nature portfolio

Corresponding author(s): Andreas Wagner

Last updated by author(s): Apr 28, 2022

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the 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
☐	☒ A statement on 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 statistical parameters 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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	BD FACSDiva™ (version 8.0.1), FlowJo™ (version 10.0), Tecan i-control (version 3.14)
Data analysis	R (version 3.5.2) was used for general statistical analysis. Python (version 3.9) was used for identifying neo-functionalizing variants. MEGA X was used for identifying non-synonymous variants. The code for evolutionary simulations are available at: https://github.com/dasmeh/Discrete_Time_Markov_Chain_Evolution

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Custom code used in this study is available in a public GitHub repository (https://github.com/dasmeh/Discrete_Time_Markov_Chain_Evolution). All data are available in the manuscript or the supplementary materials. SMRT sequencing data are available at NCBI with a BioProject ID PRJNA833567 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA833567)

Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

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

Study description	Directed evolution of GFP through multiple rounds of mutagenesis followed by selection was used to determine the role of protein abundance in the evolvability of a protein. Selected protein populations were sequenced at every round of evolution. Evolution of GFP was modeled to investigate the mechanism behind higher evolvability of lowly expressed proteins.
Research sample	Protein populations of GFPmut2 evolving from green to cyan fluorescence
Sampling strategy	Four independent directed evolution experiments for highly expressed and lowly expressed proteins for evolution towards new phenotype of cyan fluorescence. The experiments were conducted with and without the phase of weak stabilizing selection.
Data collection	Fluorescence data and sequence data was collected at every round of directed evolution for both the experiments (with and without weak stabilizing selection). Folding stability was measured at the end of the first experiment and at the end of phase I of the second experiment.
Timing and spatial scale	NA
Data exclusions	NA
Reproducibility	We used four replicate directed evolution experiments for deriving our conclusions.
Randomization	NA
Blinding	NA

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	E. coli cells expressing the GFP at either high or low abundance
Instrument	Sorter: BD Aria II, Analyzer: BD Fortessa
Software	BD FACSDiva™ (version 8.0.1) for the data collection, FlowJo™ (version 10.0) for data analysis
Cell population abundance	All the E. coli populations were with the GFP construct
Gating strategy	For the first experiment and second phase of the second experiments, we selected 50 thousand cells with a top 1% cyan fluorescence followed by 50 thousand cells with a top 5% cyan fluorescence (top 0.05%, in effect). For the first phase of the second experiment we selected 50 thousand cells with a top 70% green fluorescence.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.