Explainable Artificial Intelligence (XAI) and Single Cell Genomics to Understand the Cellular Complexity of the Human Brain

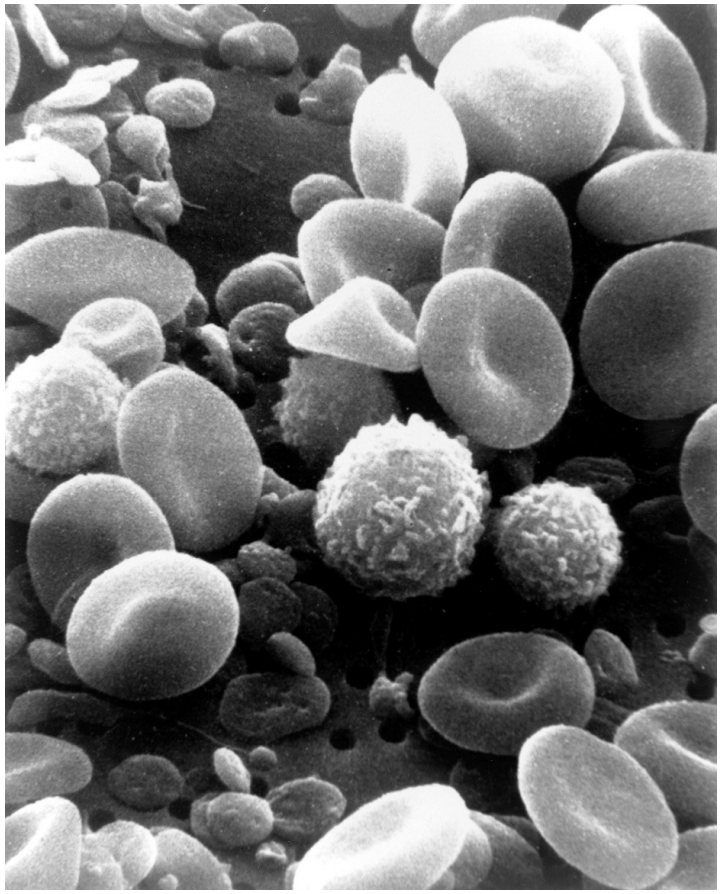
Richard H. Scheuermann, PhD
Scientific Director
National Library of Medicine

April 04, 2024

Cellular profiling

Cell Phenotype

- Cells are the most fundamental functional units of life
- Different cell types play different physiological roles in the body
- Cell identity and function (phenotype) is dictated by the subset of genes/proteins expressed
- Abnormalities in the expressed genome (disorders) form the physical basis of disease
- Understanding normal and abnormal cellular phenotypes and their genomic makeup is key for diagnosing disease and for identifying therapeutic targets



Bruce Wetzel & Harry Schaefer, National Cancer Institute, 1982
http://en.wikipedia.org/wiki/Image:SEM_blood_cells.jpg

Technology

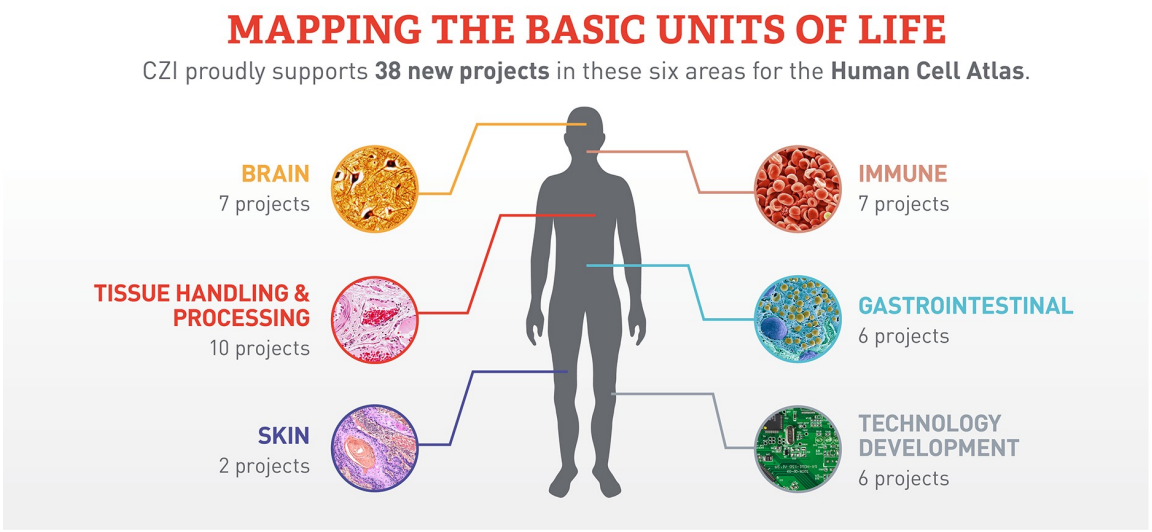
- Transcriptional profiling of bulk samples obscures the cellular complexity of tissues
- Single cell/nucleus RNA sequencing (scRNA-seq) allows us to quantify cellular phenotypes in an unbiased fashion, enabling the evaluation of both known and novel cell subsets in complex tissue samples

Data-driven

- Machine learning and Explainable Artificial Intelligence (XAI) have emerged as valuable tools to characterize this complexity

HUMAN CELL ATLAS

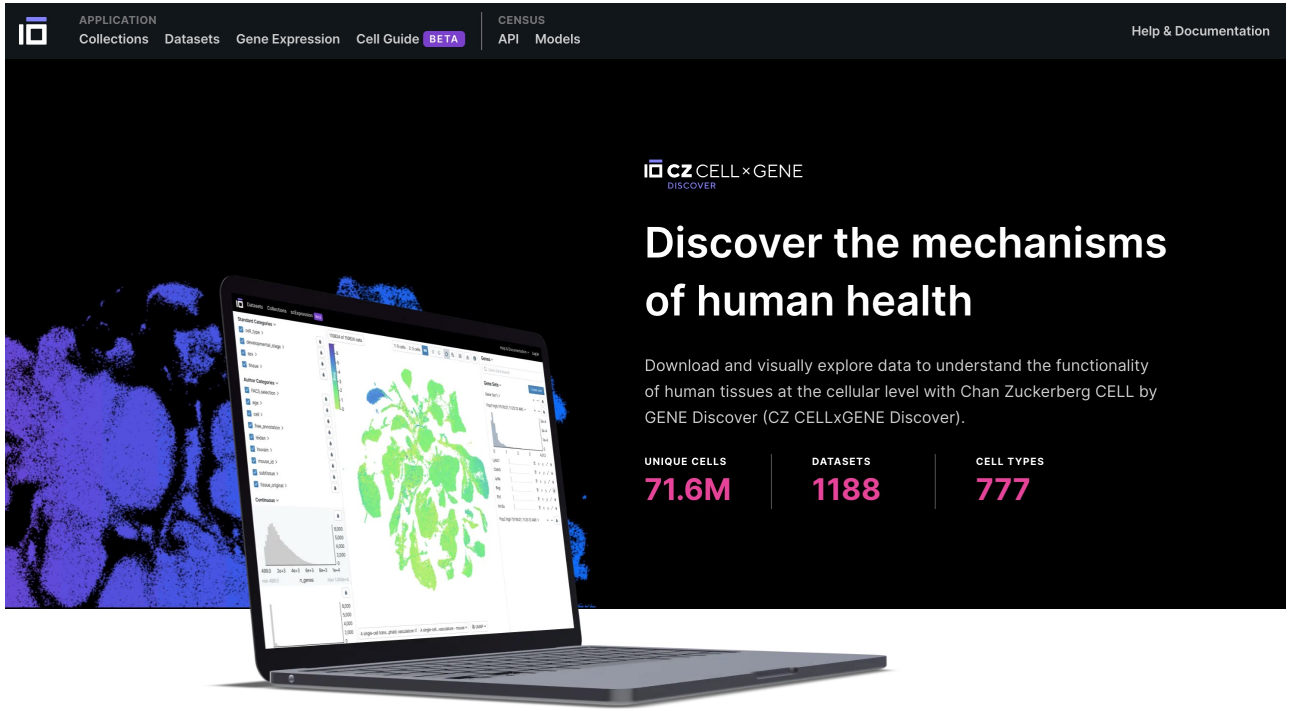
- Comprehensive reference maps of all human cells as basis for understanding fundamental human biological processes and diagnosing, monitoring, and treating disease
- Led by Aviv Regev (Broad) and Sarah Teichmann (Sanger)
- Brings together experts in biology, medicine, genomics, technology development and computation
- Utilize standardized experimental and computational methods to compare diverse cell and tissue types and samples across human communities in consistent ways

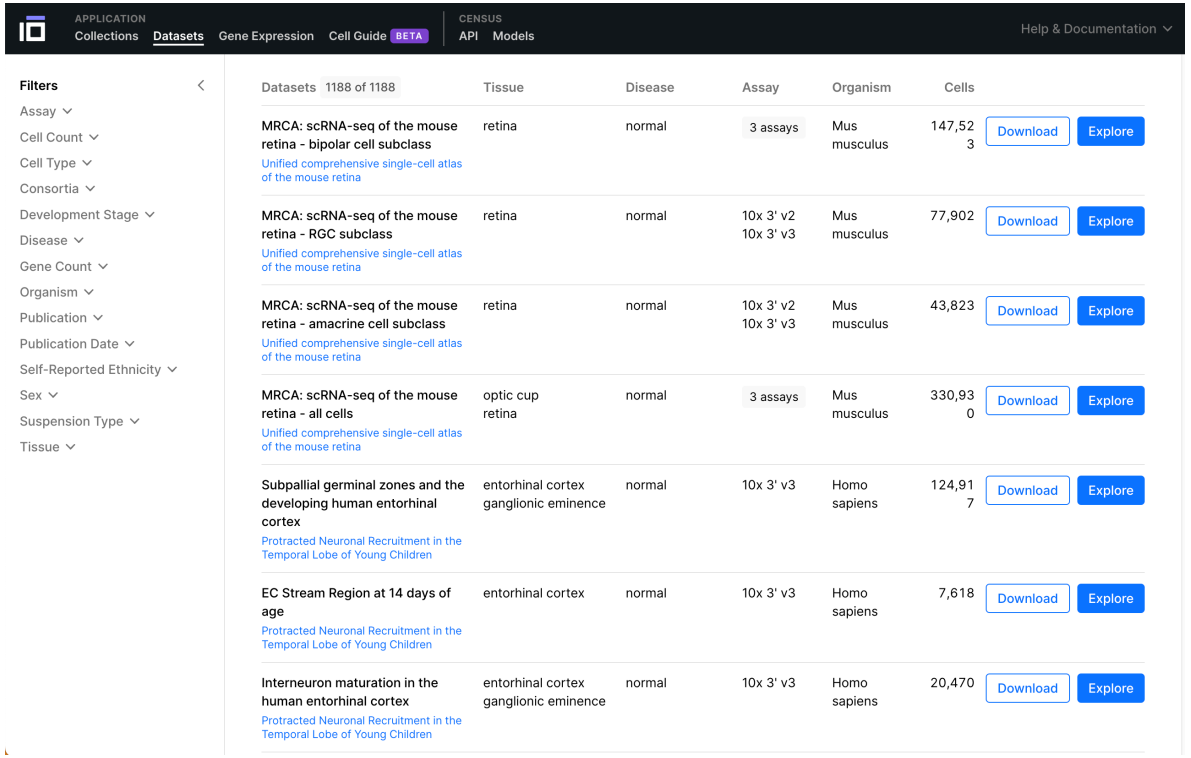


scRNA-seq publications (Pubmed January 8, 2024)

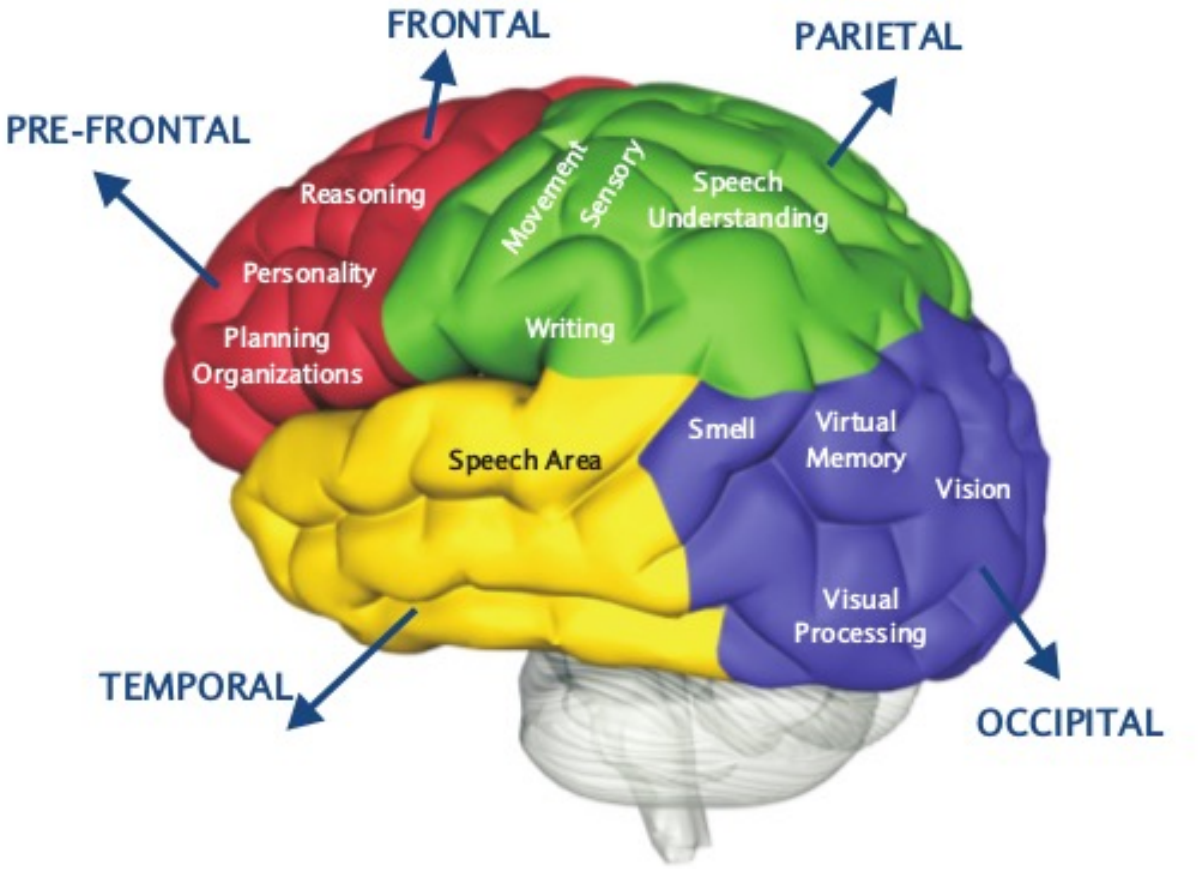
- “single cell RNA sequencing” – 11,611
- “single cell RNA sequencing” AND human – 7112
- “single cell RNA sequencing” AND human; publication date one year – 2234
- “single cell RNA sequencing” AND human; publication date one year; free full text – 1590
    - AND lung – 197
    - AND brain – 198
    - AND kidney – 73

Single Cell Data Repositories





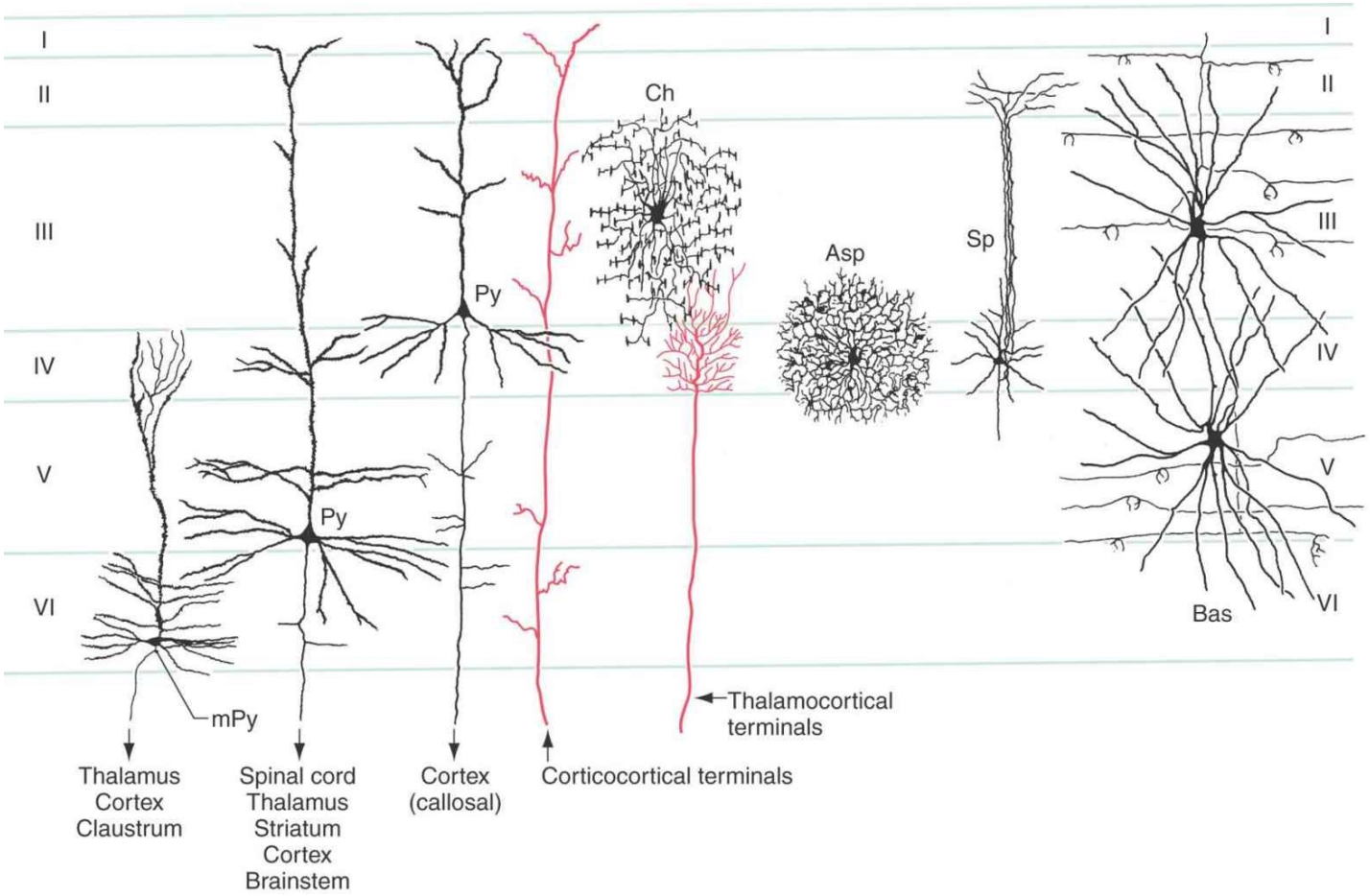
Human neocortex

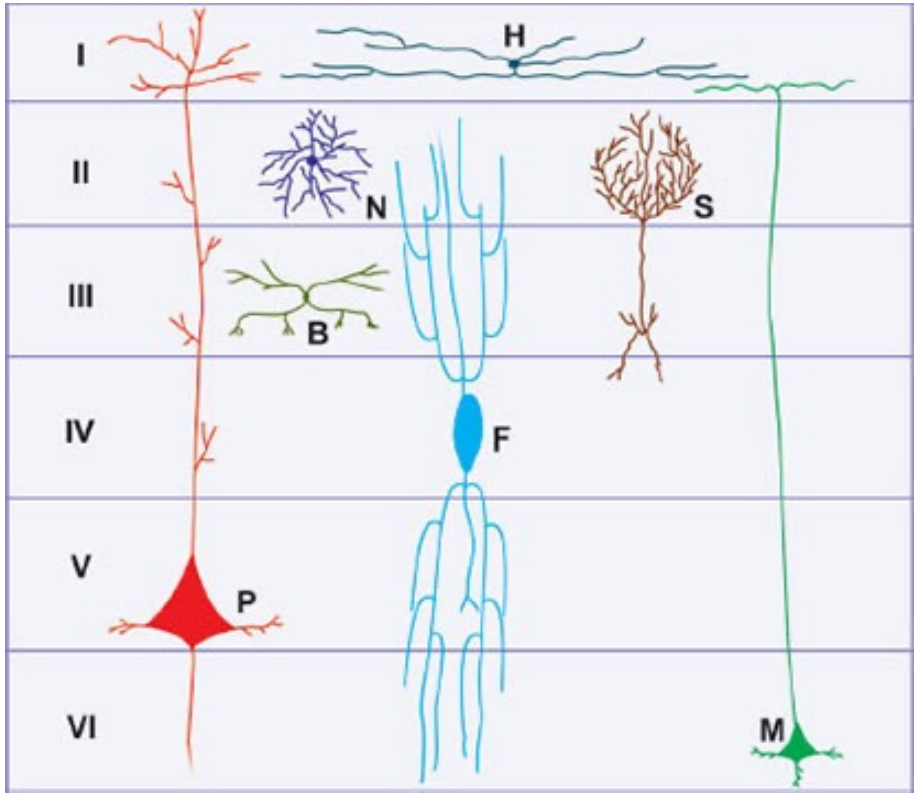


- Neocortex is part of higher vertebrate brain
- Top layer of the cerebral hemispheres, 2-4 mm thick
- Deep grooves (sulci) and wrinkles (gyri) in primates and other mammals serve to increase the area of the neocortex considerably
- In humans, accounts for about 76% of the brain's volume
- Involved in higher-level functions such as spatial reasoning, sensory perception, generation of motor commands, conscious thought, and in humans, language

ALLEN INSTITUTE for BRAIN SCIENCE

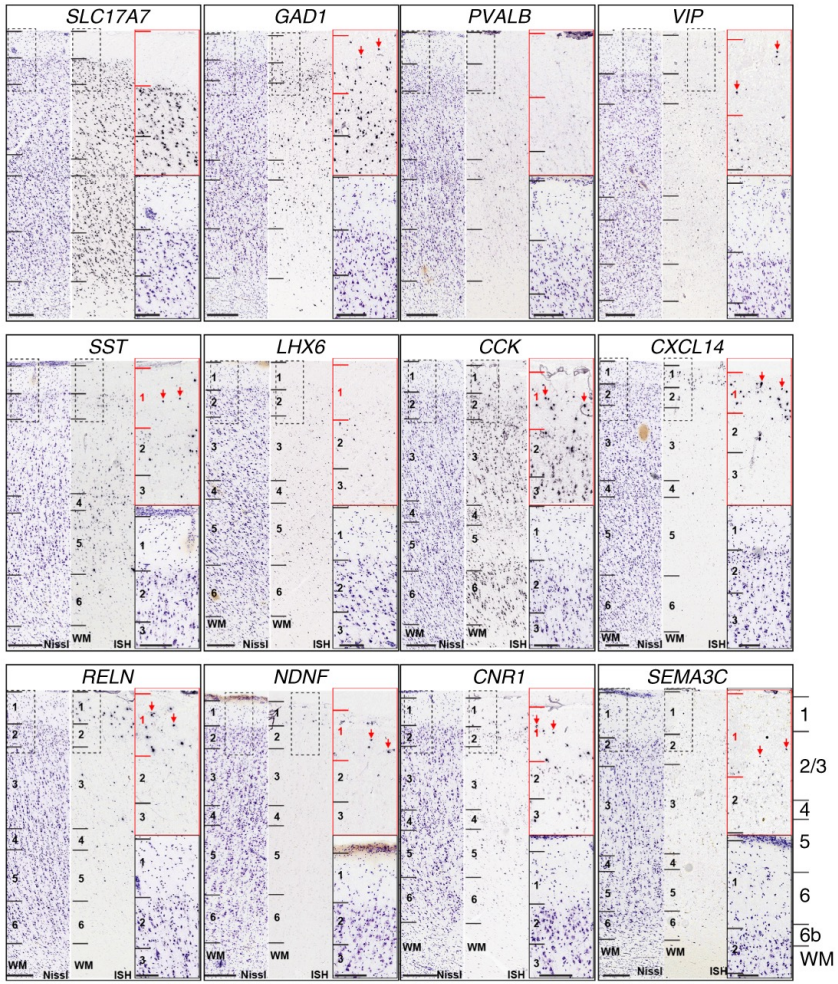
Morphological cell types

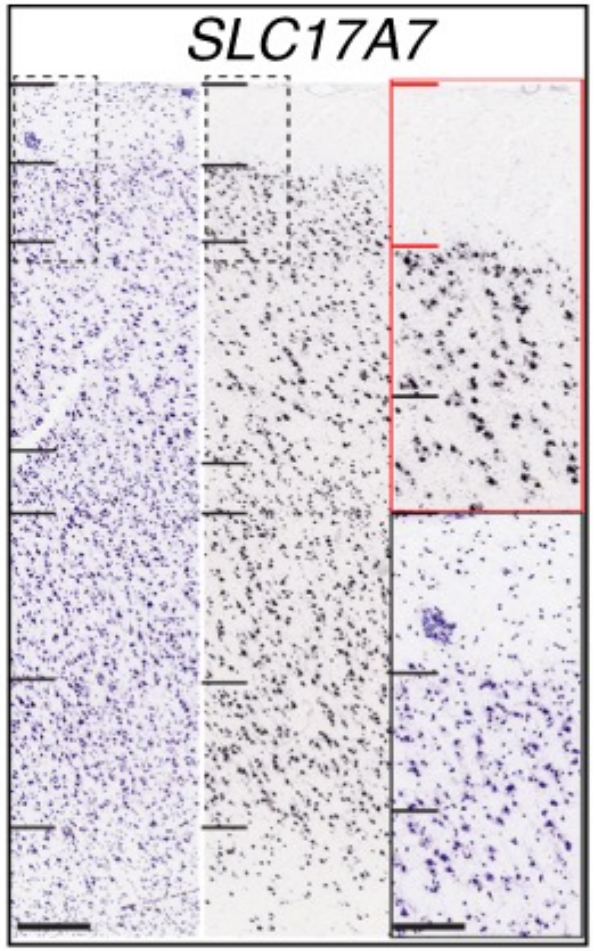


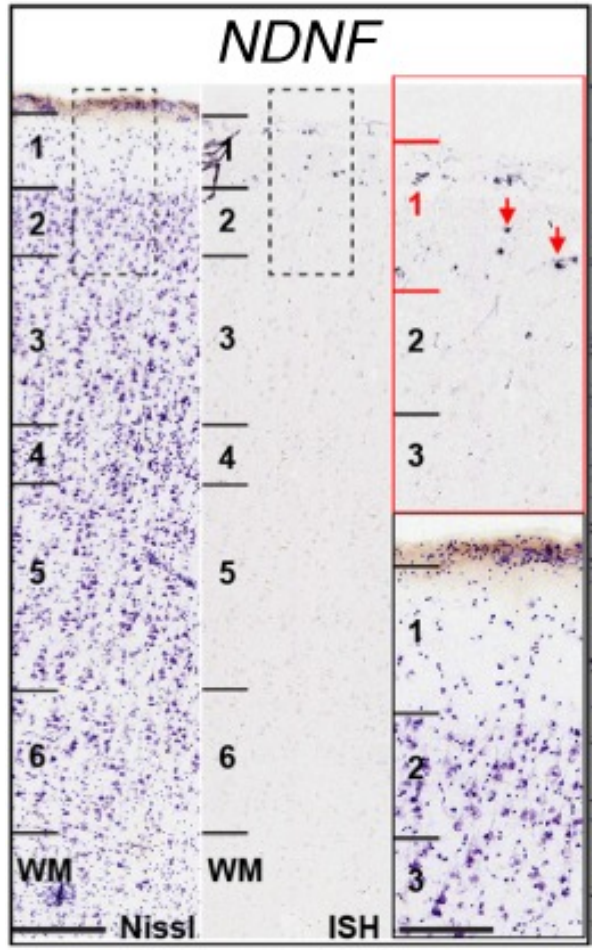


https://neupsykey.com

Gene expression complexity

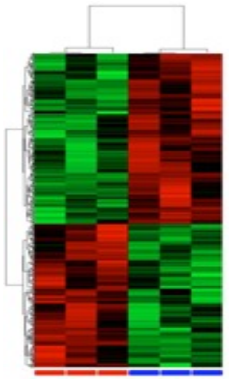






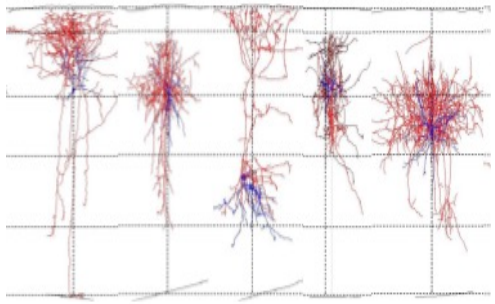
Large-scale quantitative cell phenotyping of human neocortex

Transcriptomic



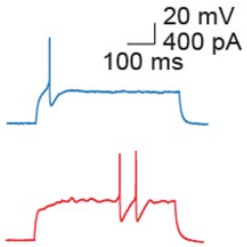
RNA-seq

Anatomical



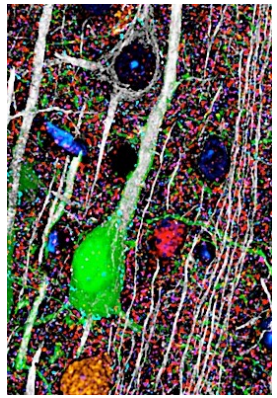
Quantitative morphology

Physiological



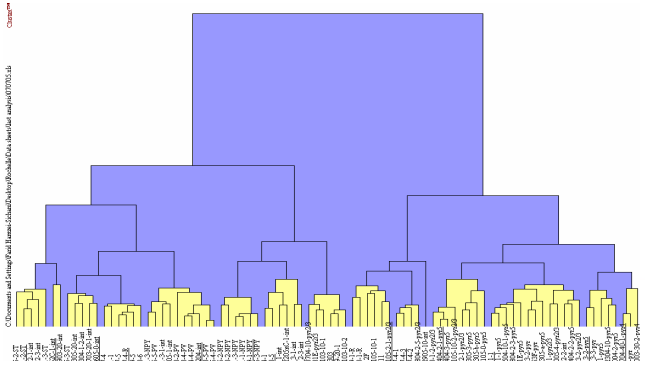
High-throughput, standardized electrophysiology

Synaptomic



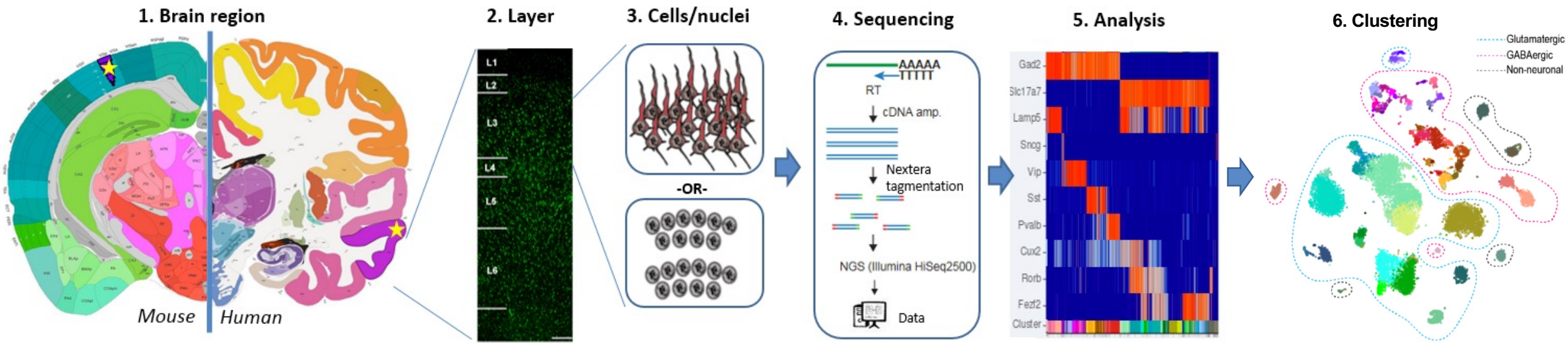
Array tomography

Data-driven gold standard taxonomy of human neocortical neuron types



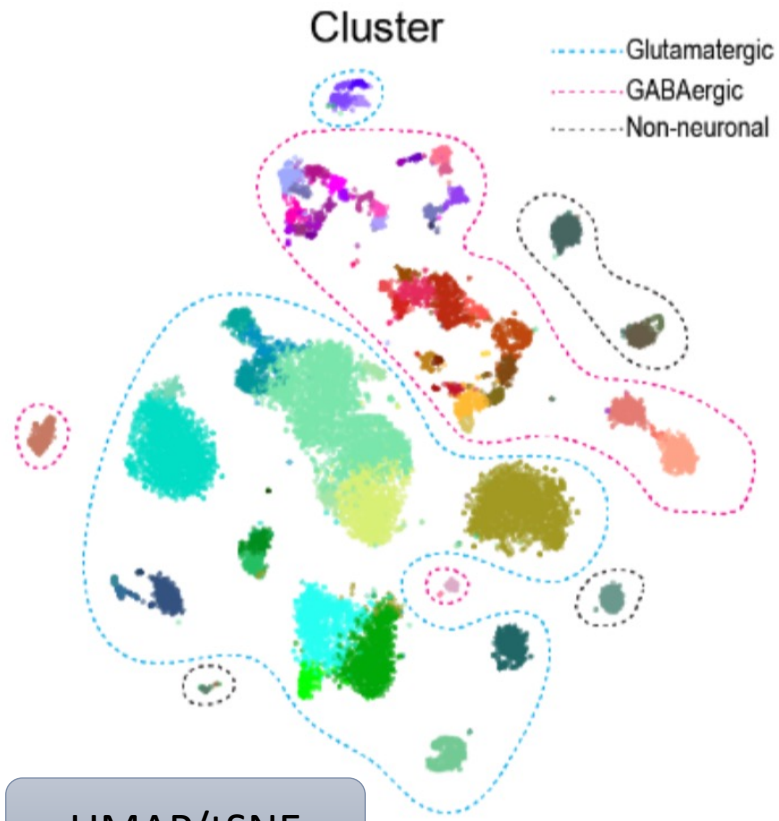
ALLEN INSTITUTE for BRAIN SCIENCE

Single cell/nucleus transcriptional profiling



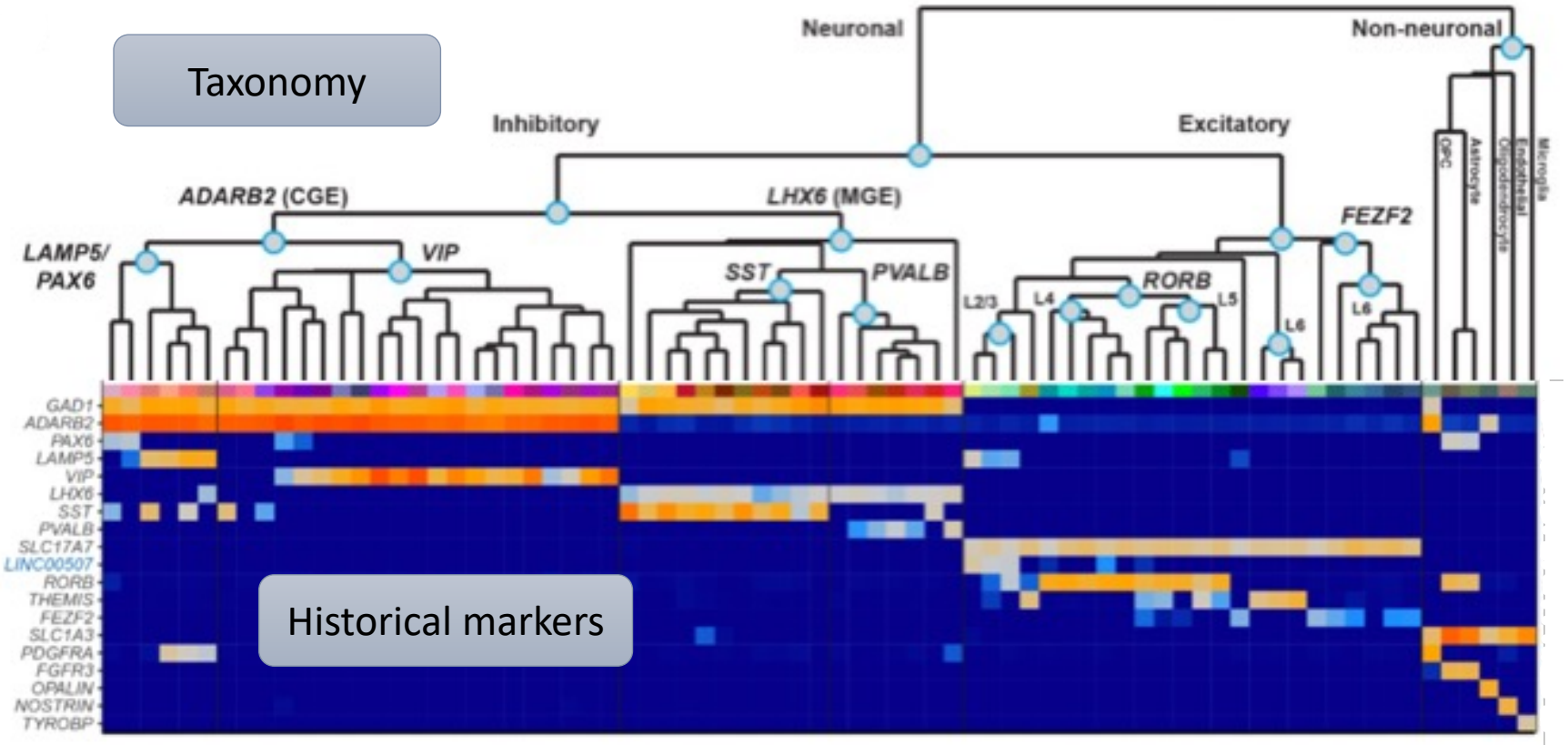
transcriptome cluster = cell type (state)

MTG cell type clusters, taxonomy & marker genes



UMAP/tSNE

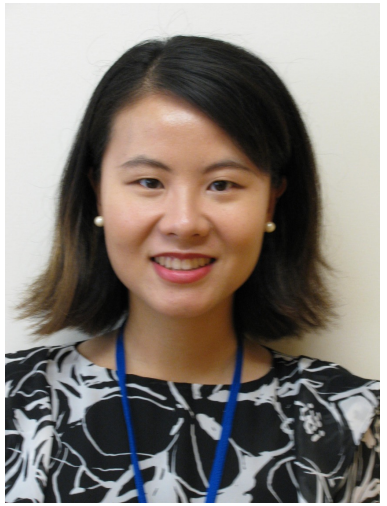
Taxonomy



Historical markers

Hodge RD, et al. (2019) ***Nature***, 573:61-68. PMID: 31435019

Brian Aevermann



Yun (Renee) Zhang

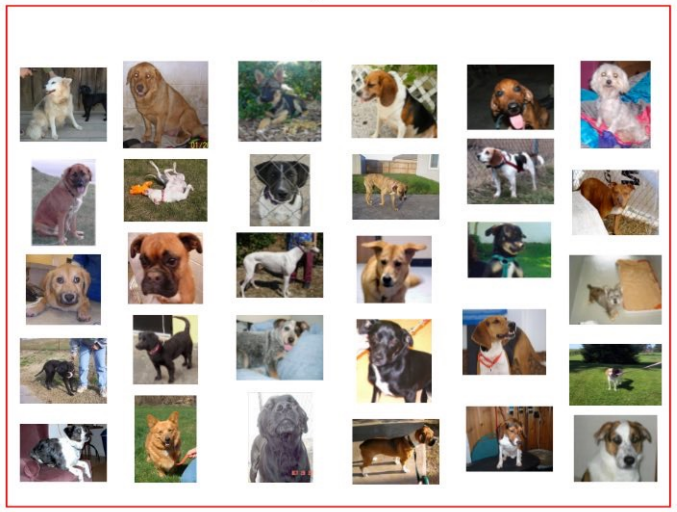
A machine learning method for the discovery of minimum marker gene combinations for scRNA-seq cell types

Aevermann B, et al. (2018) ***Human Molecular Genetics***, 27(R1):R40-R47. PMID: 29590361
Aevermann B, et al. (2021) ***Genome Research***, 31:1767-1780. PMID: 34088715

Deep Learning vs Explainable AI

Cats

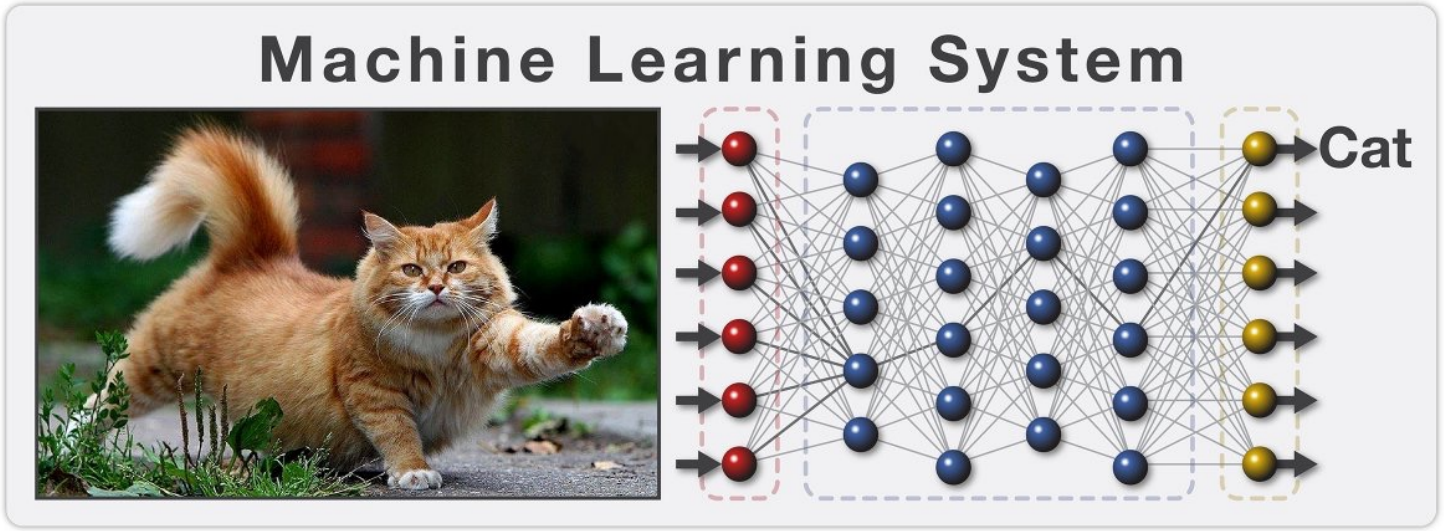
Dogs



Sample of cats & dogs images from Kaggle Dataset

Classification

Model



This is a cat.

Current Explanation

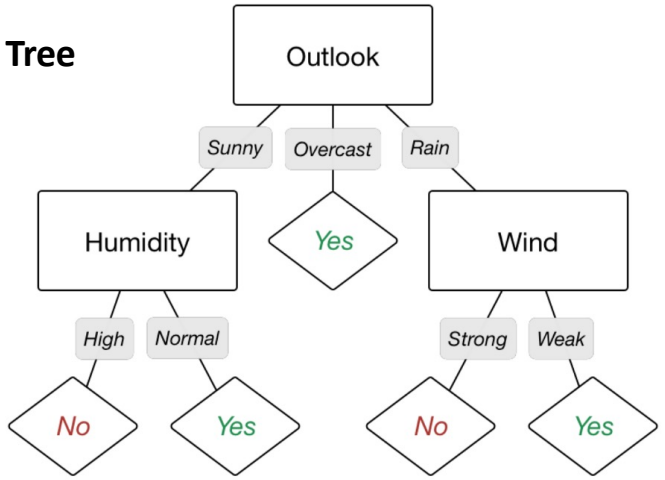
This is a cat:

- It has fur, whiskers, and claws.
- It has this feature:

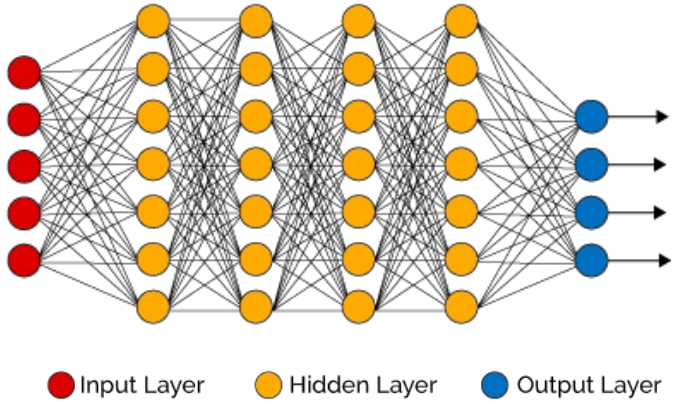
XAI Explanation

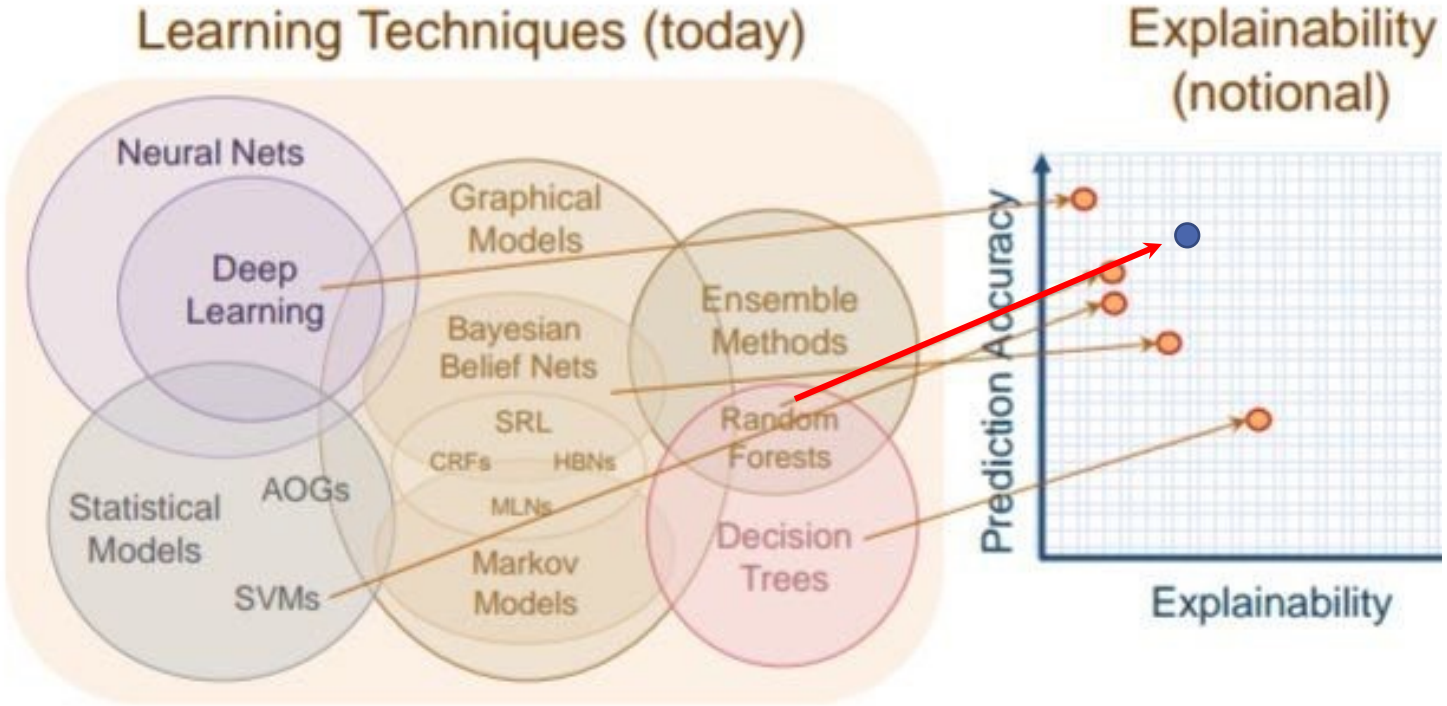
Machine Learning - predictability vs explainability

Decision Tree



Deep Learning Neural Network





Random Forest machine learning

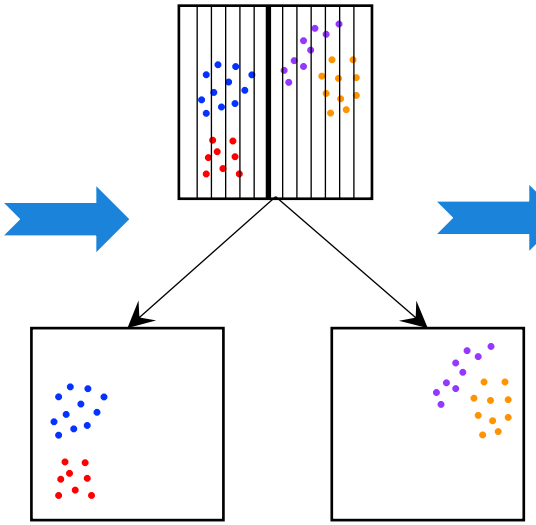
Input Data

Samples w/class labels

Features

Gene Symbol	Class 1 S14	Class 2 S15	Class 2 S16	Class 2 S17	Class 1 S18	Class 2 S19	Class 2 S20	Class 2 S21	Class 1 S22	Class 1 S23	Class 2 S24
MT-RNR2	339	613	140	218	89413	120640	139967	160362	273	163988	0
CD74	15360	13089	21823	21392	24882	20302	19912	25084	14456	16941	18638
TMSB4X	50165	18343	15620	32732	40384	16916	11749	36453	27991	30943	20930
B2M	20406	15628	19978	19748	22110	14254	8704	14587	16703	19325	18680
MT-CO1	20205	0	19549	13159	13606	15279	17448	13562	14957	12125	3
HLA-DRA	9551	12798	14487	17090	23299	13436	11494	17408	17121	12250	10376
IGKC	3777	11	14850	0	13388	4820	5	0	10	0	9569
MT-CO2	12056	4023	5702	11151	8121	0	0	0	6462	0	0
RPS27	9936	10313	10302	6110	6628	5688	8472	5843	11830	5884	10436
CD52	7617	7035	9779	5945	4362	2476	2513	6463	7762	5325	4561
EEF1A1	3661	3552	9027	3541	1511	7760	7053	3436	3022	6151	12150
RPS29	8982	7872	8448	7906	4226	4487	5214	5728	7879	6077	7171
MT-ND4	4190	3507	2974	7447	3284	2429	7711	0	3372	4382	6369
ACTB	17824	3629	3968	15718	6459	8588	4727	6444	8128	9309	7223
RPS18	11326	7980	8524	4179	3677	6585	7146	3058	5267	5173	6677
RPL21P16	3121	6470	6451	3853	3577	6080	4536	2977	7517	4224	10916
MT-CYB	0	5588	2291	11447	3082	0	5935	2635	7585	3826	1623
RPL34	5160	6945	6623	3273	2705	5042	4876	3156	5928	2903	6893
RPL41	3858	6840	6377	2908	2645	4830	2180	5758	5793	2647	5205
IGHM	0	2968	656	0	0	0	0	6537	0	0	936
RN7SL1	13005	6307	6191	4417	10911	5988	3641	8246	3950	3619	2753
RPL30	4474	3518	7651	4303	1529	4306	7066	1959	5252	5233	5595
MT-ND5	2855	2080	1917	1934	2266	3154	4741	4714	4757	2445	5466
RPL3	1914	2895	3160	2586	1126	4470	2925	715	4065	1887	5440
RPL19	4113	6488	3501	4298	2755	4893	3807	3352	5229	4730	4460
RPL11	5211	5041	3409	2218	2170	3085	3523	2031	3166	2246	3712
RPL39AP42	3549	3340	2447	3856	2334	1905	2234	2829	3443	2888	5763
TXNIP	474	2040	4779	2227	161	3030	2940	667	1181	2125	1918
RPL8	2276	3983	2088	2566	888	1327	4015	2103	4488	2159	6341
RPL39P3	2930	3843	5589	2181	3494	6272	1801	1517	6538	4656	4305
RPL5	1288	1208	2635	5087	1049	6754	7225	607	3409	1987	10201

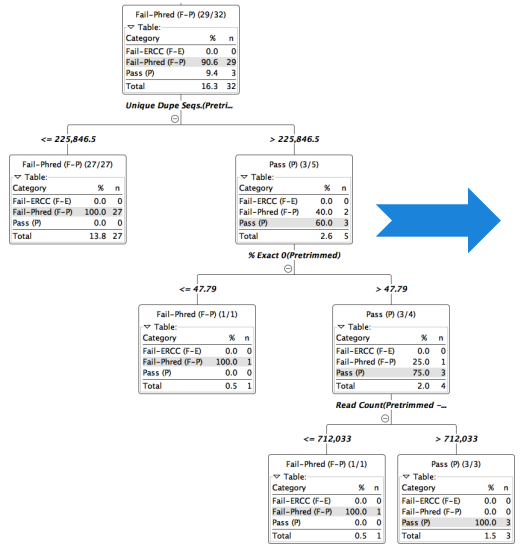
Data Partitions



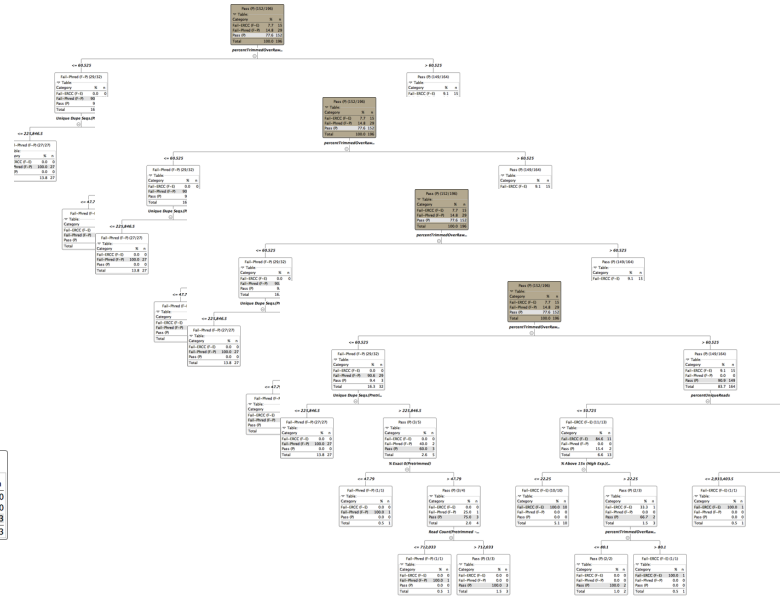
$$H(X) = -\sum_{i=1}^{n} P(x_i) log_b P(x_i)$$

IG = entropy (parent) - weighted average entropy (children)

Decision Tree



Random Forest Model



Iterate

Vote

1. Randomly select x samples/cells, where x = $\sqrt{n}$
2. Randomly select y features/genes, where y = $\sqrt{m}$
3. Calculate entropy for parent data and all partitions of each y
4. Calculate information gain from parent to all partitions of each y
5. Select y feature with largest IG
6. Repeat steps for partitioned data until all entropy = 0
7. Repeat for z bootstraps (with replacement) => z decision trees (bagging)
8. Determine how frequently a given feature is used, if available

Ranked RF gene list – Cluster e1

row ID	#splits (level 0)	#splits (level 1)	#splits (level 2)	#candidates (level 0)	#candidates (level 1)	#candidates (level 2)	Ratio1	Ratio2	Ratio3	Rank
TESPA1	18086	24430	25429	18089	36378	70833	1.00	0.67	0.36	2.03
SLC17A7	15601	16234	16083	18167	36339	70819	0.86	0.45	0.23	1.53
KCNIP1	12457	19432	21572	18042	36696	70604	0.69	0.53	0.31	1.53
SLIT3	8438	14845	17379	18388	36377	70917	0.46	0.41	0.25	1.11
SLIT3.1	6874	12440	14792	18285	36156	71322	0.38	0.34	0.21	0.93
ATP1B2	6732	10398	11850	18160	36295	71142	0.37	0.29	0.17	0.82
ZNF536	3824	11652	19314	18135	36230	70942	0.21	0.32	0.27	0.80
CUX2	7510	7808	10439	18120	36336	70460	0.41	0.21	0.15	0.78
SLIT3.2	5574	10337	12054	18376	36373	70868	0.30	0.28	0.17	0.76
DLX6-AS1	3604	11319	15088	18162	36497	70603	0.20	0.31	0.21	0.72
NECAB1	2201	11373	19263	18149	36384	70840	0.12	0.31	0.27	0.71
CUX2.1	5967	5949	7596	18131	36060	71232	0.33	0.16	0.11	0.60
IGF1	387	5976	12803	18027	35904	71120	0.02	0.17	0.18	0.37
ADARB2-AS1	865	6604	6608	18267	36231	70792	0.05	0.18	0.09	0.32
GAD2	595	6216	7760	18207	36252	70796	0.03	0.17	0.11	0.31
ADARB2-AS1.1	561	5302	5191	18342	36744	70985	0.03	0.14	0.07	0.25
GAD2.1	385	4980	6245	18147	36186	70982	0.02	0.14	0.09	0.25
BTBD11	299	2366	6124	18208	36548	70715	0.02	0.06	0.09	0.17
PROX1	0	2737	6310	18104	36474	71310	0.00	0.08	0.09	0.16
LINC01105	39	1943	4630	18168	36547	70871	0.00	0.05	0.07	0.12
COL21A1	0	1134	4016	18068	36419	70383	0.00	0.03	0.06	0.09
DLX1	1	1489	3142	18258	36574	71176	0.00	0.04	0.04	0.08

Marker gene identification using NS-Forest

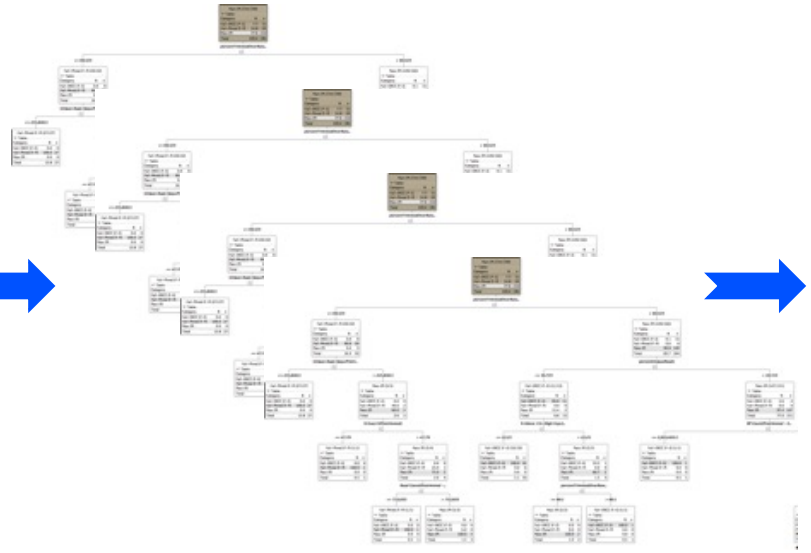
A computational strategy to identify the minimum number of necessary and sufficient features (marker genes) to define a class (cell type)

	Cluster 1					Cluster 2			
	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7	Cell 8	Cell 9
Gene 1									
Gene 2									
Gene 3									
Gene 4									
Gene 5									
Gene 6									
Gene 7									
Gene 8									

Expression matrix

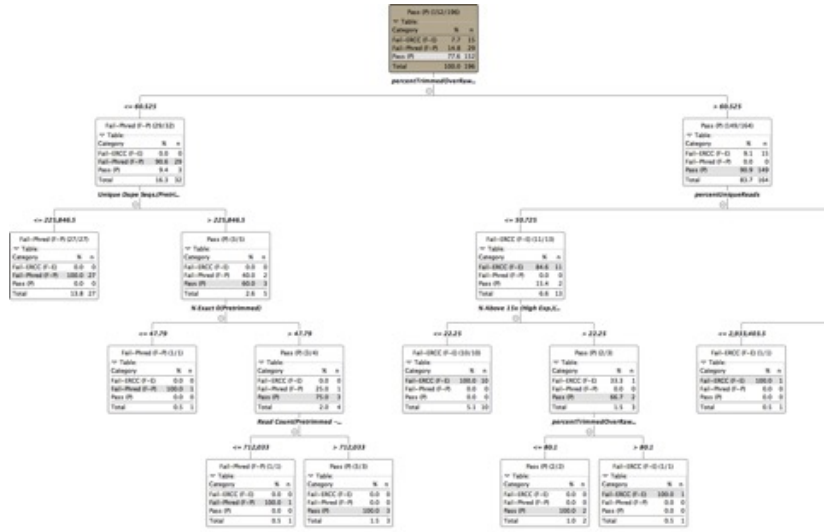
Cluster membership & gene expression data

Random forest
Cluster X vs non-Cluster X



Ranked list of useful genes

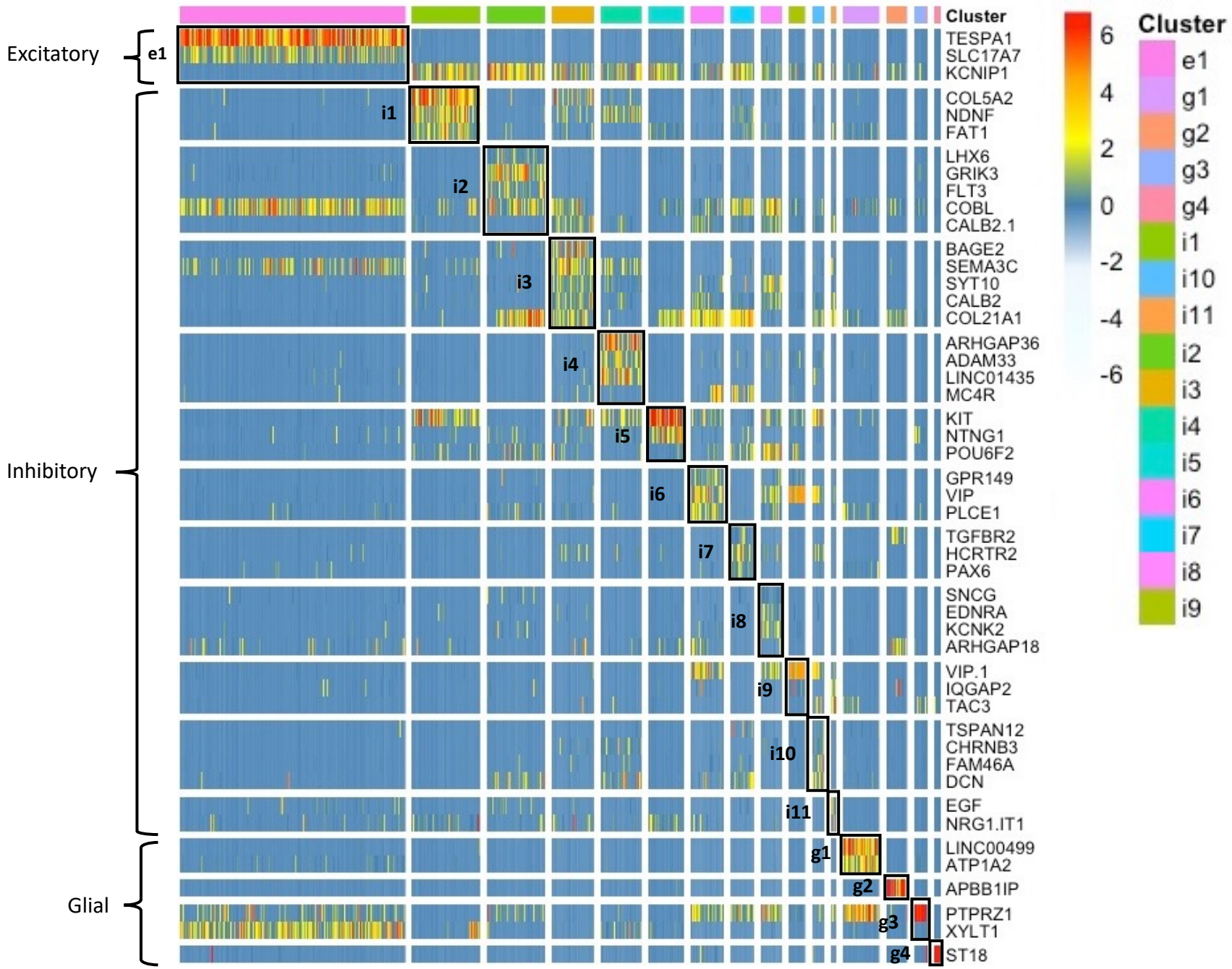
Decision trees 1,2,3…



Maximize classification accuracy (F1)
Necessary and sufficient genes =
Cluster-specific marker genes

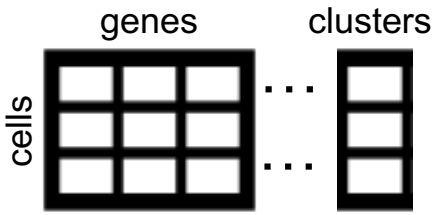
Aevermann B, et al. (2018) **Human Molecular Genetics**, 27:R40-R47.

Marker gene expression patterns grouped by cluster

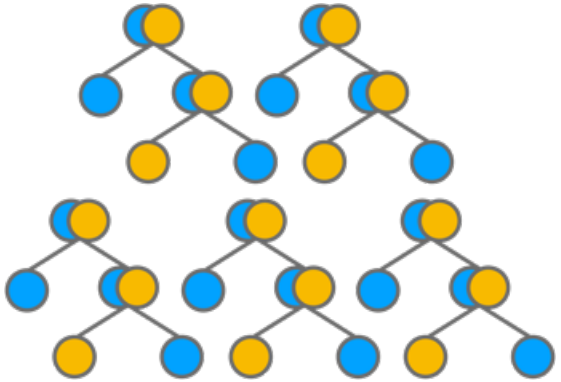


NS-Forest

Data input



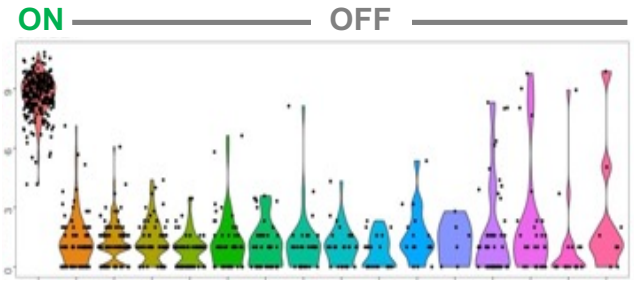
Random forest



Features ranked by information gain

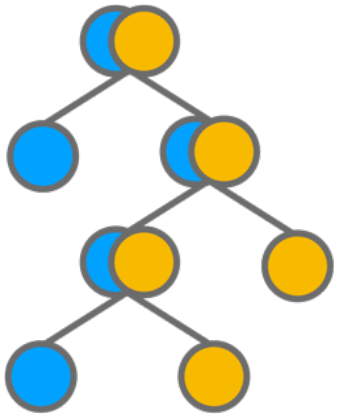
Binary score

$$\text{Score}_{gT} = \frac{\sum_{i=1}^{n}\left(1 - \frac{y_i}{y_T}\right)^{+}}{n-1}$$



Features ranked by binary score

Decision tree



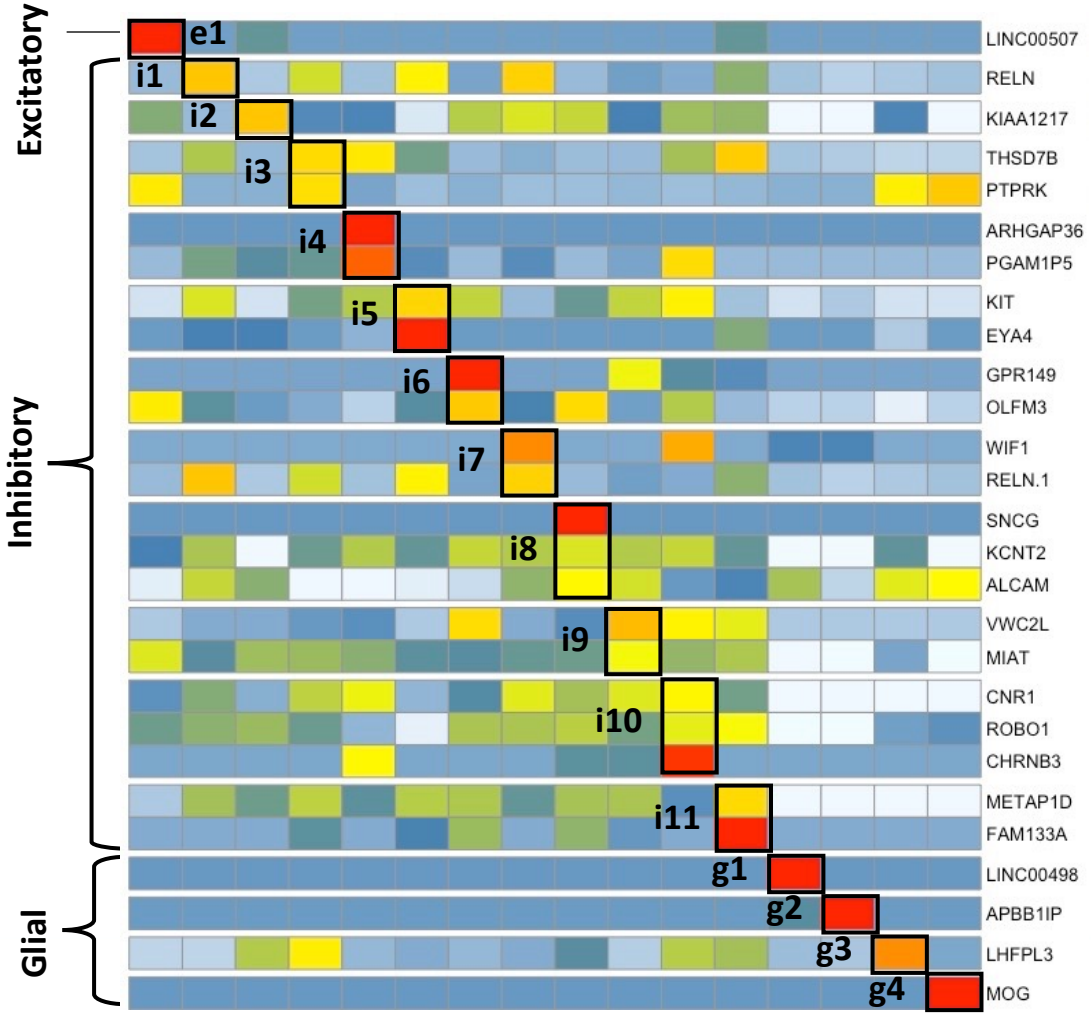
F-beta score evaluation

Aevermann B, et al. (2021) ***Genome Research***, 31:1767-1780. PMID: 34088715

https://github.com/JCVenterInstitute/NSForest

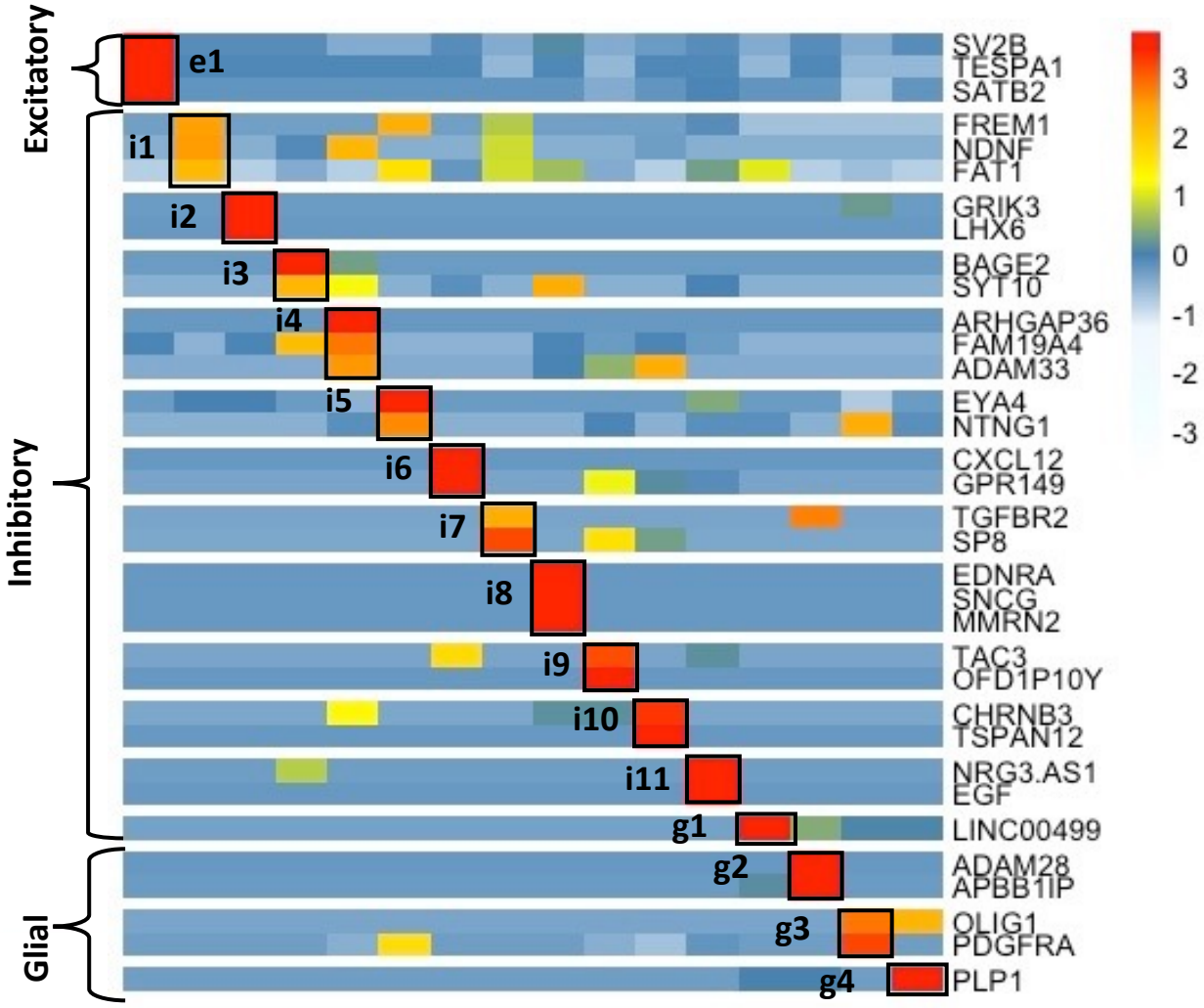
NS-Forest Layer 1 MTG results

Version 1.3

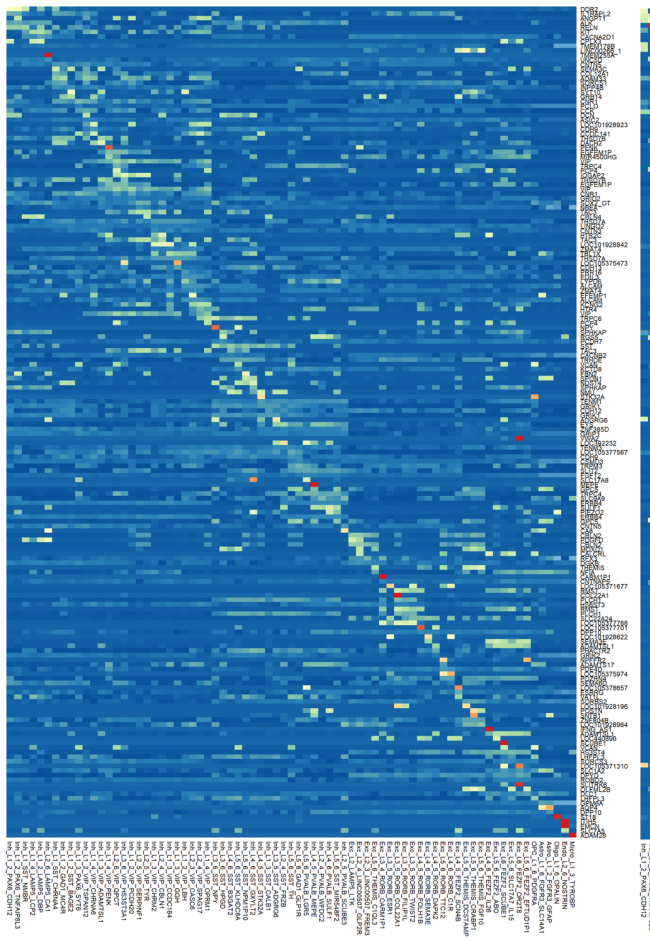


27 markers

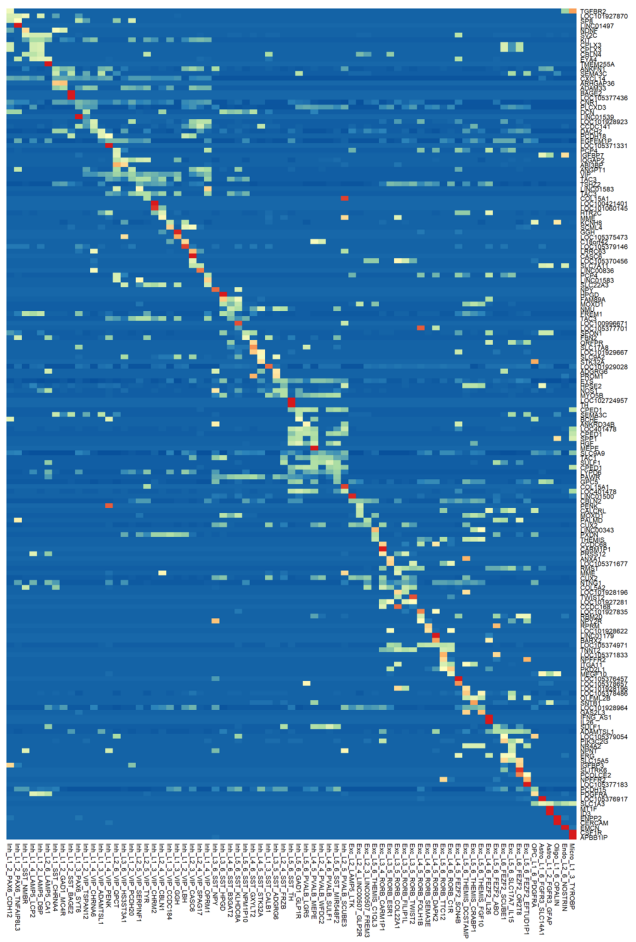
Version 2.0



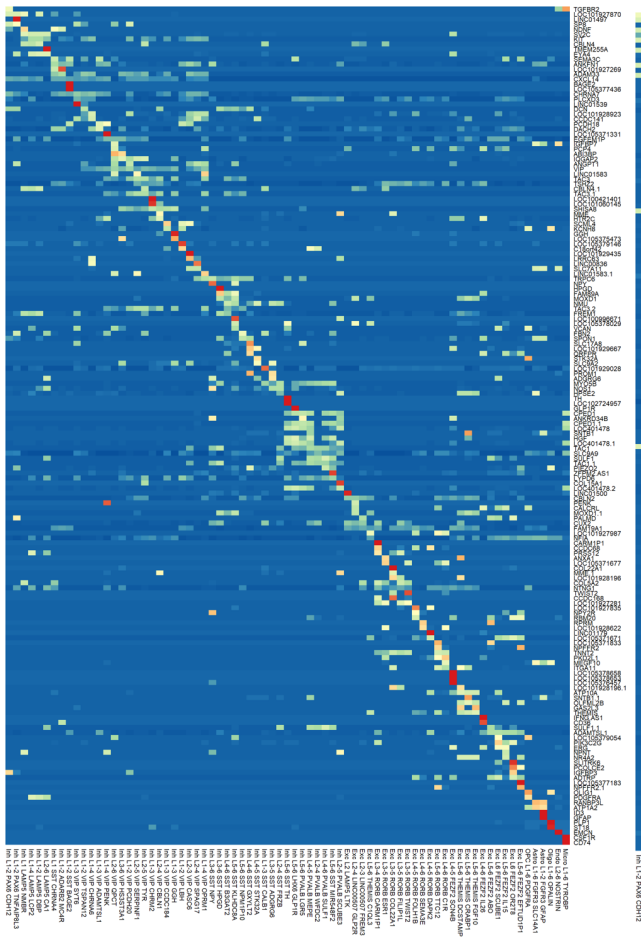
34 markers



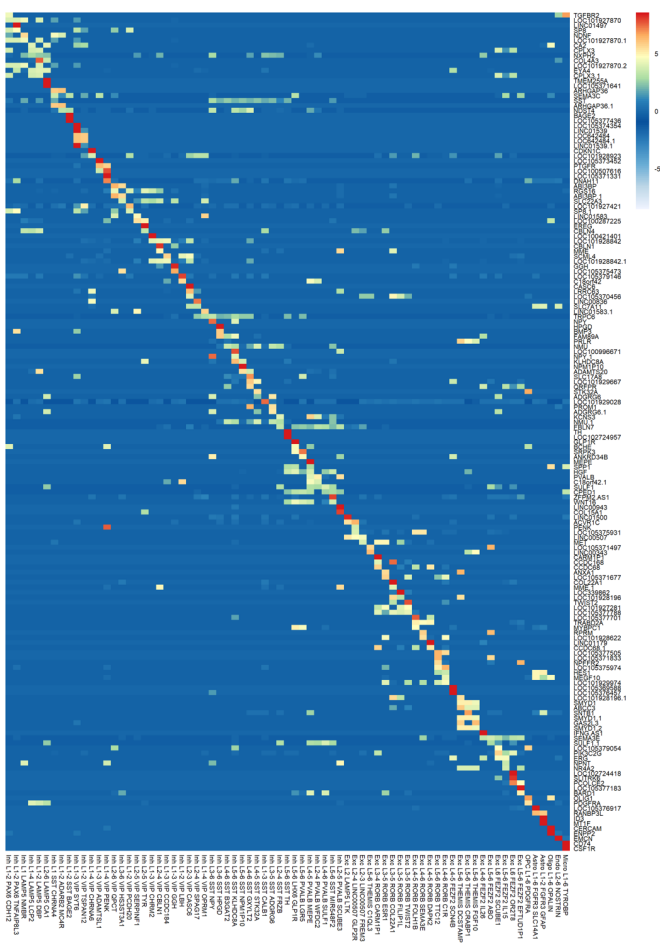
NS-Forest v1.3



NS-Forest v2.0



NS-Forest v3.9

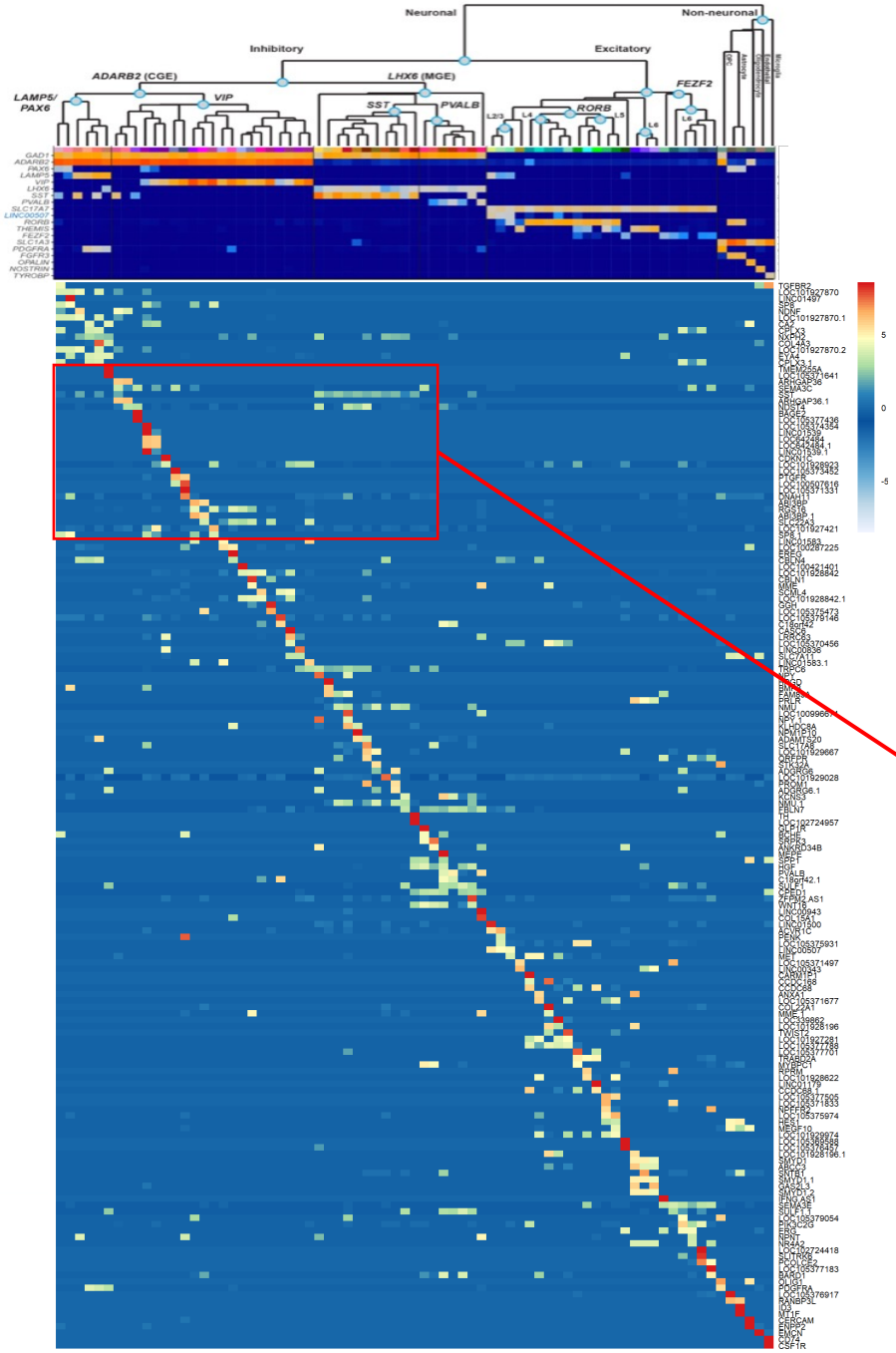


NS-Forest v4.0

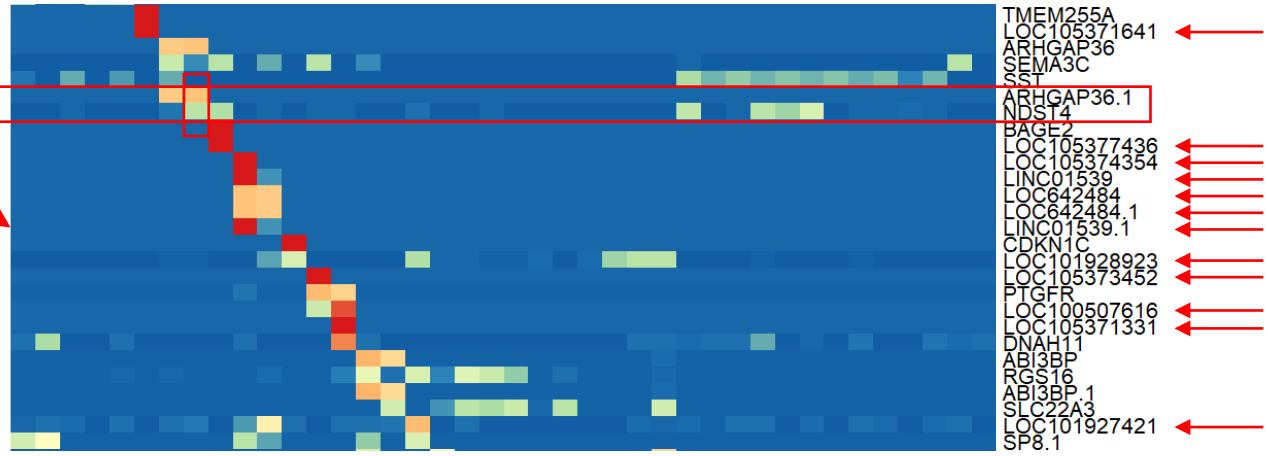
	NS-Forest v1.3	NS-Forest v2.0	NS-Forest v3.9	NS-Forest v4.0
Median f-beta	0.63*	0.68	0.69	0.68
Median PPV	NA	NA	0.85	**0.89**
On-target Ratio	NA	NA	0.18	**0.42**

* F1 score was used in v1.3.

Human MTG NS-Forest results



75 clusters
157 marker genes
1 - 4 markers/cluster
Average 2.4/cluster
Median F-beta score = 0.68



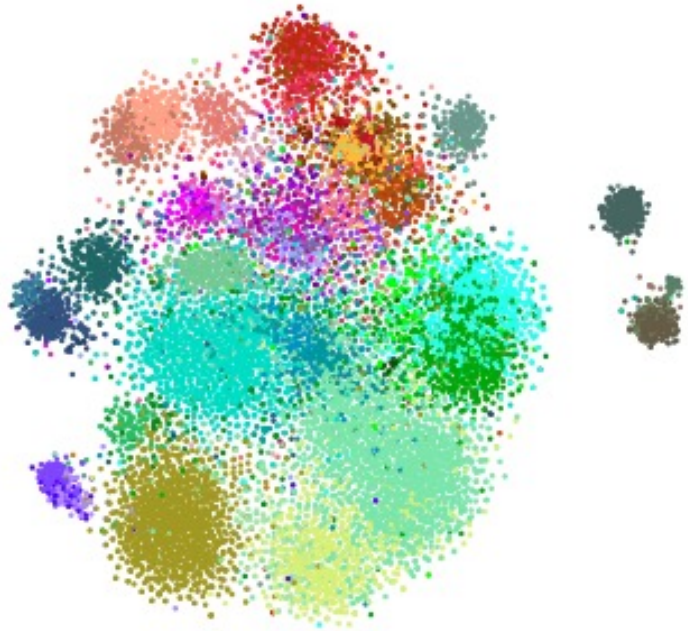
Hodge RD, et al. (2019) ***Nature***, 573:61-68. PMID: 31435019

75 MTG cell type clusters – Cytosplore

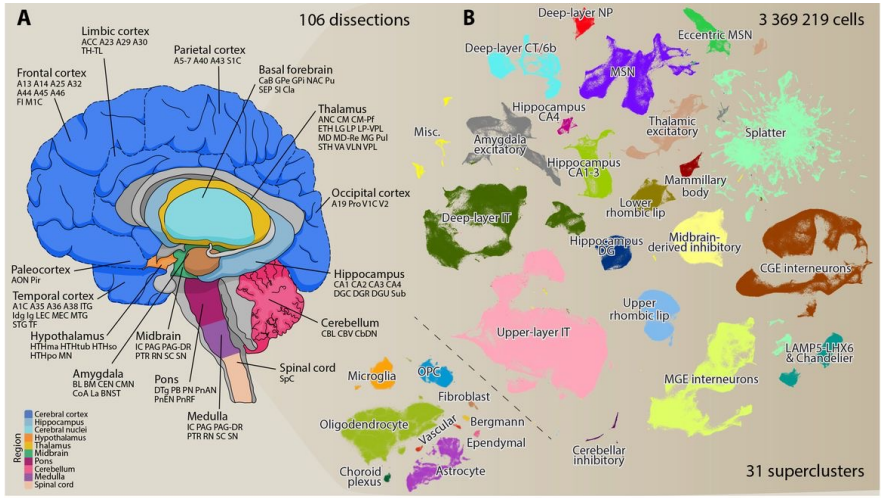
5574 variable genes

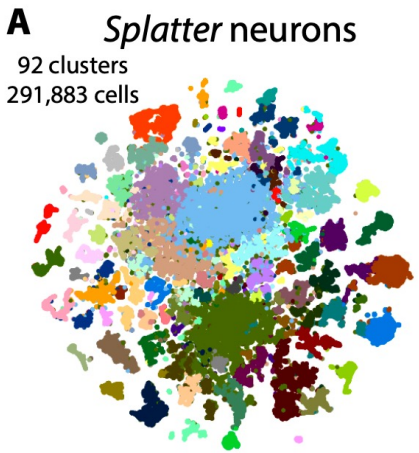
157 NS-Forest marker genes

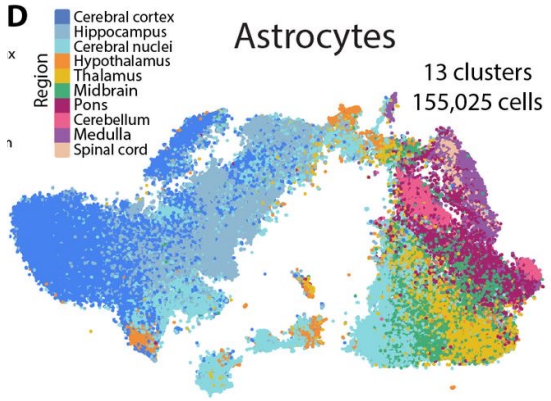
157 randomly chosen variable genes



Human (WHB)

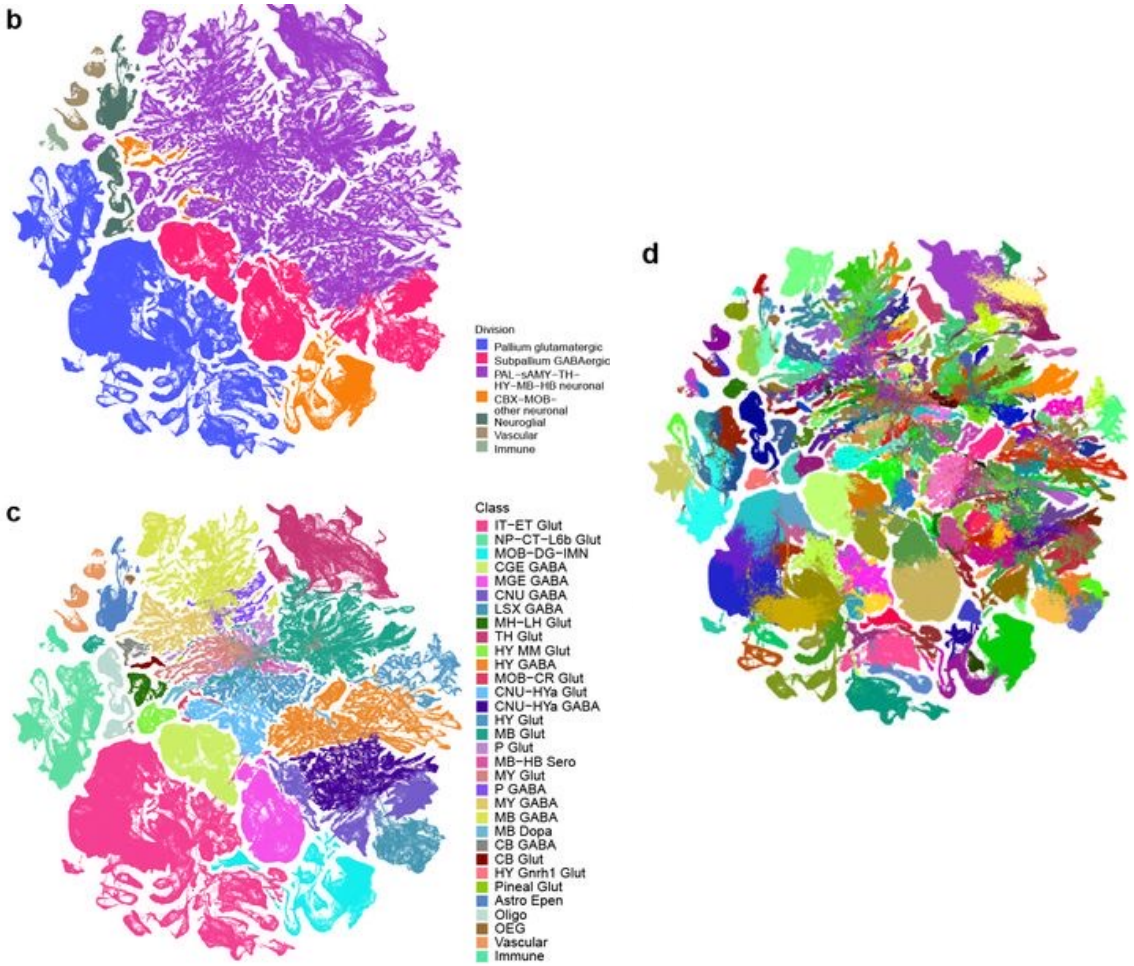






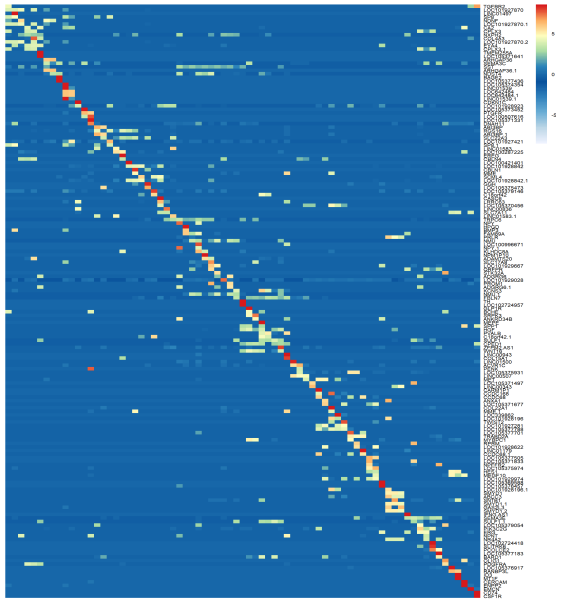
Siletti K, et al." Transcriptomic diversity of cell types across the adult human brain" (2022)
https://www.biorxiv.org/content/10.1101/2022.10.12.511898v1

Mouse (WMB)



Yao Z, et al., "A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain" (2023)
https://www.biorxiv.org/content/10.1101/2023.03.06.531121v1

Use of necessary and sufficient marker genes



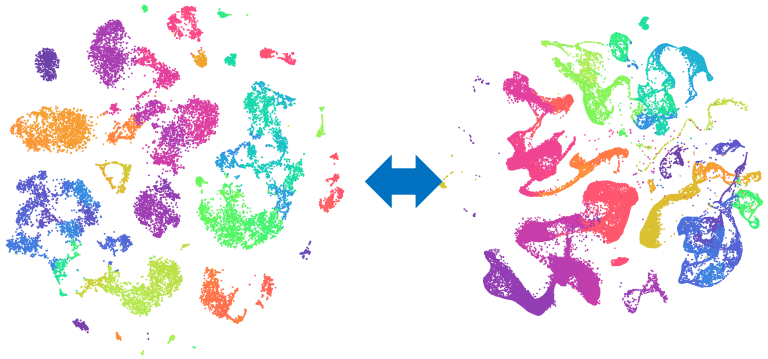
Necessary and sufficient marker genes

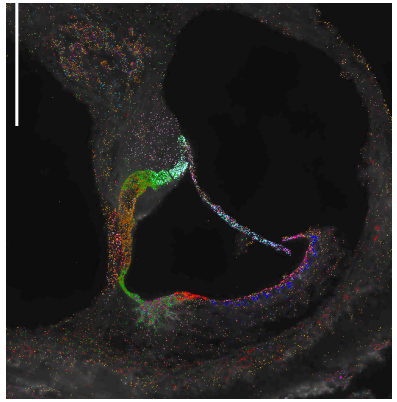
Reference dataset for statistical comparison

Spatial Tx

Probes for mFISH

Semantic definitions & knowledge representation





Annotations

rdfs:label [language: en]

rosehip neuron

Description: 'rosehip neuron'

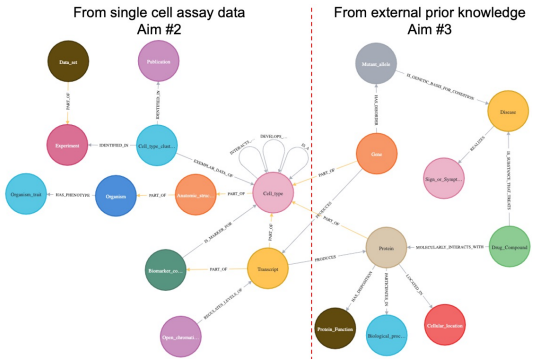
Equivalent To

interneuron
and ('has soma location' some 'cortical layer I')
and ('has soma location' some 'middle temporal gyrus')
and ('capable of' some 'gamma-aminobutyric acid secretion, neurotransmission')
and (expresses some KIT)
and (expresses some NTNG1)
and (expresses some POU6F2)

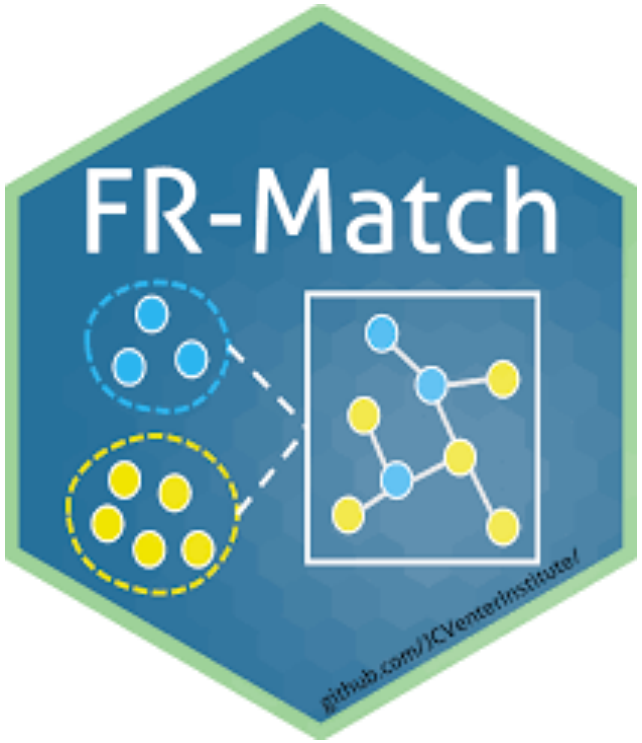
SubClass Of

'synapsed to' some 'layer III pyramidal neuron'

'cerebral cortex GABAergic interneuron'



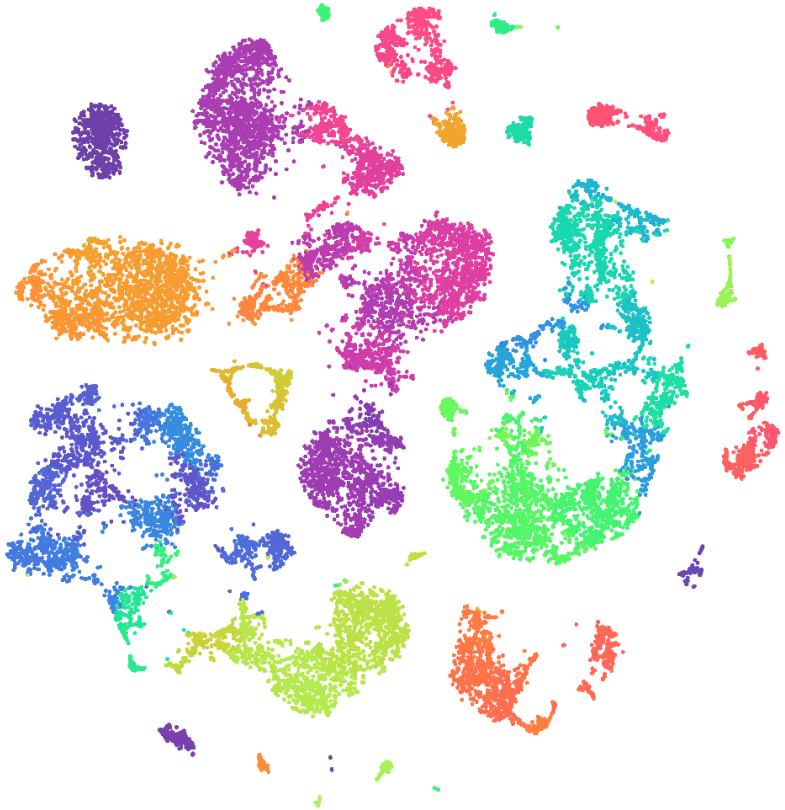
Cell type matching using FR-Match



Zhang Y, et al. (2021) ***Briefings in Bioinformatics***, 22:bbaa339. PMID: 33249453
Zhang Y, et al. (2022) ***Scientific Reports***, 12:9996. PMID: 35705694

Cell type cluster matching

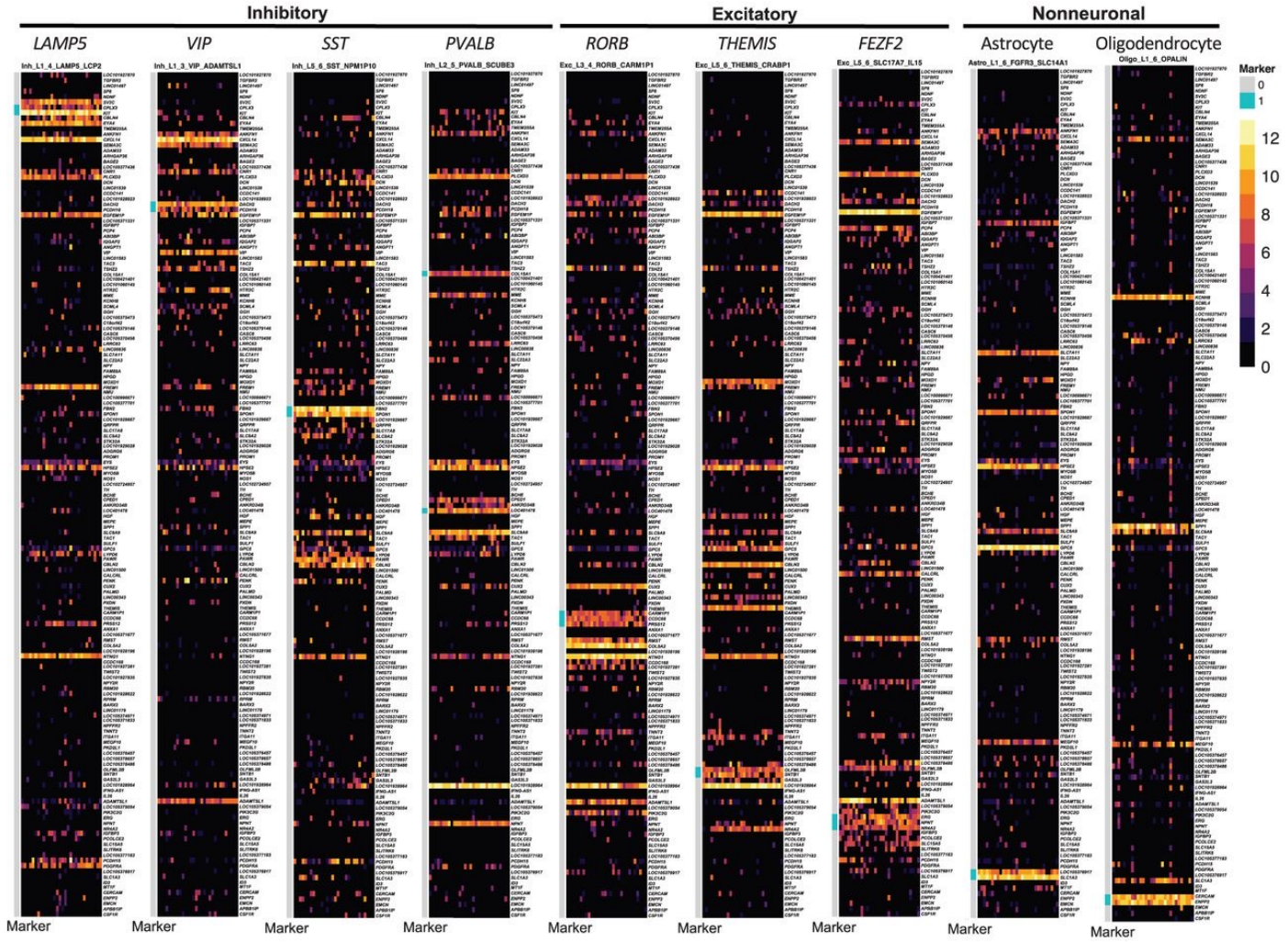
Experiment 1



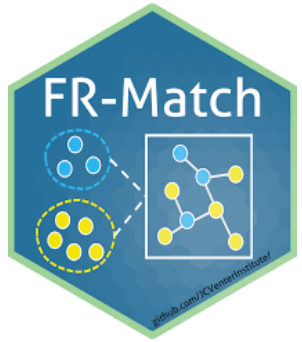
Experiment 2

Cell type barcode

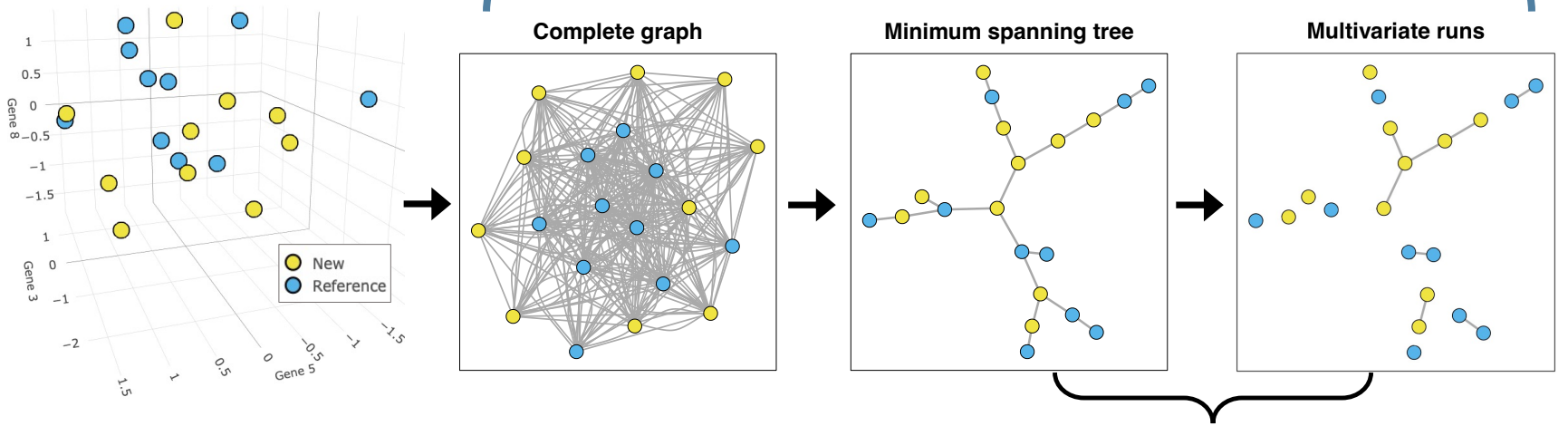
The **combination of NS-Forest markers** provide unique gene expression patterns to characterize different cell types



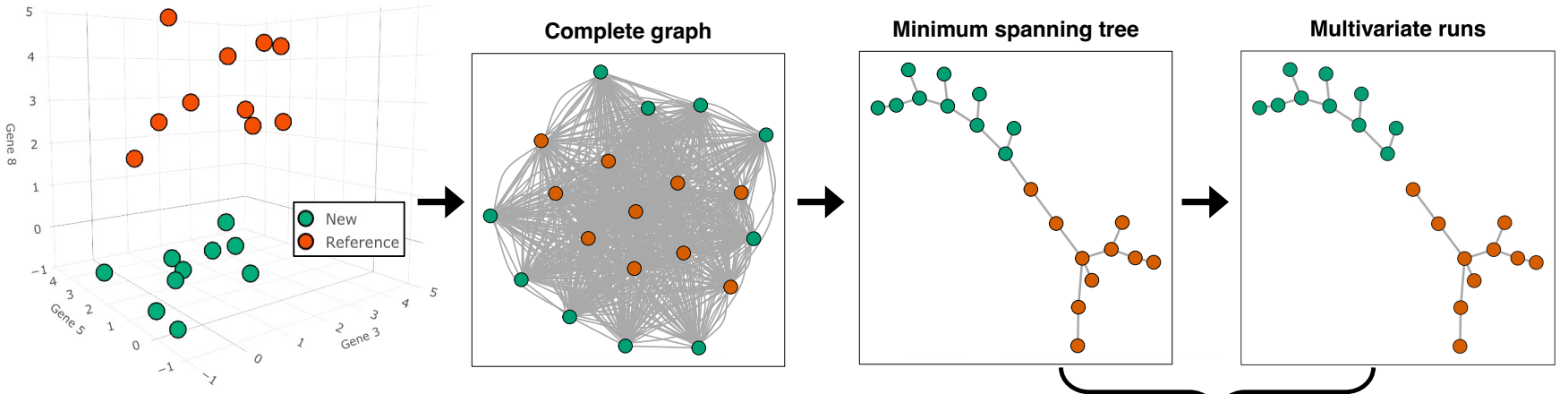
Cell type matching using FR-Match



Graph theoretical modeling of minimum spanning tree



H_0: Common cell type



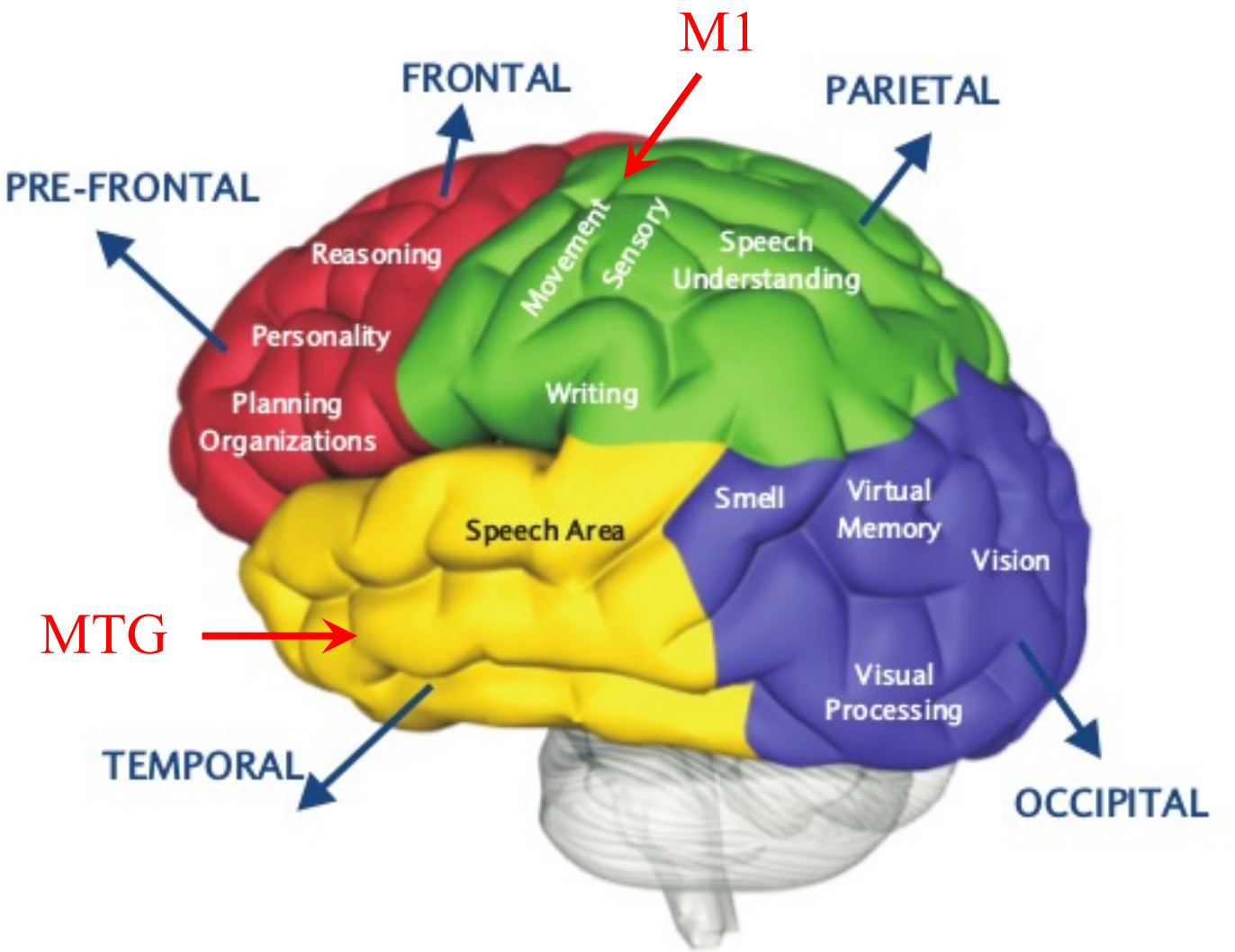
H_1: Novel cell type

Friedman-Rafsky (FR) multivariate nonparametric test

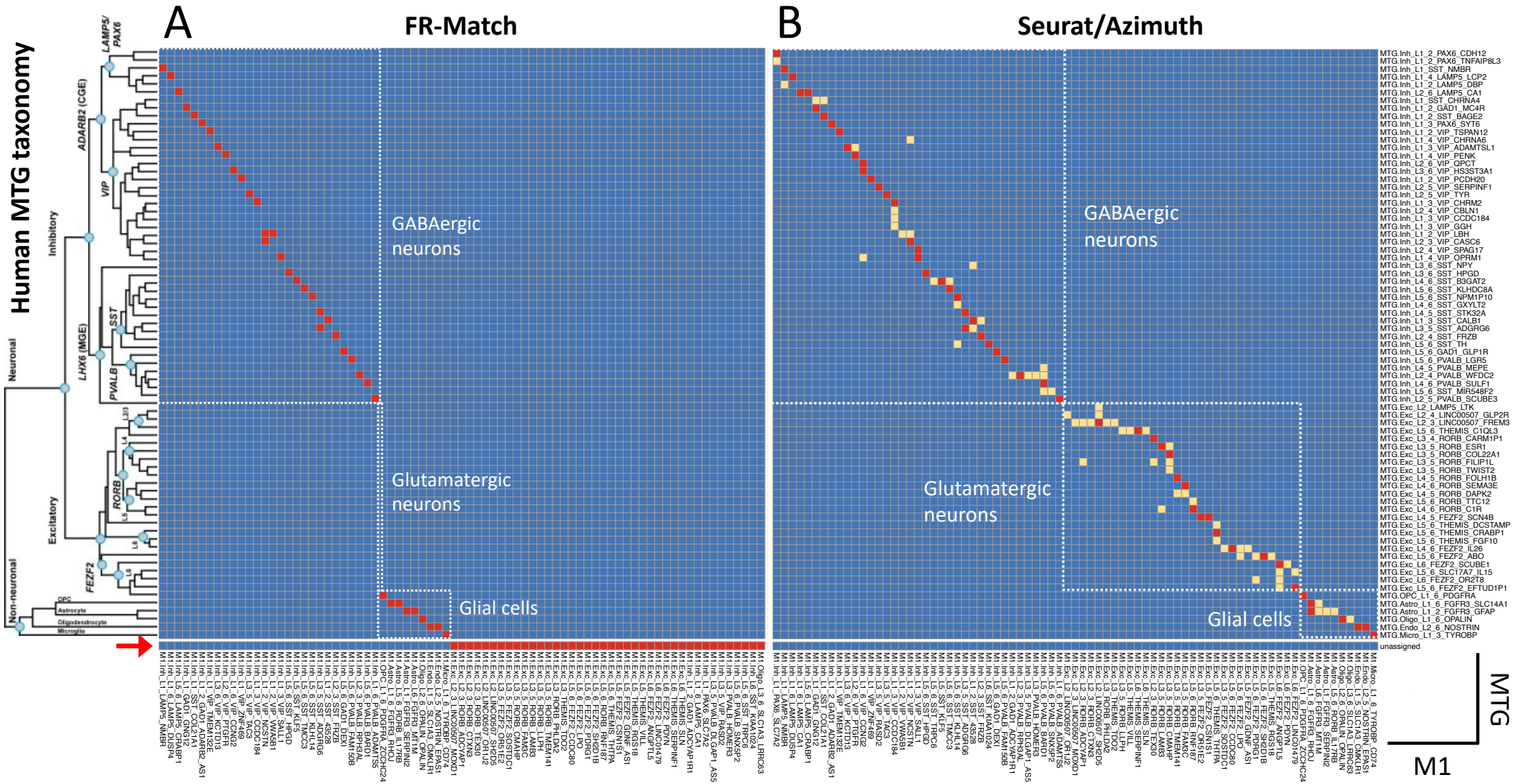
$$W = \frac{R - E(R)}{\mathrm{Var}(R)^{1/2}} \sim N(0,1)$$

Accept or reject null hypothesis H_0

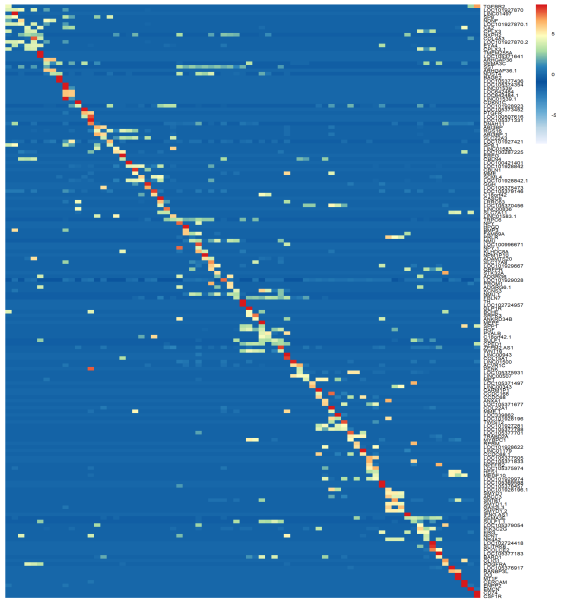
Human neocortex



FR-Match across brain regions



Use of necessary and sufficient marker genes



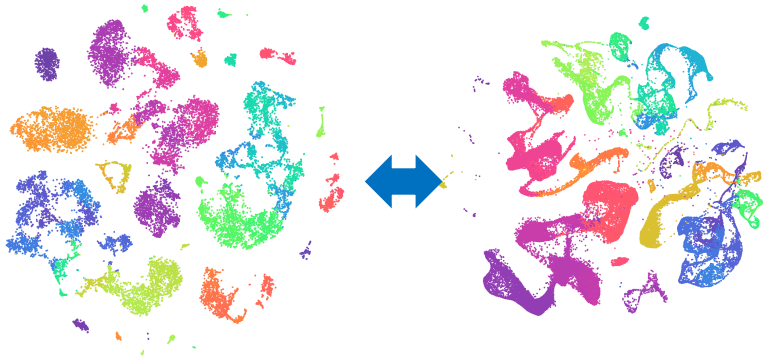
Necessary and sufficient marker genes

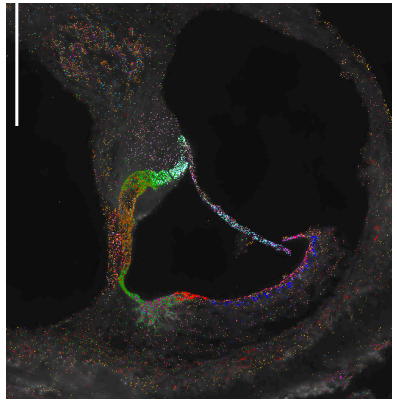
Reference dataset for statistical comparison

Spatial Tx

Probes for mFISH

Semantic definitions & knowledge representation





Annotations

rdfs:label [language: en]

rosehip neuron

Description: 'rosehip neuron'

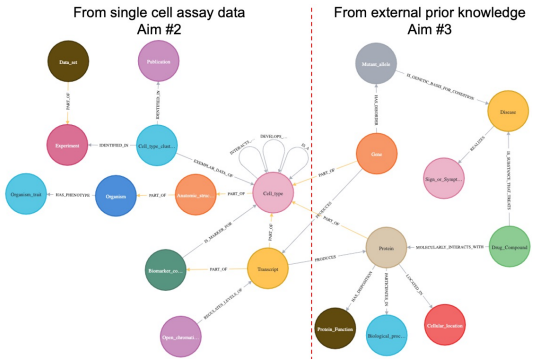
Equivalent To

interneuron
and ('has soma location' some 'cortical layer I')
and ('has soma location' some 'middle temporal gyrus')
and ('capable of' some 'gamma-aminobutyric acid secretion, neurotransmission')
and (expresses some KIT)
and (expresses some NTNG1)
and (expresses some POU6F2)

SubClass Of

'synapsed to' some 'layer III pyramidal neuron'

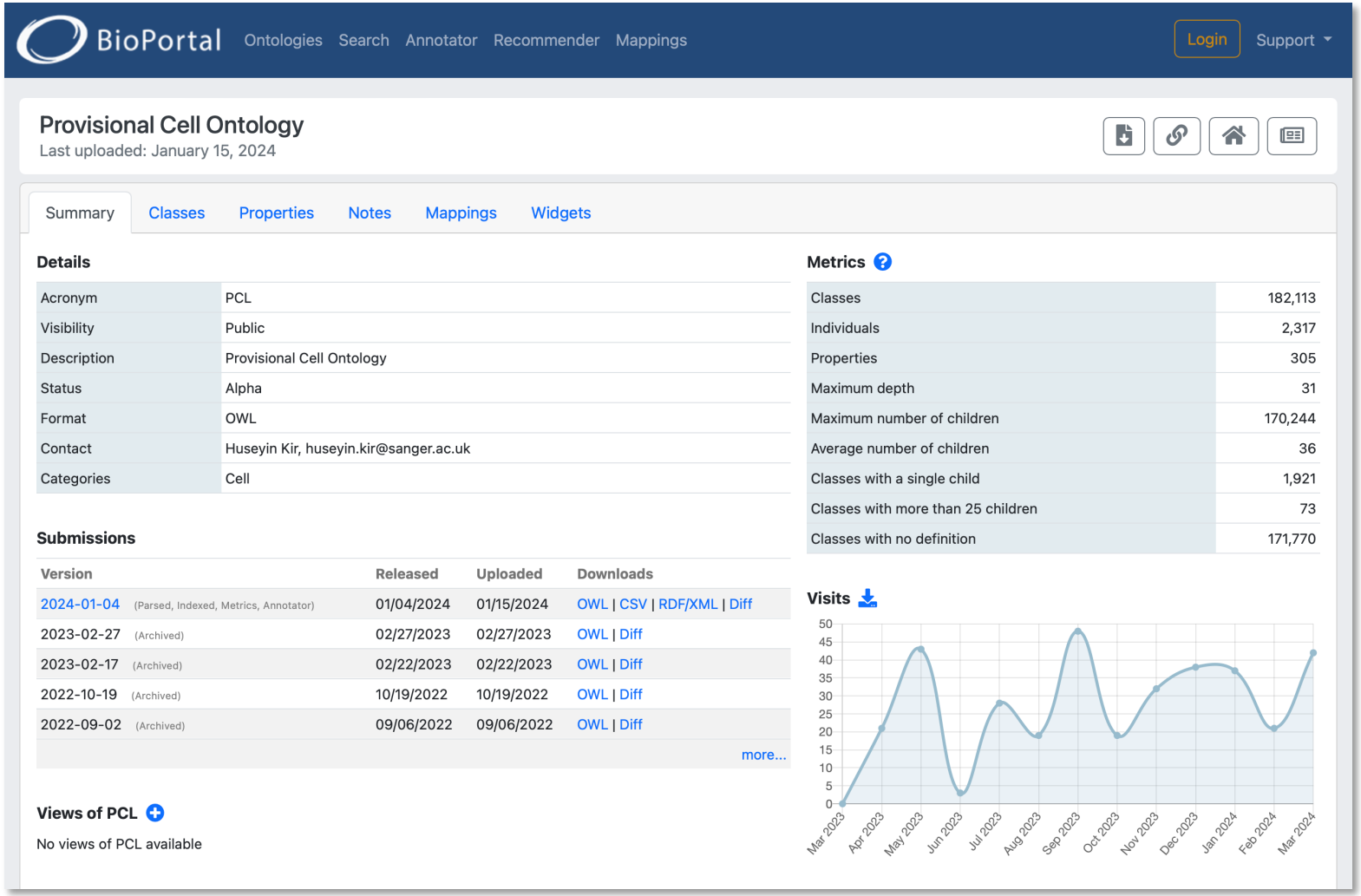
'cerebral cortex GABAergic interneuron'



Cell type definition and knowledge representation

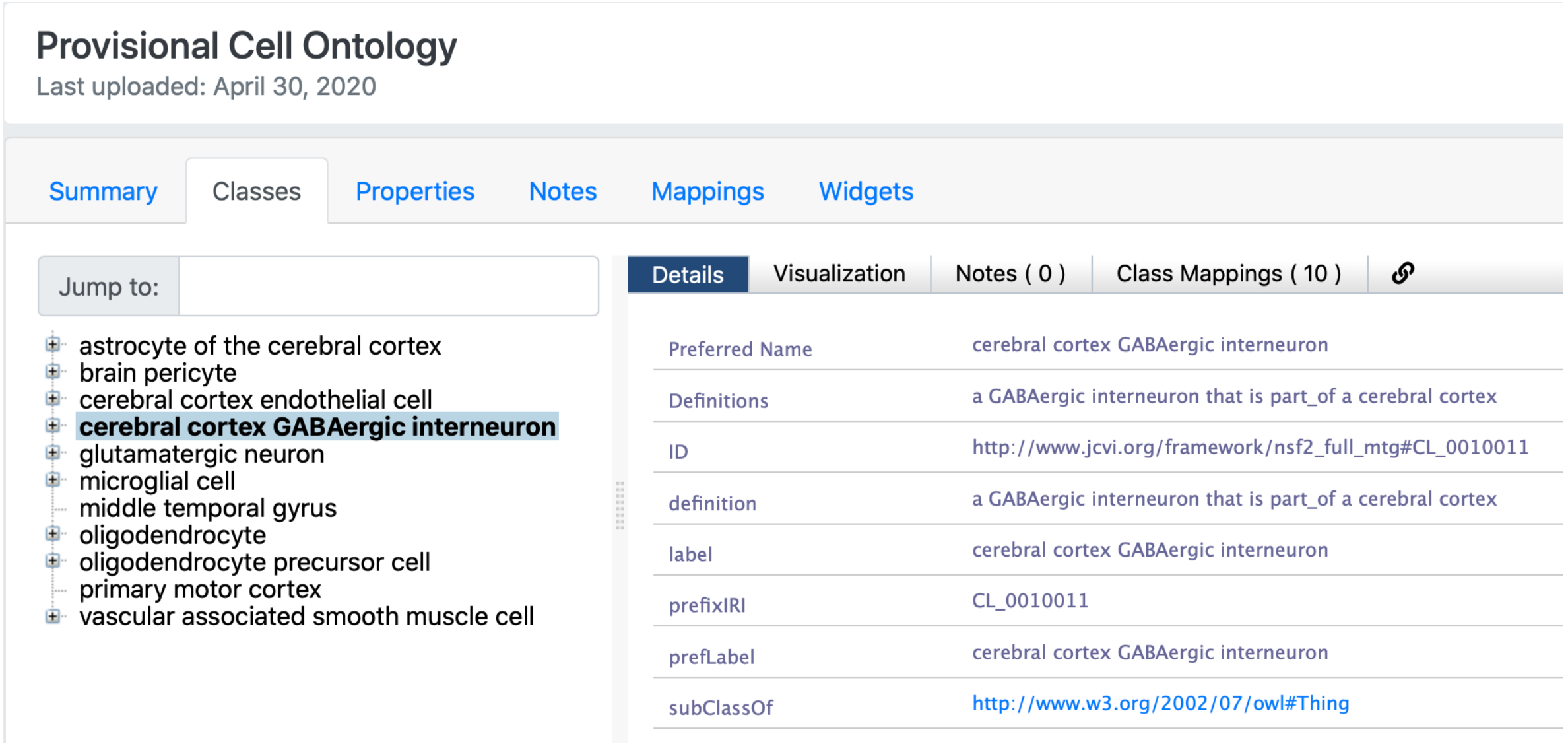
- Ontological representation
  - Structured vocabularies, including rigorous definitions, that capture domain knowledge by asserting formal relationships between terms that represent entities in reality
  - Cell Ontology (CL) currently contains ~2200 cell type terms
  - Cell type-subtype relationships (is_a), cell type developmental relationships (develops_from), cell type-tissue source relationships (part_of anatomy term), cell type-functional relationships (has_disposition)
- Define a cell type by combining metadata and necessary and sufficient conditions:
  - ***specimen source description*** (anatomic structure + species)
  - ***parent cell class*** in the Cell Ontology, and
  - ***necessary and sufficient marker genes*** selectively expressed by the cell type,
  - e.g., *A human middle temporal gyrus cortical layer 1 | GABAergic interneuron | that selectively expresses KIT, NTNG1, and POU6F2 mRNAs*
- Knowledge representation
  - Using semantic web standards rdf/owl
  - Triplet assertions: subject-predicate-object
  - Stored in graph databases, e.g., Neo4j
  - Explored using SPARQL or Cypher queries

Provisional Cell Ontology

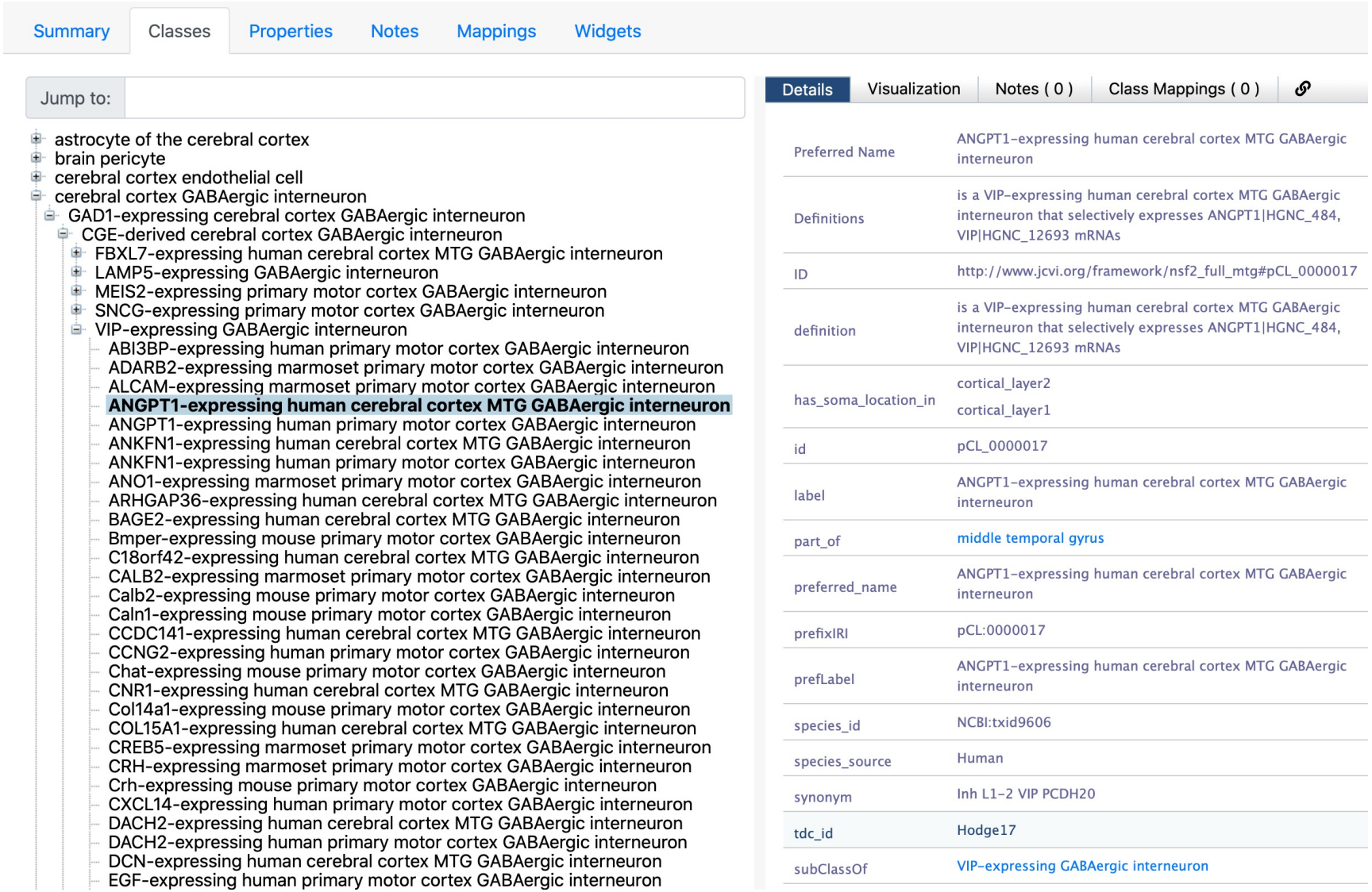


https://bioportal.bioontology.org/ontologies/PCL

Upper nodes from CL provide hooks

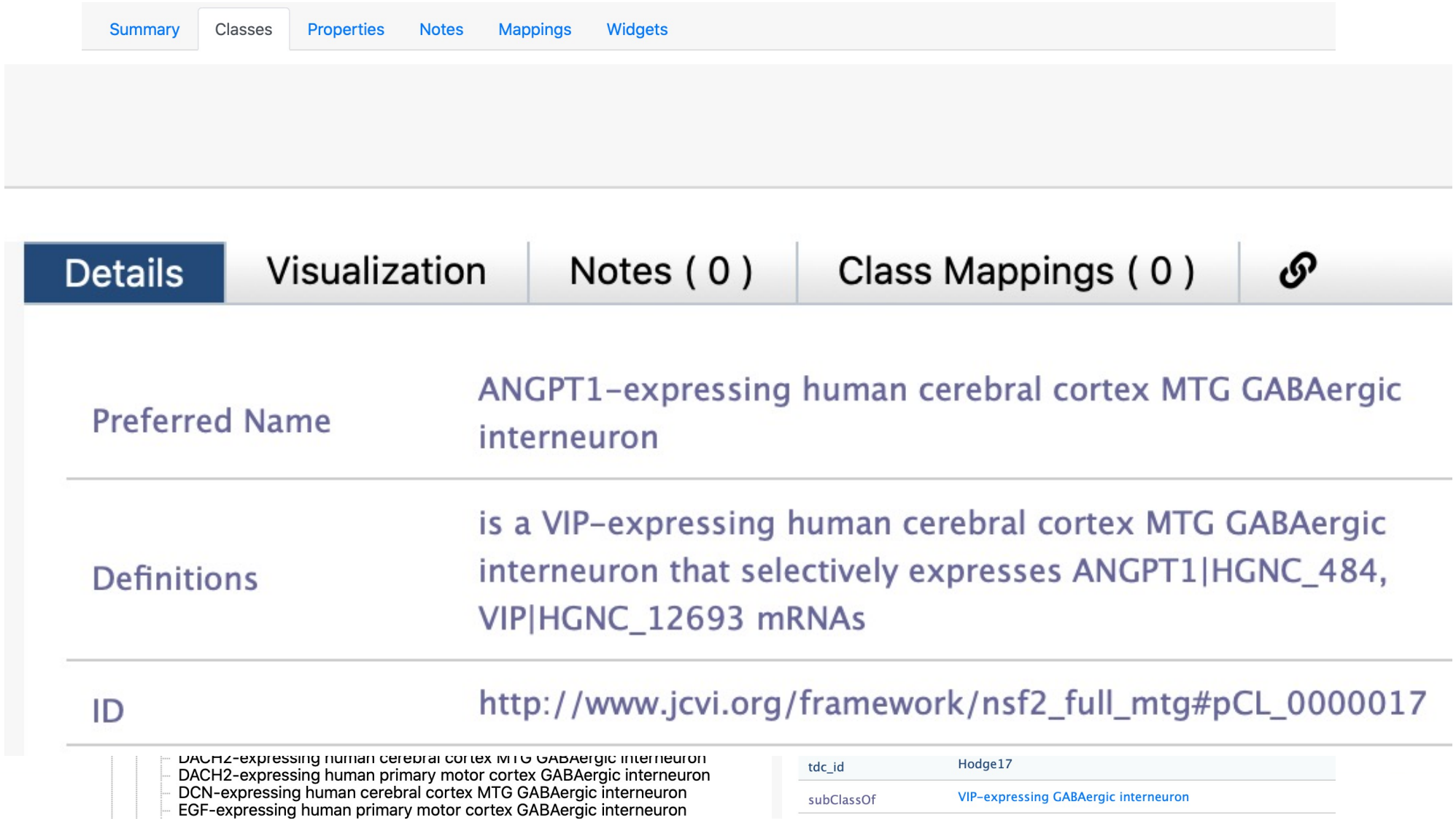


Taxonomy of cell types

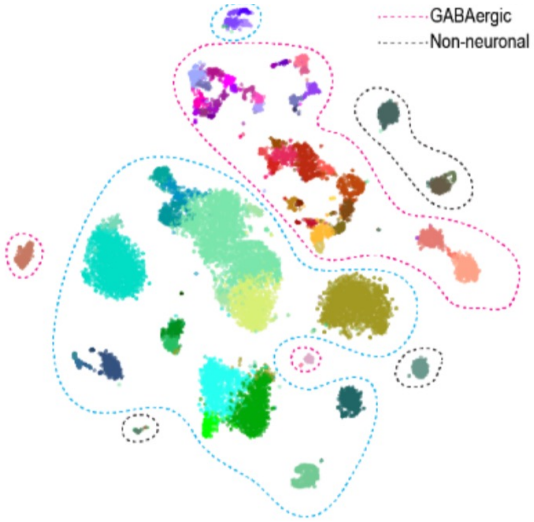


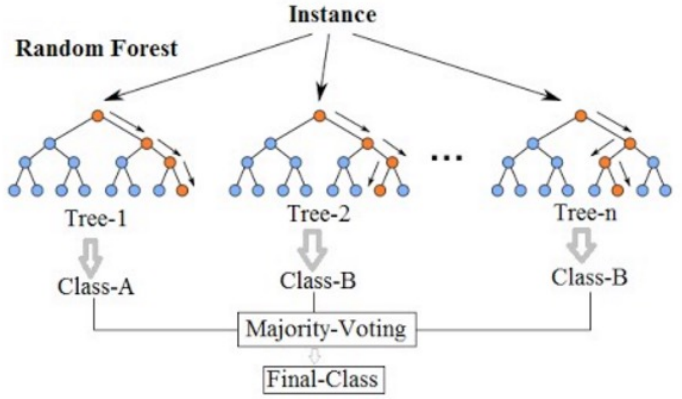
https://bioportal.bioontology.org/ontologies/PCL

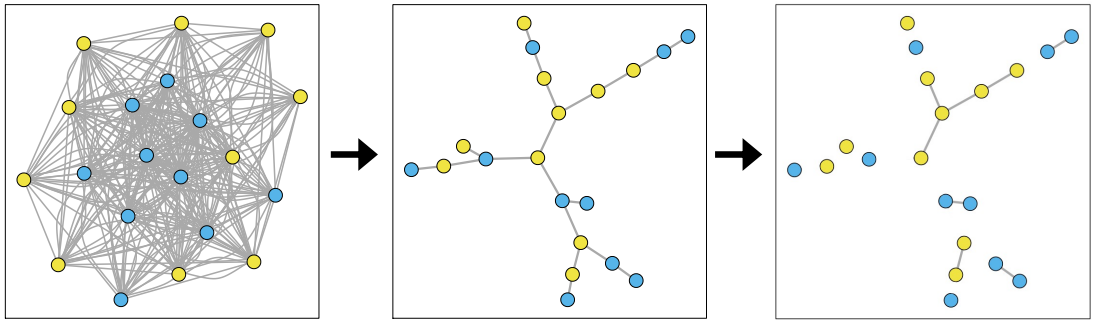
Taxonomy of cell types



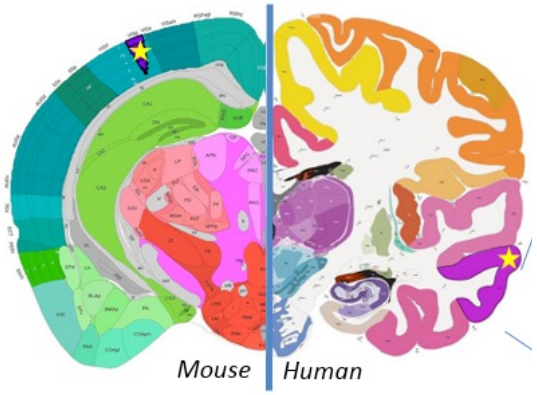
https://bioportal.bioontology.org/ontologies/PCL

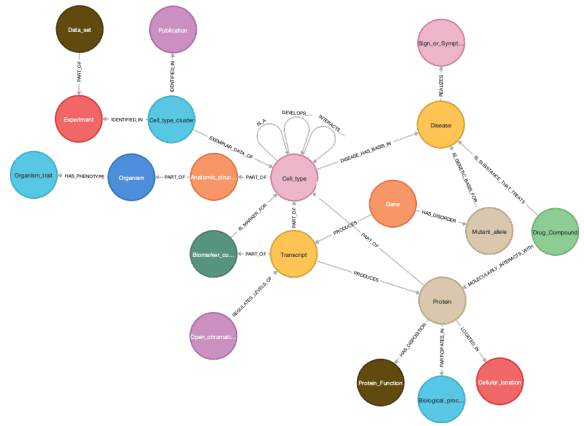


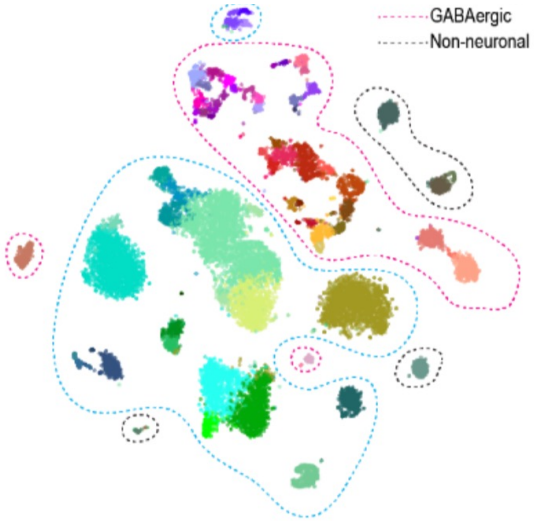


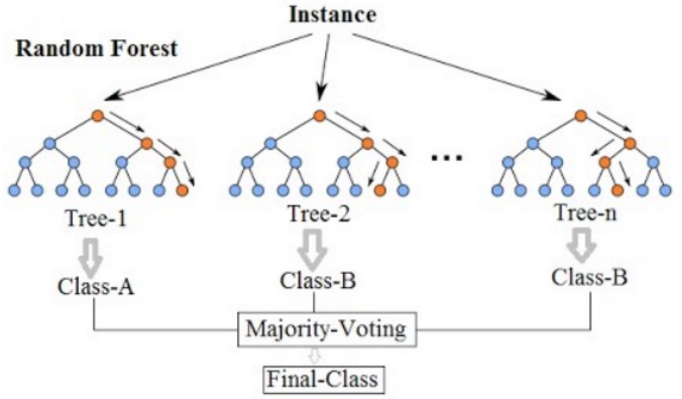


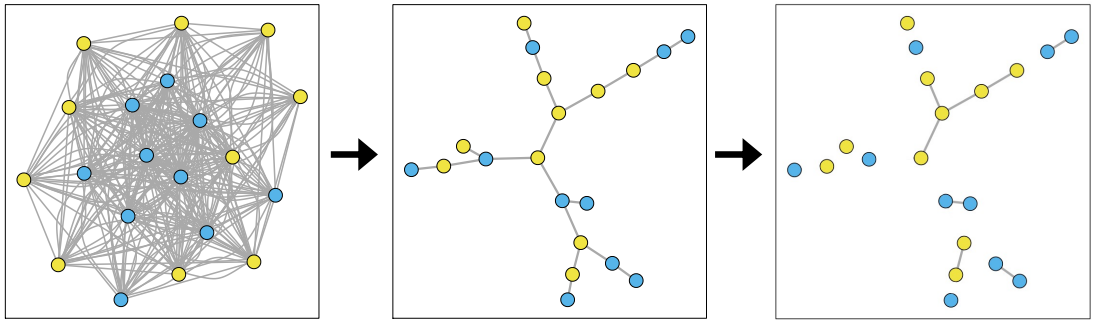
Data ⟶ Knowledge



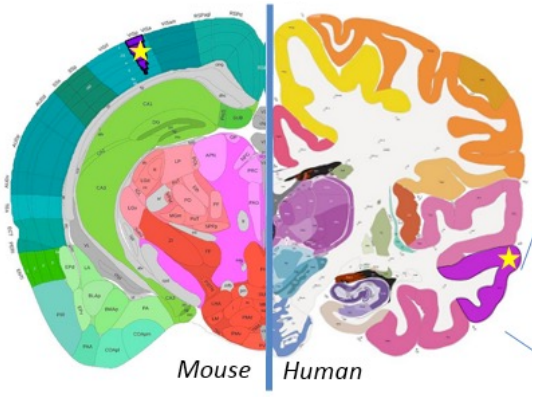


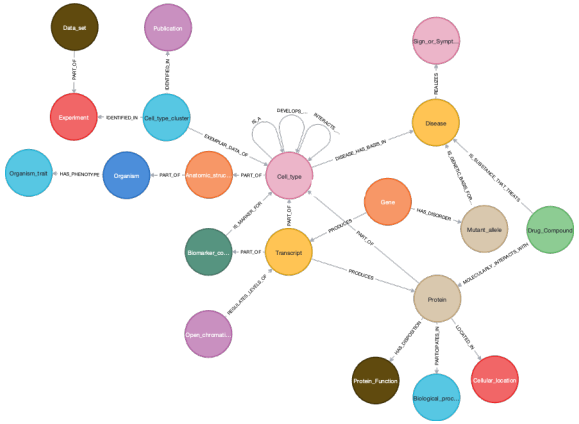




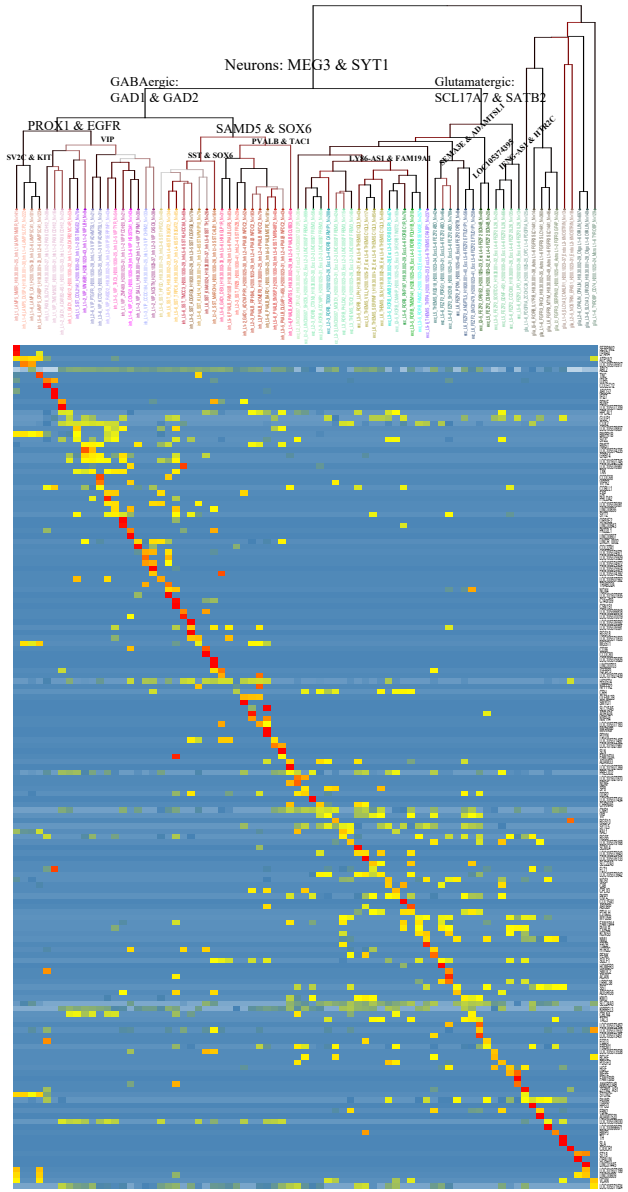


Big Data ⟶ Computable Knowledge



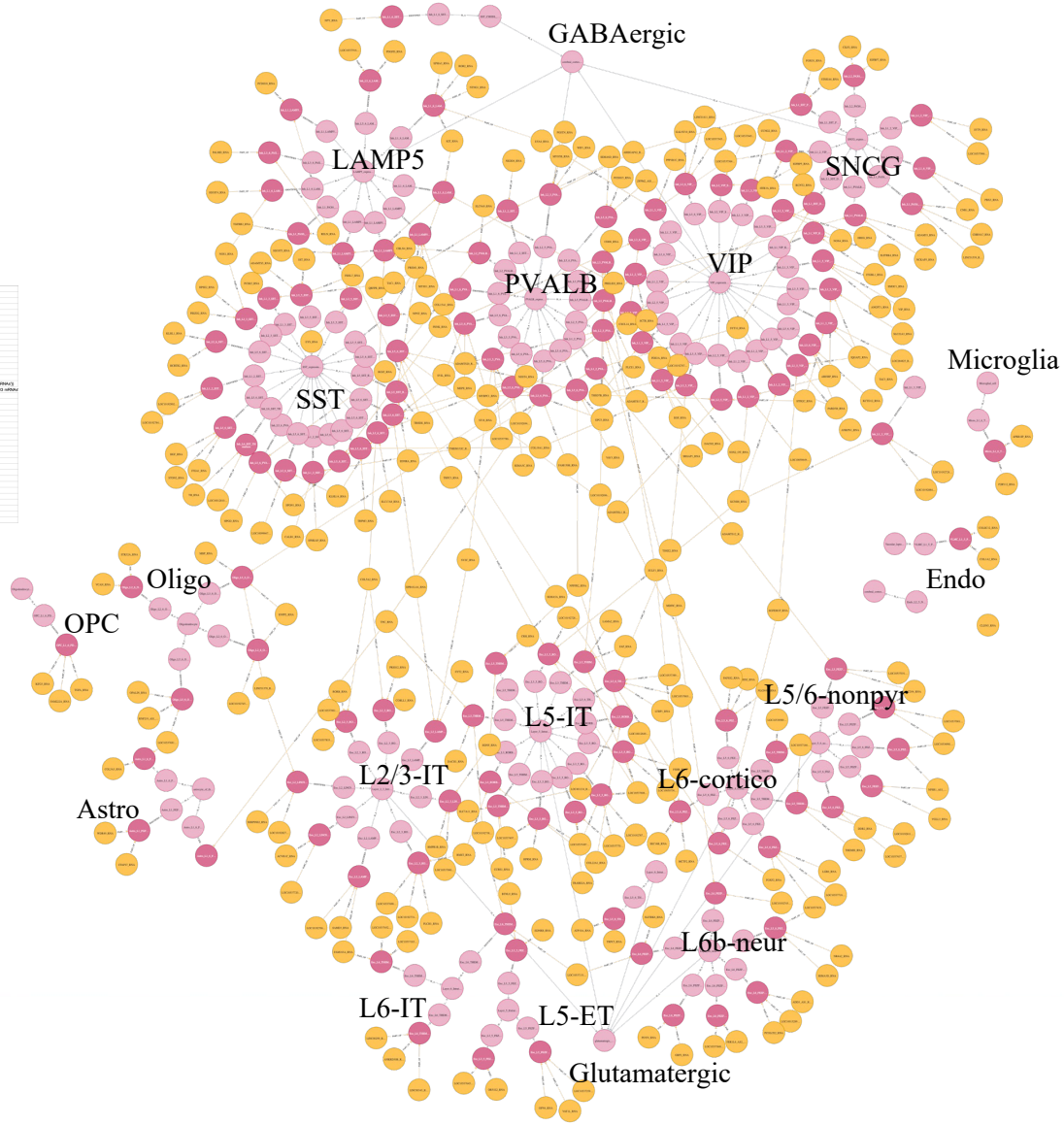


Cell Type Knowledge Graph



Cipher Create

Neo4j



Nodes and Edges

Table 1. Semantic node types

Node_type	Ontology_source
Cell_type	CL/pCL
Cell_type_cluster	from_project
Anatomic_structure	Uberon
Organism	NCBI_Taxonomy
Organism_traits	PATO
Transcript	Ensembl
Biomarker_combination	from_project
Gene	Entrez_Gene
Protein	PRO_or_Uniprot
Open_chromatin_region	SO
Protein_Function	GO-MF
Biological_process/pathway	GO-BP_and_Reactome
Cellular_location	GO-CC
Drug_Compound	DrugBank
Disease	DO
Sign_or_Symptom	MeSH_ontology
Data_set	from_project
Experiment	from_project
Publication	PubMed_ID/doi
Mutant_allele	OMIM

Table 2. Edge types and SSS assertions

	Subject_node	Predicate_relation	Object_node	Source
Aim 2 - scRNAseq & other experiments	Cell_type	PART_OF	Anatomic_structure	metadata
	Organism	HAS_PHENOTYPE	Organism_traits	metadata
	Cell_type_cluster	EXEMPLAR_DATA_OF	Cell_type	scRNA-seq
	Transcript	PART_OF	Cell_type	scRNA-seq
	Transcript	PART_OF	Biomarker_combination	NS-Forest
	Biomarker_combination	IS_MARKER_FOR	Cell_type	NS-Forest
	Data_set	PART_OF	Experiment	metadata
	Cell_type_cluster	FOUND_IN	Experiment	metadata
	Cell_type_cluster	IDENTIFIED_IN	Publication	metadata
	Open_chromatin_region	REGULATES_LEVELS_OF	Transcript	ATACseq
	Protein	PART_OF	Cell_type	proteomics
	Cell_type	INTERACTS_WITH	Cell_type	SpaceTx
Aim 3 -external knowledge sources	Cell_type	DEVELOPS_FROM	Cell_type	CL
	Cell_type(subtype/child)	IS_A	Cell_type(type/parent)	CL/PCL
	Anatomic_structure	PART_OF	Organism	Uberon
	Gene	PRODUCES	Transcript	inferred
	Transcript	PRODUCES	Protein	inferred
	Protein	HAS_DISPOSITION	Protein_function	GO
	Protein	PARTICIPATES_IN	Biological_process/pathway	GO
	Protein	LOCATED_IN	Cellular_location	GO
	Disease	REALIZES	Sign_or_Symptom	DO
	Gene	HAS_DISORDER	Mutant_allele	OMIM
	Mutant_allele	IS_GENETIC_BASIS_FOR_CONDITION	Disease	OMIM
	Drug_Compound	MOLECULARLY_INTERACTS_WITH	Protein	DrugBank
	Drug_Compound	IS_SUBSTANCE_THAT_TREATS	Disease	DrugBank

Integrated Semantic Knowledge Graph

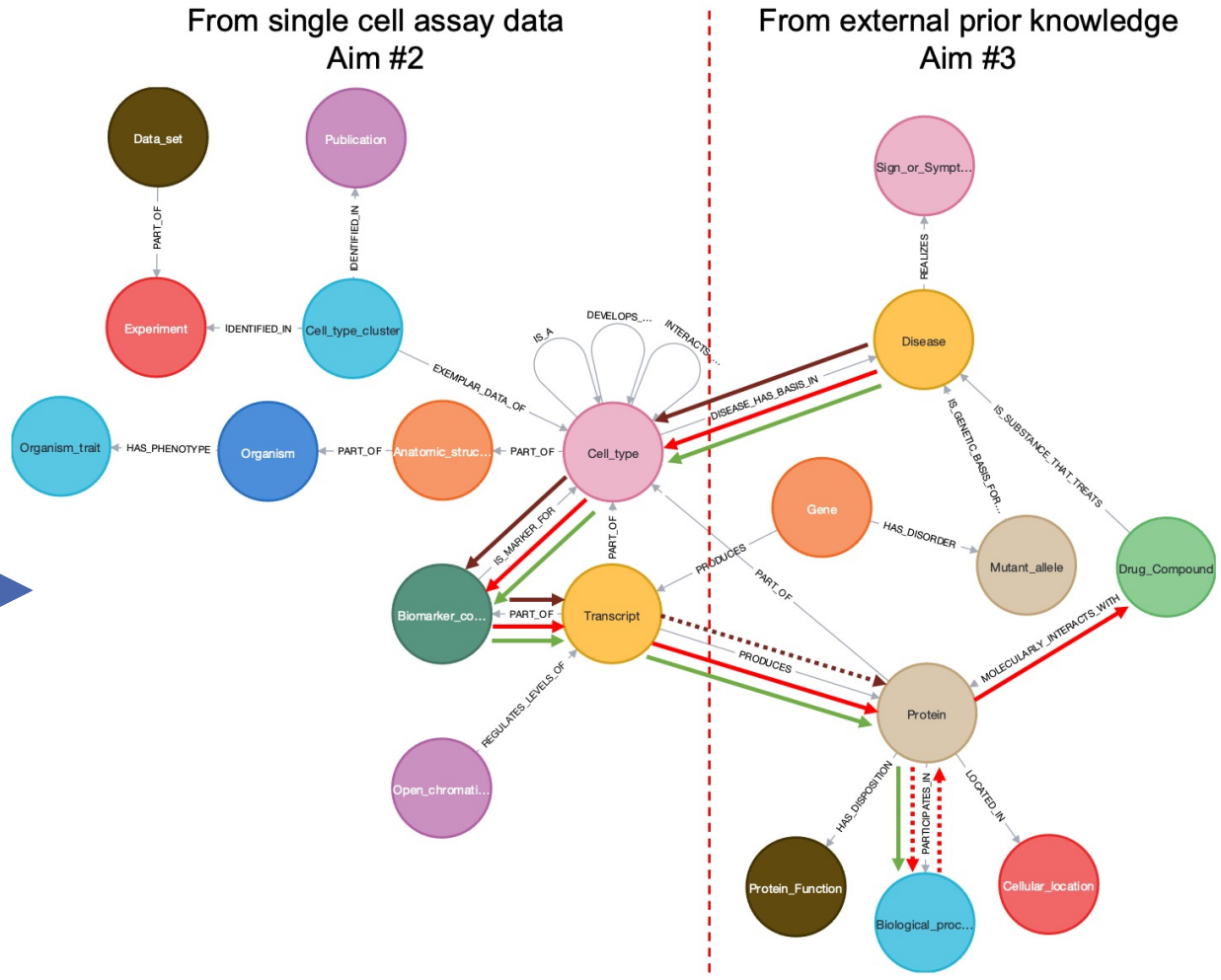
OMIM
GWAS Catalog
ChemBL
DrugBank
Disease Ontology
Gene Ontology
Reactome
Uberon

→ SPOKE

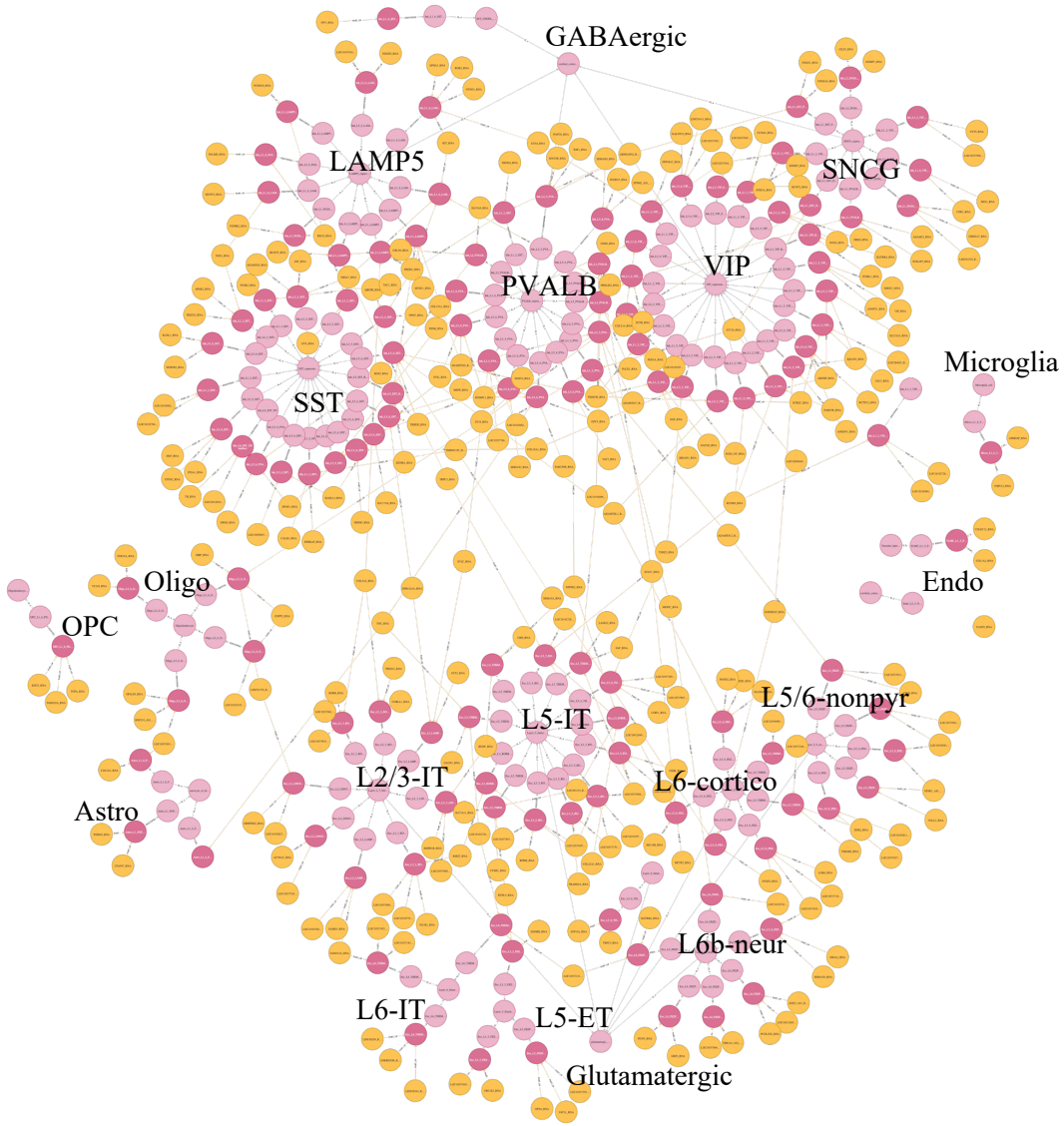
Table 2. Edge types and SSS assertions

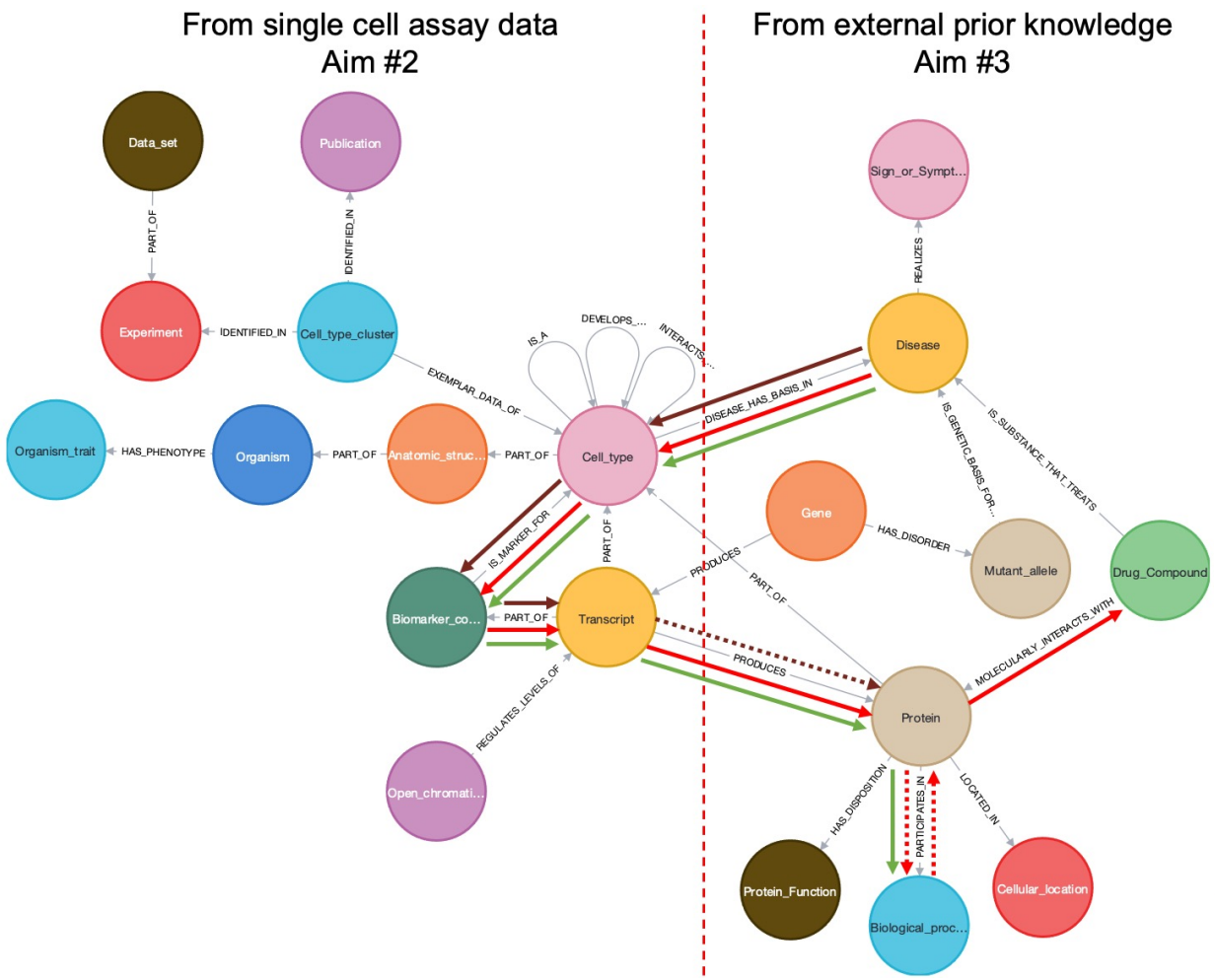
Subject_node	Predicate_relation	Object_node
Cell_type(subtype/child)	IS_A	Cell_type(type/parent)
Cell_type	DEVELOPS_FROM	Cell_type
Cell_type	PART_OF	Anatomic_structure
Cell_type	INTERACTS_WITH	Cell_type
Anatomic_structure	PART_OF	Organism
Organism	HAS_PHENOTYPE	Organism_traits
Gene	PART_OF	Cell_type
Transcript	PART_OF	Cell_type
Protein	PART_OF	Cell_type
Gene	PRODUCES	Transcript
Transcript	PRODUCES	Protein
Open_chromatin_region	REGULATES_LEVELS_OF	Transcript
Protein	HAS_DISPOSITION	Protein_function
Cell_type_cluster	EXEMPLAR_DATA_OF	Cell_type
Cell_type_cluster	FOUND_IN	Experiment
Cell_type_cluster	IDENTIFIED_IN	Publication
Transcript	PART_OF	Biomarker_combination
Biomarker_combination	IS_MARKER_FOR	Cell_type
Protein	PARTICIPATES_IN	Biological_process/pathway
Protein	LOCATED_IN	Cellular_location
Drug_Compound	MOLECULARLY_INTERACTS_WITH	Protein
Gene	HAS_DISORDER	Mutant_allele
Mutant_allele	IS_GENETIC_BASIS_FOR_CONDITION	Disease
Disease	REALIZES	Sign_or_Symptom
Drug_Compound	IS_SUBSTANCE_THAT_TREATS	Disease
Data_set	PART_OF	Experiment

Cipher →



Integrated Semantic Knowledge Graph

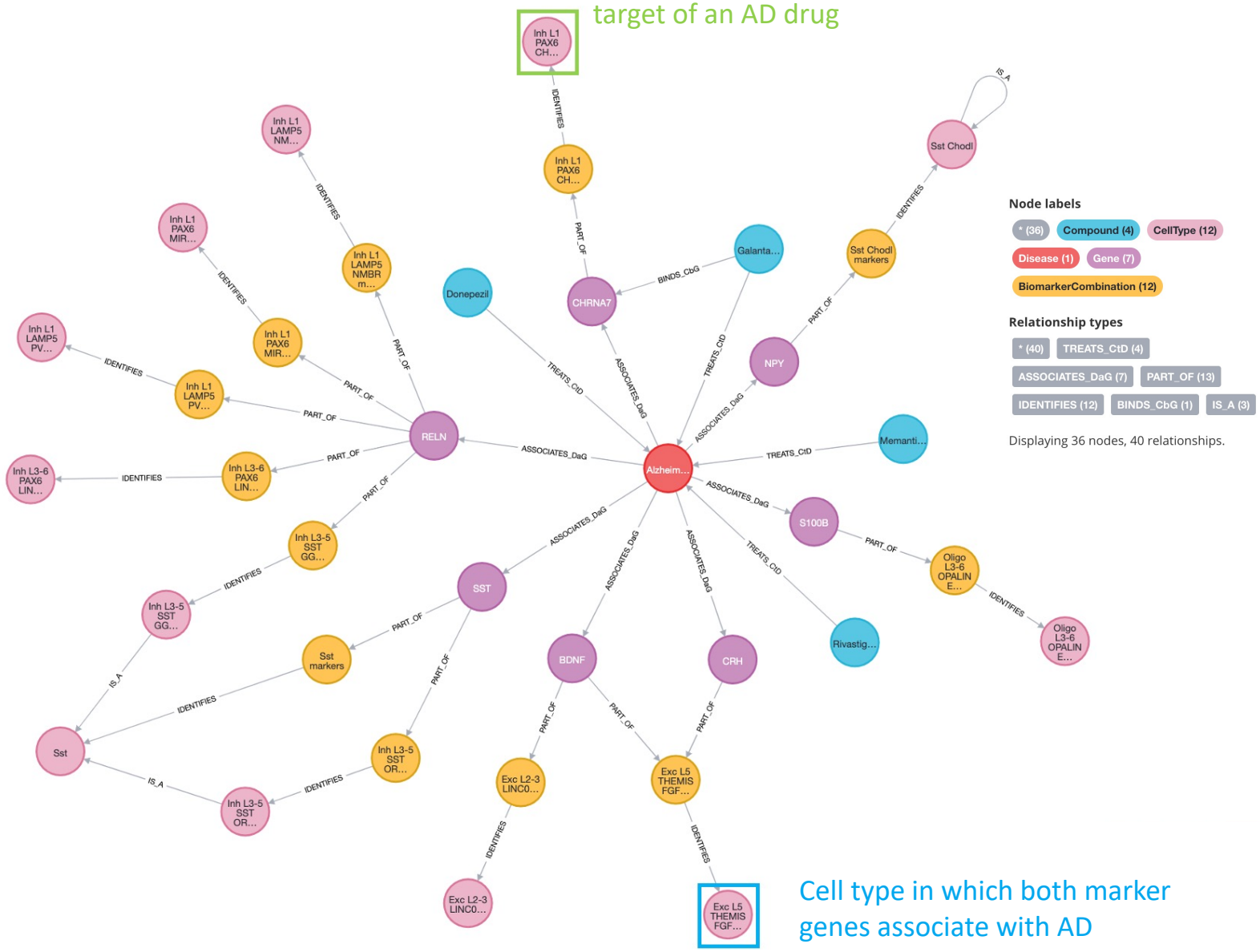




Neuronal Cell Types and Alzheimer's

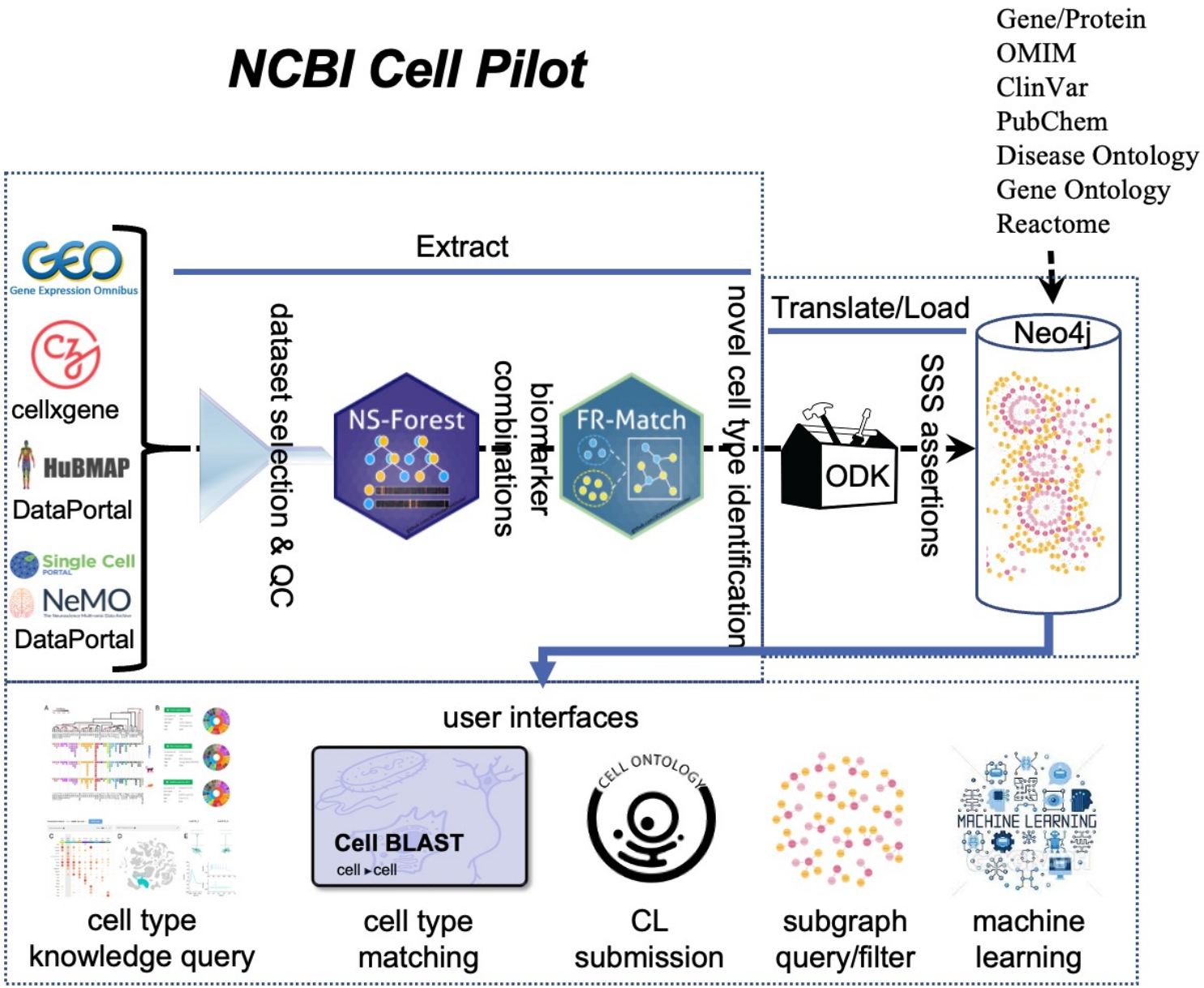
```
MATCH p = (c:Compound)--
(d:Disease)--(g:Gene)--
(b:BiomarkerCombination)--
(ct:CellType) WHERE d.name =
"Alzheimer's disease" RETURN p
AS path
```

Cell type in which marker gene associates with AD and is target of an AD drug



Cell type in which both marker genes associate with AD

NCBI Cell Pilot



Article

A molecular single-cell lung atlas of lethal COVID-19

114 | Nature | Vol 595 | 1 July 2021

https://doi.org/10.1038/s41586-021-03569-1

Received: 16 November 2020

Accepted: 19 April 2021

Published online: 29 April 2021

Check for updates

Johannes C. Melms[1,2,33], Jana Biermann[1,2,33], Huachao Huang[3,4,5,33], Yiping Wang[1,2,33], Ajay Nair[5,33], Somnath Tagore[6,33], Igor Katsyv[7,33], André F. Rendeiro[8,9,33], Amit Dipak Amin[1,2,33], Denis Schapiro[10,11], Chris J. Frangieh[11,12], Adrienne M. Luoma[13], Aveline Filliol[5], Yinshan Fang[3,4,5], Hiranmayi Ravichandran[9,14,15], Mariano G. Clausi[16], George A. Alba[17], Meri Rogava[1,2], Sean W. Chen[1,2], Patricia Ho[1,2], Daniel T. Montoro[18,19], Adam E. Kornberg[2], Arnold S. Han[2], Mathieu F. Bakhoum[20], Niroshana Anandasabapathy[9,21,22], Mayte Suárez-Fariñas[23,24], Samuel F. Bakhoum[25,26], Yaron Bram[27], Alain Borczuk[28,29], Xinzheng V. Guo[16], Jay H. Lefkowitch[7], Charles Marboe[7], Stephen M. Lagana[7], Armando Del Portillo[7], Emily J. Tsai[5], Emmanuel Zorn[2], Glen S. Markowitz[7], Robert F. Schwabe[5,30], Robert E. Schwartz[27,34], Olivier Elemento[8,9,15,34], Anjali Saqi[7,34], Hanina Hibshoosh[7,34], Jianwen Que[3,4,5,31,34] ✉ & Benjamin Izar[1,2,31,32,34] ✉

Impaired alveolar epithelial regeneration

Within the epithelial compartment, we identified alveolar epithelial cells (AT1 and AT2 cells; $n = 20{,}949$), airway epithelial cells (basal, ciliated, club, goblet, and mucous cells; $n = 7{,}223$), a cluster characterized by the expression of inflammatory and cell cycle genes, including *IRF8*, *B2M*, *MKI67* and *TOP2A* ('cycling epithelium'; $n = 609$), and a cluster showing high expression of the extracellular matrix (ECM) components *COL6A3*, *COL1A2*, and *COL3A1* ('ECM^{high} epithelium'; $n = 1{,}179$) (Fig. 3a, b, Extended Data Fig. 8a–c, Supplementary Tables 2, 7).

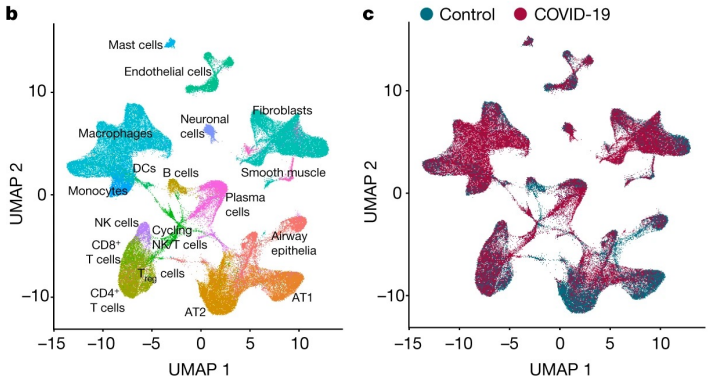
AT2 cells serve as progenitors for AT1 cells during lung regeneration[22]. AT2 and T1 cells in control lungs formed distinct clusters (Fig. 3a, b) and demonstrated the expected changes in differential gene expression (DGE) analysis, including expression of the lineage markers *SFTPC* and *SFTPB* in AT2 cells, and *CLIC5* and *AGER* in AT1 cells (Fig. 3c, Supplementary Table 7). By contrast, clustering of AT2 and AT1 cells in COVID-19 lungs was less discrete, with a substantial portion of cells not overlapping with their control counterparts (Fig. 3b). Both AT2 and AT1 cells from COVID-19 lungs showed decreased overall expression of defining markers (Fig. 3c). COVID-19 AT2 cells displayed decreased expression of *ETV5* (Fig. 3d), a transcription factor that is required for maintaining AT2 cell identity. Decreased *ETV5* expression is associated with differentiation towards AT1 cells[23], indicating that AT2 cells had

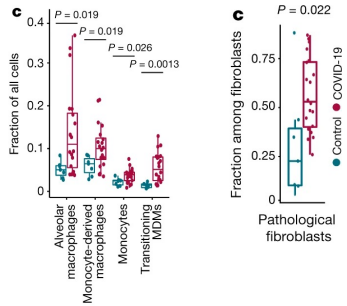
Study design:

- 19 individuals who died of COVID (rapid autopsy); 7 controls
- 116,000 nuclei => snRNA-seq
- 41granular cell types

Results:

- Increase in monocyte/macrophage populations
- Loss of AT1 and AT2 pneumocytes
- Aberrant AT2 => AT1 transitioning
- Increase in pathological fibroblasts



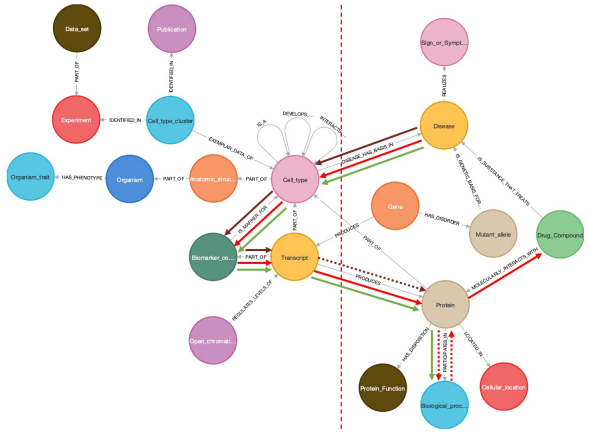


LLM/NLP

Structured Semantic Assertions

AT1_cell IS_A epithelial_cell
AT2_cell IS_A epithelial_cell
AT1_cell DEVELOP_FROM AT2_cell
AT1_cell PART_OF lung
AT2_cell PART_OF lung
CLIC5_transcript PART_OF AT1_cell
AGER_transcript PART_OF AT1_cell
SFTPC_transcript PART_OF AT1_cell
SFTPA1_transcript PART_OF AT1_cell
AT1_cell ASSOCIATED_WITH lethal_COVID

Cypher



NCBI Cell Pilot

Gene/Protein
OMIM
ClinVar
PubChem
Disease Ontology
Gene Ontology
Reactome

Extract

Gene Expression Omnibus
cellxgene
HuBMAP DataPortal
Single Cell Portal
NeMO DataPortal

dataset selection & QC

NS-Forest

biomarker combinations

FR-Match

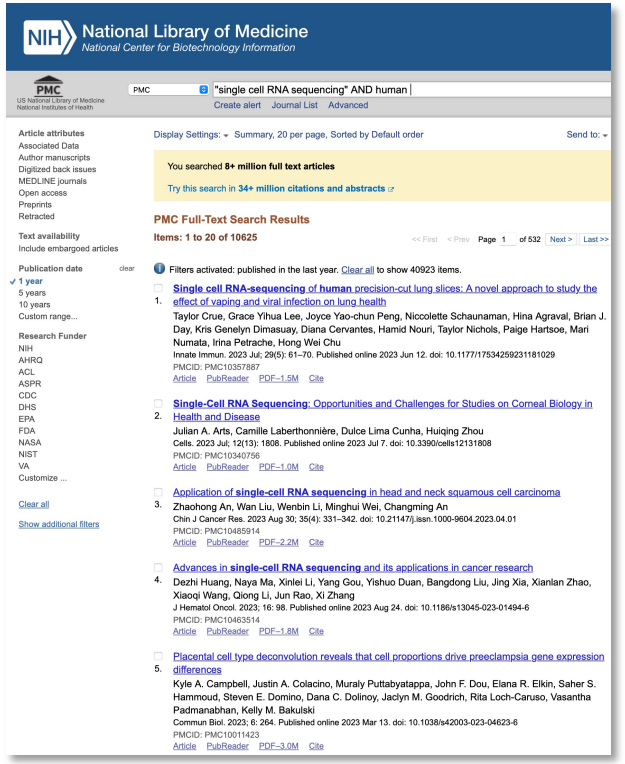
novel cell type identification

Translate/Load

ODK

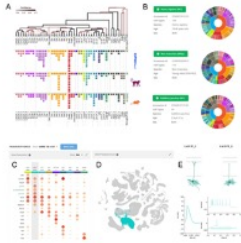
SSS assertions

Neo4j



PMC

user interfaces



cell type knowledge query

cell type matching

CL submission

subgraph query/filter

machine learning

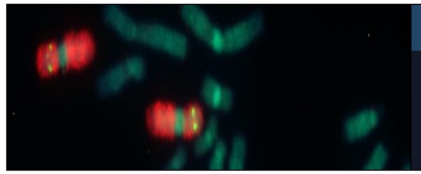
National Library of Medicine
National Center for Biotechnology Information

Log in

Cell

Cell | Search

Advanced | Help



Cell

Cell integrates information from a wide range of species. A record may include cell type nomenclature and synonyms, marker genes, gene and protein expression phenotypes, developmental trajectories, enriched pathways, disease associations, perturbation responses, etc. with links to other NCBI and external resources.

Using Gene

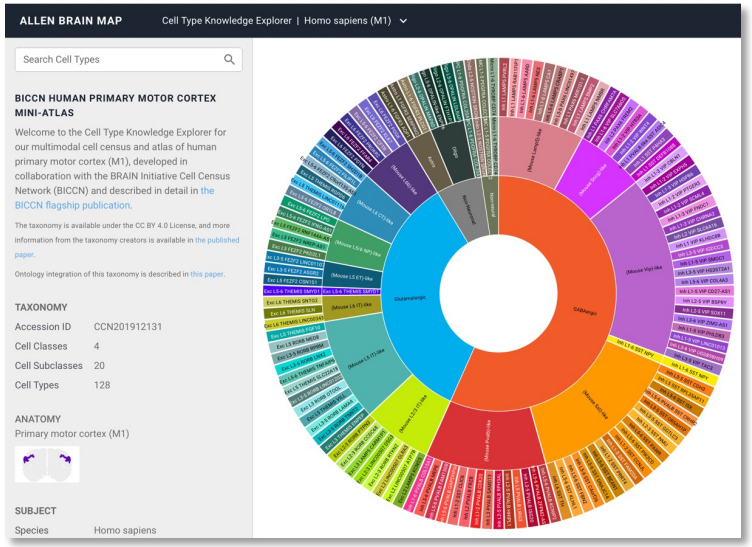
- Cell Quick Start
- FAQ
- Download/FTP
- RefCell mailing list
- Cell News
- Factsheet

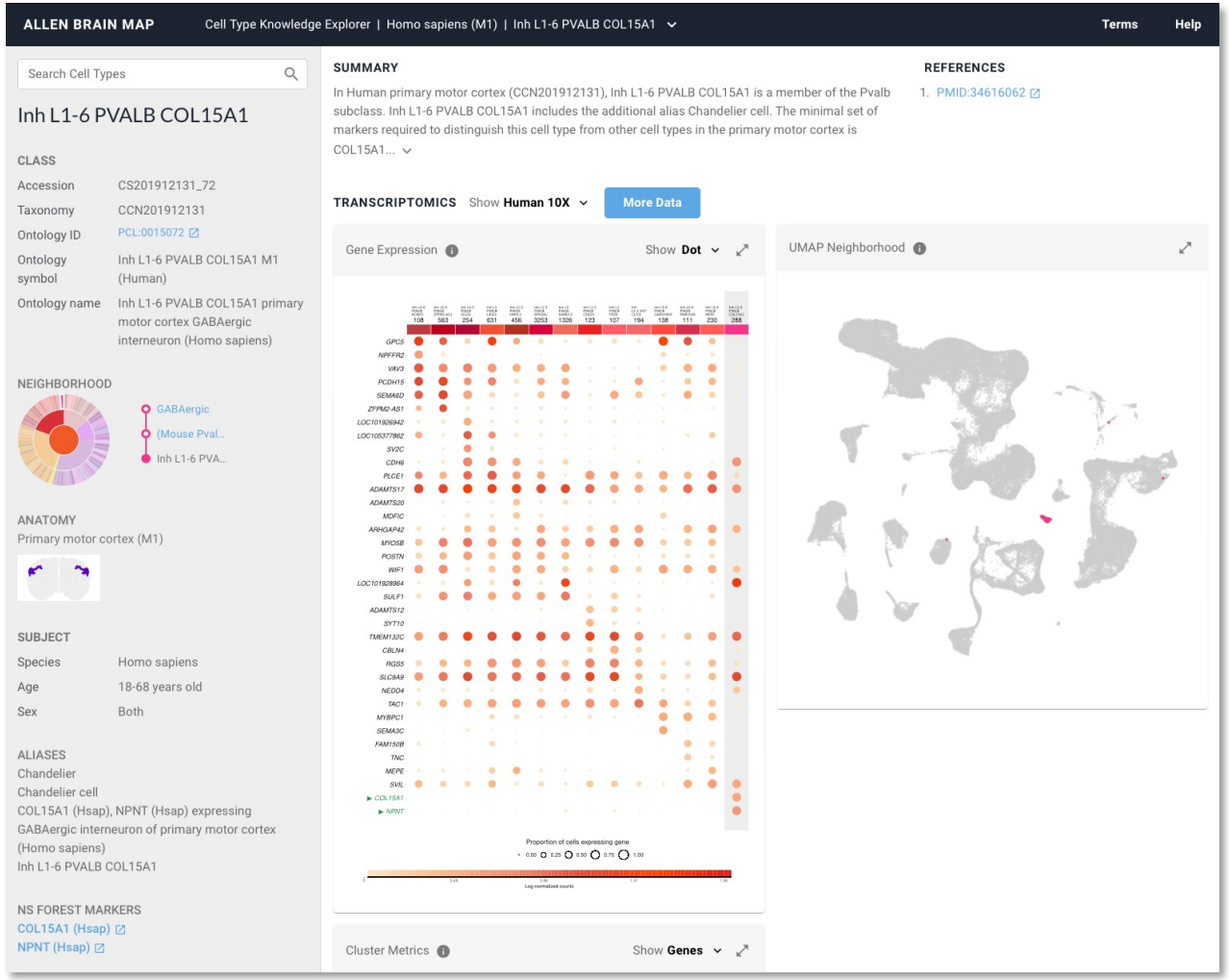
Gene Tools

- Submit GeneRIFs
- Submit Correction
- Statistics
- Cell BLAST
- Cell Workbench
- Splign

Other Resources

- OMIM
- RefCell
- RefSeqGene
- Protein Clusters





Conclusions

- Single cell transcriptomics is revolutionizing our understanding of the cellular complexity of human tissues
- Explainable AI approaches offer significant advantages over deep learning methods by keeping track of useful classification features
- Features that are useful for classification can be used as molecular biomarkers, and frequently point to important underlying biological phenomena
- NS-Forest is an effective dimensionality reduction method for picking out the minimum set of necessary and sufficient marker genes (NS-genes) to identify and define discrete cell types in a defined specimen context
- NS-genes can be used as:
    - a condensed reference data matrix for assessing cell type identity in new datasets using FR-Match
    - characteristics for semantic cell type definitions => Provisional Cell Ontology Knowledgebase
    - probes for multiplex FISH and targets for qPCR assays to identify specific cell types
- Translation of omics data into computable forms of knowledge (e.g. semantic knowledge graphs) can facilitate the integration with other types of knowledge for translational discovery

Resources

- Publications
  - NS-Forest v1.0 - Aevermann B, et al. (2018) ***Human Molecular Genetics***, 27(R1):R40-R47. PMID: 29590361
  - NS-Forest v2.0 - Aevermann B, et al. (2021) ***Genome Research***, 31:1767-1780. PMID: 34088715
  - FR-Match v1.0 - Zhang Y, et al. (2021) ***Briefings in Bioinformatics***, 22:bbaa339. PMID: 33249453
  - FR-Match v2.0 - Zhang Y, et al. (2021) ***Scientific Reports***, 12:9996. PMID: 35705694
  - Cortical layer 1 cell types - Boldog E, et al. (2018) ***Nature Neuroscience***, 21: 1185-1195. PMID: 30150662
  - MTG human cell types - Hodge RD, et al. (2019) ***Nature***, 573:61-68. PMID: 31435019
  - M1 human, mouse, marmoset – Bakken T, et al. (2021) ***Nature***, 598:111-119. PMID: 34616062
- Source Code
  - NS-Forest source code is available at https://github.com/JCVenterInstitute/NSForest
  - FR-Match source code is available at https://github.com/JCVenterInstitute/FRmatch
- Protocols
  - NS-Forest protocol is available at https://www.protocols.io/view/ns-forest-version-2-un7evhn
  - FR-Match protocol is available at https://www.protocols.io/view/fr-match-cell-type-matching-for-scrnaseq-data-bmyfk7tn
- Ontology
  - PCL is available through the BioPortal - https://bioportal.bioontology.org/ontologies/PCL

Acknowledgements

J. Craig Venter Institute

Renee (Yun) Zhang
Brian Aevermann
Mohamed Keshk
Mark Novotny
Steven Lee
Roger Lasken
Nik Schork
Joyce Hsiao
Neil Tedeschi
Kavya Chegireddy

UC San Diego

Dan Carrillo
Eric Duong
Aditi Gandhi
Huy Le
Joshua Lau
Seunghyun Lee
Yueshan Liang
Tian Liu
Christopher Lin
Beverly Peng
Noura Tbeileh
Janelle Uy

European Bioinformatics Institute

David Osumi-Sutherland
Shawn Tan
Huseyin Kir

National Library of Medicine

Ajith V. Pankajam
Vinh Nguyen
Noam Rotenberg
Zhiyong Lu
Ling Luo
Don Comeau
Robert Leaman

Allen Institute for Brain Science

Ed Lein
Trygve Bakken
Jeremy Miller
Rebecca Hodge
Mike Hawrylycz

University of Leiden

Boudewijn Lelieveldt