Supplementary Materials for

Joint reconstruction of *cis*-regulatory interaction networks across multiple tissues using single-cell chromatin accessibility data

Kangning Dong[1,2] and Shihua Zhang[1,2,3*]

[1]NCMIS, CEMS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;
[2]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China;
[3]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China.

*To whom correspondence should be addressed. Tel/Fax: +86 01 82541360; Email: zsh@amss.ac.cn.

Outline

Supplementary Methods

Stability analysis of grouping cell process

Supplementary Figures

Supplementary Tables

Stability analysis of grouping cell process

Two random factors in cell grouping process (Step 1 of JRIM) could affect the results. (1) The stopping criterion. JRIM stops the grouping process until no new group can be created, which means no new group could overlap any existing one by less than a percentage of cells (80% by default). Obviously, this stopping criterion is directly related to the number of groups where lower stop criterion will lead to fewer number of groups (Figure S15A). (2) The randomness of dimensionality reduction algorithm. JRIM constructs a *k*-nearest neighbor graph based on the t-SNE embedding. However, t-SNE has a certain randomness.

In order to analyze the stability of cell grouping process, we first run JRIM using all the cells from 13 tissues in chromosome 1 with different stopping criteria (70%, 80% and 90%) (Figure S15A). Note that the number of peaks in each 500kb local genomic window is generally less than 200, while the number of groups in cerebellum (tissue with fewest cells) is 241 for 70% and 459 for 80%. Thus, in principle, the results of JRIM are stable because the sample covariance matrix is nonsingular. We set the inferred interactions of 80% as reference and calculated the ROC curves. As expected, the AUC score is high (>0.85 between 80% and 90%, >0.78 between 80% and 70%) in tissues with different number of cells (Figure S15B, S15C) and the rate of common interactions between results of 80% and results of 90% is about 87% (Figure S15D), suggesting that the inferred networks of JRIM is robust in terms of the stopping criterion.

We then set the stopping criterion as 80% and run JRIM with different random seeds which affecting the t-SNE results in cell grouping process. As shown in Figure S15B and S15C, the impacts of randomness in t-SNE is greater than the impacts of different stopping criteria (AUC=0.73 in cerebellum and AUC=0.66 in heart). And the rate of common interactions between results of different random seeds is about 71% (Figure S15D). Although we think this degree of difference is acceptable because the number of candidate pairs of peaks is huge, JRIM still needs further optimization to alleviate its impact. One possibility is to adopt a more robust dimensionality reduction method or integrate prior knowledge such as the protein labeling of cells when perform dimensionality reduction. Another approach we more recommended is to run JRIM with different random seeds and voted for the final results. An example of common interactions between results using three different random seeds are shown in Figure S15E. The rate of common interactions in both three resulting networks is about 60%.

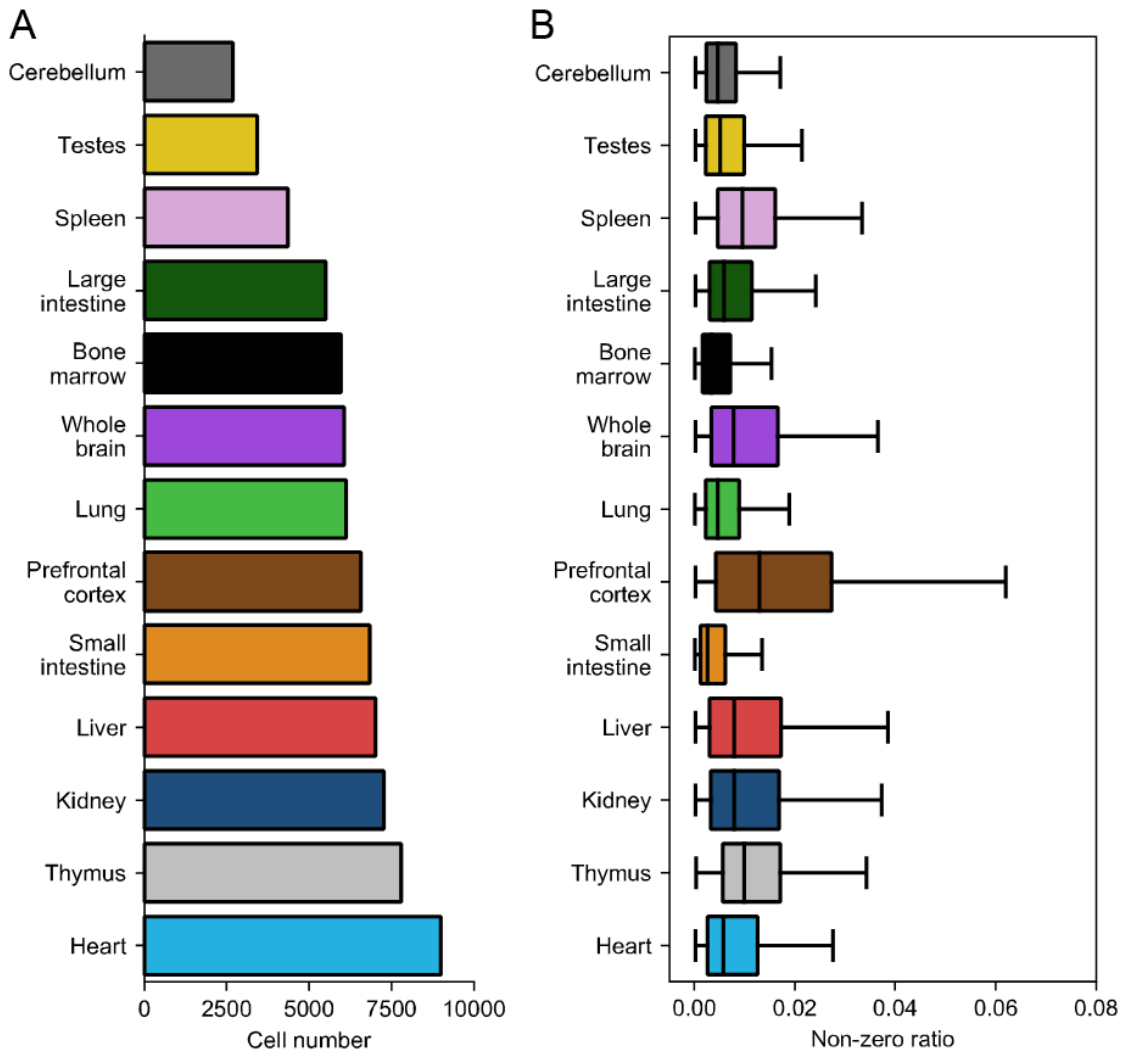

Figure S1. Cell number (A) and data sparsity (B) of single-cell ATAC-seq data for each tissue.

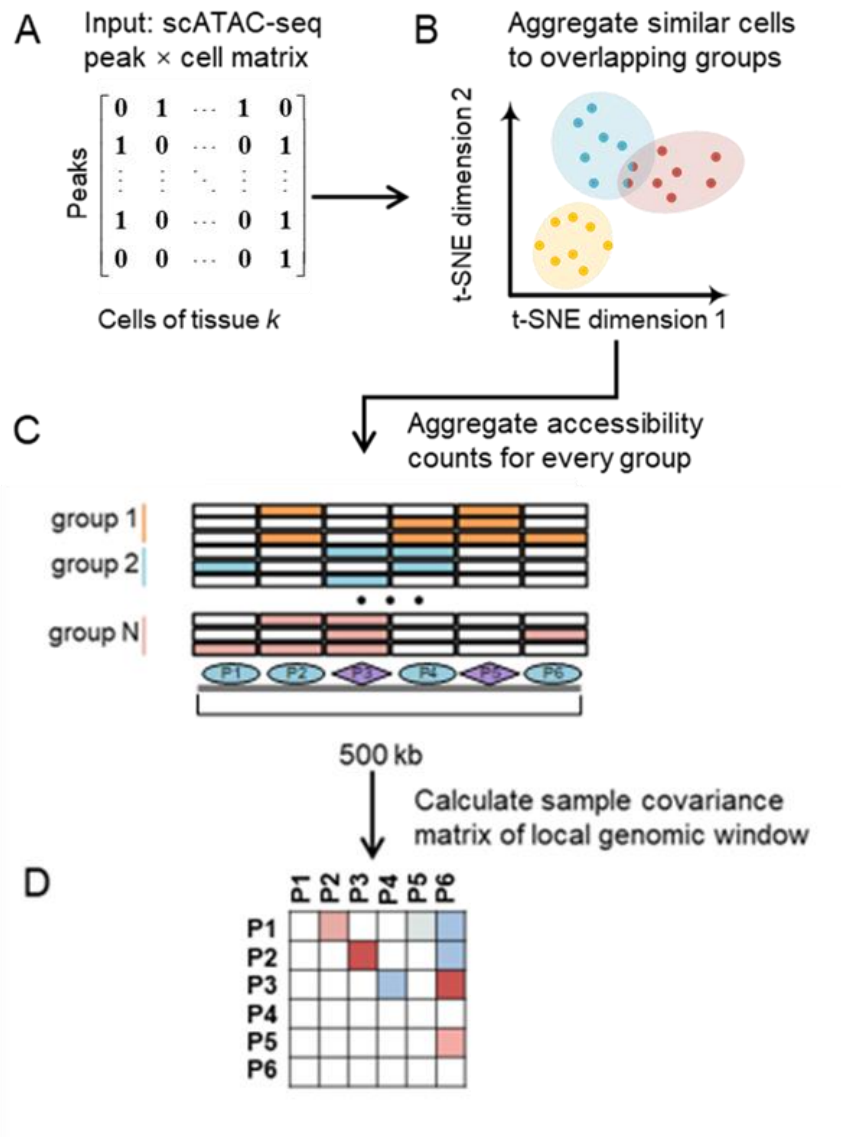

Figure S2. Illustration of the workflow of data processing for each tissue. (A) JRIM takes the peak × cell matrix of each tissue as input. (B) JRIM first maps single cells into low dimensional spaces using t-SNE and aggregates single cells into overlapping groups. (C) After clustering similar cells into overlapping groups, JRIM aggregates their accessibility counts to construct grouped matrix. The purple diamond represents peaks overlapped with one gene promoter and the ellipse represents the remaining ones. (D) Finally, JRIM calculates sample covariance matrix of grouped matrix for each local genomic window. The resulting sample covariance matrices are adopted to joint graphical lasso model to jointly reconstruct *cis*-regulatory interaction networks (Figure 1).

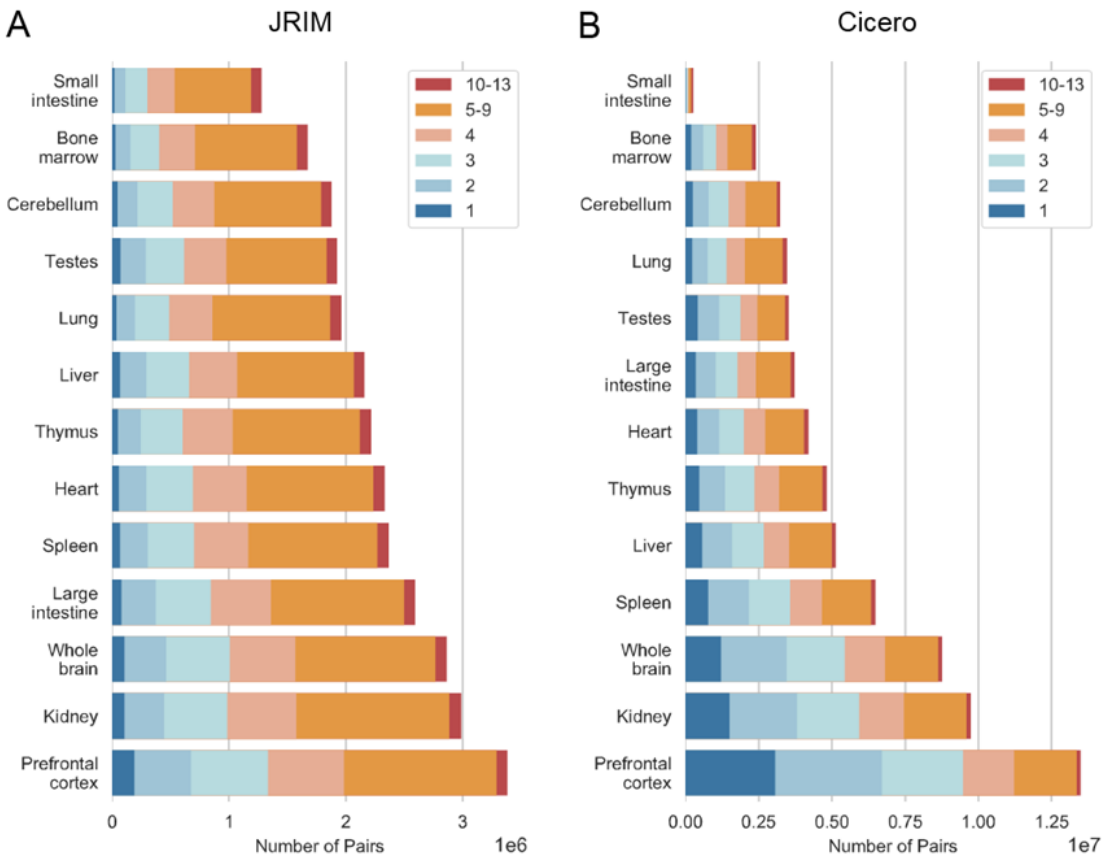

Figure S3. Number of interacting co-accessible DNA element pairs (i.e., *cis*-regulatory interactions) in each tissue estimated by JRIM (A), and by Cicero (using the same sparsity parameter λ=0.25 for each tissue respectively) (B).

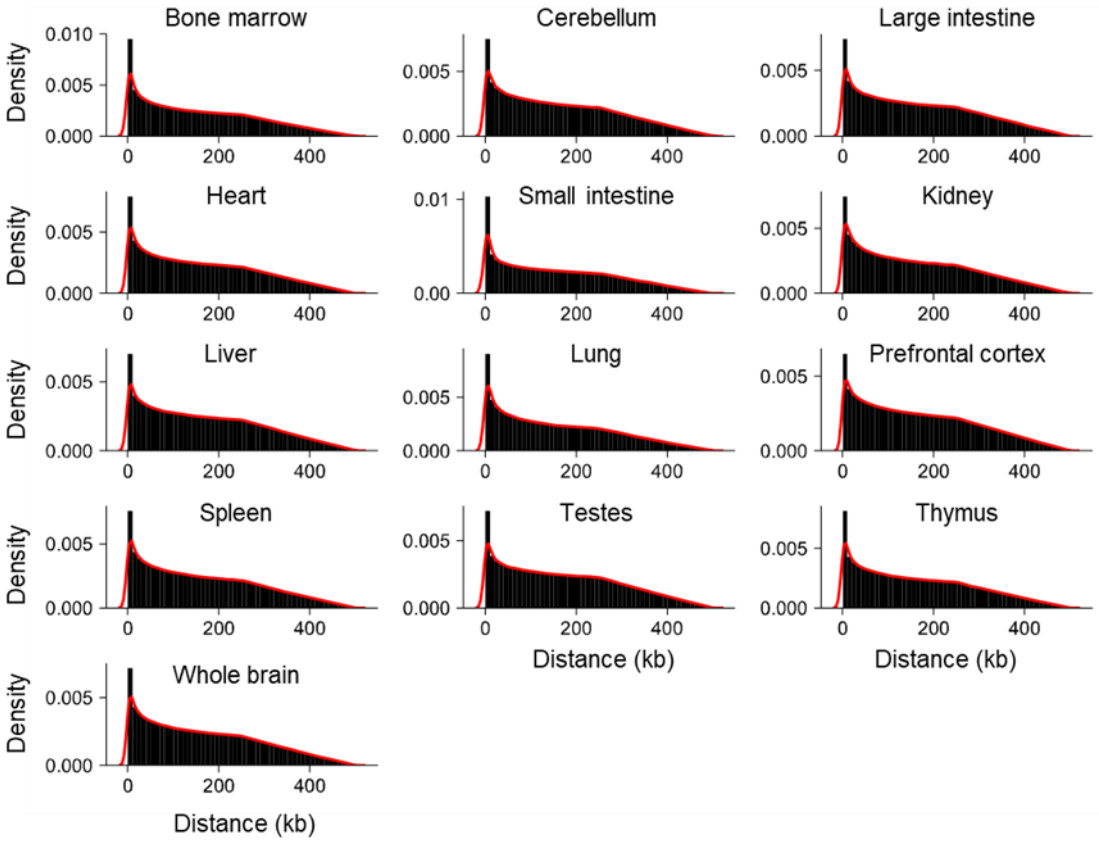

Figure S4. Distribution of distance between interacting co-accessible DNA elements. The red line is the estimated probability distribution function by python function `seaborn.Distplot()`.

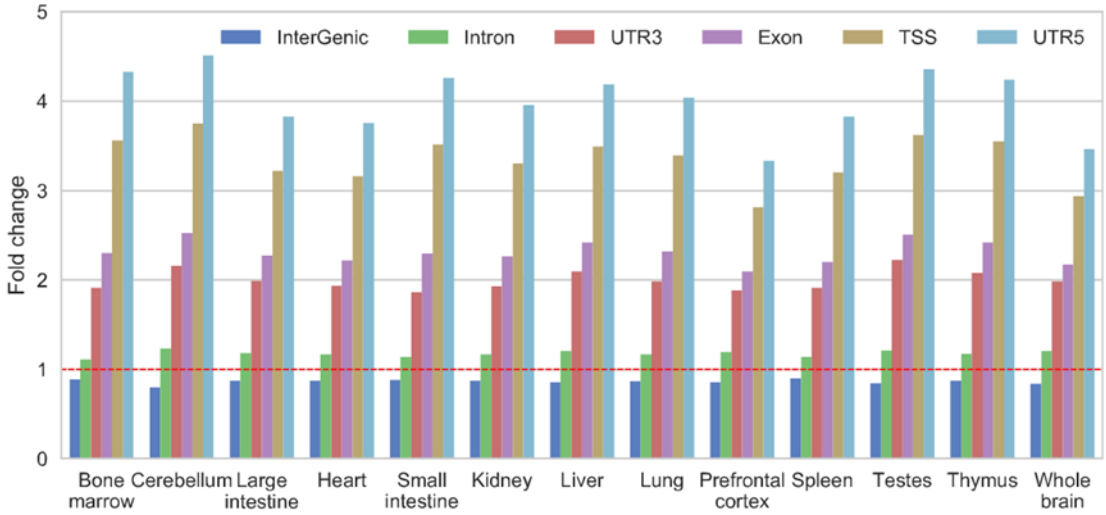

Figure S5. Enrichment degrees (fold changes) of the number of *cis*-regulatory interactions in different genomic regions.

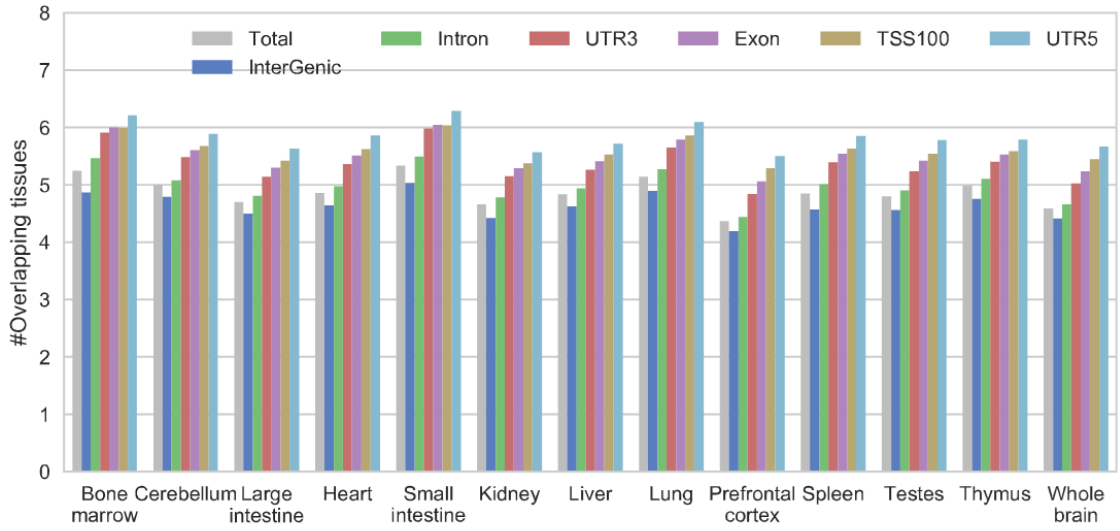

Figure S6. Mean number of occurring tissues of six types of genomic region related interactions.

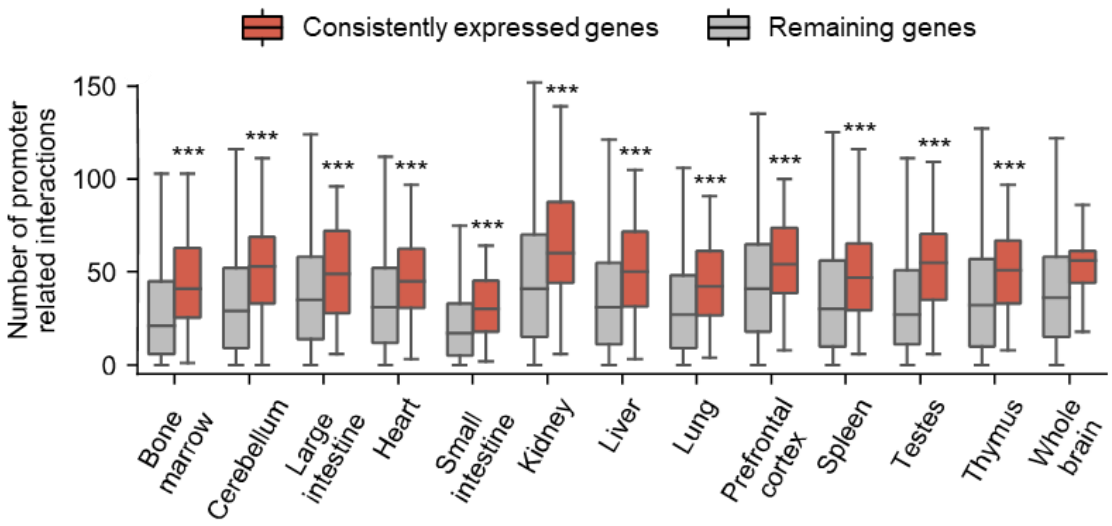

Figure S7. Comparison of the number of promoter-related interactions of 27 consistently and highly expressed genes obtained from [1] and remaining genes. Statistical significance of the difference was calculated using two-sample Wilcoxon tests with $P < 0.01$ in all 13 tissues.

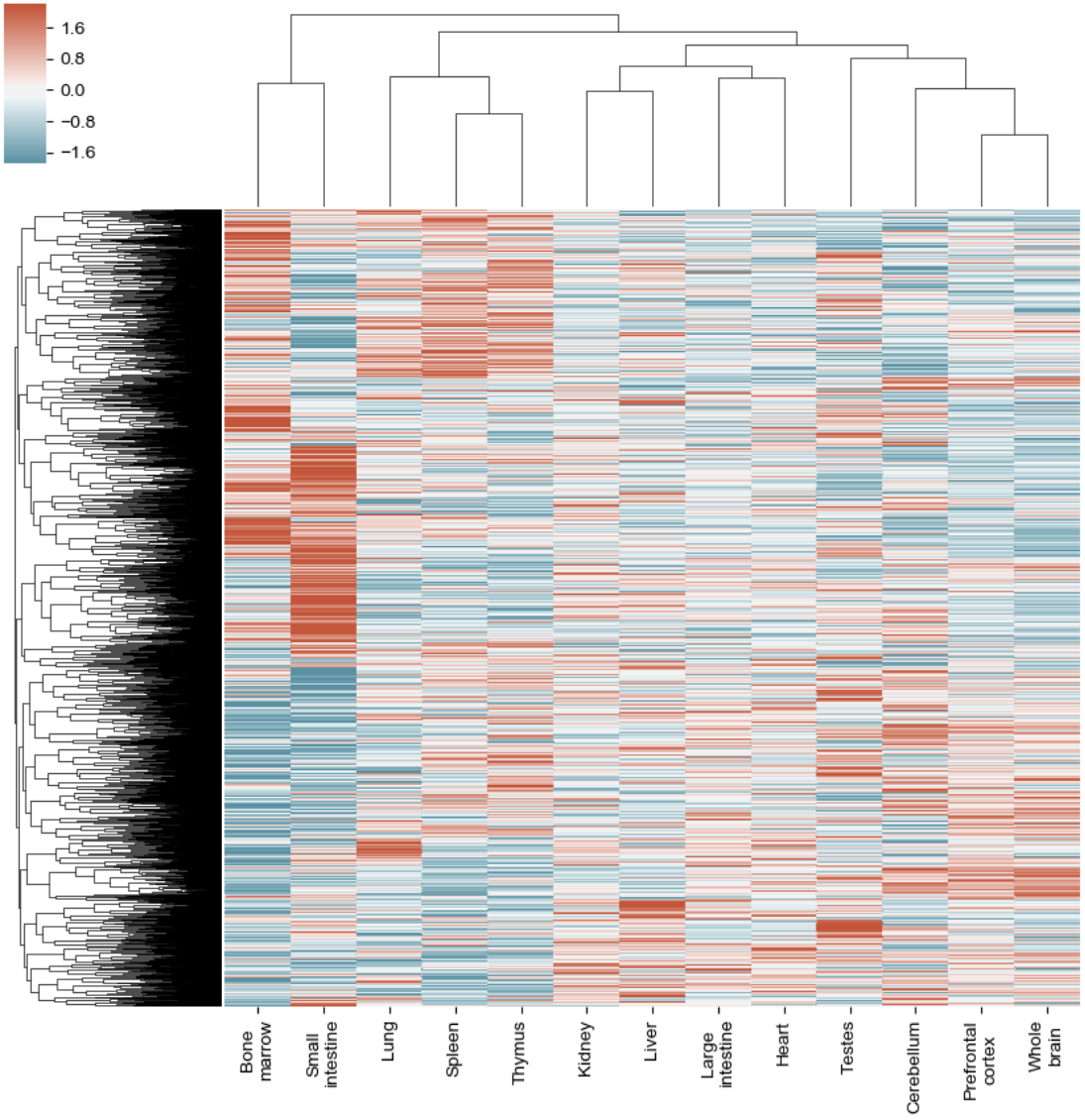

Figure S8. Hierarchical clustering of the 13 tissues in terms of gene activity scores.

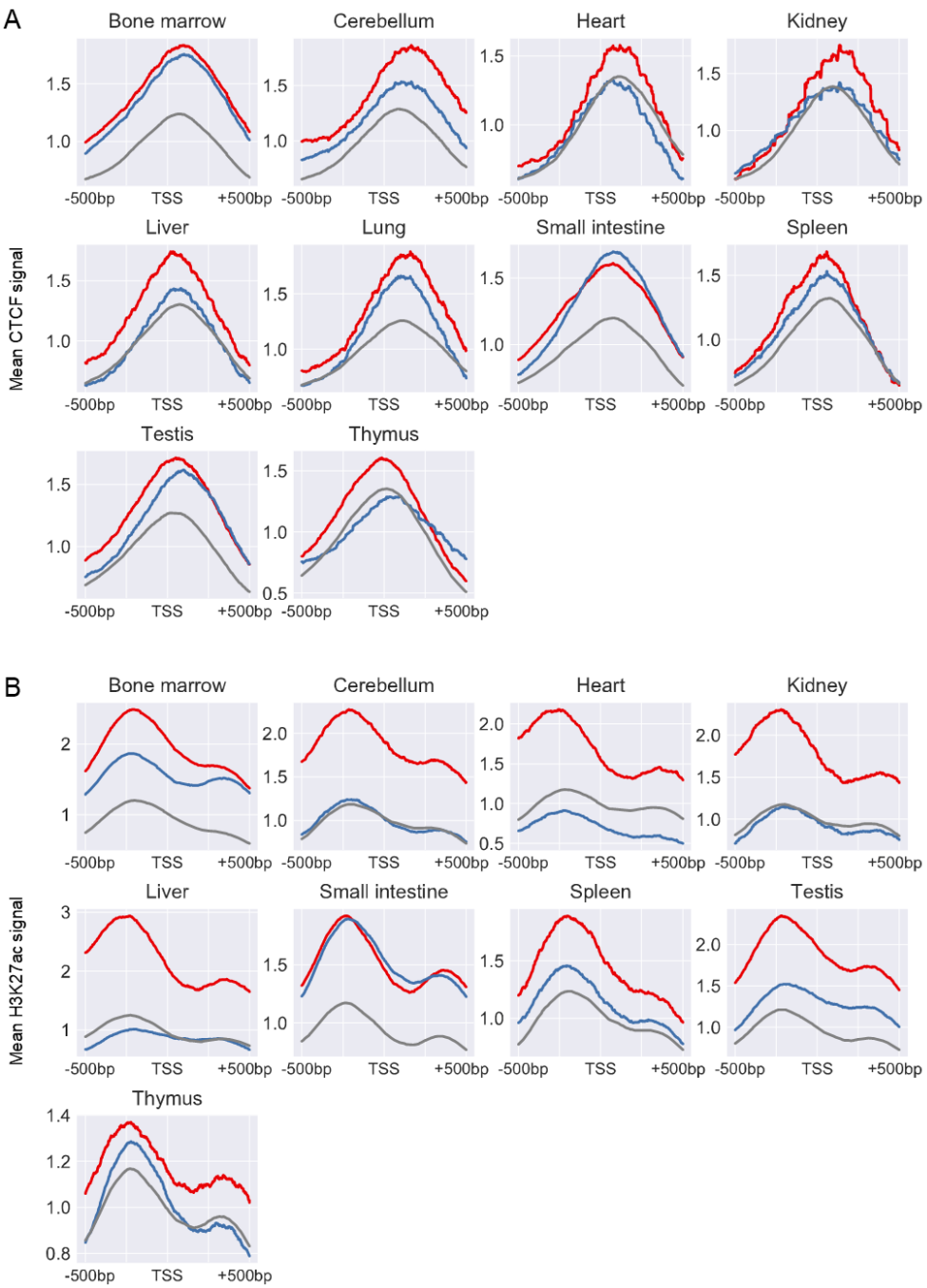

Figure S9. Enrichment analysis of chromatin modification mark around TSSs of tissue-specific differential activity genes (DAGs). (A) Enrichment analysis of CTCF around TSSs of DAGs. The CTCF signal around TSSs of DAGs and remaining genes are labeled as red and gray colors, respectively. The blue line is the mean CTCF signals around TSS of DAGs in other tissues. (B) Enrichment analysis of H3K27ac mark around TSSs of DAGs. The colors of lines are the same as (A).

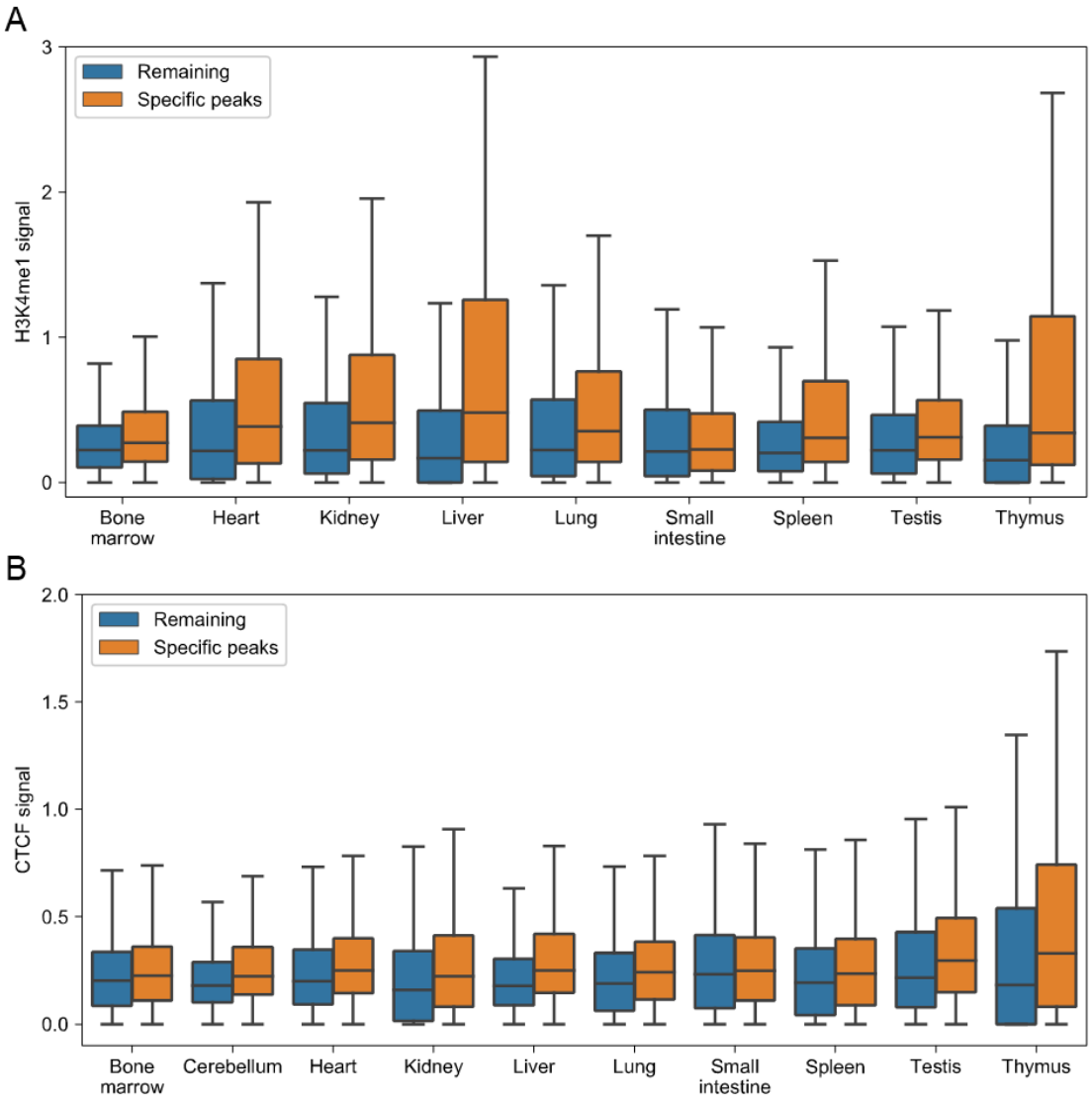

Figure S10. The H3K4me1 signal (A) and CTCF signal (B) of tissue-specific functional peaks compared to those of other peaks.

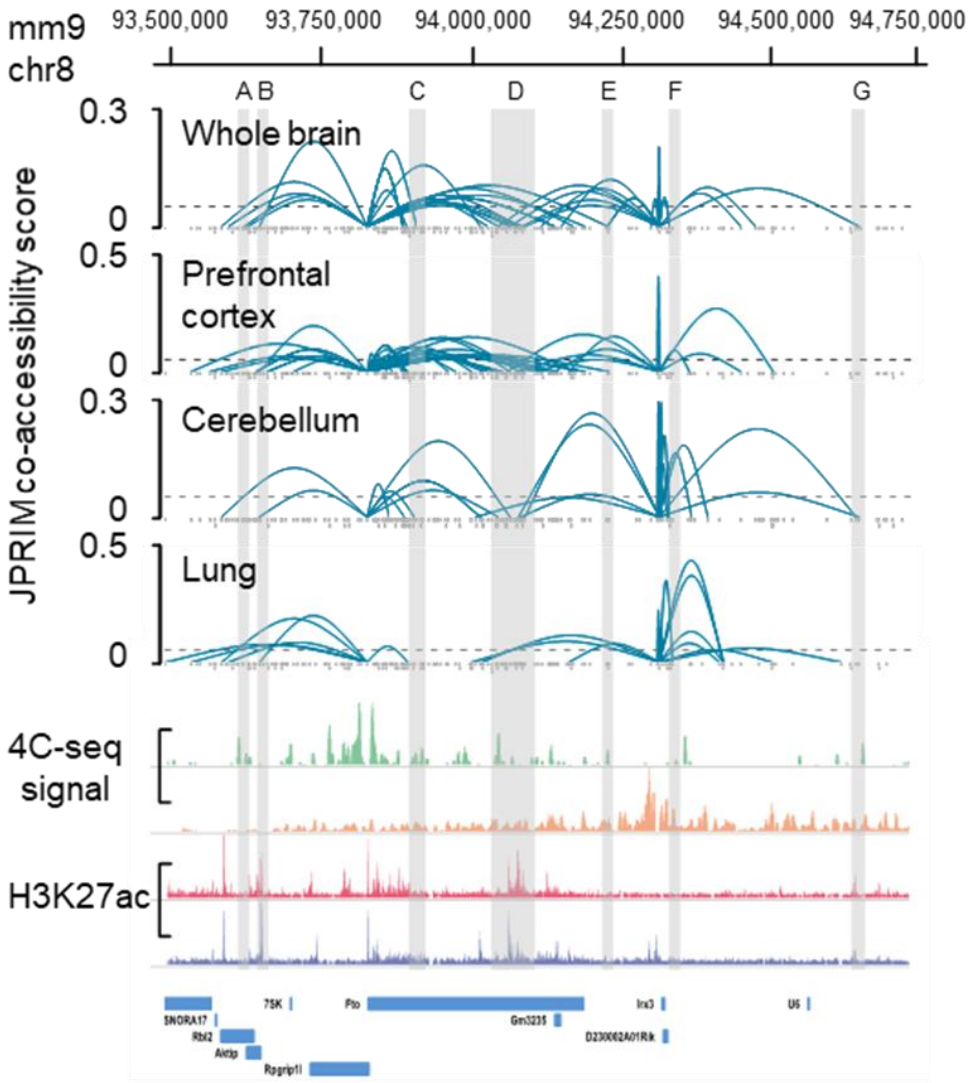

Figure S11. Reconstructed regulatory networks around *Fto* gene and *Irx3* gene (chr9: 93,500,000-94,750,000). The green line and yellow line represent the contact frequencies with *Fto* TSS and *Irx3* TSS respectively. The red line and blue line are the ChIP-seq signal of H3K27ac in cortex and cerebellum respectively. *Irx3* is highly expressed in brain and lung. *Fto* is highly expressed in brain only. JRIM identifies the relatively high activity of *Fto* in brain and prefrontal cortex. And it has been reported [2] that enhancers located in *Fto* region regulate the expression of *Irx3* in brain as shown in the region D.

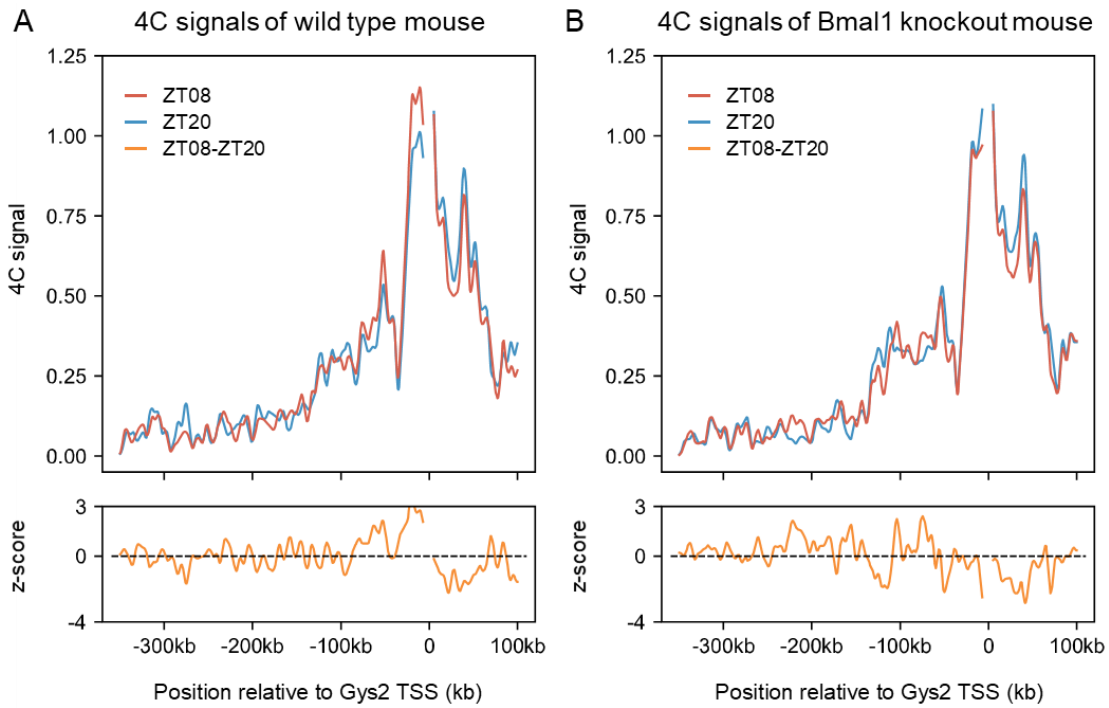

Figure S12. Illustration of 4C-seq data of Gys2 gene in liver at different time points in wild-type mouse (A) and clock-deficient Bmal1 knockout mouse (B). The 4C-seq data was generated in [3].'ZT08' and 'ZT20' indicate that the 4C-seq experiment is performed at zeitgeber time 8 and zeitgeber time 20. 'ZT08-ZT20' indicates the difference between signals at ZT08 and ZT20.

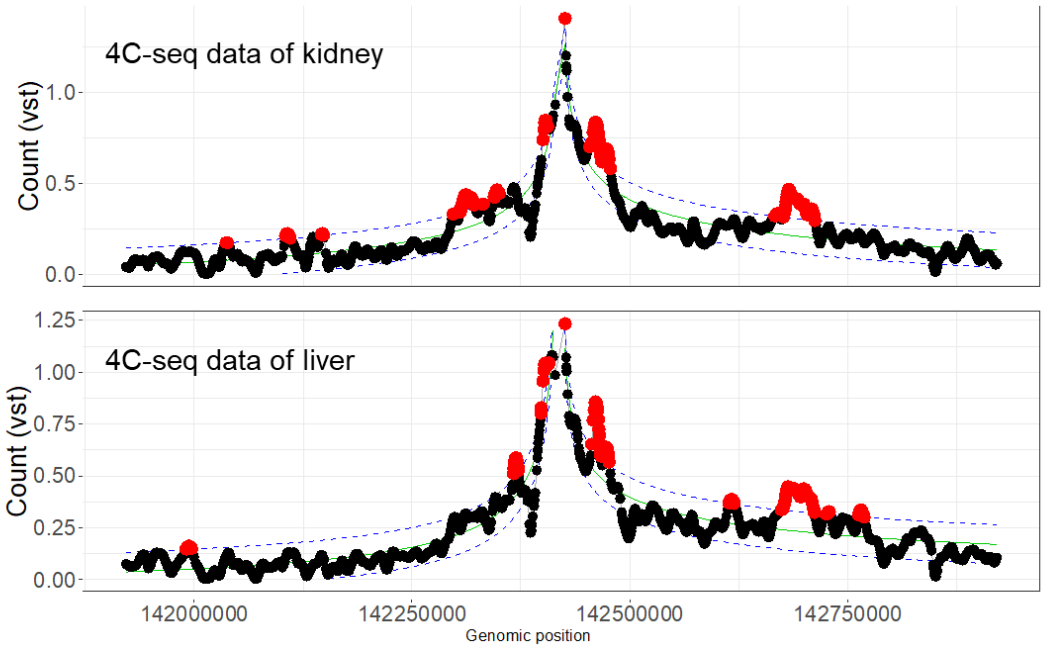

Figure S13. Spatial regulatory loci interacting with *Gys2* TSS in liver and kidney estimated by 'FourCSeq' method [4] from a 4C-seq data with z-score >1.5.

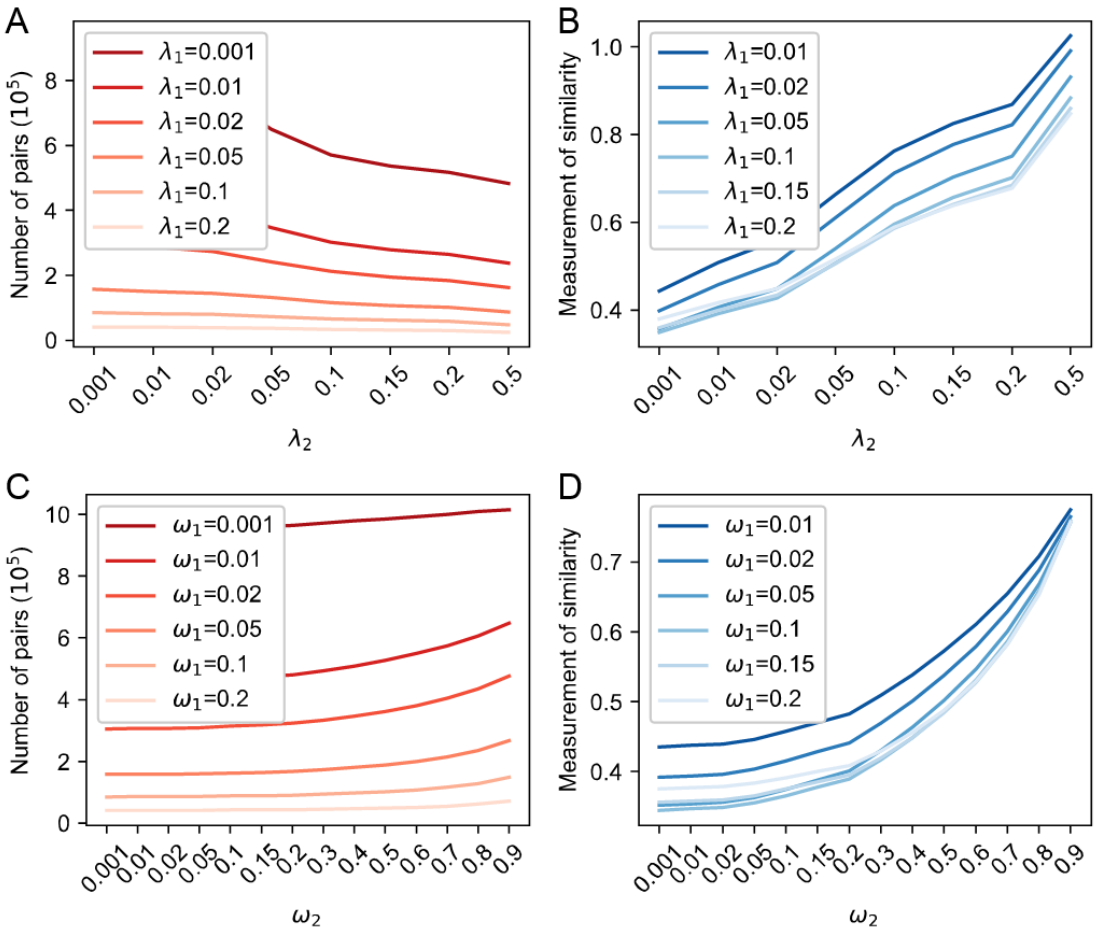

Figure S14. Changes of sparsity and similarity of *cis*-regulatory interaction networks with respect to parameter tuning: λ_1 and λ_2 in (A, B) and ω_1 and ω_2 in (C, D). This experiment was performed in four tissues (kidney, liver, heart and thymus).

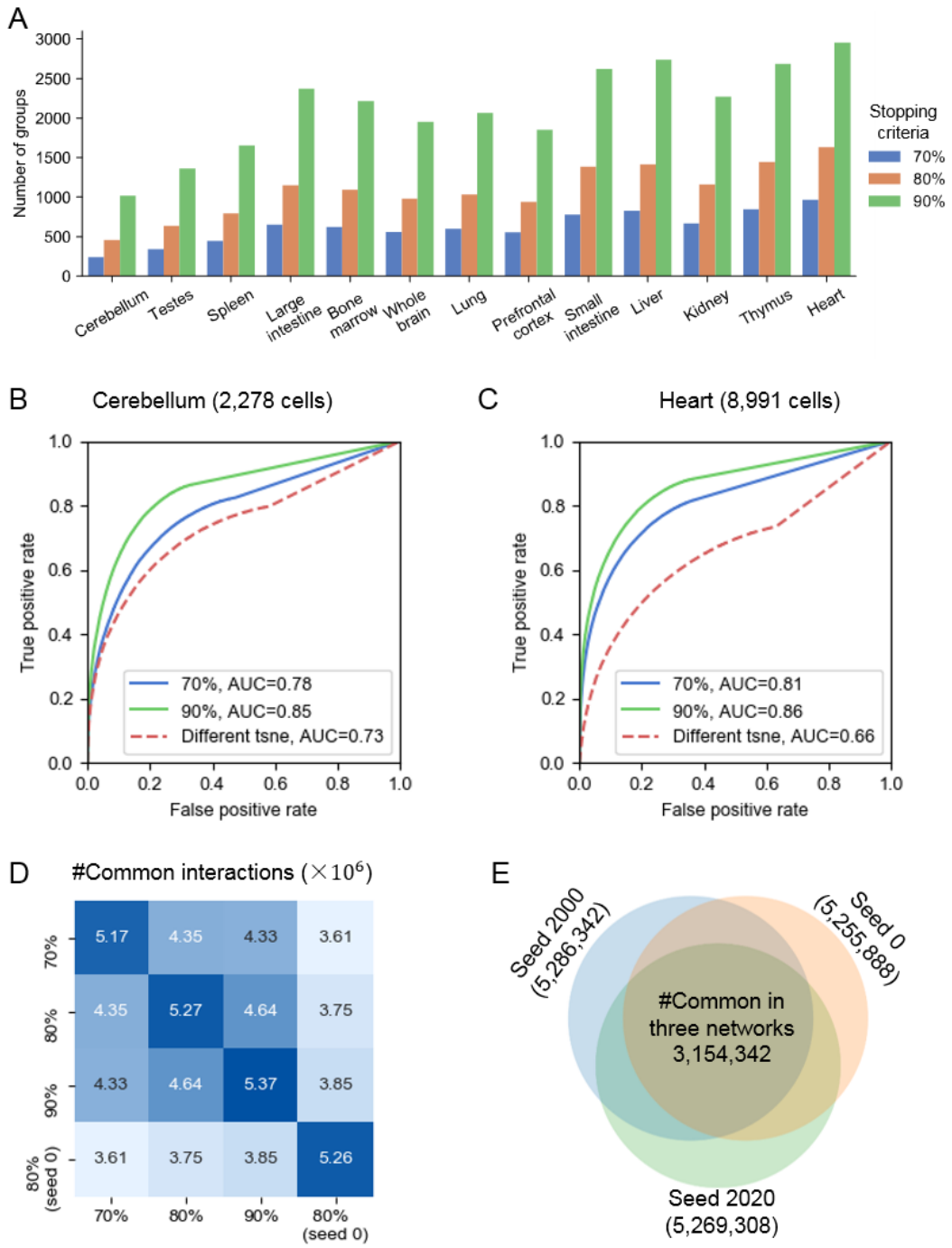

Figure S15. Stability analysis of cell grouping process. (A) The number of groups using different stopping criteria in cell grouping process. Tissues are sorted by the number of cells. (B, C) ROC curves of the consistency with JRIM results using stopping criterion 80% in Cerebellum (tissue with the fewest cells) and Heart (tissue with the most cells). We set the results of stopping criterion 80% as the reference (i.e. the true label) when plotting the ROC curves. (D) The number of common interactions in all 13 tissues between inferred networks using different stopping criteria 70%, 80% and 90% with random seed 2020. (E) Venn diagram about the number of common interactions in three JRIM networks using three different random seeds (i.e., 0, 2000, 2020).

Table S1. Examples of Homer motif analysis for tissue-specific functional peaks. Top three enriched motifs of each tissue are shown.

Tissue	Motif	*P*-value	TF	Brief description	Reference
Bone marrow		1e-45	RUNX1	RUNX1 is essential for the development of normal hematopoiesis and Involved in lineage commitment of immature T cell precursors.	18258917
		1e-44	GATA6		
		1e-43	MAX	MAX is overexpressed in peripheral blood mononuclear cells, CD4 T cells, and monocytes.	17072327
Cerebellum		1e-105	NF1		
		1e-100	NEUROD1	NEUROD1 associates with chromatin to enhancer regulatory elements in neurogenesis regulation genes.	18007592
		1e-99	ATOH1	ATOH1 plays a role in the differentiation of subsets of neural cells by activating E box-dependent transcription.	10648228
Large intestine		1e-229	KLF5	Tissue-specific expression in digestive track. Highest expression in adult mouse colon.	25409824
		1e-217	HOXA11		
		1e-211	KLF14		
Heart		1e-161	MEF2B	MEF2B may be involved in muscle-specific and growth factor-related transcription.	9443808
		1e-153	TBX5	TBX5 regulates the transcription of ion channel genes and is essential for heart development.	20133910
		1e-149	THRB		
Small intestine		1e-77	OLIG2		
		1e-67	CRX		
		1e-65	ATOH1	Express specificity in adult epithelial cells of the gastrointestinal tract.	
Kidney		1e-282	PPARA	PPARA is key regulator of lipid metabolism and regulates the peroxisomal beta-oxidation pathway of fatty acids.	12955147
		1e-277	HNF4A	HNF4A binds fatty acids and may be essential for development of the liver, kidney and intestine.	25409824
		1e-248	RARA		
Liver		1e-313	RARA		
		1e-283	HNF4A	HNF4A binds fatty acids and may be essential for development of the liver, kidney and intestine.	25409824
		1e-272	ERRA		
Lung		1e-119	NKX2.1	NKX2.1 forms a regulatory loop with GRHL2 that coordinates lung epithelial cell morphogenesis and differentiation.	22955271
		1e-114	NKX2.5		
		1e-108	NKX3.1		

Tissue	Motif	*P*-value	TF	Brief description	Reference
Prefrontal cortex		1e-156	OLIG2	OLIG2 is required for oligodendrocyte and motor neuron specification in the spinal cord.	11955448
		1e-143	NEUROD1	NEUROD1 associates with chromatin to enhancer regulatory elements in neurogenesis regulation genes.	18007592
		1e-123	ATOH1	ATOH1 plays a role in the differentiation of subsets of neural cells by activating E box-dependent transcription.	10648228
Spleen		1e-136	ELF4	ELF4 plays a role in the development and function of NK and NK T-cells.	12387738
		1e-128	ETV2		
		1e-123	ETS1	ETS1 controls the differentiation, survival and proliferation of lymphoid cells.	11909962
Testes		1e-102	MYB	MYB plays an important role in the control of proliferation and differentiation of hematopoietic progenitor cells.	20484083
		1e-88	AMYB	AMYB acts as a master regulator of male meiosis by promoting expression of piRNAs.	21750041
		1e-85	THRB		
Thymus		1e-247	ETS1	ETS1 controls the differentiation, survival and proliferation of lymphoid cells.	11909962
		1e-226	ETV2		
		1e-213	FLI1	FLI1 is involved in erythroleukemia induction by Friend murine leukemia virus (F-MULV).	2044959
Whole brain		1e-164	NEUROD1	NEUROD1 associates with chromatin to enhancer regulatory elements in neurogenesis regulation genes.	18007592
		1e-163	OLIG2	OLIG2 is required for oligodendrocyte and motor neuron specification in the spinal cord.	11955448
		1e-151	NEUROG2	NEUROG2 is Involved in neuronal differentiation.	14697366

REFERENCES

1. Li, B., et al., *A comprehensive mouse transcriptomic BodyMap across 17 tissues by RNA-seq.* Scientific reports, 2017. **7**(1): 4200.
2. Smemo, S., et al., *Obesity-associated variants within FTO form long-range functional connections with IRX3.* Nature, 2014. **507**(7492): 371-375.
3. Mermet, J., et al., *Clock-dependent chromatin topology modulates circadian transcription and behavior.* Genes & development, 2018. **32**: 347-358.
4. Klein, F.A., et al., *FourCSeq: analysis of 4C sequencing data.* Bioinformatics, 2015. **31**(19): 3085-3091.