Cancer Immunotherapy Data Science Challenge 3

Proposing a metric for ranking effective perturbations

Irene Bonafonte[1], Artur Szalata[1], Myriam Lizotte[2], and Benjamin Schubert[1]

[1]Helmholtz Center Munich

[2]Mila - Quebec AI Institute

In this report we introduce our proposal for a metric that ranks a perturbation in terms of their capacity of inducing a desired shift on the cellular state. Our proposal is based on a set of observations regarding the challenge dataset, which we introduce in Section 1. We continue by proposing a statistic $s(\cdot)$ to summarise the gene expression distribution P_i induced by knocking out gene i (Section 2). We finally propose a scoring function $f(s(P_i), P_0, Q))$ —with P_0 the unperturbed cells gene expression and Q the desired cellular state— that evaluates how effective the perturbation of gene i is (Section 3).

1 Observations

1.1 Variability between non-targeting guides

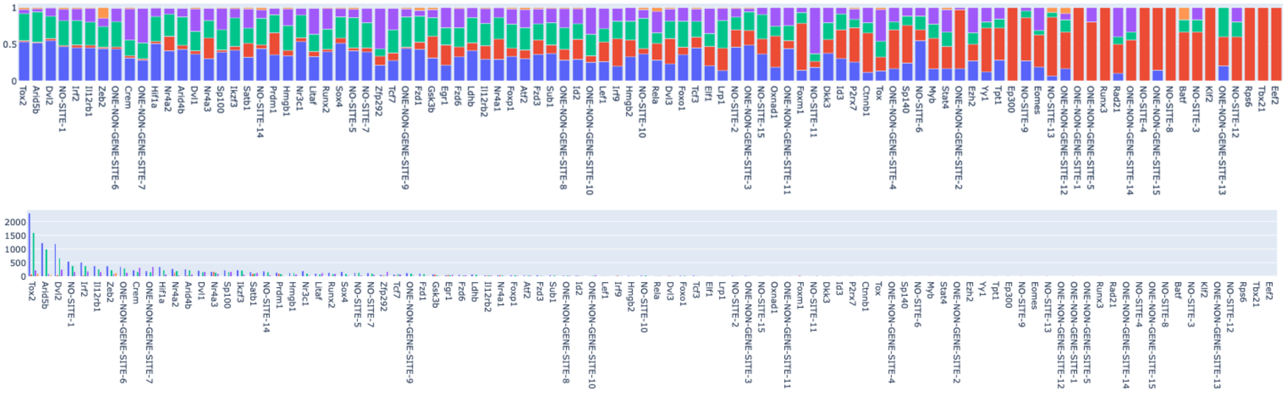

Figure 1: Cell type proportions (top) **and counts** (bottom) **per perturbation and non-targeting guide**. Red: progenitor cells, blue: cycling cells, green: terminal exhausted cells, purple: effector cells, orange: other cells. Perturbations ordered by total number of cells.

Unperturbed cells are cells where one of 30 different guide RNA sequences that do not target the genome has been introduced. Theoretically, they should all have the same (null) effect. However, if we analyse the proportion and cell count distribution between different guides, we can observe a strong variability, likely to be generated by a combination of biological variability and technical noise. As observed in figure 1, non-targeting guides tend to result in a low number of cells and a high percentage of progenitor cells. However, some of them show opposed characteristics, being among the top 10 guides with higher total number of cells, and having an important percentage of effector, cycling and terminal exhausted cells.

1.2 Correlation between total number of cells and cell proportions

We observe in figure 1 that one of the strongest signals distinguishing perturbed and unperturbed cells is the total number of cell counts. In fact, the total number of cells per perturbation has a -0.86 correlation with the proportion of progenitor cells in the logarithmic scale (Figure 4). After dividing perturbations in four different clusters, based uniquely in the total number of cells, one can predict cell proportions with a 0.28 leave-one-out L1-loss by simply predicting the cluster's mean proportions.

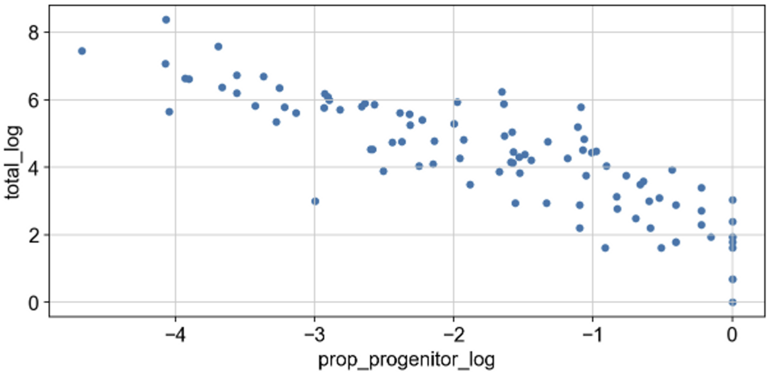

Figure 2: Correlation between total number of cells and proportion of progenitor cells.

We do not regard this correlation to be a technical factor, but a consequence of two different manifestations of the underlying cellular state. Perturbations that result in cells that do not expand will have low cell numbers and a high proportion of cells in the early stages of their life cycle (progenitor cells). Perturbations that react to the tumour and are able to expand will have high cell numbers and high proportions of cells either directly dividing (cycling T cells) or on the latter stages of their life cycle (effector and exhausted T cells). These two related gradients can be observed in Figure 1.

There is a high level of uncertainty in the true proportions for perturbations with only a few observed cells. However, as presented, this same factor is intrinsically associated to the biological signal that we are aiming to predict. Therefore, a metric that gives less importance to perturbations with a low-cell-number and a high uncertainty will effectively ignore all those perturbations that result in cells that do not significantly attack the tumour. We believe that such an approach would unfortunately lead to unreliable results.

1.3 Soft limits between cell clusters

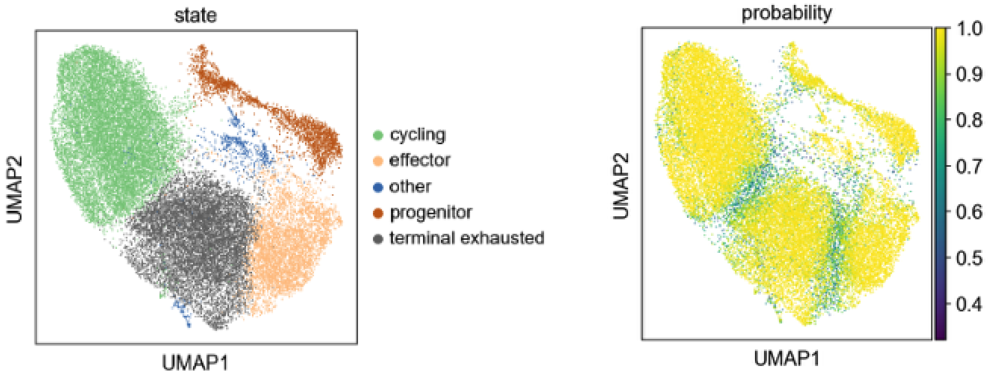

Figure 3: Study cells UMAP coloured by cell type (left) and by maximum cell type probability in a gradient boosting model (right).

Cell assignments are not based on protein surface markers but on the cell gene expression. This defines a gradient of cell states, where a discrete division between distinct cell types is not observed (Figure 3). We have trained a gradient boosting model to predict cell type using the first 20 gene expression principal components. This model can predicts cell type labels with a 0.96 accuracy in a held out test set. By colouring cells with their maximum cell type probability (Figure 3), we can unsurprisingly observe that the confidence for cells in the intersection of two cell clusters can get remarkably low, with values close to 0.3. Thus, it does not seem reasonable to give all cells the same weight.

1.4 Perturbation viability is reflected by number of cells per cell type

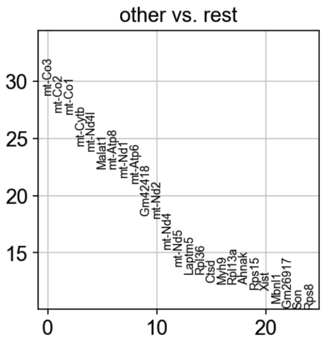

Figure 4: Differentially expressed genes in the other T cell cluster.

As presented above, ineffective perturbations result in a low total number of cells that remain undifferentiated. Additionally, we observe that the "other" T cell cluster is characterised by a high expression of mitochondrial genes. In single cell experiments, a high expression of mitochondrial genes is generally related to apoptotic cells or lysing cells. Therefore, a high proportion of “other” cells is likely to be reflective of a non-viable perturbation.

1.5 Knock out gene expression in perturbed cells

It is known that PerturbSeq experiments do not have a 100% efficacy. In line with this, we observe that some of the perturbed cells have high expression levels of their knock out target gene. In particular, $4,419$ out of $23,719$ perturbed cells show higher target gene expression than the mean expression in unperturbed cells.

2 Statistic proposal

Based on the previous observations, we propose the following pipeline to summarise the gene expression P_i of n_i cells under perturbation i into a single vector S_i:

- Train a classifier to predicts cell type labels based on gene expression (e.g. using principal components).
- Predict the cell type probability vector x_j for each of the perturbed cells j, e.g. $x_j = (0.1, 0.1, 0.8, 0, 0)$.
- Exclude cells that have higher target gene expression than the mean of the unperturbed cells. Possibly, do this conditioned to the cell type (i.e. comparing to the unperturbed mean from the corresponding cell type cluster).
- Average cell type probabilities for the remaining cells per perturbation. This is, do not transform x_j to a hard assignment classification vector $t_j = (0, 0, 1, 0, 0)$. Instead, compute the perturbation statistic S_i by directly averaging soft cell type assignment vectors of all cells, $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_j$.
- Concatenate the proportions vector with a second statistic indicating the total number of cells per perturbation, n_i.
- For the reasons expressed above, we do not consider it to be necessary to include a score that directly captures if a perturbation decreases cell viability or induces a senescent state.

The final vector for perturbation i is $S_i = (\bar{x}_{i,a}, \bar{x}_{i,b}, \bar{x}_{i,c}, \bar{x}_{i,d}, \bar{x}_{i,e}, n_i)$, where a, b, c, d, e correspond each to a different cell type.

3 Scoring function proposal

Based on the previous observations, we propose the following system to score the effectivity of a perturbation using S_i, P_0 and Q. Our scoring system is based on using the cell type distribution for non-perturbing guides to derive the distribution of the parameters for non-effective perturbations. We will model not only the non-targeting guides mean proportions but also their dispersion.

To do this, we assume that our observed unperturbed cell count vectors $Y_0 = \{y_1, ..., y_M\}$ are multinomial distributed with a Dirichlet prior:

$$y_m \sim \text{multi}(N, \theta), \quad \sum_{k=1}^{K} \theta_K = 1$$
$$\theta \sim \text{Dir}(\alpha), \quad \alpha_1, ..., \alpha_K > 0$$

We begin by estimating the model parameters θ posterior probability for the unperturbed cell distribution, $p(\theta|Y_0)$. Once the parameters for the unperturbed samples distribution have been fitted, we can sample from the posterior predictive distribution for each of the different observed perturbations. Using these samples, we can calculate, per each perturbation, the probability of the observed cell type numbers based on the unperturbed data, $p(Y_i|Y_0)$. More precisely, we compute the frequency in which the observed number is greater or lower than that of the samples from the posterior predictive distribution. With this, we get an estimation of how significantly the perturbation diverges from the unperturbed distribution. The code for this can be found in the file `unperturbed_divergence.ipynb`.

If we use as n_i the actual number of cells observed in each case, this would lead to giving more weight to guides with a high number of cells —those guides we are more certain about would have a stronger impact in the inferred distribution. However, we have already exposed that this approach would hide an important part of the signal we are interested in detecting. Therefore, we suggest to normalise the observations to a constant total number of $n_i = 100$ cells.

A similar approach can be followed to measure how much the total cell counts number n_i diverges from the unperturbed distribution. In this case, we suggest to model n_i as a poisson distribution.

The final score for the relevance of a perturbation is a weighted sum of the logarithm of the significance of the change induced in cell type numbers as well as in the total cell number, multiplied each by 1 or -1 based on whether an increase or decrease is induced. Weights can be used to give more importance to cell types of interest or to require an increase or a decrease based on Q. We generally suggest, based on the previous observations, to give a high weight to an increase in the total number of counts.