Comparative Study of Predictive Models for Mutations Generated by Repair of Cas9-induced Double-strand Breaks

Jideofor Ezike
jezike@mit.edu

Frederik Heymann Lassen
lassen@mit.edu

I. Introduction

Crispr-Cas9 is a powerful gene editing tool that leverages components of the bacterial immune defense system to effectively and quickly target DNA sequences. It has enabled researchers to more easily edit their favorite genes compared to preexisting genome editing tools (TALENs, Zinc Finger Nucleases). While other techniques require a more laborious process to generate the constructs that recognize target DNA, the Crispr-Cas9 system only requires the presence of a PAM sequence flanking the target sequence, making this system a robust form of editing or disrupting gene function. Upon being led to the target sequence via a complementary guide RNA, the Cas9 nuclease cuts the target 3 bp upstream of the PAM sequence, resulting in activation of endogenous cellular DNA repair mechanisms[1].

The DNA repair pathways induced upon Cas9 nuclease cutting fall into two general camps: Homologous Directed Repair (HDR) and Non-homologous end joining (NHEJ). There is also a third branch of pathways, integral to our analysis, called microhomology mediated endjoining (MM-EJ) which refers to when the presence of homologies in both strands of DNA mediate repair of a double strand break (DSB). For the remainder of the paper we will refer to MM-EJ to include alternative end joining (a-EJ) and single strand annealing (SSA). These various repair pathways in eukaryotes are summarized in Figure 1A. When cutting occurs without a homologous template of DNA present, the more error prone processes of NHEJ / MM-EJ dominates, resulting in insertions or deletions (indels) within the target gene sequence of interest [2]. Certain factors within the cell also dictate which DNA repair pathway dominates, such as stages of the cell cycle as well as presence of particular enzymes[3]. These pathways are well understood and are relied upon to introduce disruptions within genes. However, until recently little effort has been put into understanding the actual distribution of indels that can result from DNA repair. Given that there is a large feature space of indel classes that can occur for any guide-target sequence pair, being able to predict the distribution with high accuracy would be useful for experimental design in the lab space as well as in therapeutic contexts in which you may want to put a gene back into frame.

In this comparative study we test, compare and interpret three models that predict mutational outcomes of template-free Crispr Repair: Indelphi, Lindel and ForeCast [4][5][6]. The models are slightly different and employ different formulations of feature engineering, model architecture, training data and objective functions. We aim to display the strengths and shortcomings of each method and propose a model that builds off the lessons learned from each approach. Furthermore, the primary goal is to investigate their qualitative and quantitative differences when tested on new unseen data. Here, we discuss the effect of using different cell types in order to delineate potential differences in prediction mechanisms. We hope that these results will generate further insight into the models and guide the selection of models depending on the task at hand.

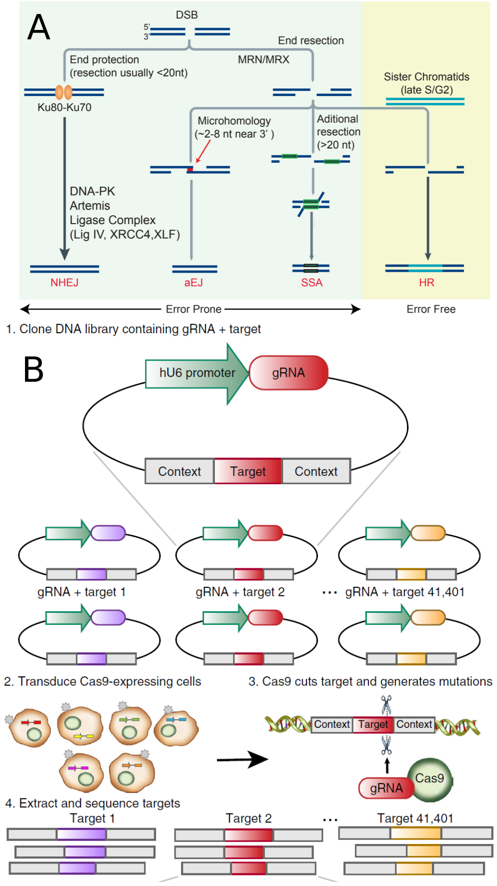
II. Methods

In the following section, we first describe the experimental setup in acquiring the data used for testing these models. Then we briefly describe how we collect and process the data. Lastly, we describe the different approaches employed to conduct a thorough analysis of their performance.

A. *Experimental Design and Data Acquisition*

The experimental design in generating the data used for these models were very similar. Firstly, randomly generated oligonucleotides are mass produced; these serve as the target sequences. These sequences are then cloned into DNA constructs that contain the following: a human promoter that drives expression, as well as the target sequence's complementary guide-RNA. This construct is then introduced to Cas9 expressing cell-lines via lentiviral transfection. This procedure can be accomplished in a host of cell lines. After 5-7 days post infection (to allow for ample cutting events), the DNA from these cells are then deeply sequenced, generating the raw data. The raw data comes in the form of the DNA sequencing reads of the gene products that arise from Cas9 cutting. We also have access to pre-processed ground truth data that summarizes the indel mutational profile for each guide RNA-target pair. This data was provided by the authors of the Lindel study[5], and was extended to the analysis of the other two models. This analysis is based on randomly sampling 3000 oligo target sequences for each cell type.

When deciding which data to test the pre-trained models on, we aimed to select data from cell types that were unseen to the models during training. This would allow us to make a claim regarding the relative generalizability of the three models. InDelphi was trained primarily with mouse embryonic stem cells (mESCs), FORECasT with K562 cells and Lindel with HEK293T cells. Therefore we tested the pretrained models' ability to predict mutational profiles specific to Chinese Hamster Ovary (CHO) and Retinal Pigmented Epithelial 1 (RPE1) cells. We coded three platforms that allow for both processing of target sequences using the respective model prediction pipelines (found on github) as well as generation of accuracy/performance metrics.



jnnn

Fig. 1: ***A****: Illustrates repair mechanisms in mammalian cells. Double stranded breaks can be repaired through different repair pathways in mammalian cells and yield unique products. NHEJ operates by end resection and ligation of blunt ends. A-EJ pathway operates with microhomology and can result in insertions or deletions. SSA is analogous to A-EJ but results in deletions of greater length. Homologous Recombination (HR) is the only pathway that is error free. (Adapted from 7.58 Lecture Slides).* ***B****: General Experimental Design. Introducing guide RNA/target sequence pairs into Cas-9 expressing cell lines followed by deep sequencing of resulting DNA products. (Adapted from Allen et al.[4])*

B. Repair Profile Analysis

For each model, we collected the normalized distribution of actual and predicted repair profiles. Each mutation contained a unique identifier describing the mutation type and position with respect to the cut-site, i.e. how far away from the cut site the mutation occurred. We decided to compare the difference between the actual and predicted repair profile using the KL-divergence illustrated in equation 1,

$$D_{KL}(P\|Q) = \sum_i P_i \; log\frac{P_i + \epsilon}{Q_i + \epsilon} \tag{1}$$

where i corresponds to the different indel genotypes, and P_i and Q_i are the relative proportions of the actual and predicted mutations respectively. We decided to add a small pseudocount ϵ in order to avoid division by zero. To simplify the analysis, we decided to define ”indel genotype”, or ”indel type” to be the length of the indel, independent of positioning (e.g. D1 is an example of a genotype with a deletion of length one).

The frequency of unique indels across the different models was assessed. Since the models struggle to effectively represent specific mutations, and certain mutations occur with very low frequency in actuality, we decided to only consider genotypes of deletions of 1-30 in length and insertions of 1-10 in length.

Furthermore, in order to assess the utility of these models for potential gene perturbation experiments, we investigated their ability to faithfully capture the proportion of mutations that yielded an in-frame product. Specifically, indels whose length were a multiple of 3 were defined to be in-frame or non-frameshift mutations.

Lastly we decided to define and use our own makeshift accuracy metrics. We thought that since normally a few indel genotypes dominate in a given mutational profile, while the rest drop off very quickly in prevalence, we would reward models that were able to predict to some extent the top few indel types. We employ a 0-1 loss in three metrics: 1) Assign a model an accuracy of 1 for predicting the most highly represented indel, and zero otherwise (T1 metric). 2) Assign a model an accuracy of 1 for predicting at least 1 out of the top 2 indels correctly, and zero otherwise (T2 metric). 3) Assign a model an accuracy of 1 for predicting at least 1 out of the top 3 indels correctly, and zero otherwise (T3 metric). These accuracy metrics for each model are summarized in Table 2.

III. Models

In this section we briefly detail the individual model architecture, objective functions and training schemes relevant for performance comparison.

A. FORECasT

FORECasT (Favored outcomes of repair events at Cas9 targets) is a computational predictor of CRISPR cas9 mutational profiles for a given guide RNA. The model has been trained using logistic regression while minimizing the L2-regularized symmetric KL-divergence between actual and predicted mutation profiles. The model was trained on guide-target pairs from K562 cells.

Comprehensive feature engineering was implemented in order to formulate a machine learning problem. Firstly, candidate indels were generated by considering all possible and relevant deletions and insertions relative to the cut size. For each of these candidates a set of 3633 binary features were generated containing information regarding the sequence composition, relative nucleotide position, and microhomology. Additionally, in order to consider interaction effects, multiple features were combined into one single feature. In other words, more than one feature would have to be present in conjunction, in order to contribute to the output.

$$p_k = \frac{exp(\theta x_j)}{\sum_i exp(\theta x_i)} \tag{2}$$

The probabilities of all outcomes were modelled using logistic regression according to equation 2.

The optimal L2 regularized hyperparameter was obtained using grid search. We have that x_j is the feature vector and θ is a vector of weights. This approach is analogous to softmax, in which one obtains discrete probabilities for each of the feature engineered possible mutation events that sum up to one.

B. InDelphi

InDelphi is another computational framework used to predict the frequency of major editing outcomes from distinct cut sites. The model is compromised of three major components that work in concert to predict MH deletions, MH-less deletions or 1-basepair insertions. The first component of inDelphi simulates the microhomology mediated repair mechanism (MM-EJ). This yields multiple unique deletion mutations in which different parts of the homology have been considered. Each homology candidate is subsequently processed by a neural network that scores a candidate deletion mutation by its GC content and length. Scores are normalized and the relative frequency of each candidate can therefore be modelled. The choice of features was decided by the authors through experimentation with different architectures, in which it was observed that strong microhomologies were usually long and with a high GC content. Deletion events that can not be adequately modelled by homology directed repair are shaped by another module. This secondary component models ”microhomology less” deletions using the minimum required resection length as an input feature. This module was set up with inspiration from MM-EJ which usually results in deletions[6]. Again, this feature choice was determined through experiments and observations indicating that microhomology-less depending deletions are inversely proportional to sequence length. The third and final component models 1 bp insertions. This component is based on k-nearest neighbour indicating a high influence of local sequence context.

The microhomology dependent and micro homology less neural networks were trained jointly on target DNA that was incorporated into mamallian cells in a multi-task problem using backpropagation. Due to slight differences in cellular repair mechanism, individual models were trained on mESC, HCT116, HEK293, K562 thus making the model specific towards a given cell type. In the following, we restrict the analysis to the model trained on mESC, as the author claims it have the best predictive performance on previously unseen cell types[6].

C. Lindel

The Lindel model first defines the number of possible indel event classes based on their fixed criteria of considering deletion events less than or equal to 30 in size that overlap with a -3/+2 window around the cleavage site and all possible 1-2 bp insertion events at the DSB. Across all experimentally constructed targets they identified 584 event classes, of which only 563 were considered for downstream analysis. Since the majority of Crispr/MM-EJ mediated indels fall into one of 584 event classes, they framed their machine learning task as predicting for an arbitrary sequence the relative frequency of each of these 584 event classes. The eventual 563 event classes that they decided to include resulted from disregarding large insertions as these are rarely observed.

Subsequently to defining the mutational outcomes, they delineated the feature space characterizing the target sequence using binary one-hot encoded vectors. They defined a total of 2,962 binary features broken down as such: 1) 384 binary features corresponding to one-hot encoded sequence, including: 80 for single nucleotide content (4 nucleotides * 20 positions) and 304 for dinucleotide content (16 dinucleotides * 19 positions); 2) 2,578 binary features corresponding to MH tracts; they defined 5 binary features corresponding to MH lengths in the range 1-5. (5 bp * 563 deletion event classes = total 2,815 binary features). They excluded 237 feature vectors that were not observed in the training data, resulting in the final 2578 binary features used for encoding MH tracts.

Lastly, for their model architecture the authors tested both neural nets and L1 regularized logistic regression, and found that the logistic regression outperformed the neural net in performance. They formulated their optimization problem as minimizing cross entropy loss for the training set.

IV. Results

The KL divergence of each model was determined across the 3000 oligos for RPE1 cells and the 3000 oligos for CHO cells. The results are summarized in table I and further illustrated in figure 2. For the CHO cells, Lindel had the lowest KL-divergence average of 0.77 (median = 0.63), followed by InDelphi which had an average of 1.21 (median 0.77). FORECasT produced the overall highest KL-divergence with a mean of 2.66 (median 2.48). The distributions were skewed to the right. Although slightly higher, these results were generally in concordance with reported values in literature[6][5].

		FORECasT	InDelphi	Lindel
CHO	Mean	2.66	1.21	0.77
	Median	2.48	0.77	0.63
RPE	Mean	3.20	2.12	0.87
	Median	3.02	1.51	0.73

TABLE I: *The raw performance (KL-Divergence) of FORECasT, InDelphi and Lindel on either CHO or RPE1 cells.*

The source of the KL-divergence was also investigated for the RPE1 cells. We wanted to get a sense of the models' performance in predicting specific indel genotypes. This data is illustrated in Figure 3. For each model it can be seen that 1-bp insertions generally contributed the most to KL divergence, indicating that insertions were generally harder to predict than other mutations. Interestingly, FORECasT seems to have the most difficulty predicting low bp deletions, while Lindel struggled more so than the other models in predicting 1-bp insertions. Lastly, InDelphi's low bp deletion predictions (D1-D8) seemed to have a more evenly distributed contribution to KL divergence than the other models.

		FORECasT	InDelphi	Lindel
CHO	T1 Pred	50%	37%	57%
	T2 Pred	81%	82%	86%
	T3 Pred	93%	95%	94%
RPE1	T1 Pred	48%	34%	58%
	T2 Pred	80%	81%	85%
	T3 Pred	90%	94%	93%

TABLE II: *The raw performance metrics captured by the different models. The top consensus refers to whether the highest predicted mutation could be found within the top i of actual mutations.*

We further decided to investigate the accuracy of each model using our own accuracy metrics defined above. Firstly, we investigated how often there was an overlap between the most frequent predicted and actual mutation (T1). As shown in table II, none of the models did well in predicting the topmost ground-truth genotype, resulting in the worst accuracy out of the three accuracy metrics. Lindel appeared to outperform FORECast by producing both lower KL divergences as well as higher accuracy metrics. Lindel and FORECast had accuracies of 90-94% as defined by our T3 accuracy metric.

We wanted to see whether the frequency of mutations was faithfully reproduced by the different models. This is illustrated for the CHO and RPE1 cells in figure 2B. The following observations can be seen from the figure: Firstly, all models underestimated the true frequency of 1 bp insertions. However, this tendency was less significant in CHO cells compared to the RPE1 cells. Secondly, all models appeared to effectively capture the true frequency of deletions. However, it's interesting to note that InDelphi tended to overestimate the amount of 3-9 bp deletions while FORECasT tended to overestimate the amount of 25-30 bp deletions. Furthermore, while FORECasT did not capture any substantial evidence of 2-bp insertions, Lindel overestimated this number in both the CHO and RPE1 cells.

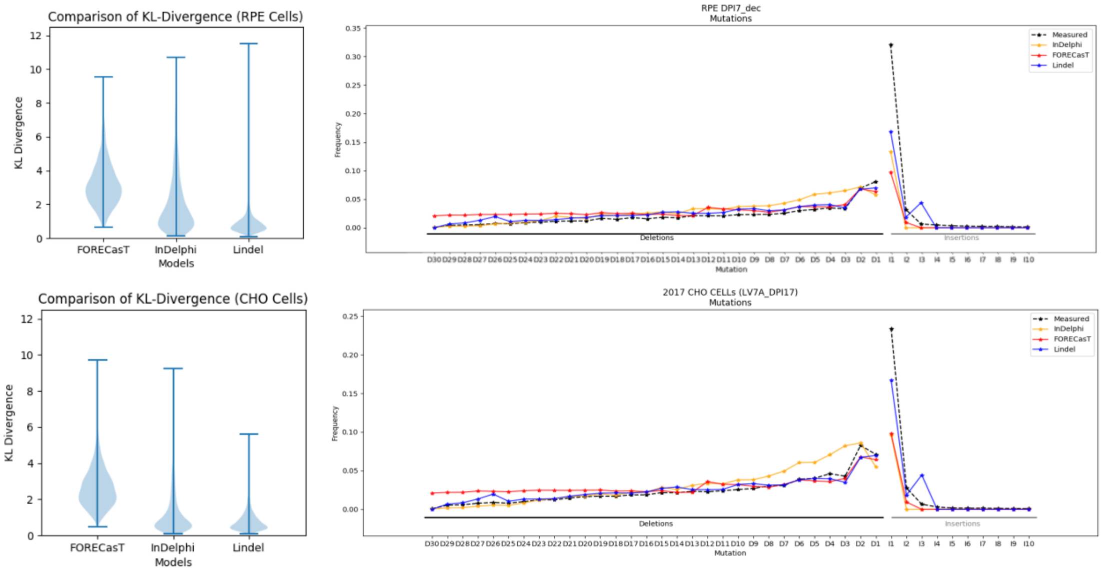


Fig. 2: ***Left Panels****: Illustrate the distribution of KL divergence across the different models for CHO (bottom) and RPE1 cells (top). It can be seen that Lindel and InDelphi generally achieves the highest performance.* ***Right Panel****: The frequency of the measured and predicted mutations for CHO (bottom) and RPE1 cells (top). It can be seen that the predicted mutations were more accurately represented in the CHO cells.*

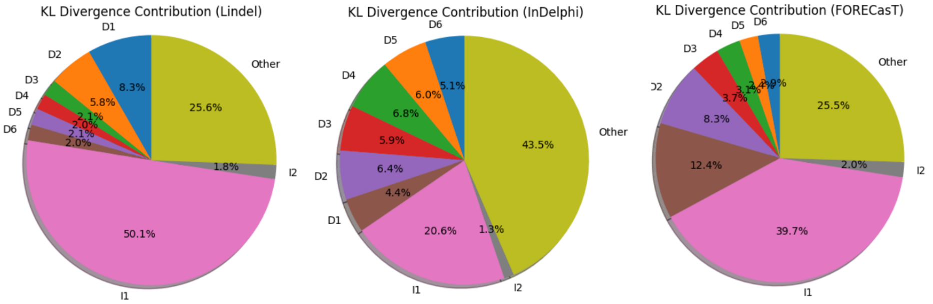


Fig. 3: *The KL-divergence contribution by mutations in the RPE1 cells only. As illustrated by the charts, the majority of KL-divergence came from an inability to adequately model insertions of length 1.*

Finally, the frequency of frame-shift mutations were analyzed. The frequency of predicted and actual in-frame mutations are depicted in figure 4. As seen by the figure, the predicted and actual outcomes were generally correlated with a Pearson correlation coefficient above 0.4 for all models. In general, the models achieved better accuracy on the RPE1 data compared to the CHO cell data. One exception to this was FORECasT which performed not only better on the CHO cell data, but produced the overall highest Pearson correlation of 0.65. Lindel performed significantly worse on the CHO cells (R=0.402) compared to the RPE1 (R=0.586). An interesting property of the models is that they are all slightly biased towards predicting a higher in-frame mutation frequency compared to the ground truth. This is seen in the correlation plots as the majority of data points are generally situated below the unity line.

V. Discussion

Traditionally, the 'gold-standards' for comparing probability distributions are either through 1) cross-entropy, which measures the propagation of information or 2) KL-divergence, which a distance function that measures the distance between two probability distributions. In fact, by fixing P in the KL divergence formulation, optimizing KL divergence is equivalent to optimizing cross entropy loss. That being said, either metric would have sufficed but we opted for KL-divergence as it naturally "measures faithfulness with which Q's distribution models P's" [7], providing a similarity measure, which is in essence what we want to observe.

Firstly, it was interesting to observe that the model with the objective to minimize KL divergence, FOReCasT, consistently had the highest KL divergence. This could be due to the way in which FORECasT had been trained; relative to the other models it may have over-fitted more to it's training data due to the properties of it's objective function. This result could also be explained by the lack of diverse training data representative of other cell types. Lastly, one could hypothesize that the logistic regression approach is not expressive enough for modelling mutational profiles and further extrapolating to newer cell types. However, this notion contradicts the decent performance of Lindel which is also based on logistic regression. Given more time, we would have liked to predict mutational profiles using the same cell types as in the original paper, thus allowing us to determine whether the high KL-divergence is actually due to an inability to generalize.

Interestingly, we observed that Lindel produced the lowest overall mean and median KL-divergence on both CHO and RPE1 cells, indicating that this model was better at generalizing for the two cell types. Given that we used the InDelphi model pretrained on mESCs, we also expected InDelphi to perform the best on CHO cells. Due to evolutionary similarity between the organisms from which CHO cells and mESCs are derived, we hypothesized that the KL-divergence for InDelphi predicting CHO cells would be smaller compared to the RPE1 cells. And in fact, the mean of the KL-divergence was doubled on RPE1 cells compared to CHO cells, indicating that the genetic background of the target cell may indeed have influence on model capacity.

From our results, we could confirm that the models were biased towards particular cell types. This was indicated by a disconnect between accuracy, median and average KL-divergence across different cell types. This was also in concordance with established theory, that the action of repair mechanism is dependent on cell specific conditions[8]. We would therefore also expect the best performance to be achieved using models trained on the exact cell type or organisms of common ancestry in which the cellular repair mechanism are conserved. Since the CHO and RPE1 cells are relatively more evolutionary distinct, we would also expect the models to achieve different performances.

We wanted to be able to understand the origin of KL divergence. So we decided to explain it by quantifying, for each model, the contribution of each indel genotype to the observed KL divergence. We found that for all of the models, insertions of size 1 were the most difficult to predict accurately. This may be explained by the fact that these models place more emphasis in encoding deletions in their feature engineering or simply a lack of training data containing insertions. Furthermore, MM-EJ is more likely to result in deletions, so the models have more deletion events to learn than insertions, which could make it biased towards predicting deletions.

Lindel RPE1 predictions for I1 genotypes con-

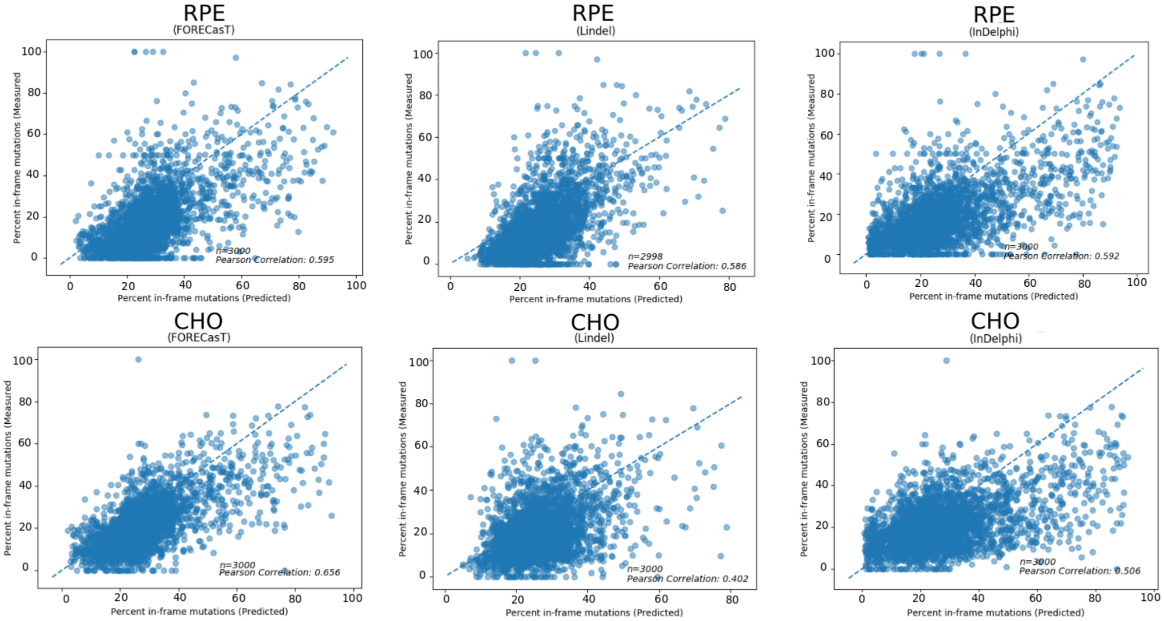


Fig. 4: *Depicts the correlations between the predicted and actual proportion of inframe mutations for 2998-3000 oligonucleotides. An inframe mutation is defined as any mutation that result in an insertion or deletion of a sequence that is a multiple of three. It can be seen that the accuracy in predicting the proportion of inframe mutation is cell specific, by comparing the difference in correlation between CHO and RPE1 cells.*

tributed more to the KL divergence relative to the other models. This could be due to a number of reasons. Firstly, the way in which we calculated the KL divergence contribution was to say that the indel with the highest absolute value of $|P_i - Q_i|$ contributed the most to the KL divergence. Since KL divergence measures distance between two distributions, we hypothesized that this would be a sound metric for measuring deviation from the ground truth distribution. However, this approach could bias towards assigning lower KL divergence contributions to lowly represented indels, as the variance for frequent indels is expected to be higher. The high KL divergence attributed to I1 in Lindel could be due to the fact that Lindel performs very well in tracking everything in the ground truth with the exception of I1 genotypes (figure 2), meaning that the contribution for individual genotypes other than I1 would be low, leaving a bigger piece of the KL contribution pie for I1 to inherit. Whereas for the other models, the subpar performance is more evenly distributed across the indel genotype landscape, resulting in the I1 KL contributions being relatively closer to the others.

Finally, we propose a few ideas to improve the models. One major source of KL-divergence in all the models originated from the 1-bp insertions as previously discussed. This is therefore an obvious target for improvement. InDelphi competitively models insertions and deletions. One potential improvement would be to change the K-Nearest Neighbour module to a more expressive non-parametric model. One could find another model that could capture local sequence context while simultaneously proving increased predictive power. One approach to achieve better performance would be to employ a support vector machine instead of the K-nearest neighbour module. This model is widely utilized in academia and would provide a relatively easy of the shelf implementation.

VI. Conclusion

Three models, InDelphi, FORECasT and Lindel were compared on CHO and RPE1 cells. The results were cell dependent, in which CHO cells produced the lowest KL-diverence. The proportion of frameshift mutations was analogous to what had previously been reported in literature. The mutation frequency was generally captured equally well between the models, with the exception of 1-bp insertions. This genotype was underestimated in both cell types. To explain this, we confirmed that a majority of the KL-divergence originated from insertions of length 1. From this knowledge, we further proposed a way to potentially improve InDelphi using another non-parametric machine learning model for modelling insertions.

VII. Code Availability

The analysis was coded in python3 using the respective github repositories for InDelphi, FORECasT and Lindel. The code for the comparative analysis can be found here.

References

[1] F. Zhang, Y. Wen, and X. Guo. Crispr/cas9 for genome editing: progress, implications and challenges. *Human Molecular Genetics*, 23(1), 2014.

[2] R. Ceccaldi, B. Rondinelli, and A.D. D'Andrea. Repair pathway choices and consequences at the double-strand break. *Trends Cell Biol*, 26(1), 2017.

[3] J. Her and S. Bunting. How cells ensure correct repair of dna double-strand breaks. *Journal of Biological Chemistry(JBC)*, (1):10502–10511, 2018.

[4] L. Crepaldi F. Allen, C. Alsinet A.J. Strong, V. Kleshchevnikov, and P. Angeli1. Predicting the mutations generated by repair of cas9-induced double-strand breaks. *Nature Biotechnology*, 37(1):64–75, 2019.

[5] W. Chen, A. McKenna, J. Schreiber, Y. Yin, V. Agarwal, W.S. Noble, and J. Shendure. Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *bioRxiv*, pages 1–30, 2018.

[6] M.W. Shen, M. Arbab, J.Y. Hsu, D. Worstell, S.J. culbertson, O. Krabbe, C.A. cassa, D.R. liu, D.K. Gifford, and R.I. Sherwood. Predicting the mutations generated by repair of cas9-induced double-strand breaks. *Nature*, 563(1):646–651, 2018.

[7] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, (1):2579–2605, 2008.

[8] F.A. Ran, P.D. Hsu, J. Wright, V. Agarwala, D.A. Scott, and F. Zhang. Genome engineering using the crispr-cas9 system. *Nature Protocols volume*, 8(1):2281–2308, 2018.