Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs

Listgarten J et. al. Nature Biomedical Engineering volume 2, pages 38–47 (2018)

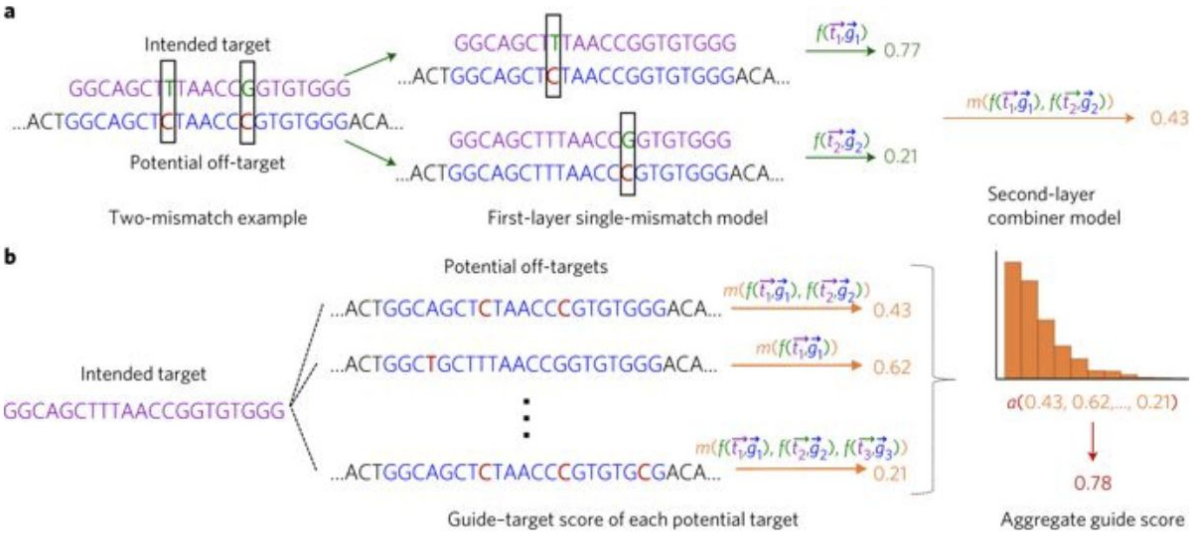
Reviewed by: Eric Wang
University of Virginia
https://qdata.github.io/deep2Read/

Three main tasks of off-target predicting

(1)
Search and filter genome wide for potential targets for one gRNA. For example, those regions of the genome that match the gRNA up to N number of nucleotide mismatches to the target site. Note, these sites are not deemed to be active off-targets until after step 2, which uses machine learning to distinguish the targets that are expected to be active from those that are not. This merely creates a short-list of potentially active sites.
(2)
Score each potential target for activity from step 1. That is, assign a numeric value that indicates how much off-target activity is expected for one gRNA–target pair.
(3)
Aggregate the scores from step 2 into a single off-target potential with which to assess the gRNA.

Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs

Overview


a
Intended target
GGCAGCTTTAACCGGTGTGGG
...ACTGGCAGCTCTAACCCGTGTGGGACA...
Potential off-target
Two-mismatch example
GGCAGCTTTAACCGGTGTGGG
...ACTGGCAGCTCTAACCGGTGTGGGACA...
$f(\vec{t_1},\vec{g_1})$ 0.77
GGCAGCTTTAACCGGTGTGGG
...ACTGGCAGCTTTAACCCGTGTGGGACA...
$f(\vec{t_2},\vec{g_2})$ 0.21
First-layer single-mismatch model
$m(f(\vec{t_1},\vec{g_1}), f(\vec{t_2},\vec{g_2}))$ 0.43
Second-layer combiner model
b
Potential off-targets
Intended target
GGCAGCTTTAACCGGTGTGGG
...ACTGGCAGCTCTAACCCGTGTGGGACA...
$m(f(\vec{t_1},\vec{g_1}), f(\vec{t_2},\vec{g_2}))$ 0.43
...ACTGGCTGCTTTAACCGGTGTGGGACA...
$m(f(\vec{t_1},\vec{g_1}))$ 0.62
...ACTGGCAGCTCTAACCCGTGTGCGACA...
$m(f(\vec{t_1},\vec{g_1}), f(\vec{t_2},\vec{g_2}), f(\vec{t_3},\vec{g_3}))$ 0.21
Guide-target score of each potential target
a(0.43, 0.62,..., 0.21)
0.78
Aggregate guide score

a, An example of how to score a gRNA–target pair with two mismatches. First, the gRNA–target pair is broken down into two single-mismatch pseudo-pairs, $(\{t_1, g_1\},\ \{t_2, g_2\})$, (where t_i and g_i, respectively, denote the target and gRNA in the *i*th pair), each of which is scored with the first-layer (single-mismatch) model, $f(\overrightarrow{t_1}, \vec{g}_1)$. Then, these scores are combined with the second-layer model, $m(f(\overrightarrow{t_1}, \vec{g}_1), f(\overrightarrow{t_2}, \vec{g}_2))$, yielding a single gRNA–target score that accounts for all mismatches. **b**, An example of how to aggregate the set of gRNA–target scores for a single gRNA into one summary off-target score for a gRNA. The aggregator model, *a*(), computes statistics of the input distribution of gRNA–target scores as features and runs them through a model, producing the aggregate score for a gRNA (for example, 0.78).

[Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs]

Individual gRNA-target pair off-target predictive modelling

- Asymmetry for off-target prediction: it is generally more consequential to mistake an active off-target site for an inactive one, rather than the other way around
  - first type of error can disrupt the cell or confound experimental interpretation
  - second may only require designing another gRNA.

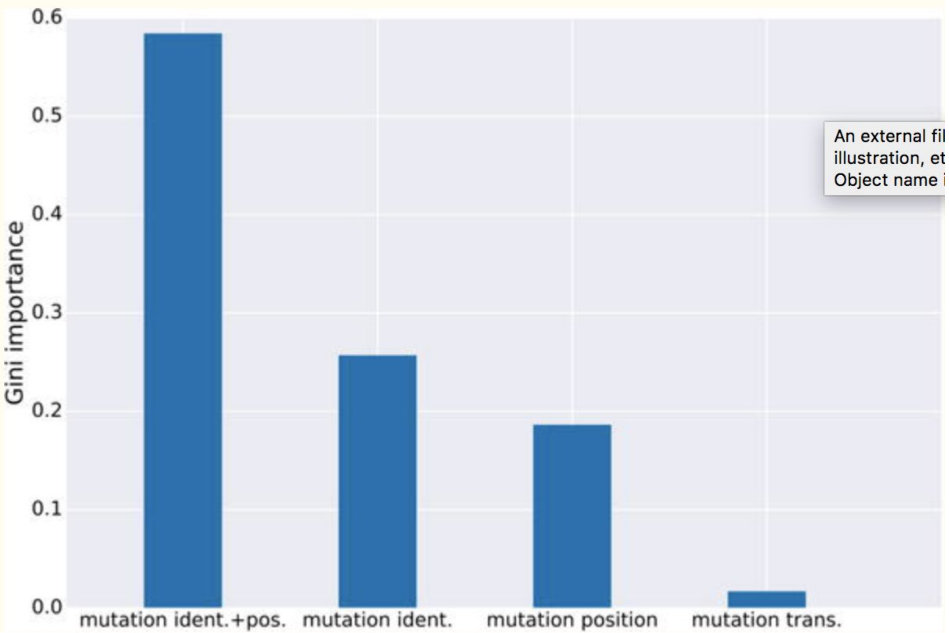
“...Consequently, we chose an evaluation measure which accounts for this asymmetry—the **weighted Spearman correlation,** where each gRNA-target pair is weighted by an amount which is a (monotonic) function of its measured activity.

Because the precise asymmetry is not *a priori* known and may vary for different applications, we varied the weight continuously between two extremes: from being directly proportional to the measured activity (such that false negatives effectively do not count), to a uniform weighting (*i.e.*, yielding standard Spearman correlation)...”

Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs

Individual gRNA-target pair off-target predictive modelling

- For first-layer (single-mismatch) model features
  - (i) the **position** of the mismatch
  - (ii) the nucleotide **identities** of the mismatch
  - (iii) the **joint** position and identities of the mismatch in a single feature
  - (iv) whether the mutation was a transition or transversion.



[Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs]

Individual gRNA-target pair off-target predictive modelling

“...It is interesting to note that using both the joint “position and mismatch nucleotide identity” features—those effectively used by CFD—are aided by additionally **decoupling** these into additional features of **position and nucleotide identity**, even though regression trees can in principle (with enough data) recover the joint features from the decoupled ones…”

[Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs]

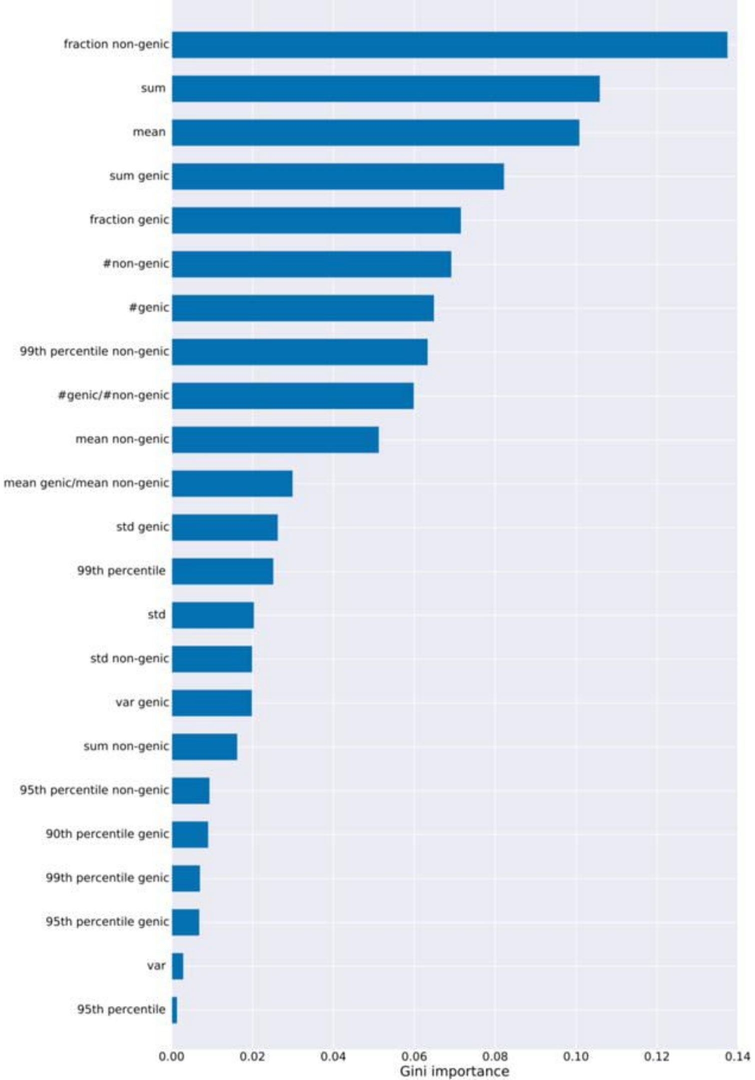
Aggregating individual off-target scores into a single gRNA summary score

- The end task, of aggregation, requires obtaining a single off-target ***summary*** *score* for a gRNA given all its individual gRNA-target scores.
- To evaluate our approach on this task we made use of two data sets with gRNAs targeting non-essential genes in viability screens, the Avana[1] and Gecko[32]libraries. Because each gRNA is designed to target one non-essential gene in these screens, the cell should be viable if no off-target effects are present.
- In particular, at least three papers have shown evidence that a cell is more likely to die when sustaining numerous DNA breaks.[1,32,33] Additionally, a fourth paper leverages this phenomenon to assess off-target cutting.[34] Therefore, there is now substantial evidence that **cell viability** is determined at least in part by the **number of DNA breaks** per cell.
- A second effect on viability could be off-target activity at an **essential gene**. However, essential genes cover merely 0.2% of the human genome and are therefore not likely to have much effect in our experiments. To further elucidate this point, we evaluate the performance of our model using only gene essentiality as a feature, which performs vastly worse than when we either ignore it altogether, or additionally use the scores from our gRNA-target model as features. This then empirically shows that **gene essentiality** is not adversely affecting our conclusions using the viability data. Hence these viability-based experiments serve as bronze standard for the combined task of scoring and aggregation.

Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs

Aggregating individual off-target scores into a single gRNA summary score

The importance of each aggregation feature



[Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs]

Conclusion

- We have introduced the first machine-learning based approach to predictive modelling of off-target effects for CRISPR-Cas9. Through systematic investigation we demonstrated that our newly developed suite of models, Elevation, performs better for each of the two main off-target-related tasks in gRNA design: gRNA-target scoring and aggregation.

- Additionally, we are the first to systematically evaluate available competing approaches on the task of summary scoring (aggregation), showing that Elevation consistently outperformed competing approaches by a substantial margin.

- We also considered how to balance errors between active and inactive gRNAs, developing a new metric to do so, based on the **weighted Spearman correlation**. This type of evaluation encapsulates a range of practical use cases, and enabled us to show that Elevation is consistently superior across the entire range.

- As data become available for a richer set of scenarios, including different endonucleases, different organisms and *in vitro* versus *in vivo*, **epigenetics** on more cell types, we will update our models and tools accordingly.

Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs

GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases

- Shengdar Q Tsai
- , Zongli Zheng
- , Nhu T Nguyen
- , Matthew Liebers
- , Ved V Topkar
- , Vishal Thapar
- , Nicolas Wyvekens
- , Cyd Khayter
- , A John Iafrate
- , Long P Le
- , Martin J Aryee
- & J Keith Joung

Nature Biotechnology volume 33, pages 187–197 (2015)

Main

- The identification of indels or higher-order rearrangements that can occur anywhere in the genome is a challenge that is not easily addressed, and sensitive methods for unbiased, genome-wide identification of RGN(RNA-guided nuclease/CRISPR-Cas9)-induced, off-target DSBs in living cells have **not yet been described**

- Whole genome resequencing has been used to attempt to identify RGN off-target alterations in edited single-cell clones[14,15], but the **exceedingly high projected cost** of sequencing very large numbers of genomes makes this method impractical for finding low-frequency events in cell populations.

GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases

Main cont.

- Focused deep sequencing: identify indel mutations at potential off-target sites identified either by **sequence similarity** to the on-target site or by ***in vitro* selection** from partially degenerate, binding-site libraries. However, these approaches are **biased** because they assume that off-target sequences are closely related to the on-target site and, as a result, may miss potential off-target sites elsewhere in the genome.

- ChIP-seq has also been used to identify off-target binding sites for gRNAs complexed with **catalytically dead Cas9** (dCas9), but the majority of published work suggests that very few, if any, of these sites represent off-target sites of cleavage by active Cas9 nuclease.

GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases

Results

- GUIDE-seq consists of two stages
    - In stage I, RGN-induced DSBs in the genomes of living human cells are tagged by integration of a blunt, double-stranded oligodeoxynucleotide (dsODN) at these breaks by means of an end-joining process consistent with NHEJ (nonhomologous end-joining).
    - In stage II, dsODN integration sites in genomic DNA are precisely mapped at the nucleotide level using unbiased amplification and next-generation sequencing.

GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases

Summary

- State of the art off-target evaluation: based on direct base-pair mismatch, data on cell viability, whole-genome sequencing of edited cells - high cost.
  - WGS data limited by price and efficiency
  - Cell viability data may not reveal subtle but clinically critical off-target mutations

- Problem: finding a dataset that’s large, easy to generate/replicate, and provides reliable labeling of off-target events, e.g.
  - Did cleavage occur?
  - Did insertion/deletion take place at the DSB?
  - Are the off-target effects harmful or benign?