



Supplementary Material

Correcting Cross-Cell UMI Bias for Improved Single-Cell Transcriptomics

Rulin Hua¹, Wenjing Yan², Xiaoyan Liu³, Suyang Xuan⁴, Xuejuan Li⁵, Shafiq Ur Rahman¹, Dekang Lv^{4*} and Feng Zheng^{1,5*}

¹Advanced Institute for Medical Sciences, Dalian Medical University, Dalian, 116044, China

²Dalian Center for Disease Control and Prevention, Dalian, 116021, China

³Department of Nephrology, the Second Hospital of Dalian Medical University, Dalian, 116023, China

⁴Institute of Cancer Stem Cell, Dalian Medical University, Dalian, 116044, China

⁵Wuhu Hospital and Health Science Center, East China Normal University, Shanghai, 200241, China

Rulin Hua and Wenjing Yan contributed equally.

Supplementary Table I. Overview of public scRNA-seq data for analysis.

Reference	GEO accession no.	Protocol	Cell number	Species	Description
Garrido-Trigo <i>et al.</i> , 2023	GSE214695	10X	46,698	Human	Inflammatory bowel diseases colon
Lake <i>et al.</i> , 2018	GSE97930	snDrop-Seq	10,319	Human	Frontal cortex
The Tabula Sapiens Consortium, 2022	GSE201333	10X	456,101	Human	Tabula Sapiens
The Tabula Muris Consortium, 2018	GSE109774	10X	55,656	Mouse	Tabula Muris
Trzebanski <i>et al.</i> , 2024	GSE261003	10X	4,705	Mouse	Bone marrow myeloid cells
Zheng <i>et al.</i> , 2017	SRX1723926	10X	68,579	Human	Peripheral blood mononuclear cells
Gan <i>et al.</i> , 2024	GSE223309	10X	31,657	Mouse	Brain metastatic microenvironment

References can be found in the main article.

Supplementary Table II. Runtimes for different normalization methods.

Dataset	log (CP10K+1)	Scrans-form	log (LC-UMI+1)
68k PBMC	29 s	703 s	506 s
32k brain metastatic cells	21 s	294 s	138 s

The table presents the runtimes of different normalization methods on the Zheng 68k PBMC dataset and the mouse brain metastatic microenvironment dataset. All runtimes were measured on a server with 256 GB RAM and 64 CPU threads at 2.4 GHz.

* Corresponding author: dekanglv@126.com, Fzheng@hsc.ecnu.edu.cn
0030-9923/2026/0004-1835 \$ 9.00/0



Copyright 2026 by the authors. Licensee Zoological Society of Pakistan.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Supplementary Table III. Comparison of the number of marker genes identified using log(CP10K+1) and log(LC-UMI+1) normalization across different fold change thresholds. This table summarizes the number of marker genes identified using log(CP10K+1) and log(LC-UMI+1) normalization in the 68k PBMC and brain metastatic microenvironment datasets across a range of log2 fold change thresholds. Marker genes were classified into CP10K-specific (identified only by log(CP10K+1)), LC-UMI-specific (identified only by log(LC-UMI+1)), and Shared (identified by both methods). In the cell type column, “combined” represents the total number of marker genes across all cell types in each dataset.

68k PBMC				
log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type
0.25	552	276	1120	Combined
0.5	304	93	816	Combined
1	105	24	552	Combined
2	26	4	193	Combined
3	2	1	84	Combined
4	0	0	29	Combined
5	0	0	13	Combined
6	0	0	5	Combined
0.25	252	2	77	CD14+ Monocyte
0.5	214	2	76	CD14+ Monocyte
1	76	1	66	CD14+ Monocyte
2	19	0	20	CD14+ Monocyte
3	0	0	10	CD14+ Monocyte
4	0	0	2	CD14+ Monocyte
5	0	0	2	CD14+ Monocyte
6	0	0	0	CD14+ Monocyte
0.25	28	0	63	CD19+ B
0.5	10	0	46	CD19+ B
1	3	0	30	CD19+ B
2	2	0	6	CD19+ B
3	2	0	4	CD19+ B
4	0	0	1	CD19+ B
5	0	0	0	CD19+ B
6	0	0	0	CD19+ B
0.25	32	1	414	CD34+
0.5	32	0	411	CD34+
1	15	0	349	CD34+

Table continues on next column.....

log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type
2	4	0	137	CD34+
3	0	0	57	CD34+
4	0	0	23	CD34+
5	0	0	11	CD34+
6	0	0	5	CD34+
0.25	3	16	39	CD4+ T cell
0.5	0	1	9	CD4+ T cell
1	0	0	0	CD4+ T cell
2	0	0	0	CD4+ T cell
3	0	0	0	CD4+ T cell
4	0	0	0	CD4+ T cell
5	0	0	0	CD4+ T cell
6	0	0	0	CD4+ T cell
0.25	206	0	165	CD56+ NK
0.5	45	0	116	CD56+ NK
1	10	0	52	CD56+ NK
2	1	0	15	CD56+ NK
3	0	0	7	CD56+ NK
4	0	0	1	CD56+ NK
5	0	0	0	CD56+ NK
6	0	0	0	CD56+ NK
0.25	30	1	71	CD8+ T cell
0.5	2	0	12	CD8+ T cell
1	0	0	2	CD8+ T cell
2	0	0	0	CD8+ T cell
3	0	0	0	CD8+ T cell
4	0	0	0	CD8+ T cell
5	0	0	0	CD8+ T cell
6	0	0	0	CD8+ T cell
0.25	1	256	291	Dendritic
0.5	1	90	146	Dendritic
1	1	23	53	Dendritic
2	0	4	15	Dendritic
3	0	1	6	Dendritic
4	0	0	2	Dendritic
5	0	0	0	Dendritic
6	0	0	0	Dendritic
Mouse brain metastatic microenvironment				
0.25	3588	3631	9757	combined
0.5	3444	3597	9680	combined
1	2627	3292	9123	combined

Table continues on next page.....

log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type
2	1165	1561	7148	combined
3	534	569	4920	combined
4	212	205	3328	combined
5	79	75	2199	combined
6	25	21	1403	combined
0.25	367	19	507	ASTRO
0.5	332	19	501	ASTRO
1	245	16	462	ASTRO
2	110	11	384	ASTRO
3	49	3	273	ASTRO
4	2	1	146	ASTRO
5	0	0	52	ASTRO
6	0	0	19	ASTRO
0.25	78	1	56	BAM
0.5	77	1	56	BAM
1	54	1	55	BAM
2	25	1	42	BAM
3	13	0	30	BAM
4	4	0	25	BAM
5	2	0	19	BAM
6	0	0	9	BAM
0.25	103	0	52	BC
0.5	103	0	52	BC
1	100	0	52	BC
2	75	0	52	BC
3	51	0	50	BC
4	26	0	40	BC
5	14	0	38	BC
6	9	0	31	BC
0.25	102	168	378	BG
0.5	97	164	372	BG
1	58	130	351	BG
2	11	75	252	BG
3	6	40	177	BG
4	1	1	96	BG
5	0	0	43	BG
6	0	0	19	BG
0.25	52	242	233	BMDM
0.5	51	242	233	BMDM
1	26	226	227	BMDM
2	4	115	131	BMDM

Table continues on next column.....

log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type
3	1	55	80	BMDM
4	1	31	56	BMDM
5	1	18	46	BMDM
6	0	6	33	BMDM
0.25	26	411	645	CPEC
0.5	23	411	645	CPEC
1	12	404	645	CPEC
2	0	165	494	CPEC
3	0	60	223	CPEC
4	0	23	130	CPEC
5	0	9	91	CPEC
6	0	3	71	CPEC
0.25	163	190	1079	EpC
0.5	151	189	1079	EpC
1	109	145	1053	EpC
2	33	83	883	EpC
3	11	42	717	EpC
4	5	16	557	EpC
5	1	7	436	EpC
6	0	3	345	EpC
0.25	252	8	436	GN
0.5	252	8	436	GN
1	247	8	436	GN
2	139	7	404	GN
3	66	5	349	GN
4	31	4	291	GN
5	11	1	216	GN
6	1	0	144	GN
0.25	131	56	371	HpC
0.5	131	56	371	HpC
1	123	55	371	HpC
2	69	36	316	HpC
3	34	20	204	HpC
4	14	10	141	HpC
5	7	2	104	HpC
6	3	1	65	HpC
0.25	171	111	950	MG
0.5	137	109	922	MG
1	55	90	789	MG
2	8	39	475	MG
3	0	8	261	MG

Table continues on next page.....

log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type
4	0	1	94	MG
5	0	0	29	MG
6	0	0	1	MG
0.25	561	0	169	NB
0.5	561	0	169	NB
1	531	0	169	NB
2	229	0	161	NB
3	101	0	134	NB
4	41	0	111	NB
5	12	0	81	NB
6	2	0	42	NB
0.25	191	0	111	NEUT
0.5	191	0	111	NEUT
1	189	0	111	NEUT
2	158	0	111	NEUT
3	81	0	105	NEUT
4	36	0	85	NEUT
5	20	0	60	NEUT
6	10	0	51	NEUT
0.25	1	1476	1263	NIPC
0.5	1	1476	1263	NIPC
1	0	1466	1263	NIPC
2	0	674	1067	NIPC
3	0	170	499	NIPC
4	0	52	286	NIPC
5	0	17	158	NIPC
6	0	1	39	NIPC
0.25	94	27	197	NK
0.5	94	27	197	NK
1	83	27	197	NK
2	28	26	172	NK
3	11	18	138	NK
4	7	14	119	NK
5	2	9	98	NK
6	0	6	84	NK
0.25	107	29	283	OEC
0.5	103	28	283	OEC
1	79	21	271	OEC
2	42	11	233	OEC
3	24	1	188	OEC
4	9	1	125	OEC

Table continues on next column.....

log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type
5	2	0	79	OEC
6	0	0	49	OEC
0.25	40	516	699	OLIG
0.5	34	516	699	OLIG
1	12	470	699	OLIG
2	1	190	544	OLIG
3	0	77	370	OLIG
4	0	27	253	OLIG
5	0	9	152	OLIG
6	0	0	94	OLIG
0.25	419	178	943	OPC
0.5	396	159	909	OPC
1	226	88	661	OPC
2	51	41	330	OPC
3	19	24	241	OPC
4	11	9	162	OPC
5	4	1	104	OPC
6	0	0	56	OPC
0.25	85	6	95	PER
0.5	85	6	95	PER
1	85	6	95	PER
2	56	5	91	PER
3	27	4	66	PER
4	9	1	40	PER
5	1	0	23	PER
6	0	0	11	PER
0.25	432	6	472	VEC
0.5	429	6	472	VEC
1	276	6	463	VEC
2	85	4	381	VEC
3	24	0	296	VEC
4	9	0	211	VEC
5	0	0	137	VEC
6	0	0	90	VEC
0.25	85	145	417	VLC
0.5	69	138	414	VLC
1	28	94	362	VLC
2	9	51	295	VLC
3	6	27	251	VLC
4	4	11	191	VLC
5	2	2	131	VLC

Table continues on next page.....

log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type					
6	0	1	92	VLC	2	32	27	330	VSMC
0.25	128	42	401	VSMC	3	10	15	268	VSMC
0.5	127	42	401	VSMC	4	2	3	169	VSMC
1	89	39	391	VSMC	5	0	0	102	VSMC
					6	0	0	58	VSMC

Supplementary Table IV. Marker genes identified by log(CP10K+1) and log(LC-UMI+1) normalization in two scRNA-seq datasets. This table presents representative marker genes identified by log(CP10K+1) and log(LC-UMI+1) normalization in the 68k PBMC and brain metastatic microenvironment datasets using a log2 fold change (log2FC) threshold > 0.5. For each cell type, marker genes were classified into CP10K-specific (identified only by log(CP10K+1)), LC-UMI-specific (identified only by log(LC-UMI+1)), and Shared (identified by both methods). When the number of marker genes in a given category exceeded 20, the top 20 genes ranked by log2FC were retained. “None” indicates that no marker genes met the specified criteria in that category.

68k PBMC			
Cell type	CP10K-specific	LC-UMI-specific	Shared
CD14+ Mono-cyte	FCER1G, SPI1, IFI30, CTSS, LGALS1, S100A11, PYCARD, CFP, S100A4, LGALS2, TSPO, IGSF6, BLVRB, GSTP1, COTL1, MS4A7, HCK, S100A6, NPC2, CEBPD	FCGR3A, IFITM2	S100A9, S100A8, AIF1, LST1, TYROBP, FCN1, SAT1, CFD, TYMP, PSAP, FTL, SERPINA1, TMEM176B, RP11-290F20.3, CD14, IFITM3, STXBP2, NEAT1, PILRA, LINC00936
CD19+ B	HLA-DRA, CD74, IRF8, FCGR2B, PKIG, LAT2, HHEX, STX7, LTB, LAPTM5	None	CD79A, CD79B, IGLL5, TCL1A, MS4A1, CD37, LINC00926, HVCN1, VPREB3, FCER2, NCF1, EAF2, MZB1, SPIB, CYB561A3, BANK1, PPAPDC1B, HLA-DOB, BLK, CD40
CD34+	ARPC1B, TIMP1, CD68, FTH1, TUBA1B, MYL6, ACTB, SNX3, TMSB4X, ALDOA, GAPDH, SUPT4H1, C7orf50, POLR2E, LYN, VASP, ATF4, MAP1LC3B, PARK7, RTN4	None	PPBP, PF4, GNG11, SDPR, CLU, GPX1, HIST1H2AC, ACRBP, NRG1, NCOA4, TUBB1, GP9, SPARC, RUFY1, TREML1, MYL9, PTCRA, CMTM5, ITGA2B, FERMT3
CD4+ T cell	None	LTB	CD4, IL32, AQP3, GSTK1, CORO1B, PLP2, TNFRSF4, TRADD, MAL
CD56+ NK	FCGR3A, RHOC, ABI3, CD300A, ITGB2, IFITM2, ID2, CLIC1, MIR142, AOA1, CTSC, PFN1, FAM49B, ARPC2, PPP1CA, DBI, ANXA1, OSTF1, CXXC5, CDK2AP2	None	GNLY, NKG7, GZMB, FGFBP2, GZMH, CST7, PRF1, CLIC3, GZMA, CCL5, CTSW, SPON2, HOPX, PTGDS, CCL4, CD63, PRSS23, IGFBP7, MATK, CD247
CD8+ T cell	IL7R, LDHB	None	GZMK, CD8B, CD27, CD8A, DUSP2, CD3E, NOSIP, CCR7, CD3D, TMEM66, LEF1, PIK3IP1
Dendritic	ITM2C	HLA-DRA, CD74, FTH1, FCER1G, COTL1, GSTP1, CTSS, S100A4, LGALS1, CYBA, SPI1, S100A11, MYL6, S100A6, PYCARD, ARPC1B, CFP, ACTB, NPC2, GAPDH	LYZ, CST3, HLA-DRB1, HLA-DPB1, HLA-DPA1, HLA-DQA1, HLA-DRB5, FCER1A, HLA-DQB1, HLA-DQA2, HLA-DMA, CPVL, GRN, FCGRT, ANXA2, CLEC10A, HLA-DMB, C1QA, AMICA1, LY86

Table continues on next page.....

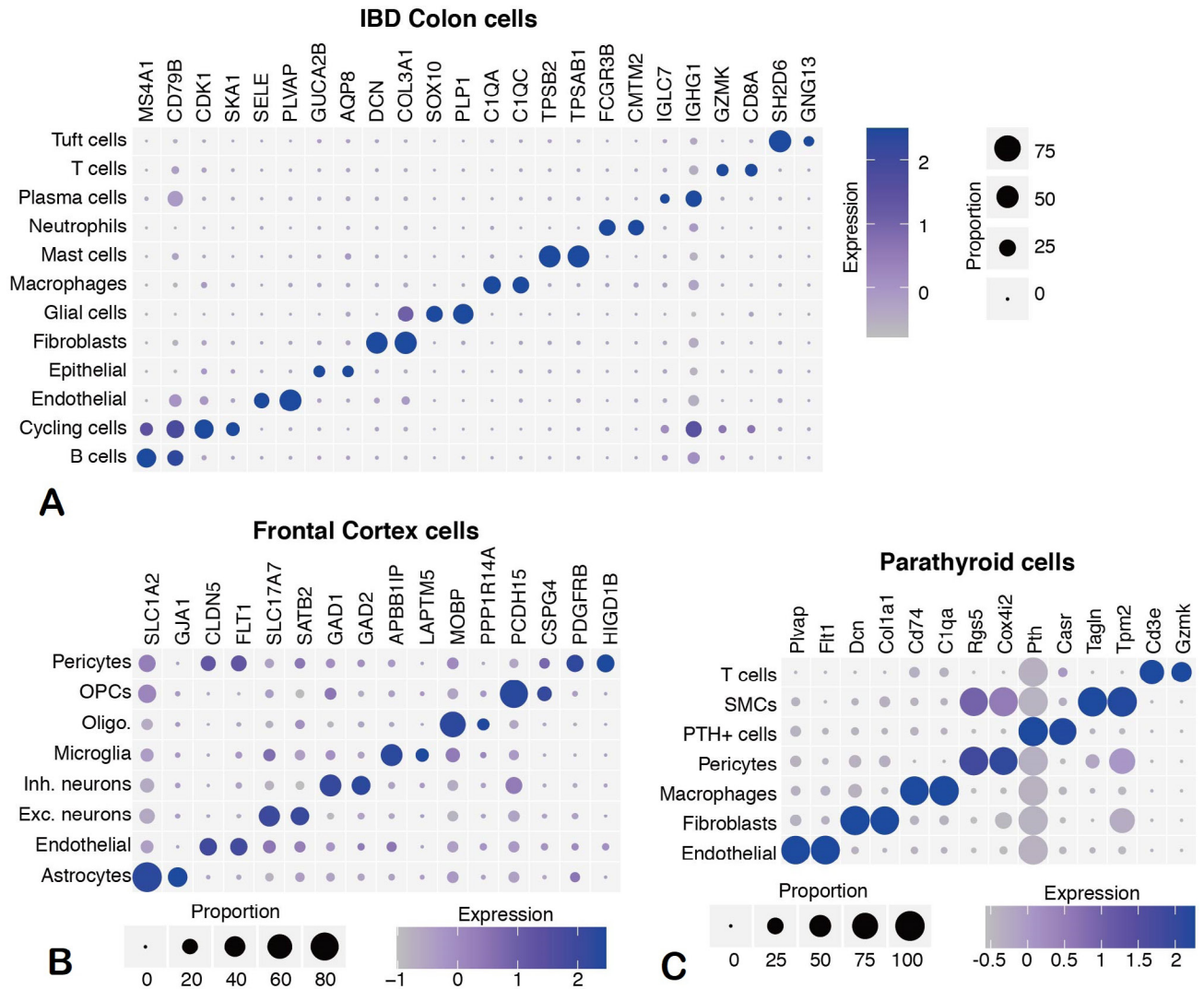
Cell type	CP10K-specific	LC-UMI-specific	Shared
Mouse brain metastatic microenvironment			
ASTRO	CAR8, KCNK1, GABRB1, NAT8, ALDOC, HHATL, TMEM229A, ADRA1A, RFX4, KCNJ16, NR-CAM, PAMR1, BTBD17, FOLH1, SLC14A1, SCG3, ADCY2, PIPOX, NPAS3, USP51	RGS7, SLC30A10, CHIL1, SEMA4A, 6430573F11RIK, SORL1, NRXN2, EFR3B, COLGALT2, GPD2, TMEM136, MAGI3, 2900076A07RIK, MALAT1, ANKS1, MCUR1, DENND4C, NFAT5, DOCK4	C030018K13RIK, SLC6A11, PDZPH1, GPC5, HAPLN1, 2900052N01RIK, GRIN2C, FEZF2, PPP1R3G, IGSF1, CYP4F15, KIRREL2, PRSS50, NR2E1, LHX2, GABBR2, GM266, DIO2, SL-C1A2, GM45650
BAM	H2-EB1, H2-AA, CD74, H2-AB1, CXCL1, TGFB1, ARHGAP19, TRPV4, COL14A1, DSE, HPSE, MSR1, C2, MS4A6B, MDFIC, TMEM8, IFI204, APOBEC1, CLEC4A2, SMAGP	ACP5	LYVE1, MRC1, CD163, FCNA, CCL8, PF4, PLA2G2D, F13A1, CLEC10A, MS4A4A, LILRA5, CBR2, CLEC4N, FOLR2, SAA3, CCL24, MS4A7, SER-PINB8, AOA, CCL7
BC	FCRLA, LY6D, SPIB, P2RY10, CD2, IL12A, GPR171, CIITA, STAP1, B3GNT5, CLEC2I, CHST3, RHOF, IFI203, ARHGAP15, MNDAL, SER-PINB1A, CD69, AC125149.3, KYNU	None	IGLC1, FCMR, MS4A1, IGLC2, FCER2A, IGKC, CD79A, GM43291, PAX5, 4930597A21RIK, CD19, POU2AF1, BLK, CD79B, MZB1, GM8369, IGLC3, CXCR5, VPBEB3, IGHD
BG	GM9945, SEPT4, ABLIM2, SOX1, GPM6B, ADRA1B, KLHL13, SCD2, CDK5RAP2, WIP1, NDST1, TPMT, NEK8, IRS1, IQSEC1, ATRN, TSPAN33, IMPAD1, HAUS3, TB-C1D7	KCNC3, BTBD17, EN2, NAT8, FOLH1, GABRB1, ALDH1A1, ALDOC, NEGR1, SLC14A1, RFX4, SCG3, ADCY2, CYP2J9, PHYHIPL, KCNJ16, PFKM, ADRA1A, NRCAM, FXYD7	GM3455, CD70, GDF10, TDO2, GRIA4, DAO, SSTR2, MYBPC1, SYNDIG1, GRIA1, GREB1L, WIF1, MSX2, CPS-F4L, SOWAHA, FRRS1L, A2M, SYT10, CDH22, COX7B2
BMDM	ACE, E2F2, HAAO, OBFC1, ALD-H3B1, LPL, SGK3, VRK1, FEM1C, ACTG1, GM4799, RNH1, SCAND1, SHFM1, TAF10, HSPA8, AKT1, TFEB, EIF3K, UBE2A	SIRPB1C, LILRA6, MEFV, SIRPB1B, PIRA2, FPR2, FLT3, SLPI, GDA, STAP1, DPP4, EMILIN2, CIITA, H2-AA, LSP1, PILRB2, H2-EB1, MCEMP1, H2-AB1, CD300LB	EAR1, EAR2, CHIL3, MS4A8A, CD300E, ADGRE4, TREML4, IFITM6, CD177, GM9733, GPR141, NAPSA, CD209A, RYR1, F10, GPR132, CCR2, LTB4R1, CD300LD, ATP8B4
CPEC	FAM92A, ENY2, CBWD1, ACAT1, POLR2K, PPIA, SUMO3, DNLZ, SEC61G, RARS2, PSMA2, ADSL, BTBD1, NDUFA10, HBS1L, ZRSR2, PSMG1, SLIRP, ANAPC4, FKBP8	ATP2B3, PLK5, MSX1, RBP1, SLC29A4, ACE, CIDEA, CLDN1, 2810433D01RIK, KRT1, CNTNAP2, SEMA3B, FAM161A, IGFBP2, OCIAD2, KCNK1, PRKCQ, FAM160A1, TRPV4, MYRIP	PRR32, KCNE2, SULT1C2, TTR, DEFB9, INS2, DEFB11, COX8B, HTR2C, IGFN1, WDR72, KL, RP24-233B16.2, SOSTDC1, MRAP2, CRB3, WDR86, SLC16A8, 9330185C12RIK, EPN3
EpC	CIDEA, FAM161A, TPPP3, SEM-A3B, CHAF1A, PSMC3IP, CCDC34, B4GALNT4, 2310030G06RIK, EZR, AIF1L, CCDC24, FUOM, ARL6, KCTD15, NRBP2, CEP89, HIST1H4I, BOK, ETFA	FAM216B, HDC, ACBD7, RIT2, WNT4, CCDC114, KCTD8, 2010300C02RIK, LY6H, DZANK1, CYB561, FAM84B, KCNMB4, GM42517, NNAT, TSPYL5, SGSM1, PTPRU, PLTP, WDR34	SNTN, AOC1, STOML3, TMEM212, SLC26A3, KRT15, CCDC153, LYG2, GOT1L1, GM44109, GM20661, GM21149, SLC14A2, FRMPD2, DNAH3, UOX, AKAP14, GM12695, CCDC33, ZFP474
GN	ATP2B3, KCTD8, RIT2, DLX6, EPHA3, DLGAP3, ARFGEF3, GALNT13, KCNH7, EN2, TRNP1, GM42372, DLG2, TMEM108, PNMAL2, 2010300C02RIK, JPH4, CNTNAP2, 2900011O08RIK, BEX2	MEG3, ST8SIA3, SYT6, CSRNP3, AUTS2, IPCEF1, DMXL2, RAPGEF2	SCGN, CPA6, RPRML, KCNF1, GM42912, CALB2, GRIN2A, ZMAT4, GABRA5, SYT1, GABRA6, FRMD7, SL-C17A7, NRSN2, GRM2, NPPA, CBLN1, DLGAP2, SNHG11, CBLN3

Table continues on next page.....

Cell type	CP10K-specific	LC-UMI-specific	Shared
HpC	FAM216B, MSX1, ACBD7, CLDN1, TMEM150C, CCDC114, SLC29A4, PLK5, CDON, CYB561, OCIAD2, PTPRU, BEX4, TSPYL5, FAM84B, SERINC2, WDR34, TUBB4B, GALM, KIF21A	3632451O06RIK, DBNDD1, EMB, SGMS2, ARFGEF3, GM42372, PNMAL2, BC067074, A730017C20RIK, SMIM5, BEND7, PNCK, CACNB2, NDN, NRG2, PCGF2, 4430402I18RIK, CCDC24, ANXA2, SDC2	SPDEF, SPP2, MYH2, FERD3L, SSPO, BC048502, RP23-477K6.1, LDLRAD2, D930020B18RIK, RASEF, CAR3, MYH1, WFDC2, NPSR1, SCT, SULT6B2, CREB3L4, MAB21L2, CUZD1, CAR1
MG	SGK1, FMNL1, EVI2A, RBM47, ATP8A2, GARNL3, ELMO1, ZCWPW1, DOCK10, SLC15A4, NRM, FAM134B, KHK, PIK3R6, ARL10, UAP1L1, FKBP5, KLF10, NEK6, SCOC	H2-OB, TRPM2, CSF3R, CD22, SLAMF6, NFKBID, PIL-RA, CCRL2, C5AR1, RAC2, NFAM1, IGSF6, GLIPR1, SNX20, GPR65, SLC16A3, MIR142HG, PTAFR, LILRB4A, LRRC25	CCR6, HSPB3, SIGLECF, ADORA3, SLC2A5, IL1A, 2610528A11RIK, AFP, CXCL13, P2RY12, CST7, CRYBB1, TNF, GPR34, TMEM119, SIGLECH, GPR84, CRYBA4, LAG3, HEXB
NB	SLIT1, ELAVL4, TOX3, FOXO6, LMO3, MYCN, DBNDD1, YPEL1, ST8SIA3, TTC9B, GNG3, MEG3, HMGB3, PAK7, PAK3, DBN1, CDCA7, FOXP2, MFAP2, STMN1	None	IGFBPL1, ISLR2, PROKR2, DCX, SP9, TUBB3, SOX11, TIAM2, SP8, DLX5, DRAXIN, DLX1, RNF165, ELAVL2, DLX2, SCRT2, STMN2, ZFP57, MGAT5B, SRRM4
NEUT	SIRPB1B, LILRA6, MEFV, GDA, SELL, HDC, FPR2, SIRPB1C, LTB, SLPI, CD300LB, PIRA2, IL18RAP, MCEMP1, PILRB2, CD44, EMILIN2, PSTPIP1, FGR, LSP1	None	BC100530, STFA2L1, MRGPRA2B, RETNLG, ASPRV1, GM5483, STFA2, IL1F9, CXCR2, NLRP12, S100A9, DHRS9, S100A8, 4833407H14RIK, WFDC21, TACSTD2, IL1R2, KLRA17, HP, 4732465J04RIK
NIPC	DHPS	SLIT1, TTC9B, MYCN, EPHA3, SAPCD2, HMGB3, FOXO6, CDCA7, DLGAP3, PAK3, DLX6, ELAVL4, KCNH7, B3GNT5, TOX3, MFAP2, LMO3, FRAT2, FOXG1, YPEL1	HELT, GSX2, GSX1, GM29260, NPTX2, INSM1, MXD3, ASCL1, DLL3, PSRC1, FAM64A, LRR1, RRM2, NEIL3, TTK, PBK, SKA1, 2810417H13RIK, HIST-1H2AF, PIF1
NK	FLT3, DPP4, THY1, RUNX3, PRK-CQ, CORO2A, ITGA4, OSBPL3, D16ERTD472E, A530032D15RIK, IFI47, CKM, FAM117A, FOXP4, B4GALNT1, SEPT9, VOPP1, PTGER4, REL, PHF11A	SPIB, LY6D, FCRLA, P2RY10, CD2, GPR171, SELL, LTB, IL18RAP, RHOF, CLEC2I, MNDAL, IFI203, SERPINB6B, KYNU, CD69, DUSP2, ST-8SIA4, HCST, IL2RG	CMA1, KLK1B27, KLK1, D13ERT-D608E, DNNT, PRF1, NCR1, SH2D1A, GZMA, IL2RB, GZMB, CLNK, NKG7, KLRG1, KLRI2, KLRB1C, KLRC1, GATA3, TRBC1, XCL1
OEC	ERBB3, PRIMA1, 2810433D01RIK, APOD, NCALD, SP5, BC067074, FIGN, SLC2A13, UBE2QL1, ZDH-HC2, SCRN1, AHCYL2, SLC25A13, RAP1GAP, PLEKHF1, TSPAN15, FXYD7, NCAM2, PHGDH	HMGCS2, PLSCR2, 9330159F19RIK, GADD45A, CEBPD, ATP8A2, CASKIN2, BCAR3, NECTIN3, STARD13, FGF2, ISG15, OASL2, TMX4, IRGM2, DDX60, DENND1B, SELENOP, PIGH, RNF215	GM45385, NKX3-1, FOXD3, CLCA3A1, GM13499, GM12688, GPD4, OLFR32, FRZB, AQP1, SLC30A2, NPY, PRSS56, GM44691, IL13RA2, PSMB11, MYOZ1, COL28A1, CDH3, FNDC1
OLIG	RNF7, POP4, DCPS, TRIM35, GABARAPL2, MED30, MIEN1, FAM204A, PFDN6, TADA2A, MRFAP1, DNAJC19, DCTN2, PCNP, TBCC, CWC15, PPP2CA, TRAPPC2, TMEM60, FXR2	IL12A, POU3F1, PRIMA1, TMEM150C, JPH4, SERPIN-B1A, BFSP2, ELAVL3, SOX10, ERBB3, KLHL4, TPPP3, APOD, FIGN, PAK7, GM9945, NAP1L5, PPP1R14A, ANKS1B, SEPT4	C030029H02RIK, ERMN, KLK6, OPA-LIN, LY6G6F, CYP2J12, TMEM125, GM45624, MOG, GJC2, ITGAD, MOBP, FA2H, HAPLN2, CLDN11, GM4593, NIPAL4, MYRF, NFE2L3, TNNI1

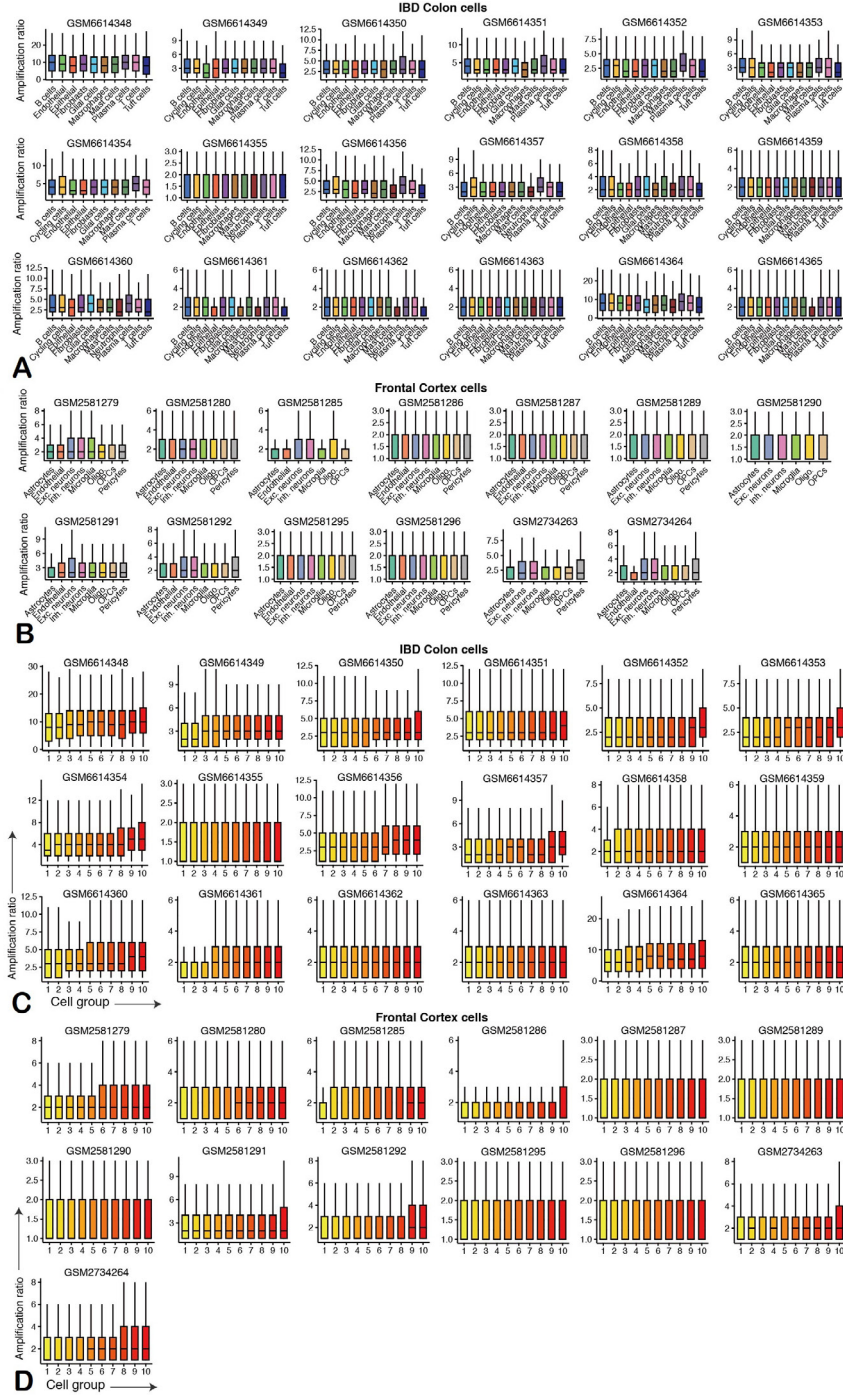
Table continues on next page.....

Cell type	CP10K-specific	LC-UMI-specific	Shared
OPC	3632451O06RIK, SAPCD2, BFSP2, POU3F1, SOX10, KCNK10, ANKS1B, KRT1, RP24-286J21.7, GM42517, EPHB2, GPR45, SH3D19, UNC5C, MYBL1, ZFP462, ZWILCH, MANEAL, DPYD, MAP6	GNG3, GFRA2, CHST3, TMEM108, 2900011O08RIK, CNIH2, NCALD, CAR8, ZDHHC2, NETO2, TMEM35A, REEP2, ASIC1, RAP1GAP, D630045J12RIK, ARXES1, WASF1, PLPPR2, XKR6, NAV1	PYY, PNLIP, GPR17, IQCJ, NEU4, 3110035E14RIK, VGF, SSTR1, ANKRD34C, A930009A15RIK, C1QL2, A930003O13RIK, TNFR, HMGA2, FAM19A2, PRKG2, ZFP488, LHFPL3, CHRNA4, 5730559C18RIK
PER	CASQ2, HIC1, GFRA2, COX4I2, LAMC3, FAP, CHP2, TMEM45A, MND1, NID1, SLC30A10, ARH-GEF17, KCTD1, IMPA2, GRTP1, SL-C19A1, INPP4B, HRCT1, GUCY1A2, HSPA12A	THY1, CXCL1, BMP1, OLFML2B, IL1R1, SERPINE2	ABCC9, KCNJ8, NODAL, ACE2, SERPINA1D, SERPINA1B, ATP13A5, ENPEP, PLA1A, REM1, ART3, GGT1, CYGB, VTN, EPHX3, SLC5A5, SL-C16A12, RGS5, COL5A3, CD248
VEC	STRA6, SERPINB6B, SGMS2, SL-C16A9, MYLPP, CRISPLD1, MYRIP, KLHL4, PLTP, HMGCS2, BEND7, ST8SIA4, SH3BP4, NDRG1, NET1, LIMS2, TANC1, PRRG1, SOCS2, NHSL2	PRKD2, DOCK6, SH2D3C, SMAGP, SWAP70, ETS2	AU021092, GM694, TMEM252, SCGB3A1, SLC38A5, SOX18, MYCT1, ADGRL4, RP23-381H23.2, LY6C1, ROBO4, CYR1, SOX17, PTPRB, ADGRF5, TEK, SLC01A4, FOXQ1, SHE, FLT1
VLC	RBP1, WNT4, IGFBP2, FAM160A1, ALDH1A1, PARD6G, ITGA3, GSTT1, COPZ2, FDX1, NEDD4, CMC2, CKAP4, AIFM2, ARF2, ARF-GAP3, ACADSB, TBGR1, TBC1D8B, TGDS	GALNT13, FOXP2, CRISPLD1, DLG2, LAMC3, FAP, HIC1, CHP2, SLC16A9, SLC2A13, STRA6, TMEM45A, DDR2, PAMR1, NID1, ANXA1, DACT1, UNC5C, COL14A1, GPX8	PRG4, SLC47A1, PRDM6, PTGDR, LUM, DCN, WNT16, CARD14, ASGR1, GM5860, MPZL2, MFAP5, LYPD2, SLC22A6, ALDH1A2, SLC26A7, VSX1, AOX3, FAM180A, NOV
VSMC	SMIM5, PPP1R14A, DDR2, CACNB2, MPP7, PNCK, MYO1D, TTLL7, RGS7BP, OLFML2B, PCP4L1, SDC2, DACT3, FBXL2, OAZ2, EVA1C, HSPB6, PRKAG2, TUBA1C, FXYD1	COX4I2, CASQ2, EBF1, MND1, CACNA2D1, ARHGEF17, SLC19A1, HRCT1, LIMS2, SCN3A, INPP4B, PEAR1, GUCY1A2, NET1, PRKAR2B, RGS6, UTRN, CAV1, SYDE1, PLP2	RERGL, TAGLN, PLN, MAP3K7CL, MYH11, ACTA2, RP23-37D14.4, LMOD1, XIRP1, GPR20, MYOCD, FBXL22, TPM2, CNN1, NXNL2, MUSTN1, MYL9, OLF558, SCN4B, HSPB7

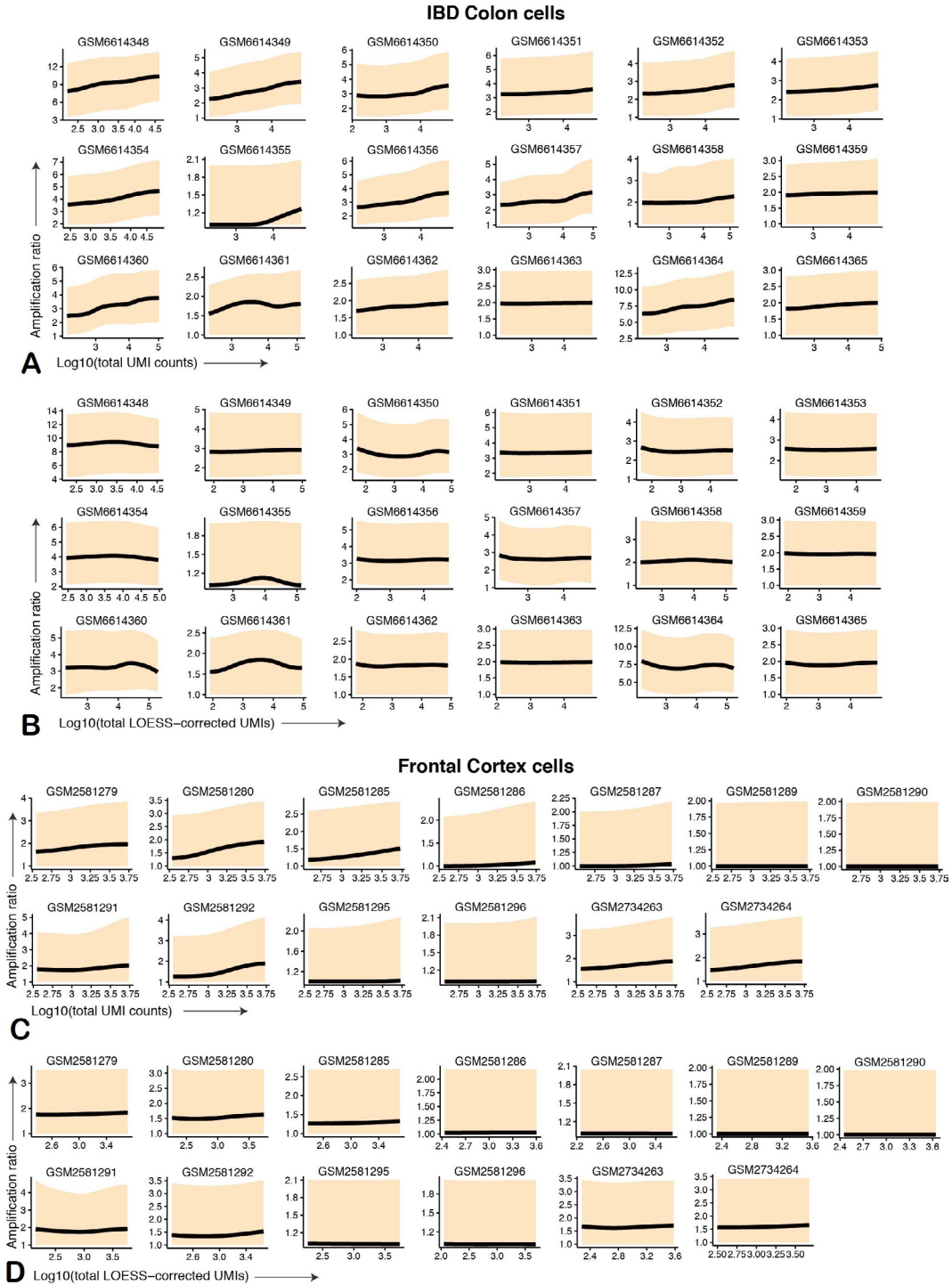


Supplementary Fig. 1. Marker gene expression across cell types in the human IBD colon, human frontal cortex, and rat parathyroid datasets.

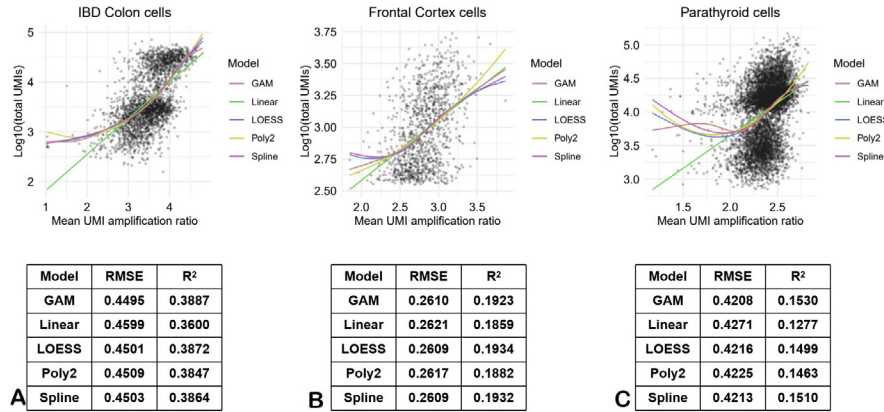
Bubble plots illustrating the expression levels and proportions of canonical marker genes across cell types in the (A) human IBD colon, (B) human frontal cortex, and (C) rat parathyroid datasets. Bubble size denotes the proportion of cells expressing each gene, while color intensity reflects the scaled average expression level.



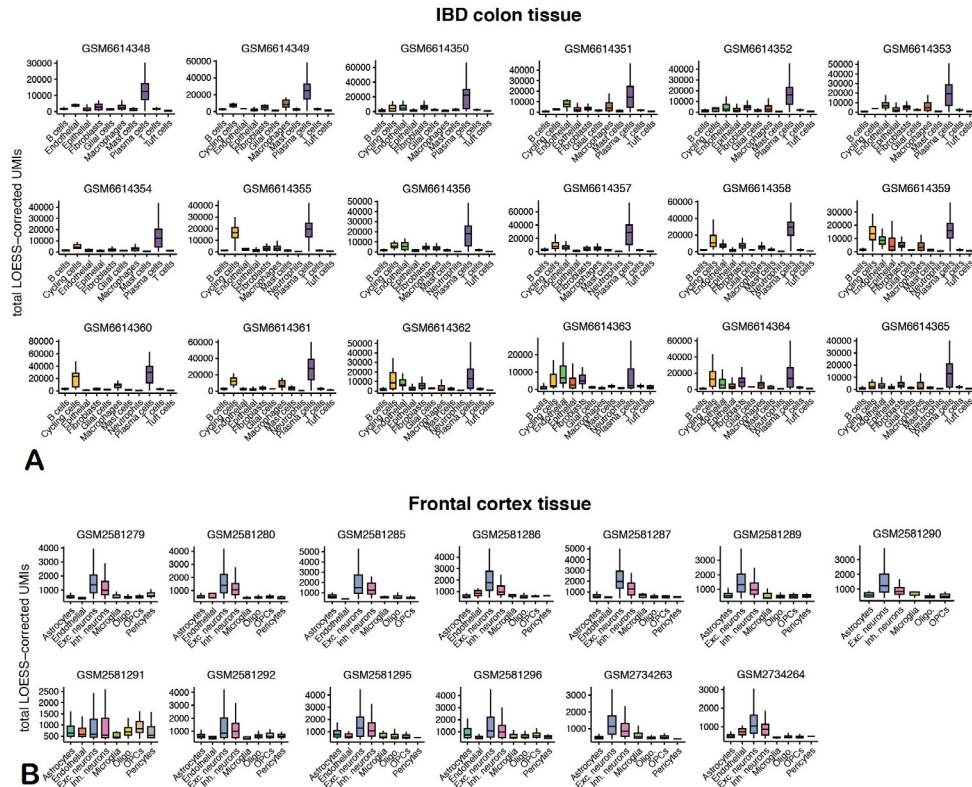
Supplementary Fig. 2. UMI amplification ratio distribution across samples in the human IBD colon and human frontal cortex datasets. (A, B) Distribution of UMI amplification ratio across different cell types in individual samples from the human IBD colon dataset (A) and human frontal cortex dataset (B). Each boxplot represents the UMI amplification ratio distribution within a specific cell type for a given sample. (C, D) UMI amplification distribution across cell groups stratified by total UMI counts in individual samples from the human IBD colon dataset (C) and human frontal cortex dataset (D). Within each sample, cells were divided into ten equal-sized groups based on their total UMI counts, and boxplots depict the distribution of UMI amplification ratio across these groups. The x-axis represents cell groups ordered from lowest to highest total UMI counts, illustrating how UMI amplification ratio varies with total UMI counts.



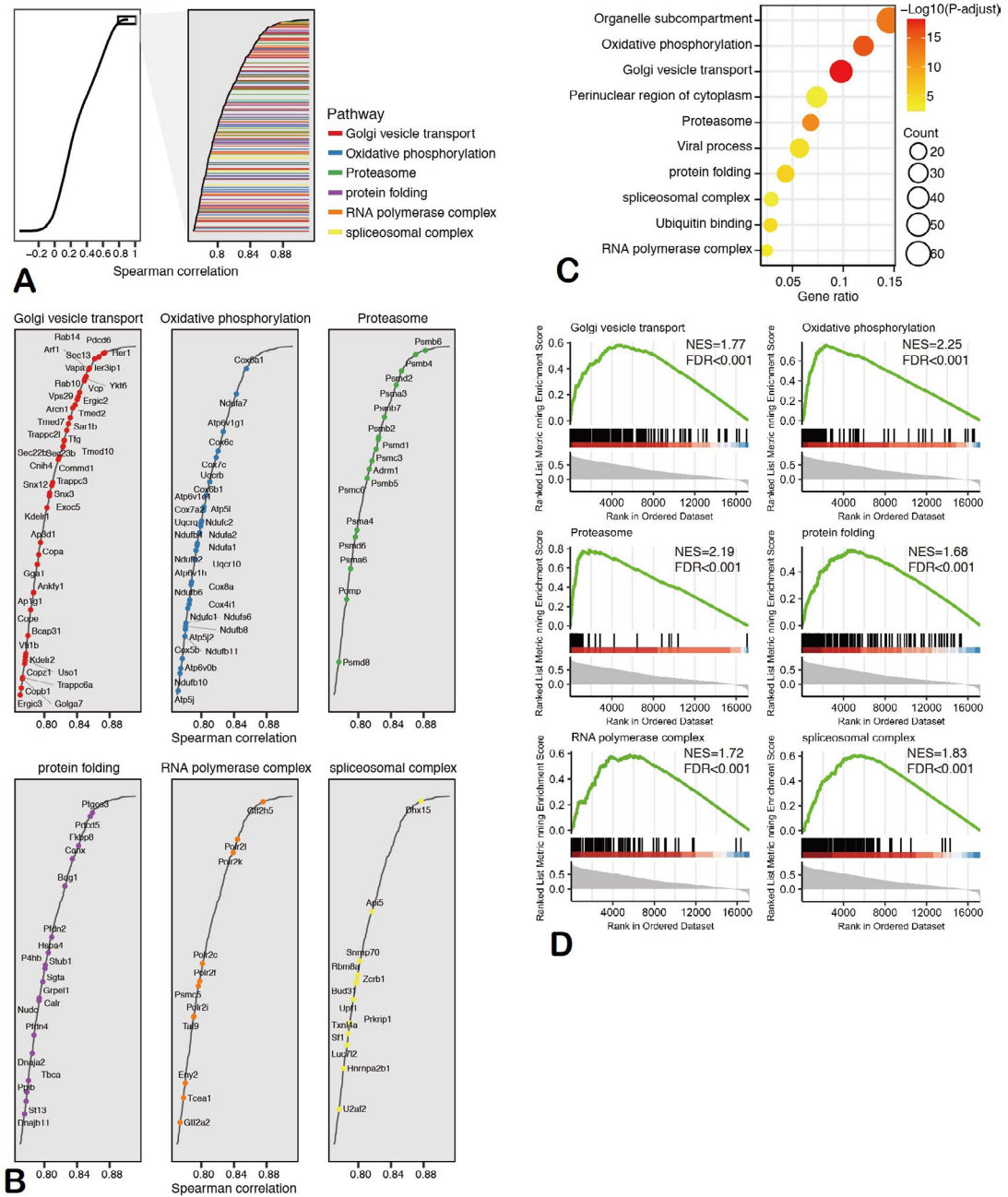
Supplementary Fig. 3. Relationship between cellular total UMI counts (before and after LOESS correction) and UMI amplification ratio. (A) Relationship between $\log_{10}$ -transformed cellular total UMI counts and UMI amplification ratio in individual samples from the human IBD colon dataset. Each panel represents a distinct sample, with a smoothed line indicating the median UMI amplification ratio and the colored region representing the interquartile range. (B) Same as (A), but using total LC-UMIs instead of total raw UMI counts. (C, D) Equivalent analyses performed on individual samples from the human frontal cortex dataset, following the same structure as (A, B).



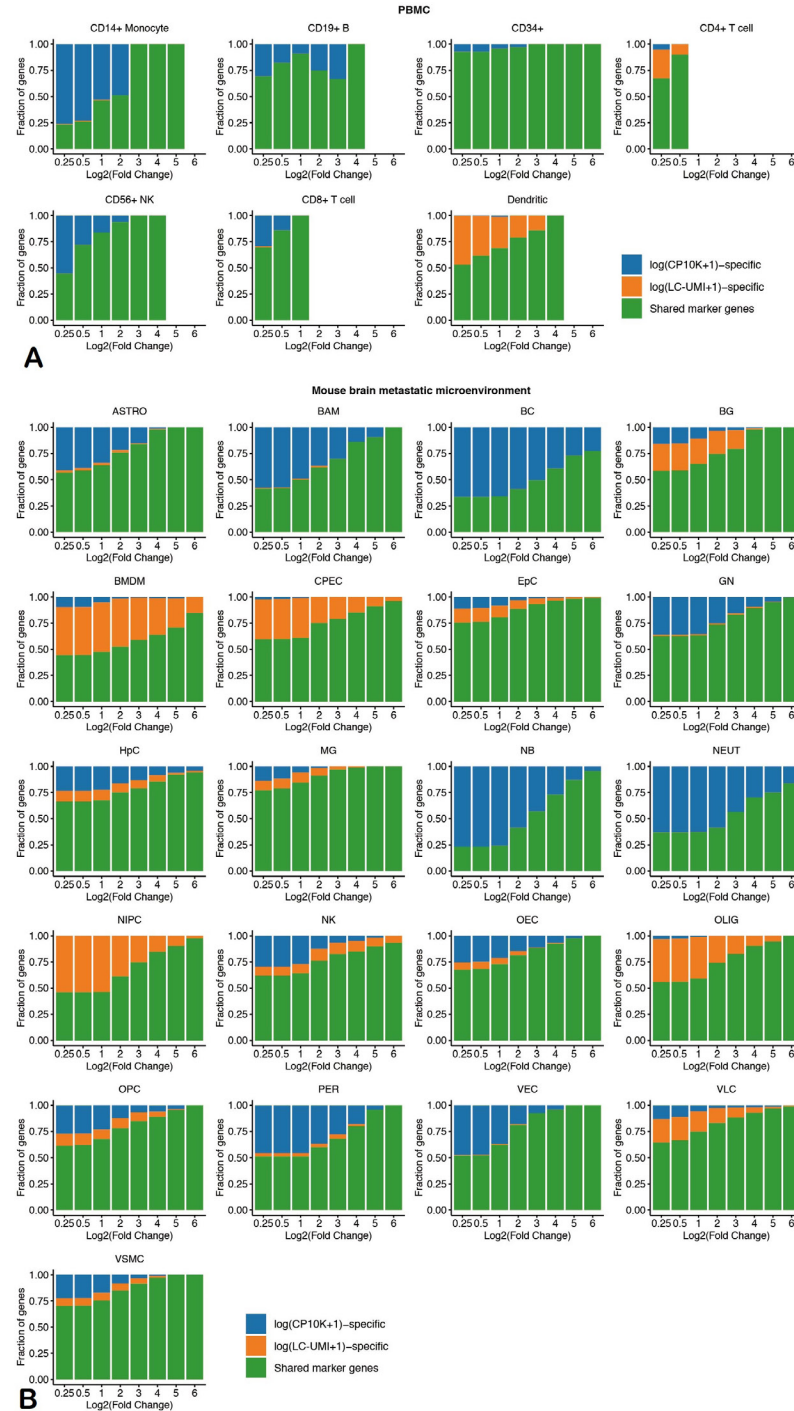
Supplementary Fig. 4. Comparison of regression models for modeling the relationship between UMI amplification ratio and total UMI counts in representative scRNA-seq samples. Scatter plots show data from representative samples of the three datasets presented in **Figure 1**: (A) human IBD colon (GSM6614349), (B) human frontal cortex (GSM2581279), and (C) rat parathyroid. Each point represents a single cell, with the x-axis indicating the mean UMI amplification ratio (reads per UMI) and the y-axis showing the log₁₀-transformed total UMI count. Five regression models linear, second-degree polynomial (Poly2), natural spline (df = 4), LOESS (span = 0.75), and generalized additive model (GAM) were fitted to each sample. Fitted curves are overlaid in color. Model performance was evaluated by root mean square error (RMSE) and coefficient of determination (R²), summarized in the tables below each panel. While GAM achieved marginally better fit metrics, LOESS produced smoother and more stable trends without overfitting, supporting its utility for correcting UMI amplification bias.



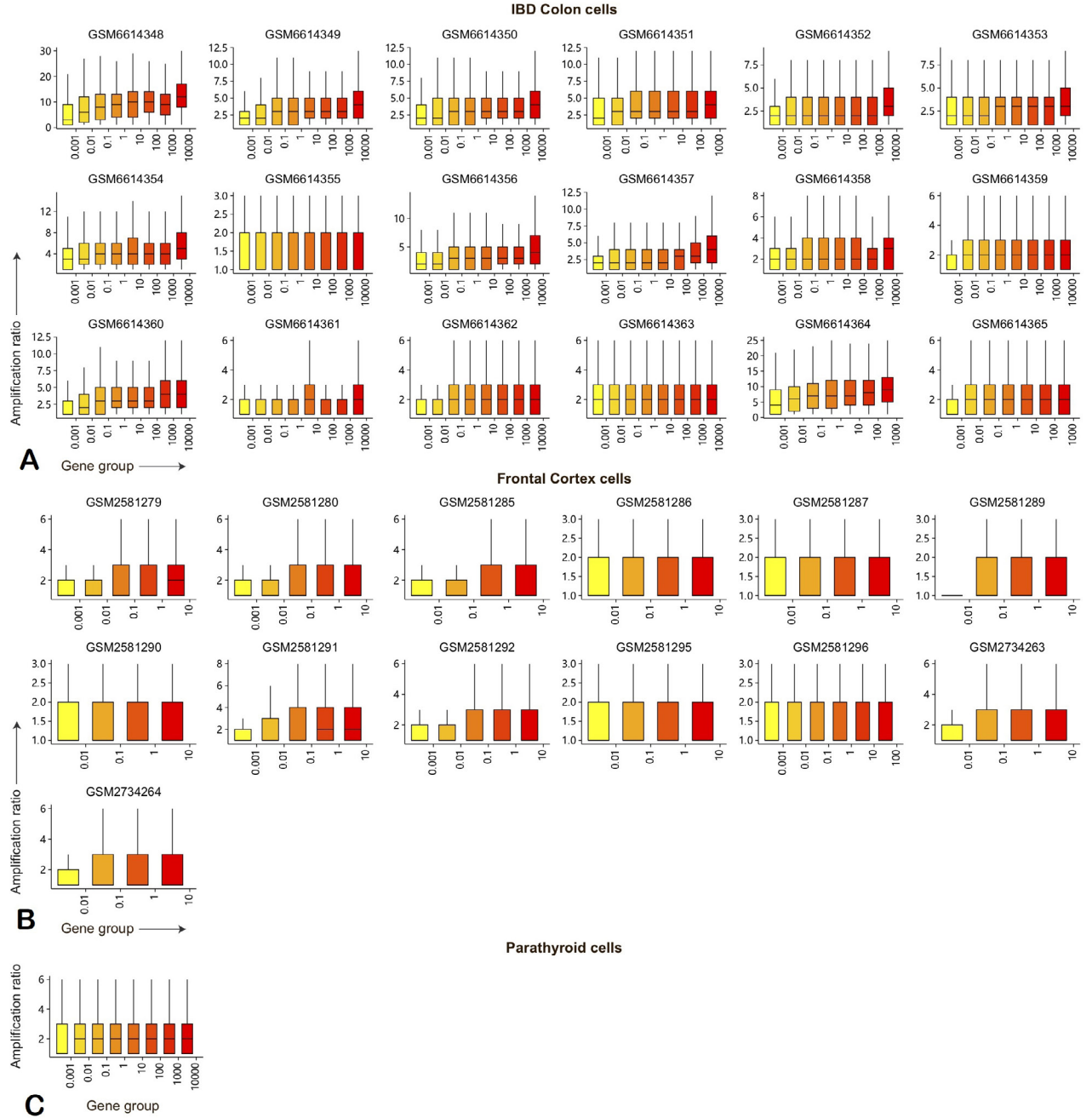
Supplementary Fig. 5. Total LC-UMIs across cell types in the human IBD colon and human frontal cortex datasets. Boxplots showing the distribution of total LC-UMIs across cell types in individual samples from the (A) human IBD colon and (B) human frontal cortex datasets.



Supplementary Fig. 6. Correlation between total LC-UMIs and gene expression, and pathway enrichment in the Tabula Muris dataset. (A) Spearman correlation between total LC-UMIs and gene expression levels across 72 cell types in the Tabula Muris dataset. The 500 genes with the highest correlation coefficients are highlighted on the right, with transcription-related pathway genes color-coded by pathway identity. (B) Detailed breakdown of transcription-related pathways identified in (A), with each pathway displayed separately. Genes within each pathway that rank among the top 500 most correlated genes are highlighted in distinct colors. (C) Top 10 enriched pathways among the top 500 most correlated genes from (A). The x-axis represents the gene ratio, defined as the proportion of pathway-associated genes relative to the total analyzed genes. Bubble size indicates the number of genes per pathway, while color intensity corresponds to statistical significance ($-\log_{10}$ adjusted p-value). (D) GSEA results for transcription-related pathways. Each panel displays the enrichment score (green line) for a given pathway across the ranked gene list ordered by Spearman correlation coefficients from (A). The normalized enrichment score (NES) and false discovery rate (FDR) are shown.



Supplementary Fig. 7. Comparison of marker gene identification using log(CP10K+1) and log(LC-UMI+1) normalization in PBMC and mouse brain metastatic microenvironment datasets. (A, B) Stacked bar plots illustrating the proportion of marker genes identified at progressively increasing log₂FC thresholds for each cell type in the (A) PBMC and (B) mouse brain metastatic microenvironment datasets. Marker genes were identified using Wilcoxon tests (adjusted p-value < 0.05), required to have the highest expression in the corresponding cell type compared to all others, and further filtered across a range of log₂FC thresholds. Shared marker genes identified by both normalization methods (log(CP10K+1) and log(LC-UMI+1)) are shown in green, while those uniquely detected by log(CP10K+1) and log(LC-UMI+1) are shown in blue and orange, respectively.



Supplementary Fig. 8. Distribution of UMI amplification ratios across gene expression groups in the human IBD colon, human frontal cortex, and rat parathyroid datasets. (A) Human IBD colon, (B) human frontal cortex, and (C) rat parathyroid datasets. Each panel represents a single sample from the corresponding dataset. Within each sample, genes were grouped by average expression level, defined as the mean UMI count per gene across all cells. Expression bins were determined using $\log_{10}$ -scale thresholds. For each group, boxplots show the distribution of UMI amplification ratios, calculated as the number of reads per UMI molecule. Genes with higher expression levels tend to display increased amplification ratios.