Optimally choose barcode targets for CRISPR/Cas9 to enrich DropSynth perfect gene assemblies

Mindy H. Tran[1,2], Calin Plesa[2], Natanya K. Villegas[1,2], Yukiko R. Gaudreault[1,2], James A. Stapleton[2]

[1]University of Oregon, [2]Department of Biology

[2]University of Oregon, Phil and Penny Knight Campus for Accelerating Scientific Impact, Eugene, OR

DropSynth: a multiplex gene assembly method

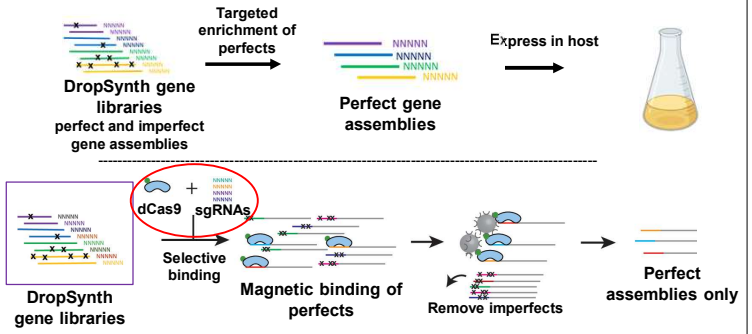


Figure 1: Approach for CRISPR-dCas9 targeted enrichment

Using deactivated Cas9 (dCas9)

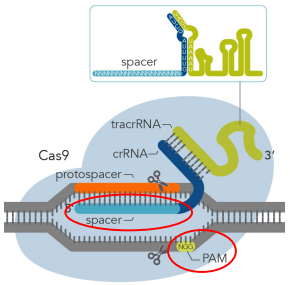


Figure 2: Bound Cas9/sgRNA complex

- sgRNA: single-guide RNA
- PAM: Proximal adjacent motif
- Spacer: 20 nucleotides upstream of the PAM
- By identifying the target sequences, we can encode RNA for dCas9 to locate the perfect assemblies

Seed collisions: part of the spacer that has more weight in triggering the binding of Cas9

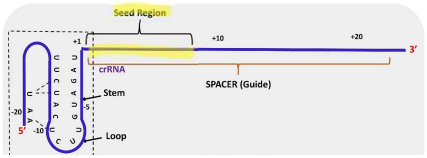
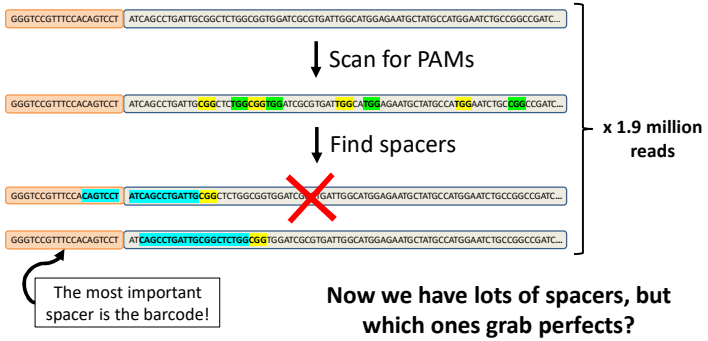


Figure 3: seed region of the sgRNA

If we can identify unique spacers of perfect assemblies, we can make sgRNA and enriches the perfect assemblies.

Identify Spacers



Now we have lots of spacers, but which ones grab perfects?

Compile Perfects

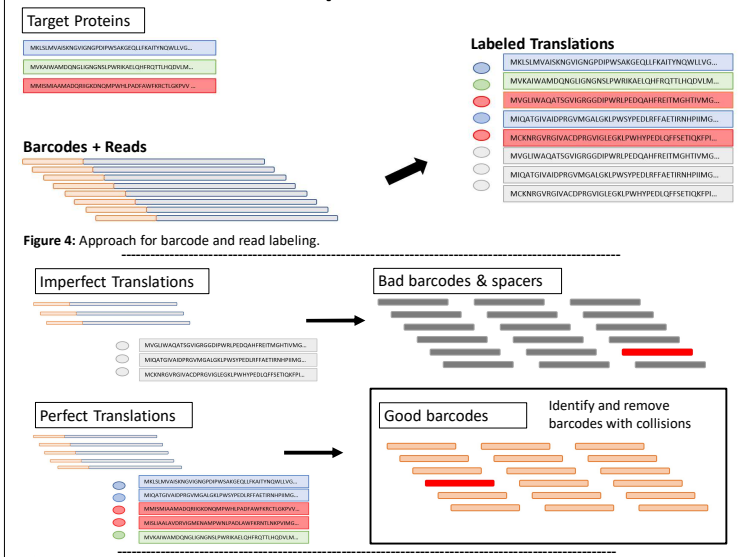


Figure 4: Approach for barcode and read labeling.

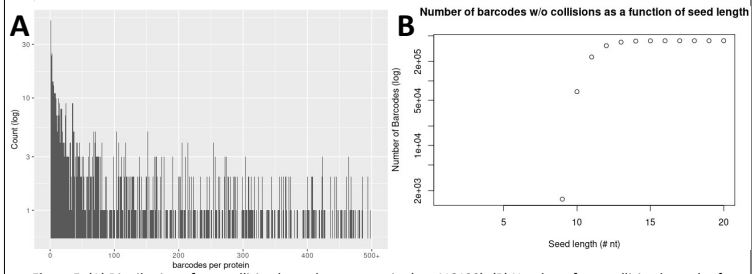


Figure 5: (A) Distribution of non-collision barcodes per protein. (n = 418120). (B) Number of non-collision barcodes for different seed lengths. For longer seeds, the barcode may seem more unique than it is, risking off-target effects.

Normalize Reads

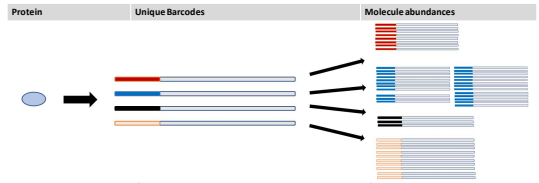


Figure 6: Example of the molecule abundance distribution for one protein.

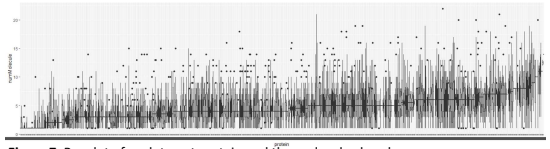


Figure 7: Boxplot of each target protein and the molecule abundances

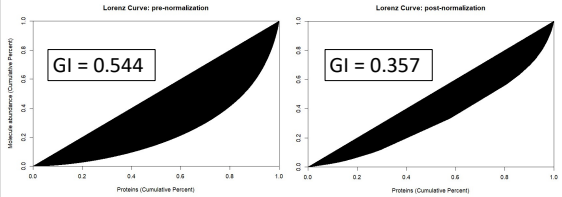


Figure 8: Lorenz Curve and GINI coefficient prior to and following normalization for 40k sequences. Reduced GINI coefficient by approximately 34%.

Selecting barcodes that are close to the median of median molecule abundances.

- We can reduce the number of outliers
- Avoid proteins with high median abundances from being overrepresented when expressed in host

Conclusions

- Developed modular data pipeline architecture that simplified >1.9M data points and identified target DNA sequences in R
- Performed data labeling, data cleaning, and data modeling to optimize and reduce DNA libraries by approximately 80%
  - Reduced >40k sequences to approximately 8k barcodes
  - Reduced >1.9 million sequences to appx. 418k
- Given a set of target proteins, we can modularly select barcodes and create a secondary library encoding for RNA for use with dCas9.

Future Directions

- Off target effects
  - Barcodes that are have high similarity the target barcodes are incorrectly picked up
- Algorithm to score barcodes/spacers
  - Score spacers on probability of having off-target hits
- Speed of code and maintaining architecture that reduces running time
  - Goal: complete barcode selection and normalization overnight (7-10 hours total)

References

1. Integrated DNA Technologies. 2014. What is the PAM sequence for CRISPR and where is it? https://www.idtdna.com/pages/Support/FAQs/what-is-a-pam-sequence-and-where-is-it-located
2. Plesa C, Sidore AM, Lubock NB, Zhang D, and Kosuri S. 2018. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science* 359: 343–47. DOI: 10.1126/science.aao5167
3. Sidore AM, Plesa C, Samson JA, Lubock NB, Kosuri S. 2020. DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions. *Nucleic Acids* Res. 48: e95. DOI: 10.1093/nar/gkaa600
4. Wu X, Kriz AJ, Sharp PA. Target specificity of the CRISPR-Cas9 system. Quant Biol. 2014 Jun;2(2):59-70. doi: 10.1007/s40484-014-0030-x. PMID: 25722925; PMCID: PMC4338555.

Acknowledgments

Phil and Penny Knight Campus for Accelerating Scientific Impact

Knight Campus Undergraduate Scholars Program