NANYANG TECHNOLOGICAL UNIVERSITY

SCHOOL OF BIOLOGICAL SCIENCES

BS6203 Story-telling with Graphics and Visualisations

Final Project: Graphics and Visualization as aid for Feature Engineering

Denise Ng Li Yi

Muhammad Irfan Bin Hajis

10 October 2023

Introduction

Feature engineering is indispensable to create a high-quality machine learning model in the context of biology. Feature engineering involves selection, transformation, and curation of relevant features from raw data to improved performance of models. Graphics and visualisation are critically important to aid feature engineering. Visualisation techniques, such as scatter plots, histograms, and box plots, allow researchers to understand the distribution of the data, recognize patterns or correlations, and identify outliers. Whereas, correlation matrices, scatter plots, and heatmaps reveal which features are highly correlated or have a strong influence on the target variable. Visualisation is a powerful tool to communicate findings and insights with stakeholders or readers who may not be familiar with the technical details. Consequently, with the aid of visualisations, researchers can incorporate domain knowledge into the analysis by selecting relevant biological features and discarding irrelevant ones. Additionally, feature engineering might reveal unexpected patterns or meaningful relationships in the data, leading to new biological insights or discoveries.

In the context of Type VI CRISPR (clustered regularly interspaced short palindromic repeats) enzymes, Wessel et al. observed that CRISPR Cas13d knockdown efficacy is driven by target site context and gRNA-specific features. Single mismatches were shown to generally reduce knockdown to a modest degree, but spacer nucleotides 15–21 are largely intolerant of target site mismatches. Wessel had to narrow down from 100s of thousands to selecting 35 features for their Random Forest Model. (Wessels et al., 2020)

Type VI CRISPR enzymes are programmable RNA-guided, RNA-targeting Cas proteins with nuclease activity. CRISPR Cas13 allows for target gene knockdown without changing the genome. Cas13 proteins are directed to their target RNAs by a single CRISPR RNA (crRNA). A single crRNA consists of direct repeat (DR) stem loop and a spacer sequence that mediates target recognition by RNA-RNA hybridization. Cas13 enzymes are known to exert some nonspecific collateral nuclease activity on activation. Cas13 enzymes also have drastically reduced off-target activity in cultured cells compared with RNA interference (RNAi). (Wessels et al., 2020)

In this report, we use graphics and visualisation to storytell a feature engineering journey from data exploration, data cleaning and choosing relevant features to generalise a model to predict for effective CRISPR Cas13a knockdown on any RNA target sequence. CRISPR Cas13a is well-suited to be used as a molecular diagnostics tool. (Kellner et al., 2019) We utilised Cas13a human transcript knockdown expressions from Abudayyeh et al and sought to sieve relevant features to develop a model that predicts for CRISPR Cas13a knockdown efficacy. (Abudayyeh et al., 2017) We model our approach based on Wessels' methods and visualise key relevant features to highlight biologically relevant features. We also highlighted how visualisations may lead to different method selections (eg. Pearson vs Spearman Correlations) and decisions (eg. feature1 vs feature2).

Methods

Data Collection. We collected a total of 555 gRNAs across 4 knockdown experiments using engineered variants of LwaCas13a from Abudayyeh et al. (178 from Gaussia luciferase, 93 from Cypridinia luciferase, 93 from PPIB 93 from KRAS-1 and 93 from MALAT1). Due to the experimental noise native to the data source methodology, a significant amount of the experimental replicates for a specific guide differed significantly in transcript expression, often by more than 25% (as observed in Figure2). We performed Grubbs' Test to detect any outlier, with a significance level of $\alpha = 0.05$. To select and remove outliers, an outlier is defined when a triplicate normalized expression (NE) score has a Coefficient of Variance > 0.1 and has a Z-score > 1 for the particular replicate. If two replicates have a z-score > 1, we select replicates with the smaller difference for downstream analysis.

We used the mean for each triplicate. We then normalized n=555 available gRNAs across the 5 genes to the same scale before training the model as per the methods in Wessel. To do so, for each dataset D, we computed the upper and lower quartiles of the guide -log2(NE) (UQD and LQD, respectively), as well as the corresponding quartiles for the -log2(NE) among all the datasets pooled together (UQP and LQP). We then updated each fold change, x, as follows: $\hat{x} = ((x - \text{LQD}) / (\text{UQD} - \text{LQD}) \times (\text{UQP} - \text{LQP}) + \text{LQP})$. By centering on quartiles, this procedure normalized the normalized-expression distributions in a way that was less susceptible to the influence of outliers of a single screen. (Wessels et al., 2020)

Predicting RNA secondary structures and RNA–RNA hybridization energies. The crRNA secondary structure and MFEs were derived using RNAfold [–-gquad] on the full-length gRNA (DR + guide) sequence. Direct Repeat (DR) Sequence, GATTTAGACTACCCCAAAAACGAAGGGGACTAAAAC, is used in our downstream analysis as indicated in Abudayyeh's knockdown expression experiments.

Target RNA unpaired probability (accessibility) was calculated using RNAplfold [-L 40 -W 80 -u 50]. We performed a grid-search calculating the RNA accessibility for each target nucleotide in a window of −25 bases downstream of the target site to +25 bases upstream of the target site, assessing the unpaired probability of each nucleotide over 1–50 bases for all perfect match (PM) guides. Then, we calculated Pearson’s correlation coefficient between the unpaired probabilities and the observed gRNA knockdown expression for each point and window relative to the gRNA.

RNA–RNA hybridization between the gRNA and its target site was calculated using RNAhybrid [-s -c]. We calculated the RNA-hybridization MFE for each gRNA nucleotide position p over the distance d to the position p + d with its cognate target sequence. All measures were either directly correlated with the observed gRNA knockdown expression or used partial correlation to account for the crRNA-folding MFE.

Assessing target RNA context. To assess the target RNA context, we calculated the nucleotide probability at each position (p) over a window (w) of 1–50 nt centred around the position of interest (for example, p = −18 with w = 11 summarises the nucleotide probability in a window from −23 to −13, with +1 being the first base of the gRNA). We evaluated p for all positions within 50 nt upstream and downstream of the gRNA. The nucleotide probability of each point was then correlated with the observed gRNA knockdown expression for all gRNAs, either directly or using partial correlation accounting for crRNA-folding MFE. In each case, we used both Pearson’s correlation and Spearman's Correlation Coefficient rank.

Plotting and Data Availability. The analysis was performed using Python3. We used Scipy v1.7.3 to perform correlation analysis. Plotting was performed using R 4.2.2 ggplot2 package. The source code, raw data and results are available at https://github.com/micro-irfan/Cas13AgRNAtor under the GPLv3 licence.

Results and discussion

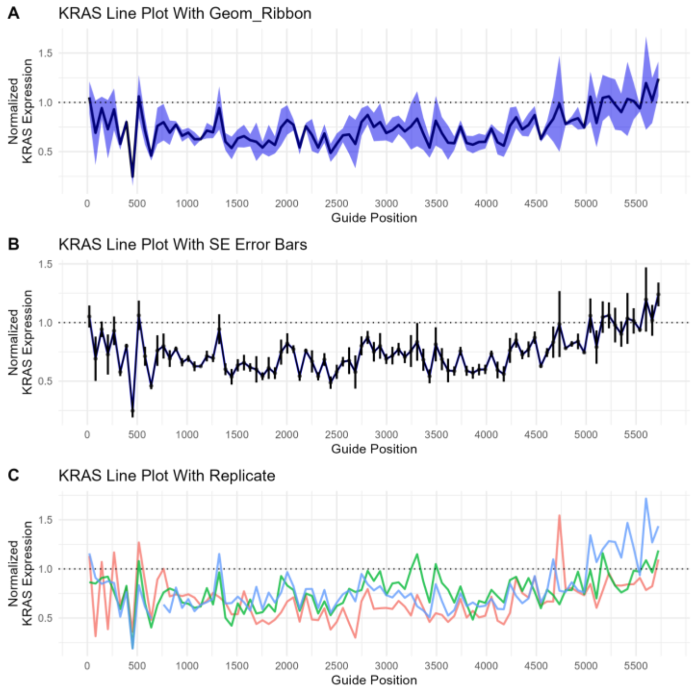
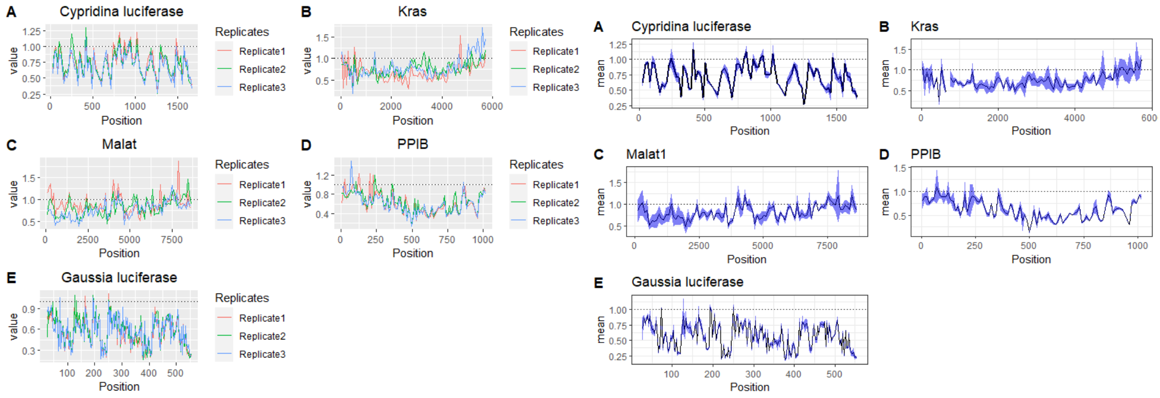


Figure 1. Line graphs depicting the distribution of gRNAs targets along the transcripts and normalized KRAS expression using 3 different methods, (a) ggplot2 geom ribbon (b) Standard error bar and (c) all replicates (n=93)



(2a)

(2b)

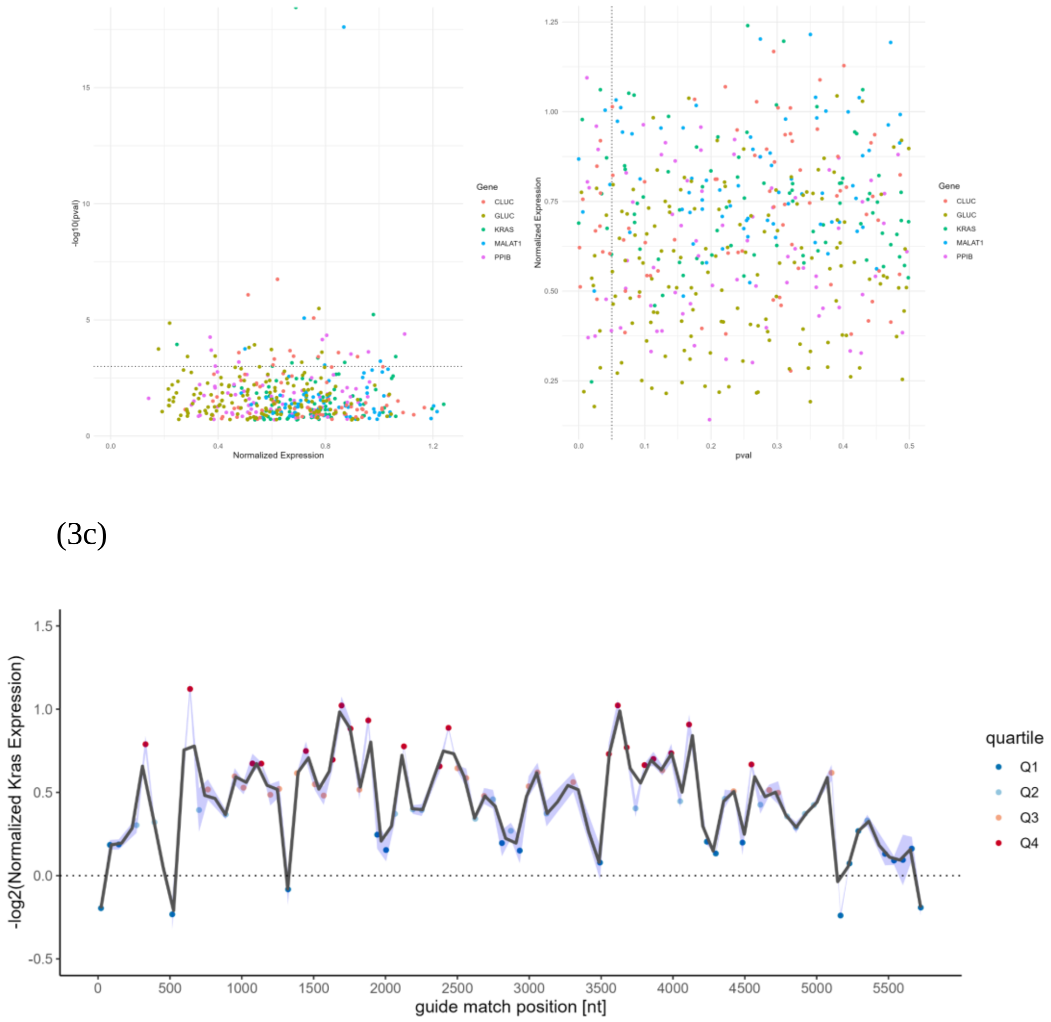
Figure 2. (a) Line graph depicting the distribution of gRNAs targets along the transcripts and their normalized expression, each line representing each replicate. Plot A-E shows gRNA targeting different transcripts. (b) Line graph depicting the distribution of gRNAs targets along the transcripts and their normalized expression. Plot A-E shows gRNA targets different transcripts.

Raw Data Analysis. The normalized expressions were plotted against the distribution of the gRNAs along the transcript. A lower normalized expression value indicates a better performing gRNA at a particular position along the transcript. From the plot, better performing gRNA seem to appear sporadically along the 5 transcripts. For example, gRNA that targets along position 4000 – 6000 are worst as observed from the general rise in normalized expression. Additionally, we can add that the standard deviation (SD) along the region is higher. While gRNAs that target PPIB within position 500 – 750 have lower normalized expressions instead.

The plotting of these data can be rather different depending on what information is preferred to be shown. Figure 1a shows the mean normalized expression of KRAS (y-axis) is plotted against the position of the transcript (x-axis) as a black-coloured line with the SD ranges of minimum and maximum values of the 3 replicates as a blue ribbon overlaid on the line. This makes the plots much cleaner and clearly shows the variance of the normalized expressions across all 3 replicates without messy lines. The general trend of gRNAs along KRAS is thus easily deduced with a quick glance. Similar to Figure 1a, Figure1b shows the variance in another form, using Standard Error (SE) Bar.

On the other hand, from Figure 1c, plots the guide scores of every replicate against the position, with each replicate represented by a different coloured line. This appears messy as the lines start to intercept each other and can be confusing to understand. However, the plotting of individual replicates meant that outliers or exceptions can be identified quickly. We observed that KRAS replicates 3 seems to be diverging away from replicate 1 and 2 between position 5000 and 6000 while replicate 1 appears sporadic in the first 500 positions. We can deduce if a particular replicate has gone wrong if almost all points are not clustered together with other replicates visually. All 5 experiments are plotted in Figure 2.

(3a) (3b)



(3c)

Figure 3. Grubbs' Test (a) -log10(pval) against the mean Normalized Expression (b) the mean Normalized Expression against pval (n = 555) (c) Line graph depicting the distribution of gRNAs targets along the transcripts and -log2 transformed normalised KRAS expression (n = 93). The gRNAs are separated into targeting efficacy quartiles Q1–Q4, with Q4 containing guides with the best knockdown efficacy.

Outlier Detection. Based on Figure 1 & 2, however, we cannot define if a replicate at a particular position really contains an outlier. We conducted a Grubbs' test to identify potential outliers within each triplicate. We established the null hypothesis (H0) as the absence of outliers within the triplicates, employing a significance level of α = 0.05. The Grubbs' test results indicate that the knockdown experiments in each transcript exhibit a range of 5 to 11 outliers. From Figure 3a and 3b, we learnt that data transformation can lead to difference in the plots' interpretation. From Figure 3a, we can see that the majority of the experiments have a pval > 0.05. This gives the visual indication that most of the experiments and its replications are fine. From Figure 3b, we observed that there is no bias for outliers towards any genes as points are not clustered based on any of the genes. Additionally, we noticed that some genes generally have a lower normalised knockdown expression (GLUC). Outliers are subsequently removed for the downstream analysis (explained in methods).

After adjusting for outliers, from Figure 3c, we observed that variance has reduced drastically. We also -log2 transformed the normalised expressions to make the plots more intuitive (ie. effective gRNAs will have better scores). We added a colour scheme to indicate which quartile each point belongs too. Visually, we can see that Q4 quartiles (effective gRNAs) cluster around position 3800 of the transcript.

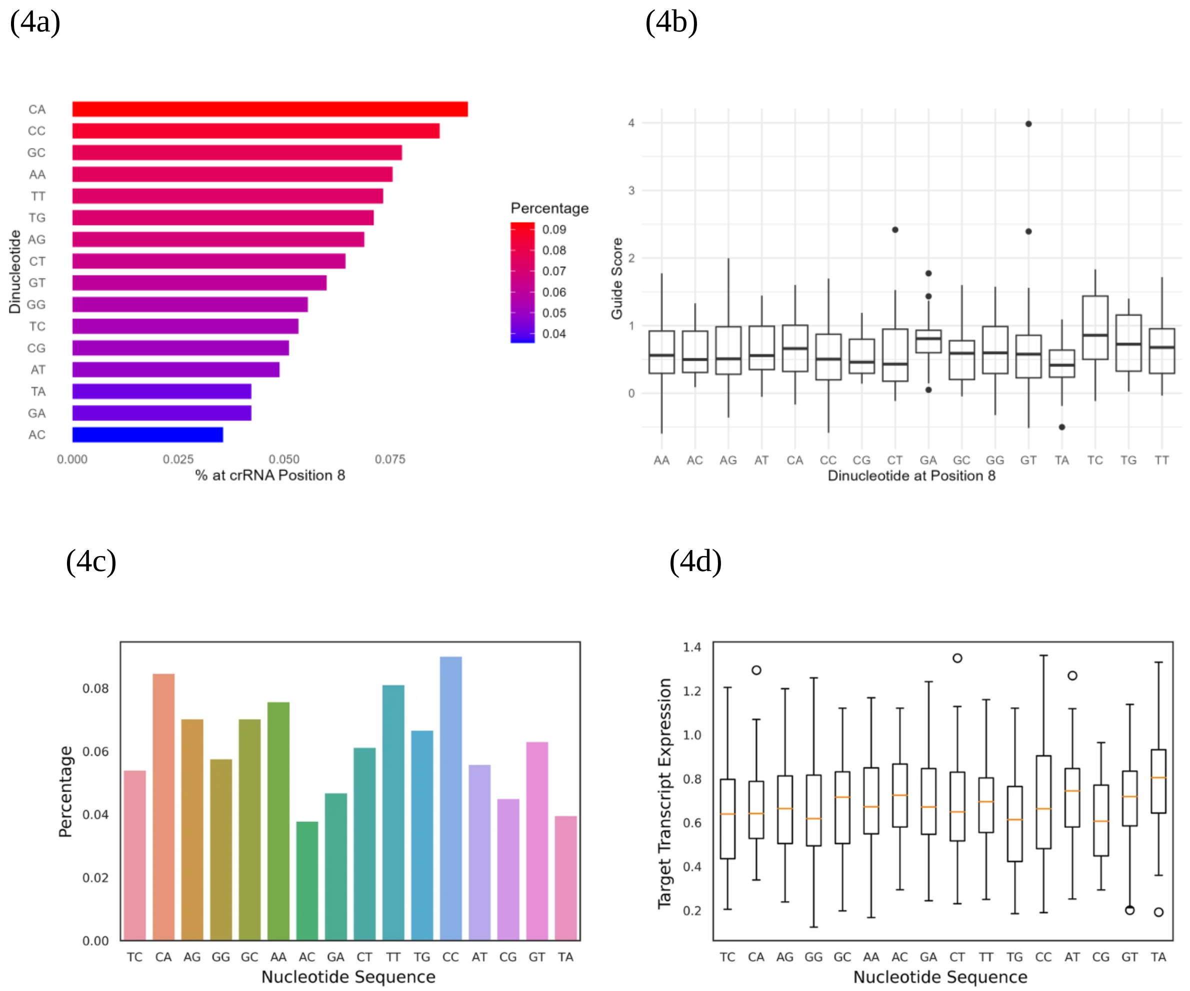


Figure 4. Bar plot of the population of gRNAs that contain a specific dinucleotide at position 8 for (a) ours and (c) Krohannon. Box plot of target transcript expression values as a function of the nucleotide at position 8 for (b) ours and (d) Krohannons'

RNA Nucleotide Composition Analysis. Maximising Information, Minimising Confusion. Krohannon et al. adopted a slightly different method. Krohannon obtained a list of over and under-represented k-mers (chi-squared test) at each location across the gRNA. gRNAs were partitioned into distinct groups based upon their nucleotide composition at a specific location; in order to perform a Kruskal Wallis test. (Krohannon et al., 2022) We compared the bar and box plots plotted in Krohannon and ours in figure 4. We noticed that Krohannon order of dinucleotide is random (Fig 4c, d) whilst ours is based on the highest percentage to the lowest (Fig 4a) and ordered lexicographically (Fig 4b). Our choice of order makes it easy for readers to match the dinucleotide. Colour scheme on Krohannon is rendered useless as it is uninformative whilst ours followed a colour scheme based on the percentage of dinucleotide observed at position 8 of the gRNA. However, these plots are quite uninformative in the context of feature engineering since plots in Figure 4 only inform the dinucleotide percentage / gRNA performance at a single position (bp 8). In order to find relevant features from thousands of features, we would need to plot in various ways that maximises information while minimising confusion.

In our report, we plotted ScatterPlot, Lollipop Plot, Barplot and Heatmaps to identify potential features. In Figure5a, we plotted for Cas13a gRNA nucleotide preferences that may influence gRNA efficacies for all experiments. We measured the effect-size (Δ nucleotide probabilities), and P values of observing the conditional probability of a guide in the top 20% under the null distribution examined at every position including the 3 nucleotides 5’ and 3’ of the gRNA target site. The P values were calculated from the binomial distribution with a baseline probability estimated from the full-length mRNA target sequence with all perfect match gRNAs (n = 451). We adjusted the P values using Bonferroni-correction. However, we did not observe any nucleotides at any position of the target that had any significant difference.

From Figures 5a and 5b, we can start to curate multiple potential features. We can infer which nucleotide/s is/are favoured on the target sequence across 34 positions (or more) and 4 nucleotides. For example, C is disfavoured at position 14 of the gRNA on the target sequence, U is favoured at position 25 of the gRNA on the target sequence. These favoured / disfavoured nucleotides at *i*th position could be used as features to build models to predict for effective Cas13a gRNAs. From these two plots alone, we’ve considered and plotted for 136 (34 x 4) potential features.

(5a)

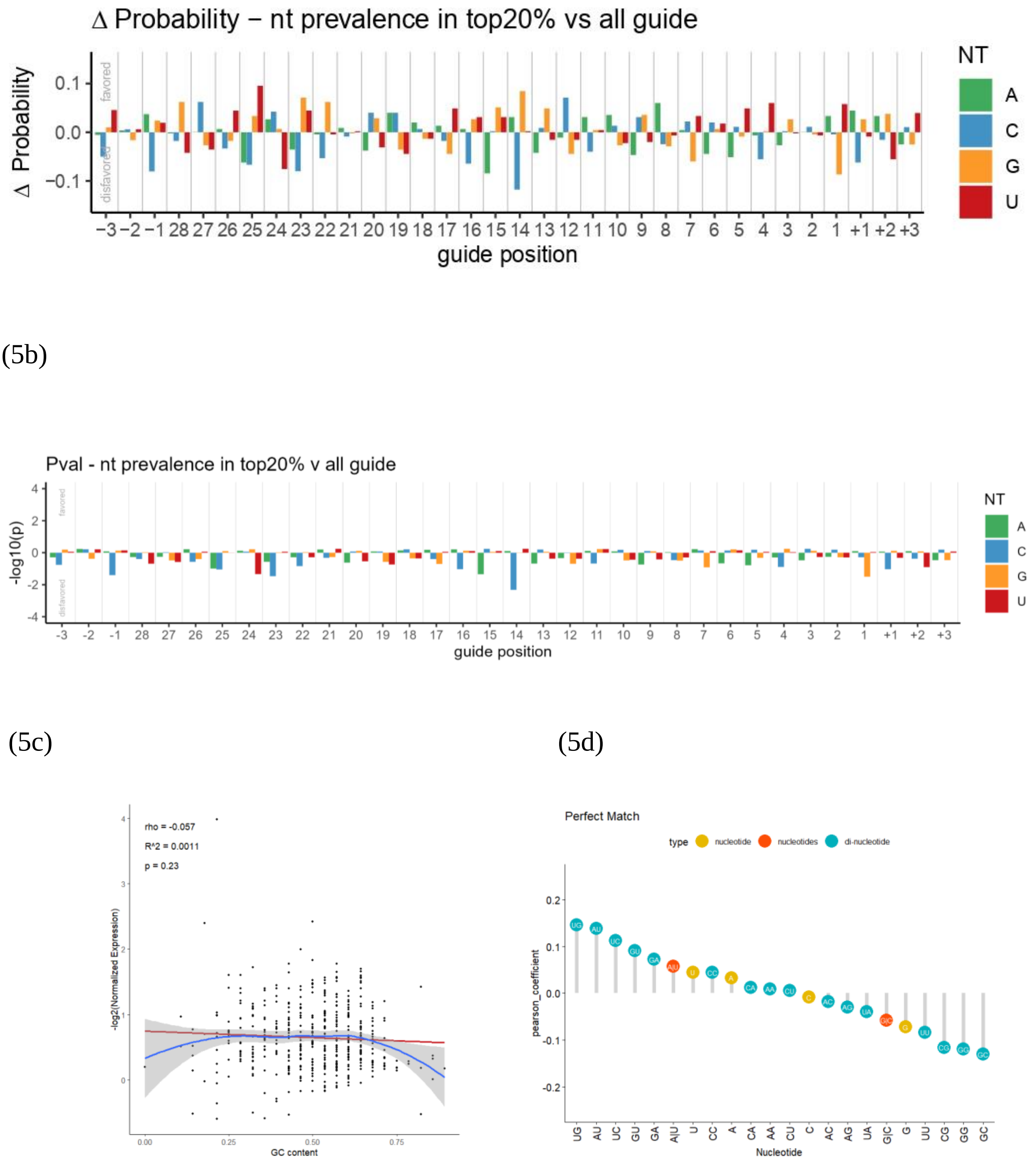


Figure 5. . (a) (top) Effect size (Δ nucleotide probabilities), (b) P values of observing the conditional probability of a guide in the top 20% under the null distribution examined at every position including the 3 nucleotides 5’ and 3’ of the gRNA target site. (c) Scatterplot depicting the gRNA -log2(NE) and gRNA GC-content as a fraction (d) As in c, but showing the Pearson’s correlation coefficient (rp) between gRNA -log2(NE) and all guide single nucleotides, di-nucleotides, and G|C and A|U-content.

From Figure 5c, we plot for Pearson's correlation coefficient (rp) between gRNA -log2(NE) and all guide RNA single nucleotides, di-nucleotides, and G|C and A|U-content. In Figure 5d, we plot a Scatterplot depicting the gRNA -log2(NE) and gRNA GC-content as a fraction (n = 452). The red line indicates the linear relationship between both values (Pearson's correlation coefficient rp, p = .23). The blue line indicates a LOESS fit. (Grey shading denotes LOESS fit confidence interval). From the lollipop plot, although not many of the nucleotides and di-nucleotides have a high correlation, we identified that di-nucleotide UG is most favoured in effective gRNAs whilst di-nucleotide GC is the most disfavoured. Di-nucleotides GG, CG and GC are the 3 most disfavoured, indicating a dislike towards nucleotide G (which also has a negative correlation). This is consistent with Wessels Cas13d results where the gRNA has a general dislike towards nucleotide G. (Wessel et al., 2020) From Figure 5d, we hope to observe a trend. For example, a relatively lower / higher GC% content may indicate better guide scores. However, due to the low correlation (rho), there are no trends observed and we cannot find any local optimum for GC content as well. From the LOESS fit graph, we noticed that gRNAs with GC% content at the extreme ends have lower guide scores. A LOESS fit graph is better than the Pearson Correlation to fit for non-linear trends.

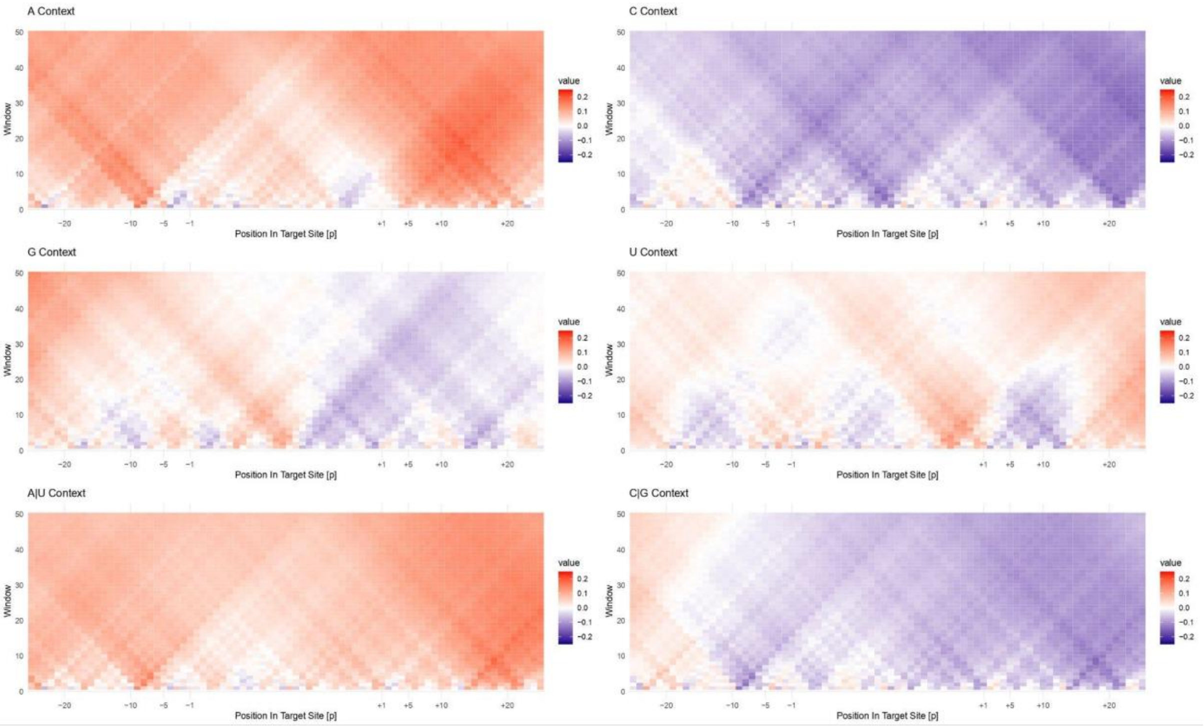
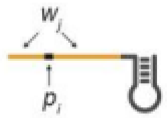
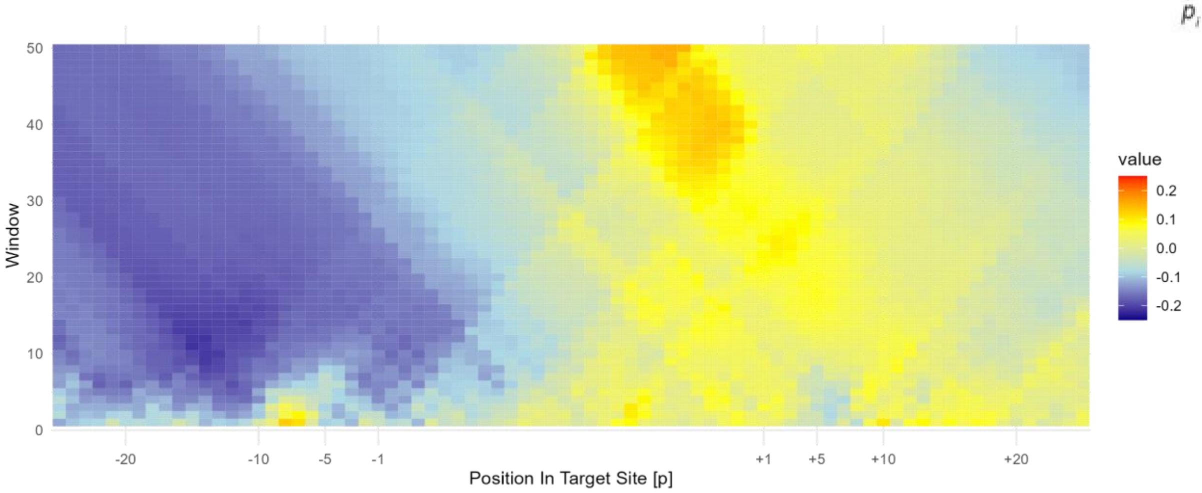


Figure 6. Heatmaps depicting the Pearson's correlation coefficient (rp) between the local target nucleotide-contexts (A, C, G, U, A|U and G|C) and observed -log2(NE) relative to gRNA match positions.

Beyond gRNA nucleotide composition, we wondered if the context features of the gRNA target site affected target knock-down. Figure 6 are heat maps depicting the Pearson’s correlation coefficient (rp) between the local target nucleotide-contexts (A, C, G, U, A|U and G|C) and observed -log2(Normalized Expression) relative to gRNA match positions. We performed a grid-search correlating the observed gRNA efficacies with the summarised target nucleotide density across a window of 1 nt up to 50 nt at every point 25 nt 5’ of the target site to 25 nt 3’ of the target site (n = 399 for each cell). From the heatmaps above, we observed that there’s a preference for A nucleotide and a dislike for C nucleotide across the target. Negative Correlation for C nucleotide on the target corresponds to the results above from the lollipop plot and bar plot (Figure 5a) where there’s a negative correlation for G nucleotide on the gRNA sequence and C nucleotide on the target sequence at position 14 of the gRNA. Heatmaps are visually helpful to bring the reader's attention to the important features. For example, heat map for U-content seems quite uninformative as the Pearsonr Correlation is hovering between -0.1 and 0.1 and readers may choose to skim over it. The heat map for A-content highlights potential features from the stronger and darker red in the upstream (5’) of the gRNA target region (around position +14 and window size of 20) with a relative correlation of 0.2. From the heatmap above, we’re able to identify potential features from 23400 (78 Positions * 50 Window Length * 6 Nucleotide Combination) features. We can increase the number of features by counting the occurrences of di-nucleotide (8), tri-nucleotide (64) for each position in the target sequence for each window size.



(7a)



(7b)

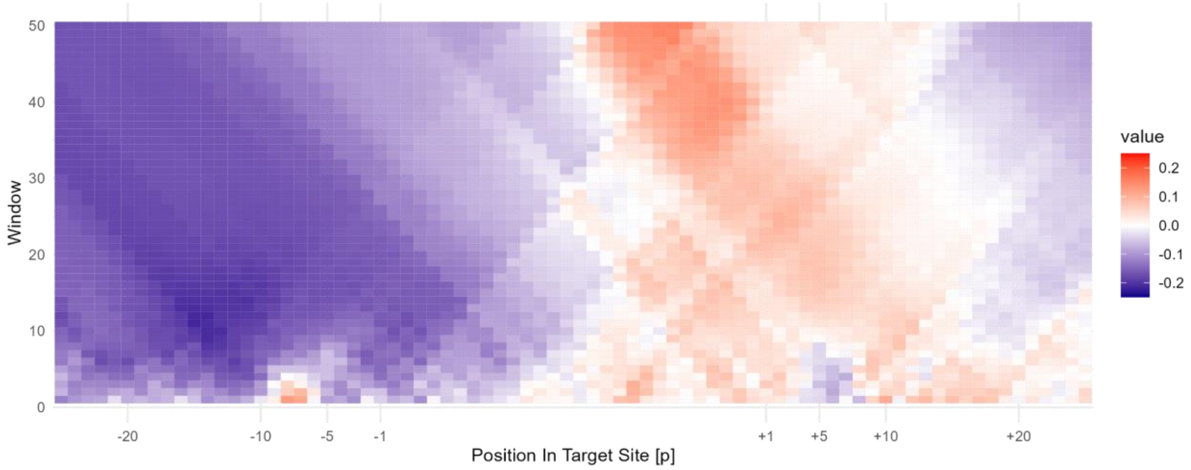


Figure 7. Heatmap depicting the Pearson's correlation coefficient (rp) between the local target site accessibility (unpaired probability) and the observed -log2(NE) relative to gRNA match positions with different colour schemes (a) Red-yellow-lightblue-Darkblue and (b) Red-White-Darkblue

Target Site Accessibility. Creating Gradients. We also assessed whether the target site accessibility influences knock-down by correlating the observed gRNA efficacies with the target site accessibility. We define target site accessibility as the probability that the target RNA is unpaired. Figure 7 are heat maps depicting Pearson's correlation coefficient (rp) between the local target site accessibility (unpaired probability) and the observed -log2(NE) relative to gRNA match positions. We performed a grid-search correlating the observed gRNA efficacies with the unpaired probability in a window (w) of 1 nt up to 50 nt at every point 25 nt 5' of the target site to 25 nt 3' of the target site. We found a weak positive correlation with increased target site accessibility is centred on the 5' end of the spacer RNA. This is opposite of the target-RNA accessibility preferences in Cas13b and Cas13d where the increased target site accessibility is centred on the 3'-end of the spacer RNA. In the heatmap, darker colours indicate stronger correlation, red for positive correlation and blue for negative correlation. We observe the gradual change from Positive Correlation to Negative Correlation towards the 3' end of the target sequence. Due to the different colour scheme used, we are able to identify the boundaries between positive and negative correlation better in 7b. This may have an impact if the reader is trying to define boundaries between positively correlated gRNAs and negatively correlated gRNAs to the unpaired probability of the position and window.

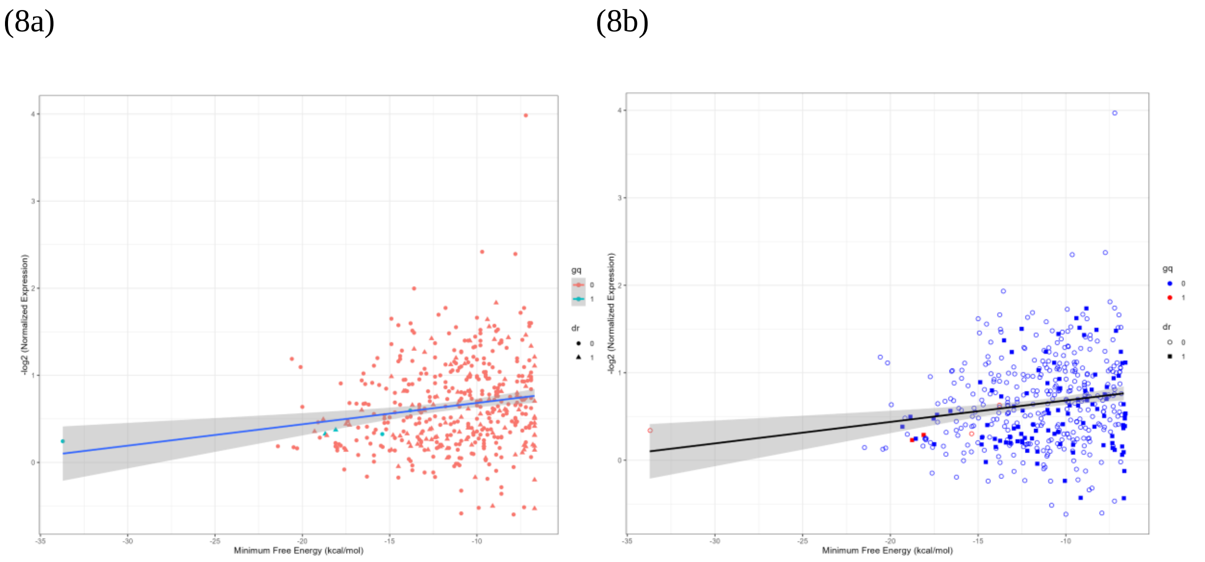


Figure 8. (a) Scatterplot of the guide scores plotted against the minimum free energy (MFE) of the gRNAs. Gq is the G-quadruplex structure of the gRNA and dr represents direct repeat represented in binary where 1 indicates presence of the expected Direct Repeat (DR) folding. G-quadruplex and presence of DR folding are differentiated by colour for gq and shape for DR. The blue line indicates the correlation between guide score and MFE. (b) An alternate representation of scatterplot guide score vs minimum free energy (MFE). Gq is the G-quadruplex structure of the gRNA and dr represents direct repeat. 0 or 1 is the binary indication of variables and is differentiated by colour for gq and shape for dr. The black line indicates the correlation between guide score and MFE.

Predicting RNA secondary structures using RNAFold. Overplotting and distinguishing variables. The scatterplot of the guide score as Y axis plotted against minimum free energy as the X-axis was done in order to understand the relationship between the efficacy of the gRNA and thermodynamic stability of the secondary structure of the gRNA which is denoted as the minimum free energy and is influenced by variables such as specific conformation such as G-quadruplex structure, and the interaction between the direct repeats sequence and the gRNA in the crRNA sequence. A lower MFE indicates better stability of the gRNA. The linear line confirms this as the -log2 normalized expression generally increases with increasing MFE values. The presence of G-quadruplex structure may potentially inhibit gRNA binding to the target sequence. However, we found no strong correlation as only 4 gRNA were found to exhibit G-quadruplex structure. The Direct Repeat sequence confers structural stability and recognition by CRISPR Cas13a enzyme. However, similar to the G-quadruplex structure, we found no strong correlation between the presence of DR folding and guide score.

The scatterplot seen in Figure 8a is an example of overplotting due to the representation of multiple variables (0 or 1 for gq and 0 or 1 for dr). The data points for these binary indicators started to blend together due to the selected shapes (solid circles and triangles for dr) and colours (red and blue for gq). Such that data points with differing dr status become indistinguishable as they both share the same variable of being negative gq as is the majority of the data points. Figure 8b is an example of how elements from Gestalt's principles can be used to plot a graph with better readability. The element similarity and grouping were used to differentiate the dissimilarity between dr binary indicators by having a shaded square for 1 and a non-shaded circle for 0. Thus, despite having the same colour (blue for 0 gq), 0 and 1 dr data points are well differentiated. Similarity and salience is also seen in the use of strong contrasting colours for gq which allows the easy identification of the binary indicators. A darker blue for the majority of data (gq-0) and a bright contrasting red for uncommon data points (gq-1). The use of black instead of default blue for correlation helps to prevent the obscuring of data points near the line. Jittering was also applied in Figure 4 to avoid the overcrowding of data points and adjusted to avoid a misrepresentation of the data.

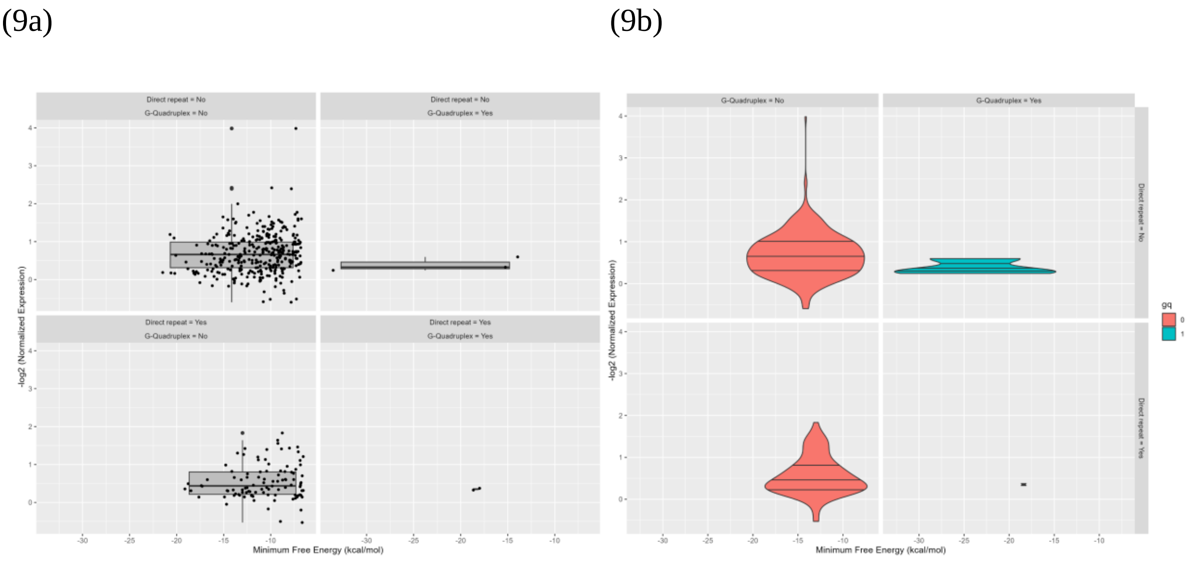


Figure 9 (a) Boxplot of the guide scores vs MFE. Each faucet is shown with a different combination of the 2 variables, direct repeat and G-quadruplex structure. (b) Violin plot of the guide scores vs MFE. Each faucet is shown with a different combination of the 2 variables, direct repeat and G-quadruplex structure.

The distribution of the data seen in the scatterplot (fig 8 and 9) was further analysed by plotting boxplots and violin plots. The boxplot visualises the summary statistics such as the median, shown as a line within the box, and interquartile ranges, shown as the upper and lower end of the boxplots, which can indicate potential outliers. The individual points in the box plot shown provide a visualisation of the distribution of the data for each combination of variables but crowding of the data points may pose a distraction and obscures part of the boxplot. The violin plots on the other hand enhance the visualisation of distribution by including a kernel density plot in addition to the summary statistics which is seen as the curve of the violin plot, symmetrical on either side. The summary statistics such as median, upper and lower interquartile ranges are seen as lines within the violin. The use of differentiating colours for one of the variables accompanying the labels helps to distinguish the plots. A violin plot may be preferred if the inclusion of data distribution is informative in addition to summary statistics.

(10a) (10b)

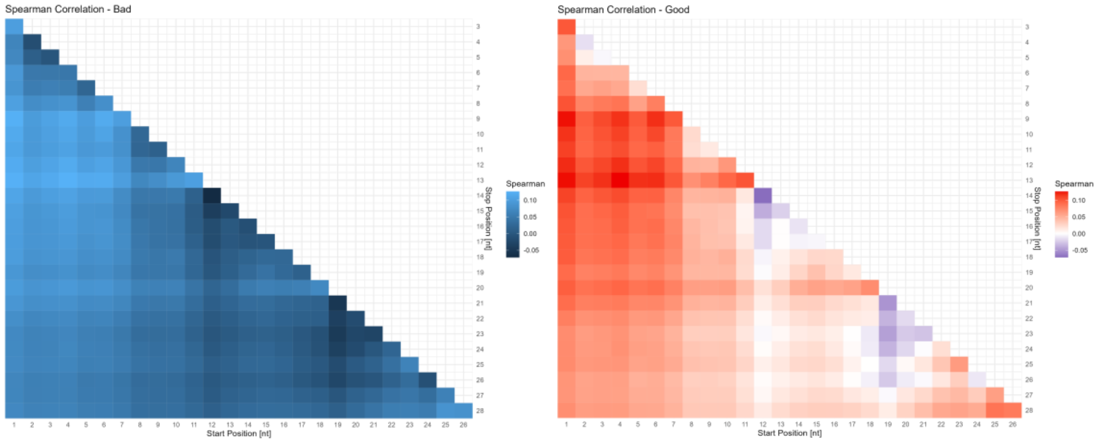


Figure 10. (a) Spearman correlation heat map of guide score and hybridization minimum free energy of gRNA nucleotide across the gRNA nucleotide position. (b) Spearman correlation heat map of guide score and hybridization minimum free energy of gRNA nucleotide across the gRNA nucleotide position.

Target Site Accessibility using RNAHybrid. Colour choices. Spearman correlation analysis was done between the quartile and the MFE from the hybridization of the gRNA nucleotide to the target site sequence for each window slice between 2 positions (start and stop). This was calculated across the positions of the gRNA nucleotides. An increase or hotspots of high correlation at specific regions indicates that the efficacy of binding between specific regions of the gDNA and target regions may be highly correlated to guide scores or efficacy of the gRNA. We notice that there is a slight correlation (~0.12) between all start positions to position 13 of the gRNA. The window length and gRNA positions correspond to the results generated from

RNAplfold in Figure 7 where gRNA position 1-15 exhibit higher unpaired probabilities for a window size of 45-50. CRISPR Cas13a gRNAs possibly bind from the 5’ end sequence. Our findings are roughly inline with Metsky results’ where weaker guide–target pairs are relatively likely to contain mismatches in positions 6–11 of the spacer, concordant with the known region, and there is a higher tolerance for mismatches on the 3′ end of the spacer. (Metsky et al., 2022)

Figure 10a is a poor attempt to visualise the correlation as a heat map due to the default colour scale which utilises a range of hues. As the correlation values are rather close together and colour hues have little contrast, it is rather difficult to recognise areas of high or low correlations, which is of interest. In order to improve this, a heatmap with strong contrasting colours for high, mid and low correlation scores has been used instead. The resulting plot is seen in Figure 10b where high correlation scores are denoted as red, mid scores as white and low scores as blue, these colours are still arranged in a colour bar with gradual changes in hue. Both axes on the plot have also been brought closer to the heatmap and axis ticks were adjusted to emphasize on the nucleotide positions. With these changes, the heatmap was able to distinctly show hot spots of high and low correlation scores in relation to the nucleotide positions.

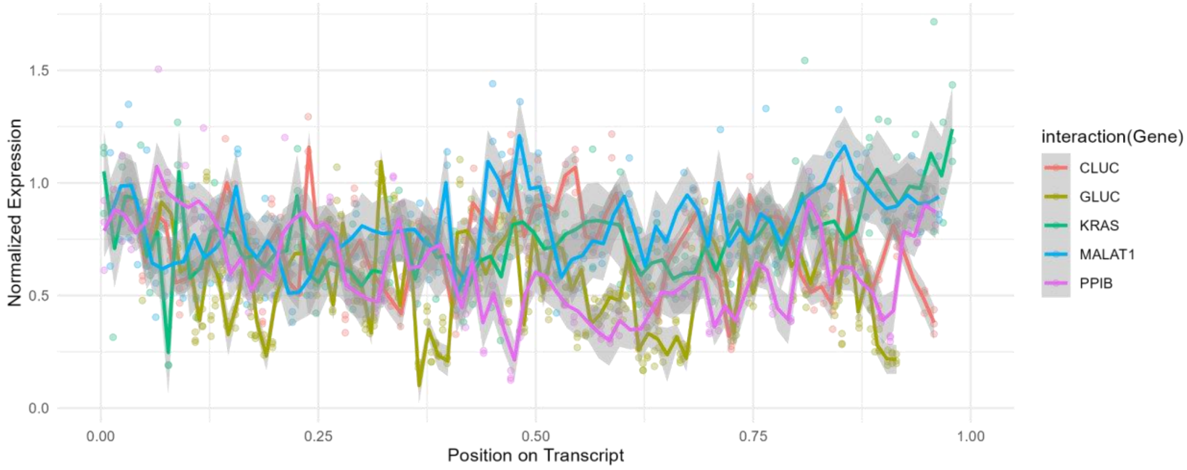


Figure 11. Distribution of gRNAs along the relative transcript position and their normalised expression (n = 555). The mean behaviour for each gene is highlighted using a LOESS fit.

Ad-hoc Plot: Relative Transcript Position. Deriving Information from your plots. Krohannon added relative guide target position as a feature for their machine learning model based on a belief that the ends of transcripts, both 5’ and 3’ ends are highly structured, both to protect the transcript from degradation and to facilitate movement to different cellular compartments. (Krohannon et al., 2022) In order to validate this, we plot a line graph, for each gene, indicating LOESS fit, based on the raw data on the same plot. From Figure 11, we can generally observe that the 5' and 3' ends have higher normalized expression indicating ineffective gRNAs. However, for the middle, the normalized expression appears sporadic with inconsistent patterns across different genes. For example, at 0.45 relative position on transcript, we see that PPIB and GLUC have lower normalized expression whilst CLUC and Malat1 have higher normalized expression.

We can also observe that for each gene, the mean normalized expression *appears* to follow a normal distribution. We conducted an assessment of the normality of mean normalized gene expression data for each gene in our study. Utilizing the Shapiro-Wilk Test with a predetermined significance level (α) of 0.01, we evaluated the normality of the data distributions. Our analysis revealed that all genes under investigation exhibited p-values greater than 0.01, thus failing to reject the null hypothesis, suggesting that the respective normalized expression followed a normal distribution. Consequently, we proceeded to normalize the data based on the mean values in each gene (gene_mean - triplicate_mean). Subsequent analysis of the Deep Learning Model trained in BS6204, using the mean-normalized scores instead of using scores from centering on quartiles (explained in methods), indicated slight improvement in learning performance (0.5 to 0.6 in terms of accuracy). It is worth noting that we did not use the mean-normalized scores due to time constraint. These factors and methods were only identified toward the conclusion of the allocated time frame.

Conclusion

The aid of graphics and visualization helps in feature selection for effective gRNAs significantly. The features can then be studied to understand its influence on the efficacy of gRNA and thus influence the efficacy of the CRISPR Cas13a system as a whole. The data generated is highly complex and requires good data visualisation, from employing various types of plot (scatterplot, heat maps, and even lollipop plots) to choosing colour schemes and shapes for labels. Visualization allows us to better understand the analysis of our data, efficiently present the findings and choose proper subsequent methods. From our graphics and visualization, we were able to pinpoint multiple features, albeit having a low correlation (between 0.1 - 0.2), that can potentially be used as features to train machine learning models for predicting effective CRISPR Cas13a gRNAs. Low correlations in biology are quite common due to the diverse noise that are unaccounted for. For example, global RNA structures in genes or RNA viruses, or a binded protein that may potentially inhibit any sort of binding to CRISPR Cas Module. Wessels CRISPR Cas13d gRNA Random Forest Model has previously shown that the strongest correlation was only a mere 0.5. (Wessels et al., 2020) In our report, we have also shown that, many a time, we tend to miss trends if we did not plot in a particular manner which highlights the importance for graphics and visualisation.

Poor visualisation can also lead to confusion and misinterpretation of the analysis. This can stem from overplotting especially when we can derive multiple features leading to many dimensions and/or multiple variables. The plots in this report have primarily shown that plots with data represented with poor colour contrasts or similar shapes are prone to overplotting. The selection of the type of plots for data analysis should also be carefully considered to avoid the presentation of irrelevant data which can confuse readers or cause misinterpretation of the plots. Our plots are not without limitations and can be improved with better colours or shapes to represent the data better. The highlights of important data points and separation of data can also help in drawing attention to critical conclusions.

References

Abudayyeh, O., Gootenberg, J., Essletzbichler, P. *et al.* RNA targeting with CRISPR–Cas13. *Nature* **550**, 280–284 (2017). https://doi.org/10.1038/nature24049

Kellner, M.J., Koob, J.G., Gootenberg, J.S. *et al.* SHERLOCK: nucleic acid detection with CRISPR nucleases. *Nat Protoc* **14**, 2986–3012 (2019). https://doi.org/10.1038/s41596-019-0210-2

Krohannon, A., Srivastava, M., Rauch, S. *et al.* CASowary: CRISPR-Cas13 guide RNA predictor for transcript depletion. *BMC Genomics* **23**, 172 (2022). https://doi.org/10.1186/s12864-022-08366-2

Metsky, H.C., Welch, N.L., Pillai, P.P. *et al.* Designing sensitive viral diagnostics with machine learning. *Nat Biotechnol* **40**, 1123–1131 (2022). https://doi.org/10.1038/s41587-022-01213-5

Wessels, HH., Méndez-Mancilla, A., Guo, X. *et al.* Massively parallel Cas13 screens reveal principles for guide RNA design. *Nat Biotechnol* **38**, 722–727 (2020). https://doi.org/10.1038/s41587-020-0456-9

Contributions

Irfan contributed to the conception of the report scope, methods and data generation for plotting. Irfan and Denise contributed and plotted for the subsection on Raw Data Analysis. Irfan contributed and plotted for the subsection on Outlier Detection, RNA Nucleotide Composition Analysis and Target Site Accessibility. Denise contributed and plotted for the subsection on Target Site Accessibility using RNAHybrid, Predicting RNA secondary structures using RNAFold and Conclusion. Irfan and Denise contributed to the writing and final edits of the report.