nature research

Corresponding author(s): Oded Béjà & Hideki Kandori

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	□	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	*Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.*
Data analysis	BLASTX, PSI-BLAST, CD-blast,HHpred, Phyre2,MUSCLE, TMHMM, Phobius, Philius, and SPOCTOPUS. DIAMOND v0.9.10, Python v3.6.4, Python packages: pandas (v0.22.0) and seaborn (v.0.8.1), Jupyter notebooks: diamond_processing-RPKM.ipynb and diamond_processing-RPKM.ipynb (available at https://github.com/BejaLab/heliorhodopsin).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Fosmid KIN48C12 sequence was deposited in the GenBank with accession #MF737519. All the scripts and notebooks used to process the metagenomic data analysis are available at https://github.com/BejaLab/heliorhodopsin.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used metagenomics is some parts of the manuscript. Number of reads per sample is between 20 M to 100 M reads and was the maximal in each sample. No data was excluded. Otherwise not relevant.
Data exclusions	No data was excluded
Replication	Each experiment was repeated at least twice (see figure legends) except for the biophysical measurments with the Escherichia coli cells. This includes the topology experiments as well as the biophysical measurements. E. coli transport assays and spectroscopic experiments were performed only one time (n = 1). Each clone and mutation was verified by sequencing. all attempts at replication were successful.
Randomization	No randomization was applied
Blinding	No blinding was applied

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Rat/mouse hybrid. The cell line is commercially available from Merck or DS Pharma Biomedical (Japan).
Authentication	Only for research purpose.
Mycoplasma contamination	No.
Commonly misidentified lines (See ICLAC register)	Not relevant.