Supplementary methods

Emma Dann, Michael D. Morgan

22 October, 2020

Description of *Milo*

Building the KNN graph

Similarly to many other tasks in single-cell analysis, *Milo* uses a KNN graph computed based on similarities in gene expression space as a representation of the phenotypic manifold in which cells lie. While *Milo* can be used on graphs built with different similarity kernels, here we compute the graph as follows: for a gene expression matrix of N cells is projected onto the first d principal components (PCs) to obtain a $N \times d$ matrix X_{PC}. Then, for each cell i, the euclidean distances to its k nearest neighbors in X_{PC} are computed and stored in a $N \times N$ adjacency matrix. Then, D is symmetrized, such that cells i and j are nearest neighbors (i.e. connected by an edge) if either i is nearest neighbor of j or j is nearest neighbor of i. The KNN graph is encoded by the undirected symmetric version of $\tilde{D}$ of D, where each cell has at least K nearest neighbors.

Definition of cell neighbourhoods and index sampling algorithm

We define the neighbourhood n_i of cell i as the group of cells that are connected to i by an edge in the graph. Formally, a cell j belongs to neighbourhood n_i if $\tilde{D}_{i,j} > 0$. We refer to i as the index of the neighbourhood.

In order to define a representative subset of neighbourhoods that span the whole KNN graph, we implement a previously adopted algorithm to sample the index cells in a graph (Gut et al. 2015; Setty et al. 2016). Briefly, we start by randomly sampling $p \cdot N$ cells from the dataset, where $p \in [0, 1]$ (we use $p = 0.1$ by default). Given the reduced dimension matrix used for graph construction X_{PC}, for each sampled cell we consider its k nearest neighbors $j = 1, 2, ..., k$ with PC profiles $x_1, x_2, ..., x_k$. We measure the mean PC profile $\bar{x}$ for the j cells and search for the cell i such that the euclidean distance between x_i and $\bar{x}$ is minimized. This yields a set of $M \leq p \cdot N$ index cells that are used to define neighbourhoods.

Testing for differential abundance in neighbourhoods

Milo builds upon the framework for differential abundance testing implemented by *Cydar* (Lun, Richard, and Marioni 2017). In this section, we briefly describe the statistical model and adaptations to the KNN graph setting.

Quasi-likelihood negative bionomial generalized linear models

We consider a neighbourhood n with cell counts y_{ns} for each sample s. The counts are modelled by the negative binomial (NB) distribution, as it is supported over all non-negative integers and can accurately model both small and large cell counts. For such non-normally distributed data we use generalized-linear models (GLMs) as an extension of classic linear models that can accomodate complex experimental designs. We therefore assume that

$$y_{ns} \sim NB(\mu_{ns}, \phi_n),$$

where μ_{ns} is the mean and ϕ_n is the NB dispersion parameter. The expected count value for neighbourhood n in sample s μ_{ns} is given by

$$\mu_{ns} = \lambda_{ns} N_s$$

where λ_{ns} is the proportion of cells belonging to sample s in n and N_s is the total number of cells of s. In practice, λ_{ns} represents the biological variability that can be affected by treatment condition, age or any biological covariate of interest. We use a log-linear model to represent the influence of the biological condition on the expected counts in neighbourhoods:

$$log\ \mu_{ns} = \sum_{g=1}^{G} x_{sg}\beta_{ng} + log\ N_s$$

where $x_s g$ is the covariate vector indicating the condition applied to sample s and β_{ng} is the regression coefficient by which the covariate effects are mediated for neighbourhood n.

Estimation of β_{ng} for each n and g is performed by fitting the GLM to the count data for each neighbourhood, i.e. by estimating the dispersion ϕ_n that models the variability of cell counts for replicate samples for each neighbourhood. Dispersion estimation is done using the quasi-likelihood method in `edgeR`(Robinson, McCarthy, and Smyth 2010), where the dispersion is modelled from the GLM deviance and stabilized with empirical Bayes shrinkage, to stabilize the estimates in the presence of limited replication.

Adaptation of Spatial FDR to neighbourhoods

To control for multiple testing, we adapt the Spatial FDR method introduced by *Cydar* (Lun, Richard, and Marioni 2017). The Spatial FDR can be interpreted as the proportion of the union of neighbourhoods that is occupied by false-positive neighbourhoods. This accounts for the fact that some neighbourhoods are more densely connected than others. To control spatial FDR in the KNN graph, we apply a weighted version of the Benjamini-Hochberg (BH) method. Briefly, to control for FDR at some threshold α we reject null hypothesis i where the associated p-value is less than the threshold

$$\max_i p_{(i)} : p_{(i)} \leq \alpha \frac{\sum_{l=1}^{i} w_{(l)}}{\sum_{l=1}^{n} w_{(l)}}$$

Where the weight $w_{(i)}$ is the reciprocal of the neighbourhood connectivity c_n. As a measure of neighbourhood connectivity, we use the euclidean distance to the kth nearest neighbour of the index cell for each neighbourhood.

References

Gut, Gabriele, Michelle D. Tadmor, Dana Pe'er, Lucas Pelkmans, and Prisca Liberali. 2015. "Trajectories of Cell-Cycle Progression from Fixed Cell Populations." *Nature Methods* 12 (10): 951–54. https://doi.org/10.1038/nmeth.3545.

Lun, Aaron T. L., Arianne C. Richard, and John C. Marioni. 2017. "Testing for Differential Abundance in Mass Cytometry Data." *Nature Methods* 14 (7): 707–9. https://doi.org/10.1038/nmeth.4295.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Setty, Manu, Michelle D. Tadmor, Shlomit Reich-Zeliger, Omer Angel, Tomer Meir Salame, Pooja Kathail, Kristy Choi, Sean Bendall, Nir Friedman, and Dana Pe'er. 2016. "Wishbone Identifies Bifurcating Developmental Trajectories from Single-Cell Data." *Nature Biotechnology* 34 (6): 637–45. https://doi.org/10.1038/nbt.3569.