Clustif: Motif-based clustering of short peptides for protein engineering applications

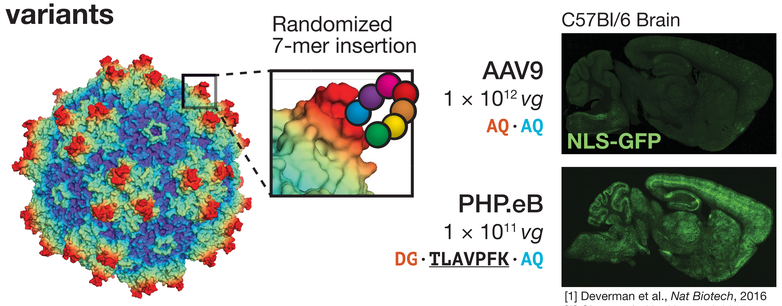
Albert Tian Chen[1,2], Yujia Alina Chan[1], Benjamin E. Deverman[1,*], Fatma Elzahraa-Eid[1]

1 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA
2 Department of Bioengineering, Northeastern University, Boston MA

Summary

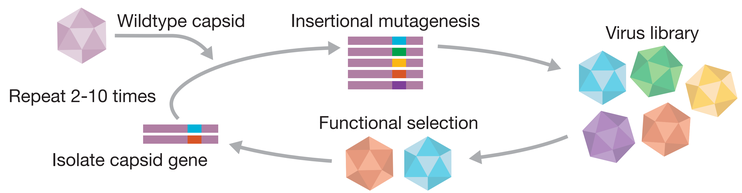
- Protein engineering with gene libraries enables researchers to rapidly design and screen new functional variants of proteins
- Typically, promising variants are selected through time-consuming sequential rounds of selection
- Clustif is a computational tool that identifies patterns among promising peptides from screens containing millions of variants, as identified by NGS
- Clustif accelerates variant discovery and enables subsequent mutagenesis and protein engineering efforts

Protein engineering enables discovery of new functional variants

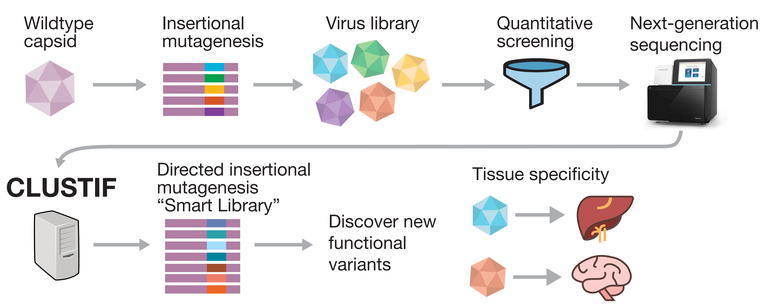


[1] Deverman et al., *Nat Biotech*, 2016
[2] Chan et al., *Nat Neurosci*, 2017

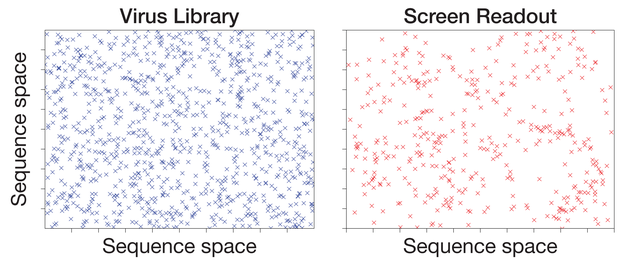
Directed evolution via sequential selection is time-consuming and labor-intensive



Can we use a single round of high-throughput screening to identify functional variants/patterns?

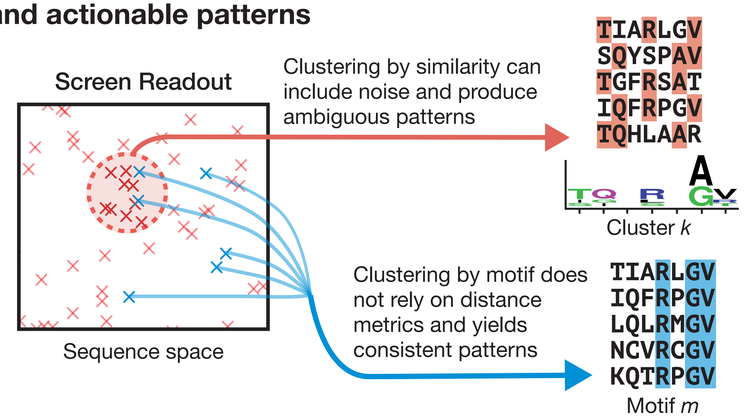


NGS results are dominated by noise and false positives

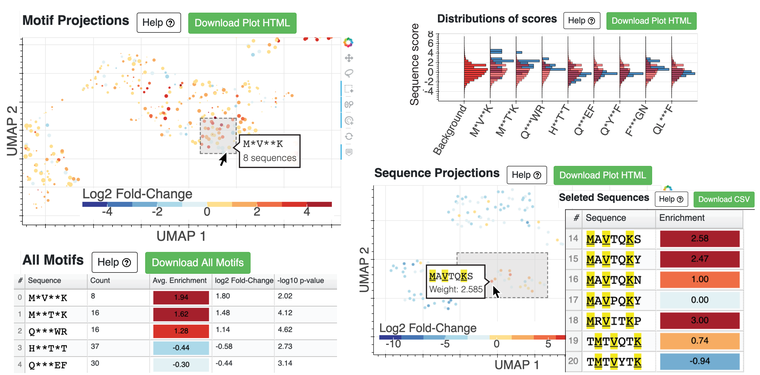


Identifying motifs enriched by screening will enable subsequent targeted mutagenesis studies.

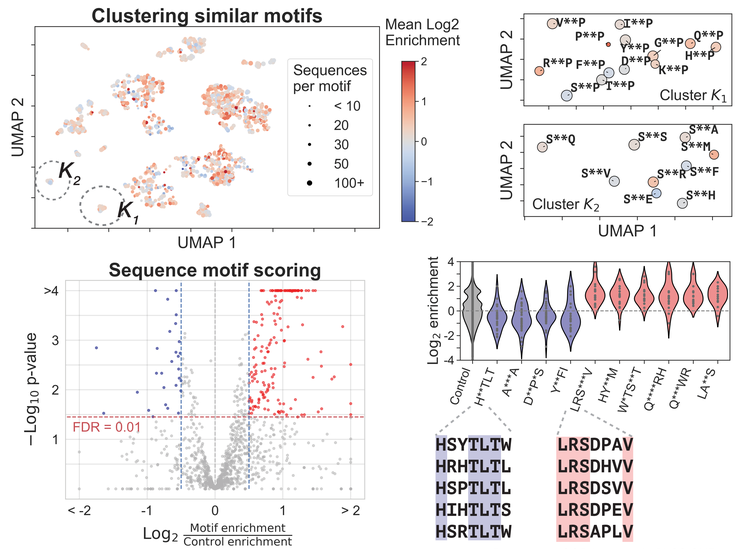
Clustering by sequence motifs produces interpretable and actionable patterns



Integration of quantitative screening data within an interactive web application

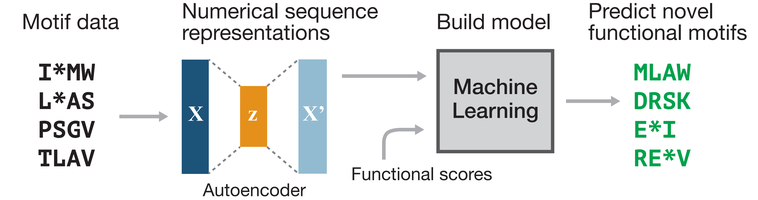


Clustif identifies functionally enriched motif clusters from a publicly-available phage display dataset



Data source: [3] Brinton et al., *PLOS ONE*, 2016

Motif–score associations enable *in silico* predictions of novel functional motifs



Conclusion

- Single-round quantitative screens can uncover many promising variants and patterns, but are dominated by noise and false positives
- Clustering by sequence motifs removes noise and allows association of patterns with functional scores
- Clustif enables targeted second-round screens for identification of functional variants in protein engineering applications

Acknowledgements

We thank K. Chan and the Vector Engineering Team for feedback and discussion. This project was supported by a Broad Shark Tank award, the Stanley Center for Psychiatric Research, and an NIH Somatic Cell Genome Editing Consortium grant (NINDS 1UG3NS111689).

References

[1] Deverman et al., *Nat Biotech*, 2016. [2] Chan et al., *Nat Neurosci*, 2017. [3] Brinton et al., *PLOS ONE*, 2016