Human neocortex cross-areal NS-Forest marker gene analysis

Renee Zhang
Beverly Peng
Richard Scheuermann

Aug 5, 2024

Datasets

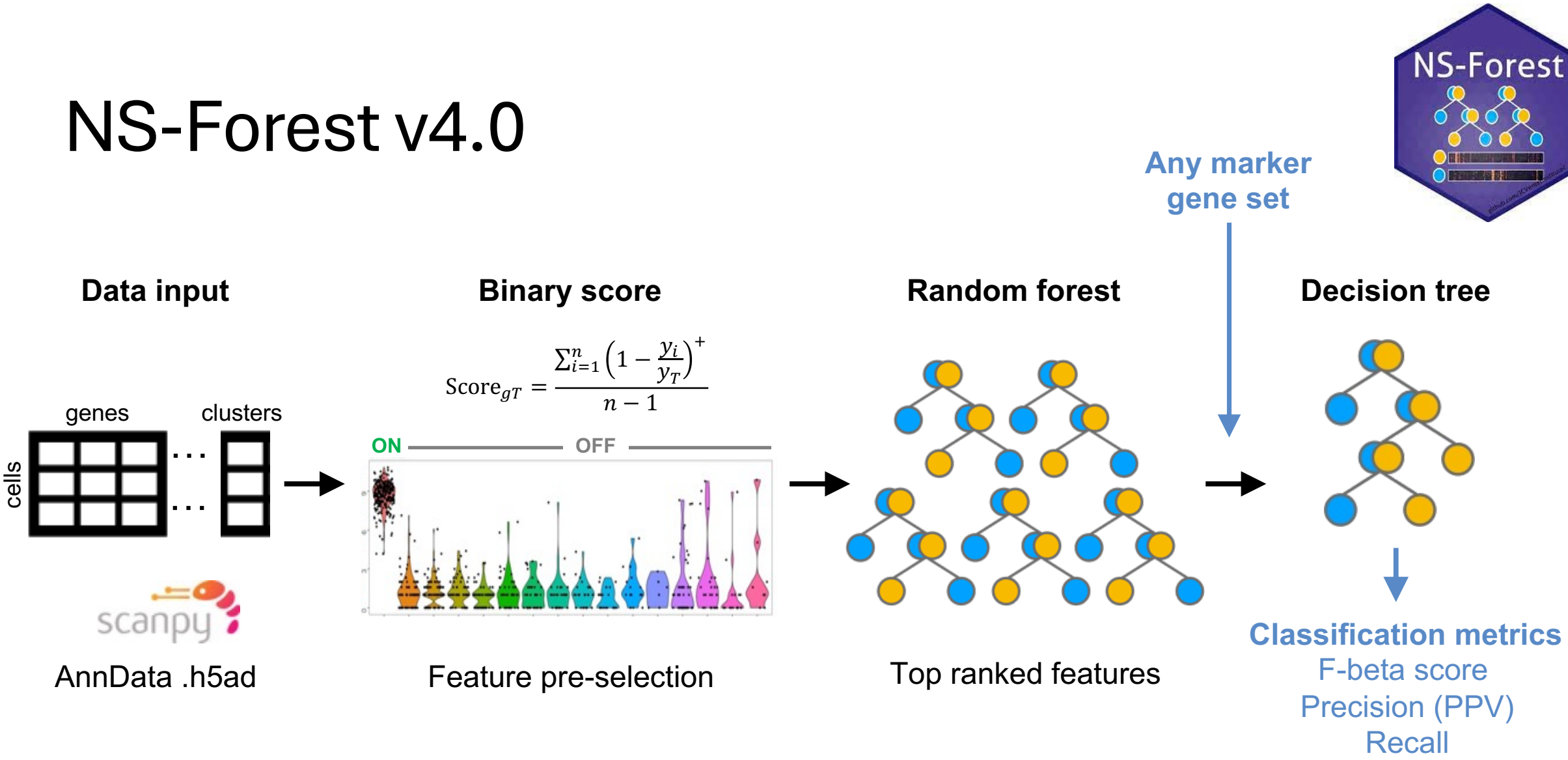
- Paper: Jorstad et al. (2023) Science
- Downloaded “Supercluster” datasets from CELLxGENE: https://cellxgene.cziscience.com/collections/d17249d2-0e6e-4500-abb8-e6c93fa1ac6f
- 5 Superclusters: 2 for interneurons, 2 for excitatory neurons, and 1 for non-neuronal
- Cross-area cell type stats:
  - 24 subclasses and 153 clusters
  - The size of subclasses ranges from 765 – 288,049 nuclei
  - The size of clusters ranges from 71 – 207,759 nuclei
  - The interneuron superclusters have substantially more clusters (48, 49 vs. 22, 19, 15)

CELLxGENE link									
Dataset	**Genes**	**Cells**	**Dataset_id**	**# CrossArea_subclass**	**max. size**	**min. size**	**# CrossArea_cluster**	**max. size**	**min. size**
Supercluster: IT-projecting excitatory neurons	30265	638,941	Supercluster_01_IT	5	288049	26616	22	207759	902
Supercluster: MGE-derived interneurons	29507	185,477	Supercluster_02_MGE	4	97963	1218	48	48491	293
Supercluster: CGE-derived interneurons	29413	129,495	Supercluster_03_CGE	5	67384	6195	49	15120	71
Supercluster: Non-neuronal cells	29330	108,940	Supercluster_04_NN	6	51816	765	15	49396	109
Supercluster: Deep layer (non-IT) excitatory neurons	29348	92,969	Supercluster_05_DL	4	39267	8811	19	25547	133
		Total		Total			Total		
		1,155,822		24			153		

Design and terminology

- The focus of this NS-Forest analysis is to identify marker genes for cell types across brain regions in human neocortex. Therefore, in this deck,
  - subclass = **CrossArea_subclass** and cluster = **CrossArea_cluster**
- Workflow
  - Define NS-Forest for each subclass by:
    - generating a sampled "global" data file in which we merge and sample each cluster (up to 500 nuclei) in each Supercluster file such that we would have roughly equal representation of all 153 clusters ==> 73,144 nuclei
    - use this to identify the NS-Forest markers for each subclass ("global" subclass markers)
  - Run NS-Forest on each Supercluster file separately to produce NS-Forest markers at the cluster level ("local" cluster markers)
  - Best combinatorial search for "global" *combinatorial* markers for clusters
    - Idea: each cluster would naturally inherit the subclass and cluster markers from the "global" analysis producing a biomarker set determined by the best F-beta score
    - Run NS-Forest on the "global" data at the cluster level ("global" cluster-*only* markers)
    - Run NS-Forest decision tree evaluation module to determine the best combination from a candidate set = "global" subclass markers + "global" cluster-*only* markers for each cluster
  - After comparing global and local markers on global and local data, the best overall performer is the "global" *combinatorial* markers for clusters

NS-Forest v4.0



Liu A, et al. (2024) ***bioRxiv*** [Preprint], 2024 Apr 27:2024.04.22.590194. PMID: 38712147

Subclass – NS-Forest "global" subclass markers

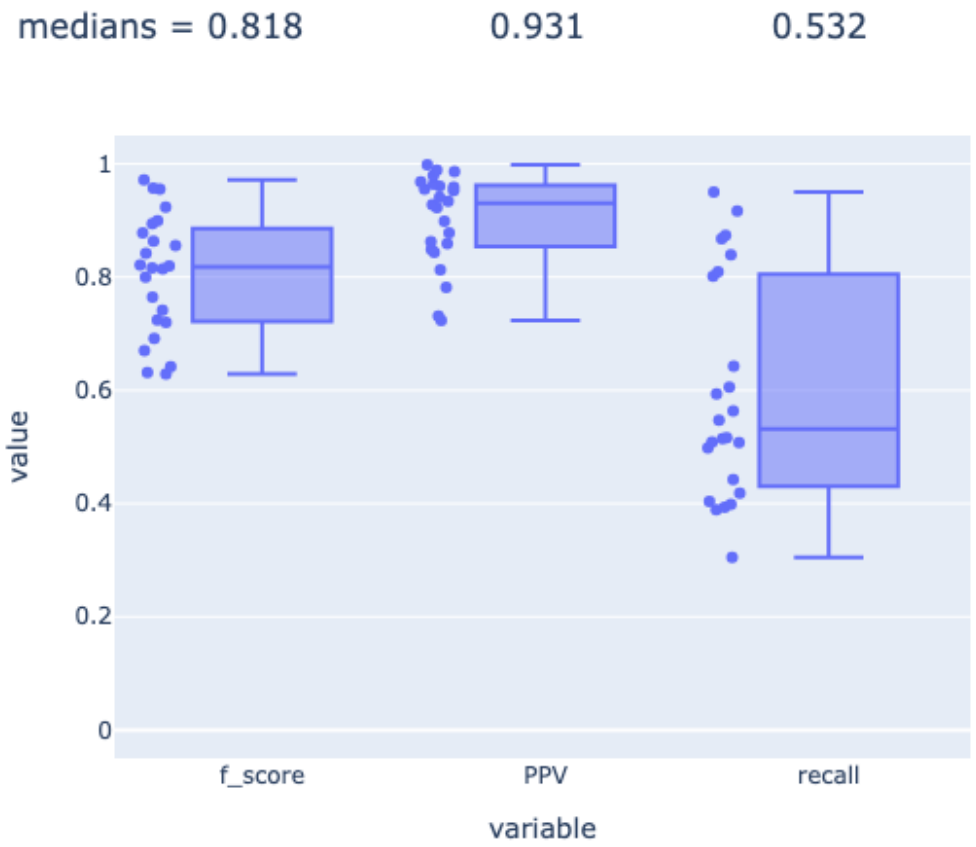
- NS-Forest produces from 1 - 3 marker genes for optimal classification of subclasses
- 46 markers for 24 subclasses
- They include well-known genes like VIP, SST, PVALB, LAMP5, THEMIS, RORB
- LINC RNAs are selected as markers

	clusterName	clusterSize	f_score	PPV	TN	FP	FN	TP	marker_count	markers	binary_genes	recall
0	Astro	2500	0.957044	0.958821	70542	102	125	2375	1	[OBI1-AS1]	['OBI1-AS1', 'AQP4', 'FGFR3', 'GFAP', 'LINC004...	0.958821
1	Chandelier	500	0.799748	0.933824	72626	18	246	254	2	[GPR149, CA8]	['COL15A1', 'NPNT', 'TAFA4', 'GPR149', 'CA8', ...	0.933824
2	Endo	609	0.923892	0.960630	72515	20	121	488	2	[FLT1, ABCG2]	['FLT1', 'HERC2P3', 'ABCG2', 'CLDN5', 'VWF', '...	0.960630
3	L2/3 IT	3000	0.670150	0.813232	69873	271	1820	1180	2	[ONECUT2, COL5A2]	['TPBG', 'ONECUT2', 'RTKN2', 'SLC38A11', 'COL5...	0.813232
4	L4 IT	3000	0.720109	0.781637	69685	459	1357	1643	2	[RORB, ST8SIA4]	['MET', 'VIPR2', 'RORB-AS1', 'KCNH8', 'RORB', ...	0.781637
5	L5 ET	1636	0.815891	0.955732	71469	39	794	842	2	[CHST8, POU3F1]	['CHST8', 'POU3F1', 'CACNA1H', 'SLC66A1L', 'PC...	0.955732
6	L5 IT	3000	0.764261	0.963874	70097	47	1746	1254	3	[ARSJ, RORB, KCNK2]	['TLL1', 'MKX', 'ARSJ', 'RORB', 'PCP4', 'KCNK2...	0.963874
7	L5/6 NP	2633	0.899288	0.998820	70509	2	940	1693	2	[NPSR1-AS1, TLL1]	['NPSR1-AS1', 'FER1L6-AS2', 'GDNF-AS1', 'HTR2C...	0.998820
8	L6 CT	1500	0.631898	0.862524	71571	73	1042	458	3	[SEMA5B, ADAMTSL1, SULF1]	['SYT6', 'VWA2', 'SEMA5B', 'ADAMTSL1', 'SEMA3E...	0.862524
9	L6 IT	1000	0.641696	0.723404	71975	169	558	442	2	[LINC00507, THEMIS]	['PXDN', 'LINC00507', 'ADAMTS3', 'ART3', 'THEM...	0.723404
10	L6 IT Car3	1000	0.877752	0.988581	72137	7	394	606	2	[SMYD1, THEMIS]	['SMYD1', 'POSTN', 'NTNG2', 'NR4A2', 'OLFML2B'...	0.988581
11	L6b	2951	0.819941	0.968367	70144	49	1451	1500	2	[CCN2, ROS1]	['MDFIC', 'CCN2', 'ROS1', 'CDH24', 'SEMA3D', '...	0.968367
12	Lamp5	3500	0.814688	0.952532	69554	90	1694	1806	2	[LINC00298, KIT]	['CPLX3', 'LINC00298', 'KIT', 'LAMP5', 'PDGFD'...	0.952532
13	Lamp5 Lhx6	500	0.821678	0.927632	72622	22	218	282	2	[CHST9, LAMP5]	['SFTA3_ENSG00000229415', 'SFRP1', 'PDGFD', 'C...	0.927632
14	Micro/PVM	500	0.955106	0.979684	72635	9	66	434	2	[APBB1IP, DOCK8]	['APBB1IP', 'P2RY12', 'DOCK8', 'ADAM28', 'TBXA...	0.979684
15	OPC	500	0.741071	0.844068	72598	46	251	249	1	[FERMT1]	['FERMT1', 'GPR17', 'OLIG1', 'STK32A', 'OLIG2'...	0.844068
16	Oligo	1529	0.971856	0.986629	71596	19	127	1402	2	[ST18, PLP1]	['ST18', 'MOBP', 'PLP1', 'ENPP2', 'MOG', 'UGT8...	0.986629
17	Pax6	2000	0.724125	0.922894	71079	65	1222	778	2	[RELN, HCRTR2]	['SP8', 'DDR2', 'RELN', 'WIF1', 'CHRNA7', 'HCR...	0.922894
18	Pvalb	4500	0.691999	0.848485	68324	320	2708	1792	3	[PVALB, TAC1, MYO5B]	['PVALB', 'TAC1', 'RNF144B', 'MYO5B', 'SULF1',...	0.848485
19	Sncg	3803	0.628944	0.730952	68776	565	2268	1535	2	[RXRG, THSD7B]	['RXRG', 'SEMA3C', 'NDST4', 'CHRNA7', 'MEGF11'...	0.730952
20	Sst	18095	0.842367	0.941027	54376	673	7356	10739	2	[SST, PLCH1]	['SST', 'PAWR', 'FLT3', 'PLCH1', 'COL25A1', 'L...	0.941027
21	Sst Chodl	500	0.894026	0.899177	72595	49	63	437	1	[NPY]	['NPY', 'NPY2R', 'LINC01014', 'CDHR1', 'ANKRD3...	0.899177
22	VLMC	639	0.863105	0.877759	72433	72	122	517	1	[CARMN]	['CARMN', 'SLC19A1', 'NOTCH3', 'PDGFRB', 'COLE...	0.877759
23	Vip	13749	0.855531	0.859622	57510	1885	2206	11543	1	[VIP]	['VIP', 'CHRNA2', 'SHISA8', 'TAC3', 'ASIC4', '...	0.859622

Subclass classification using NS-Forest “global” subclass markers

- F-beta scores range from 0.63 - 0.97 with a median of 0.82
- PPV (precision) scores range from 0.72 - 0.99 with a median of 0.93
- Recall scores range from 0.31 - 0.95 with a median of 0.53

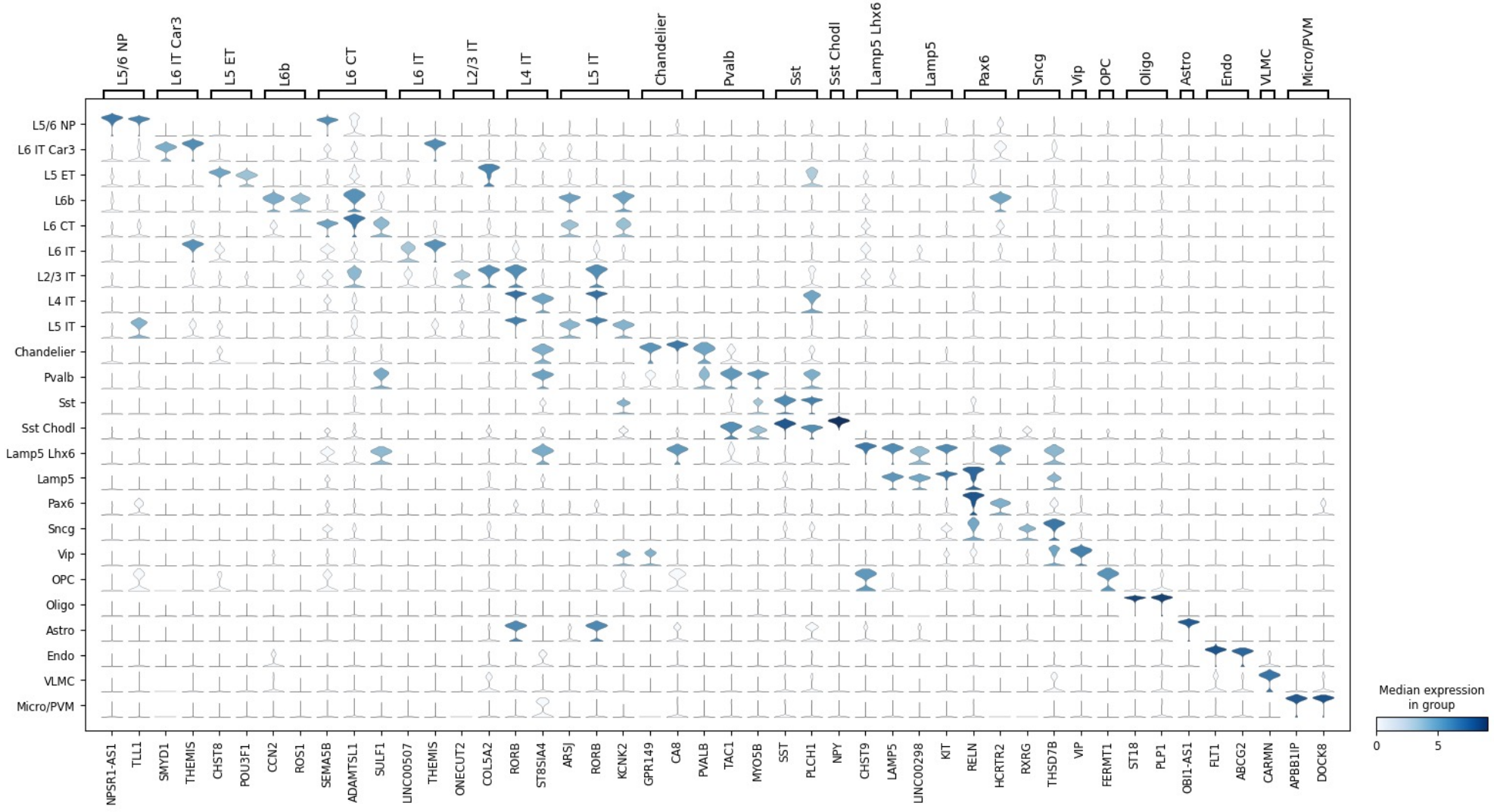
- The F-beta score is often lower than the PPV due to false negative predictions that would be impacted by the expression dropout artifact seen in snRNA-seq experiments
- Recall is largely impacted by the false negative predictions



PPV = TP / (TP + FP); recall = TP / (TP + FN)

Subclass – NS-Forest “global” subclass markers

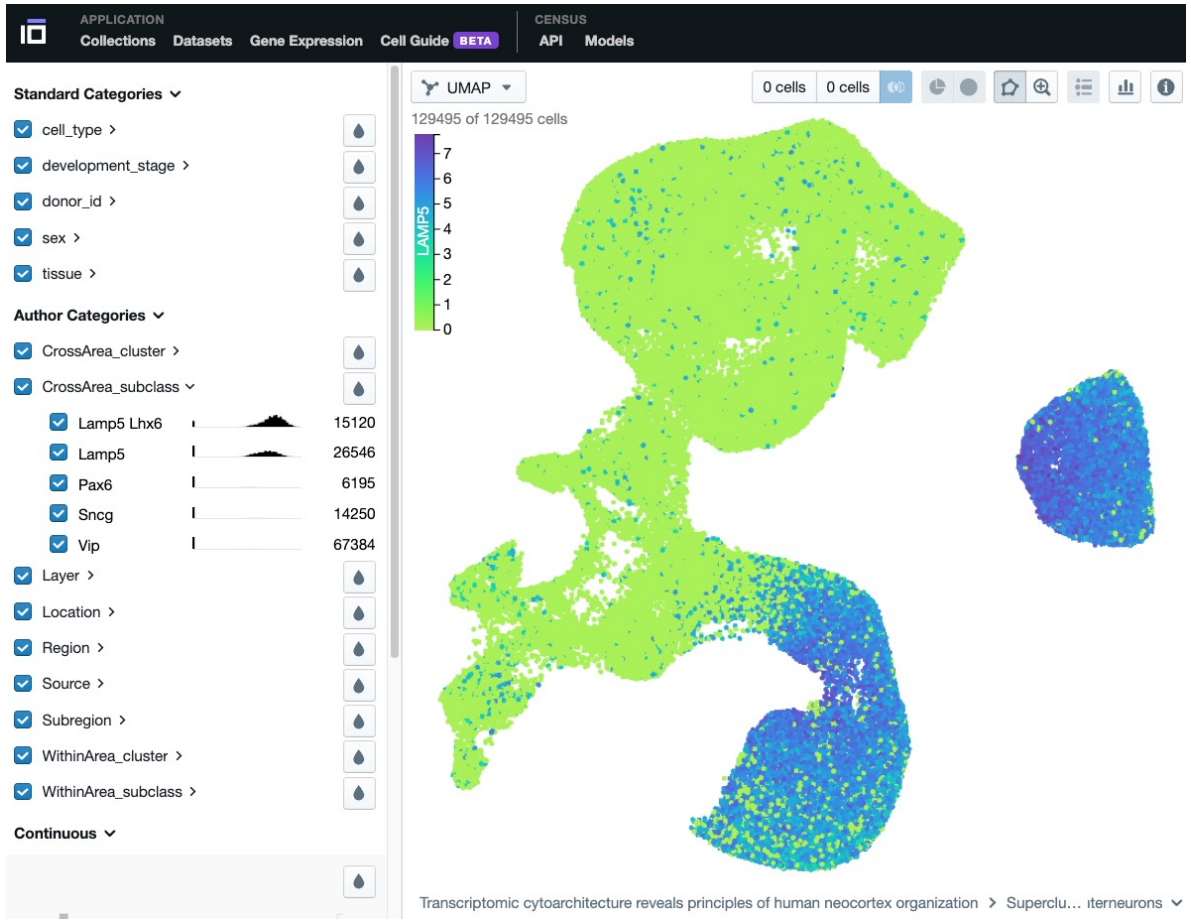
- Markers are expressed at high levels in the specific subclasses, forming along the diagonal



Subclass marker expression in subclasses and clusters

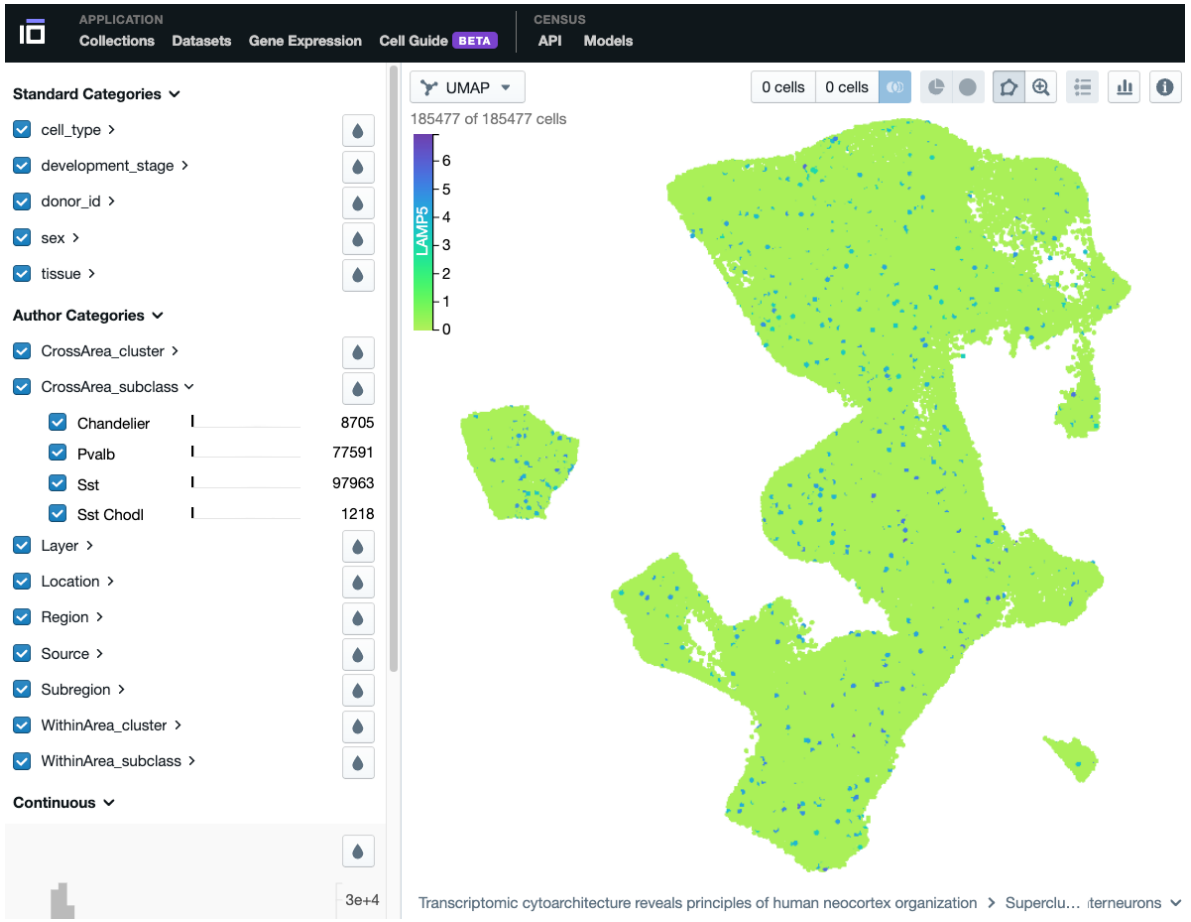
Subclass expression of Lamp5 Lhx6 subclass markers

LAMP5 expression in Supercluster:
CGE-derived interneurons



Expression in two subclasses

LAMP5 expression in Supercluster:
IT-projecting excitatory neurons



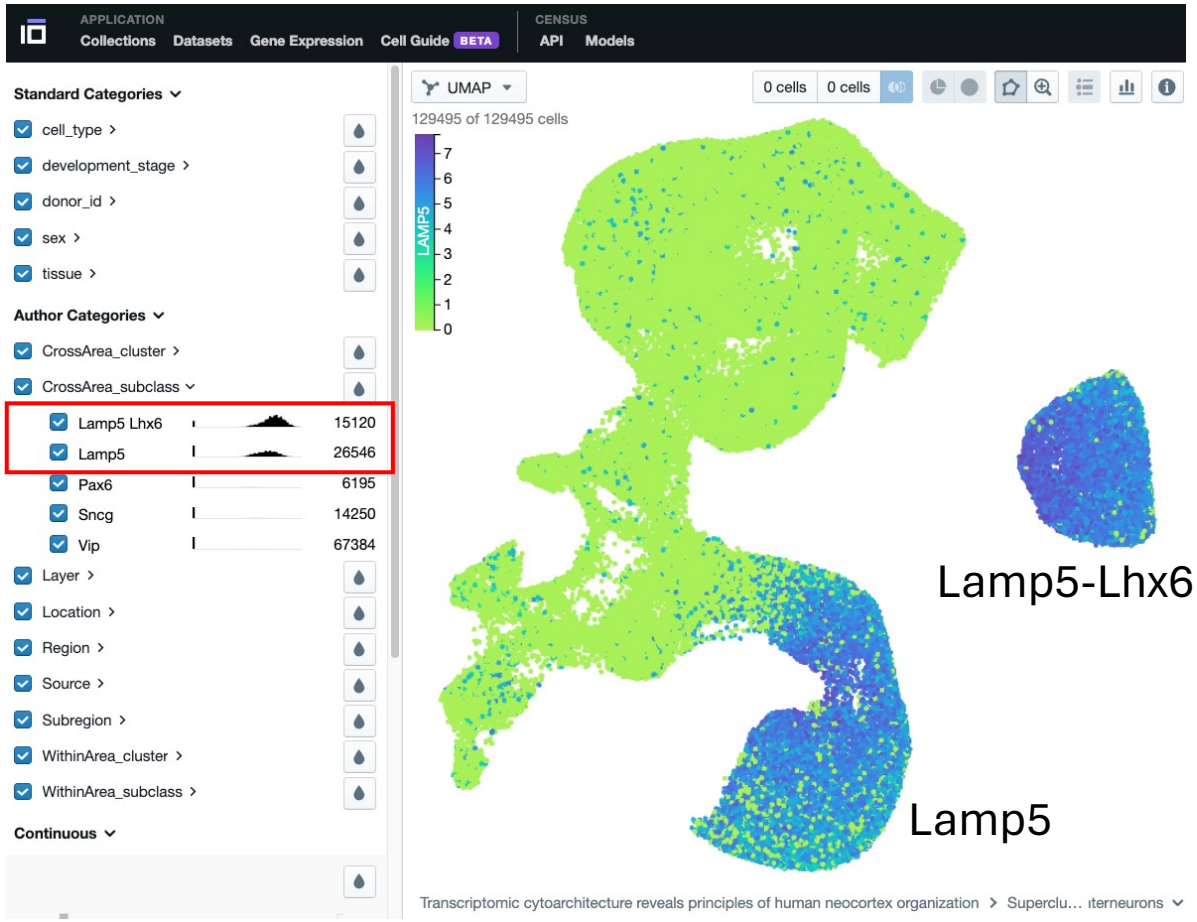
No expression

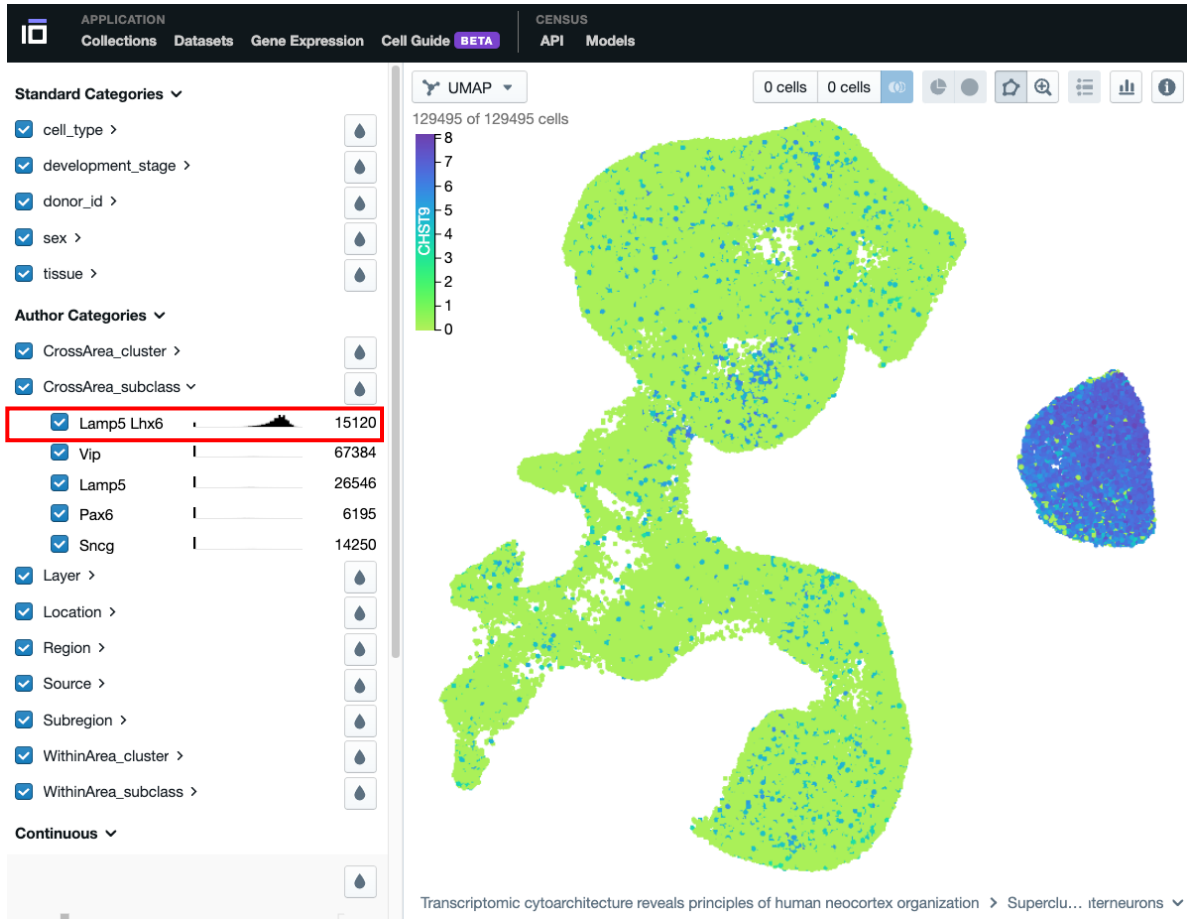
Subclass expression of Lamp5 Lhx6 subclass markers

LAMP5

Expression in Supercluster: CGE-derived interneurons

CHST9





LAMP5 expressed in two subclasses

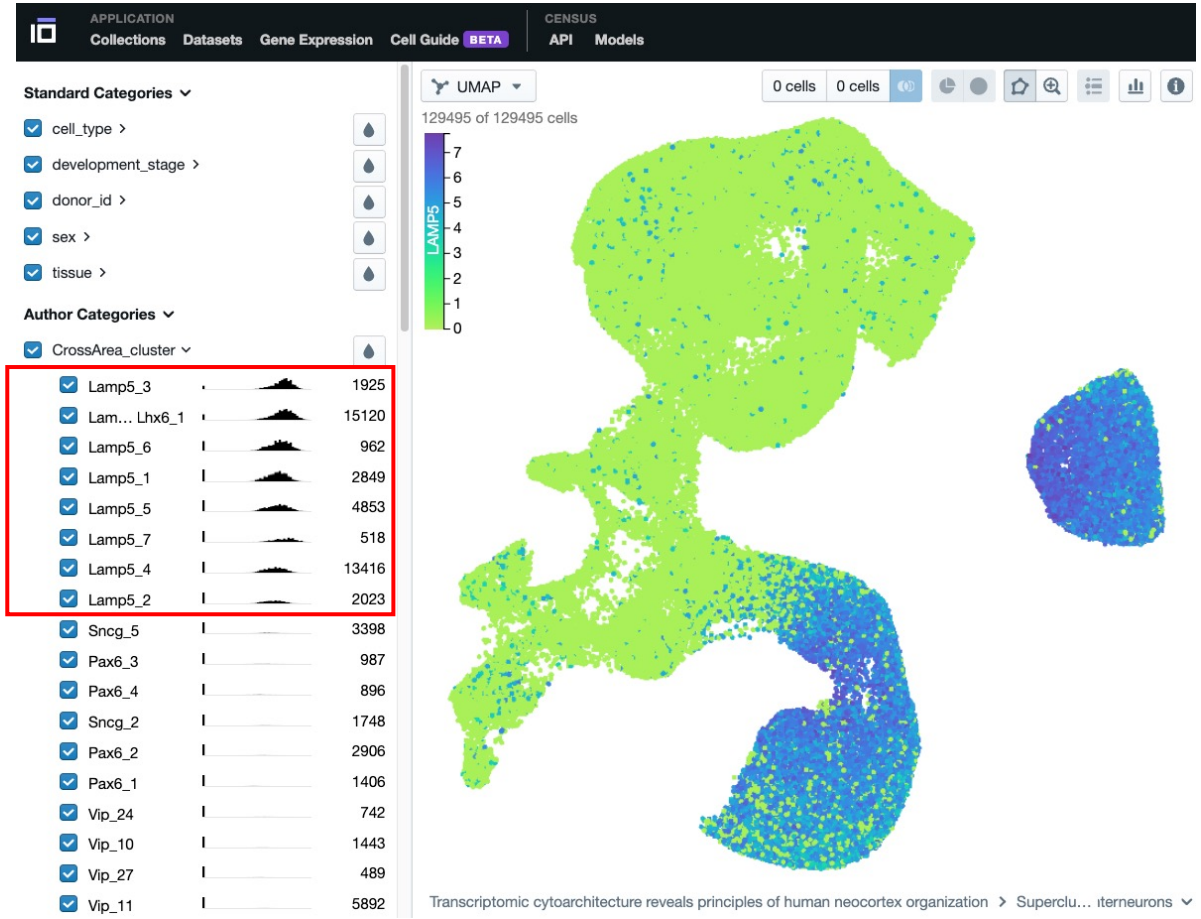
CHST9 gives specificity

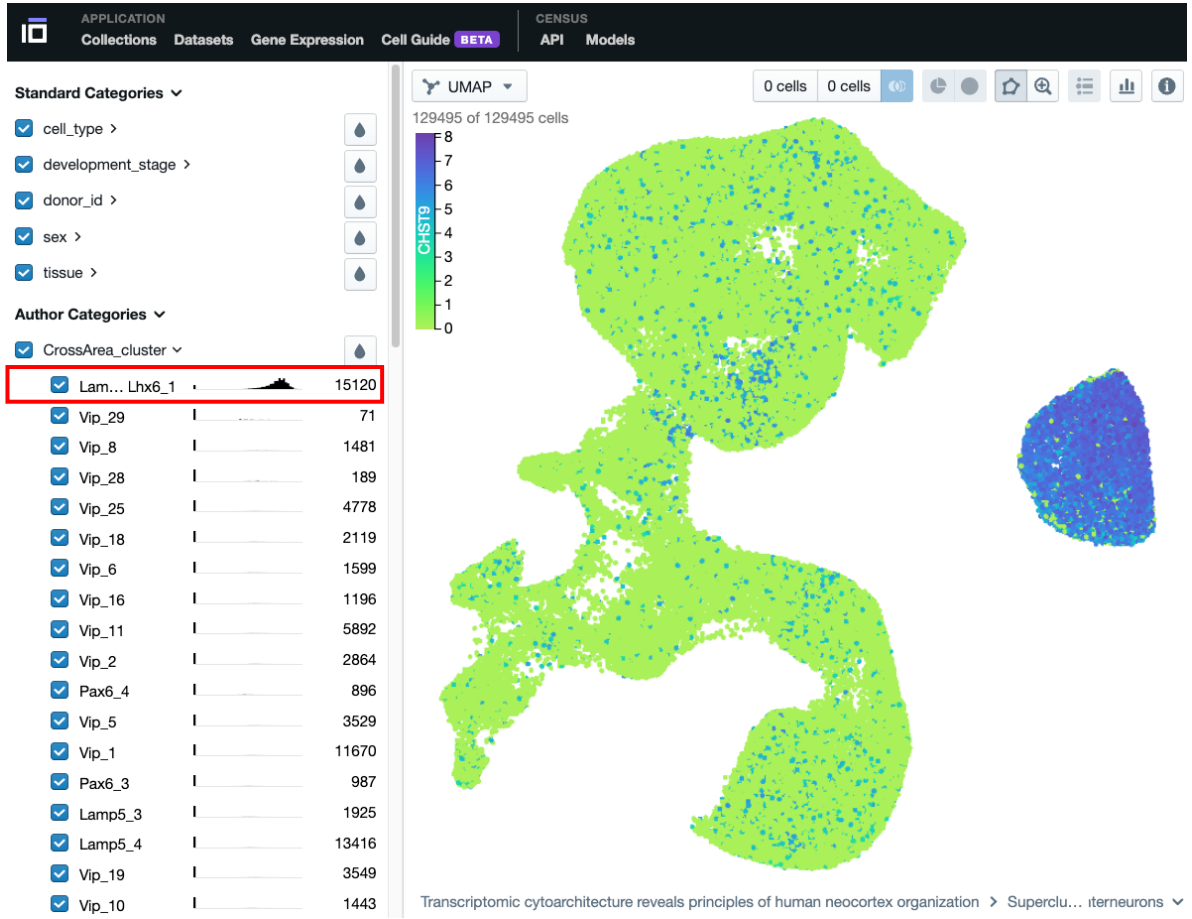
Cluster expression of Lamp5 Lhx6 subclass markers

LAMP5

Expression in Supercluster: CGE-derived interneurons

CHST9





LAMP5 expressed in quite a few clusters

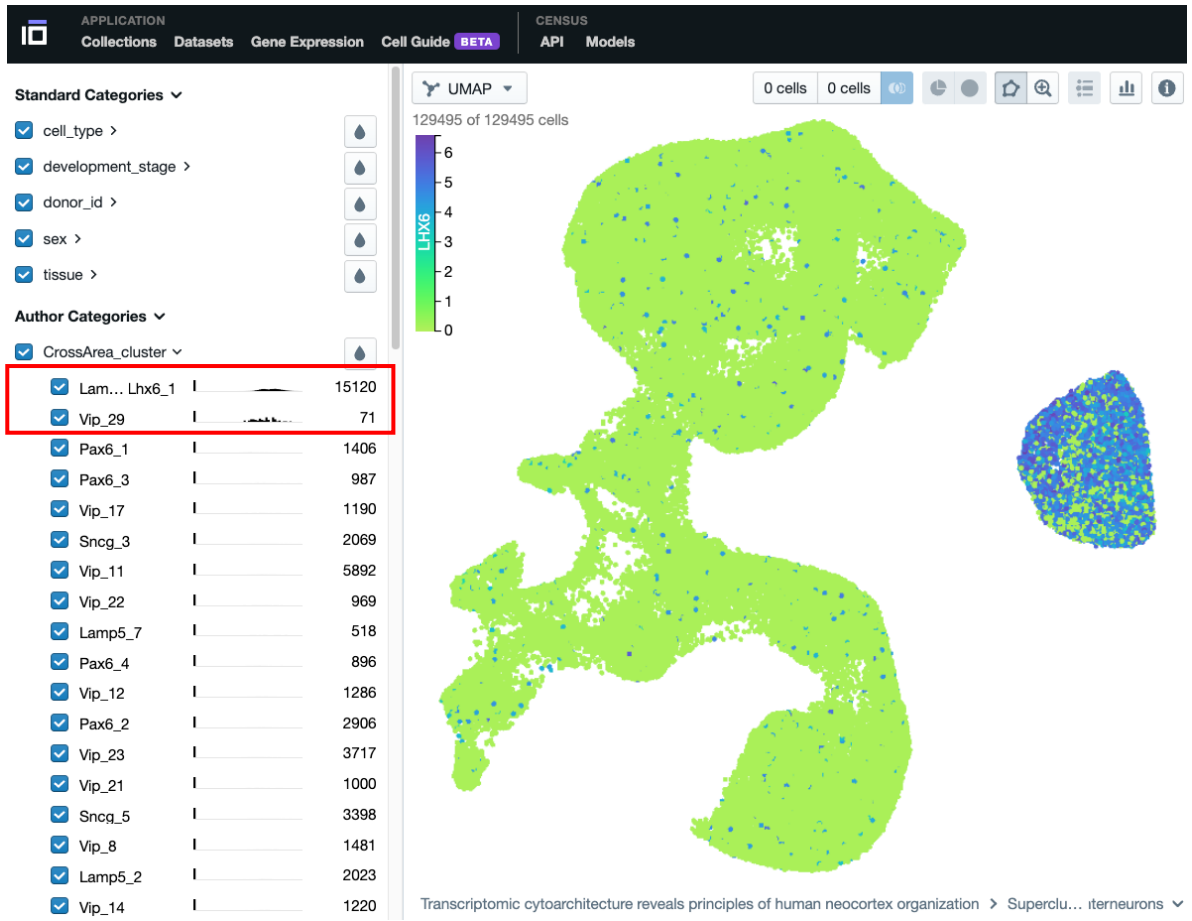
CHST9 gives specificity at cluster level

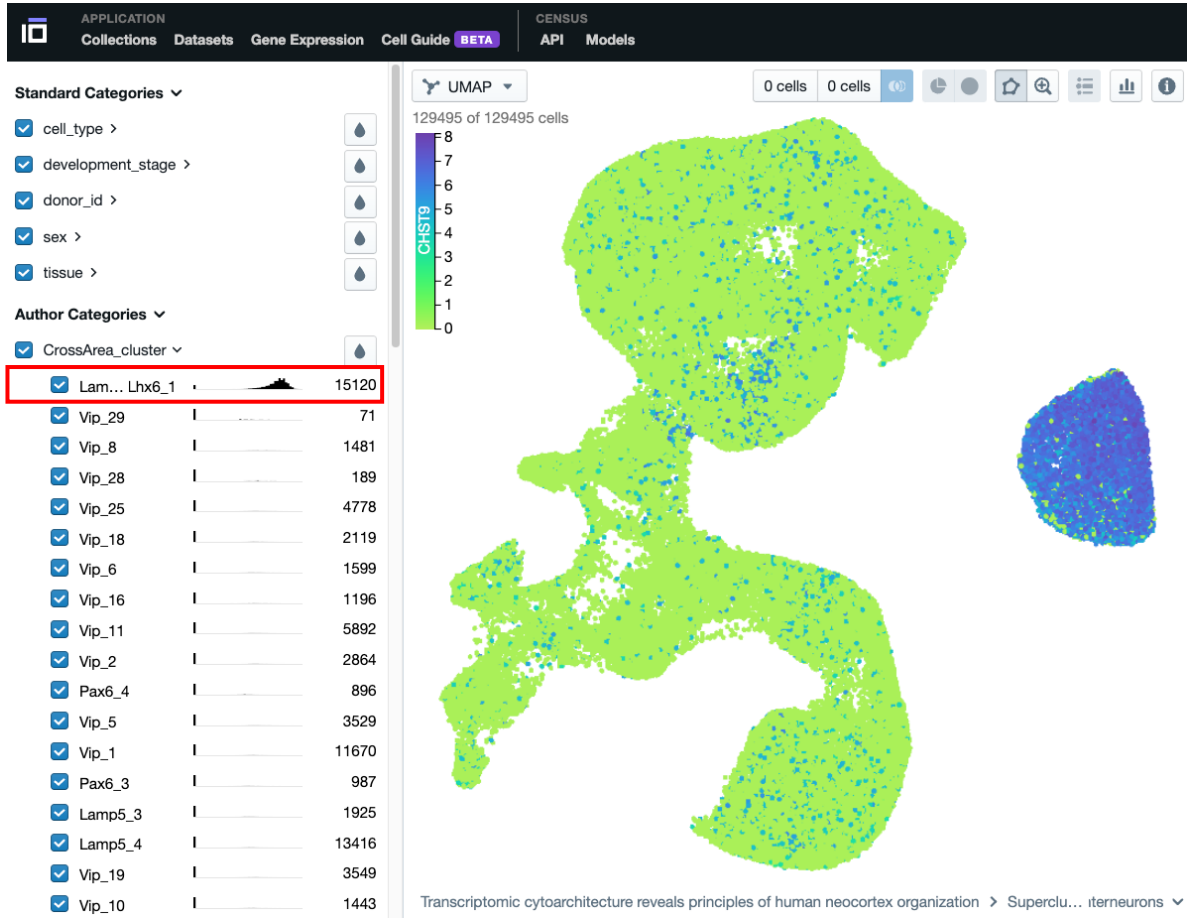
Cluster expression of Lamp5 Lhx6 subclass markers

LHX6

Expression in Supercluster: CGE-derived interneurons

CHST9





LHX6 expressed in two clusters
(N.B. LHX6 is ***not*** an NS-Forest marker)

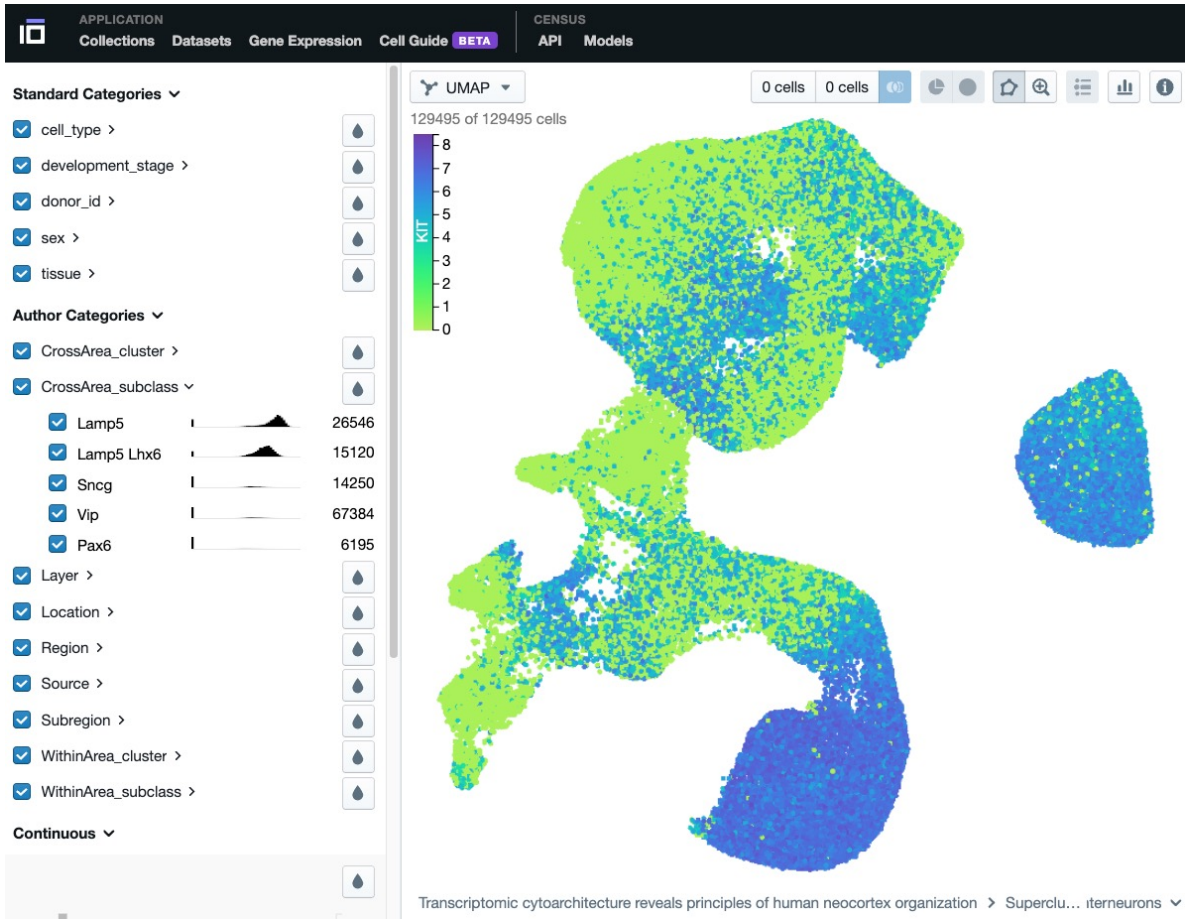
CHST9 gives specificity at cluster level

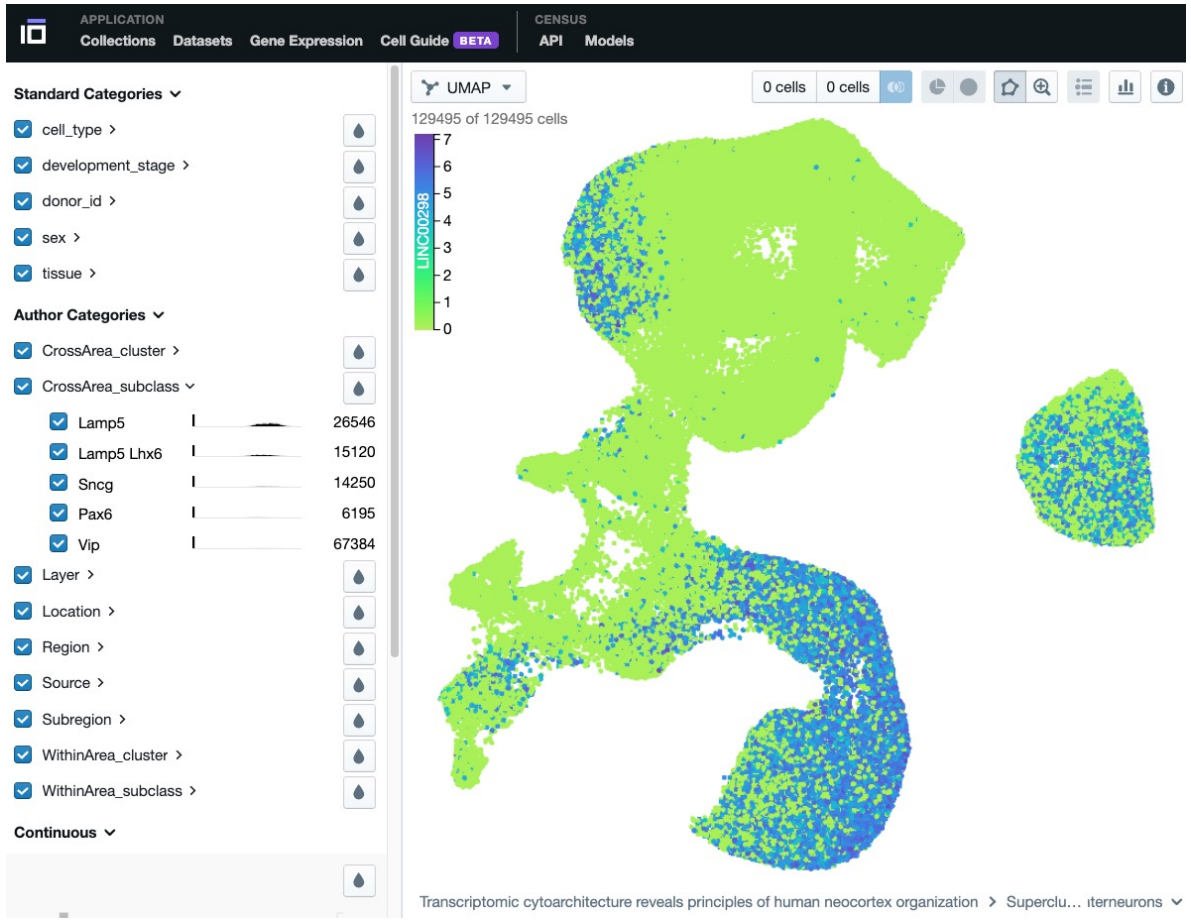
Subclass expression of Lamp5 subclass markers

KIT

Expression in Supercluster: CGE-derived interneurons

LINC00298





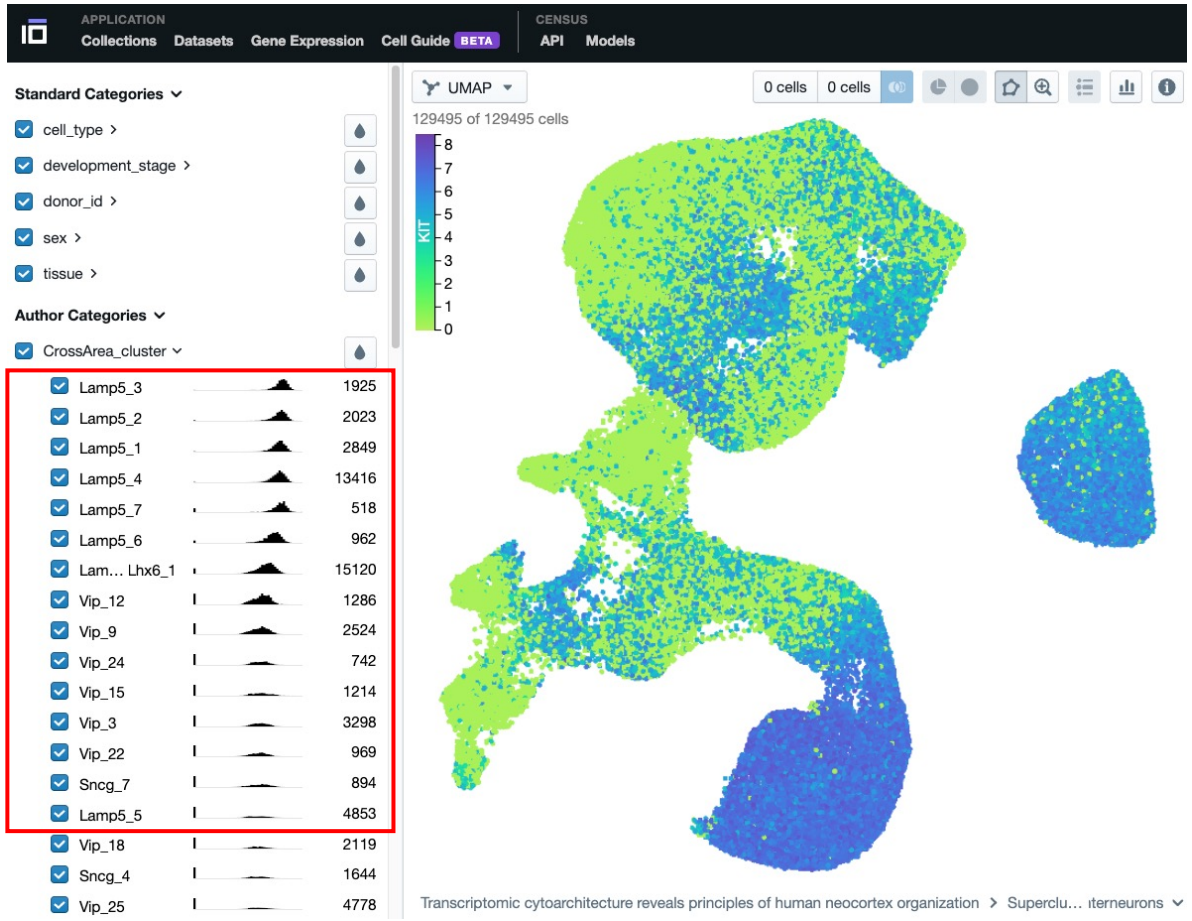
Combinatorial expression of the two LAMP5 subclass markers, KIT and LINC00298 gives specificity

Cluster expression of Lamp5 subclass markers

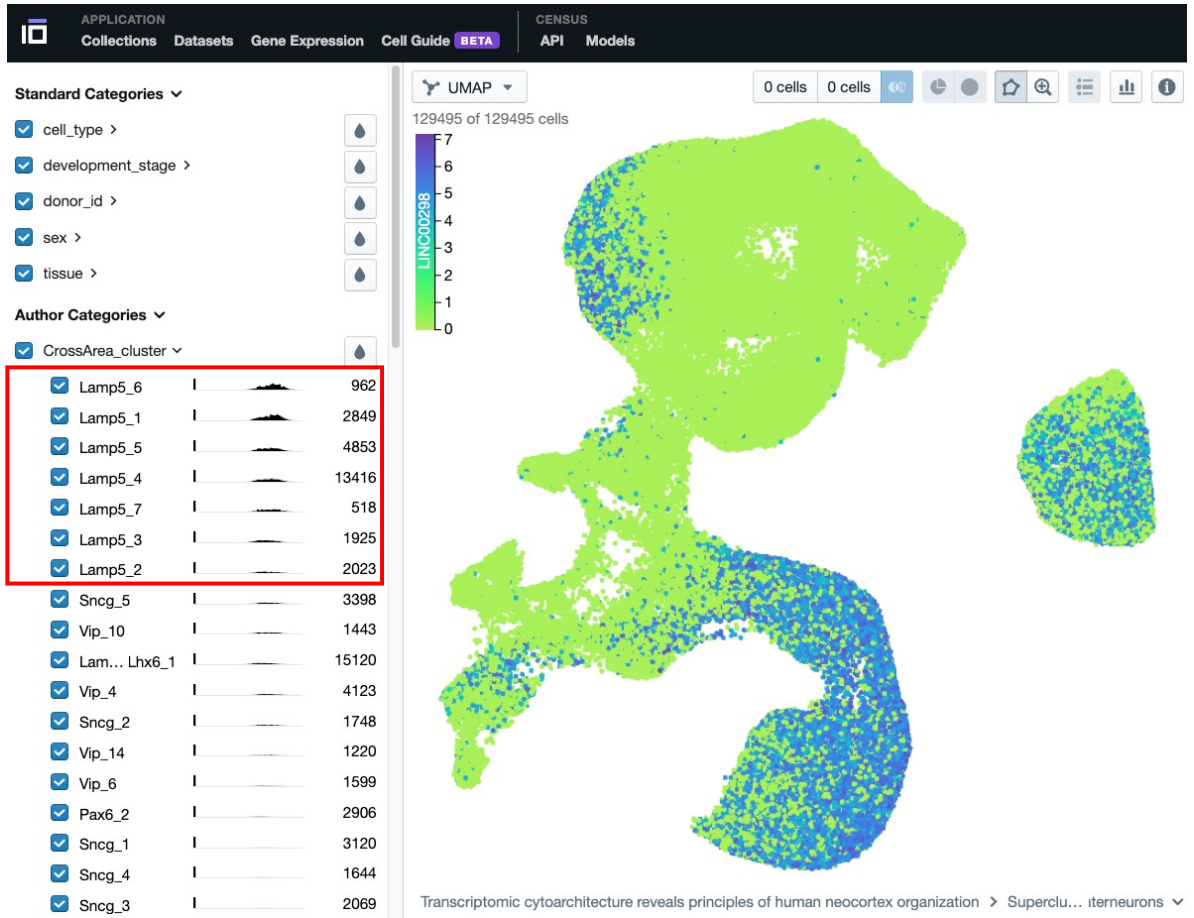
KIT

Expression in Supercluster: CGE-derived interneurons

LINC00298



KIT is highly expressed in many clusters

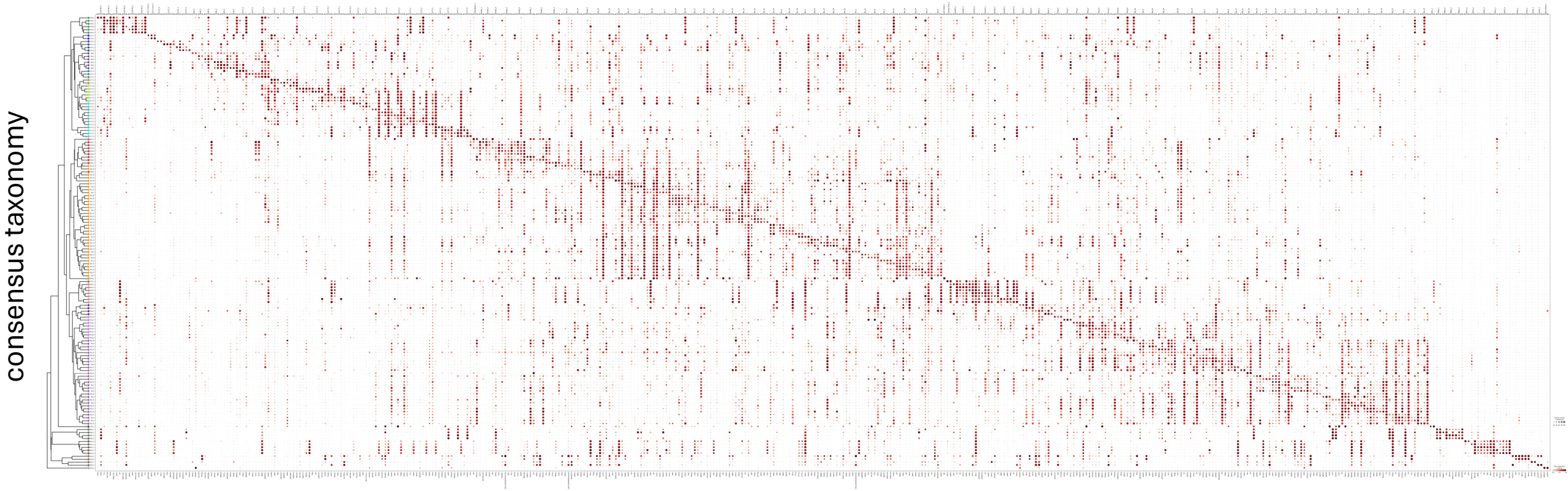


The LINC RNA LINC00298 is highly expressed in all clusters of the LAMP5 subclass

Cluster – NS-Forest “global” vs. “local” markers in all clusters

Cluster – “global” markers performance

global combinatorial markers evaluated on global data ==> 460 markers in total

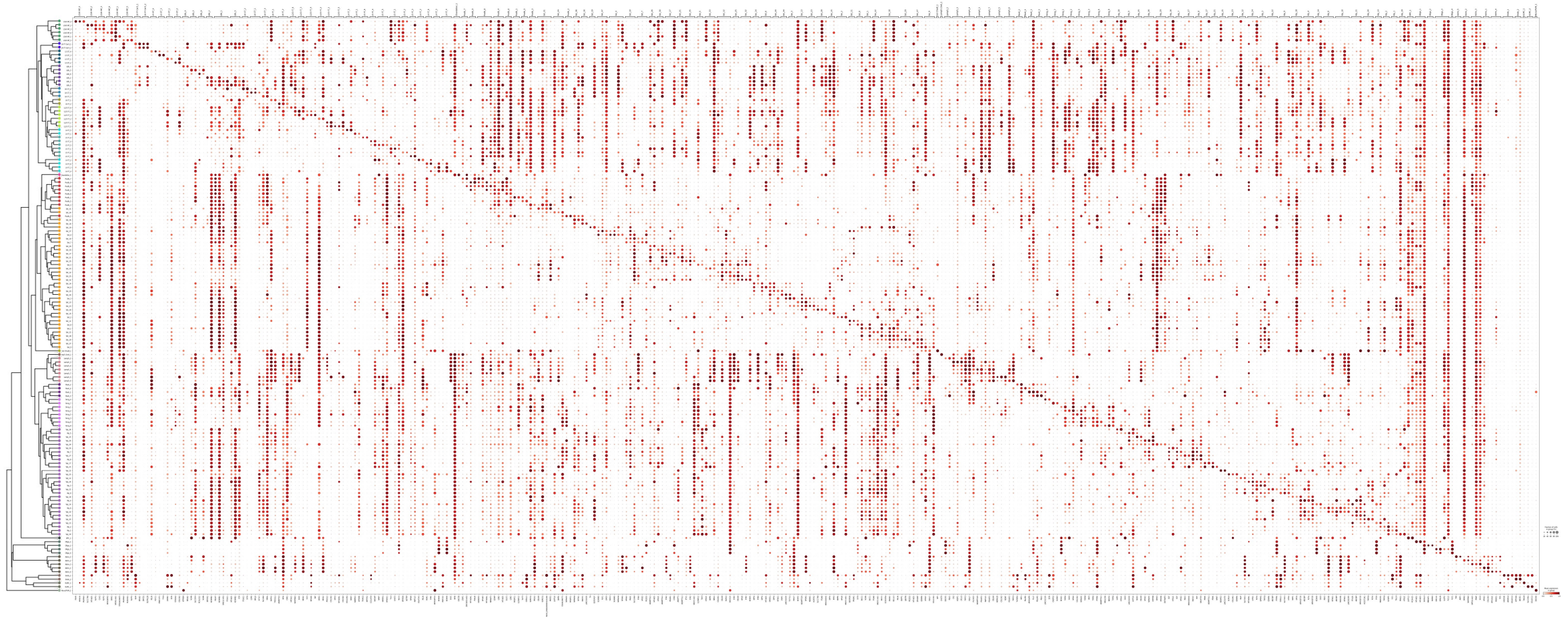


- High expression along or close to the diagonal; sparse expression in the off-diagonal supercluster branches
- For some clusters, the subclass markers are selected as the combinatorial markers, forming vague blocks along the diagonal

Cluster – “local” markers performance

local combinatorial markers evaluated on local data

consensus taxonomy



- Local markers show massive expression in the off-diagonal supercluster branches

Cluster – NS-Forest “global” combinatorial markers

- NS-Forest produces from 1 - 5 marker genes for optimal classification of subclasses
- 460 markers for 153 clusters

- LINC RNAs are selected as markers
- Canonical/parent/subclass markers are selected as needed determined by the algorithm, e.g., VIP

	clusterName	clusterSize	f_score	recall	PPV	TN	FP	FN	TP	marker_count	markers	onTarget
0	Astro_1	500	0.679688	0.348000	0.892308	72623	21	326	174	2	[SLC1A3, HPSE2]	0.088745
1	Astro_2	500	0.556680	0.220000	0.901639	72632	12	390	110	2	[LINC01411, OBI1-AS1]	0.380457
2	Astro_3	500	0.571172	0.382000	0.651877	72542	102	309	191	5	[HEPN1, F3, AGT, GJA1, AQP4]	0.226393
3	Astro_4	500	0.472798	0.292000	0.559387	72529	115	354	146	4	[PPP1R3C, ETNPPL, ID4, OBI1-AS1]	0.246185
4	Astro_5	500	0.751263	0.476000	0.878229	72611	33	262	238	2	[TNC, CD44]	1.000000
5	Chandelier_1	500	0.799748	0.508000	0.933824	72626	18	246	254	2	[CA8, GPR149]	0.180236
6	Endo_1	109	0.831354	0.642202	0.897436	73027	8	39	70	4	[COLEC12, VWF, FLT1, ABCG2]	0.470640
7	Endo_2	500	0.827088	0.618000	0.903509	72611	33	191	309	2	[HERC2P3, PRKCH]	0.528635
8	L2/3 IT_1	500	0.551771	0.324000	0.669421	72564	80	338	162	4	[PRSS12, ROS1, COL5A2, SLC38A11]	0.122426
9	L2/3 IT_2	500	0.662114	0.446000	0.753378	72571	73	277	223	3	[THSD4, TESPA1, TRPC3]	0.060345
10	L2/3 IT_3	500	0.675118	0.458000	0.765886	72574	70	271	229	4	[COL27A1, ONECUT2, MOXD1, BMPR1B]	0.244858
11	L2/3 IT_4	500	0.633705	0.364000	0.777778	72592	52	318	182	4	[TPBG, FAP, ONECUT2, COL5A2]	0.225247
12	L2/3 IT_5	500	0.506114	0.298000	0.613169	72550	94	351	149	3	[GPX3, ADCYAP1, RMST]	0.541111
13	L2/3 IT_6	500	0.550987	0.268000	0.748603	72599	45	366	134	3	[LINC00507, ADAMTS3, LY86-AS1]	0.259655
14	L4 IT_1	500	0.612903	0.304000	0.821622	72611	33	348	152	3	[TPBG, RP11-197K6.1, ST8SIA4]	0.138592
15	L4 IT_2	500	0.773810	0.546000	0.863924	72601	43	227	273	2	[C6orf141, ST8SIA4]	0.517592
16	L4 IT_3	500	0.615132	0.374000	0.733333	72576	68	313	187	3	[TMEM215, KCNH8, RORB]	0.110912
17	L4 IT_4	500	0.652624	0.378000	0.797468	72596	48	311	189	3	[STAC, KCNH8, SYT2]	0.101415
18	L4 IT_5	500	0.657680	0.322000	0.889503	72624	20	339	161	2	[COL22A1, CMAHP]	0.640083
19	L4 IT_6	500	0.670391	0.384000	0.824034	72603	41	308	192	3	[PLEKHH2, CLMN, ST8SIA4]	0.067491
20	L5 ET_1	500	0.728892	0.442000	0.870079	72611	33	279	221	2	[LINC00922, POU3F1]	0.628122
21	L5 ET_2	500	0.759375	0.486000	0.883636	72612	32	257	243	3	[ATP6V1C2, LRP2, ST8SIA2]	0.283076
22	L5 ET_3	335	0.889145	0.689552	0.958506	72799	10	104	231	4	[ONECUT1, TTC6, SEMA3D, FAM189A2]	0.570144
23	L5 ET_4	301	0.729761	0.425249	0.888889	72827	16	173	128	2	[EHF, FGF11]	1.000000
24	L5 IT_1	500	0.625000	0.456000	0.688822	72541	103	272	228	3	[CHST2, TLL1, RORB]	0.098589
25	L5 IT_2	500	0.568524	0.302000	0.729469	72588	56	349	151	4	[NPY2R, PCED1B, TLL1, RORB]	0.126065

...

Cluster – NS-Forest “global” combinatorial markers

- NS-Forest produces from 1 - 5 combinatorial marker genes for optimal classification of clusters
- 460 markers for 153 clusters

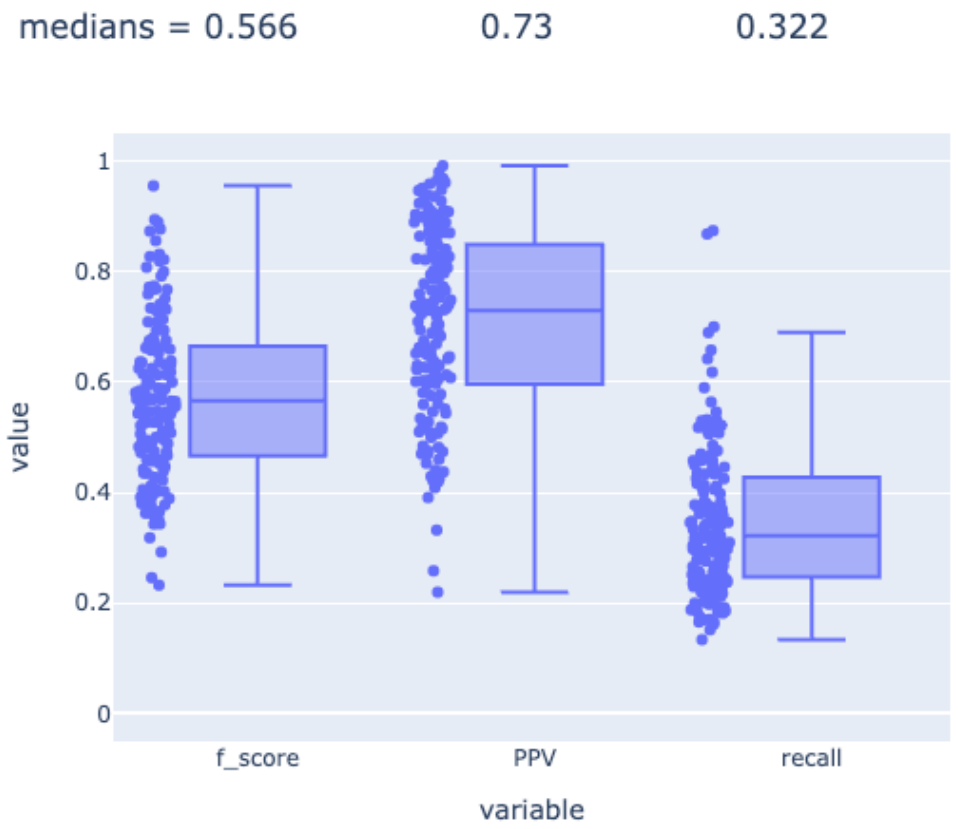
- LINC RNAs are selected as markers
- Canonical/parent/subclass markers are selected as needed determined by the algorithm, e.g., VIP

124	Vip_1	500	0.400763	0.252000	0.470149	72502	142	374	126	3	[FLT1, LRRC1, SHISA8]	0.075723
125	Vip_2	500	0.474138	0.242000	0.623711	72571	73	379	121	4	[IGFBP5, SLC22A3, PTGDS, VIP]	0.059694
126	Vip_3	500	0.402500	0.322000	0.429333	72430	214	339	161	3	[HPSE2, TRPC6, VIP]	0.043437
127	Vip_4	500	0.546875	0.238000	0.809524	72616	28	381	119	3	[WNT5A, ABI3BP, VCAN]	0.175469
128	Vip_5	500	0.480132	0.232000	0.655367	72583	61	384	116	3	[LINC01630, IQGAP2, THSD4]	0.058913
129	Vip_6	500	0.508218	0.470000	0.518764	72426	218	265	235	3	[SMOC1, IQGAP2, PARD3B]	0.066905
130	Vip_7	500	0.400593	0.216000	0.509434	72540	104	392	108	4	[CCN2, PCDH18, SEMA3C, HPSE2]	0.088324
131	Vip_8	500	0.318396	0.162000	0.419689	72532	112	419	81	3	[VWDE, WSCD1, DACH2]	0.085709
132	Vip_9	500	0.483021	0.330000	0.546358	72507	137	335	165	4	[SLC7A11, TNS3, PPP1R1C, COBLL1]	0.079606
133	Vip_10	500	0.540541	0.320000	0.653061	72559	85	340	160	3	[ABI3BP, KCNG2, PARD3B]	0.050191
134	Vip_11	500	0.435237	0.250000	0.534188	72535	109	375	125	4	[ABI3BP, MCTP2, SLC22A3, PROS1]	0.078432
135	Vip_12	500	0.588906	0.310000	0.759804	72595	49	345	155	3	[SLC7A11, HCRTR2, EPS8]	0.065852
136	Vip_13	500	0.427766	0.334000	0.460055	72448	196	333	167	3	[ST18, PTGDS, VIP]	0.048104
137	Vip_14	500	0.596847	0.318000	0.764423	72595	49	341	159	4	[ST18, KCNH8, CA10, VIP]	0.056606
138	Vip_15	500	0.617236	0.318000	0.807107	72606	38	341	159	2	[SVIL, VIP]	0.042009
139	Vip_16	500	0.380000	0.152000	0.608000	72595	49	424	76	4	[NDST4, SEMA3C, HTR2C, VIP]	0.059079
140	Vip_17	500	0.604050	0.346000	0.742489	72584	60	327	173	2	[PENK, PARD3B]	0.122873
141	Vip_18	500	0.389535	0.134000	0.744444	72621	23	433	67	3	[NOX4, RXRG, NDST4]	0.104943
142	Vip_19	500	0.535308	0.376000	0.598726	72518	126	312	188	3	[CBLN1, HTR2C, VIP]	0.060451
143	Vip_20	500	0.365854	0.240000	0.421053	72479	165	380	120	4	[CBLN1, VIPR2, PPP1R1C, VIP]	0.086078
144	Vip_21	500	0.618081	0.268000	0.917808	72632	12	366	134	2	[ALDH1A2, ROR2]	0.333300
145	Vip_22	500	0.527913	0.348000	0.606272	72531	113	326	174	3	[ABI3BP, SLC7A11, EDNRA]	0.096851
146	Vip_23	500	0.542513	0.342000	0.635688	72546	98	329	171	3	[KCNH8, EDNRA, KMO]	0.054504
147	Vip_24	500	0.362762	0.166000	0.515528	72566	78	417	83	3	[LYPD1, ENHO, CHRNA2]	0.096597
148	Vip_25	500	0.527523	0.276000	0.683168	72580	64	362	138	3	[NOX4, IQGAP2, CBLN4]	0.067031
149	Vip_26	500	0.594796	0.256000	0.888889	72628	16	372	128	4	[SEMA3C, LINC01630, HPSE2, TAC3]	0.064425
150	Vip_27	489	0.546678	0.265849	0.742857	72610	45	359	130	4	[DACH2, CHRNA7, MIR4500HG, VIP]	0.037462
151	Vip_28	189	0.376344	0.185185	0.507246	72921	34	154	35	2	[HTR2C, DCN]	0.062984
152	Vip_29	71	0.491803	0.253521	0.642857	73063	10	53	18	5	[IQGAP2, KCNJ2, TAC1, MYO5B, VIP]	0.029573

Cluster classification using NS-Forest “global” combinatorial markers

- F-beta scores range from 0.23 - 0.96 with a median of 0.57
- PPV (precision) scores range from 0.22 - 0.99 with a median of 0.73
- Recall scores range from 0.13 - 0.87 with a median of 0.32

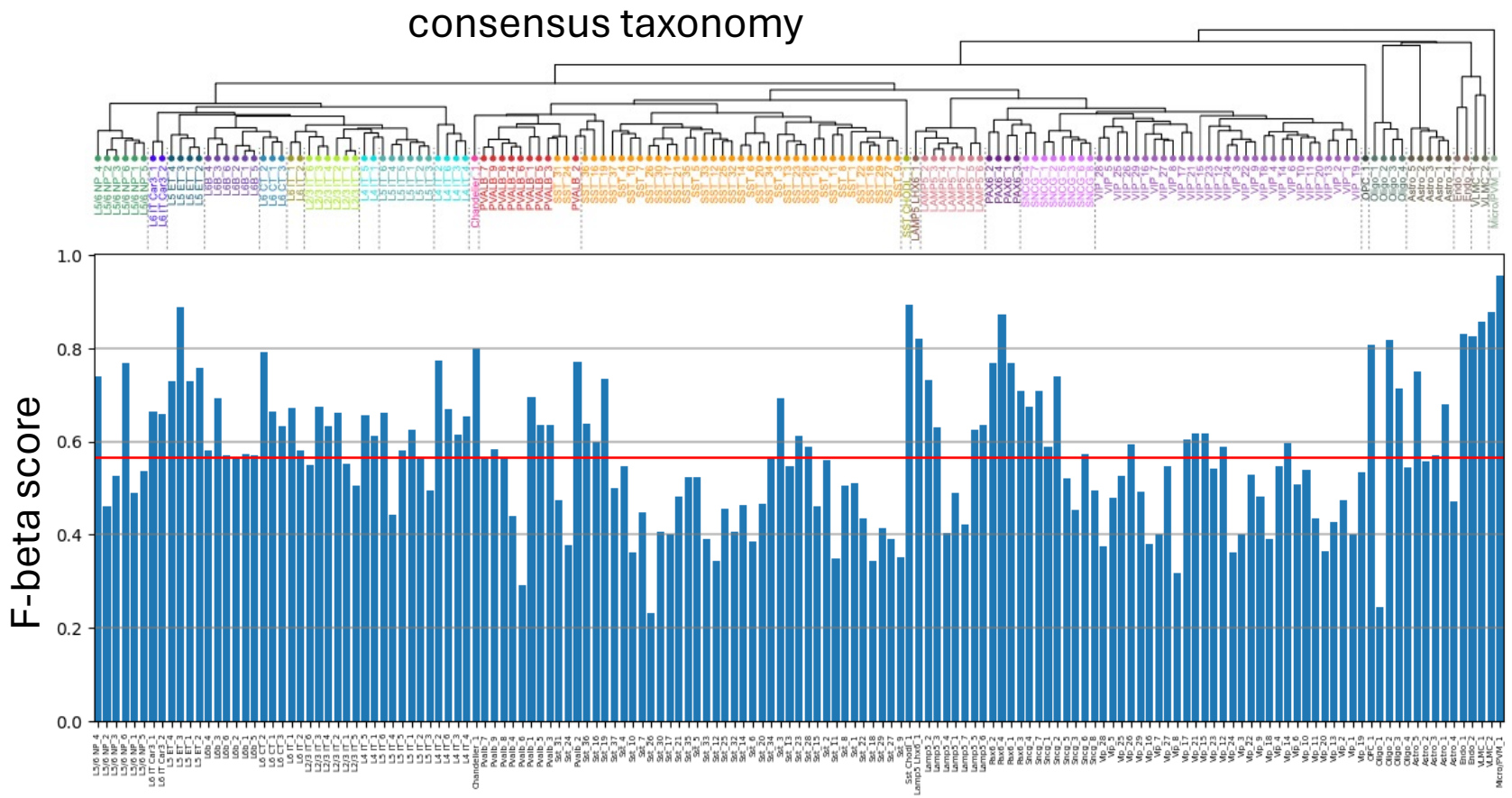
- The F-beta score is often lower than the PPV due to false negative predictions that would be impacted by the expression dropout artifact seen in snRNA-seq experiments
- Recall is largely impacted by the false negative predictions



PPV = TP / (TP + FP); recall = TP / (TP + FN)

F-beta score as a cluster quality metric

- The non-neuronal branch tends to have high F-beta scores for its clusters, except for Oligo_1
- There is a mixture of high and low F-beta scores within each major branch of inhibitory and excitatory neurons
- Low F-beta scores are suggestive of poor cluster quality due to over-partition
- High F-beta scores are suggestive of a clear segregation of those clusters on the UMAP, e.g., PVALB_2, SST_19, PAX6_4

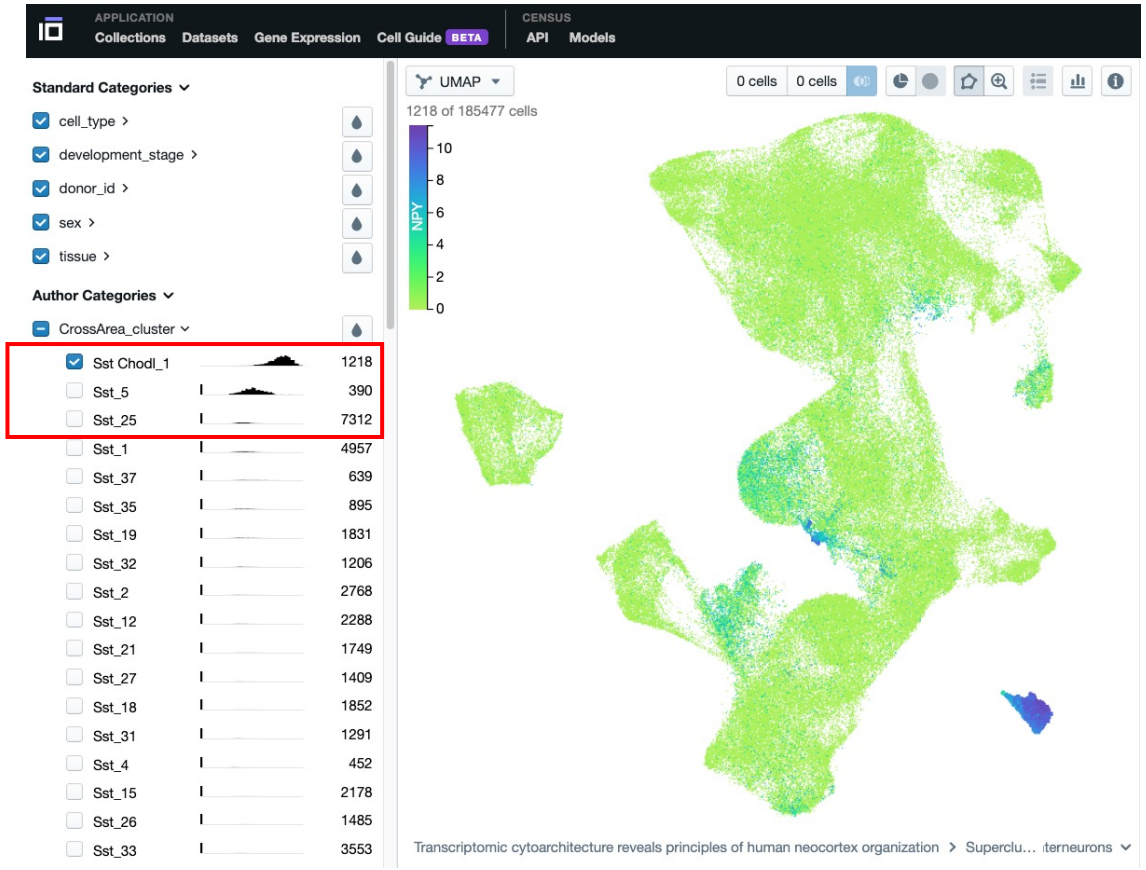


Cluster marker expression in clusters

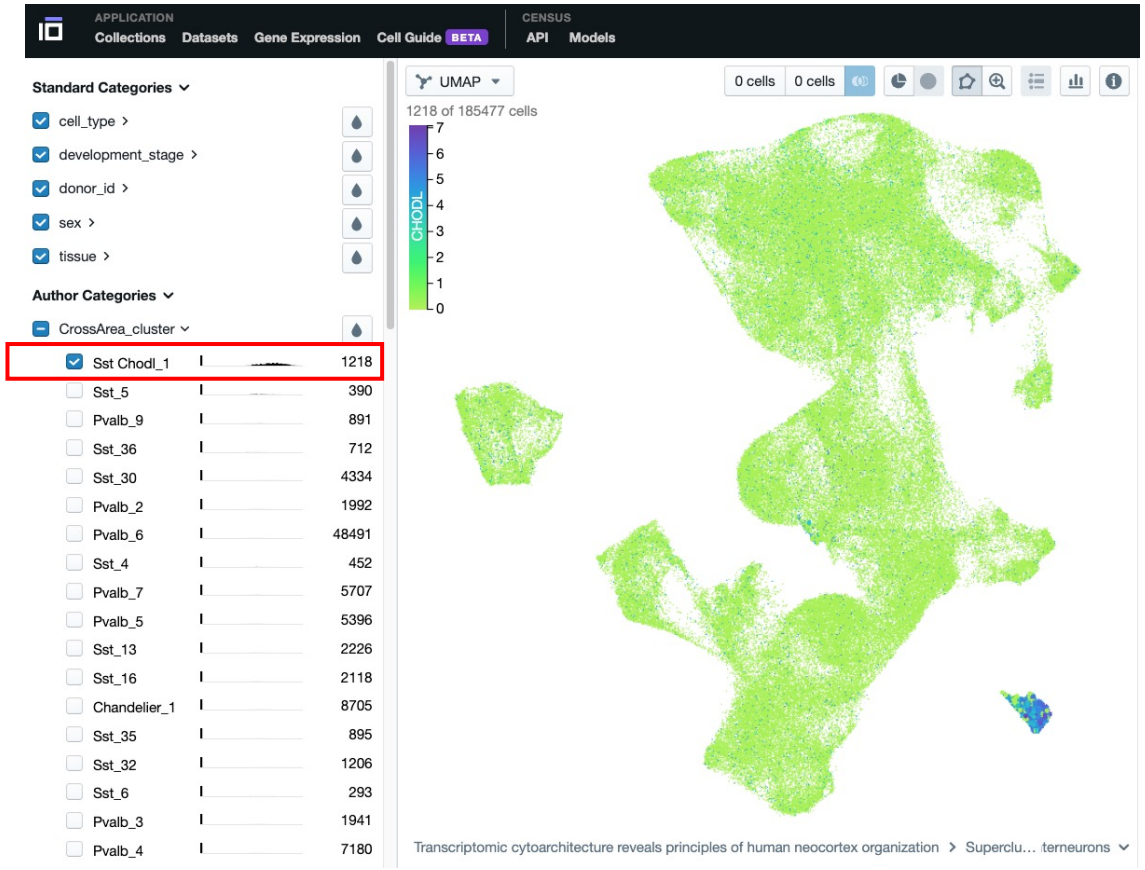
Cluster Sst Chodl_1: one highly expressed and specific marker (F-beta = 0.89)

Expression in Supercluster: MGE-derived interneurons

NPY



CHODL



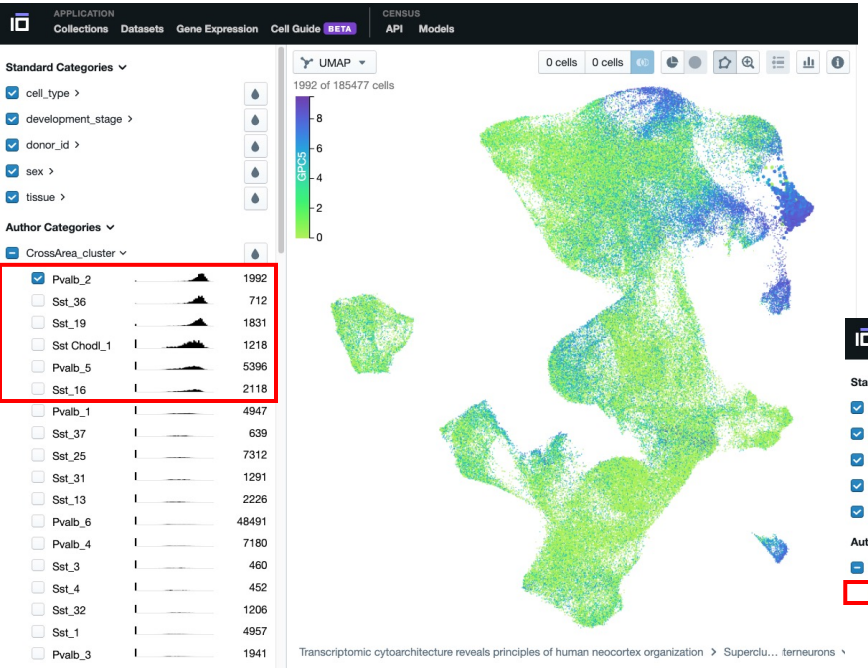
NPY is highly expressed in the target cluster and lowly expressed in a couple other clusters ==> lower FN

CHODL is highly specific and moderately expressed (N.B. ***not*** an NS-Forest marker) ==> risk of FN

Cluster Palvb_2: three markers in combination give high cluster specificity (F-beta = 0.77)

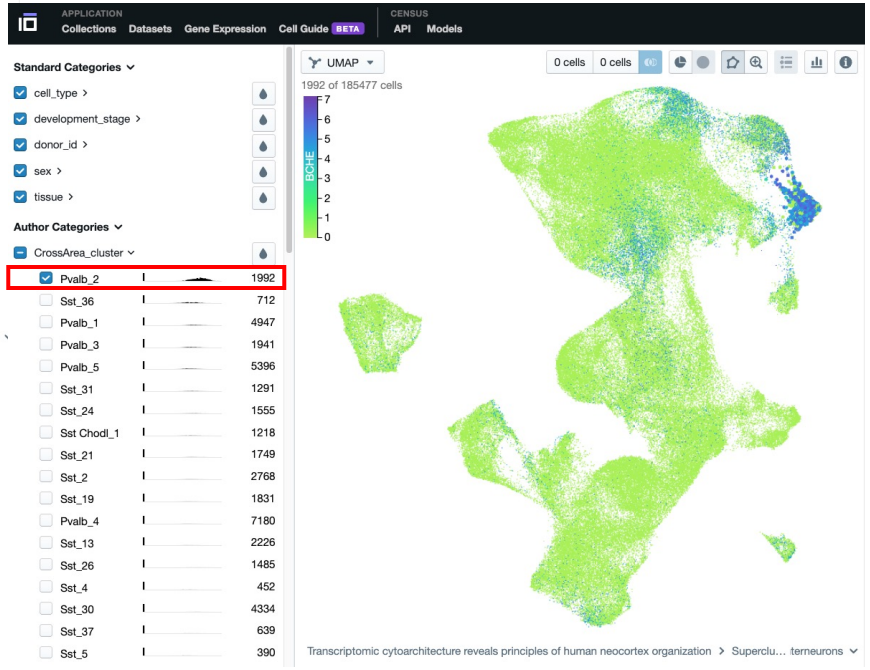
Expression in Supercluster: MGE-derived interneurons

GPC5

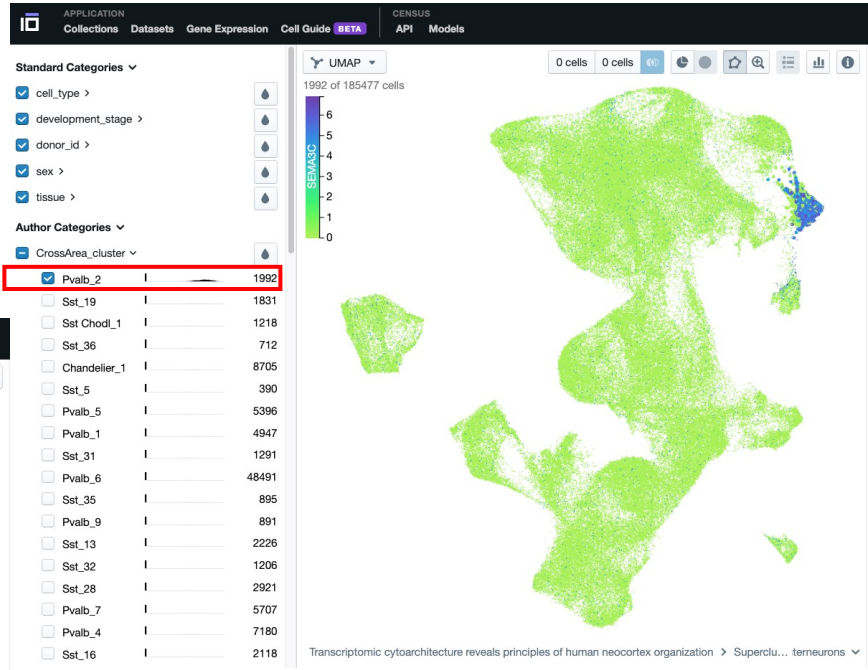


GPC5 is most highly expressed in the target cluster and expressed in several other clusters

BCHE



SEMA3C

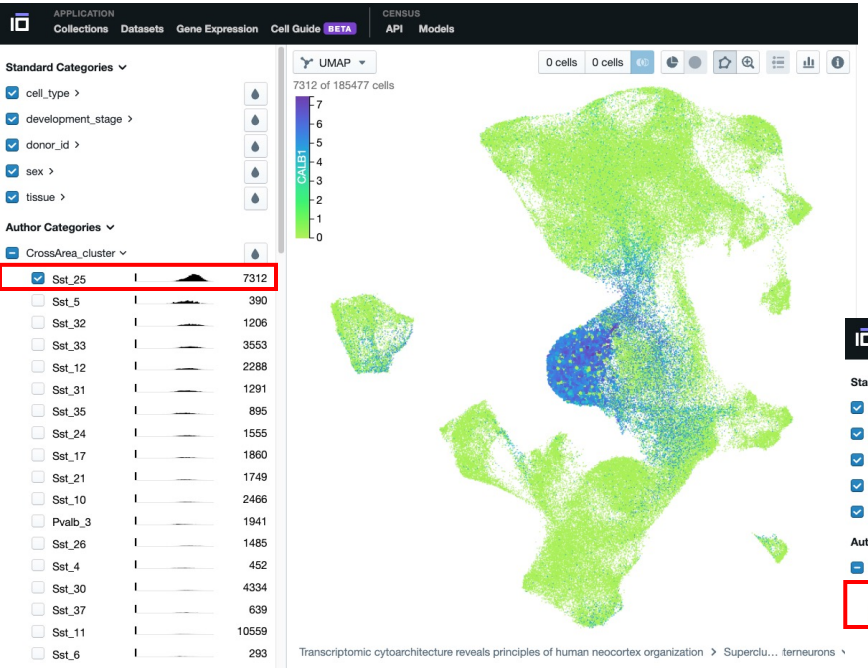


BCHE and SEMA3C in combination give the specificity of the target cluster

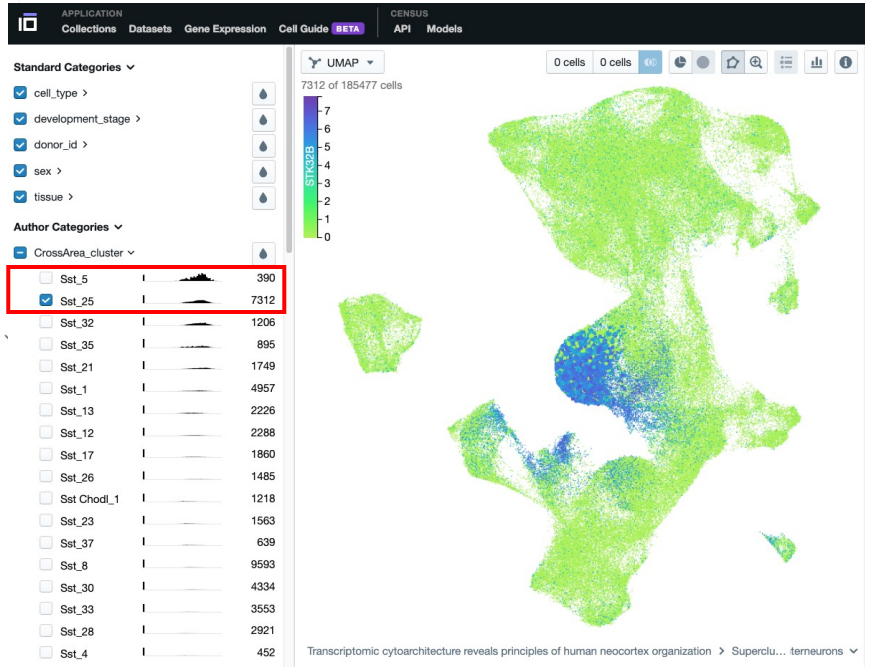
Cluster Sst_25: three markers in combination give moderate cluster specificity (F-beta = 0.46)

Expression in Supercluster: MGE-derived interneurons

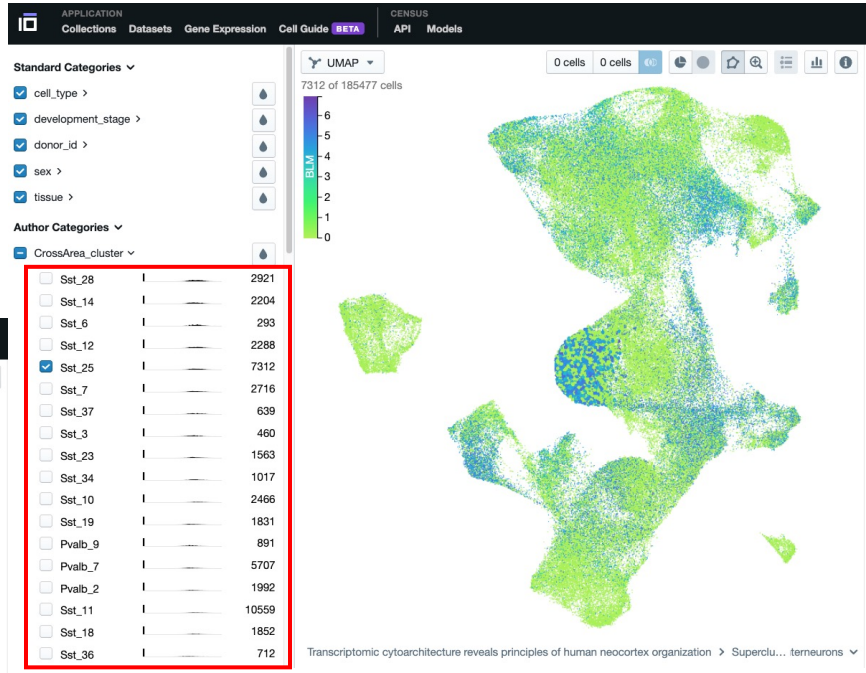
CALB1



STK32B



BLM

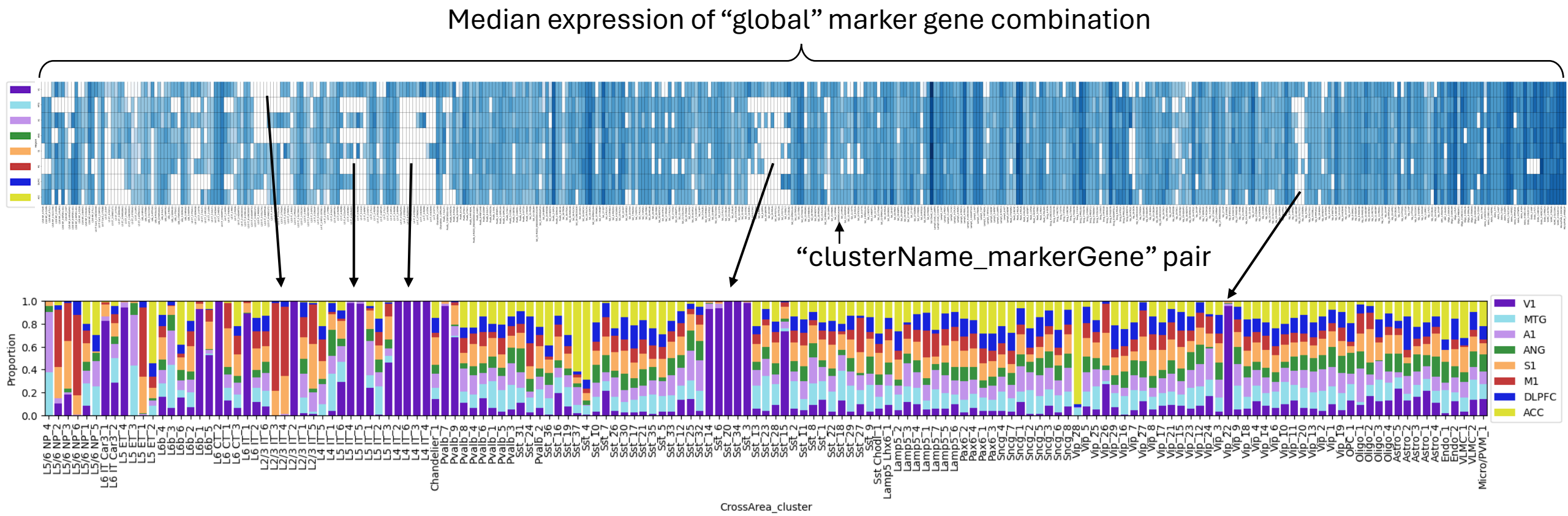


CALB1 and STK32B are highly expressed in the target cluster and scatteredly expressed in some neighboring clusters, e.g., Sst_5

BLM is selected to eliminate FPs in the neighboring clusters

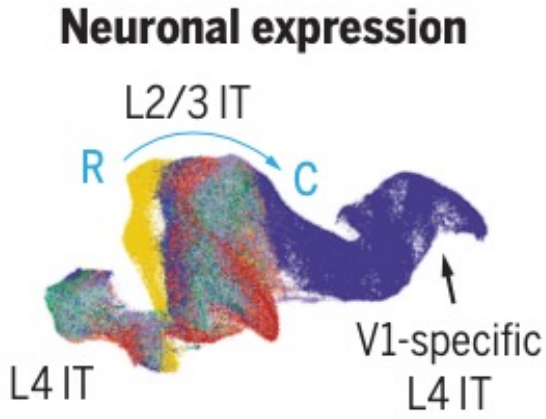
Cluster “global” marker combination expression across brain regions

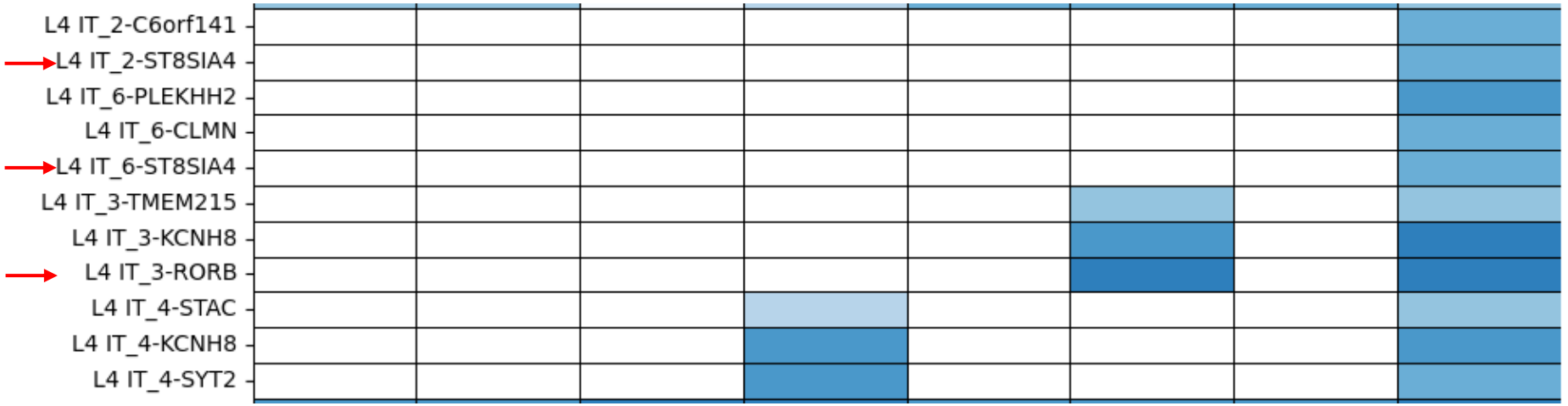
NS-Forest marker genes also show DEG-like expression across regions



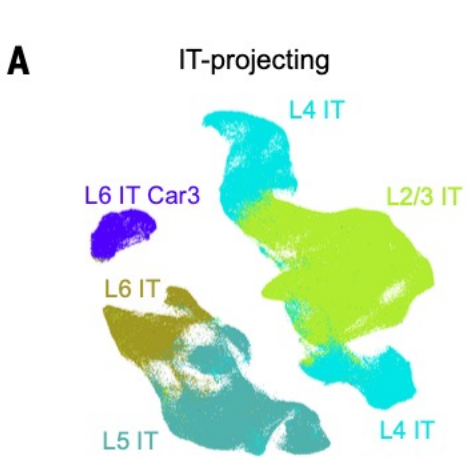
V1-specific L4 IT clusters

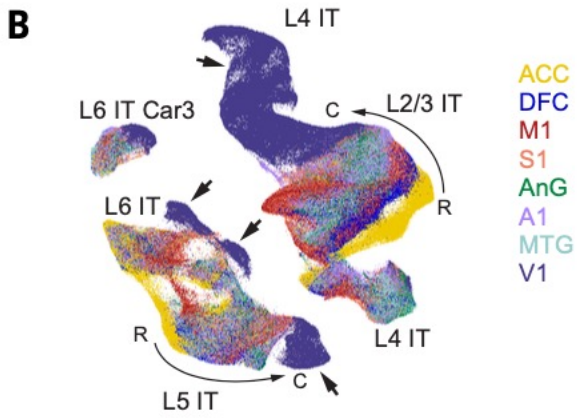
→ L4 IT subclass markers

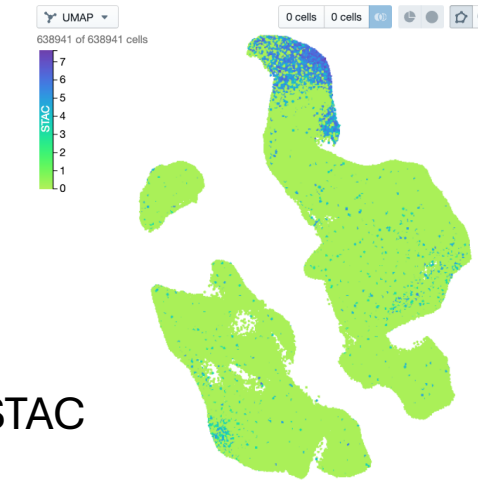




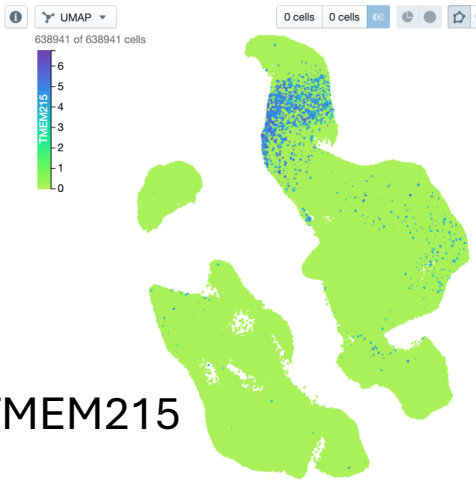
V1-specific L4 IT clusters



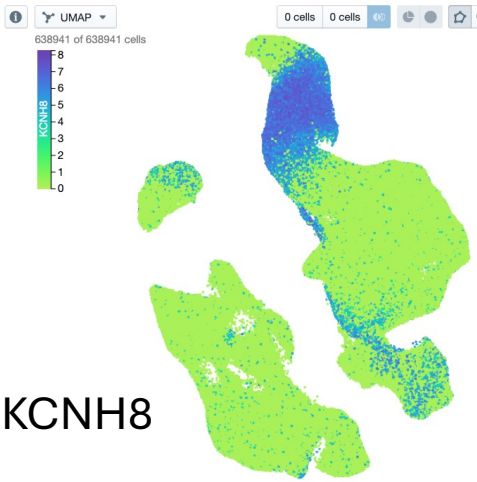




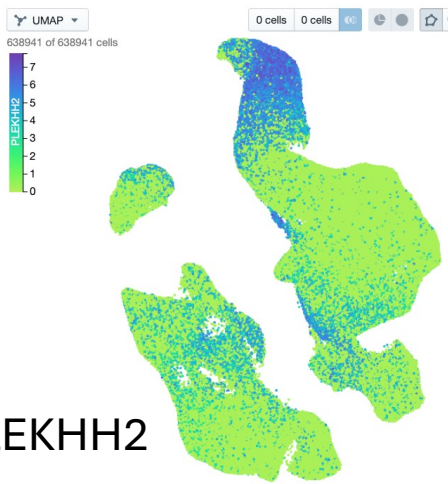
STAC



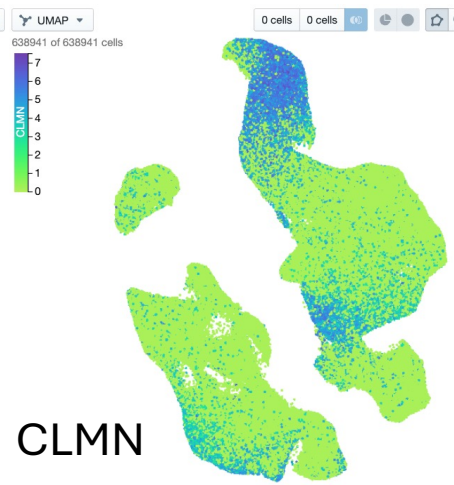
TMEM215



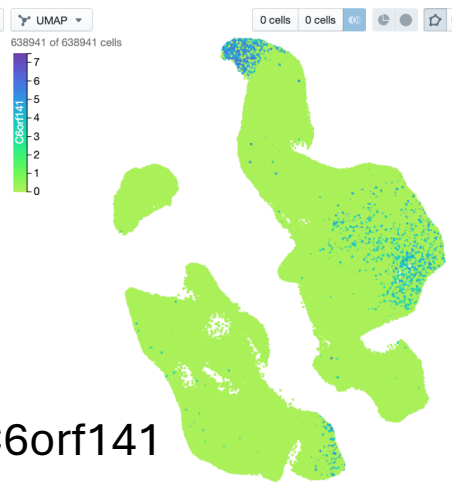
KCNH8



PLEKHH2



CLMN



C6orf141

Conclusion

- **NS-Forest markers are computational, quantitatively associated with a confidence value (F-beta score), and reproducible**
  - Zenodo doi: 10.5281/zenodo.13172641
- NS-Forest “global” marker genes for subclasses look very good
  - Good subclass specificity
  - Ubiquitous expression in each of the clusters within the subclass
- NS-Forest “global” combinatorial marker genes for clusters also look good in many clusters, indicated by decently high F-beta scores
  - The best combination of the subclass + cluster “global” marker set is algorithmically determined by the F-beta score evaluated in the NS-Forest evaluating module
- F-beta scores seem to track with the segregation of “clusters” in the UMAP space
  - Well segregated clusters = high marker scores
  - Connected/over-partitioned clusters = low marker scores
- PPV tend to be higher than F-beta and recall tend to be lower than F-beta, suggesting that FN are more of a problem than FP

Nomenclature proposal

- For example, clusterName = subclassName_clusterMarker(Combination):
  - Lamp5 Lhx6_1 → Lamp5 Lhx6_LAMP5_CHST9
  - Sst Chodl_1 → Sst Chodl_NPY
  - Pvalb_2 → Pvalb_ BCHE_SEMA3C_GPC5
  - L4 IT_2 → L4 IT_ C6orf141_ST8SIA4
  - L4 IT_6 → L4 IT_ PLEKHH2_CLMN_ST8SIA4
  - L6 IT Car3_1 → L6 IT Car3_LINC00348
  - L6 IT Car3_2 → L6 IT Car3_ SMYD1_ITGB8
- one cluster marker or marker combination (1-5)?
- if only one marker, how to choose?
  - Highest median expression level (favors recall by reducing FNs), e.g., Pvalb_2 → Pvalb_GPC5
  - Highest binary expression score (favors cluster specificity)), e.g., Pvalb_2 → Pvalb_BCHE
  - Highest single marker F-beta score, e.g., Pvalb_2 → Pvalb_???