Tim

A strategy for generating negative control pairs

I describe a strategy for generating undercover pairs for use in the undercover analysis. The method takes the following inputs:

- The number of negative control gRNAs N_{grna}. Label the NT gRNAs $1, 2, \ldots, N_{\text{grna}}$.
- The $N_{\text{cell}} \times N_{\text{gene}}$ matrix of gene expressions and the N_{cell}-dimensional gRNA-to-cell assignment vector.
- The number of pairs to generate N_{pairs}.
- The undercover group size $k \leq N_{\text{grna}}/2$.
- The minimum number of treatment cells N_{trt} and control cells N_{cntrl} used for pairwise QC.

The method constructs a set of undercover groups; each group is mapped to the same number of genes (give or take one gene). All gene-gRNA pairs have at least N_{trt} treatment cells and N_{cntrl} control cells. We proceed as follows.

Initialization step: Tabulate the number of of cells with nonzero expression for each individual NT gRNA and gene. First, we compute an $N_{\text{grna}} \times N_{\text{gene}}$ matrix M, where entry (i, j) gives the number of cells containing NT gRNA i with nonzero expression of gene j. We can easily construct this matrix either in memory or out-of-core. We then proceed in a sequence of rounds to construct the pairs.

Round 1.

Step a. We construct $r_1 := \lfloor N_{\text{NT}}/k \rfloor$ NT gRNA groups $G_1^1, \ldots, G_{r_1}^1$ such that each NT gRNA group G_i^1 contains k unique NT gRNAs. Algorithmically, we construct $G_1^1, \ldots, G_{r_1}^1$ by sampling k elements from $x = \{1, \ldots, N_{\text{grna}}\}$ without replacement $\lfloor N_{\text{NT}}/k \rfloor$ times. To prepare for a subsequent step, we also initialize an empty set D and add $G_1^1, \ldots, G_{r_1}^1$ to the set D.

Step b. Next, for a given NT gRNA group, we determine the set of genes (which we call "valid genes") to which that NT gRNA group could be paired

so that the resulting pairs pass the pairwise QC threshold. Let v_i^1 be the number of valid genes for gRNA group G_i^1. We can calculate v_i^1 by iterating over the matrix M.

Step c. We determine whether there are enough valid genes (across undercover gRNA groups) such that that the total number of undercover gRNA group-gene pairs exceeds the threshold N_{pairs}. Define $v_{\min} = \min_{i \in \{1,\ldots,r_1\}} v_i$. We check if

$$v_{\min} \geq \lfloor N_{\text{pairs}}/r_1 \rfloor + 1. \tag{1}$$

If this equation holds, then we proceed to step d; otherwise, we proceed to round 2.

Step d. Define $b = \lfloor N_{\text{pairs}}/r_1 \rfloor$, and define $l = N_{\text{pairs}} - r_1 \lfloor N/r_1 \rfloor$. Note that $l < r_1$. Define $a_1 = \cdots = a_l = b+1$, and define $a_{l+1} \ldots a_{r_1} = b$. Observe that

$$\sum_{i=1}^{r_1} a_i = l(b+1) + (r_1 - l)b = r_1 b + l = N_{\text{pairs}}.$$

Furthermore, observe that $a_i \leq b + 1 \leq v_{\min}$. Let a_i be the number of genes that we sample for undercover gRNA group G_i^1. We sample these a_i genes without replacement from the "valid" genes for G_i^1.

Round 2. Step a. We construct r_2 NT gRNA groups $G_1^2, \ldots, G_{r_2}^2$ such that each NT gRNA group G_i^2 contains k unique NT gRNAs. Importantly, $G_1^2, \ldots, G_{r_2}^2$ are constructed such that $G_i^2 \notin D$, i.e., the G_i^2s are distinct from the G_i^1s. We construct the G_i^2s in the following way.

First, we sample k elements without replacement from $x = \{1, \ldots, N_{\text{grna}}\}$ to form G_1^2. We check if G_1^2 is present in D; if so, we start the process again. If not, we proceed, sampling k elements from the elements that remain in x. We stop sampling when either fewer than k elements remain or we have sampled $\binom{N_{\text{grna}}}{k}$ total undercover gRNA groups.

Step b. For gRNA groups $G_1^2, \ldots, G_{r_2}^2$, we compute the number of valid genes $v_1^2, \ldots, v_{r_2}^2$ to which each gRNA group can be paired.

Step c. We determine if there are enough valid genes across undercover gRNA groups OR we have exhausted the total number of undercover gRNA groups. We check the equation

$$v_{\min} \geq \lfloor N_{\text{pairs}}/(r_1 + r_2) \rfloor + 1,$$

where $v_{\min}$ is taken over $G_1^1, \ldots, G_{r_1}^1, G_1^2, \ldots, G_{r_2}^2$. If this inequality is satisfied

(or if we have exhausted all possible undercover groups), we proceed to round 3. Otherwise, we proceed to step d.

Step d. Define $b = \lfloor N_{\text{pairs}}/(r_1 + r_2) \rfloor$, and define $l = N_{\text{pairs}} - (r_1 + r_2)\lfloor N/(r_1 + r_2) \rfloor$. Define $a_1 \ldots, a_l = b + 1$ and $a_{l+1} = \cdots = a_{r_1+r_2} = b$. Sample this many genes per undercover gRNA group.