Summary

Victoria Li, Manuel E. Lladser

September 2014 - May 2016

Hsu et al. [1] conducted the first systematic study of CRISPR-Cas9 off-target frequencies in a mammalian genome. Thus we shall use their data to test our model. Given the variation across different studies, we list important characteristics of Hsu et al.'s study here:

System – *S. pyogenes* CRISPR-Cas9 (wild type)

Cell – human embryonic kidney (HEK293FT)

Gene – human EMX1, 15 target sites ([1] Supp. Figure 6)

sgRNA – sgRNA(+85) (see [1] Figure 1a and Supp. Figure 5):
5' gRNA (20 nt) + direct repeat (12 nt) + artificial stem loop (4 nt) + tracrRNA (nt's 23-85) 3'

Dosage – equimolar sgRNA cassette and Cas9 plasmid, 3×10^{-10} nmol/cell

Transfection – Lipofectamine 2000

Replicates – 2 biological replicates

Sequencer – deep sequencing, Illumina MiSeq Personal Sequencer

Sequenced – regions flanking target sites ([1] Supp. Figure 6)

Filtering – For each target: MiSeq read average Q score ≥ 23, barcode and forward primer perfectly match, $\geq 85\%$ matches in Smith-Waterman alignment to 20 nucleotide (nt) target + 50 nt upstream/downstream, no part of alignment outside of MiSeq read, indel within 20 nt target + 5 nt upstream/downstream, indel Q score $> \mu - \sigma$ of negative control MiSeq read Q scores

MLE of mutation frequency

Hsu et al. [1] varied the gRNA and hence the targeted sequence, then detected occurence of indels which signified target and off-target mutations in their experiment and negative control. The negative control consisted of cells not transfected with CRISPR-Cas9. Mutation frequency p was calculated for each of two bioreplicates as a maximum likelihood estimate (MLE) from a binomial error model:

$$p = \underset{p}{\operatorname{argmax}} \binom{R(1-p)}{n-Rp} q^{n-Rp}(1-q)^{R-n} \tag{1}$$

where R is the total number of reads, n is the number of reads with observed indels, and q is the fraction of negative control reads with indels.

The resulting p estimates reported by Hsu et al. (see [1] Figure 2 and Supp. Table 5) were not all reproducible. Also, the final p estimates were not quite the average of bioreplicates. We attempted to reproduce their estimates as follows.

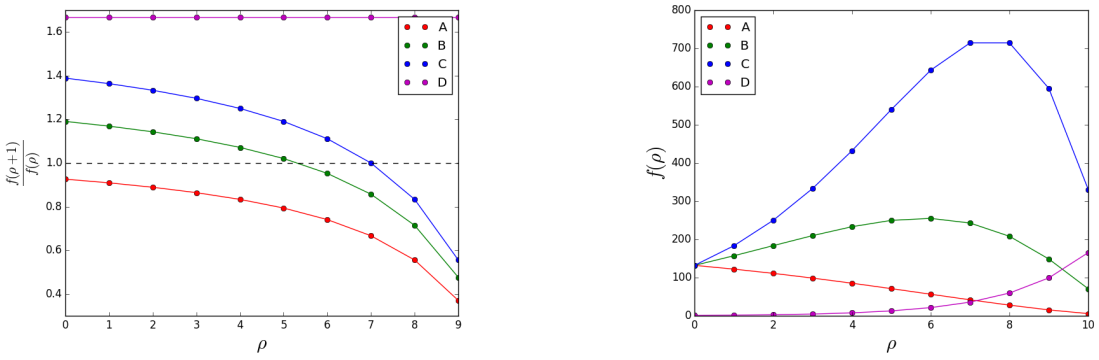

Figure 1: **Illustration of possible cases for maximization of f.** Left: plot of $\frac{f(\rho+1)}{f(\rho)}$, with $0 \leq \rho \leq n-1$, for each of the four stated scenarios. Right: plot of $f(\rho)$, with $0 \leq \rho \leq n$, for each scenario. Note that scenario C results in two ρ values that maximize f.

Forward problem

Given n, q, and R, we calculate p. Define $\rho = Rp$, which represents the estimated number of reads with true indels. We assume this takes integer values. Maximizing Equation 1 is equivalent to maximizing

$$f(\rho) := \frac{(R-\rho)!}{(n-\rho)!} q^{-\rho} \tag{2}$$

for $\rho \in \{0, 1, ..., n\}$. Moreover, since

$$g(\rho) := \frac{f(\rho+1)}{f(\rho)} = \frac{1}{q}\left(1 - \frac{R-n}{R-\rho}\right) \tag{3}$$

for $\rho \in \{0, 1, ..., n-1\}$ is strictly decreasing starting from $g(0) = n/qR$, there are four possible cases (Figure):

A. $q > \frac{n}{R}$: Maximum at endpoint

B. $q < \frac{n}{R}$, $\exists\, \rho$ s.t. $g(\rho) < 1$, $\nexists\, \rho$ s.t. $g(\rho) = 1$: Maximum at unique local max

C. $q \leq \frac{n}{R}$, $\exists\, \rho$ s.t. $g(\rho) = 1$: Maximum at two local max

D. $q < \frac{n}{R}$, $g(\rho) > 1$ always: Maximum at endpoint

Then the maximizer of f can be found either at an endpoint or by setting $g(\rho) = 1$. The resulting MLE estimate(s) of p is:

A. $p = 0$

B. $p = \frac{1}{R}\lceil \frac{n-Rq}{1-q} \rceil$ or $\frac{1}{R}\lfloor \frac{n-Rq}{1-q} \rfloor$

C. $p = \frac{n-Rq}{R(1-q)}$ and $\frac{n-Rq}{R(1-q)} + \frac{1}{R}$

D. $p = \frac{n}{R}$

Inverse problem

Since Hsu et al. reported p, q, and R but not n values in their data (see [1] Supp. Table 5), we instead calculate n given p, q, R and verify that it reproduces the p estimate. First compute $\rho = Rp$. Then the possible n values are:

$$n \in \{\lfloor \rho \rfloor, \lfloor \rho \rfloor + 1, ..., R\}.$$

For each possible n value, we solve the forward problem, obtaining 1 or 2 p estimates p_n. To check how well this reproduces the data, we calculate the error, obtaining 1 or 2 error values $\varepsilon_n = |p_n - p|$. All n values that meet the following criteria are accepted as solutions:

I. 2 p_n values $p_{n,1}$ and $p_{n,2}$:

$$p \in [p_{n,1}\,,\, p_{n,2}]$$

OR

II. 1 p_n value:

$$n = \underset{n}{\operatorname{argmin}}\, \varepsilon_n$$

Data reproduction

The above calculations were programmed in Python, and the code is attached. Many n value solutions were generated for each bioreplicate of each experiment, corresponding to the data in Supplementary Table 5 of [1]. All n and ε_n values are listed next to the corresponding Hsu et al. data in the supplement to this article (supp5emx1.1.xlsx, supp5emx1.2.xlsx, supp5emx1.3.xlsx, and supp5emx1.6.xlsx). We successfully reproduced most p estimates with error $\sim 10^{-5}$. However, several p estimates could only be reproduced with minimal error 0.5 or more. The highest minimal error was 0.83. Given that p represents a probability between 0 and 1, these errors are a serious issue.

We shall use Hsu et al.'s data for training purposes only until a reproducible dataset is available. In future, we will also try reproducing their data with continuous ρ values.

References

[1] Hsu, P.D., Zhang, F. et al. (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotech.* **31**(9), 827-832.