nature portfolio

https://doi.org/10.1038/s41586-025-09749-7

Supplementary information

Semantic design of functional de novo genes from a genomic language model

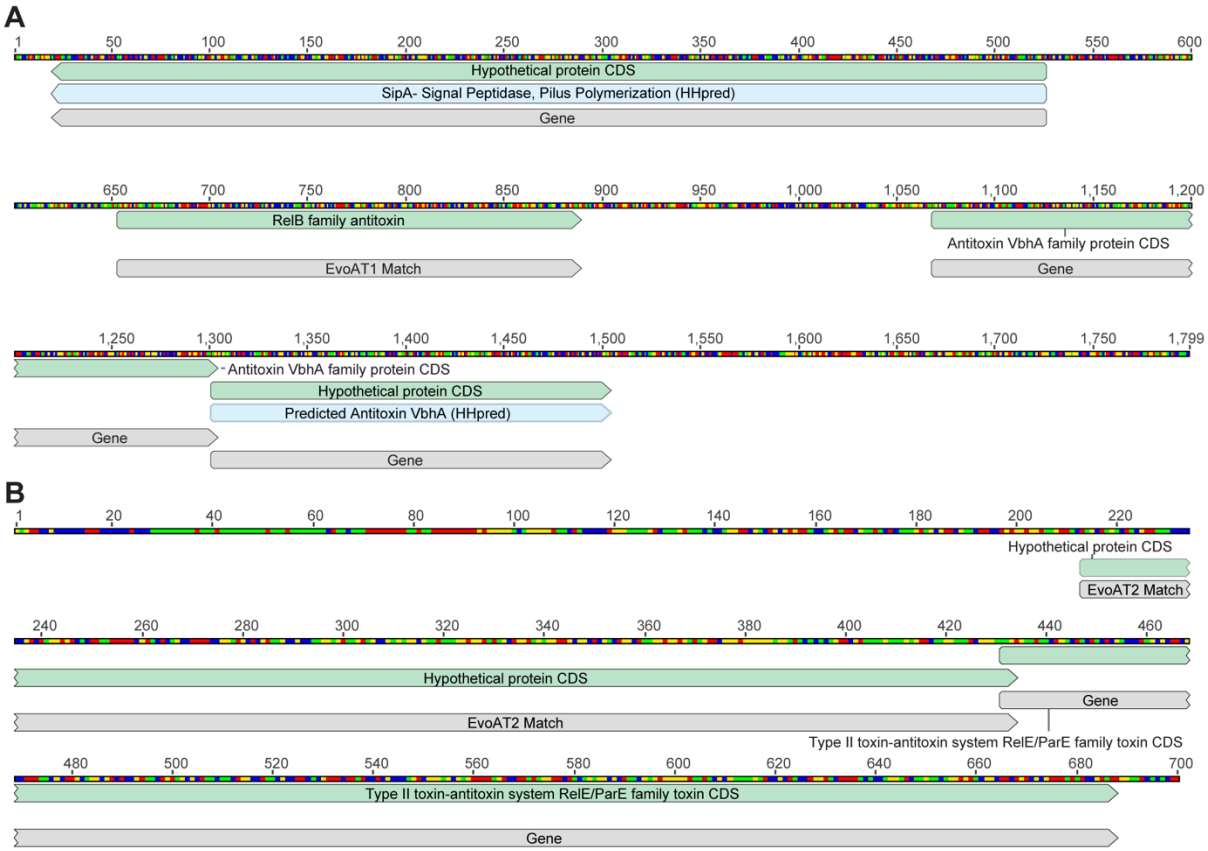
In the format provided by the authors and unedited

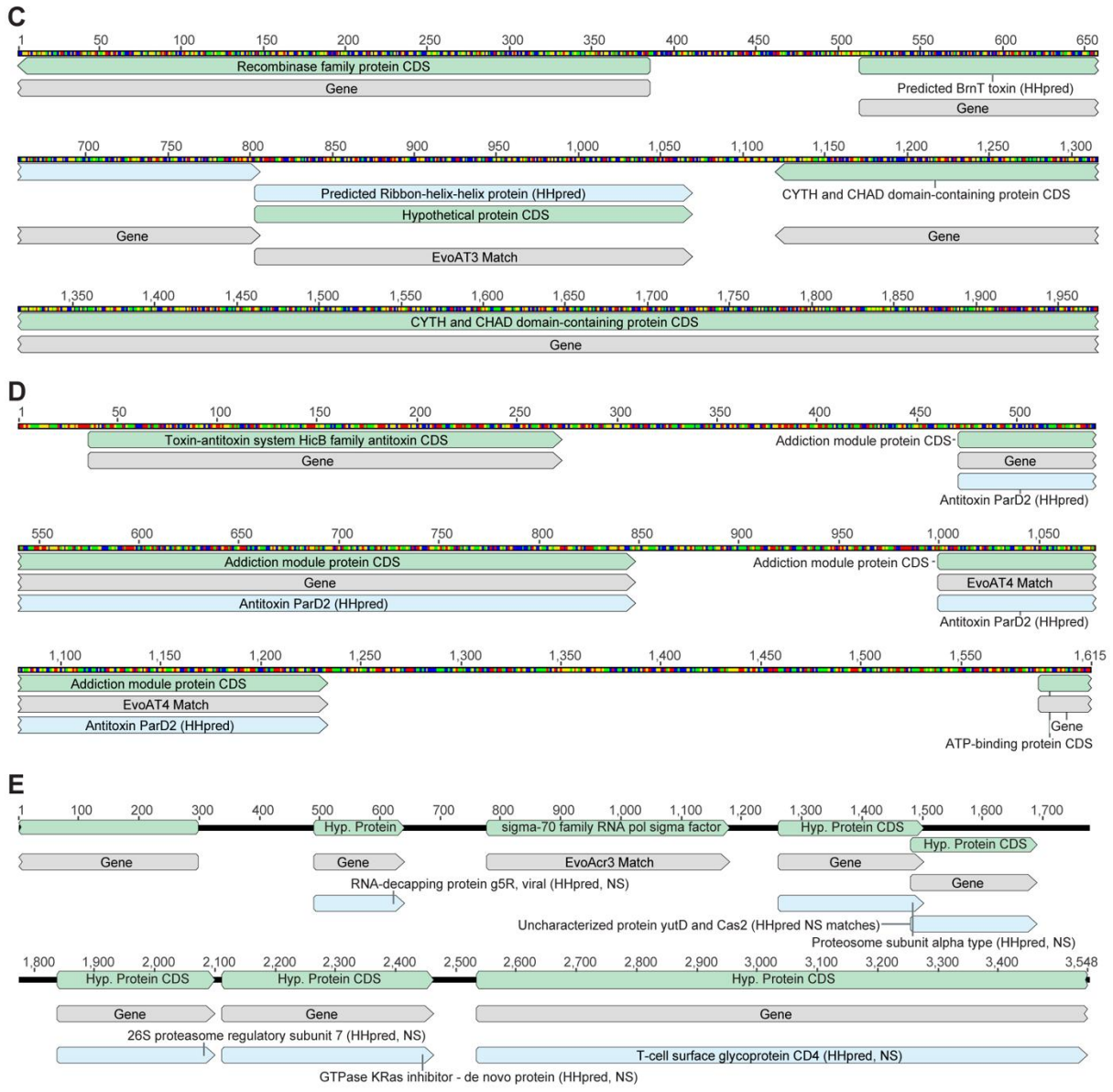
Semantic design of functional *de novo* genes from a genomic language model

Aditi T. Merchant[1,4], Samuel H. King[1,4], Eric Nguyen[1,4], and Brian L. Hie[2,3,4,*]

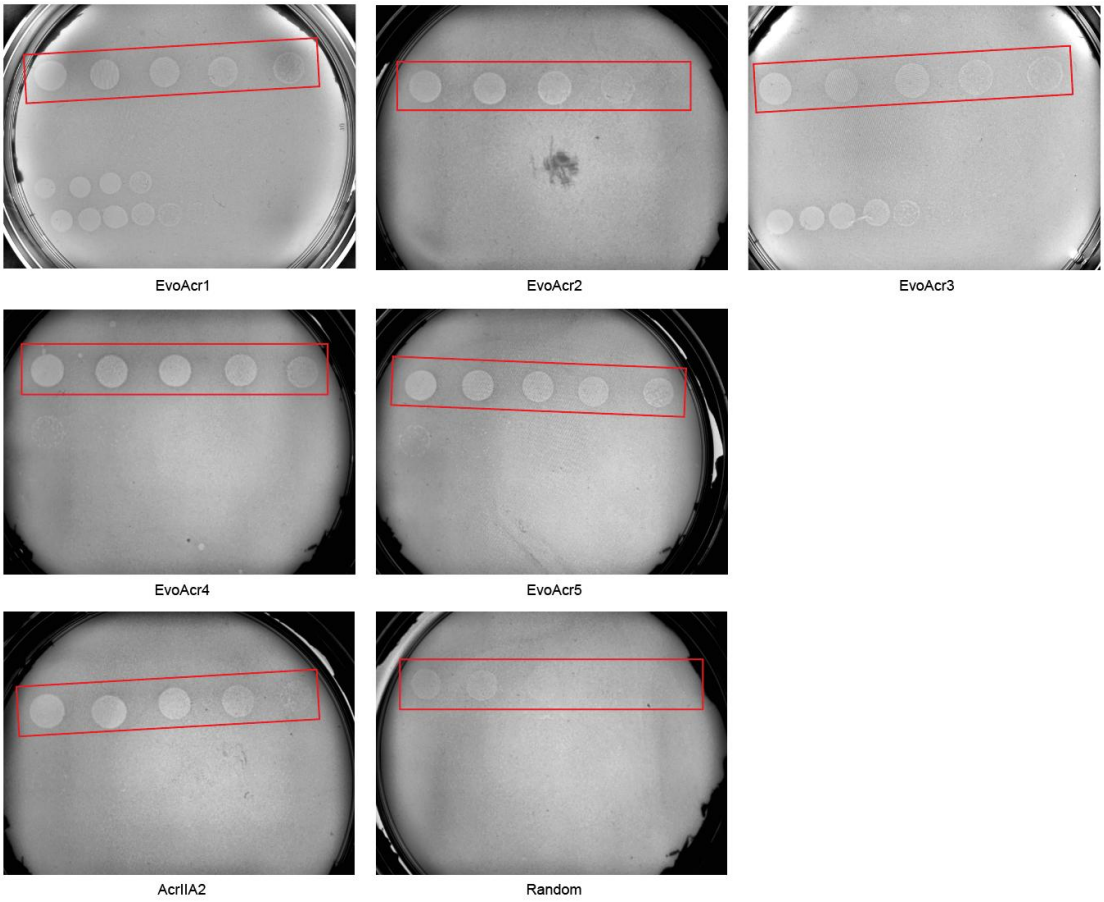
[1]Department of Bioengineering, Stanford University, Stanford, CA; [2]Department of Chemical Engineering, Stanford University, Stanford, CA; [3]Stanford Data Science, Stanford University, Stanford, CA; [4]Arc Institute, Palo Alto, CA.
*Correspondence: brianhie@stanford.edu

Supplementary information


A
1
50
100
150
200
250
300
350
400
450
500
550
600
Hypothetical protein CDS
SipA- Signal Peptidase, Pilus Polymerization (HHpred)
Gene
650
700
750
800
850
900
950
1,000
1,050
1,100
1,150
1,200
RelB family antitoxin
Antitoxin VbhA family protein CDS
EvoAT1 Match
Gene
1,250
1,300
1,350
1,400
1,450
1,500
1,550
1,600
1,650
1,700
1,750
1,799
Antitoxin VbhA family protein CDS
Hypothetical protein CDS
Predicted Antitoxin VbhA (HHpred)
Gene
Gene
B
1
20
40
60
80
100
120
140
160
180
200
220
Hypothetical protein CDS
EvoAT2 Match
240
260
280
300
320
340
360
380
400
420
440
460
Hypothetical protein CDS
Gene
EvoAT2 Match
Type II toxin-antitoxin system RelE/ParE family toxin CDS
480
500
520
540
560
580
600
620
640
660
680
700
Type II toxin-antitoxin system RelE/ParE family toxin CDS
Gene



Supplementary Figure 1: Genomic context of closest sequence matches of EvoAT1-4 and EvoAcr3. (A-D) Genes known and predicted to be related to toxin-antitoxin systems can be found in the immediate vicinity of the closest sequence identity matches for EvoAT1-4. **(E)** The closest sequence match for EvoAcr3, a sigma-70 RNA pol sigma factor, is surrounded by primarily uncharacterized proteins that are not known to be anti-CRISPRs. NS, non-significant.



Supplementary Figure 2. Uncropped phage plaque images for Acr protection assay. Uncropped phage plaque images depicting formation of plaques in response to EvoAcr1-5, AcrIIA2, and a random sequence.