Unraveling the spatial organization and development of human thymocytes through integration of spatial transcriptomics and single-cell multi-omics profiling

Yanchuan Li[1,#], Huamei Li[2,#], Cheng Peng[2], Ge Meng[3,4,5], Yijun Lu[6], Honglin Liu[7], Li Cui[8], Huan Zhou[9], Zhu Xu[6], Lingyun Sun[2], Lihong Liu[7,*], Qing Xiong[2,*], Beicheng Sun[6,*], Shiping Jiao[3,4,5,*]

[1]Department of Cell Biology, School of Basic Medical Sciences, Nanjing Medical University, Nanjing, China

[2]Department of Rheumatology and Immunology, Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School, Nanjing, China

[3]Department of General Surgery, Sir Run-Run Shaw Hospital, Zhejiang University, Hangzhou, China

[4]TCRX (KeShiHua) Therapeutics Co, Ltd. Beijing & Yunnan Pilot Free Trade Zone (Dehong Area), China

[5]Department of Oncology, Department of Rheumatology and Immunology, Ruili JingCheng Hospital, Ruili, China

[6]Department of Hepatobiliary Surgery, First Affiliated Hospital of Anhui Medical University, Hefei, China

[7]Department of Pharmacy, Organoid and Regenerative Medicine Center, China-Japan Friendship Hospital, Beijing, China

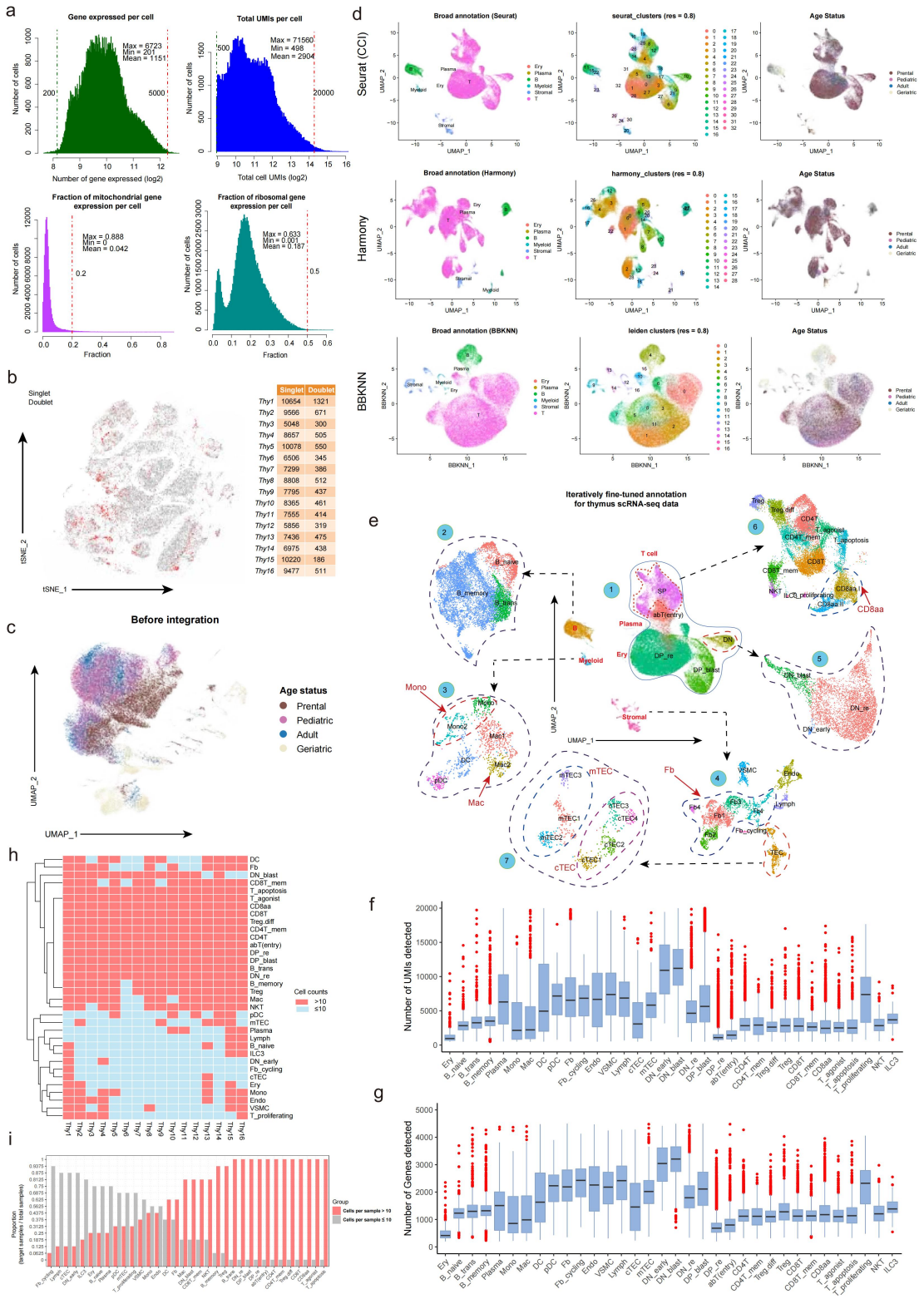
[8]Stomatological Hospital, School of Stomatology, Southern Medical University, Guangzhou, China

[9]National Institute of Drug Clinical Trial, The First Affiliated Hospital of Bengbu Medical College, Bengbu, China

[#] These authors contributed equally to the work.

* Correspondence: llh-hong@outlook.com (L.H.L.); qingx@tcrximmune.cn (Q.X.); sunbc@nju.edu.cn (B.C.S.); jiaoshp@tcrximmune.cn (S.P.J.)

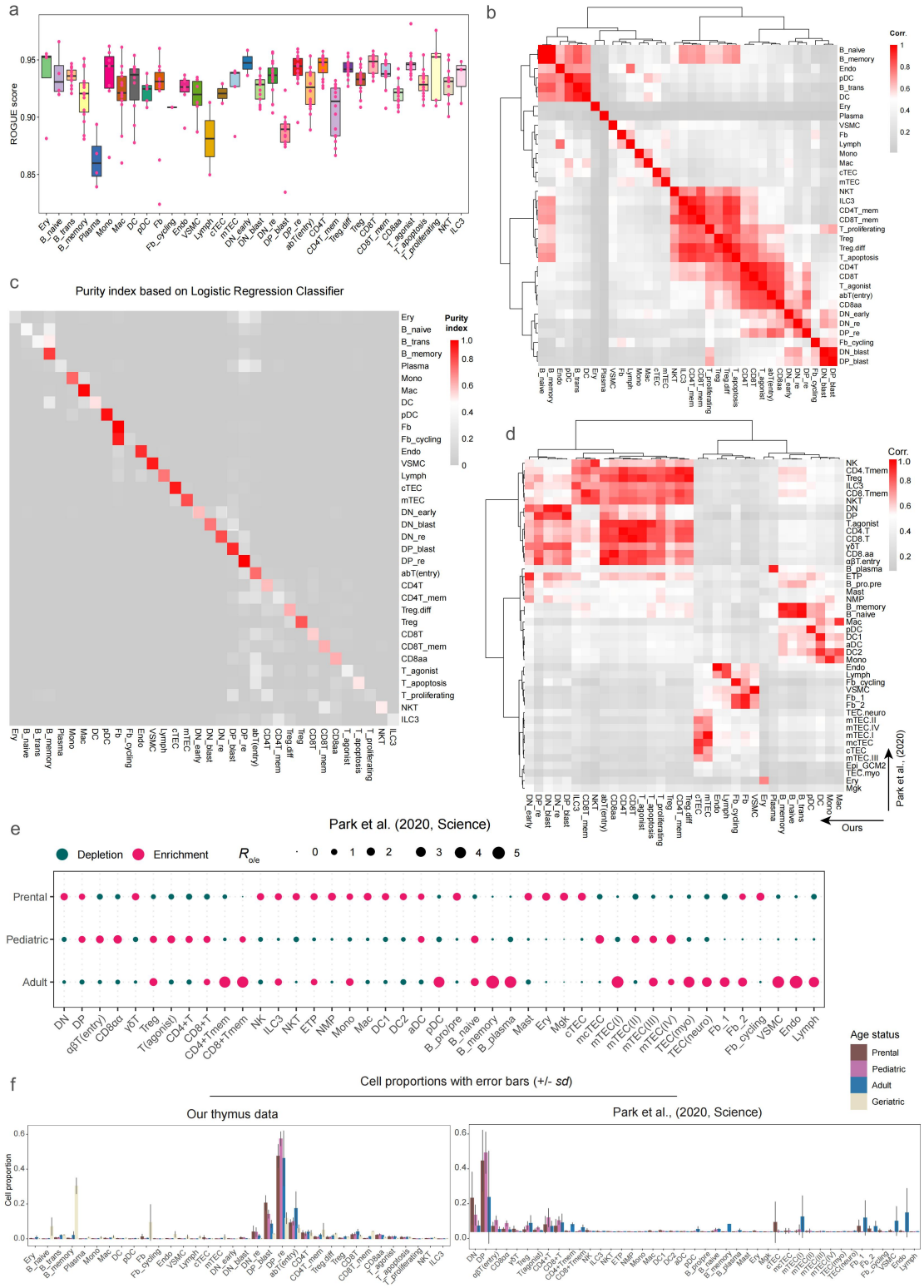
Supplementary Figures



	Singlet	Doublet
Thy1	10654	1321
Thy2	9566	671
Thy3	5048	300
Thy4	8657	505
Thy5	10078	550
Thy6	6506	345
Thy7	7299	386
Thy8	8808	512
Thy9	7795	437
Thy10	8365	461
Thy11	7555	414
Thy12	5856	319
Thy13	7436	475
Thy14	6975	438
Thy15	10220	186
Thy16	9477	511

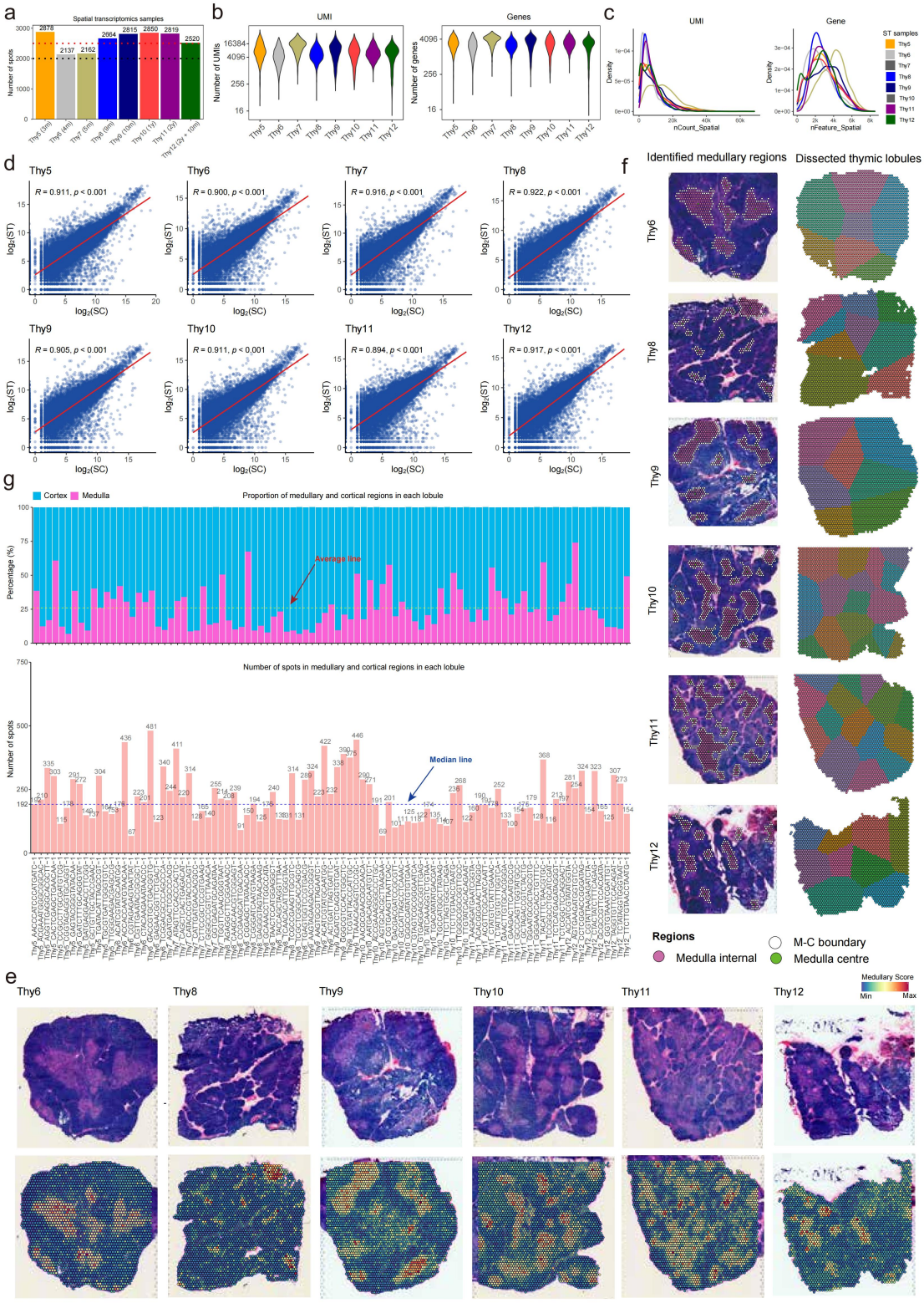
Supplementary Fig. 1 Basic characteristics and annotations of normal thymus scRNA-seq datasets. (**a**) Histograms showing the distribution of the normal thymus scRNA-seq datasets from four aspects: total UMIs (Unique Molecular Identifiers) per cell, expressed genes per cell, fraction of mitochondrial gene expression per cell, and fraction of ribosomal gene expression per cell. Dark green and red dashed lines indicate filtering thresholds. (**b**) *t*-distributed stochastic neighborhood embedding

(*t*-SNE) plot of cells from the normal thymus, color-coded by classification into singlets (grey) and doublets (red) cells following dehashing. (**c**) Uniform Manifold Approximation and Projection (UMAP) plot showing the distribution of cells across age statuses before integration for thymus samples. Each dot represents an individual cell, with age statuses marked by color codes. (**d**) Comparing the performance of different integration tools in removing batch effects from thymus scRNA-seq data across various age statuses. (**top**) Seurat; (**middle**) Harmony [1]; (**bottom**) BBKNN [2]. (**e**) UMAP plot showing annotation processes of thymus scRNA-seq data (n=130,295). Initial broad categories include T cells, B cells, Plasma cells, Myeloid cells, Stromal cells, and Ery cells. Iterative annotations were further performed on T cells, B cells, Myeloid cells, and Stromal cells. (**f, g**) Boxplots showing the number of UMIs (**f**) or genes (**g**) detected per cell type inferred from normal thymus samples. Each box represents the median (black line) and interquartile range (IQR 25th–75th percentiles), with whiskers indicating the highest and lowest values within 1.5 times the IQR. Outlier cells are marked as red dots. (**h**) Heat map showing the count of cells for annotated cell types present in each thymus sample, with hierarchical clustering conducted by rows (cell types). Cells are color-coded, with pink representing cell counts exceeding 10 and sky blue representing counts equal to or below 10. (**i**) Bar plot showing the proportion of thymus samples where the cell count for specific cell types exceeds or falls below 10, respectively, out of the total thymus samples. Source data are provided as a Source Data file.



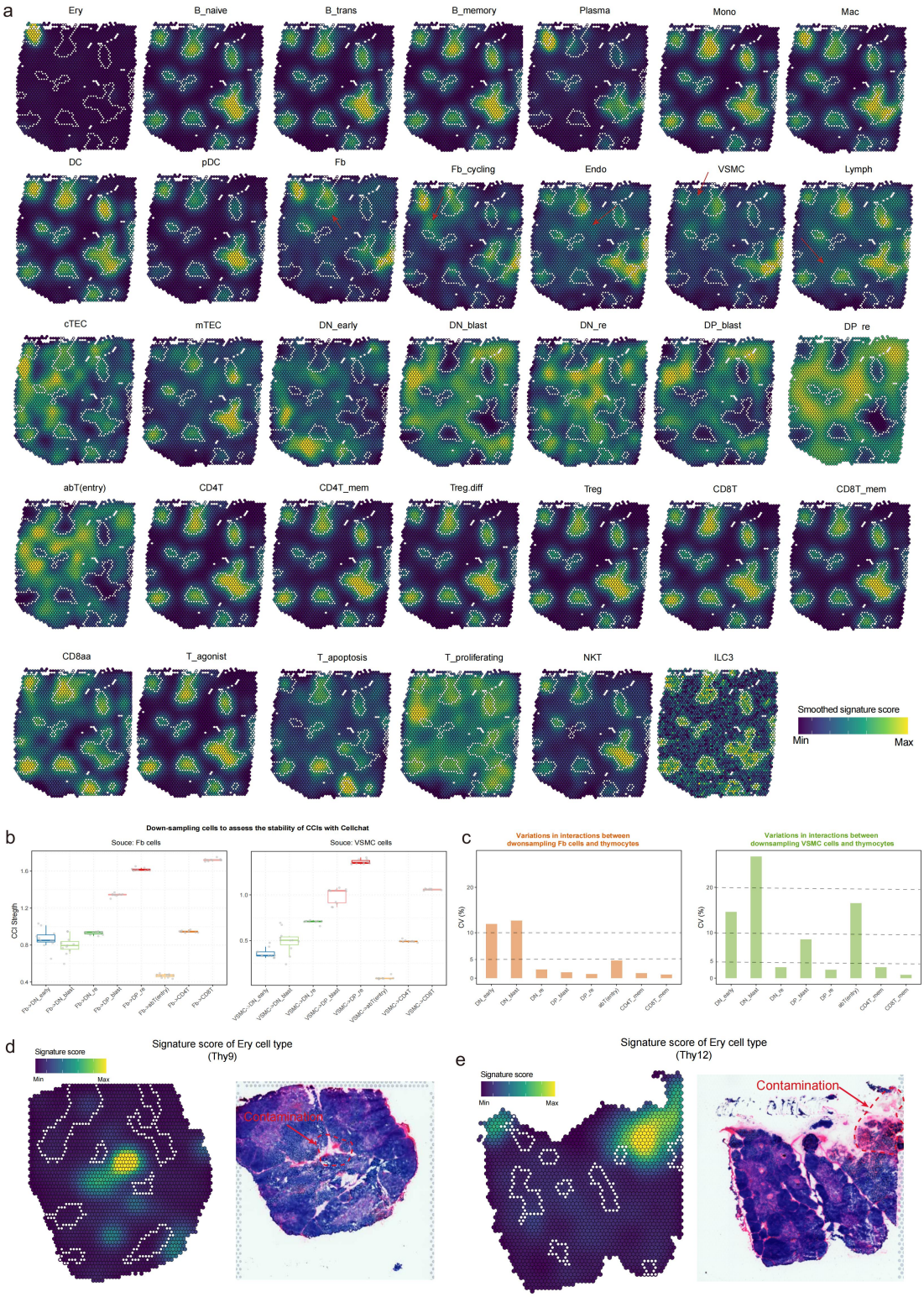
Supplementary Fig. 2 Assessment of annotation reliability for thymus scRNA-seq datasets from multiple perspectives. (**a**) Boxplot showing the ROUGE index for each cell type. Each dot represents a specific normal thymus sample. (**b**) Heat map showing the *Pearson* correlation coefficients between the averaged expression profiles of pairwise cell types. (**c**) Heat map showing the purity index defined by a logistic regression model of cell subsets. (**d**) Heat map showing *Pearson* correlation coefficients between the cell types inferred from our dataset and those inferred from the thymus dataset of Park et

al. based on expression profiles. (**e**) Age group preference of each cell type measured by the ratio of the observed number of cells to the expected number of cells at random ($R_{o/e}$), using the thymus dataset of Park et al.. (**f**) Grouped bar plots with error bars showing the distribution of cellular proportions across different age statuses, with error bars representing one standard deviation (*sd*) relative to the mean proportion of cell types. (**left**) Our thymus data; (**right**) Park et al. (2020, Science). Source data are provided as a Source Data file.



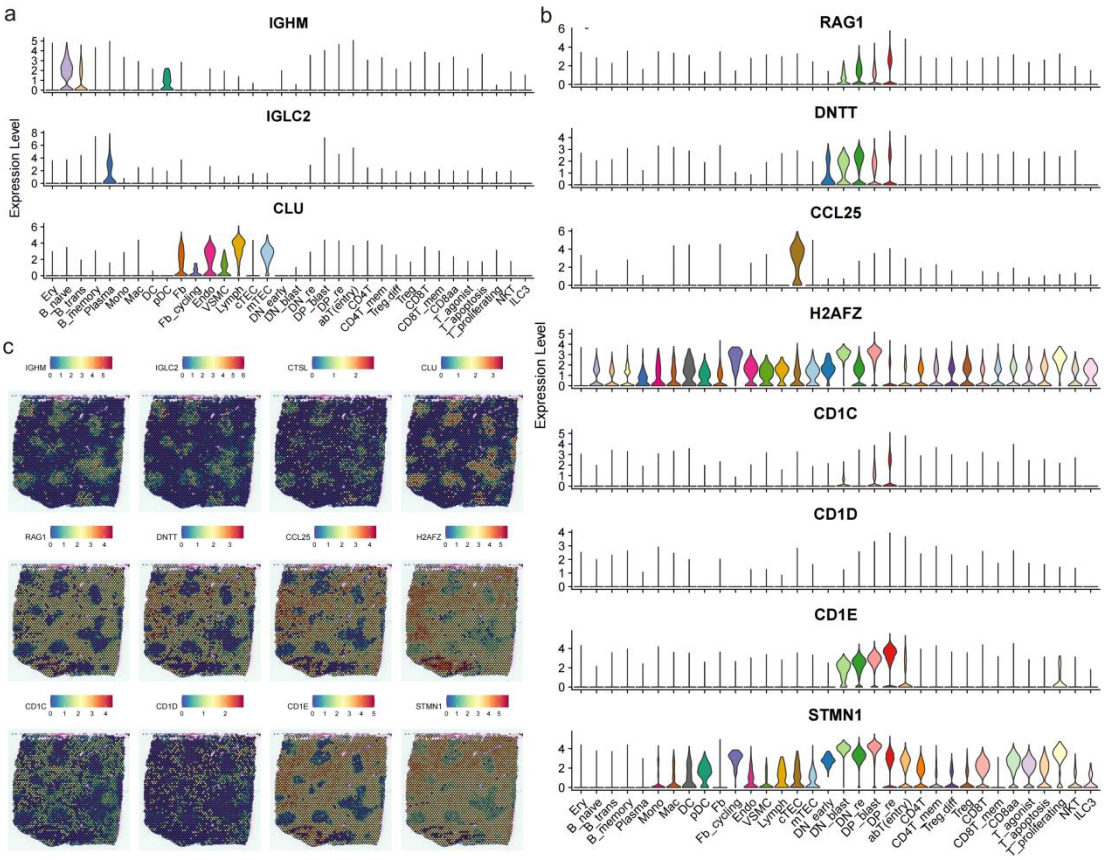
Supplementary Fig. 3 Basic characteristics of the thymus spatial transcriptomics (ST) datasets, and the critical anatomical regions determined using TSO-his. (**a**) Bar plot showing the number of spots detected in the ST samples. The red line indicates the average number of spots. (**b**) Distribution of the number of genes (**left**) and UMIs (Unique Molecular Identifiers) (**right**) detected in the spots of each ST sample. The y-axis is logarithmized with base 2. (**c**) Density plots showing the distribution of detected UMIs (**left**) and genes (**right**) in spots from each ST sample. ST samples are marked with

color codes. (**d**) Scatter plots showing the *Pearson* correlations between aggregated gene counts from cells of the SC (Single-cell) sample and aggregated gene counts from spots of matched ST sample. Each dot represents one gene. *R* represents the *Pearson* correlation coefficient. *P*-values were obtained using a two-sided *t*-test. (**e**) Hematoxylin and Eosin (H&E) staining images of six pediatric donors (**top**). Projection of signature scores based on medullary-indicator genes onto six pediatric ST sections (**bottom**). (**f**) Critical anatomical regions of six pediatric ST sections determined using TSO-his. (**left**) Cortical regions, medullary regions, Medulla-Cortex (M-C) boundaries, and Medullary centers. (**right**) Thymic lobules. (**g**) Bar plot showing the composition of cortical and medullary regions within the lobules. The proportion of medullary and cortical regions in each lobule is approximately 1:3, with the yellow line representing the average percentage (**top**). Number of spots in medullary and cortical regions in each lobule, with the blue line indicating the median number of spots (**bottom**). Source data are provided as a Source Data file.

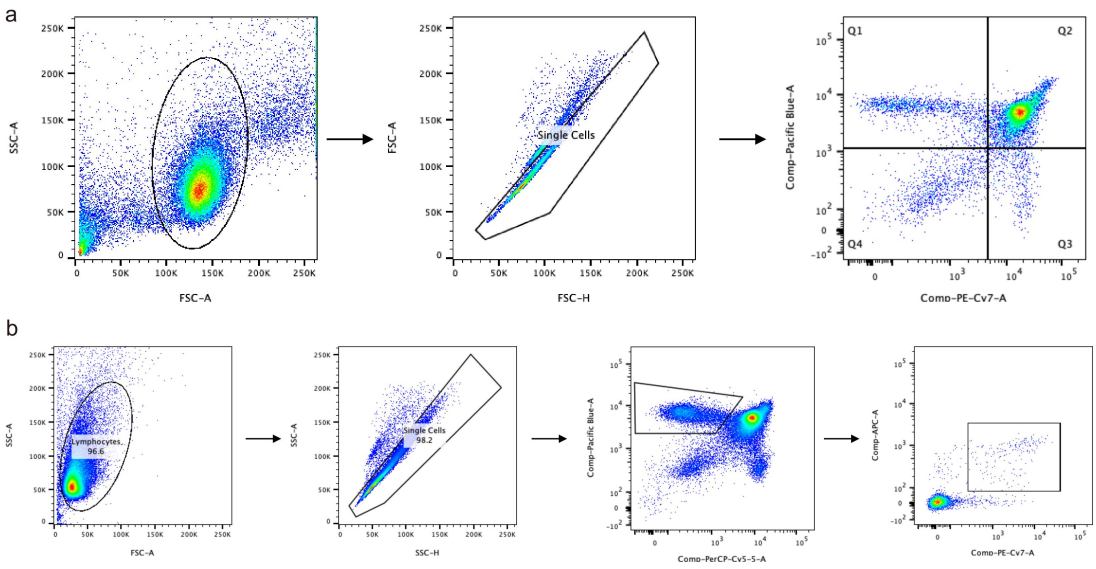


Supplementary Fig. 4 Projection of signature scores for cell types onto spatial transcriptomics (ST) sections. (**a**) Projection of signature scores for cell types onto the Thy5 ST section, with scores smoothed using the "wkde" method in the *Nebulosa* R package. The highlighted areas are clearly distinguishable in the cortical and medullary regions. (**b**) Boxplots showing the intensity of cell-cell

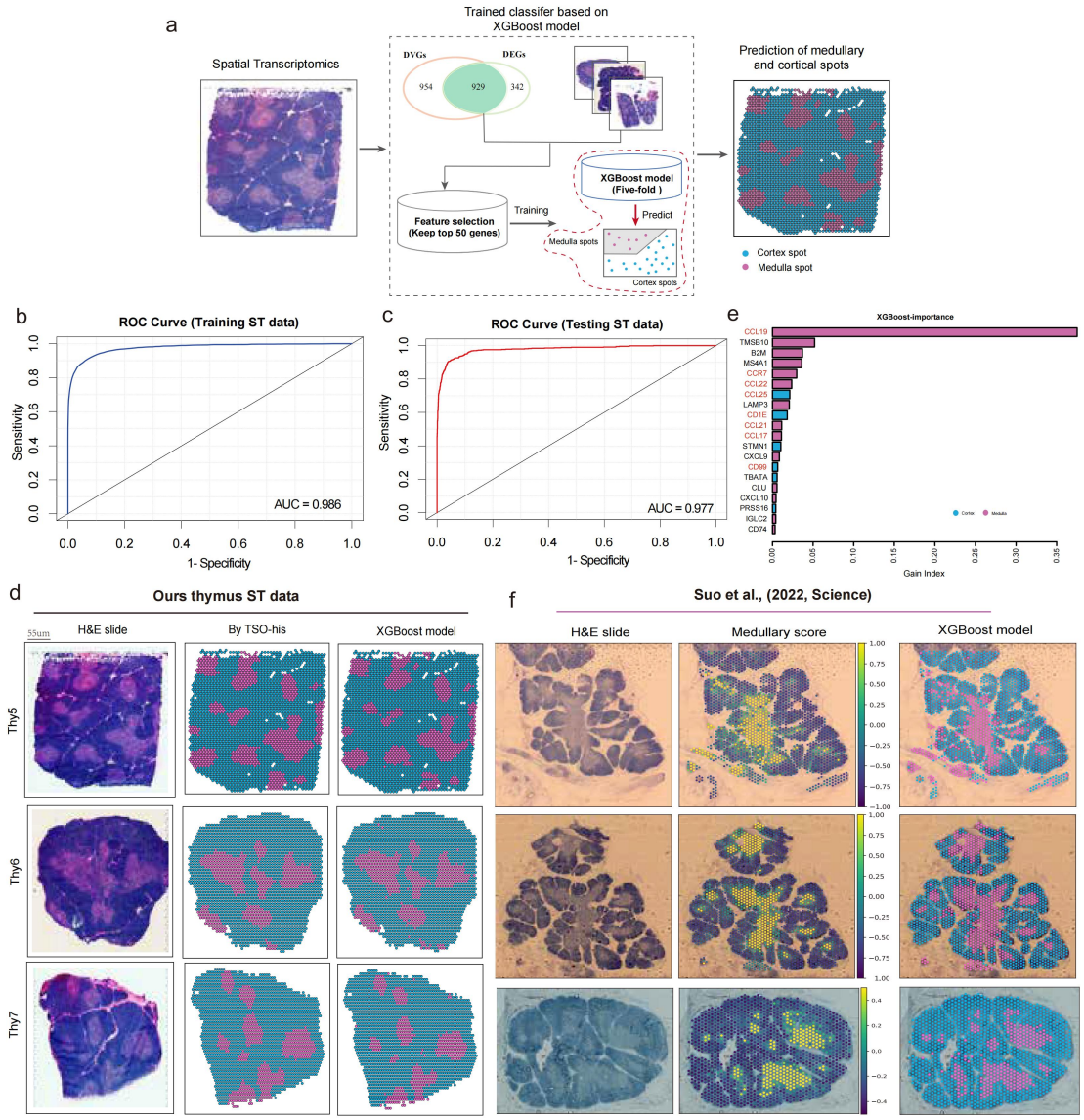
interactions (CCIs) between fibroblasts (Fb) (**left**) and vascular smooth muscle cells (VSMC) (**right**) with thymocytes at various down-sampling levels, as inferred by CellChat. Each box represents the median (internal solid line) and interquartile range (IQR 25th–75th percentiles), with whiskers indicating the highest and lowest values within 1.5 times the IQR. Outliers are marked as grey dots. Down-sampling were performed 11 times. (**c**) Bar plots showing the coefficient of variation (CV) of CCIs for each pair of cell type interactions. (**left**) Fb; (**right**) VSMC. (**d, e**) Projection of Ery signature scores onto the ST section (**left**). Hematoxylin and Eosin (H&E) image, with the region covered by a red dashed line indicating erythrocyte contamination (**right**). (**d**) Thy9; (**e**) Thy12. Source data are provided as a Source Data file.



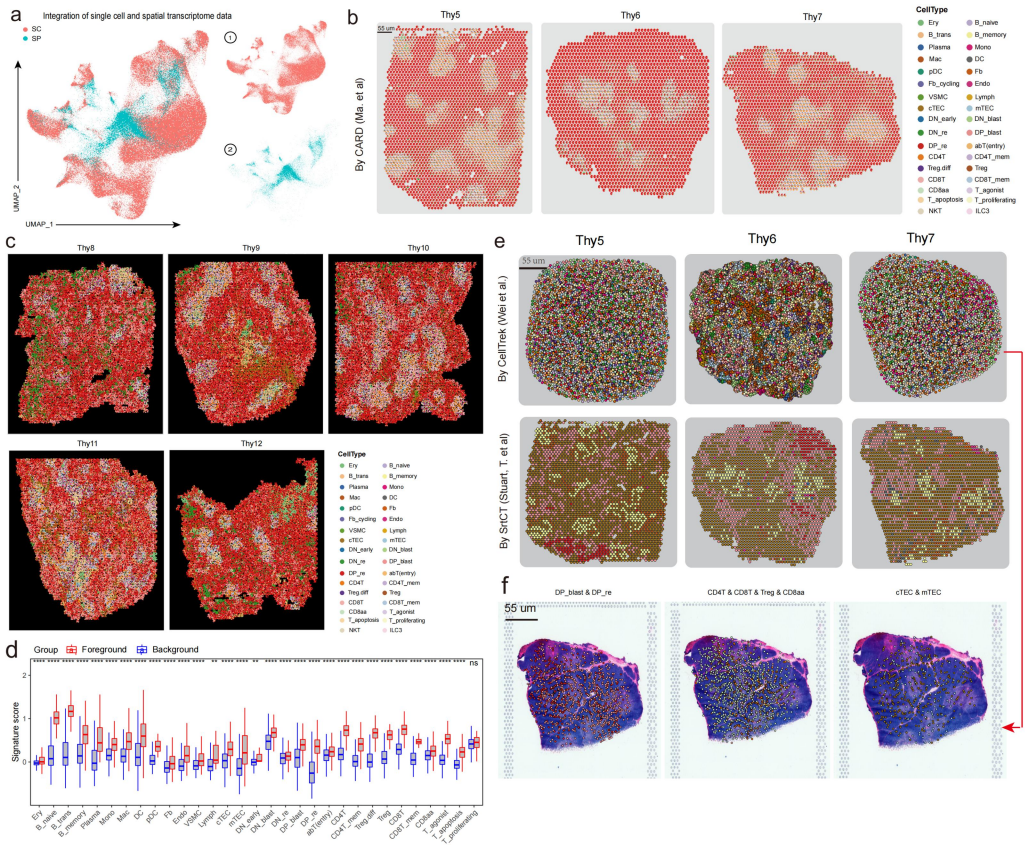
Supplementary Fig. 5 Expression and spatial localization of selected T cell development indicator genes. (**a,b**) Expression levels of selected genes in thymic cell types. (**c**) Spatial expression of selected genes in the Thy5 spatial transcriptomics (ST) sample. Source data are provided as a Source Data file.



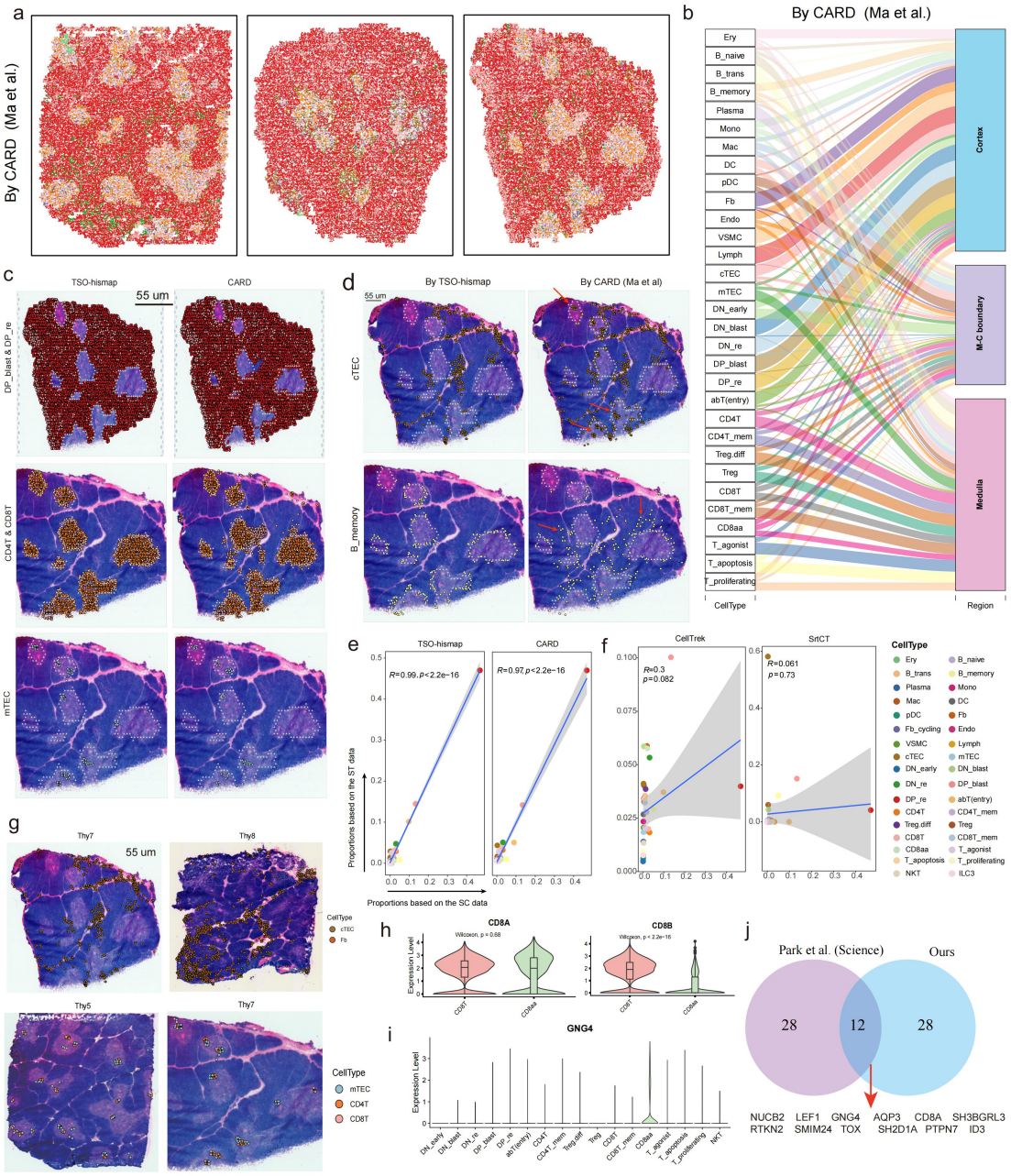
Supplementary Fig. 6 Flow cytometry gating strategies. (**a, b**) Flow cytometry gating strategies.



Supplementary Fig.7 Enhancement of TSO-his via the XGBoost model using distance-varying genes (DVGs). (**a**) Workflow for training the XGBoost model to predict candidate cortical and medullary spots using genes with overlapping DVGs and differentially expressed genes (DEGs) as input features. (**b, c**) Receiver operating characteristic (ROC) curve of the XGBoost algorithm for predicting candidate cortical and medullary spots in the training spatial transcriptomics (ST) data (**b**) and testing ST data (**c**). (**d**) Evaluation of the predictive performance of the well-trained XGBoost model using testing ST sections of Thy5 (**top**), Th6 (**middle**) and Th7 (**bottom**) samples. (**left**) Hematoxylin and Eosin (H&E) sections; (**middle**) cortical and medullary spots determined by TSO-his; and (**right**) candidate cortical and medullary spots determined by the XGBoost model. (**e**) Bar plot showing the importance of the top 20 genes inferred from the well-trained XGBoost model. Marker genes from the medulla and cortex are distinguished by color codes. (**f**) Validation of the well-trained XGBoost model performance using three normal thymus ST slices provided by Suo et al. (**left**) H&E slides; (**middle**) spatial spot scoring using medullary indicator genes, with the medullary score clearly indicating the medullary region; and (**right**) candidate cortical and medullary spots determined by the well-trained XGBoost model. Source data are provided as a Source Data file.

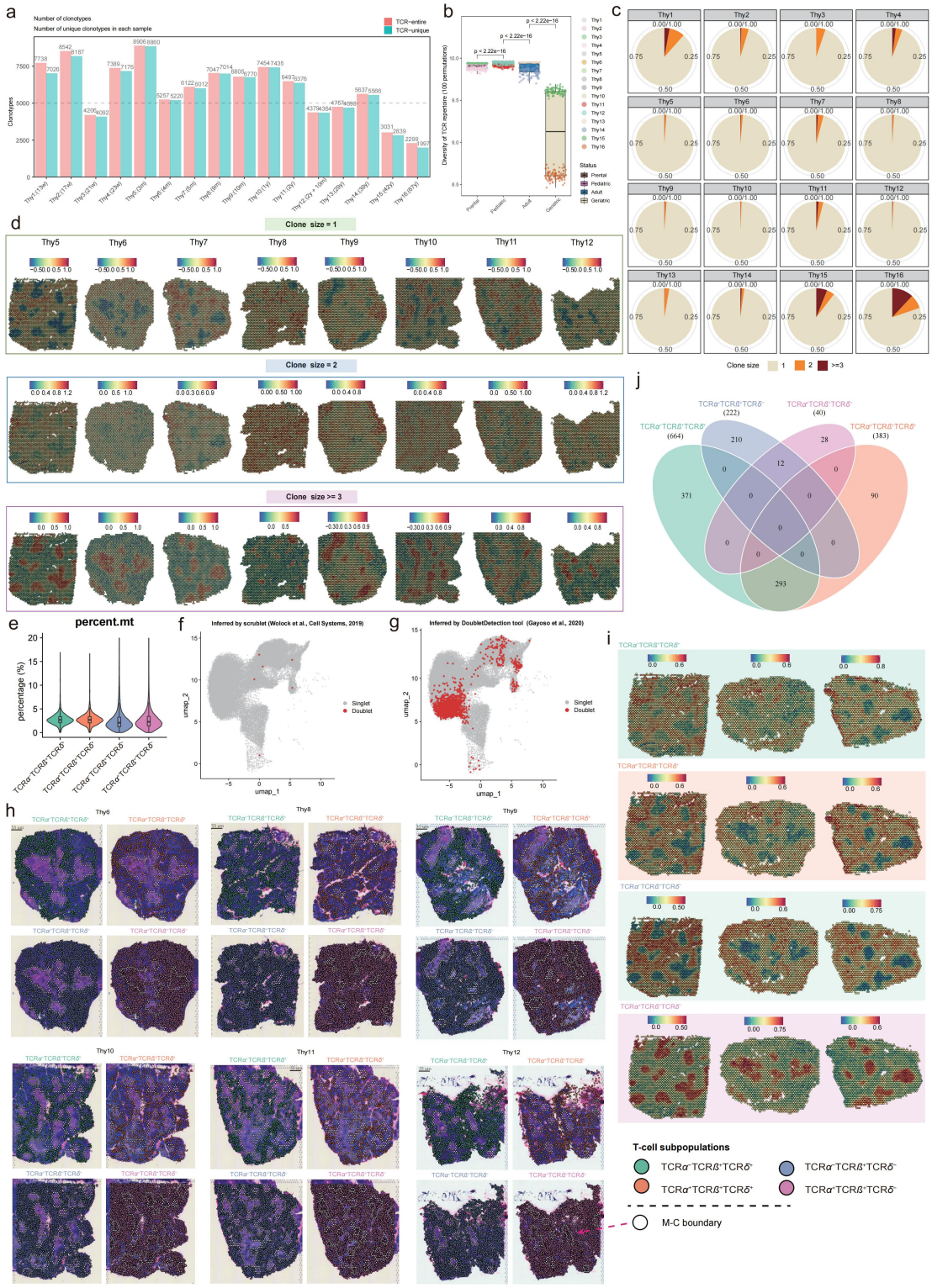


Supplementary Fig. 8 Projection of single cells onto spatial transcriptomics (ST) sections. (**a**) Uniform Manifold Approximation and Projection (UMAP) plot showing the integration of thymic single-cell (SC) data with ST data using the Canonical Correlation Analysis (CCA) method from Seurat package. Data types are colored by their source, and each dot represents a single cell or spot. (**b**) Scatter pie plots showing the cellular proportions of spots from ST sections as inferred by the CARD tool. Cell types are marked by color codes. (**left**) Thy5; (**middle**) Thy6; and (**right**) Thy7. (**c**) Projection of single cells onto ST sections using TSO-hismap. Each dot represents a single cell, with cell types marked by color codes. (**d**) Box plot showing the foreground score compared to the background score for cell types projected onto the eight ST sections. The foreground represents spots of one cell type falling on ST sections, while the background consists of spots from single cells that do not contain that cell type. Signature scores of the spots were obtained from the top 20 highly expressed marker genes of each cell type. Each box represents the median (black line) and interquartile range (IQR 25th–75th percentiles), with whiskers indicating the highest and lowest values within 1.5 times the IQR. Statistical analysis was performed using one-tailed *t*-test. ns: not significant, $^{**}p \leq 0.01$, $^{****}p \leq 0.0001$. (**e**) Projection of thymic single cells onto ST sections using CellTrek (**top**) and SrtCT (**bottom**). Each dot represents a single cell. Color codes of the cell types can be found in Supplementary Fig. 8c. (**f**) Examples of CellTrek projecting thymic single cells onto ST sections, including DP_blast (Double positive blast cells) & DP_re (Double positive rearrangement cells) (**left**), CD4 T & CD8 T & Treg (Regulatory T cells) & CD8aa (**middle**), and cTEC (Cortical thymic epithelial cells) & mTEC (Medullary thymic epithelial cells) (**right**). Source data are provided as a Source Data file.



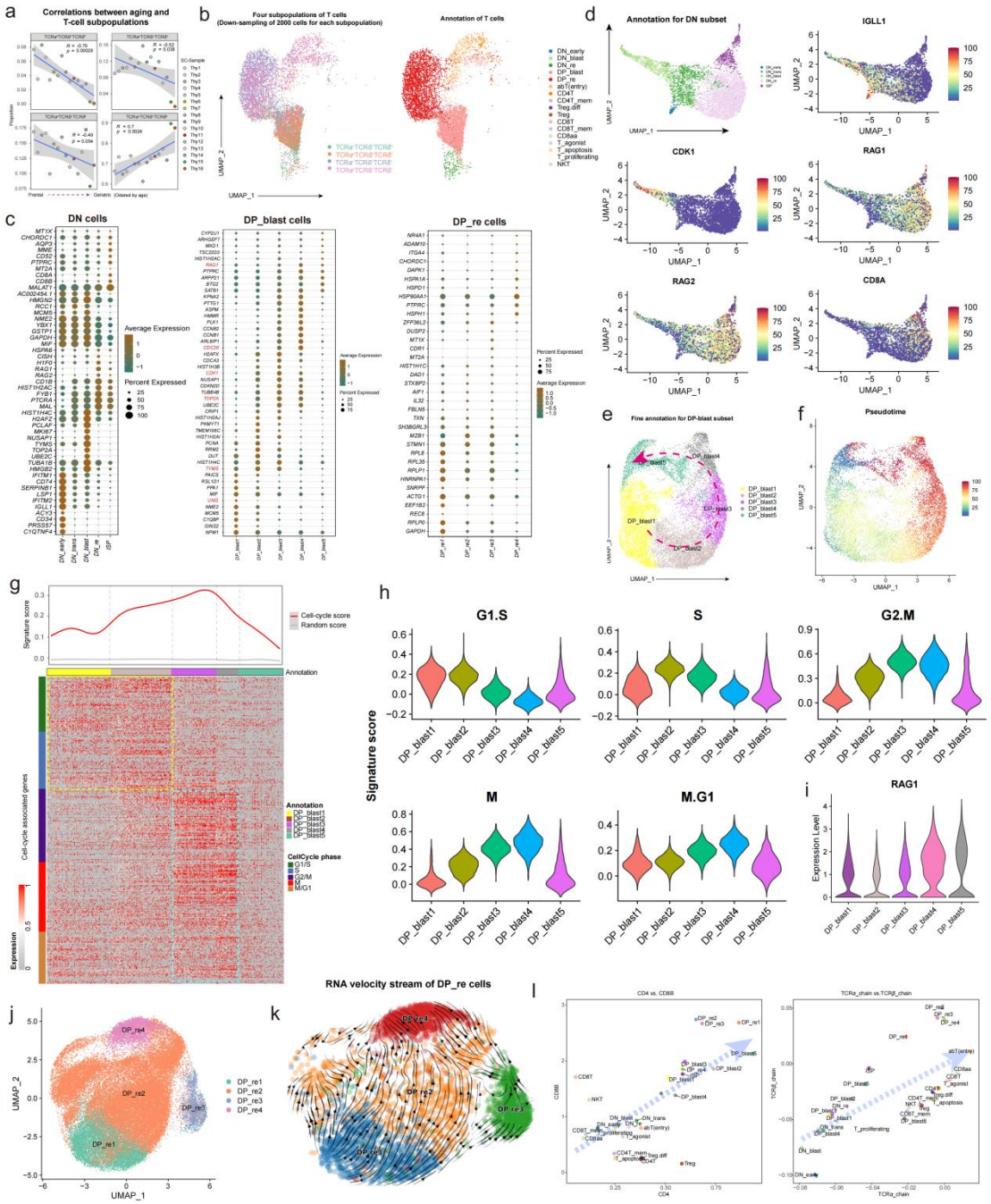
Supplementary Fig. 9 Comparison of the performance of different tools for projecting single cells onto spatial transcriptomics (ST) sections. (**a**) Projection of thymic single cells onto ST sections using CARD. Examples include Thy5, Thy6 and Thy7. Each dot represents a single cell from the normal thymus, with cell types marked by color codes. (**b**) Sankey plot showing the proportion of single cells of each cell type projected onto the medullary, cortical and Medulla-Cortex (M-C) boundary regions of ST slices using CARD. The thickness of the lines indicates the proportion of each cell type projecting to that region. (**c**) Comparison of the projection performance of TSO-hismap and CARD for specific cell types, including DP_blast & DP_re (**top**), CD4T & CD8T (**middle**), and mTEC (**bottom**). The white dashed line denotes the M-C boundary. (**d**) Distribution of cTEC and B_memory single cells projected onto thymic ST sections by TSO-hismap (**left**) and CARD (**right**). cTEC (**top**) and B_memory (**bottom**). Each dot represents a cell. (**e, f**) Comparison of the concordance between the abundance of cell types projected onto ST sections and their abundance in the single-cell (SC) samples.

Each dot represents a cell type. Cell types are marked by color code. "*R*" indicates *Pearson's* correlation coefficient, and the *p*-value obtained by two-sided *t*-test. TSO-hismap and CARD (**e**), CellTrek and SrtCT (**f**). (**g**) Example plots of co-localization of cell types in ST sections. Each dot represents a single cell, with cell types marked by color codes. cTEC vs. Fb (**top**) and mTEC vs. CD4T vs. CD8T (**bottom**). (**h**) Violin plots showing the expression levels of *CD8A* and *CD8B* in CD8 T cells (n=3,379) and CD8aa cells (n=2,717), respectively. (**i**) Violin plot showing the expression level of *GNG4* across various T cell subsets. (**j**) Venn diagram showing the overlap between the top 50 over-expressed genes within the CD8aa subset from our thymus data and the 20 signature genes of CD8aa as provided by Park et al.. DP_blast: Double positive blast cells, DP_re: Double positive rearrangement cells, cTEC: Cortical thymic epithelial cells, mTEC: Medullary thymic epithelial cells, B_memory: Memory B cells. Source data are provided as a Source Data file.



Supplementary Fig. 10 Characterization and spatial localization of T-cell receptor (TCR) clones and four T-cell subpopulations. (**a**) Bar plot showing the number of entire clonotypes and unique clonotypes per sample from the normal thymus. (**b**) Diversity of the TCR repertoire across age groups. Each point represents the TCR diversity index obtained after sampling once. Statistical analysis was performed using two-sided *t*-test. (**c**) Pie plot showing the distribution of different clone sizes in each normal thymus sample, categorized as clone size = 1, clone size = 2, and clone size ≥ 3. T cells with

clone size = 1 comprise the majority. (**d**) Scoring of spatial transcriptomics (ST) spots using marker genes from three T cell populations with different clone sizes. (**top**) Clone size = 1; (**middle**) Clone size = 2; (**bottom**) Clone size ≥ 3. (**e**) Violin plot showing the percentage of mitochondrial genes across the four T-cell subpopulations (TCRα^-TCRβ^+TCRδ^+ (n=4,165 cells), TCRα^+TCRβ^+TCRδ^+ (n=3305 cells), TCRα^-TCRβ^+TCRδ^- (n=12,691 cells), and TCRα^+TCRβ^+TCRδ^- (n=39,015 cells)). (**f, g**) Uniform Manifold Approximation and Projection (UMAP) plots showing the doublet cells identified by Scrublet (**f**) and DoubletDetection (**g**) tools, respectively. Doublets are marked in red. (**h**) Projection of the four T-cell subpopulations onto six ST sections using TSO-hismap. Each dot represents an individual T cell. (**i**) Projection of signature scores for each T-cell subpopulation generated by the *AddModuleScore* function from the Seurat package onto ST sections. Notably, only the top 20 highly expressed genes of each subpopulation were used for scoring spots. (**j**) Differentially expressed genes (DEGs) (FDR ≤ 0.01 and log2FC ≥ 0.25) identified by the *FindAllMarkers* function from the Seurat package that overlap among T-cell subpopulations. Source data are provided as a Source Data file.



Supplementary Fig. 11 Refining the annotation of T cells in the double negative (DN) and double positive (DP) stages. (**a**) Scatter plot showing the proportion of each T-cell subpopulation in relation to age. "*R*" represents *Pearson's* correlation coefficient and the *p*-value obtained using two-sided *t*-test. Each point represents one thymus sample. (**b**) Uniform Manifold Approximation and Projection (UMAP) plot colored by the four T-cell subpopulations (**left**) and cell types (**right**). The four T-cell subpopulations were down-sampled to an equal size (n=2,000). Each dot represents an individual cell, with data sources marked by color codes. (**c**) Dot plot of marker gene expression for refined cell types in the DN (**left**), DP_blast (**middle**), and DP_re (**right**) stages. The size of each dot represents the proportion of that marker expressed in a particular cell type, and the color indicates the average expression. (**d**) UMAP showing the refined annotations of DN T cells and the expression of selected marker genes. Each dot represents a single cell, with cell types marked by color codes. (**e**) UMAP plot

showing the refined annotations (i.e., DP_blast1 to DP_blast5) of DP_blast T cells, with cell types marked by color codes. (**f**) Two-dimensional representation of DP_blast cells via UMAP, as colored by the Monocle2-predicted differentiation order. Each dot represents one cell. (**g**) **Top**: Signature scores based on cell cycle-associated genes (black) and all genes (gray) for each cell. **Bottom**: Heat map of cell cycle-associated gene expression in each cell. Genes (rows) are sorted by cell cycle stages (from G1/S to M/G1). Cells are ordered from DP_blast1 to DP_blast5. (**h**) Violin plots showing the distribution of signature scores for various cell-cycle stages from DP_blast1 to DP_blast5 (n=4,609 for DP_blast1, 4,336 for DP_blast2, 3,235 for DP_blast3, 1,610 for DP_blast4, 3,180 for DP_blast5). (**i**) Violin plot showing the expression of *RAG1* from DP_blast1 to DP_blast5. (**j, k**) UMAP showing the refined annotations **(j)** and RNA velocity stream (**k**) of T cells in the DP rearrangement stage, respectively. (**l**) Scatterplots comparing the characteristics of unconventional T cells based on *CD8A* versus *CD8B* expression levels (**left**), and TCRα productive chain versus TCRβ productive chain ratio (**right**). Each dot represents a cell type, with cell types marked by color codes. DN: Double negative, DP: Double positive, DP_blast: Double positive blast cells, DP_re: Double positive rearrangement cells. Source data are provided as a Source Data file.

Supplementary Tables

Supplementary Table 1. Count of cells for each cell type present in thymic samples.

	Thy1	Thy2	Thy3	Thy4	Thy5	Thy6	Thy7	Thy8	Thy9	Thy10	Thy11	Thy12	Thy13	Thy14	Thy15	Thy16
Ery	28	46	8	6	2	0	0	0	0	1	0	1	108	1	11	5
B_naive	11	7	2	3	1	0	0	2	0	2	3	1	18	2	342	1006
B_trans	26	65	53	51	63	21	59	25	34	58	67	64	173	117	245	198
B_memory	16	41	29	39	30	6	18	11	15	24	28	15	264	35	2772	3188
Plasma	3	0	1	0	2	4	1	8	5	10	16	8	5	6	98	66
Mono	21	16	2	13	27	1	3	3	6	3	6	5	15	7	118	31
Mac	86	18	11	16	19	6	7	14	10	6	10	10	44	30	178	77
DC	48	18	5	11	29	4	6	10	10	8	1	4	14	11	47	25
pDC	0	0	2	5	22	4	4	8	3	14	8	11	3	25	20	3
Fb	365	89	13	10	7	2	0	23	7	11	3	5	15	33	1706	191
Fb_cycling	42	4	0	1	0	0	0	0	0	0	0	0	0	2	0	0
Endo	31	22	11	17	1	1	1	6	4	5	2	0	14	8	385	123
VSMC	47	21	6	11	2	0	1	10	2	0	2	1	7	2	198	39
Lymph	0	1	0	0	0	0	0	1	0	0	0	0	1	1	147	58
cTEC	275	9	3	6	4	0	0	8	1	5	4	2	20	6	3	1
mTEC	0	22	1	4	6	0	0	1	13	0	5	3	27	19	209	0
DN_early	31	2	2	11	6	1	0	6	4	4	5	7	0	0	1	0
DN_blast	164	13	24	64	50	12	13	103	34	36	32	60	8	11	1	0
DN_re	936	60	143	325	251	119	106	828	274	255	193	356	81	169	15	16
DP_blast	2044	1520	1322	1825	1655	982	1067	1379	1189	1291	867	547	517	703	282	274
DP_re	4981	5248	1954	4306	5059	4014	4182	5344	4459	4889	4624	3085	2605	4005	971	1337
abT(entry)	698	1064	516	759	1055	706	868	488	777	777	669	740	1805	740	429	704
CD4T	191	330	259	293	485	130	271	63	228	207	210	195	353	208	453	206
CD4T_mem	64	103	92	98	151	62	71	53	37	52	73	71	209	119	231	740
Treg.diff	21	69	82	111	128	22	101	31	77	68	69	54	165	80	325	252
Treg	20	62	38	30	51	7	19	12	24	39	43	24	112	54	115	92
CD8T	151	206	155	251	408	171	158	90	187	184	253	215	381	205	214	150
CD8T_mem	29	27	19	8	26	1	11	35	7	10	48	29	77	103	446	393
CD8aa	191	273	103	146	234	152	207	170	200	263	150	171	157	136	94	70
T_agonist	35	93	97	114	178	46	65	14	101	72	90	91	119	59	26	34
T_apoptosis	32	81	77	89	103	25	54	29	64	54	41	63	88	44	48	93
T_proliferating	14	11	10	19	8	1	1	7	5	3	1	3	1	2	6	13
NKT	43	20	8	15	14	6	5	26	18	13	32	15	23	26	61	75
ILC3	10	5	0	0	1	0	0	0	0	1	0	0	7	6	23	17

Supplementary Table 2. List of medullary-indicator genes for scoring spots on spatial transcriptomics (ST) sections.

hgnc_symbol	gene_biotype	chr: start-end	strand	description
CCL17	protein_coding	chr16:57404767-57416063	1	C-C motif chemokine ligand 17 [Source:HGNC Symbol;Acc:HGNC:10615]
CCL21	protein_coding	chr9:34709005-34710136	-1	C-C motif chemokine ligand 21 [Source:HGNC Symbol;Acc:HGNC:10620]
CCL22	protein_coding	chr16:57358783-57366189	1	C-C motif chemokine ligand 22 [Source:HGNC Symbol;Acc:HGNC:10621]
CCL27	protein_coding	chr9:34661880-34664048	-1	C-C motif chemokine ligand 27 [Source:HGNC Symbol;Acc:HGNC:10626]
CCR7	protein_coding	chr17:40553769-40565472	-1	C-C motif chemokine receptor 7 [Source:HGNC Symbol;Acc:HGNC:1608]
CSF2RB	protein_coding	chr22:36913628-36940439	1	colony stimulating factor 2 receptor subunit beta [Source:HGNC Symbol;Acc:HGNC:2436]
CXCL10	protein_coding	chr4:76021118-76023497	-1	C-X-C motif chemokine ligand 10 [Source:HGNC Symbol;Acc:HGNC:10637]
CXCL9	protein_coding	chr4:76001275-76007509	-1	C-X-C motif chemokine ligand 9 [Source:HGNC Symbol;Acc:HGNC:7098]
EBI3	protein_coding	chr19:4229523-4237528	1	Epstein-Barr virus induced 3 [Source:HGNC Symbol;Acc:HGNC:3129]
LAMP3	protein_coding	chr3:183122215-183163839	-1	lysosomal associated membrane protein 3 [Source:HGNC Symbol;Acc:HGNC:14582]
MS4A1	protein_coding	chr11:60455846-60470752	1	membrane spanning 4-domains A1 [Source:HGNC Symbol;Acc:HGNC:7315]
TNFRSF18	protein_coding	chr1:1203508-1206592	-1	TNF receptor superfamily member 18 [Source:HGNC Symbol;Acc:HGNC:11914]

Supplementary Table 3. Top 50 important features selected for the prediction of medullary and cortical spots.

Feature	Gain	Cover	Frequency
CCL19	0.37523767	0.0150283	0.00821337
CCL21	0.01134921	0.00999386	0.0048283
LAMP3	0.02069924	0.00712635	0.00411277
CCR7	0.02981497	0.0068516	0.00407029
MS4A1	0.03602635	0.00679345	0.00375598
IGLC2	0.00407462	0.00619404	0.0039497
CXCL9	0.00844675	0.00617161	0.00316238
TMSB10	0.0518897	0.00603766	0.0050498
CCL17	0.01105758	0.00576217	0.00280632
CCL22	0.02366576	0.00562924	0.003089
IGHG1	0.00159021	0.00535492	0.00250314
TNFRSF18	0.00267298	0.00514695	0.00410807
RPS27	0.0018709	0.00476186	0.0053528
RPS26	0.0011609	0.00470826	0.00269055
CXCL10	0.00437444	0.00447505	0.00170161
APOC1	0.00118533	0.00439136	0.00272682
RPS29	0.00178437	0.00423655	0.00249234
IGLC1	0.00096317	0.0041619	0.00280037
CSF2RB	0.00202746	0.00400322	0.00138947
IGHA2	0.00136498	0.00394037	0.00181397
RPL23A	0.00137939	0.00384199	0.0027969
EBI3	0.00184319	0.00354174	0.00146466
IGKC	0.00142291	0.00349685	0.00351412
TMEM176A	0.00072239	0.00340875	0.00118965
KRT1	0.00128035	0.00336659	0.00177135

Feature	Gain	Cover	Frequency
RPS25	0.00048247	0.00328793	0.00241362
RPS15	0.00045947	0.00327041	0.00204763
ITM2B	0.00088083	0.00323373	0.00252833
DAPP1	0.00053891	0.00319879	0.00127341
IGHG4	0.00114657	0.00317767	0.00246568
FTL	0.00061297	0.00314349	0.00245809
RPL27A	0.00057092	0.00312377	0.00213127
B2M	0.0368865	0.00301394	0.0036437
RPL41	0.00095454	0.00288931	0.00245033
RPS28	0.00132345	0.00283419	0.0025732
COL3A1	0.00046419	0.00280009	0.00130406
CD74	0.00318818	0.00279433	0.00221595
CCL27	0.00111229	0.0027722	0.00123076
RPS27A	0.00077889	0.00276776	0.00351862
RPL32	0.00076338	0.00275355	0.00280272
CST3	0.00094461	0.00273001	0.00161942
DHX58	0.00036623	0.00272153	0.00094501
HAPLN3	0.00056916	0.00271902	0.00089669
IFI30	0.00051069	0.0026639	0.00138407
KRT14	0.00051677	0.00265989	0.00138388
KRT15	0.0026463	0.00263563	0.00138255
ACTB	0.00086736	0.00263392	0.0022213
TNFSF12	0.00054237	0.00262377	0.00122168
IGHG3	0.00096267	0.00259699	0.00229099
RPS19	0.00068257	0.00258036	0.00205102

Supplementary Table 4. XGBoost model for predicting cortical and medullary spots.

Dataset	Data source	1) The spatial transcriptomics (ST) slice data from eight children contains 20,845 spots in total, with each slice having more than 2,100 spots. The data are divided into two classes of cortical and medullary spots. 2) 1,885 distance-varying genes (DVGs) and 1,271 differentially expressed genes (DEGs). 3) Three publicly available thymus ST datasets provided by Suo et al. (Science, 2022) serve as an external independent set to evaluate the model's performance.
	Dataset splits	1) ST data from our thymus Thy5 to Thy8 are taken as the training set, while Thy9 to Thy12 are used as the independent testing dataset. 2) 5-fold cross-validation is performed on the spots from training set.
	Dataset redundancy	The training and testing datasets are independent.
	Dataset availability	Yes. Additional information
Optimization	Optimization algorithm	Extreme Gradient Boosting model (XGBoost).
	Is the algorithm a meta-predictor?	No.
	Data encoding	The ST data is normalized using the *NormalizeData* function from Seurat package.
	Input parameters	max.depth = 10; objective = 'binary:logistic'; nrounds = 200; nfold = 5.
	Input features	929 genes at the intersection of DVGs and DEGs were selected as candidate features.
	Fitting method	Large number of features. Feature selection was performed to reduce the number of features used for the classification to avoid overfitting.
	Regularization method	The lasso regularization technique with 5-fold cross-validation was used to perform feature selection through least-squares regression with a penalty on the size of the estimated coefficients
	Configuration available	Yes. Additional information
Model	Interpretability	Black box. Model overview diagram refer to the Supplementary Fig. 7a.
	Output	Classification label as output: Cortical or medullary spots.
	Execution time	Less than 10 seconds in RStudio.
	Software availability	Yes. Additional information
Evaluation	Evaluation method	The model is assessed using both internal independent datasets (Thy9 to Thy12) and external thymus ST datasets (Suo et al., Science, 2022).
	Performance measures	Sensitivity/specificity/Area Under the Curve (AUC) measures for each subject/class.
	Methods comparison	No.
	Evaluation confidence	No.
	Evaluation availability	Yes. Additional information (Supplementary Fig. 7b, c)

Supplementary Table 5. sgRNA sequences for generation of knockout mice.

Target gene	sequences (5'to3') gRNA
Cd99	AGAGCTCTGACGAACGCTAC
Elovl4	GAGTTACCATGCAATCACAC
Elovl4	TTCGACATTTCGTGAGCCAC

REFERENS

1. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S: **Fast, sensitive and accurate integration of single-cell data with Harmony**. *Nat Methods* 2019, **16**(12):1289-1296.
2. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park J-E: **BBKNN: fast batch alignment of single cell transcriptomes**. *Bioinformatics* 2020, **36**(3):964-965.