

# <sup>2</sup> Supplementary Information for

- Model-free prediction test with application to genomics data
- <sup>4</sup> Zhanrui Cai, Jing Lei, Kathryn Roeder.
- 5 Kathryn Roeder

1

6 E-mail: roeder@andrew.cmu.edu

## 7 This PDF file includes:

- 8 Supplementary text
- 9 Figs. S1 to S30 (not allowed for Brief Reports)
- <sup>10</sup> Table S1 (not allowed for Brief Reports)
- 11 SI References

#### 12 Supporting Information Text

## 13 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  $\varepsilon$ . 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.

• 
$$Y = a \times 10X_1 + \varepsilon;$$

19 •  $Y = a \times 10 \sin(X_1) + \varepsilon;$ 

• 
$$Y = a \times 10 \exp(X_1) + \varepsilon;$$

• 
$$Y = a \times 100 \log(X_1 + 10) + \varepsilon;$$

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 22 1000 times and report the average results. We report the additional power curves for all the methods in Fig. S1 and S2, where 23 the type-I error rate is set as 0.01. Our methods perform very well under all settings. Moreover, the increasing dimension 24 does not have negative effects on the proposed methods. For MDC, the power decreases dramatically as the dimension of 25 X increases from 100 to 1000. As the dimension increases up to 5000, the RS-M still performs very well compared to other 26 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 27 dimension o the covariates is between 1000 and 5000. This also echos the real data analysis on the PBMC data where the high 28 29 dimensional noise may cause only a fraction of tests to lose power.

30 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 31 summarize the results in Fig. S4 where the significance level  $\alpha$  is set as 0.05. As expected, the type-I error of all methods 32 are well controlled. However, linear regression is only able to give non-trivial powers when the true model is linear. For the 33 other three models, linear regression fails to capture the nonlinear trends in the data while Xgboost successfully did. The 34 performance of the linear regression is also worse than the Xgboost under the linear model. Note that only the first column in 35 X is related to Y and all other variables are noises. One explanation is that Xgboost is able to handle the high dimensional 36 noisy variables in a more efficient way. 37

#### 38 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

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

<sup>43</sup> 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

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

49

52

Streak: We randomly choose a spot 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

$$Y_i = a \times \mathbb{I}(|X_{i,1} - X_1| \le h) + U_i$$

<sup>53</sup> We repeat the simulation 1000 times and report the average results. Besides the figures in the main paper, we also report <sup>54</sup> the power performance of all methods in Figure S5 where the type-I error  $\alpha$  is set to be 0.01.

#### 55 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 propered method is more likely to eatch the genes with more complicated structures.

<sup>59</sup> 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

<sup>62</sup> 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

<sup>67</sup> that the test could have a high false positive rate under the alternative hypothesis. Indeed, for the real data analysis, TT-S

<sup>68</sup> found 105 genes and TT-M found 115 genes, while SPARK found 143 genes and SpaDE found 124 genes. The genes detected

<sup>69</sup> by TT-S and TT-M all overlap with the findings of SPARK and SpaDE, which confirms the validity of the proposed methods. <sup>70</sup> The GO annotation reveals that the detected genes are not only related to protein bindings, but also other cell functions such

<sup>71</sup> as transcription coactivator activity.

<sup>72</sup> In terms of variable importance, %IncMSE equals 0.581 for the horizontal axis, 0.549 for the vertical axis, and is 0.665 for

<sup>73</sup> 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

<sup>75</sup> 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.

80 While all methods find less than 20 spatially variable genes for this dataset, we suspect the main reason that our method

 $_{81}$  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

<sup>84</sup> 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.

					Revigo IreeMap					
cadherin binding involved in cell-cell adhesior	cadherin binding type 3 metabotropic involved in glutamate receptor ell-cell adhesion binding		otein binding	syntaxin–1 binding	mRNA CDS binding	N6-methyladenosine-containing RNA binding	nitric–oxide synthas regulator activity	e	GTP binding	
phosphatidylinositol 3-kinase binding	ephrin receptor binding	nitric-oxide synthase binding	transmembrane transporter binding	adenylate cyclase binding	pre-mRNA intronimRNA	CDS binding	enzyme regulator activ	rity	nucleo	binding
PDZ domain binding	cadherin bindir	ng involved in cell-	cell adhesion cell adhesion molecule bindir	protein domain specific binding	5S rRNA binding	translation elongation factor activity	Structural prot constituent tra of ribosome tra		proton transmembrane transporter activity	
calcium-dependent	protein C-terminus		protein	ubiquitin protein	ATP hydrolysis activity	calmodulin-dependent protein kinase activity	phospholipid binding	protein-cont:	ining	protein binding
protein binding	binding		heterodimerizati activity	ligase binding	ATP hyo	droly575 activity phosphatase activity		complex bin	ling	proton phianing
protein N–terminus binding	kinesin binding	protein phosphatase binding	enzyme bindin	g microtubule binding	calcium-dependent protein serine/threonine kinase activity	protein kinase activity	lipid binding	guanyl	ate kir	ase activity

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

Revigo TreeMap															
cadherin binding involved in cell-cell adhesion	type 3 metabotropic glutamate receptor binding	ATP-dependent protein binding	lamin bindin	GTPase activating protein binding	DNA-binding transcr activator activity, R polymerase II-sper	ription INA ATPa: cific a	se activator ( activity	extracellularly glutamate-gated ion channel activity	P-type sodium;pctassium-exchanging transporter activity	histone kina activity (H3- specific)	rse -T6 prote C	in kinase activity	ase lactate N/ Y dehydrogenase dehydr activity ac		
Tat protein binding	glucocorticoid receptor binding	phosphatidylinositol 3-kinase binding	I MHC class protein bindir	l tau protein ng binding	nitric-oxide synthase regulator activity	itric-oxide synthase regulator activity DNA-binding transcription DNA-binding transcription DNA-binding transcription activity		histone kinase activity calmodu <mark>(H3⊈T6 specific)</mark> protein kinase activity protein		tivity c) protein	oxidoreducta	lactate ydrogenase activity			
calcium-dependent protein binding	myosin V binding	ankyrin binding	protein N-terminus binding	nitric-oxide synthase binding	potassium chai regulator activ	activity, RNA polymerase II-specific ubique standard stan		ubiquitin- transferase	i-protein se activity		activity, acting the CH–OH gro sonors, NAD or as accepto	ting on group of O or NADP eptor eptor			
PDZ domain binding	phosphoprotein binding	ephrin receptor binding	mitogen-activa protein kinase kinase kinase binding	ed protein C-terminus binding	channel regulator voltag activity chan		ge-gated ion	enzyme regula activity	ator	MHC class Il protein	nucleosoma DNA	struc consti of ribo	structural onstituent f ribosome bindir structural		
low-density lipoprotein particle receptor binding	-cadherin bindir dynein intermediate chain binding	ig involved in ce cell adhesion molecule binding	transcriptior factor bindin	chaperone g binding	glutamate binding CTP b		IP binding	BRE binding	RNA polymerase II core	MHC class complex	II protein binding	of ribo struc	tuent some tural uent of	phospholipid binding phospholipid binding	
transmembrane transporter	kinesin binding	heat shock	protein phosphatase binding	protein heterodimerization activity					sequence-specific DNA binding	actin filame	ent binding	cytosk	eleton		
binding	austavia	protein domain	identical	ubiquitin protein ligase binding	GDP binding	GDP binding ADP binding glutamate binding		pre-mFBRE binding binding		GTPase activity GTPase activity		drug binding cl		chromatin binding	
cyclase binding	binding	specific binding	binding	enzyme binding		GTP	nucleotide			phosphatas	e activity			protein	
glutamate receptor binding	SNARE binding	calmodulin binding	unfolded protein binding	cytoskeletal protein binding	NAD binding	binding	binding	RNA binding	translation elongation factor activity	peptide b	binding	lipid bindin	g complex b	<sup>nding</sup> binding	

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

					Rev	ligo Treeiviap					
Rho GDP-dissociatic inhibitor binding	ssociation cadherin binding involved in cell-cell adhesion		9 C-X:	3–C chemokine binding	5.8S rRNA binding	mRNA 3'-UTF region bir	₹ AU-rich Iding	laı subu	rge ribosomal nit rRNA binding	carboxypeptidase activity carboxypeptidase acti	proton-transporting ATP synthase vity activity, rotational
TAP binding	phosp residu	noserine binding	androge	en receptor binding		5.8S rRN.	NA binding			GTPase activity	mechanism
platelet-derived growth factor binding <mark>RI</mark>	actin monomer to GDP <sub>JTI</sub> dissocia	cell ac	hesion binding	ubiquitin protein ligase binding	RNA binding	mRNA 3'-UTR binding	mRNA	mRNA binding rRNA binding		calcium ion binding calcium ion binding	phosphatase inhibitor activity
					structural	structural	L-lact	ate		GTP binding	
SMAD binding	chaperone bin	ling unfo	lded protein binding	enzyme binding	constituent of muscle structural consti	constituent of cytoskeleton tuent of muscle	dehydrogenase activity L-lactate dehydr		peroxidase activity	RNA helicase activity	
protein C-terminus binding	protein doma specific bindi	in ng	actin filan	nent binding	structural constitu	ent of ribosome	glutathione pe		ixidase activity	protein binding	collagen binding

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

Revigo TreeMap										
cadherin binding involved in cell-cell adhesion	histone acetyltransferase binding	protease binding	identical protein binding	MHC class II protein complex binding ——MHC class II protein complex binding		structural struc cstructural constituent of mu of muscle of ribr		ural <b>scie</b> t some	DNA-binding transcription factor activity	
laminin binding <b>cadi</b>	laminin binding cadherin binding involved in c profesase binding			ribosomal small subunit binding	Diruing					
		_								
SMAD binding	actin monomer binding	unfolded protein binding	actin filament binding	ATP hydroly <b>ATP hydrol</b> y	cvsteine-type vsis activity endopeptidase activity	Iolicityi-diphosphooligosaccharide-prot giycotransferasa activity	extracellular matrix binding		CTP binding	
poly(A) binding	mRNA 3'-UTR binding	double-stranded RNA binding	mRNA binding		proton_transporting	2 iron, 2 sulfur cluste	ifur cluster binding		tein-containing mplex binding	
RNA binding	large ribosomal subunit rRNA binding	rRNA bi	rRNA binding		ne transporter activity rotational mechanism	lipid binding		pr	otein binding	

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

			Revigo Tre	eeMap			
G–protein alpha–subunit binding	ATPase binding	beta-catenin binding	galanin receptor activity	transcription coactivator activity		chloride channichloride channel a	tracellular ligand-gated activity ion channel activity
syntaxin binding G–j	clathrin binding protein alpha-subunit bindi	corticotropin-releasing hormone receptor 2 binding	gala G protein-coupled opioid receptor activity	galanin receptor activity protein-coupled opioid receptor activity nuclear receptor activity		calcium-dependent phospholipid binding calcium-dependent	neuropeptide
opioid receptor binding	g enzyme binding	opioid peptide activity	sequence-specific DNA binding	ce-specific DNA bind	transcription ing cis-regulatory region binding	phospholipid binding steroid binding	binding
signaling receptor bindi	hg	ione activity				protein binding	

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

Revigo TreeMap																								
G-protein alpha-subunit binding	ATPase binding		ATPase binding		ATPase binding		ATPase binding		ATPase binding		ATPase binding		ATPase binding		ATPase binding		ATPase binding		beta-catenin binding	transcription coactivator activity	DNA-binding transcription factor activity	peptide YY receptor activity	calcium-dependent phospholipid binding calcium-dependent phospholipid binding	sequence-specific DNA binding -sequence-specific DNA binding-
syntaxin binding G–pr	syntaxin binding clathrin binding G-protein alpha-subunit bi		corticotropin-releasing hormone receptor 2 binding ding	transc gastrin receptor activity	galanin receptor activity	ty nuclear receptor activity	steroid binding	transcription cis-regulatory region binding																
opioid receptor bindii	opioid receptor binding enzyme bindi		g opioid peptide activity	G protein-coupled opioic receptor activity	G protein-coupled receptor activity		neuropeptide binding	heparin binding																
protein heterodimerization activity		horm	none activity	- chloride channel a <b>ch</b>	Noride channel activity	ular ligand–gated hannel activity	hormone binding	protein binding																

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.



Cell type: B

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



T5.pdf

Cell type: CD4 T

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



Cell type: CD8 T



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



Cell type: DC

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



Cell type: Mono

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



Cell type: other T



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



Cell type: other

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

## 85 Details of Datasets

- <sup>86</sup> All datasets used in this paper are publicly available. We list the web source for each dataset below:
- 87 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/
- <sup>90</sup> 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

### 93 References

- 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
- <sup>97</sup> mouse hippocampus. Neuron **92**, 342–357 (2016).