performance computing of the SIB Swiss Institute of Bioinformatics.

Funding:

Comisión Sectorial de Investigación Científica grant CSIC I+D-2020-682, Uruguay (HR, MG)

Fondo para la Investigación Científica y Tecnológica grant, Argentina (PICT-2017-0854 (PSA)

Beca de Doctorado Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina (DM)

Knut and Alice Wallenberg Foundation grant 2018.0042, Sweden (LJ)

Novartis Foundation for Medical-Biological Research grant 17B111, Switzerland (CD)

Swedish Research Council grants 2016-03999 and 2020-04936 (LJ)

Swiss Leading House for the Latin American Region (CD and PSA)

Swiss National Science Foundation grant 183723 (CD)

Israel Science Foundation grants 257/17, 2462/18, 2327/19, and 178/20 (BP)

European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 844807 (NGB)

FOCEM, Fondo para la Convergencia Estructural del Mercosur grant COF 03/11 (MG)

Programa de Desarrollo de las Ciencias Básicas, Uruguay (HR, MG, and ML)

Author contributions: BP conceived the experiments; performed some imaging work; designed, supervised and analyzed cell fusion experiments. CD helped devise analysis strategies for k-mer and phylogenetic analysis. CV designed, performed and analyzed cell fusion assays. DdeS collected X-ray data, took part in structural analysis and validated metal substructure. DM carried out deep homology detection of Fsx1; 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. HR supervised the bioinformatics part of the work, estimated relative acquisition times, designed and performed, non-reversible substitution models analyses, structural phylogenetic analyses and phylogenomic surveys. KF made constructs of Fsx1 mutants. KT and JJ generated AlphaFold2 models of monomeric Fsx1. LJ supervised the biochemical and structural part of the work; collected X-ray data; solved the Fsx1 structure, refined and analyzed it; and generated AlphaFold-Multimer models of trimeric Fsx1. MG designed bioinformatic strategies, supervised bioinformatic aspects of the work, analyzed sequence and structural data. ML performed IME synteny analyses, phylogenetic and phylogenomic surveys. NGB carried out live imaging and surface biotinylation experiments, assisted with the preparation and design of plasmids; analyzed data. PSA supervised the bioinformatics part of the work, analyzed data. SN expressed, purified and crystallized Fsx1; performed SEC-MALS experiments; analyzed SAXS data; collected X-ray data and took part in structure determination, model building and structure analysis. XL carried out immunofluorescence and western blots for archaeal fusexins in mammalian cells, designed and

constructed plasmids, performed imaging work. DM, SN, XL, NGB, MG, HR, PSA, LJ and BP made figures and tables. DM, SN, XL, MG, HR, PSA, LJ and BP wrote the manuscript. All authors reviewed the manuscript.

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

Data and materials availability: Crystallographic structure factors and atomic coordinates have been deposited in the Protein Data Bank under accession code 7P4L. All relevant codes, notebooks and datasets necessary for: HHblits (v3.3.0) and HMMER searches and comparisons (Fig. 1A, fig. S1 and data S1); raw data and statistical information for functional analyses (Fig. 3, fig. S2 and data S4); sequences of *fsx1* genes synthesized (Data S5); K-mer spectra analyses (Fig. 4A and fig. S13); IMEs clustering, content and synteny analyses (Fig. 4B, figs. S14, S15 and S16, data S2 and S3), protein structure- and sequence-based comparisons (Fig. 4C and fig. S17), are available at https://github.com/DessimozLab/Archaeal-Fusexins.

Supplementary Materials

Materials and Methods

Figs. S1-S19

Tables S1-S7

References (50-145)

Movies S1-S2

Data S1-S5

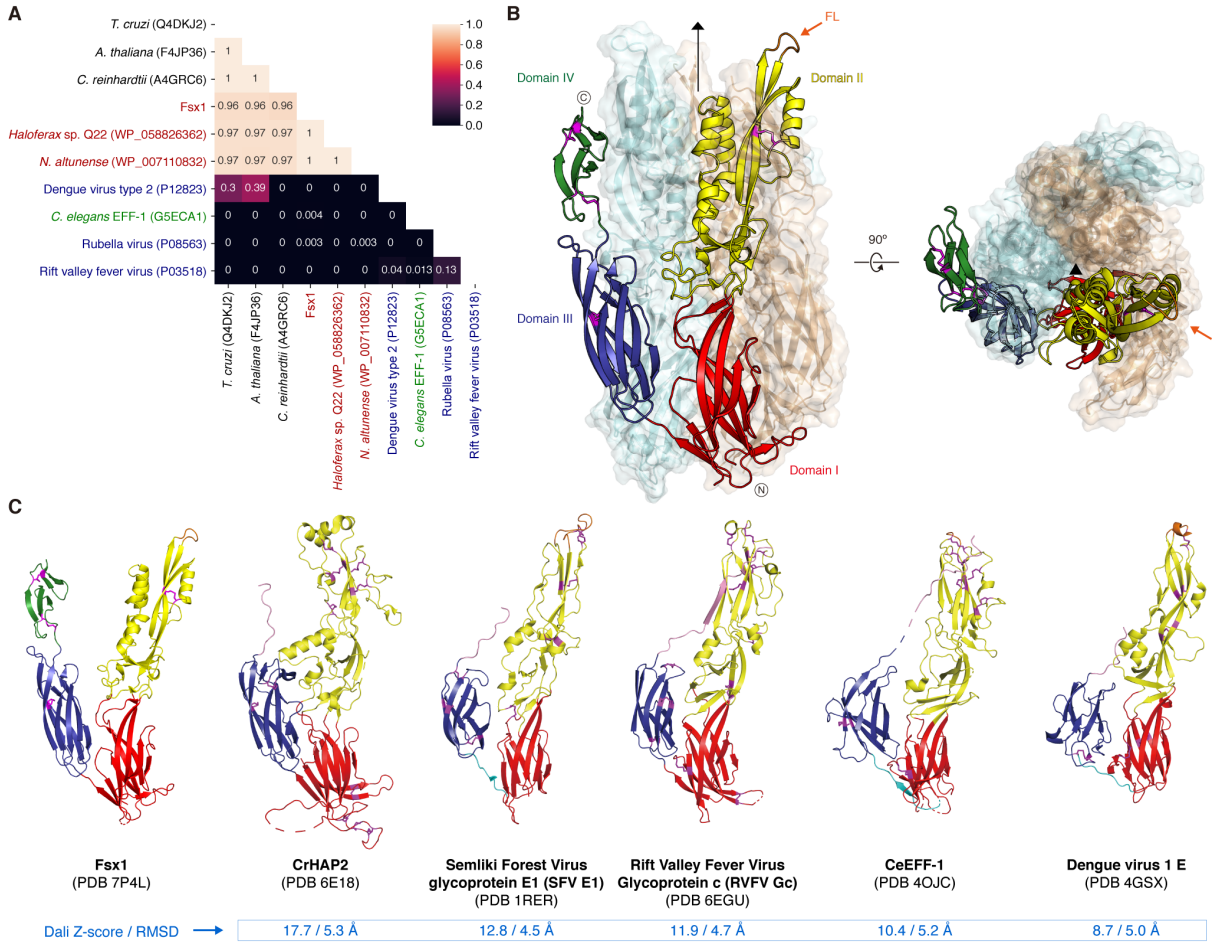

Fig. 1. Fsx1 is a member of the fusexin protein superfamily. (**A**) HMM vs HMM homology probabilities of a subset of eukaryotic, viral and archaeal fusexin ectodomains. With exception of Fsx1, which derives from a metagenomic sequence, archaeal fusexins (red), viral fusexins (blue), EFF-1 (green) and HAP2s (black) are indicated by RefSeq/UniProt identifiers. (**B**) Crystal structure of the trimeric ectodomain of Fsx1. The three-fold non-crystallographic axis is indicated. Subunit A is shown as a cartoon coloured by domain, with disulfides and the fusion loop (FL) coloured magenta and orange, respectively; subunits B, C are in mixed cartoon/surface representation. (**C**) Side-by-side comparison of $Fsx1_E$ and known class II fusogens. Elements are coloured as in panel b; the stem region and the linker between domains I and III are shown in pink and cyan, respectively.

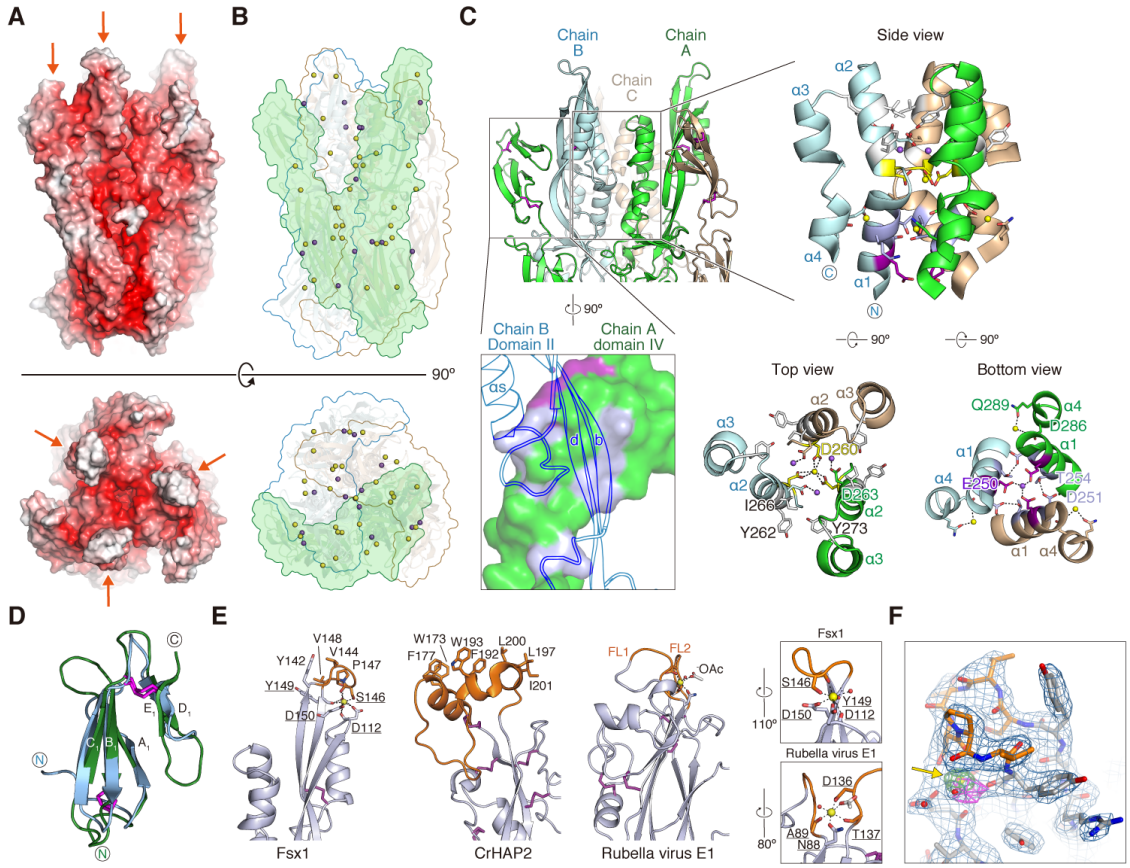

Fig. 2. Distinct structural features of Fsx1. (A) $Fsx1_E$ surface coloured by electrostatic potential from red (-10 kT/e) to blue (+10 kT/e) through white (0 kT/e). Orange arrows indicate the FLs. (B) Location of the Ca^{2+} and Na^{+} ions (depicted as yellow and purple spheres, respectively; see also figs. S5 and S7) stabilizing the $Fsx1_E$ trimer. The molecular surface of a protomer is shaded green and the outline of the other two subunits is coloured cyan and wheat. (C) Details of interactions at the level of the six-helical bundle made by domain II of the Fsx1 subunits (right subpanels) and the domain IV/domain II interface (bottom left subpanel). Selected side chains are colored by the type of inter-chain contacts in which they are involved (gray: hydrophobic interaction; blue bell: hydrogen bonding; yellow: Ca^{2+} coordination; fuchsia: Na^{+} coordination), with dashed lines indicating hydrogen bonds. Note that the helical bundle of Fsx1 is not a leucine-zipper coiled-coil structure, such as those found in class I/III viral fusion proteins or in the SNARE four-helix bundles, and see also fig. S7B. (D) Superposition of Fsx1 domain IV (green) and Der p

23 (PDB 4CZE, blue) (Dali Z-score 3.6, RMSD 2.2 Å). (E) Comparison of the Fsx1 region that includes the FL and the corresponding parts of CrHAP2 and Rubella virus E1 protein (PDB 4B3V). Residues coordinating the Ca^{2+} ion that stabilizes the Fsx1 FL are underlined, and compared to the E1 protein Ca^{2+}-binding region in the boxed panels on the far right. (F) The Fsx1 FL adopts a highly ordered conformation stabilized by a Ca^{2+} ion. Presence and identity of the latter, indicated by a yellow arrow, are supported by two other maps shown in addition to the *2mFo-DFc* map (blue mesh, contoured at 1.0 σ): a difference map calculated upon omitting all metal ions from the model (thick green(+)/red(-) mesh, 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, 3.2 σ).

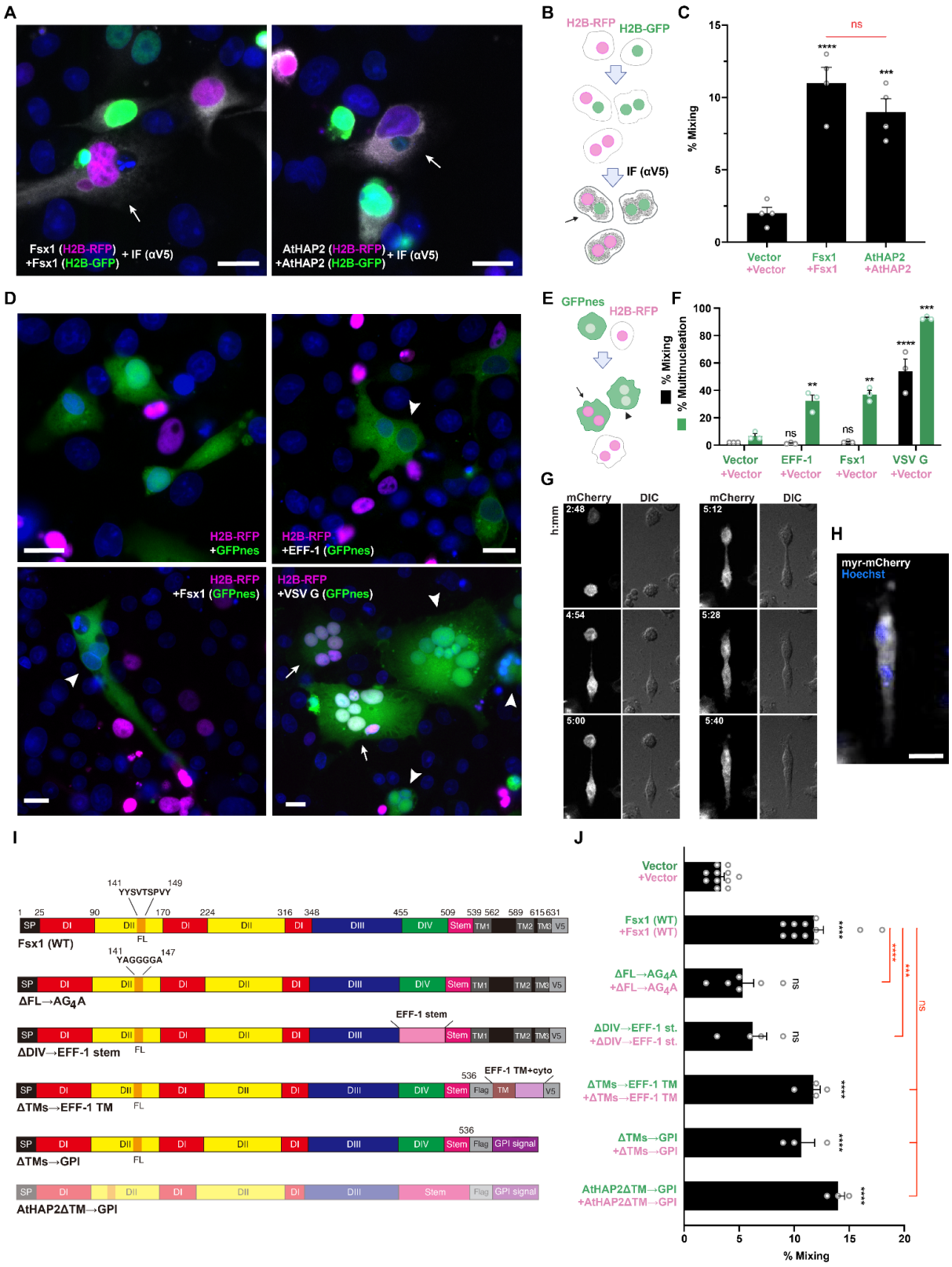

Fig. 3. Fsx1 mediates bilateral cell-cell fusion. (A to C) Cell-cell fusion was measured by content-mixing, indicated by the appearance of multinucleated cells containing green nuclei (H2B-

GFP) and magenta nuclei (H2B-RFP). Immunofluorescence against the V5 tag was also performed (gray). (A) Images of mixed cells. DAPI, blue. See also fig. S9A. (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 to 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 Fsx1, EFF-1 or VSV G. (D) Images of cells transfected with empty::GFPnes vector, Fsx1::GFPnes, EFF-1::GFPnes or VSV G::GFPnes. Fsx1 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. See also fig. S9E. (E) Scheme of experimental design. (F) Quantification of content-mixing experiments in which only the GFP population of cells express Fsx1, 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) Spinning disk microscopy time-lapse images indicating the merging of two cells expressing myr-mCherry and Fsx1. Time in hours:minutes. The red channel (mCherry, white) and the DIC are shown. Refer to Movie S1. (H) For the last point a Z-projection showing the myr-mCherry fluorescence (white) and the nuclei Hoechst (blue; Movie S2). Scale bar, 20 µm. (I to J) Structure function analysis of Fsx1. (I) Schematic diagram of wild type Fsx1 four mutants and AtHAP2ΔTM→GPI. SP, signal peptide; FL, fusion loop (Fig. 2E). (J) Quantification of content-mixing in populations of cells expressing vectors (n=7), Fsx1 (wt) (n=7), its mutants (ΔFL→AG4A (n=6), ΔDIV→EFF-1 stem (n=4), ΔTMs→EFF-1 TM (n=4), ΔTMs→GPI (n=3), or AtHAP2ΔTM→GPI (n=3) (see fig. S10). Bar chart showing means ± SEM. Comparisons by one-way ANOVA followed by Bonferroni's test

against the vector (black) and against Fsx1 (red). ns=non-significant, *** $p < 0.001$, **** $p < 0.0001$.

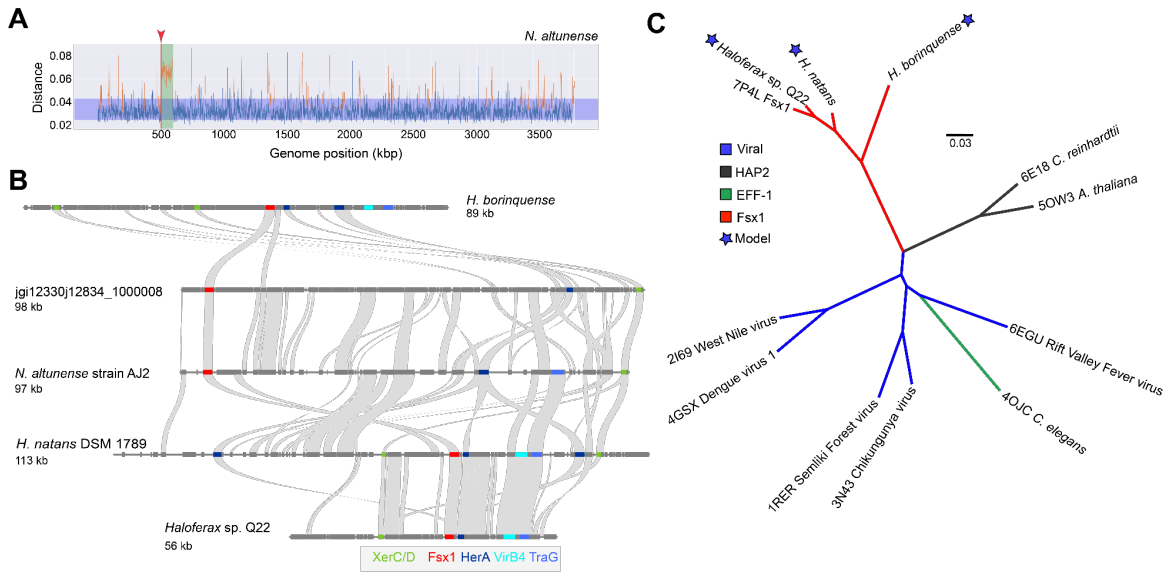

Fig. 4. Genomic features of Fsx1s and evolutionary history of fusexins. (A) Fsx1s are embedded in integrated mobile elements (IMEs). *Natrinema altunense* AJ2 complete genome k-mer spectrum deviation from centroid. Blue region shows the standard deviation. Locus of *fsx1* is indicated by the red arrowhead; the mobile element containing *fsx1* is in green. (B) Four IMEs from PCG sequences and the metagenomic sequence from which the crystallized Fsx1 was obtained (second from top) are represented by their annotated genes (thick segments). Gray links are drawn between homologs. The *fsx1* gene is marked in red and selected ORFs homologous to IME signature genes are labeled and color-coded. XerC/XerD recombinases from family integrase (green); HerA helicase (dark blue); VirB4, Type IV secretory (T4SS) pathway (cyan); TraG/TraD/VirD4 family enzyme (light blue). See also fig. S16, and tables S1 and S3. (C) Computational models based on the $Fsx1_E$ crystal structure and experimental structures of other fusexins were compared using flexible structural alignment to build a minimum evolutionary tree. Scale bar represents distance as $1\text{-}TM_{score}$ (see Methods). These distances indicate that all structures are homologous.