PNAS
www.pnas.org

Supplementary Information for

Model-free prediction test with application to genomics data

Zhanrui Cai, Jing Lei, Kathryn Roeder.

Kathryn Roeder
E-mail: roeder@andrew.cmu.edu

This PDF file includes:

Supplementary text
Figs. S1 to S30 (not allowed for Brief Reports)
Table S1 (not allowed for Brief Reports)
SI References

Supporting Information Text

Details of Simulations in Predictability Test

We let sample size $n = 200$. $X \in \mathbb{R}^{1000}$ is generated from the standard normal distribution and $\varepsilon \in \mathbb{R}^1$ is drawn from the standard Cauchy distribution. Y is set to be only related to X_1 (the first element in X) and ε. We consider very simple models, as those are most effective in illustrating the power performance of all methods. Xgboost is implemented as the regression algorithm.

- $Y = a \times 10X_1 + \varepsilon$;
- $Y = a \times 10\sin(X_1) + \varepsilon$;
- $Y = a \times 10\exp(X_1) + \varepsilon$;
- $Y = a \times 100\log(X_1 + 10) + \varepsilon$;

The signal level a is set to vary from 0 to 1 to fully investigate the size and power of each test. We repeat the simulation 1000 times and report the average results. We report the additional power curves for all the methods in Fig. S1 and S2, where the type-I error rate is set as 0.01. Our methods perform very well under all settings. Moreover, the increasing dimension does not have negative effects on the proposed methods. For MDC, the power decreases dramatically as the dimension of X increases from 100 to 1000. As the dimension increases up to 5000, the RS-M still performs very well compared to other methods, only with slightly decrease in the power. In fact, the power of RS-M varies between 0.95 to 1 (Fig. S3) when the dimension o the covariates is between 1000 and 5000. This also echos the real data analysis on the PBMC data where the high dimensional noise may cause only a fraction of tests to lose power.

To demonstrate the superior performance of the machine learning algorithm (Xgboost), we also implement the linear regression as the regression algorithm. We set $n = 200$, $d = 150$, and still use the same models to compare all the methods. We summarize the results in Fig. S4 where the significance level α is set as 0.05. As expected, the type-I error of all methods are well controlled. However, linear regression is only able to give non-trivial powers when the true model is linear. For the other three models, linear regression fails to capture the nonlinear trends in the data while Xgboost successfully did. The performance of the linear regression is also worse than the Xgboost under the linear model. Note that only the first column in X is related to Y and all other variables are noises. One explanation is that Xgboost is able to handle the high dimensional noisy variables in a more efficient way.

Details of Simulations in Spatially Variable Genes Detection

To compare the power of each test under both null hypothesis and alternative hypothesis, we generate the signals according to patterns, and add a random noise that follows uniform distribution on $[0, 1]$ to each spot. Specifically, denote the signal as $f(X_i)$, where X_i is the spatial information at location i. The gene expression at location i is generated by

$$Y_i = a \times f(X_i) + U_i, \quad U_i \sim U(0, 1).$$

For all three patterns, we describe the model details as follows.

Hotspot: We randomly choose a spot $\widetilde{X}$ whose horizontal and vertical axis both follow a uniform distribution between the range of $\{X_1, \ldots, X_n\}$. Let d_i denote the Euclidean distance between X_i and $\widetilde{X}$, then we set

$$Y_i = a \times \frac{\max_j(d_j) - d_i}{\max_j(d_j) - \min_j(d_j)} + U_i.$$

Gradient: Let $X_{i,1}$ denote the horizontal axis of spot i and X_m be the smallest horizontal axis, i.e., $X_m = \min_j(X_{j,1})$. Then we set

$$Y_i = a \times \frac{X_{i,1} - X_m}{\max_k\{X_{k,1} - X_m\}} + U_i.$$

Streak: We randomly choose a spot $\widetilde{X}$ whose horizontal axis is between the 0.4 and 0.6 quantile of the horizontal axis of $\{X_1, \ldots, X_n\}$. Let h be a tuning parameter adjusting the width of the streak. Then

$$Y_i = a \times \mathbb{I}(|X_{i,1} - \widetilde{X}_1| \leq h) + U_i,$$

We repeat the simulation 1000 times and report the average results. Besides the figures in the main paper, we also report the power performance of all methods in Figure S5 where the type-I error α is set to be 0.01.

Additional Real Data Analysis

A. Mouse Olfactory Bulb Data and Human Breast Cancer Data. We further investigate the genes only uniquely identified by our method (TT-M) in comparison with the genes only identified by SPARK. We found that the genes only identified by our methods in general are less sparse, have higher expression values and higher standard deviation (Fig. S8). This indicates that the proposed method is more likely to catch the genes with more complicated structures.

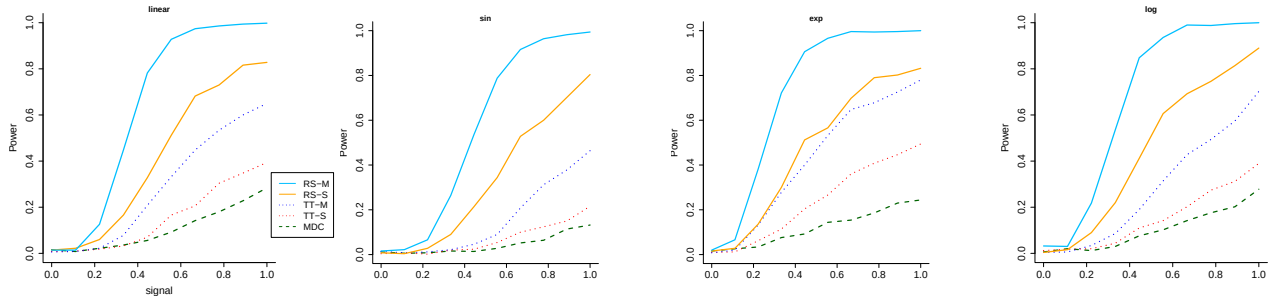

Fig. S1. The power versus signal for the all the methods when the response has heavy-tailed distribution and $\alpha = 0.01$. The RS stands for rank-sum test and TS stands for two sample t-test. M means multiple split while S represents single split.

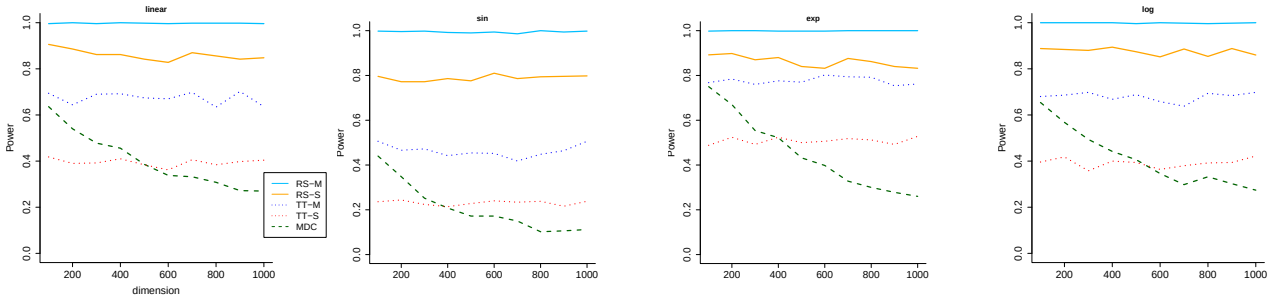

Fig. S2. The power versus dimension for the all the methods when the response has heavy-tailed distribution and the dimension of the covariates increases from 100 to 1000. $\alpha = 0.01$. The RS stands for rank-sum test and TS stands for two sample t-test. M means multiple split while S represents single split.

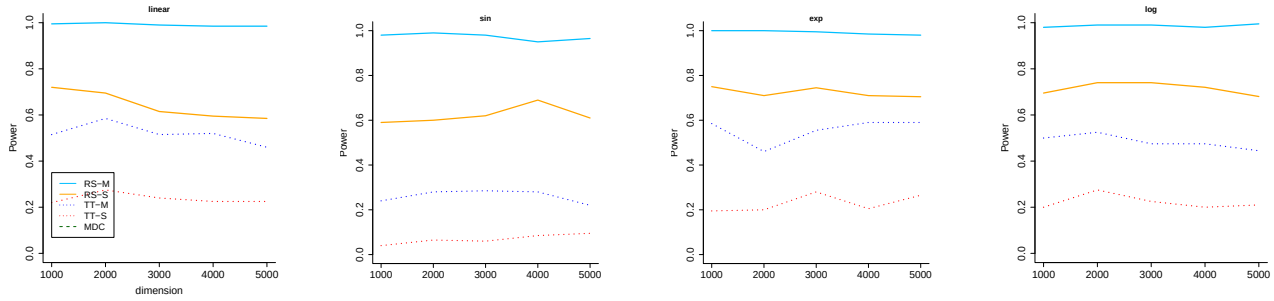

Fig. S3. The power versus dimension for the all the methods when the response has heavy-tailed distribution and the dimension of the covariates increases from 1000 to 5000. $\alpha = 0.01$. The RS stands for rank-sum test and TS stands for two sample t-test. M means multiple split while S represents single split.

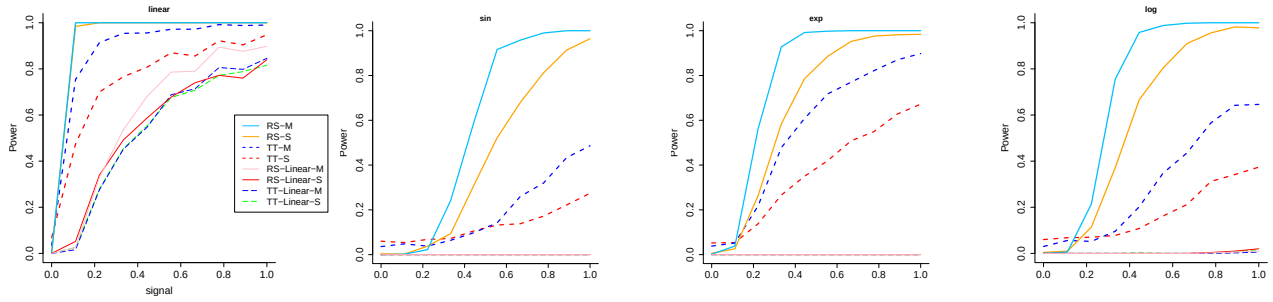

Fig. S4. The power versus dimension for the all the methods when the response has heavy-tailed distribution and the dimension of the covariates increases. $\alpha = 0.05$. The RS stands for rank-sum test and TS stands for two sample t-test. M means multiple split while S represents single split. Linear represent the regression algorithm being linear regression.

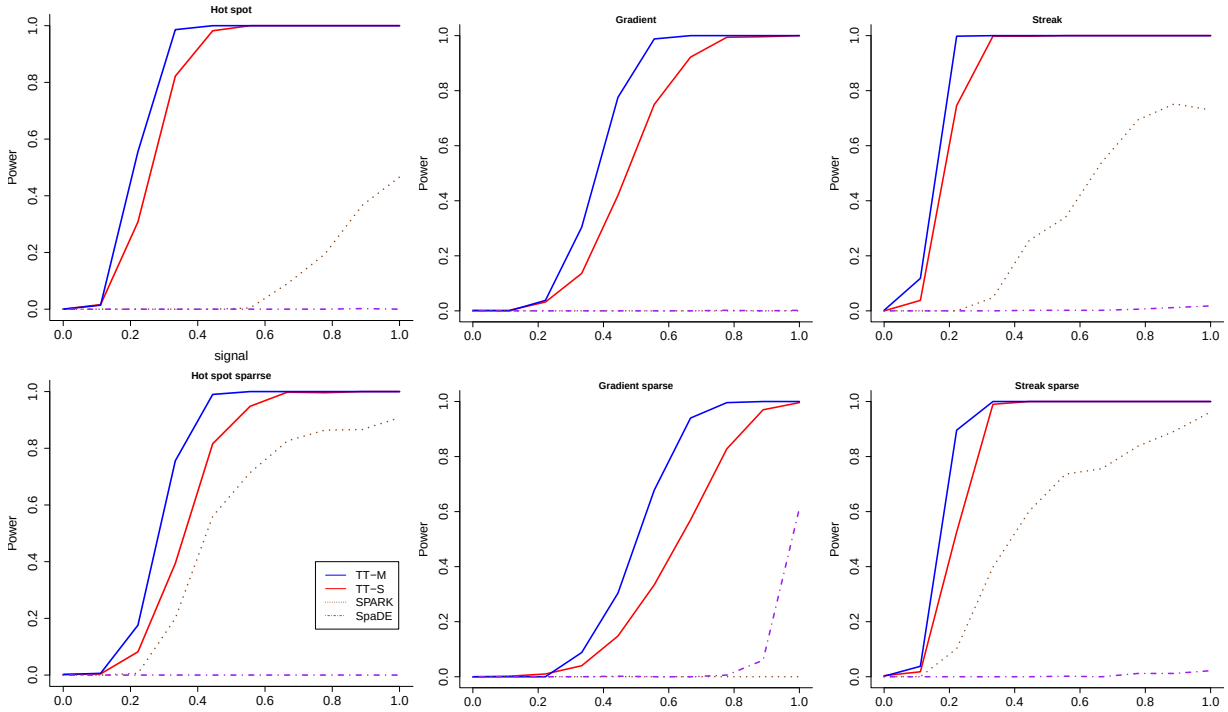

Fig. S5. The power versus signal for the all the methods when $\alpha = 0.01$. The first row are from the three patterns with non-sparse setting, and the second row are from the three patterns with sparse setting. The blue solid line and red solid line denote the multiple splits (TT-M) and single splits (TT-S). The green dashed line denote SPARK, while the purple dotted line denotes SpaDE, respectively.

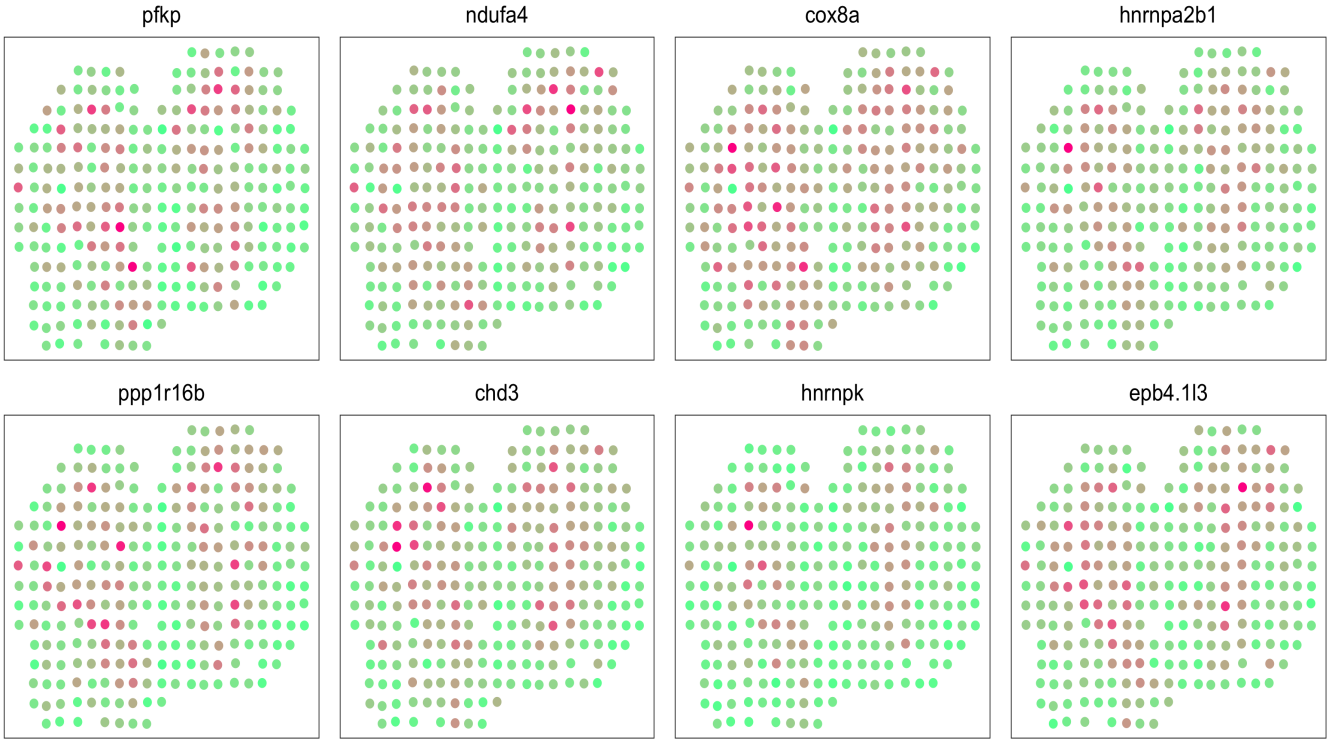

Fig. S6. The 8 genes that have the smallest p-values uniquely detected by TT-M in the mouse olfactory bulb data.

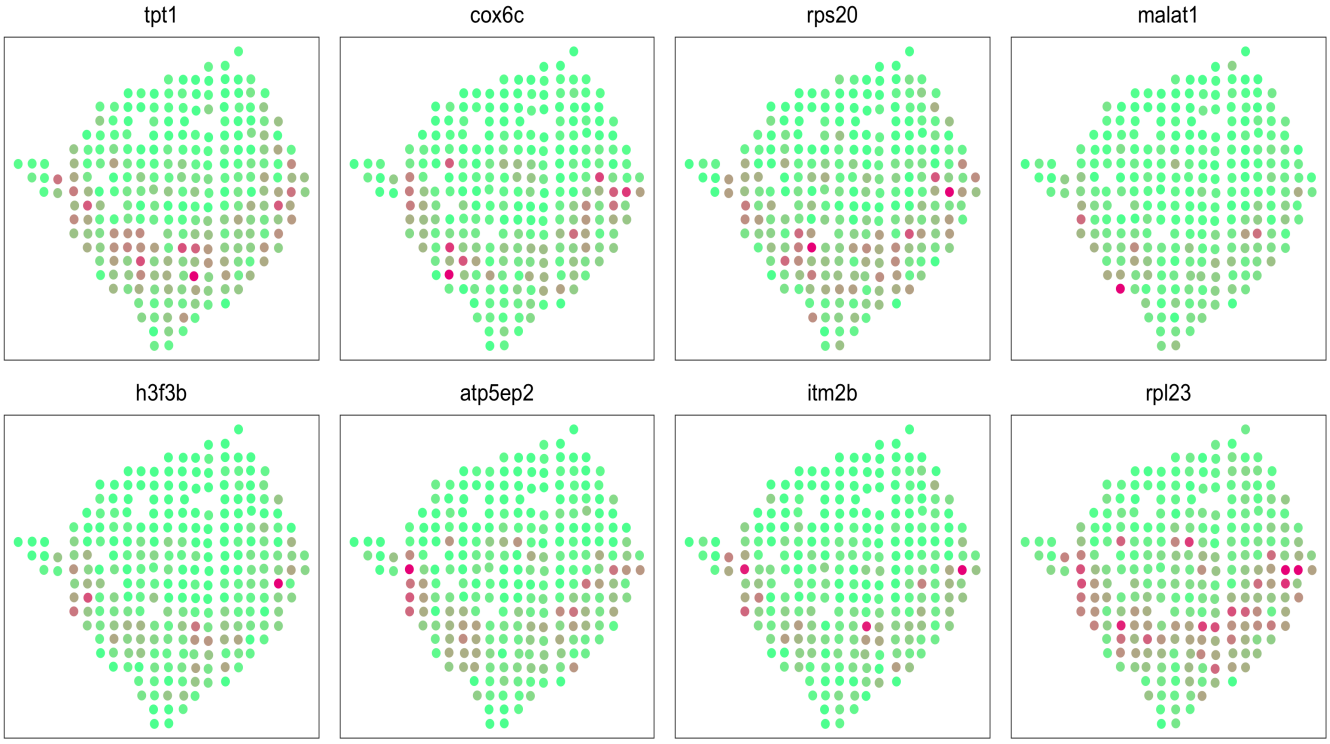

Fig. S7. The 8 genes that have the smallest p-values uniquely detected by TT-M in the human breast cancer data.

Hypothalamus Data. In this section, we analyze a MERFISH dataset collected on the preoptic area of the mouse hypothalamus (1). The dataset contains 160 genes measured in 4,975 single cells. In the original study, 155 of the 160 genes were either labeled as markers of distinct cell populations or are relevant to various neuronal functions of the hypothalamus. Thus a large

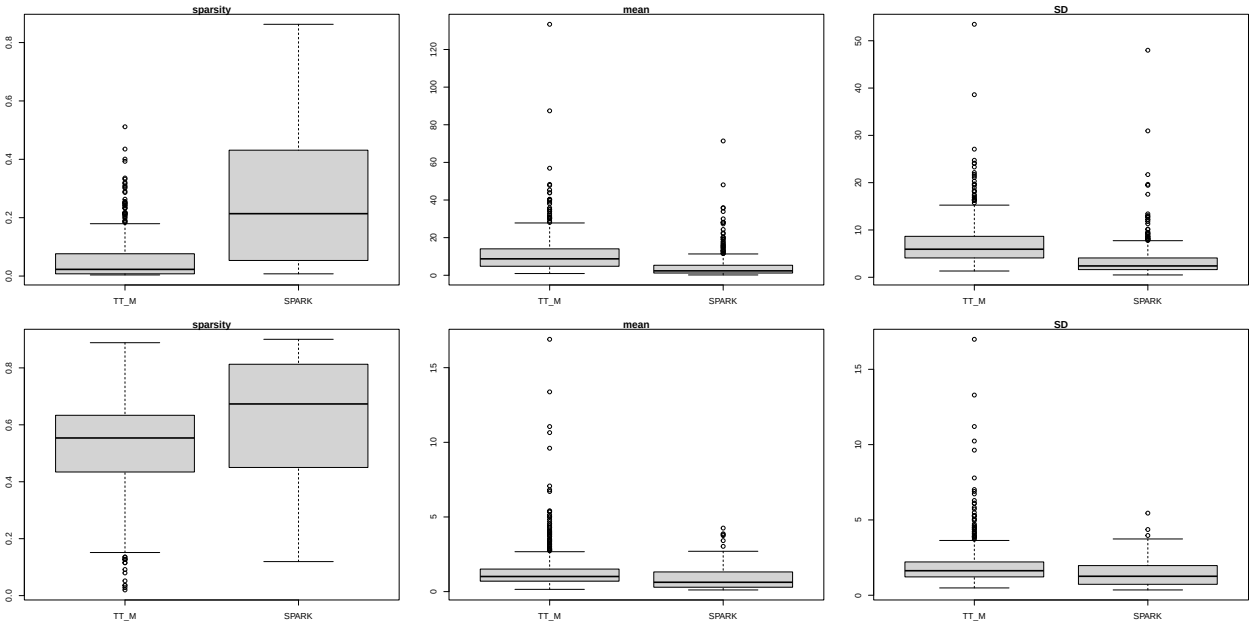

Fig. S8. Comparison of the genes only identified by our method and the genes only identified by SPARK in the mouse olfactory bulb data and the human breast cancer data. Three criterion are considered: sparsity (ratio of elements equal to zero), mean expression, standard deviation of the expression. The first row is based on the mouse olfactory bulb data and the second row is based on the human breast cancer data.

number of genes might be selected as spatially variable genes. We report the analysis result in Fig.S9. As we expected, the empirical distribution of the p-values for our proposed methods are valid under the null condition, since their distributions are clearly below the diagonal line, see Fig.S9.(a). On the other hand, SpaDE also produces uniformly valid p-values, while SPARK does not. The empirical distribution of p-values obtained by SPARK crosses the diagonal line several times, which indicates that the test could have a high false positive rate under the alternative hypothesis. Indeed, for the real data analysis, TT-S found 105 genes and TT-M found 115 genes, while SPARK found 143 genes and SpaDE found 124 genes. The genes detected by TT-S and TT-M all overlap with the findings of SPARK and SpaDE, which confirms the validity of the proposed methods. The GO annotation reveals that the detected genes are not only related to protein bindings, but also other cell functions such as transcription coactivator activity.

In terms of variable importance, %IncMSE equals 0.581 for the horizontal axis, 0.549 for the vertical axis, and is 0.665 for the interaction effect of the horizontal and vertical axis. Thus the spatial patterns are more equally spread across both the vertical and horizontal axis. This fact can also be observed from the 8 genes with the smallest p-values detected by TT-M, as illustrated in Fig.S10.

Hippocampus Data. The last dataset was a small seqFISH dataset with 249 genes measured on 131 single cells in the mouse hippocampus (2). SPARK found 17 genes and SpaDE found 11 genes. The results are reported in Figure S11. As expected, all methods produce valid p-values under the null condition. For the real analysis, TT-S did not find any genes while TT-M identified 3 genes (*lyve, mog, myl14*, shown in Fig.S11c) and all of them overlap with the previous two approaches.

While all methods find less than 20 spatially variable genes for this dataset, we suspect the main reason that our method identifies the least amount of genes is that the sample size of 131 is too small for our method. The proposed method is based on sample splitting, thus the effective sample size for both regression and testing is at most 66 in this case. These samples are far from enough to train a proper machine learning model. As expected, we did not perform as well as the existing parametric models (SPARK or SpaDE).

Table S1. Cell type and protein names where the top 5,000 genes reject H_0 but the marker genes fail to reject

cell type	protein
CD8 T	TCR-2
other	Siglec-8
Mono	CD284
CD8 T	CD49a
CD4 T	TCR-V-7.2
CD4 T	CD177
other	CD8a
other	CD324
NK	CLEC2
other	CD3-2

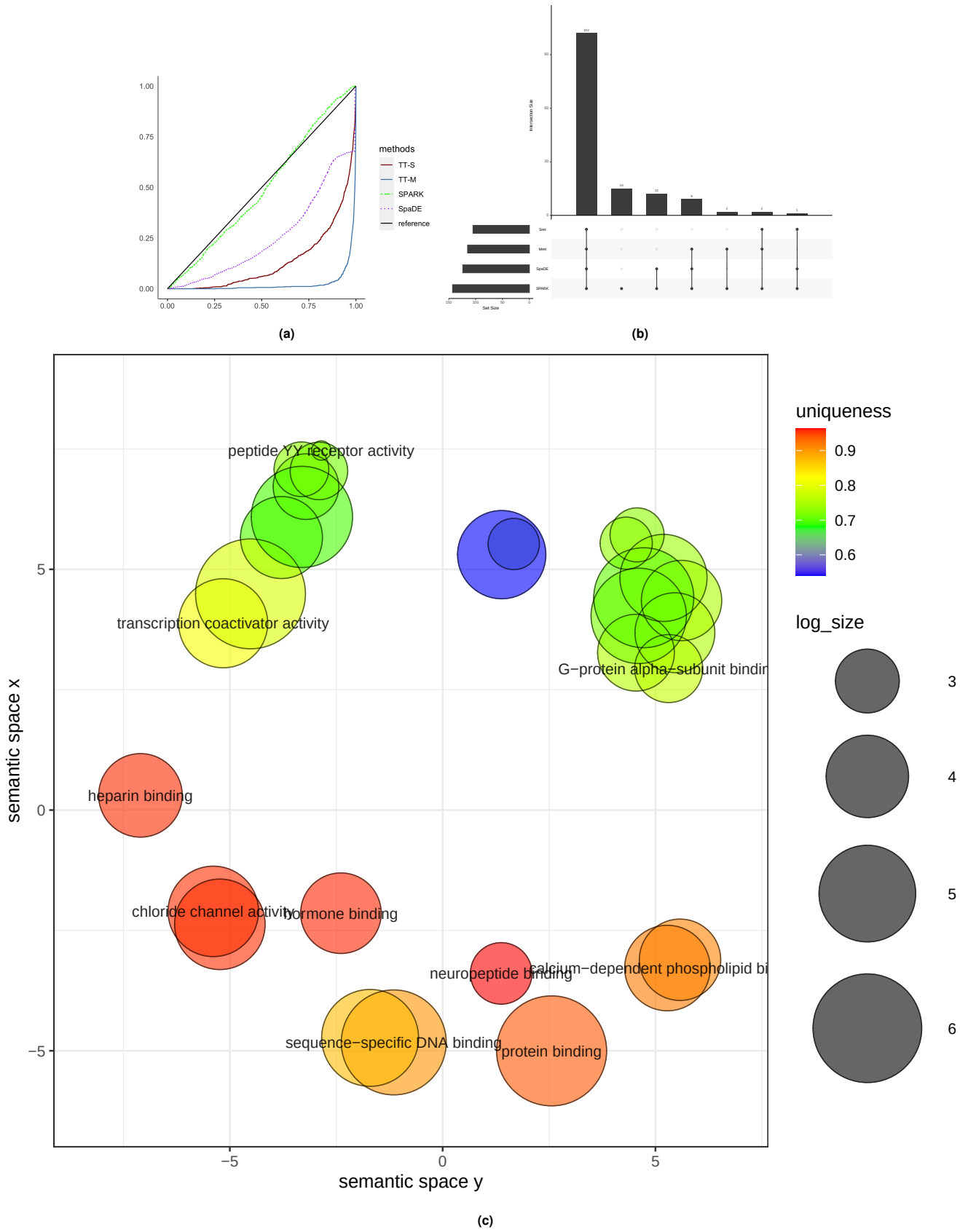

Fig. S9. Analysis for the MERFISH dataset. (a): The empirical distribution of the p-values under the null condition in the permuted data. The blue solid line and red solid line denote the multiple splits (TT-M) and single splits (TT-S). The green dashed line denote SPARK, while the purple dotted line denotes SpaDE, respectively. (b): The upset plot shows the overlap of genes for TT-S and TT-M compared with SPARK and SpaDE. (c): The clustering of GO annotations for the genes detected by TT-M.

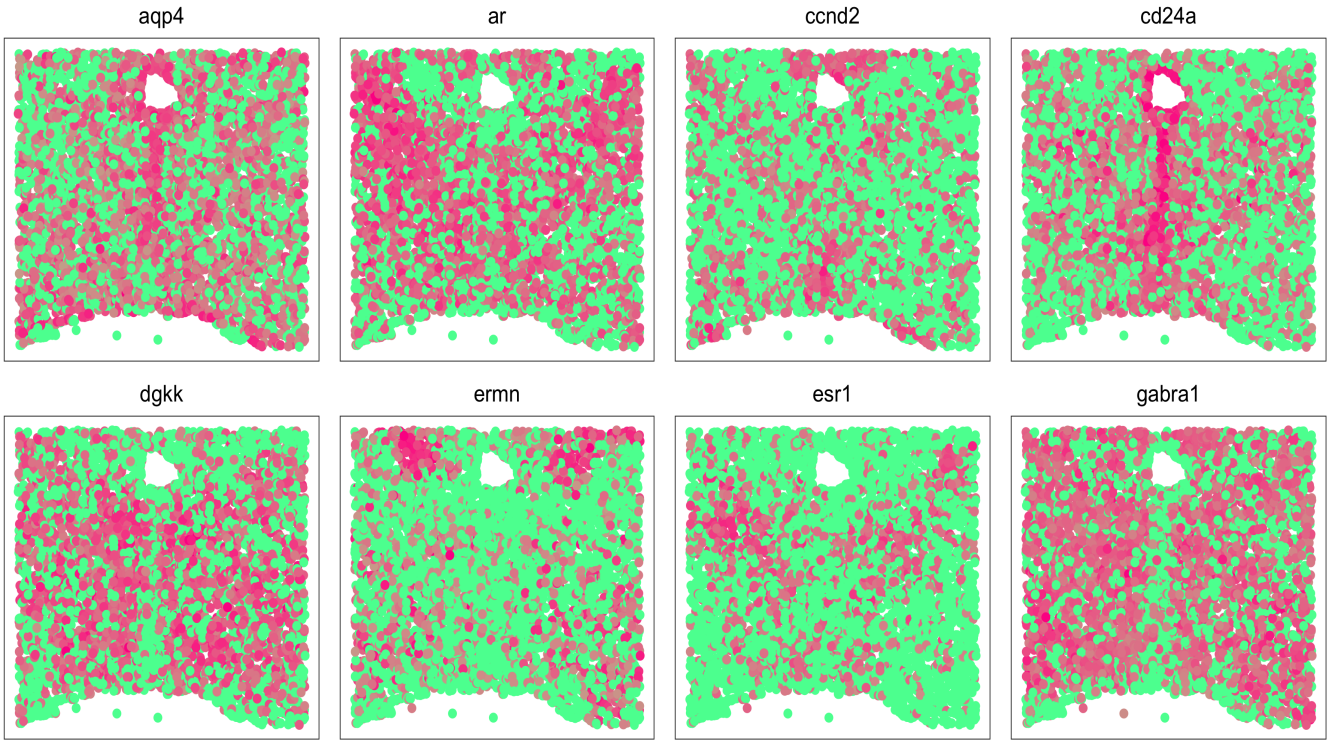

Fig. S10. The 8 genes that have the smallest p-values detected by TT-M in the mouse hypothalamus data .

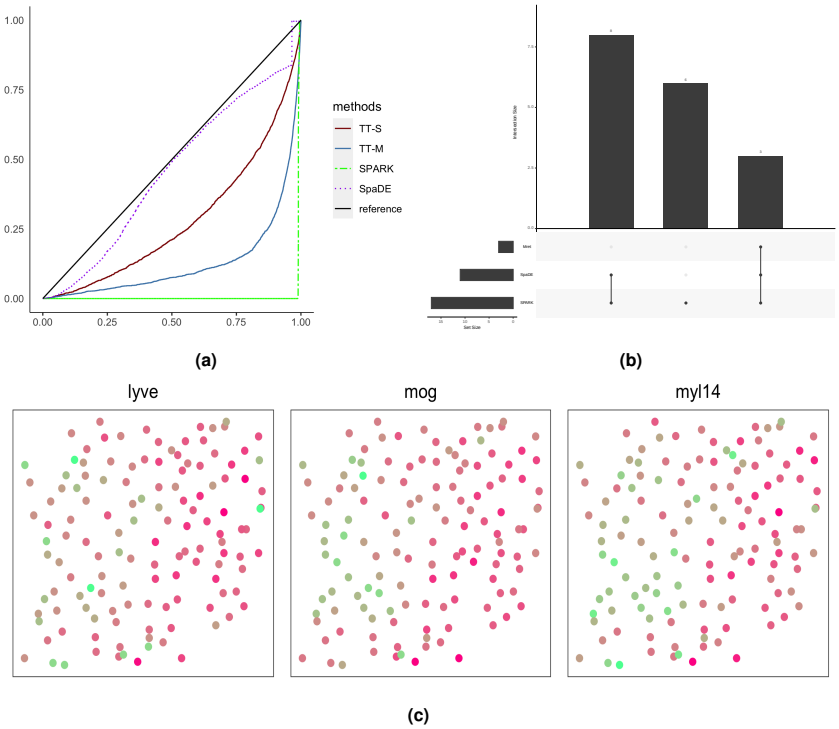

Fig. S11. Analysis for the seqFISH dataset. (a): The empirical distribution of the p-values under the null condition in the permuted data. The blue solid line and red solid line denote the multiple splits (TT-M) and single splits (TT-S). The green dashed line denote SPARK, while the purple dotted line denotes SpaDE, respectively. (b): The upset plot shows the overlap of genes for TT-M compared with SPARK and SpaDE. (c): The 3 genes detected by TT-M in the mouse hippocampus data.

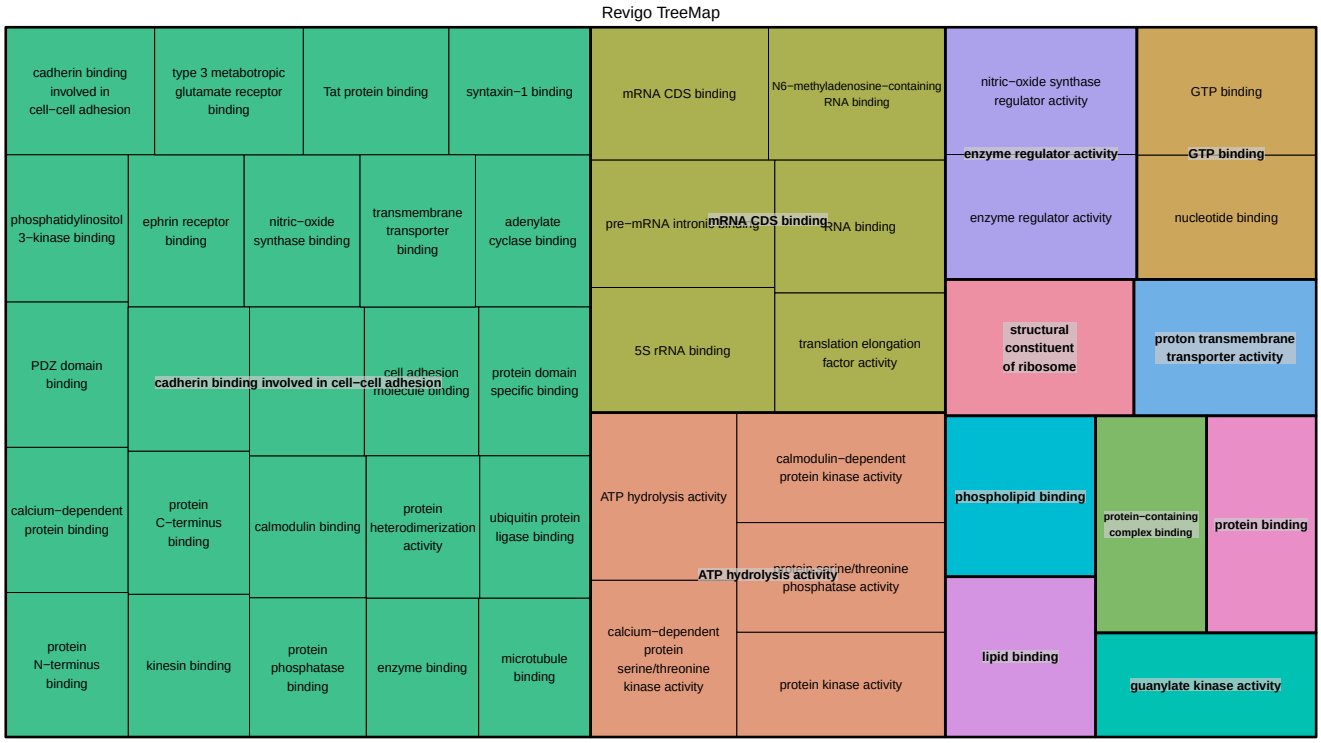

Fig. S12. The mouse olfactory bulb data: Revigo tree map for the GO annotations based on genes detected by TT-S.

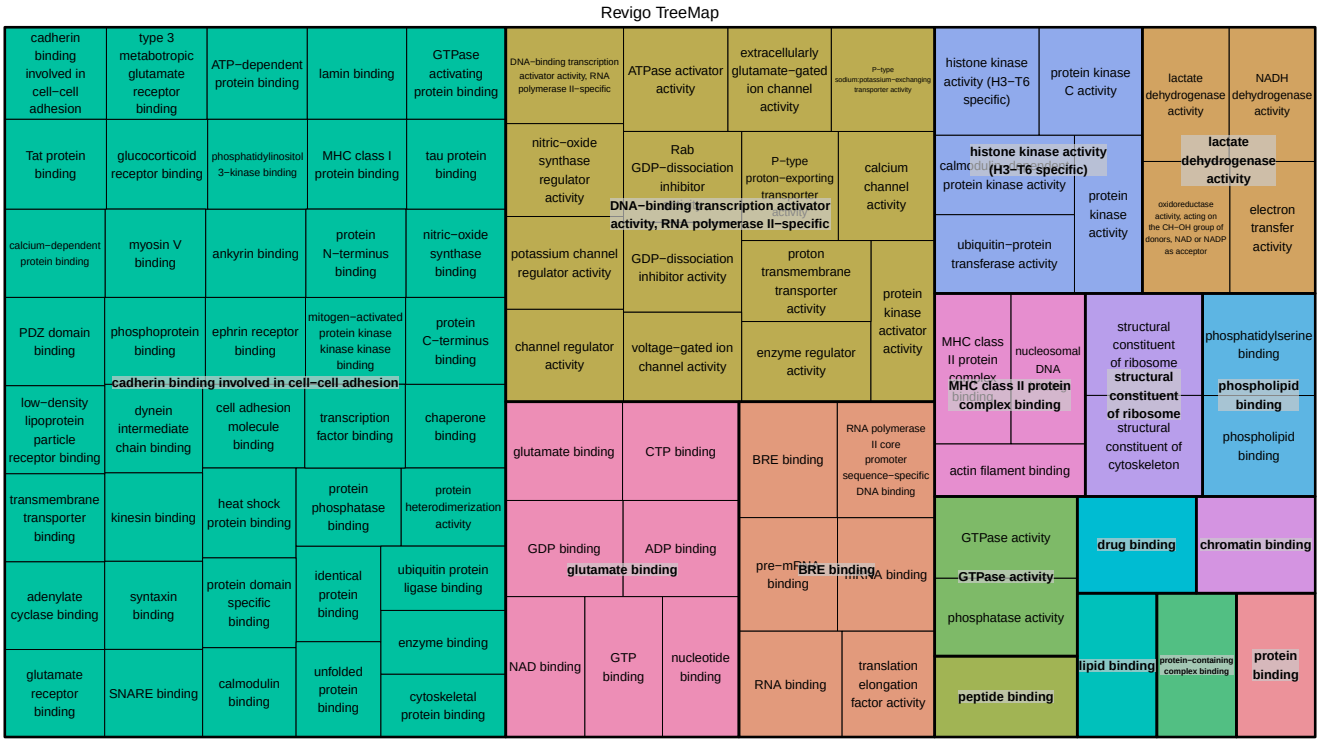

Fig. S13. The mouse olfactory bulb data: Revigo tree map for the GO annotations based on genes detected by TT-M.

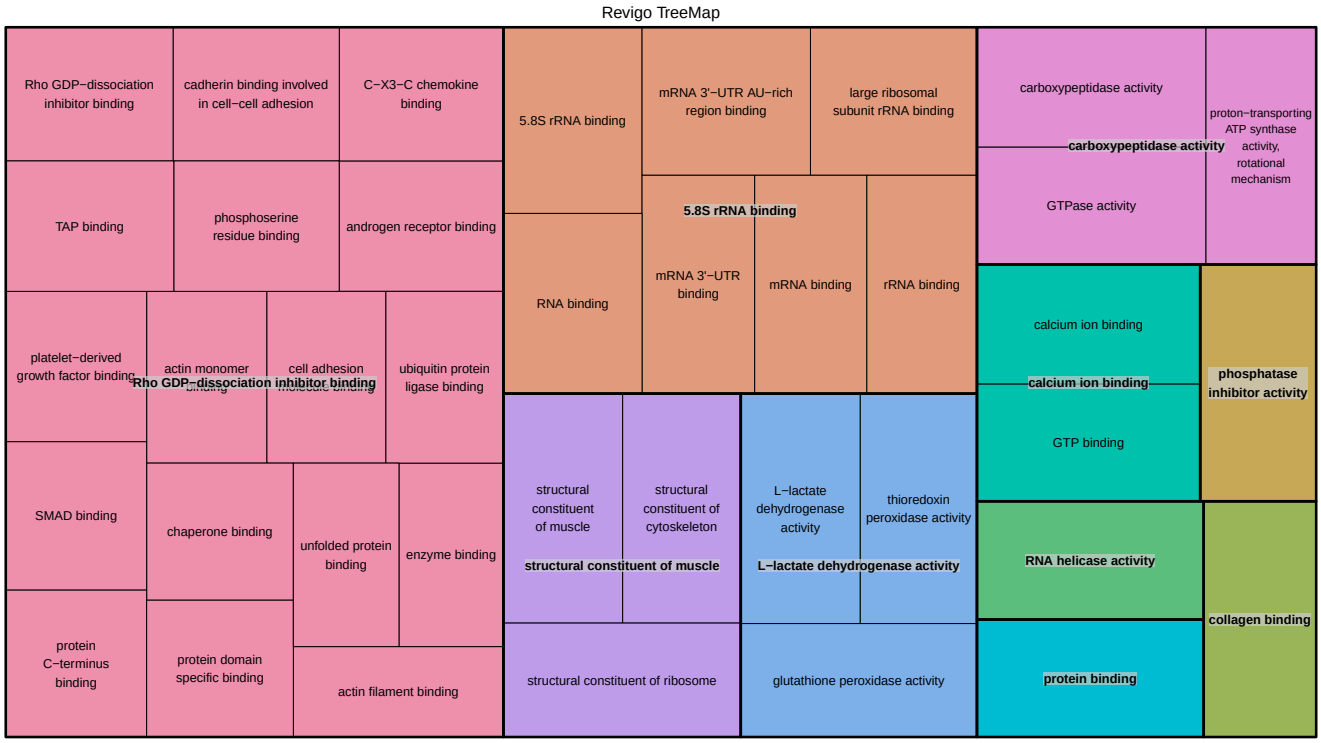

Fig. S14. The breast cancer data: Revigo tree map for the GO annotations based on genes detected by TT-S.

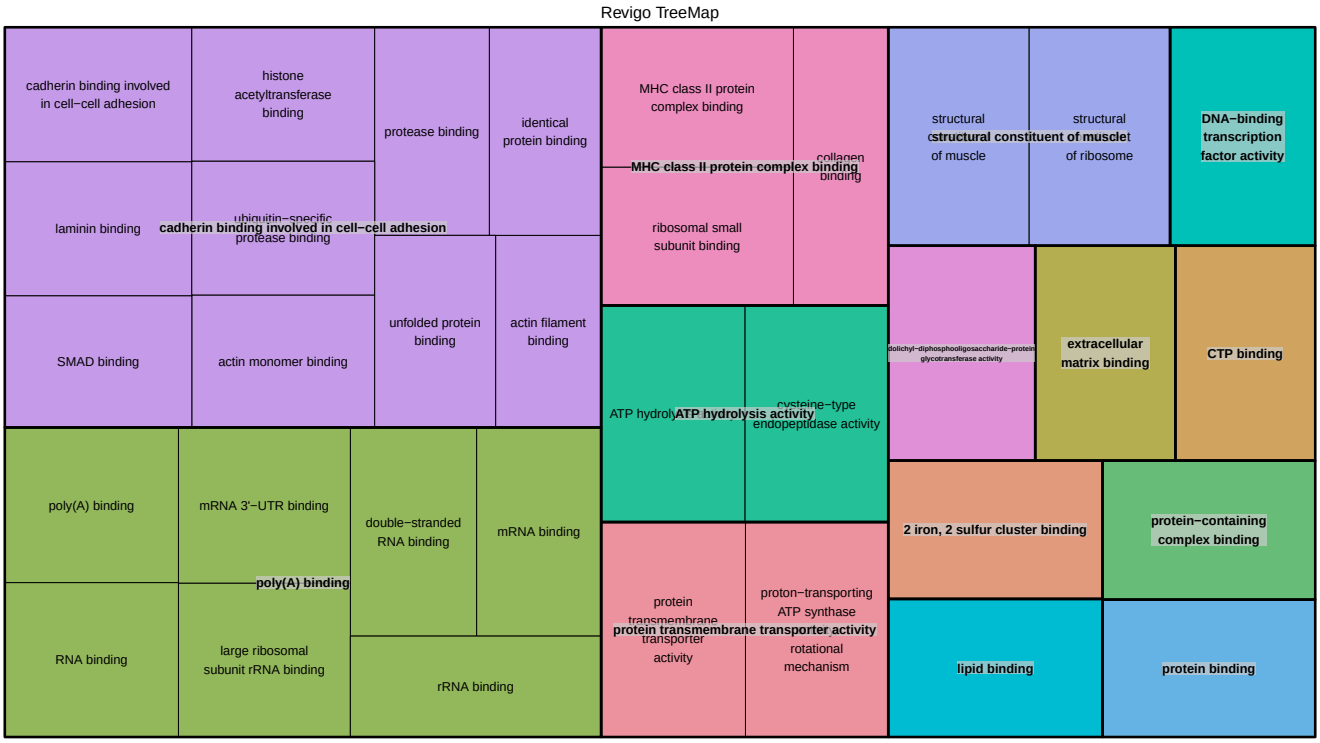

Fig. S15. The breast cancer data: Revigo tree map for the GO annotations based on genes detected by TT-M.

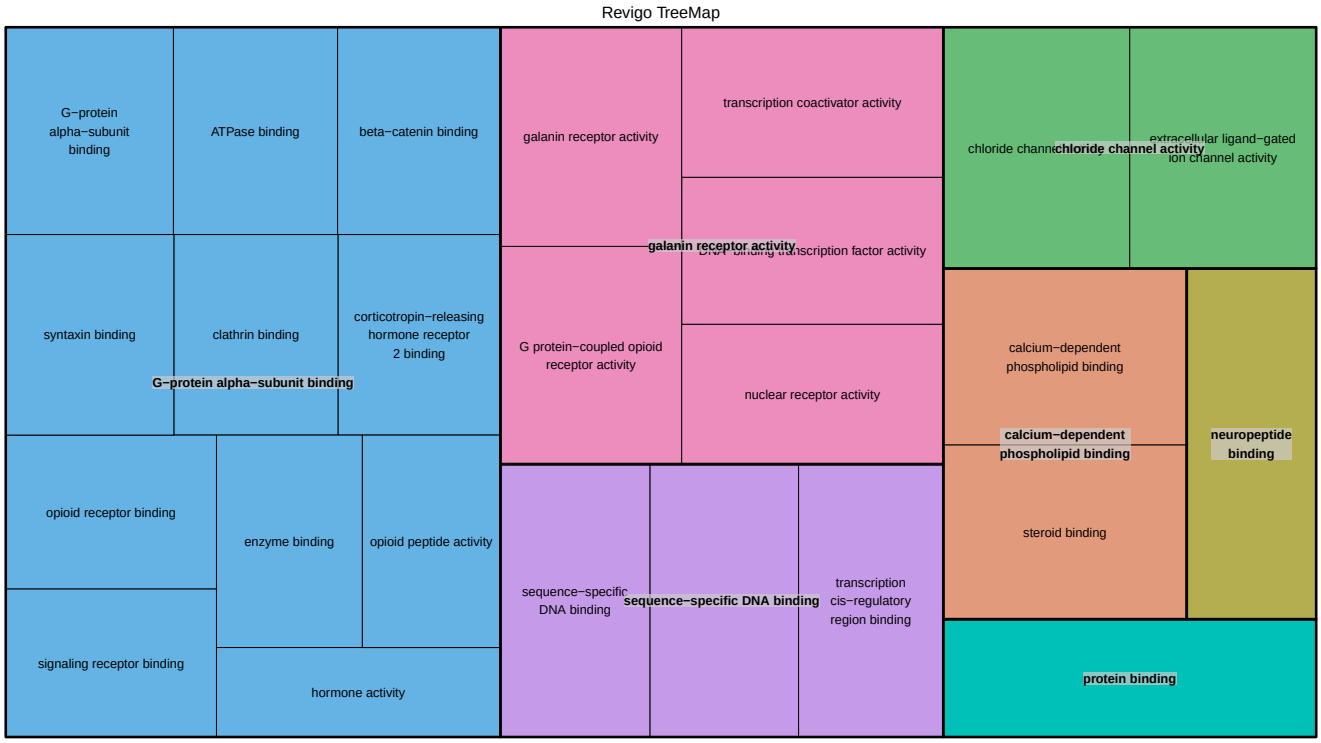

Fig. S16. The mouse hypothalamus data: Revigo tree map for the GO annotations based on genes detected by TT-S.

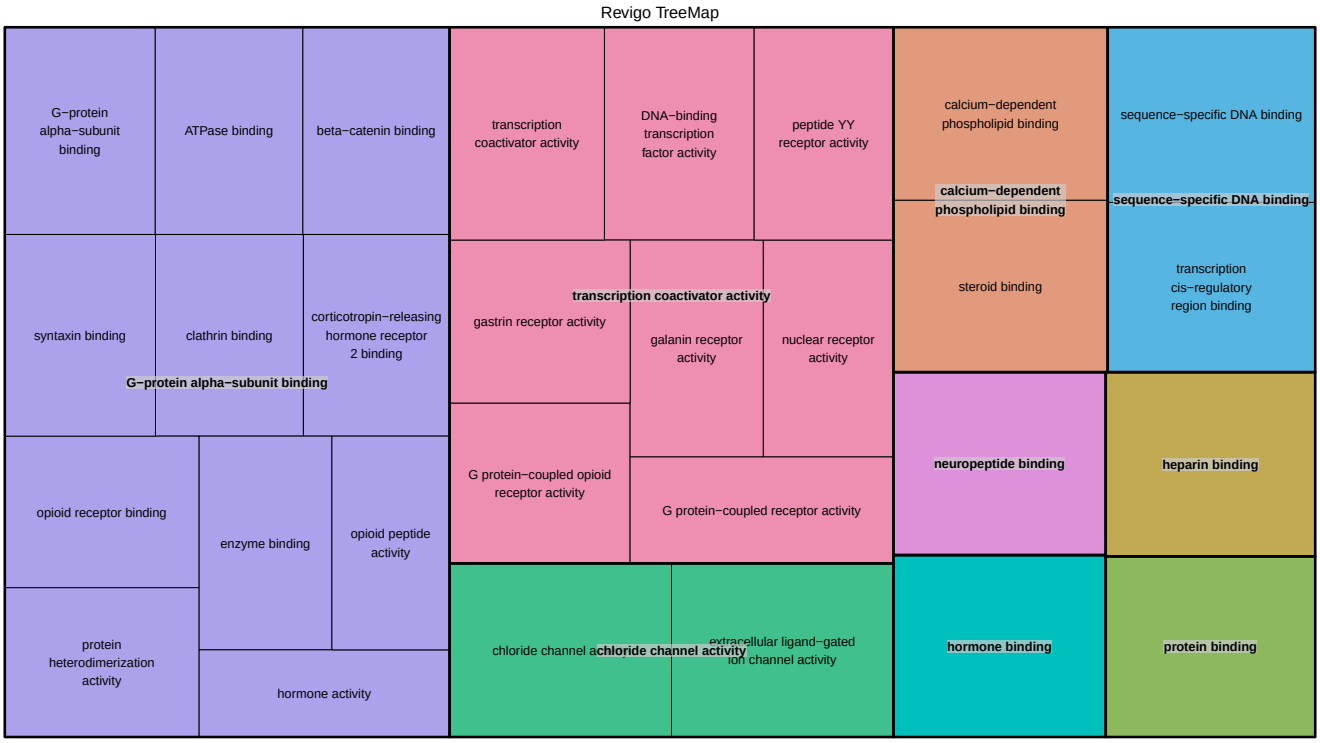

Fig. S17. The mouse hypothalamus data: Revigo tree map for the GO annotations based on genes detected by TT-M.

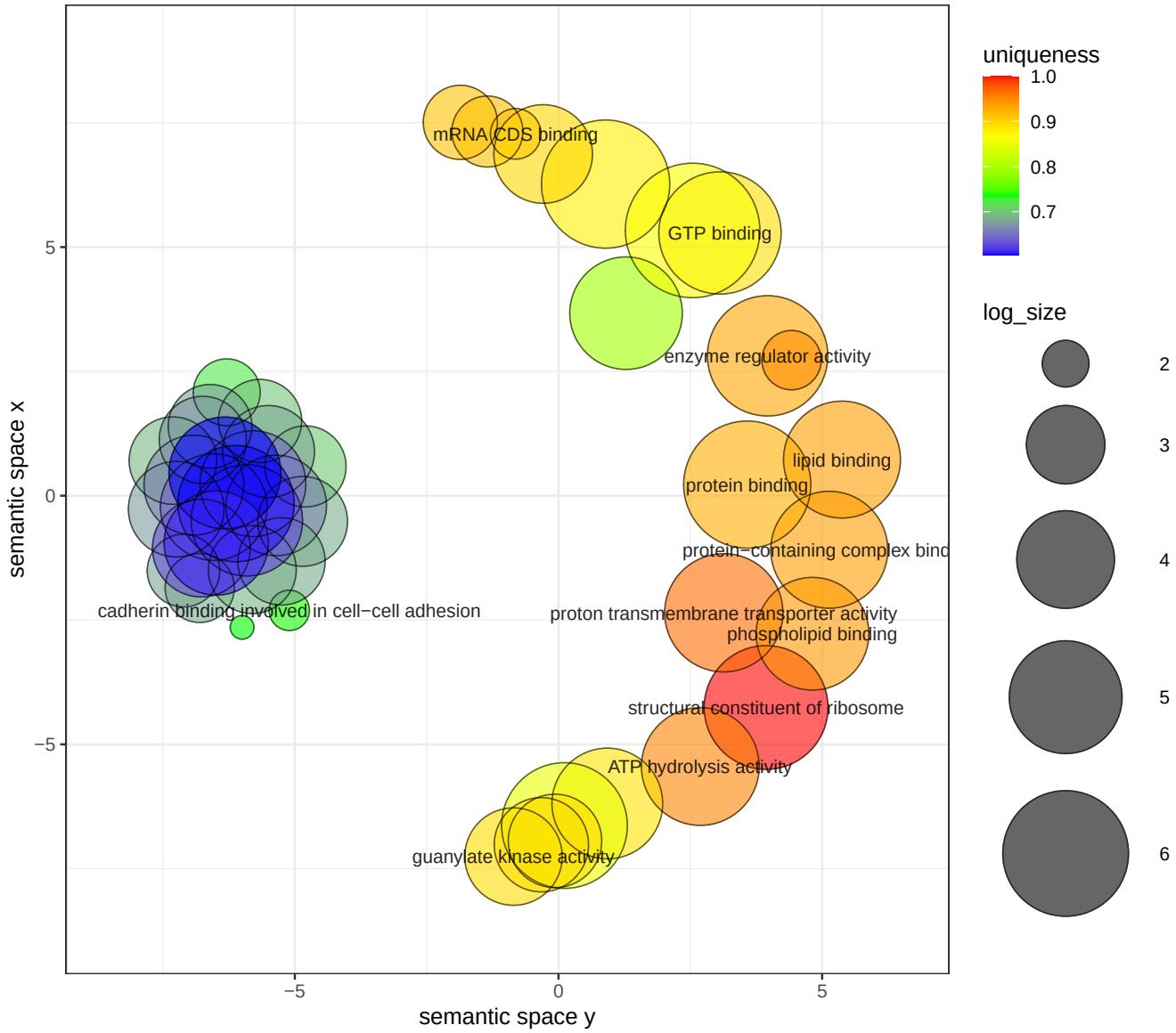

Fig. S18. The mouse olfactory bulb data: clustering of GO annotations for the genes detected by TT-S.

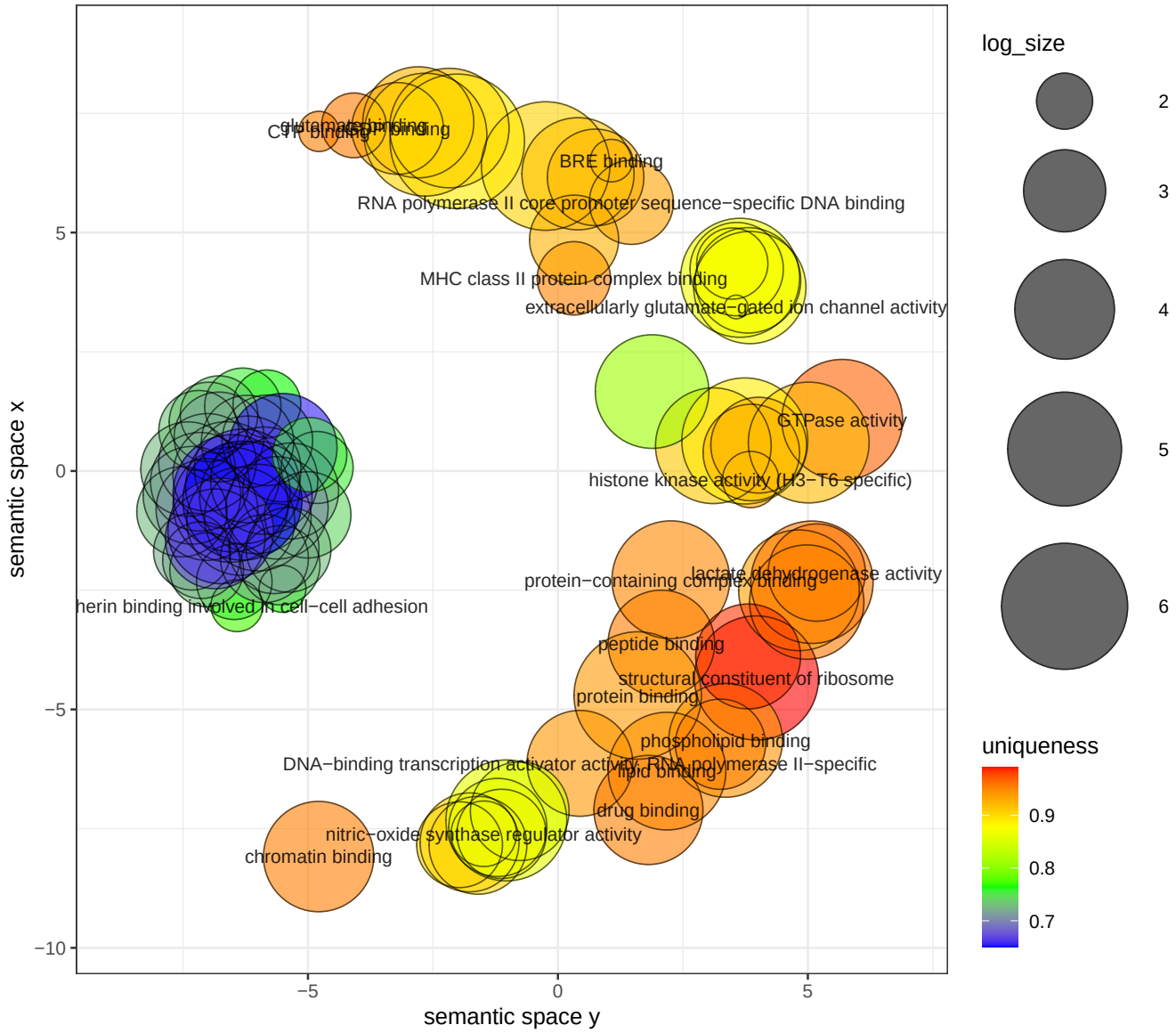

Fig. S19. The mouse olfactory bulb data: clustering of GO annotations for the genes detected by TT-M.

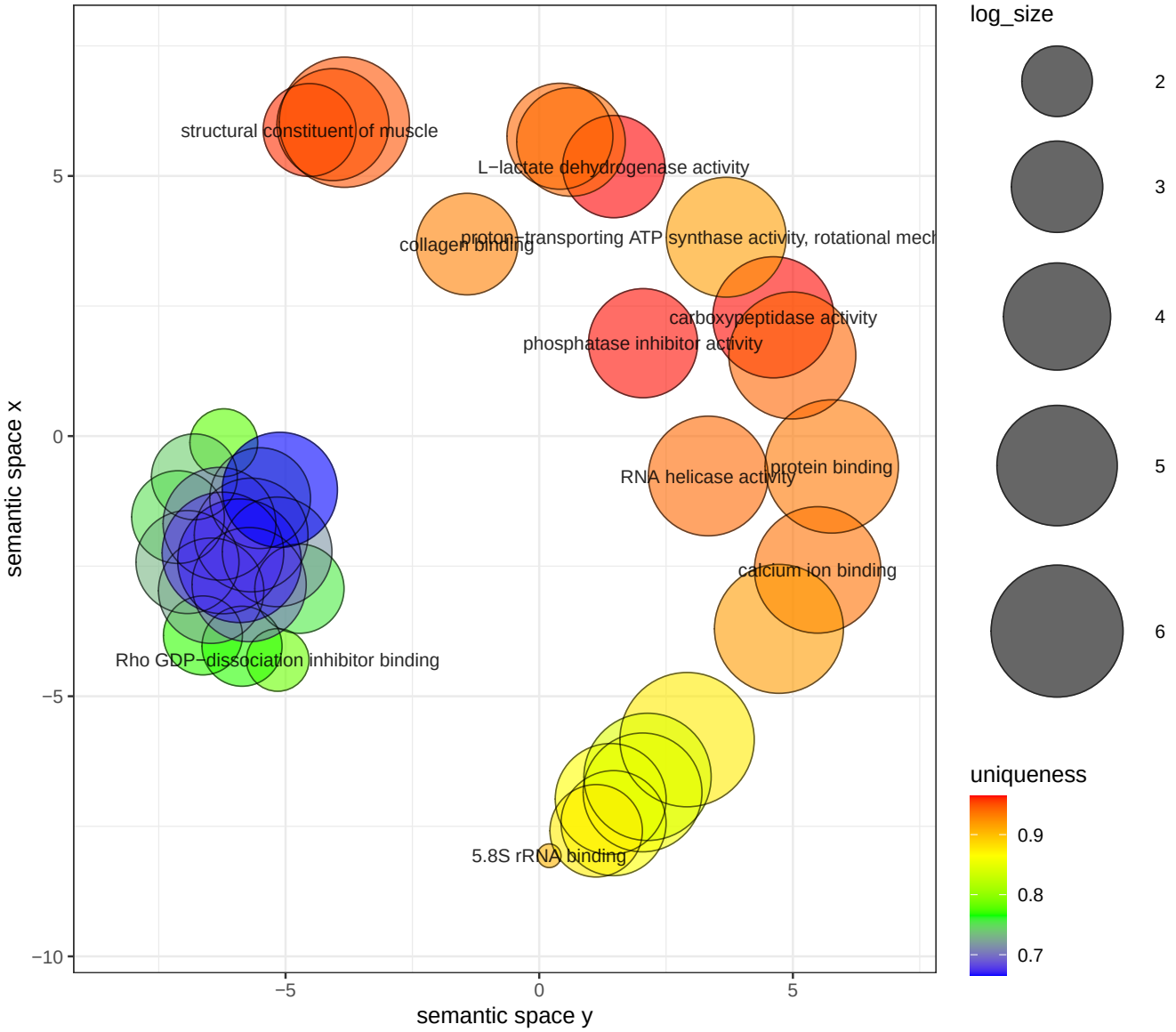

Fig. S20. The breast cancer bulb data: clustering of GO annotations for the genes detected by TT-S.

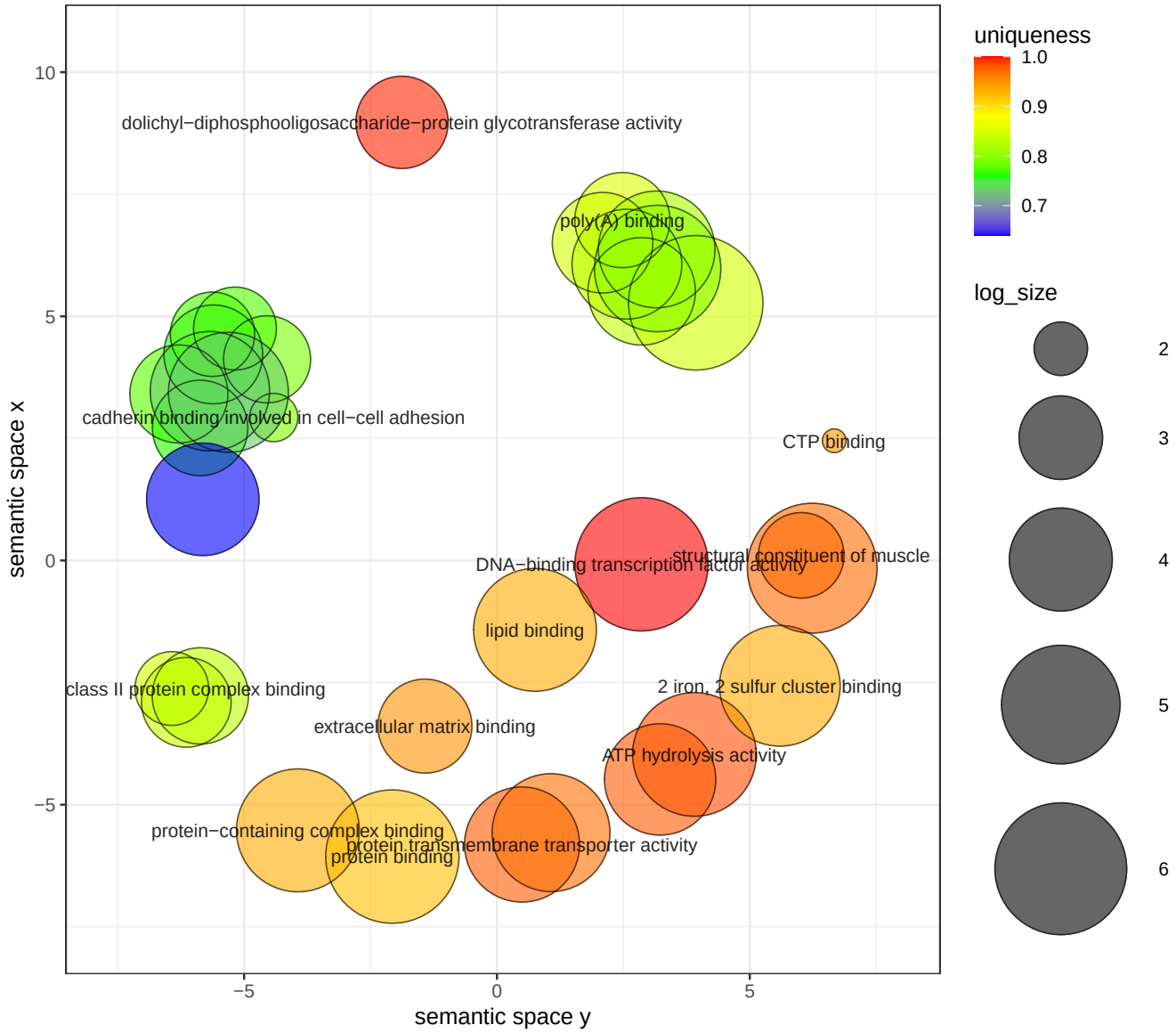

Fig. S21. The breast cancer data: clustering of GO annotations for the genes detected by TT-M.

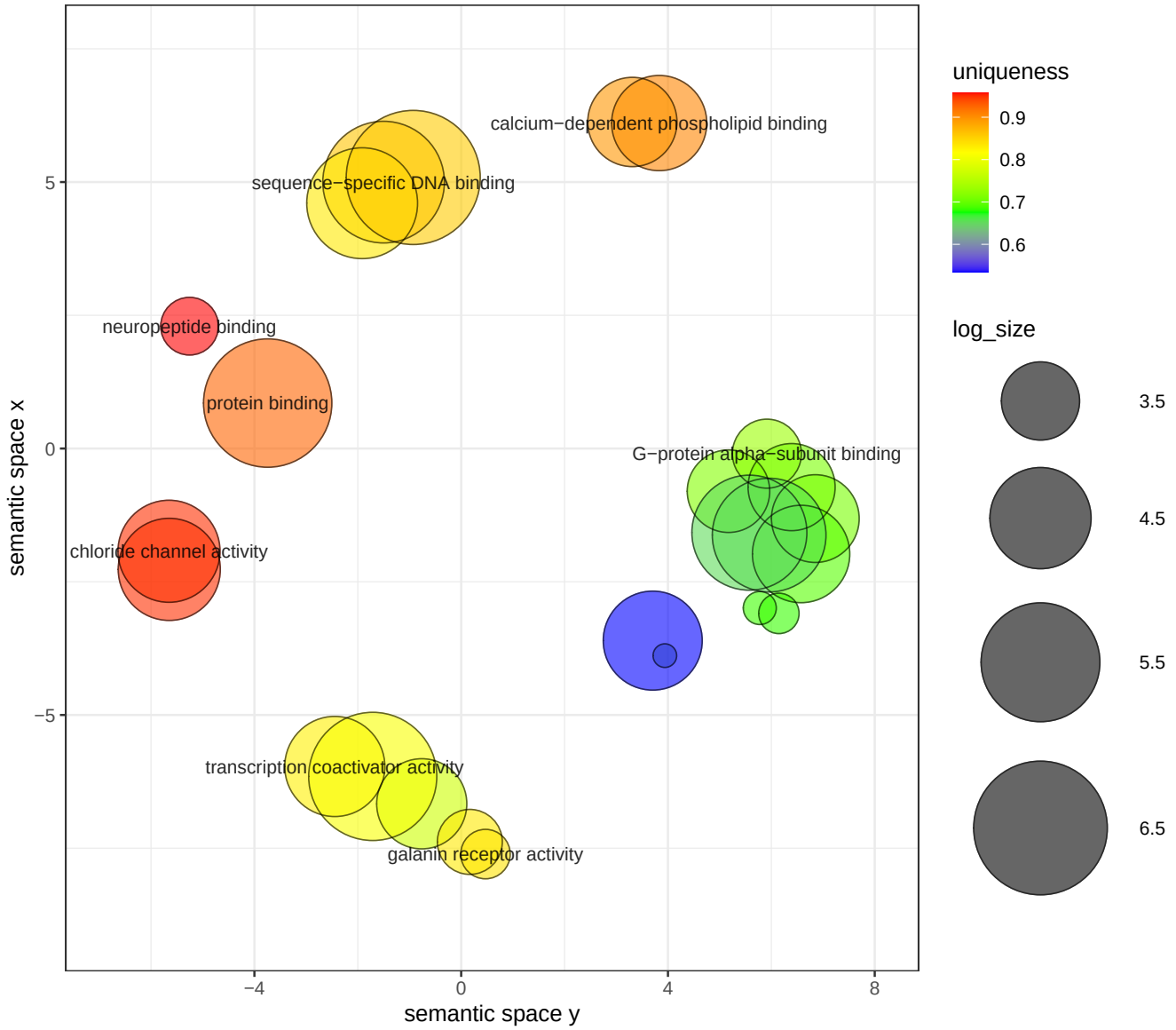

Fig. S22. The mouse hypothalamus data: clustering of GO annotations for the genes detected by TT-S.

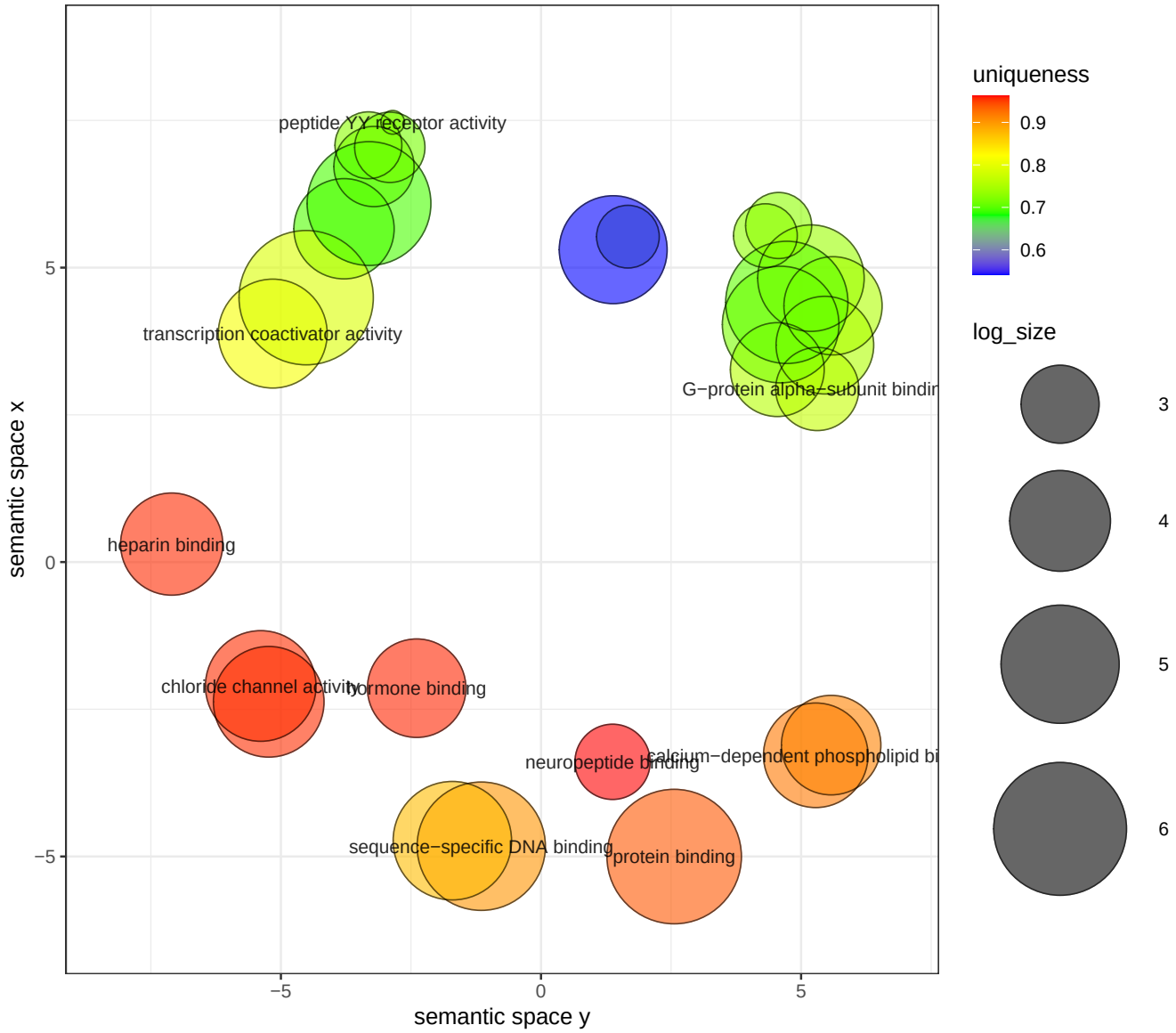

Fig. S23. The mouse hypothalamus data: clustering of GO annotations for the genes detected by TT-M.

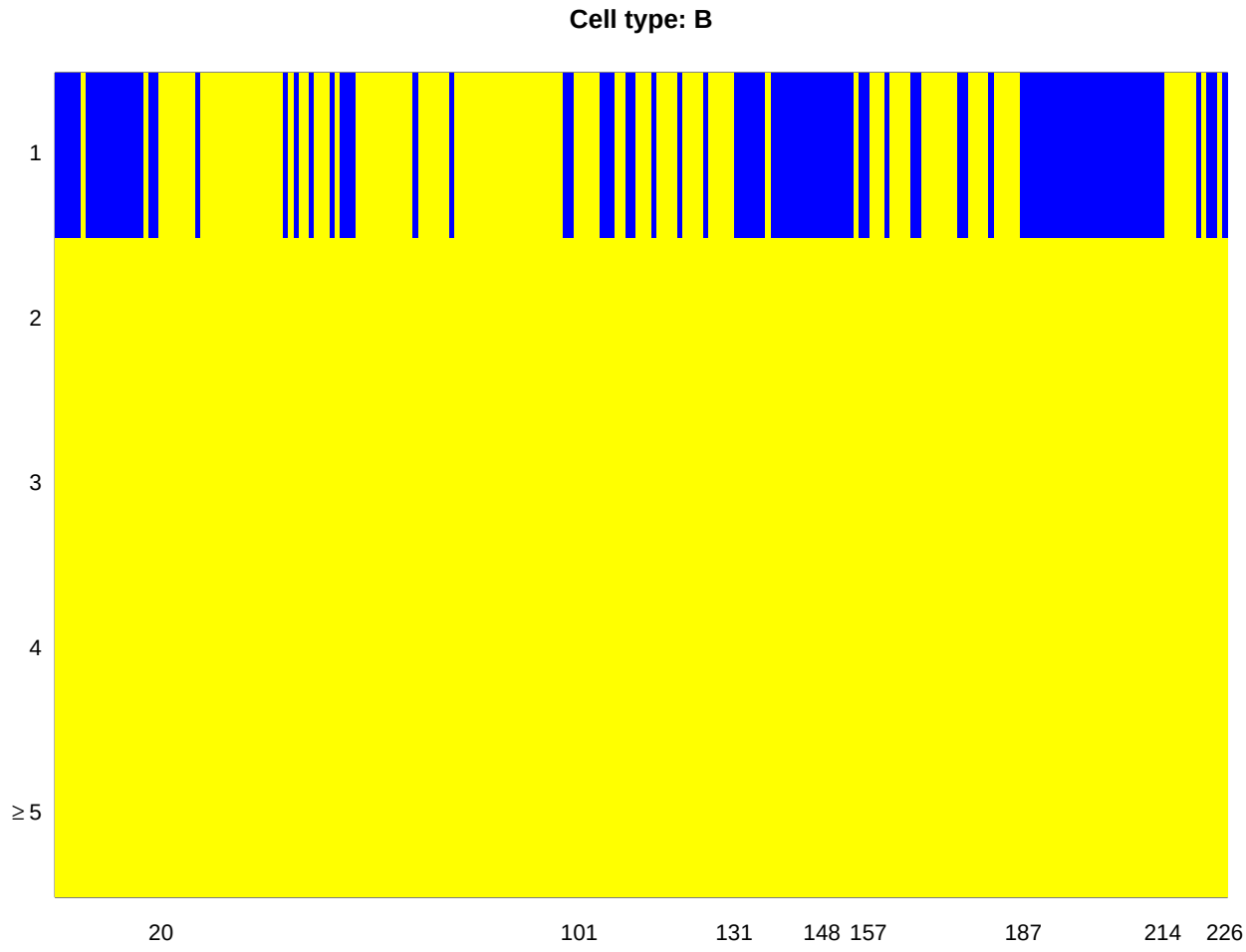

Fig. S24. The predictability test of every proteins for B cells.

T5.pdf

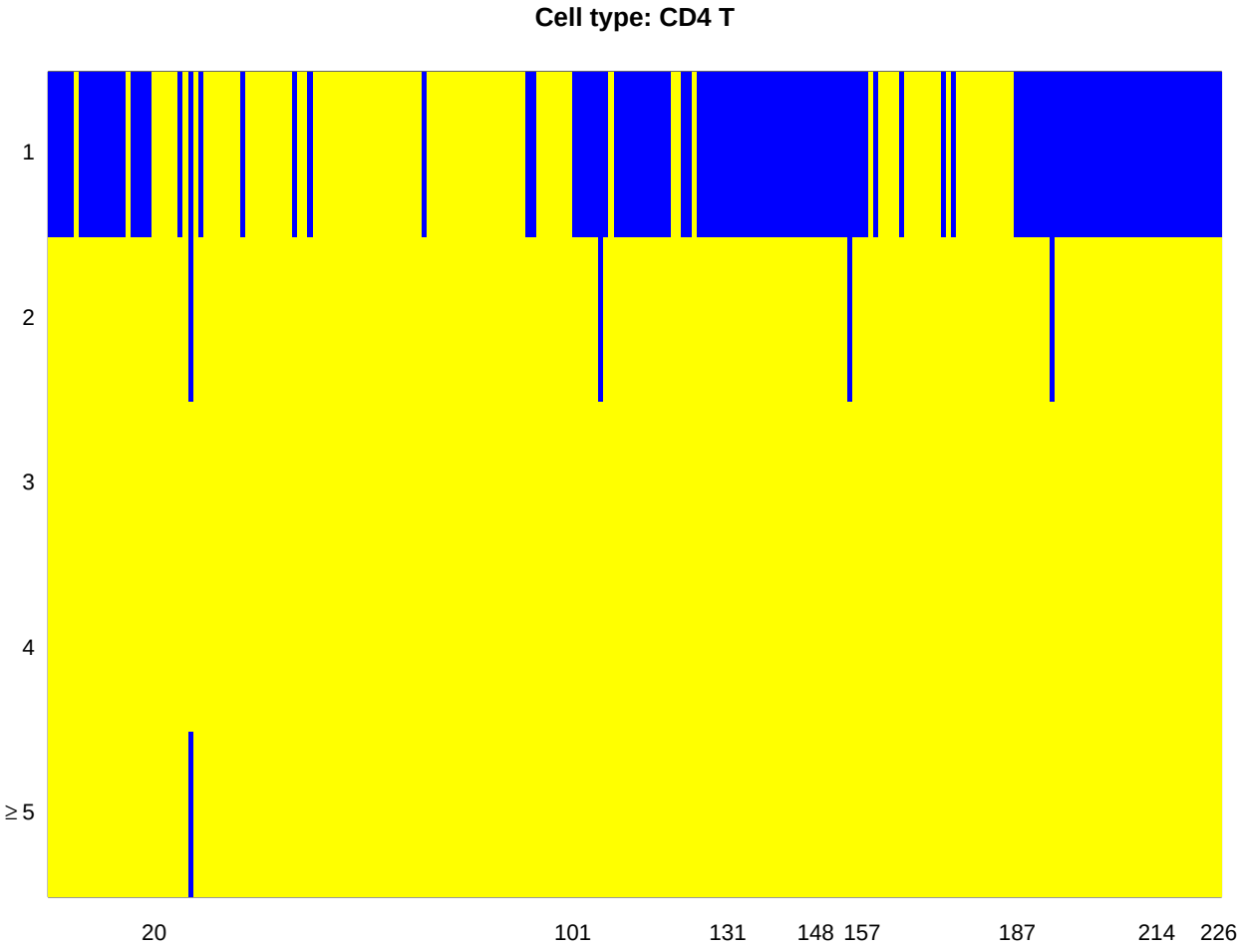

Fig. S25. The predictability test of every proteins for CD4 cells.

T5.pdf

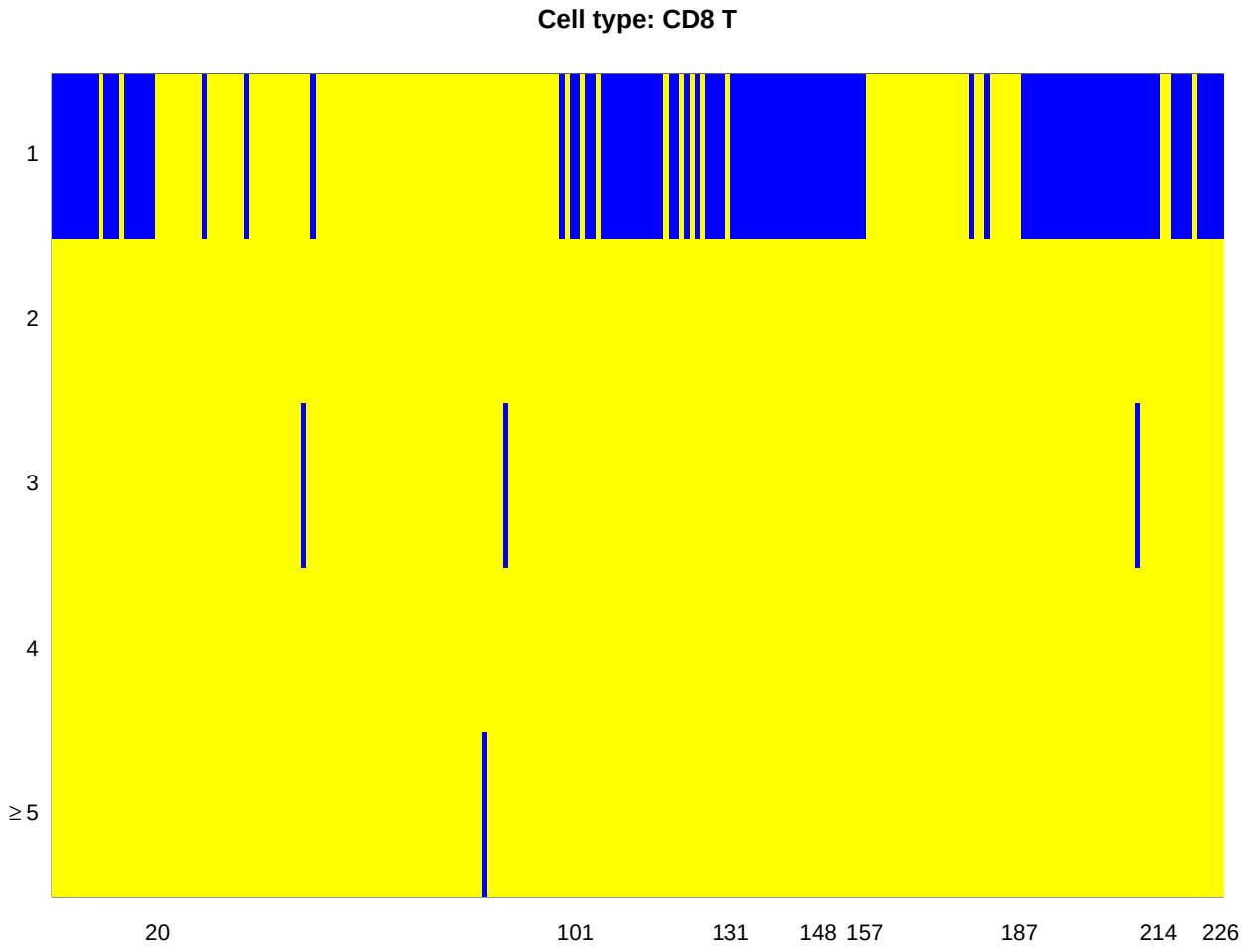

Fig. S26. The predictability test of every proteins for CD8 T cells.

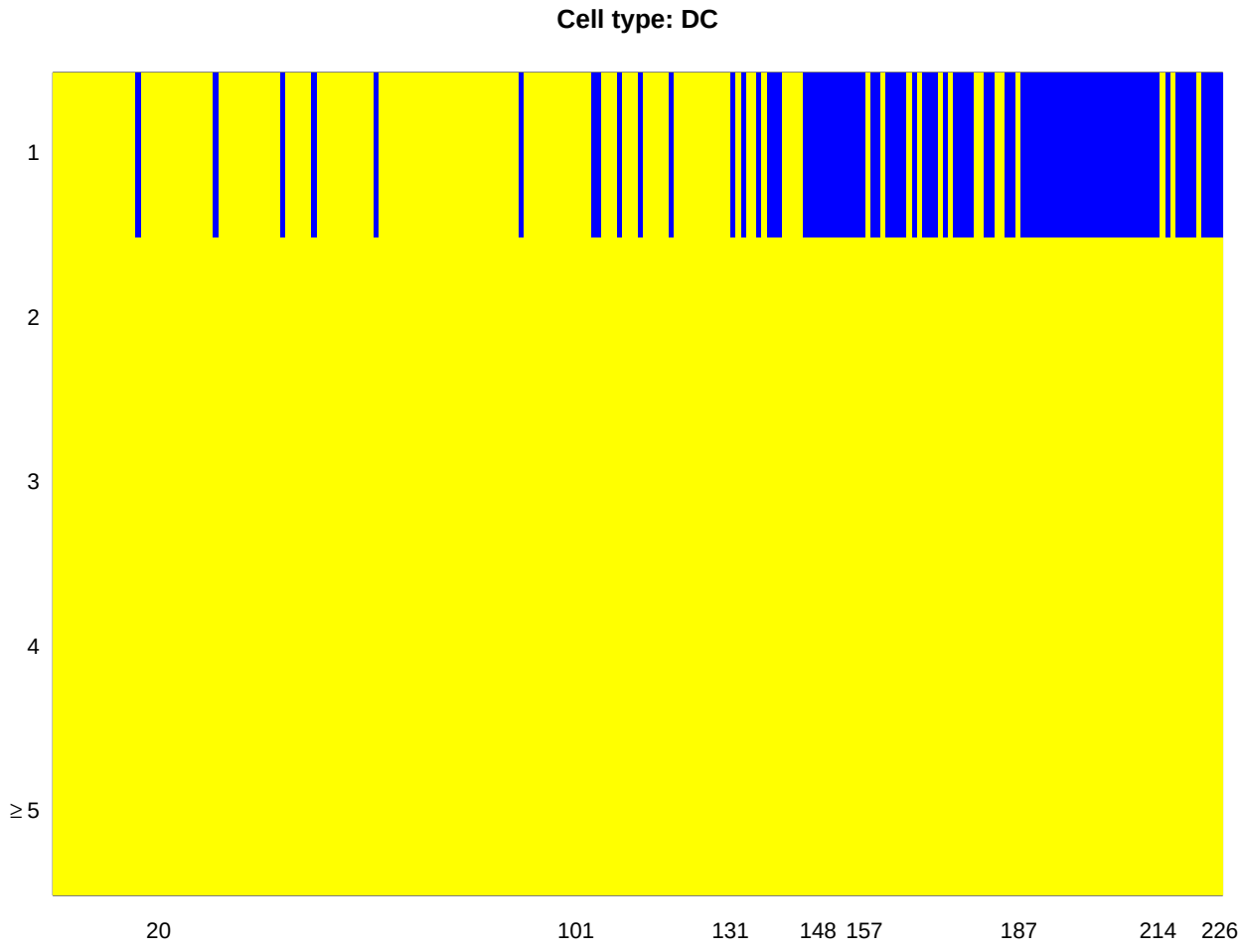

Fig. S27. The predictability test of every proteins for DC cells.

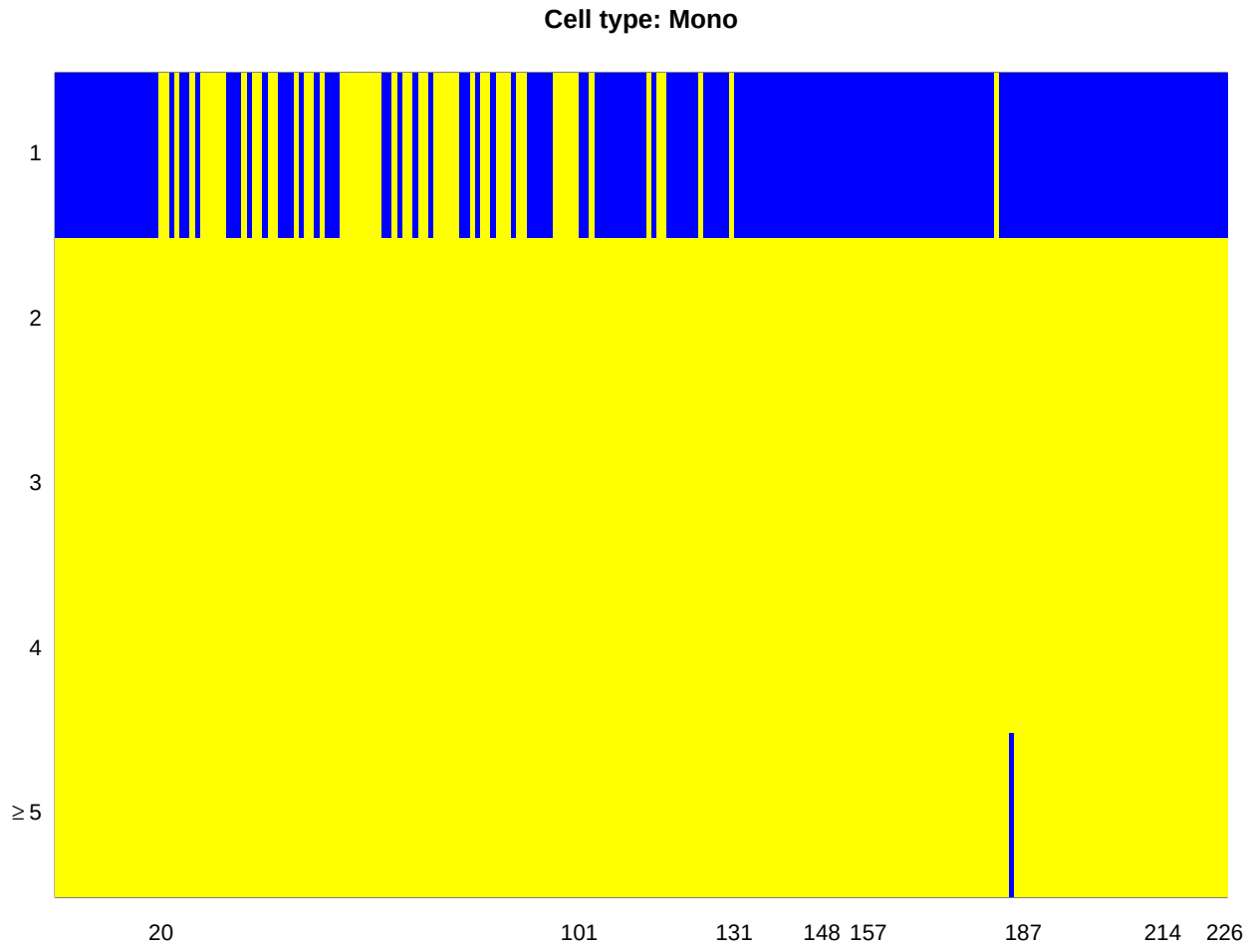

Fig. S28. The predictability test of every proteins for Mono cells.

T5.pdf

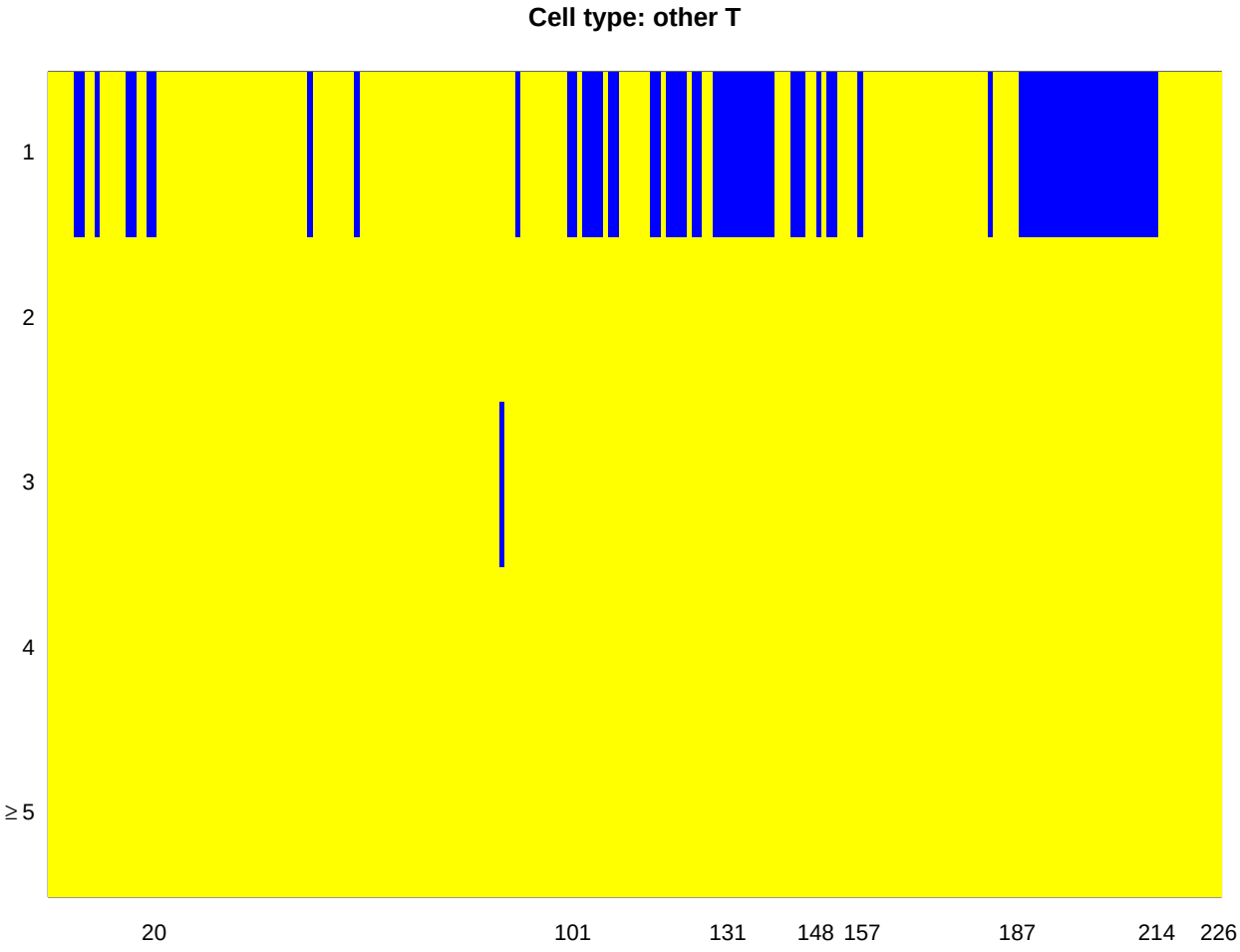

Fig. S29. The predictability test of every proteins for other T cells.

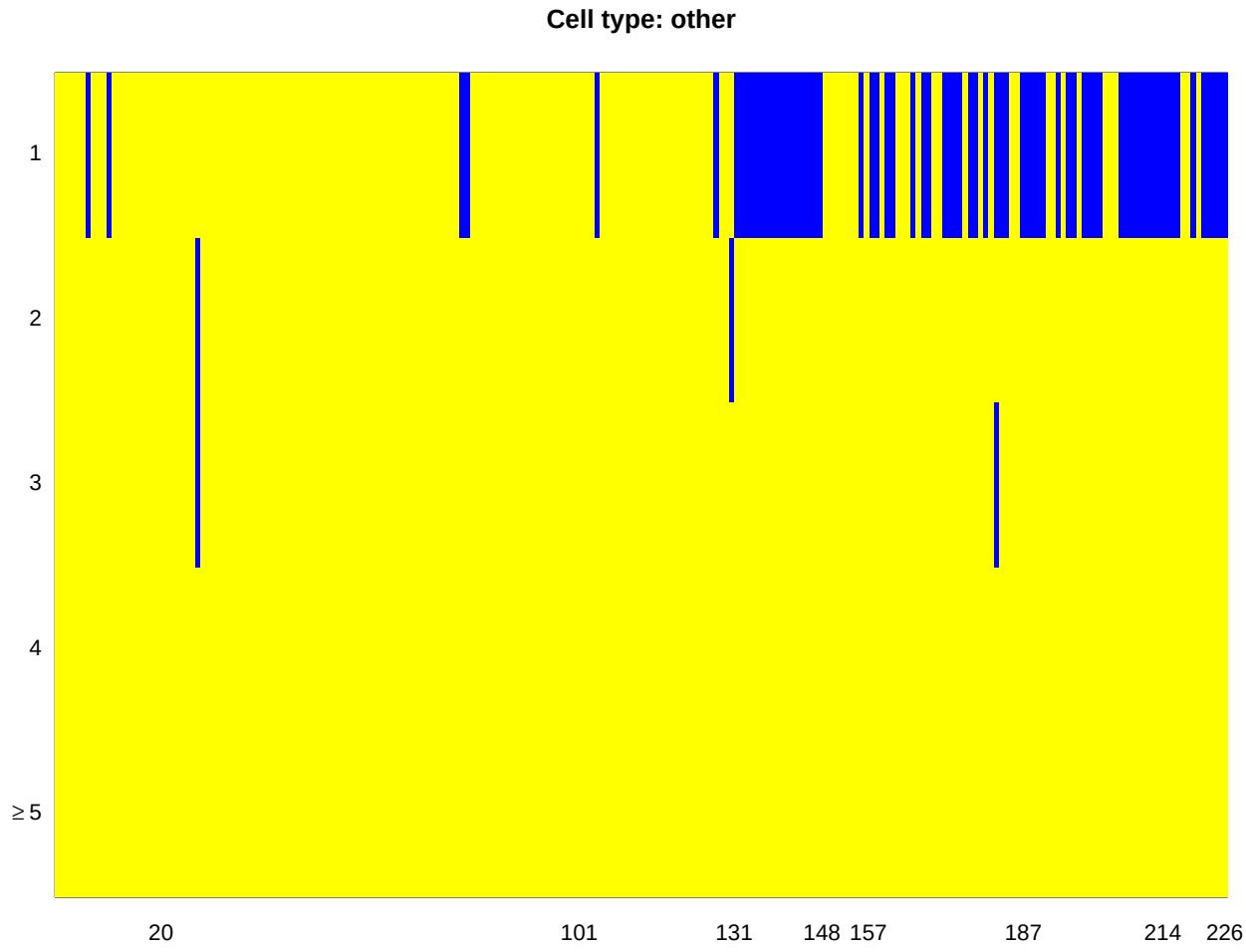

Fig. S30. The predictability test of every proteins for other cells.

Details of Datasets

All datasets used in this paper are publicly available. We list the web source for each dataset below:

CITE-seq Human PBMC data https://atlas.fredhutch.org/data/nygc/multimodal/pbmc_multimodal.h5seurat

Mouse olfactory bulb data https://www.spatialresearch.org/resources-published-datasets/doi-10-1126science-aaf2403/

Human breast cancer data https://www.spatialresearch.org/resources-published-datasets/doi-10-1126science-aaf2403/

Hypothalamus Data https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248

Hippocampus Data https://www.cell.com/cms/10.1016/j.neuron.2016.10.001/attachment/759be4dc-04a6-4a58-b6f6-9b52be2802db/mmc6.xlsx

References

1. JR Moffitt, et al., Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
2. S Shah, E Lubeck, W Zhou, L Cai, In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357 (2016).