Supervised Learning-based Receptor Abundance Estimation using STREAK: An Application to the 10X Genomics human extranodal marginal zone B-cell tumor/mucosa-associated lymphoid tissue (MALT) dataset

Azka Javaid and H. Robert Frost

Load the STREAK package

STREAK is a supervised receptor abundance estimation method that depends on functionalities from the Seurat (Hao et al. 2021; Stuart et al. 2019; Butler et al. 2018; Satija et al. 2015), SPECK (Frost and Javaid 2022), VAM (Frost 2021) and Ckmeans.1d.dp (Wang and Song 2011; Song and Zhong 2020) packages.

```
library(STREAK)
```

Receptor gene set construction using a subset of joint scRNA-seq/CITE-seq training data

STREAK performs receptor abundance estimation by leveraging expression associations learned from joint scRNA-seq/CITE-seq training data. These associations can either be manually specified using pre-existing ground truth or can be built using a subset of joint transcriptomics and proteomics data. Below, we use a subset of 1000 cells from the 10X Genomics human extranodal marginal zone B-cell tumor/mucosa-associated lymphoid tissue (MALT) scRNA-seq/CITE-seq joint dataset to build a gene set weights membership matrix for the CD3, CD4, CD8a, CD14 and CD15 receptors. Given a $m \times n$ training scRNA-seq counts matrix and a $m \times h$ CITE-seq matrix, the `receptorGeneSetConstruction()` function is utilized to learn associations between each CITE-seq ADT transcript and all scRNA-seq transcripts. The resulting gene weights membership matrix is $n \times h$.

```
data("train.malt.rna.mat")
data("train.malt.adt.mat")
receptor.geneset.matrix.out <- receptorGeneSetConstruction(train.rnaseq =
                                                             train.malt.rna.mat,
                                                           train.citeseq =
                                                             train.malt.adt.mat[,1:5],
                                                           rank.range.end = 100,
                                                           min.consec.diff = 0.01,
                                                           rep.consec.diff = 2,
                                                           manual.rank = NULL,
                                                           seed.rsvd = 1)
dim(receptor.geneset.matrix.out)
#> [1] 33538     5
head(receptor.geneset.matrix.out)
#>                      CD3        CD4        CD8a        CD14        CD15
```

```
#> MIR1302-2HG 0.072643688 0.06139920 -0.04973724 -0.04441452 -0.04178776
#> FAM138A     0.614008401 0.27427471 -0.11811688 -0.58005689 -0.59082835
#> OR4F5       0.001950812 0.25261813 -0.15629596  0.01274566 -0.07105958
#> AL627309.1  0.067616000 0.08191478 -0.09400833 -0.09638293 -0.06789212
#> AL627309.3  0.082064369 0.19007994 -0.15351573 -0.02229459 -0.11769248
#> AL627309.2  0.093341494 0.05361682  0.03018559 -0.14336654 -0.11781685
```

Receptor abundance estimation for target scRNA-seq data

Following the development of weighted gene sets, the `receptorAbundanceEstimation()` function is used to perform receptor abundance estimation. A subset of 1100 cells from the 10X Genomics MALT scRNA-seq data is used for estimation. Given a $m \times n$ target scRNA-seq counts matrix and a $n \times h$ gene set weights membership matrix, target scRNA-seq expression from top most weighted genes with each ADT transcript is used for gene set scoring and subsequent thresholding. The resulting estimated receptor abundance matrix is $m \times h$.

```
data("target.malt.rna.mat")
receptor.abundance.estimates.out <-
  receptorAbundanceEstimation(target.rnaseq = target.malt.rna.mat,
                              receptor.geneset.matrix =
                                receptor.geneset.matrix.out,
                              num.genes = 10, rank.range.end = 100,
                              min.consec.diff = 0.01, rep.consec.diff = 2,
                              manual.rank = NULL, seed.rsvd = 1,
                              max.num.clusters = 4, seed.ckmeans = 2)
dim(receptor.abundance.estimates.out)
#> [1] 1100    5
head(receptor.abundance.estimates.out)
#>                            CD3 CD4      CD8a      CD14      CD15
#> CTACCTGAGAGCGACT-1 0.0000000   0 0.9987944 0.6740526 0.7753415
#> TGGCGTGCACAGCATT-1 0.9464793   0 0.0000000 0.0000000 0.0000000
#> TAGGAGGAGCTGGCCT-1 0.0000000   0 0.0000000 0.9992784 0.9988085
#> ACTATCTCACCCTATC-1 0.0000000   0 0.9982689 0.1559718 0.2513592
#> ACGGAAGTCAATCCGA-1 0.0000000   0 0.9957439 0.5229880 0.6813975
#> AAGTACCCACAGAGCA-1 0.0000000   0 0.0000000 0.9990658 0.9985386
```

References

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. 2018. "Integrating Single-Cell Transcriptomic Data Across Different Conditions, Technologies, and Species." *Nature Biotechnology* 36: 411–20. https://doi.org/10.1038/nbt.4096.

Frost, H. Robert. 2021. *VAM: Variance-Adjusted Mahalanobis*. https://CRAN.R-project.org/package=VAM.

Frost, H. Robert, and Azka Javaid. 2022. *SPECK: Receptor Abundance Estimation Using Reduced Rank Reconstruction and Clustered Thresholding*. https://CRAN.R-project.org/package=SPECK.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, et al. 2021. "Integrated Analysis of Multimodal Single-Cell Data." *Cell*. https://doi.org/10.1016/j.cell.2021.04.048.

Satija, Rahul, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. 2015. "Spatial Reconstruction of Single-Cell Gene Expression Data." *Nature Biotechnology* 33: 495–502. https://doi.org/10.1038/nbt.3192.

Song, Mingzhou, and Hua Zhong. 2020. "Efficient Weighted Univariate Clustering Maps Outstanding Dysregulated Genomic Zones in Human Cancers." *Bioinformatics* 36 (20): 5027–36. https://doi.org/10.1093/bioinformatics/btaa613.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. "Comprehensive Integration of Single-Cell Data." *Cell* 177: 1888–1902. https://doi.org/10.1016/j.cell.2019.05.031.

Wang, Haizhou, and Mingzhou Song. 2011. "Ckmeans.1d.dp: Optimal k-Means Clustering in One Dimension by Dynamic Programming." *The R Journal* 3 (2): 29–33. https://doi.org/10.32614/RJ-2011-015.