Bioinformatics Journal Club

April, 28 2014

Ulvi Talas

Deep sequencing of large library selections allows computational discovery of diverse sets of zinc fingers that bind common targets

Anton V. Persikov1, Elizabeth F. Rowland1, Benjamin L. Oakes1, Mona Singh1,2,* and Marcus B. Noyes1,3,*

Nucleic Acids Research, 2014, Vol. 42, No. 3 1497–1508 doi:10.1093/nar/gkt1034

Deep sequencing of large library selections allows computational discovery of diverse sets of zinc fingers that bind common targets

Anton V. Persikov1, Elizabeth F. Rowland1, Benjamin L. Oakes1, Mona Singh1,2, and Marcus B. Noyes1,3,**

Princeton University, Princeton, NJ 08544, USA

Bioinformatics Journal Club

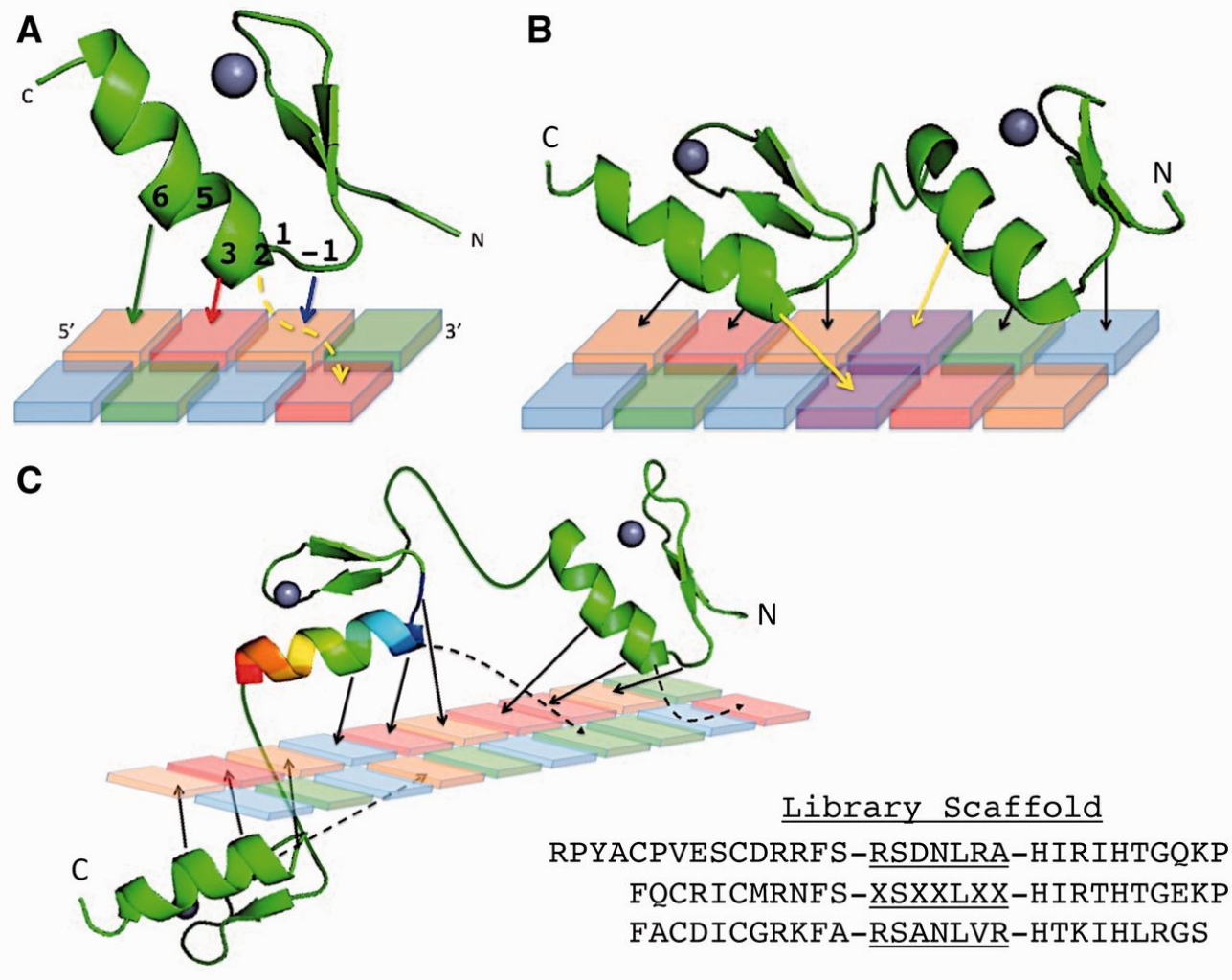
April, 28 2014

Ulvi Talas

Abstract

The Cys2His2 zinc finger (ZF) is the most frequently found sequence-specific DNA-binding domain in eukaryotic proteins. The ZF's **modular protein-DNA interface** has also served as a **platform for genome engineering applications.** Despite decades of intense study, a predictive **understanding of the DNA-binding specificities of either natural or engineered ZF domains remains elusive**. **To help fill this gap**, we **developed an integrated experimental-computational approach to enrich and recover distinct groups of ZFs that bind common targets.** To showcase the power of our approach, we **built several large ZF libraries and demonstrated their excellent diversity**. As **proof of principle**, we used one of these ZF libraries to select and recover thousands of ZFs that bind several 3-nt targets of interest. **We were then able to computationally cluster these recovered ZFs to reveal several distinct classes of proteins, all recovered from a single selection, to bind the same target.** Finally, for each target studied, we confirmed that one or more representative ZFs yield the desired specificity. In sum, **the described approach enables comprehensive large-scale selection and characterization of ZF specificities and should be a great aid in furthering our understanding of the ZF domain.**

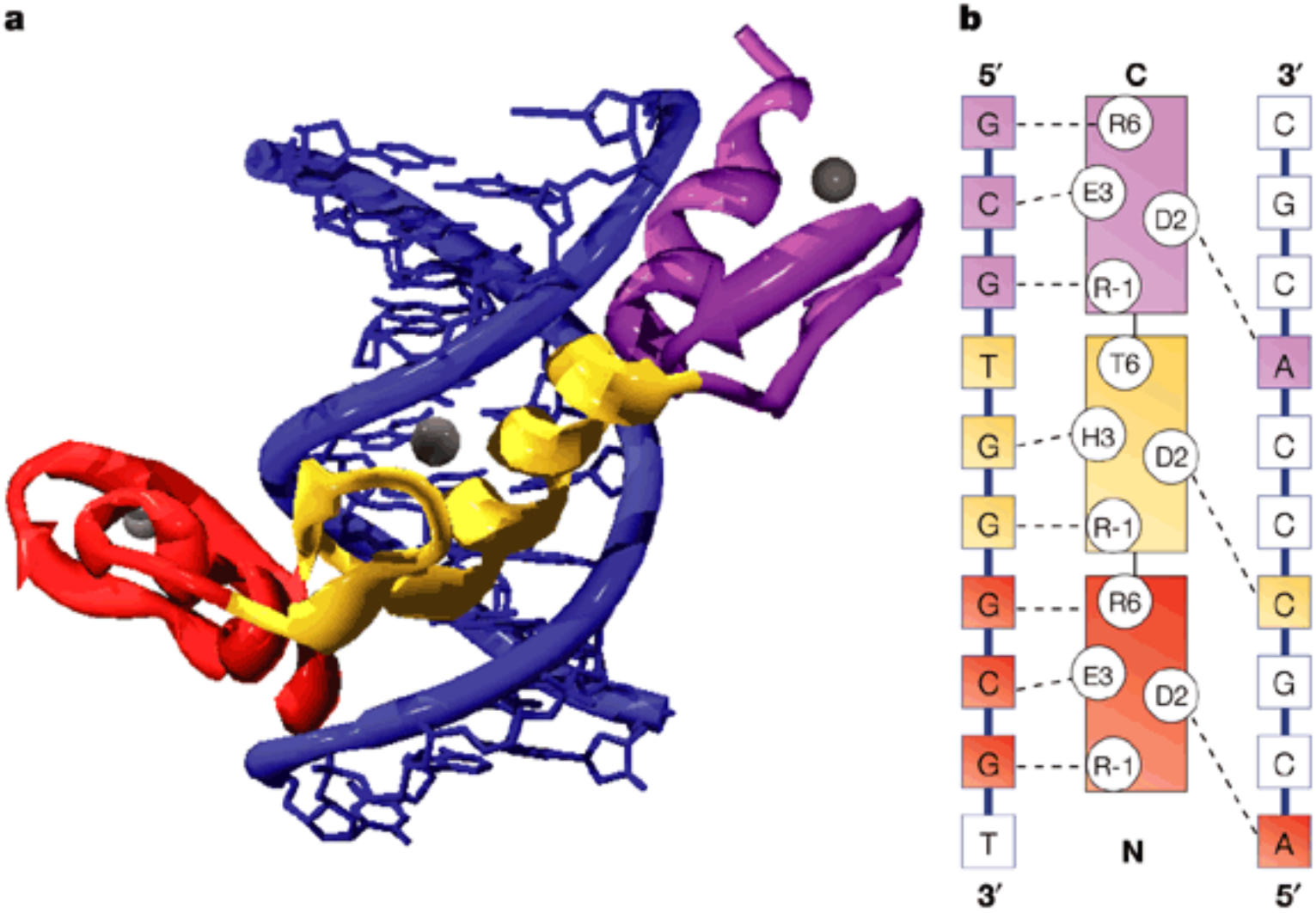
Zinc finger–DNA interactions.



Persikov A V et al. Nucl. Acids Res. 2014;42:1497-1508

Nucleic Acids Research

FIGURE | Modular interactions between zinc fingers and DNA



FROM THE FOLLOWING ARTICLE:

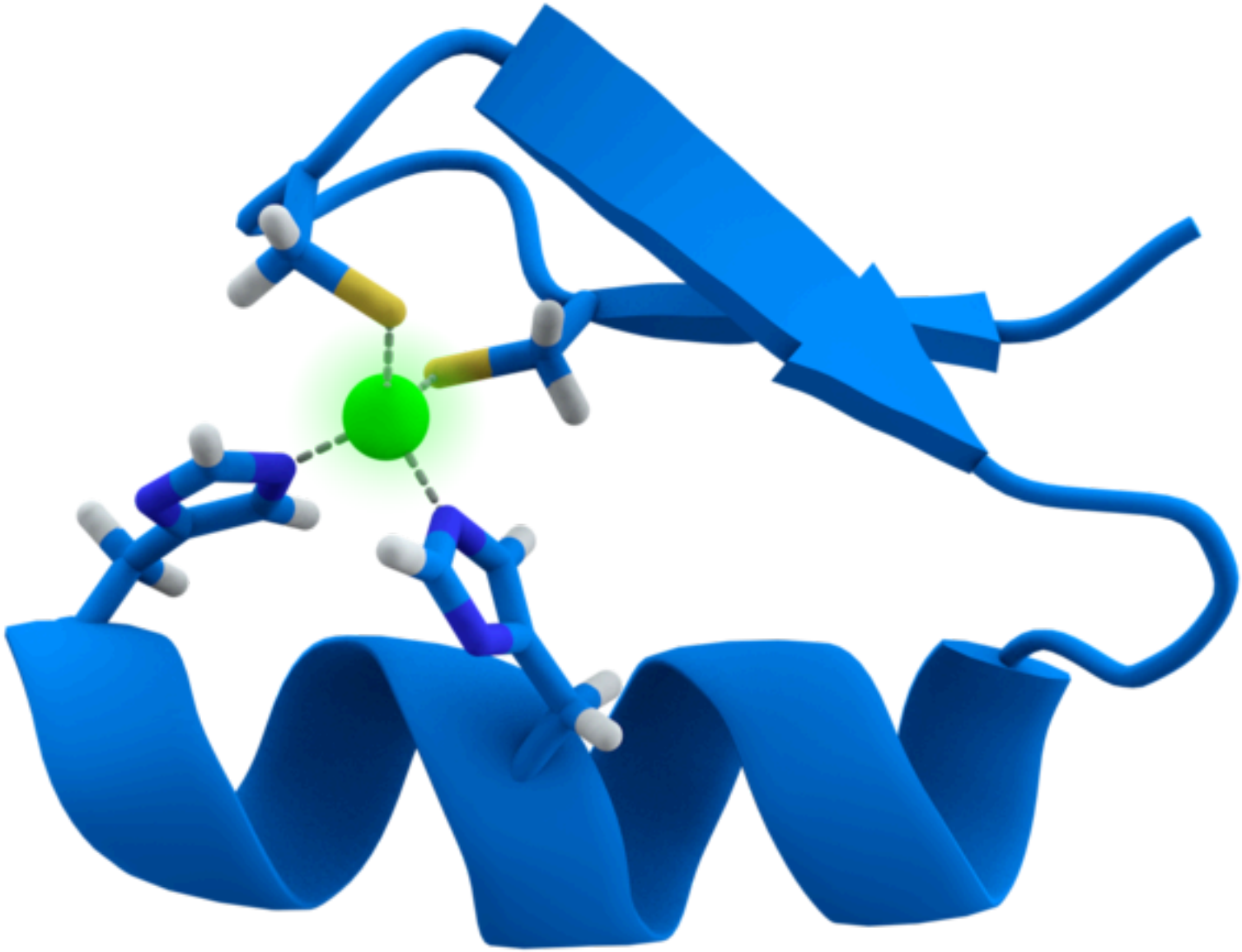
[Drug discovery with engineered zinc-finger proteins](#)

Andrew C. Jamieson, Jeffrey C. Miller & Carl O. Pabo

Nature Reviews Drug Discovery **2**, 361-368 (May 2003)

Nature Reviews | Drug Discovery

Cys2His2 Zn-finger:

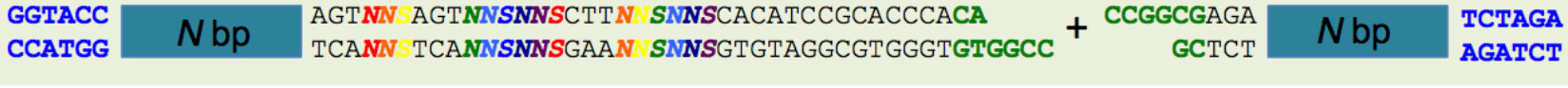


Outline:

- **First**, we optimized a polymerase chain reaction (PCR)-based cassette mutagenesis method to build libraries of ZF proteins where up to six amino acid positions are varied, all 20 amino acids are possible and theoretical DNA library sizes of $3x10^7$ and $1x10^9$ are over-sampled by at least 5-fold.
- **Second**, we derived a simple analytical formula to calculate the expected diversity of a library and showed via high-throughput sequencing that the produced libraries offer levels of diversity that approach the theoretical maximum.
- **Third**, we used one of our ZF libraries with five varying amino acid positions in conjunction with the B1H system to select and deep sequence ZFs that bind several 3 bp targets of interest.
- **Fourth**, we developed an information-theoretic approach, based on the number of ways a protein sequence may be encoded that allowed us to uncover enriched ZFs (i.e. corresponding to ZFs binding the targets of interest) from large sequence pools that may contain considerable background.
- **Fifth**, because the sequencing depth of selected ZF pools resulted in thousands of enriched ZFs for each target, we clustered them to uncover distinct classes of similar amino acid profiles.
- **Finally**, for each target studied, we confirmed that one or more of the selected ZFs offer the desired specificity and further tested a subset of these to confirm that they can act as artificial TFs in yeast.

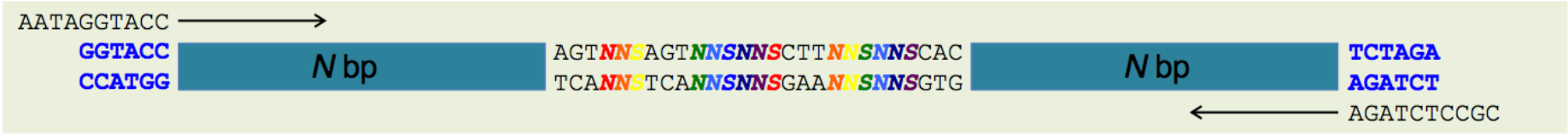
Step 1. Amplify fragments from parent template with appropriate oligonucleotide design.

Key components: 65-70ºC TM
48-96 reactions
15-20 cycles



Step 2. Digest and ligate PCR fragments.

a. Digest with pre-designed, internal restriction site (green)
b. Ligate fragments to create full-length library template



Step 3. Amplify library cassette.

Key components: 48-96 reactions
30 cycles
External restriction sites (blue) used for cloning into final vector

Diversity of the designed libraries:

For a sequence with ***i*** variable amino acid positions, encoded by the NNS codons, the total possible number of encoding DNA sequence variants (**N**) could be computed as:

$$N = (4*4*2)^i$$

6 *variable AA positions => N = 1.07 × 10^9*

5 *variable AA positions => N = 3.36 × 10^7*

We can compute the expected number of distinct (or unique) sequences **U** observed in the ideal case with perfectly uniform distribution as a function of the total number sequenced ***n*** and the total number of possible variants **N** as:

$$U = N\left(1 - \left(1 - \frac{1}{N}\right)^n\right)$$

Figure S2

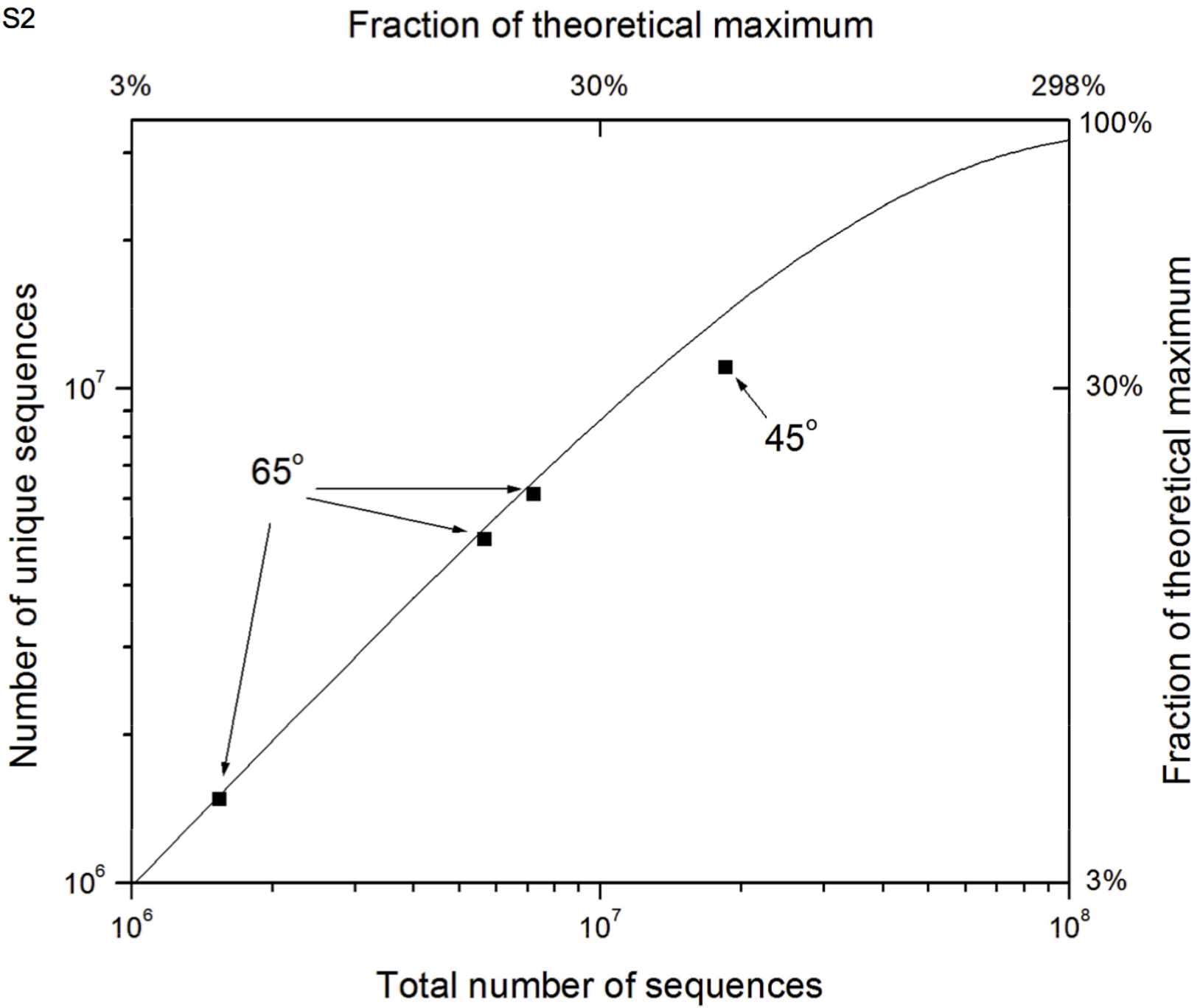
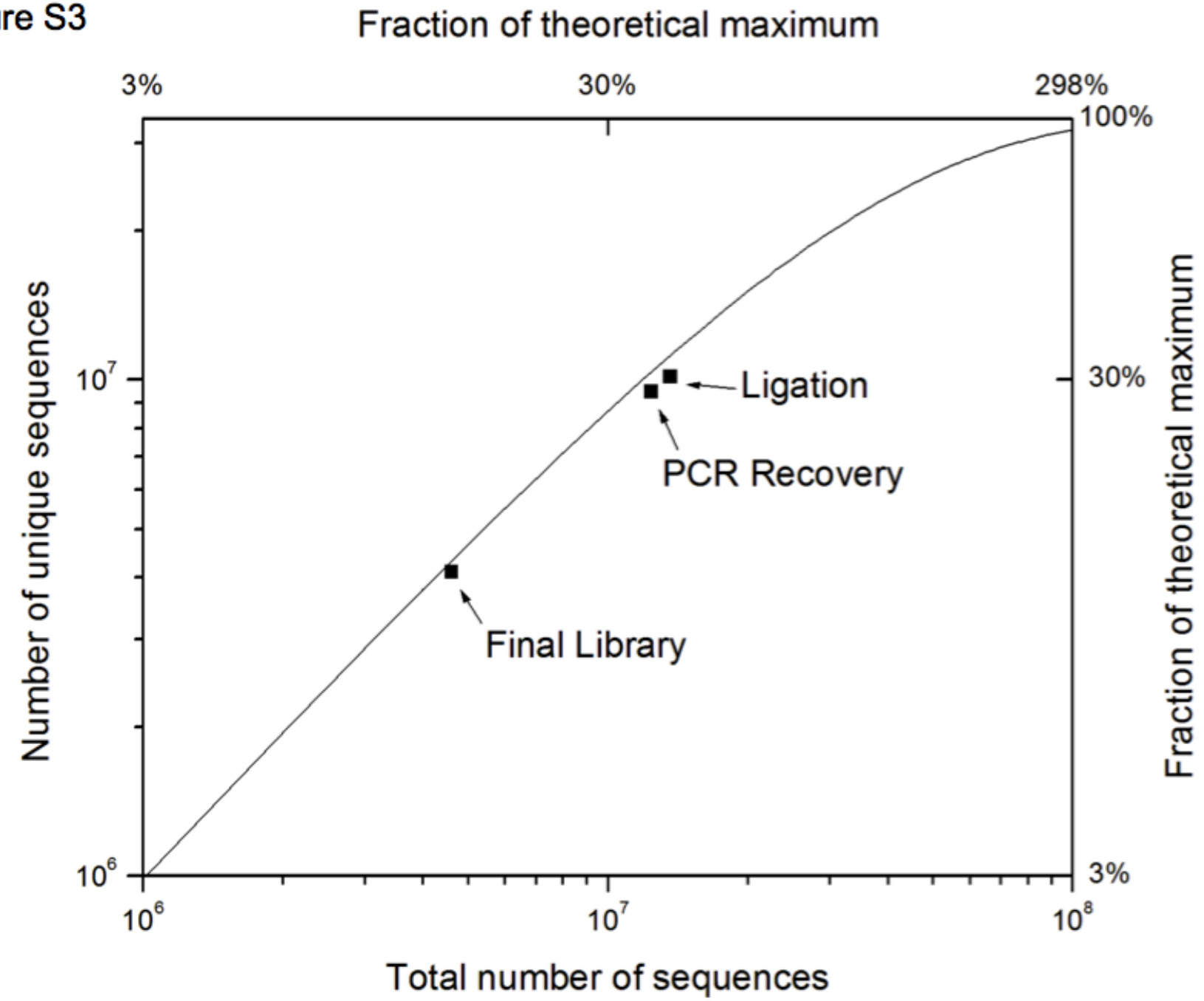
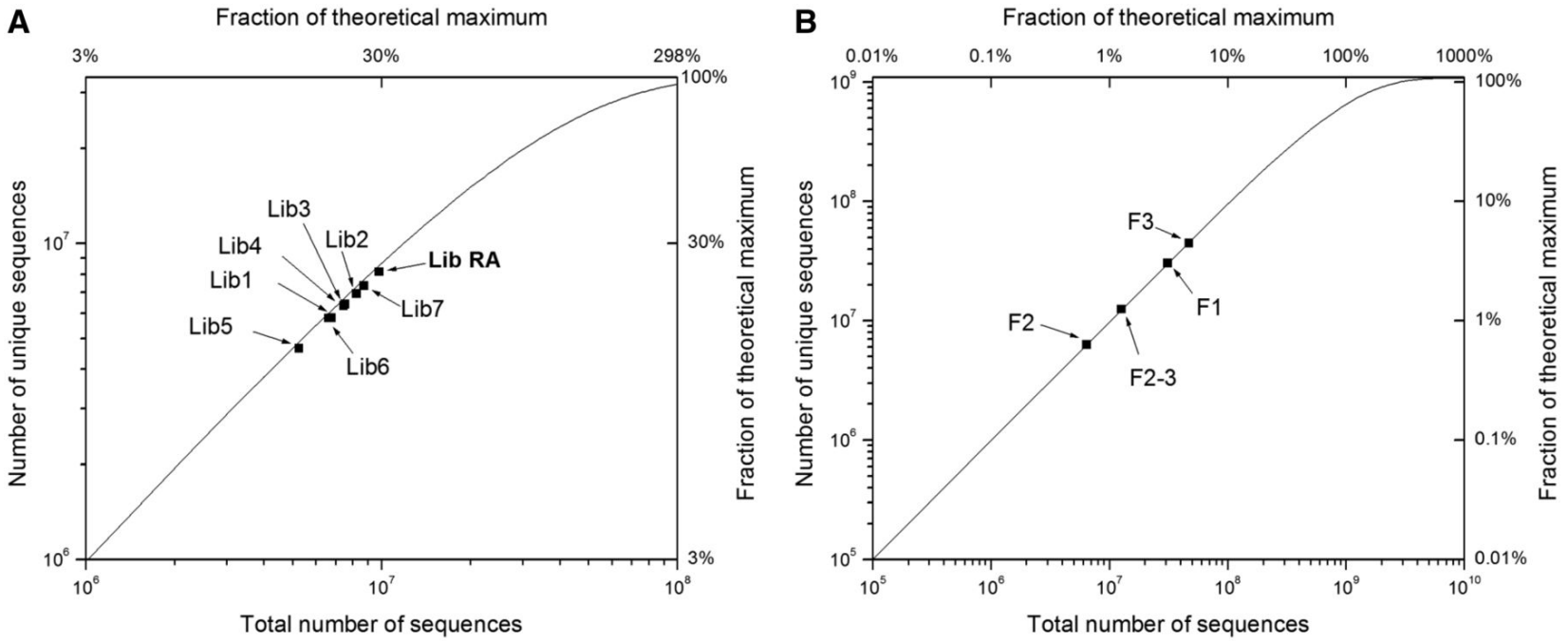

Fraction of theoretical maximum
3%
30%
298%
100%
30%
3%
10^7
10^6
10^6
10^7
10^8
Number of unique sequences
Fraction of theoretical maximum
Total number of sequences
65°
45°

Figure S3


Fraction of theoretical maximum
3%
30%
298%
100%
Number of unique sequences
10^7
30%
Ligation
PCR Recovery
Final Library
10^6
3%
10^6
10^7
10^8
Total number of sequences
Fraction of theoretical maximum

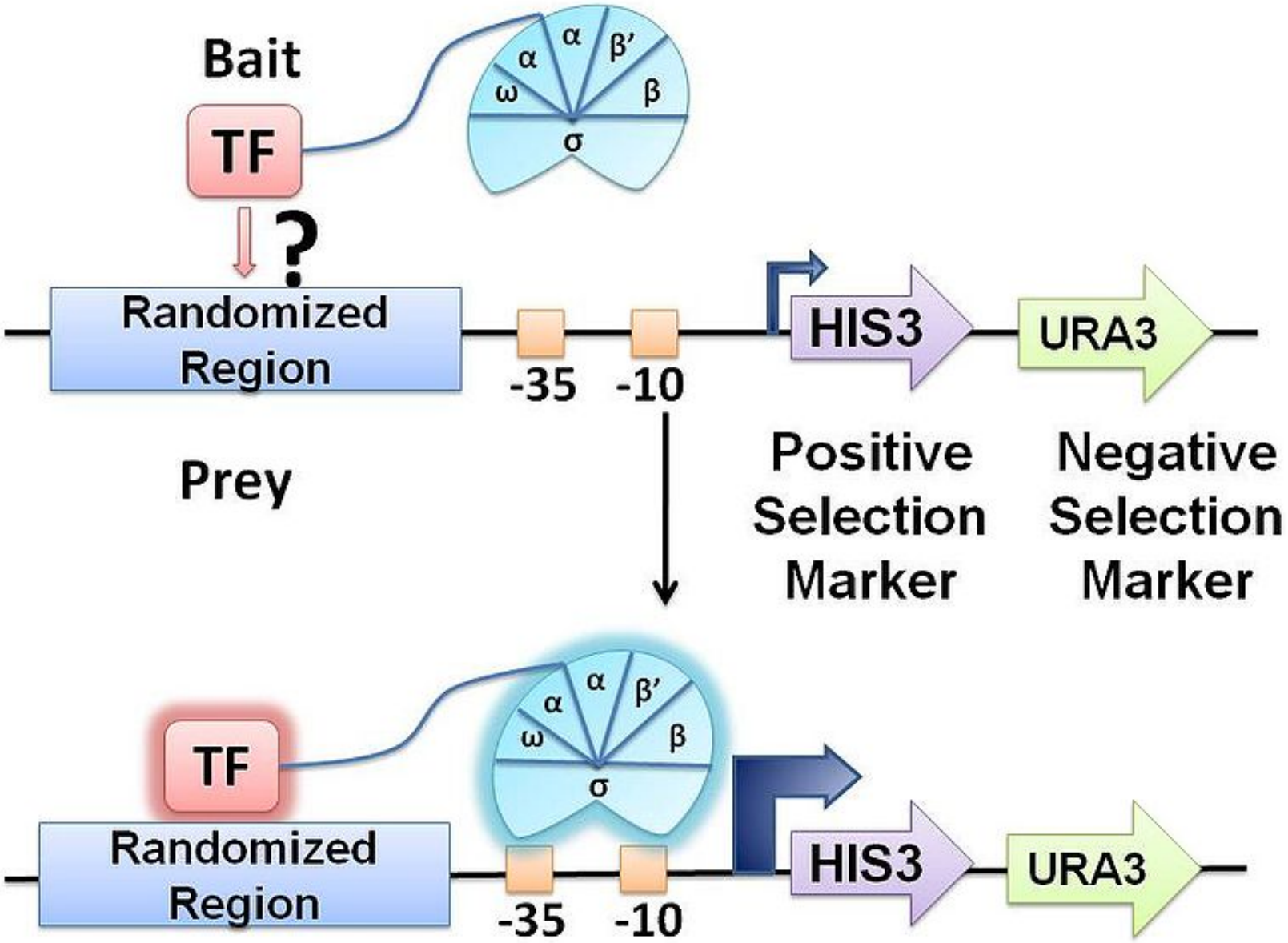
Library diversity determined by deep sequencing.



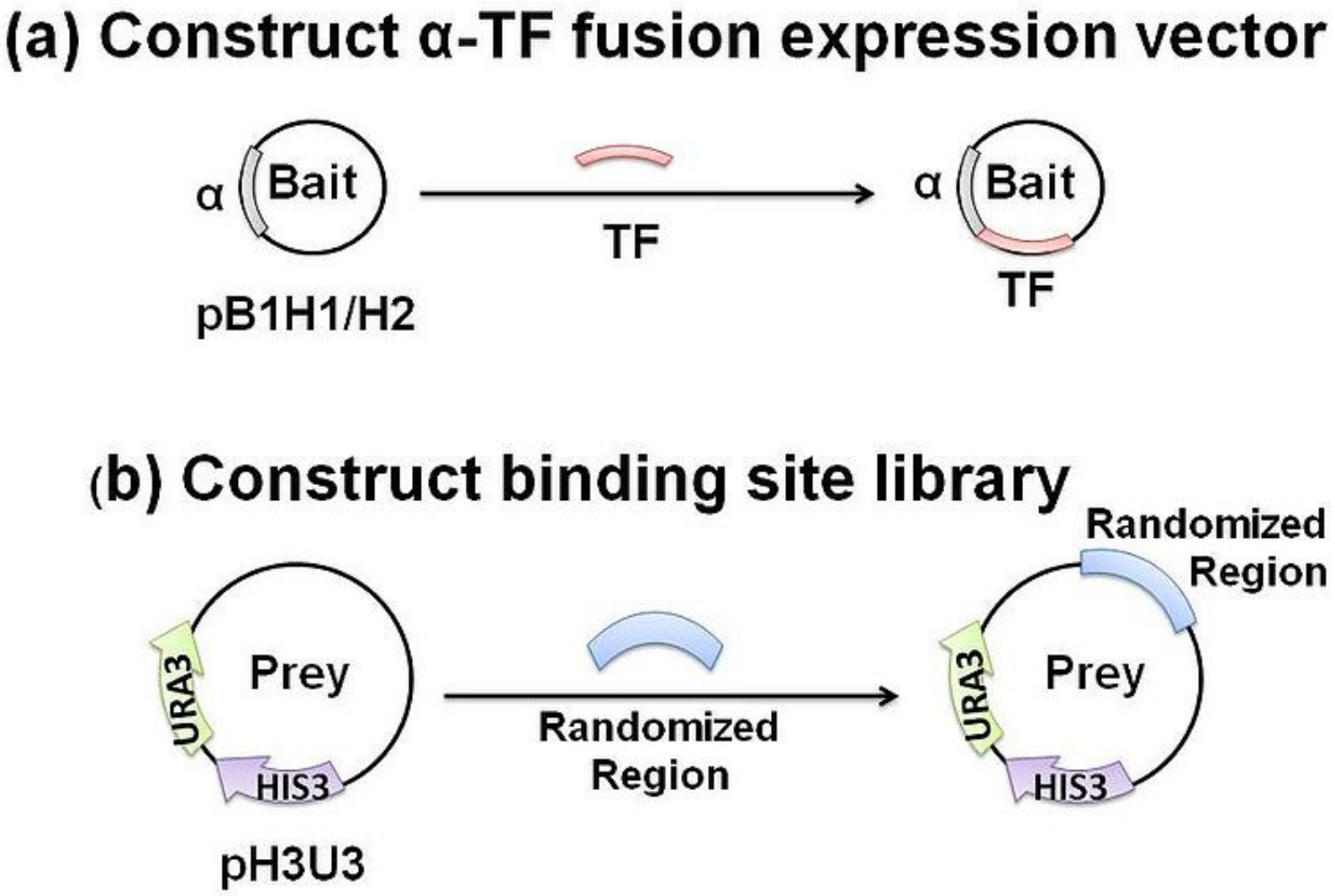
Persikov A V et al. Nucl. Acids Res. 2014;42:1497-1508

Nucleic Acids Research

Bacterial one-hybrid system (B1H)



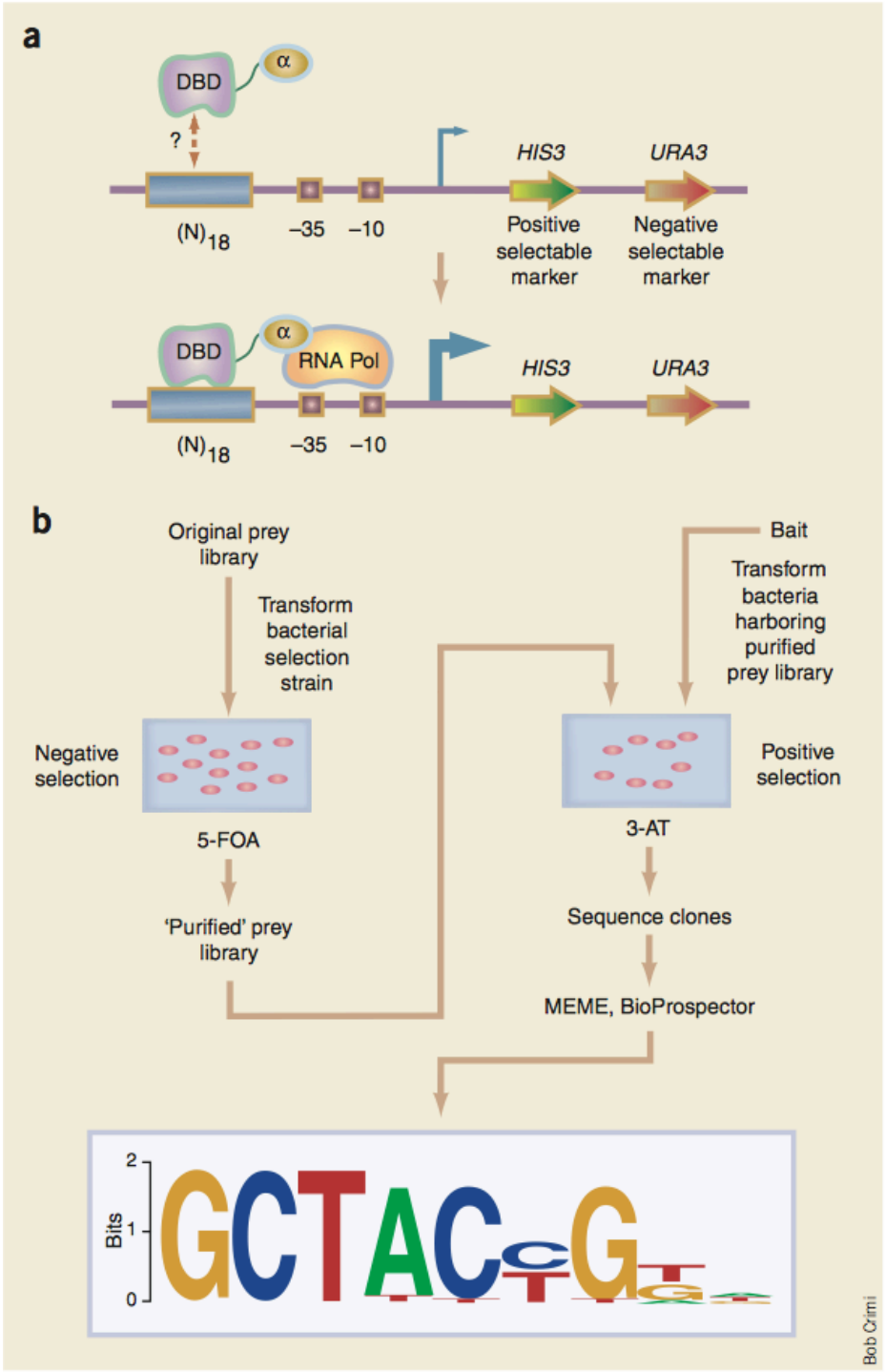
The bacteria one-hybrid system requires two customized plasmids:



Meng, X., Brodsky, M.H. & Wolfe, S.A.

A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors.

Nat. Biotechnol. **23**, 995–1001 (2005)



Processing zinc finger selection data:

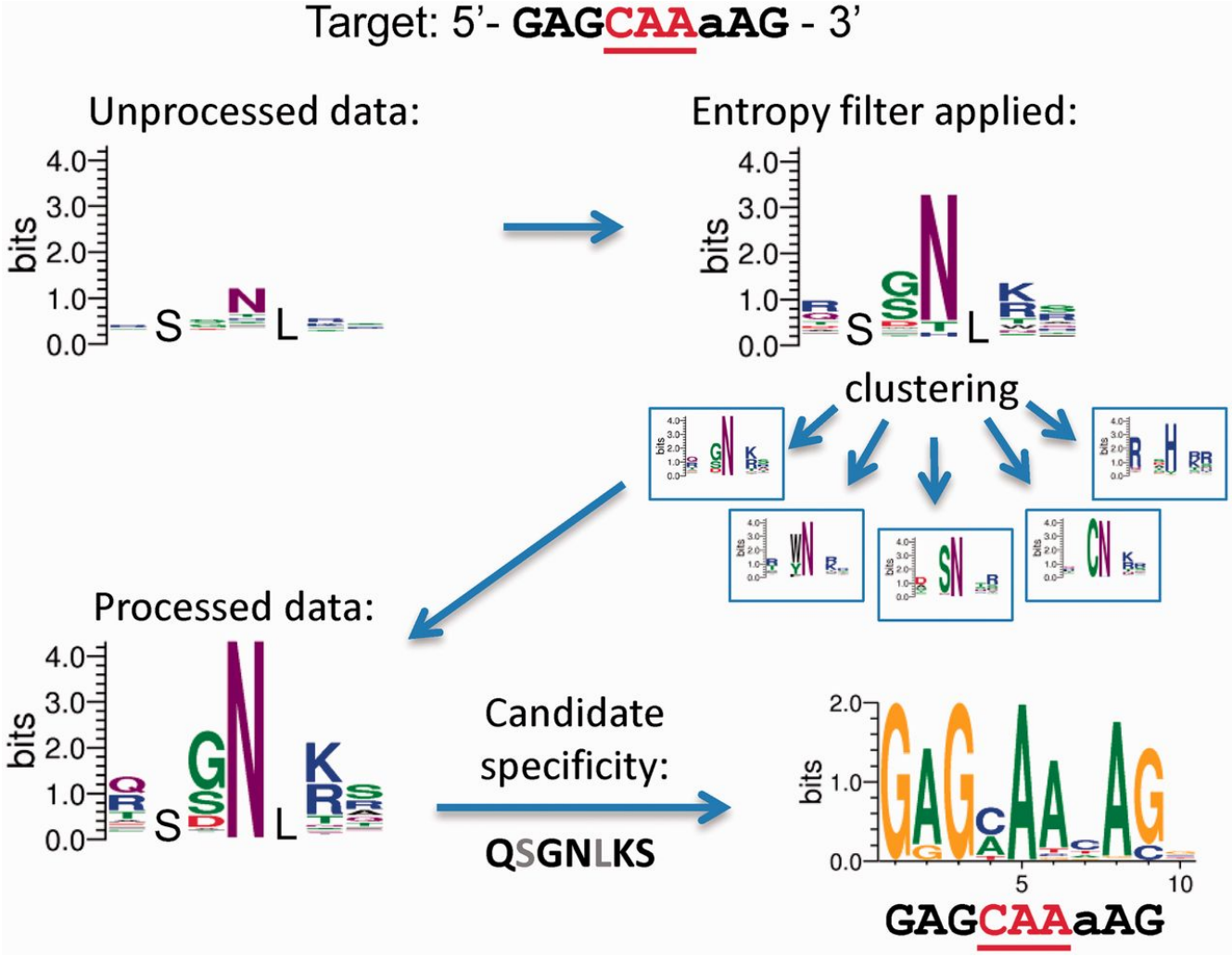
The Shannon entropy, normalized by the number of possible ways to code the amino acid sequence using NNS codons, was used as a measure of the diversity with which that particular amino acid sequence was observed:

$$E = -\frac{1}{\log_2 N} \cdot \sum_{i=1}^{N} p_i \log_2 p_i$$

where N is the total number of DNA variants to encode the current amino acid sequence and p_i the telative observation frequency of each such DNA sequence in the Illumina sequencing.

Amino acid sequences with either normalized entropy ≥0.25 or that can only have a single possible encoding with NNS codons were retained for further analysis; the rest were removed, as the lack of diversity in their underlying coding sequences suggests that they may be artifacts.

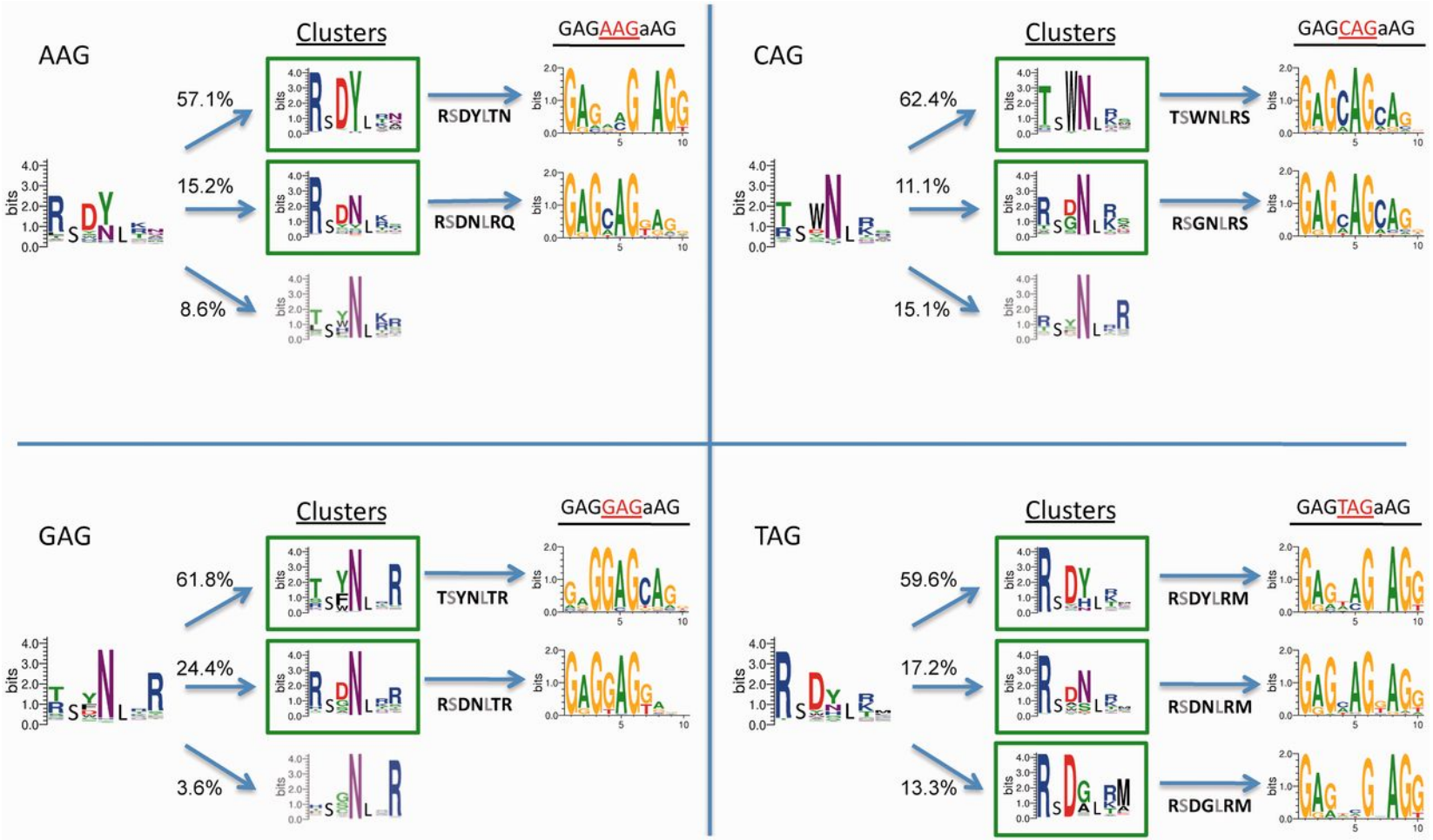
Overview of the computational pipeline for zinc finger selection analysis.



Persikov A V et al. Nucl. Acids Res. 2014;42:1497-1508

Nucleic Acids Research

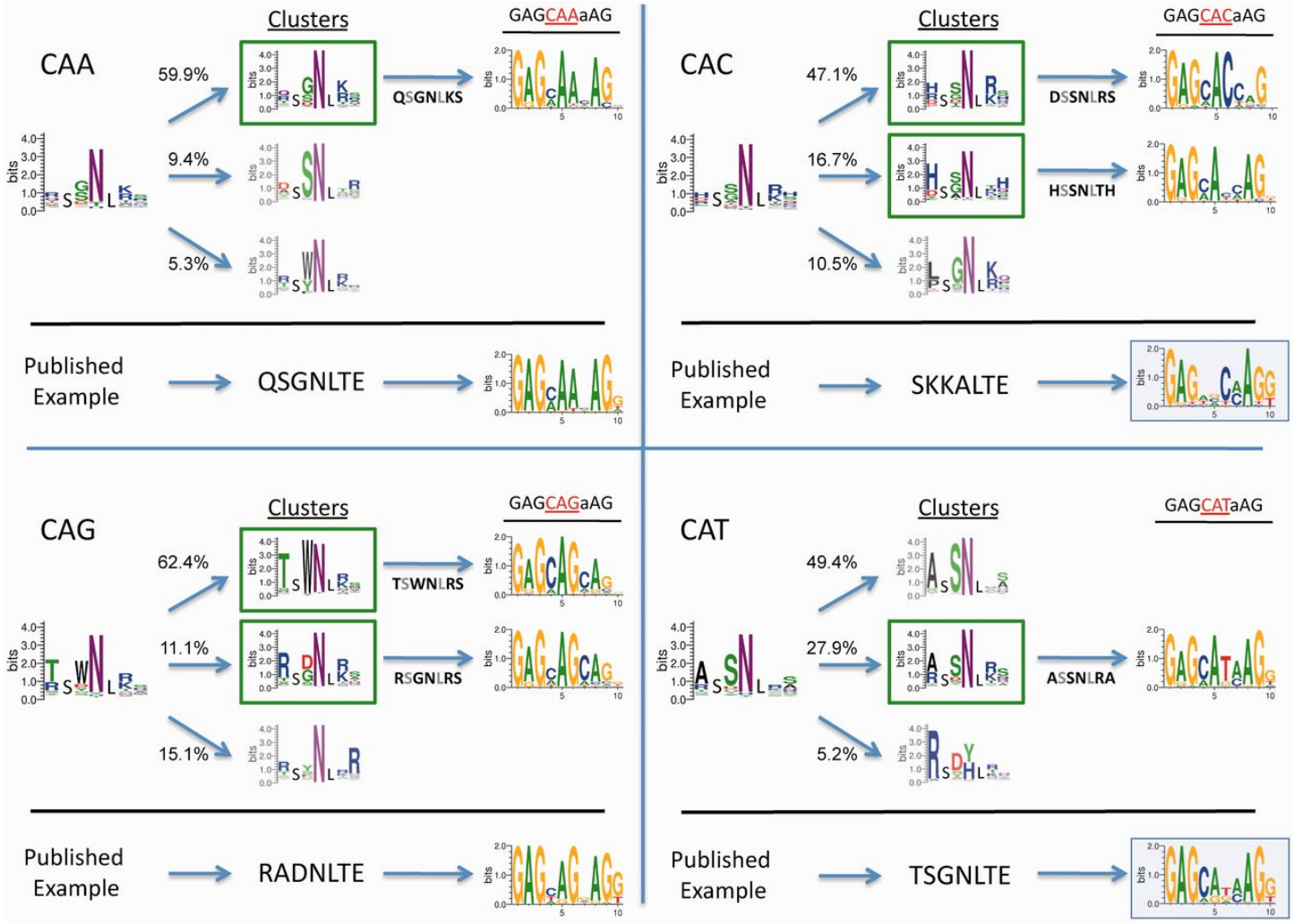
Selection of zinc fingers to bind all nAG targets.



Persikov A V et al. Nucl. Acids Res. 2014;42:1497-1508

Nucleic Acids Research

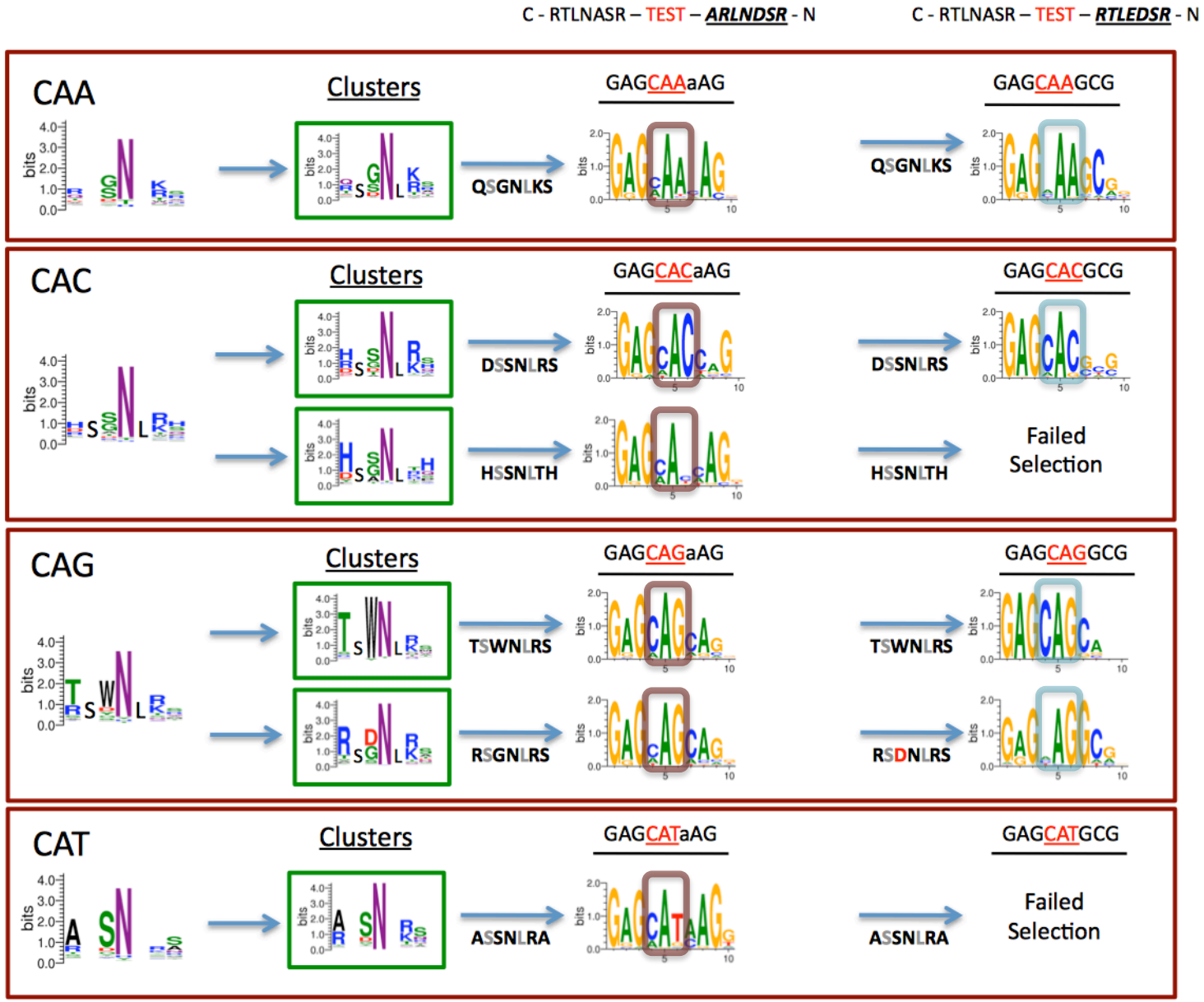
Selection of zinc fingers to bind all CAn targets.



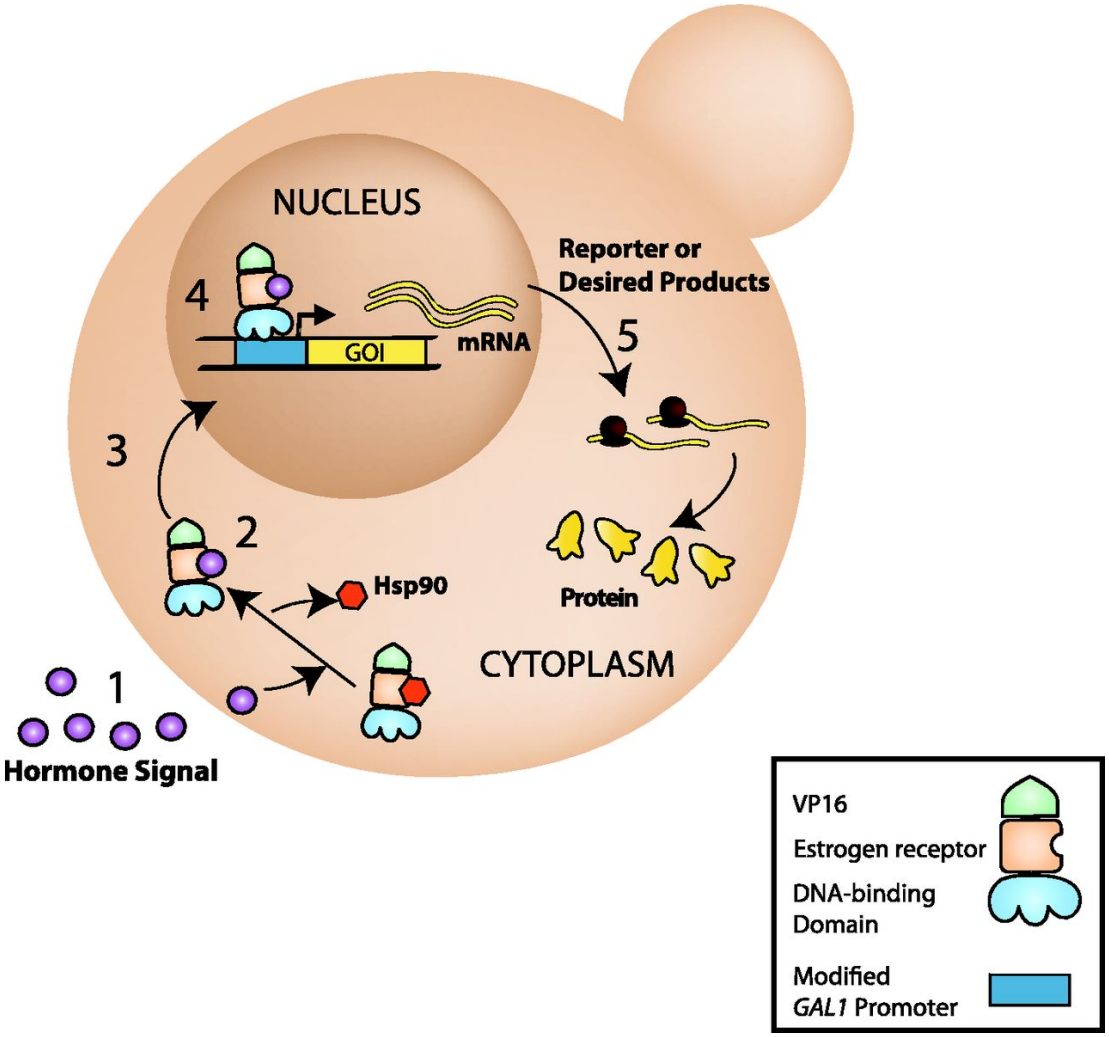
Persikov A V et al. Nucl. Acids Res. 2014;42:1497-1508

Nucleic Acids Research

Figure S4


C - RTLNASR – TEST – ARLNDSR - N
C - RTLNASR – TEST – RTLEDSR - N
CAA
Clusters
GAGCAAaAG
GAGCAAGCG
QSGNLKS
QSGNLKS
CAC
Clusters
GAGCACaAG
GAGCACGCG
DSSNLRS
HSSNLTH
DSSNLRS
HSSNLTH
Failed Selection
CAG
Clusters
GAGCAGaAG
GAGCAGGCG
TSWNLRS
RSGNLRS
TSWNLRS
RSDNLRS
CAT
Clusters
GAGCATaAG
GAGCATGCG
ASSNLRA
ASSNLRA
Failed Selection

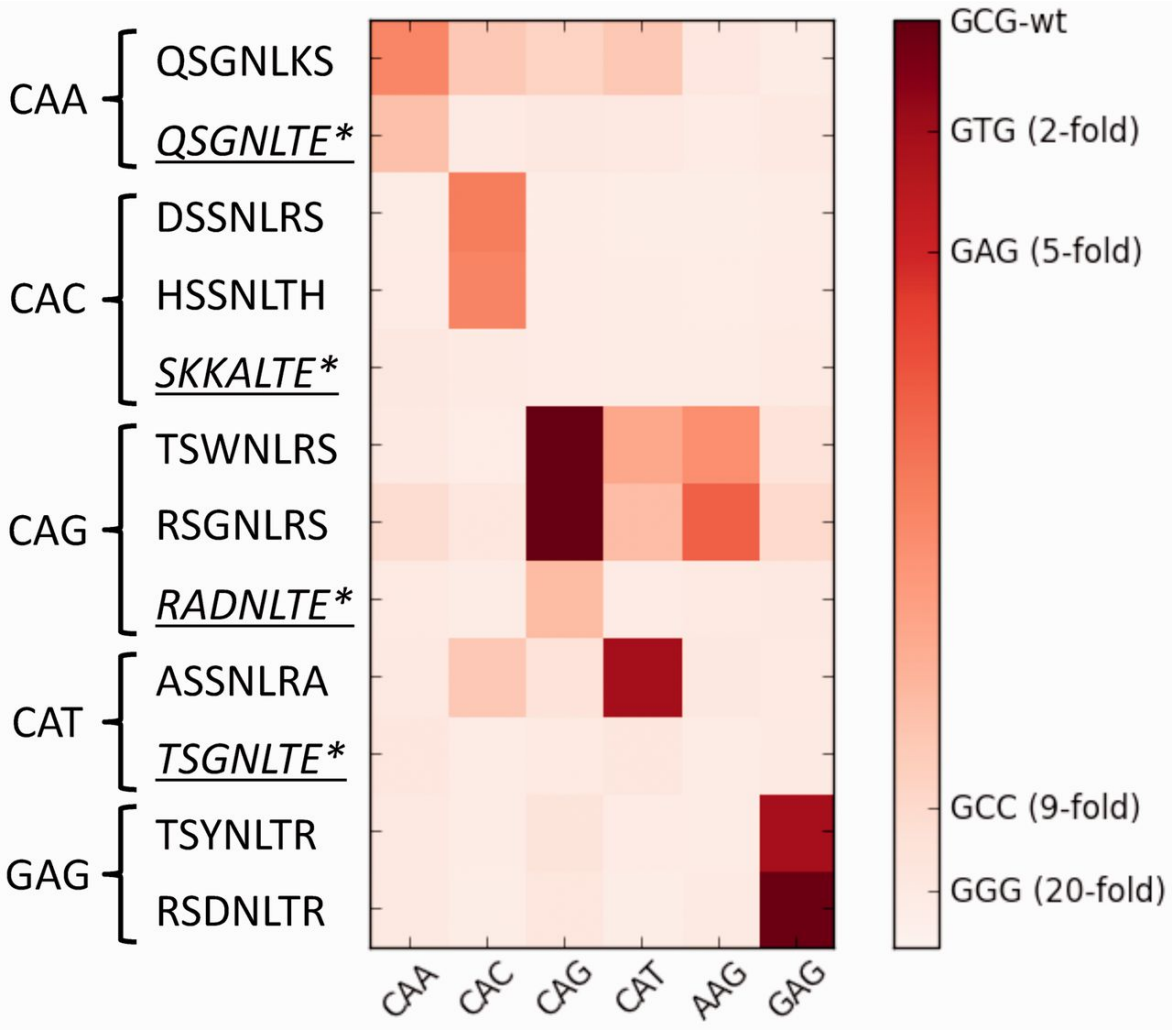
Schematic of hormone-based gene expression system.



McIsaac R S et al. Nucl. Acids Res. 2013;41:e57

Nucleic Acids Research

Artificial TF induction of GFP in yeast.



Persikov A V et al. Nucl. Acids Res. 2014;42:1497-1508

Nucleic Acids Research

Zinc finger library builds by PCR-driven cassette mutagenesis. Important points at the core of the optimized method:

- **First,** a separation of 20°C between the designed annealing temperature of the library oligo and the temperature performed in the PCR reaction minimizes any amplification advantage that one library member may have over another.
- **Second,** many reactions (48-96 per fragment) were carried out and pooled to dilute any bias that might occur in one reaction versus another.
- **Finally,** design of a restriction site at the 5′ end of the oligo allows for capture of full-length library fragments. These fragments can be captured by digestion and ligation with another fragment of DNA.

Applicability:

- The data can be leveraged **to build predictive models of zinc finger binding specificity and to assemble zinc finger proteins with desired binding specificities.** In the future, these techniques can be used to select new zinc finger domains to bind all possible targets.

- According to the authors, the high-throughput approach has the potential to greatly influence genome and protein engineering. Scaling this approach up may provide a deep understanding of the zinc finger and potentially other protein domains, **allowing for the prediction of naturally occurring proteins and the design of novel ones.**