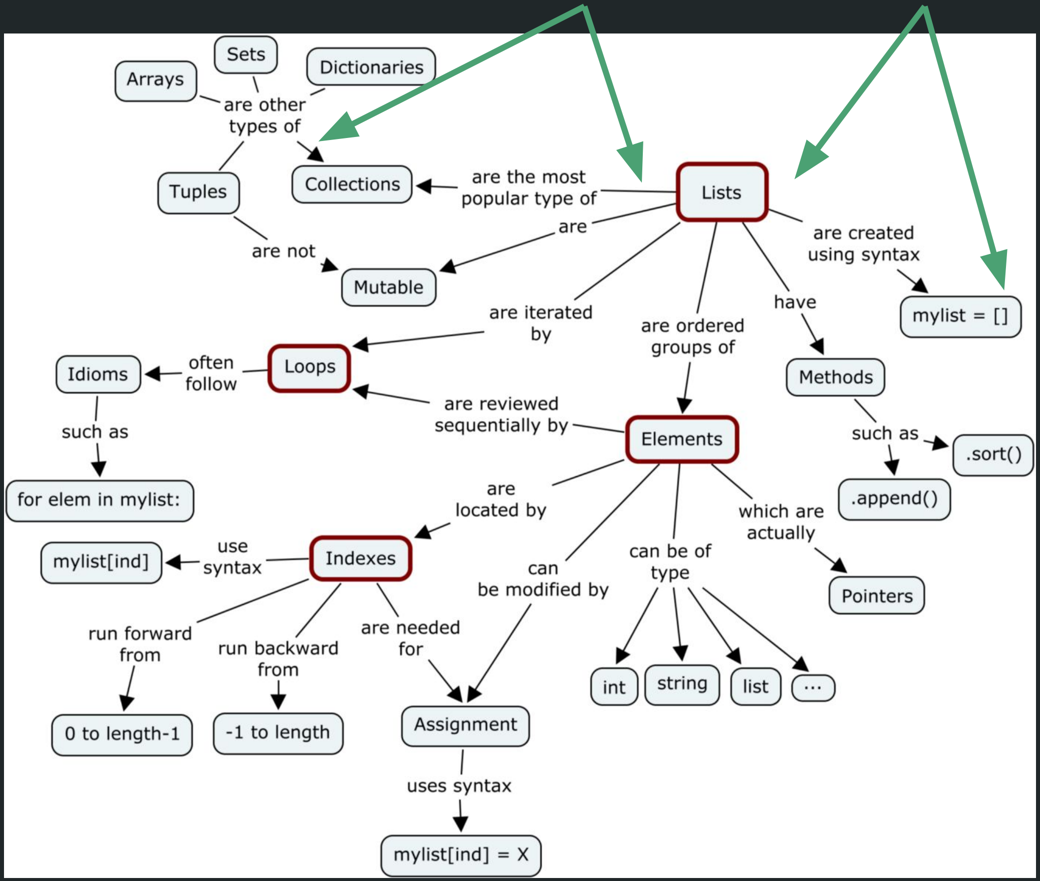
Putting this in terms of graph theory:

Concept maps

(A directed graph/network)

Use paper or tablet or https://jamboard.google.com/ or https://excalidraw.com/ to create yours and submit on Canvas

You need at least **10 nodes** and **15 edges**

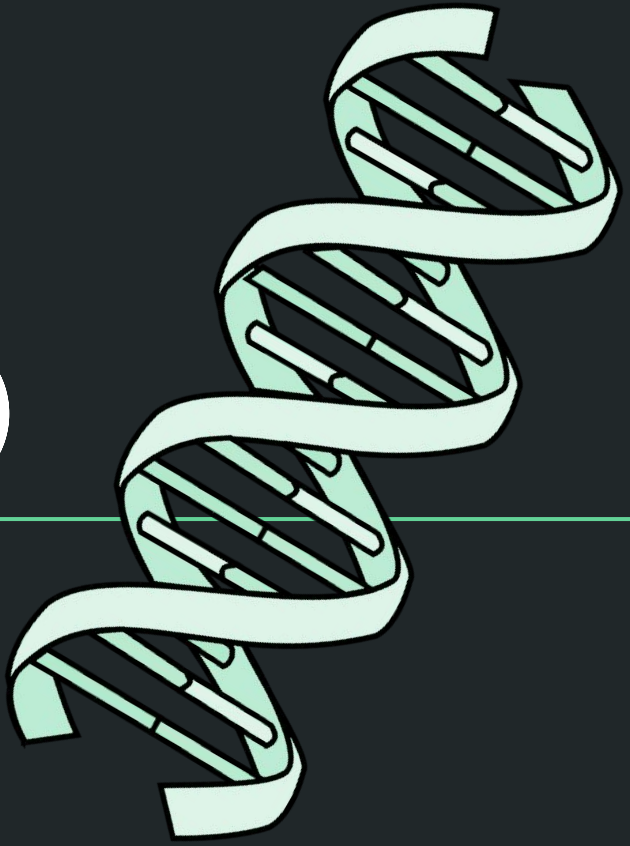


“edges”

“nodes”

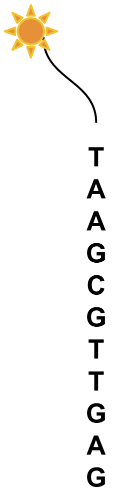
Gene Expression (RNAseq & Patch-Seq)

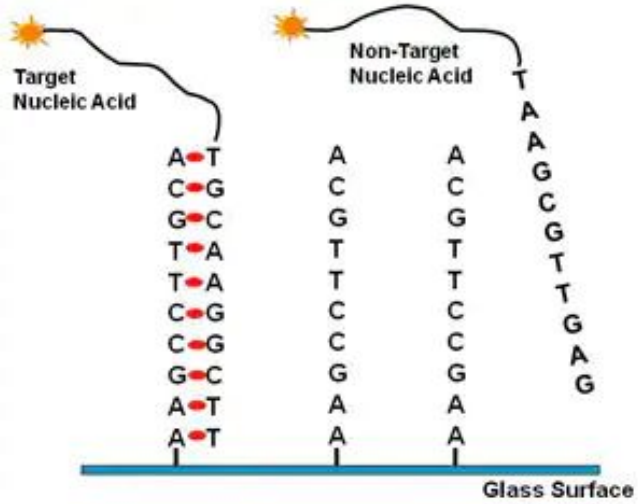
BIPN 162

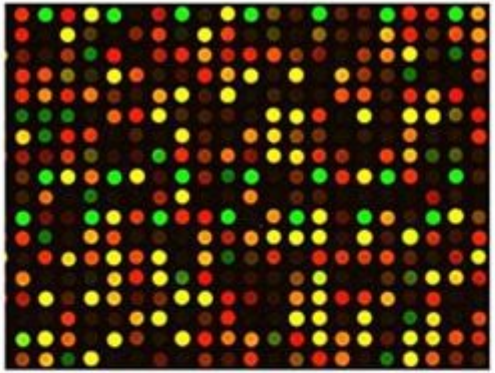


Objectives for today

- Compare and contrast DNA microarrays and RNA-seq
- Describe Patch-Seq
- Use Pandas to manipulate a Patch-Seq dataset
- Identify reasons to **transform** and **normalize** data
- Implement a log transform and normalization







DNA Microarray

Extract mRNA from samples

Create cDNA (via reverse transcriptase) **attached to fluorophore**

Bind to a microarray chip with Whole Human Genome probe set

Positive controls: Pooled RNA samples from same brain, and other brains

Negative control: nuclease free deionized water (NFdH2O)

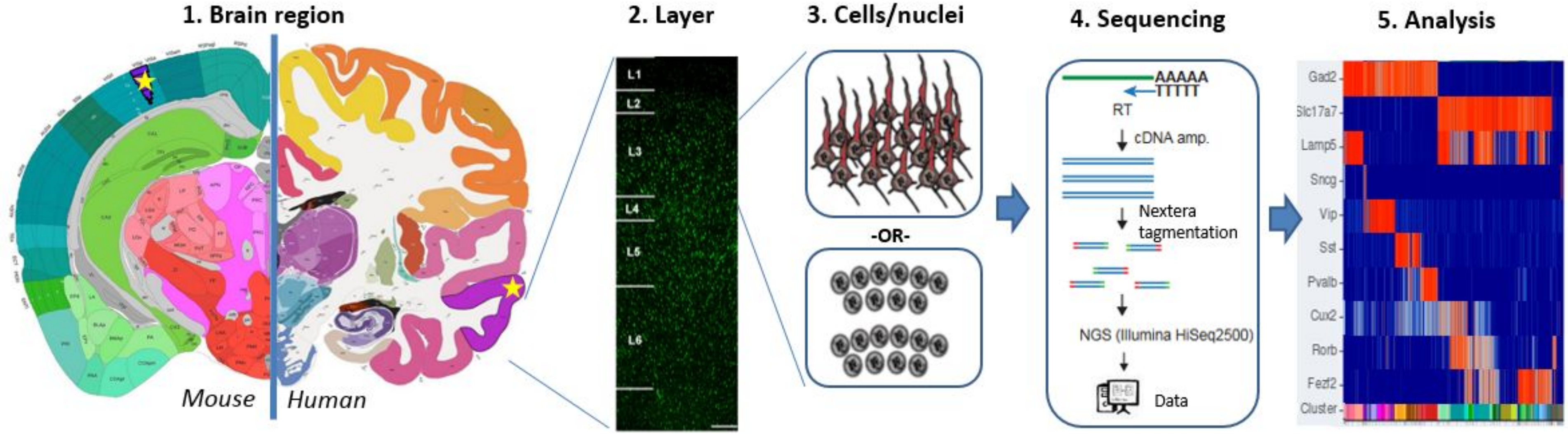
Normalize expression across all samples (from one subject)

and later, across batches of experiments.

Reminder: microarrays

Images: Agilent

RNA-seq (by brain region)



https://portal.brain-map.org/atlases-and-data/rnaseq

What is RNA-seq?

- Very sensitive assay (single nucleotide, different exons in gene) to measure gene expression
- *Constantly evolving* technology & becoming more inexpensive

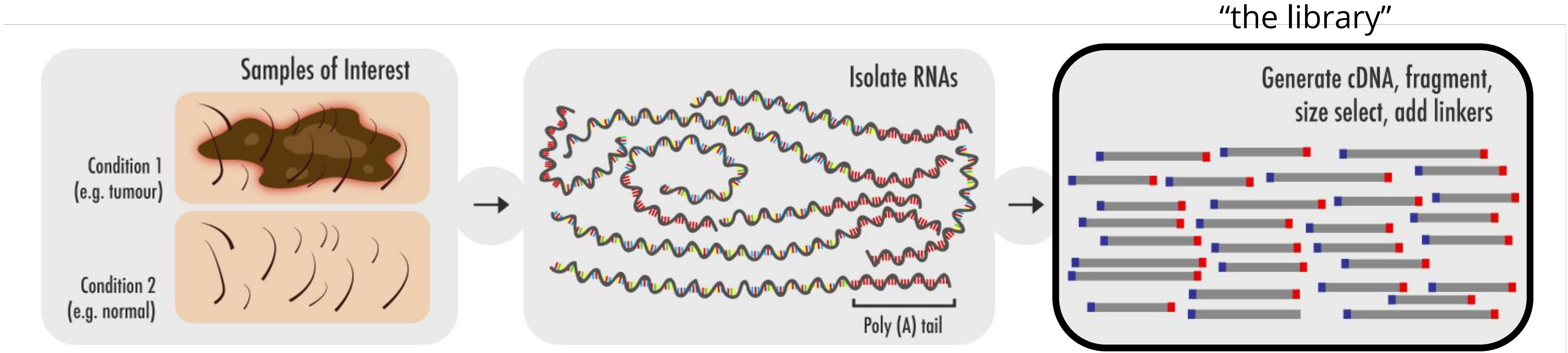
Several different kinds:

- **Bulk** or **single-cell (scRNA-seq)**
- **Single-read** (cheaper & faster) or **paired-end** (deeper reads)
- **Strand-specific** or **non-strand-specific**
- **With** or **without** a **reference genome**

Information on single-cell RNA-seq: Introduction to Single-Cell RNA Sequencing - Olsen - 2018 - Current Protocols in Molecular Biology - Wiley Online Library

Different options: Genome Sequencing: Defining Your Experiment | Columbia University Department of Systems Biology

RNA-Seq: Basics, Applications and Protocol | Technology Networks



Sequencing can be done in multiple ways.

If you're curious, see this video.

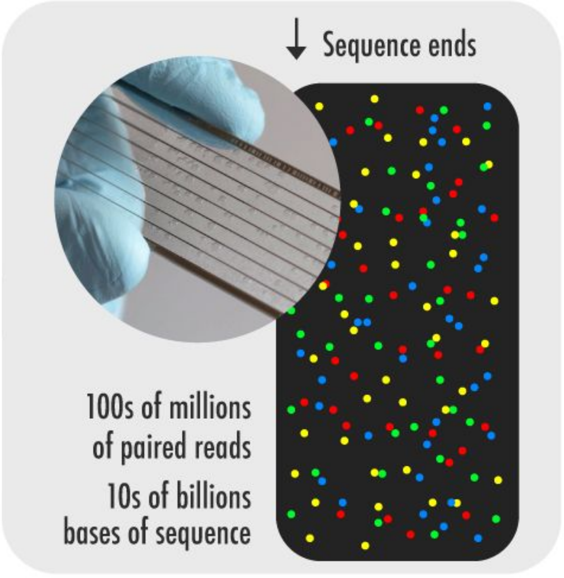
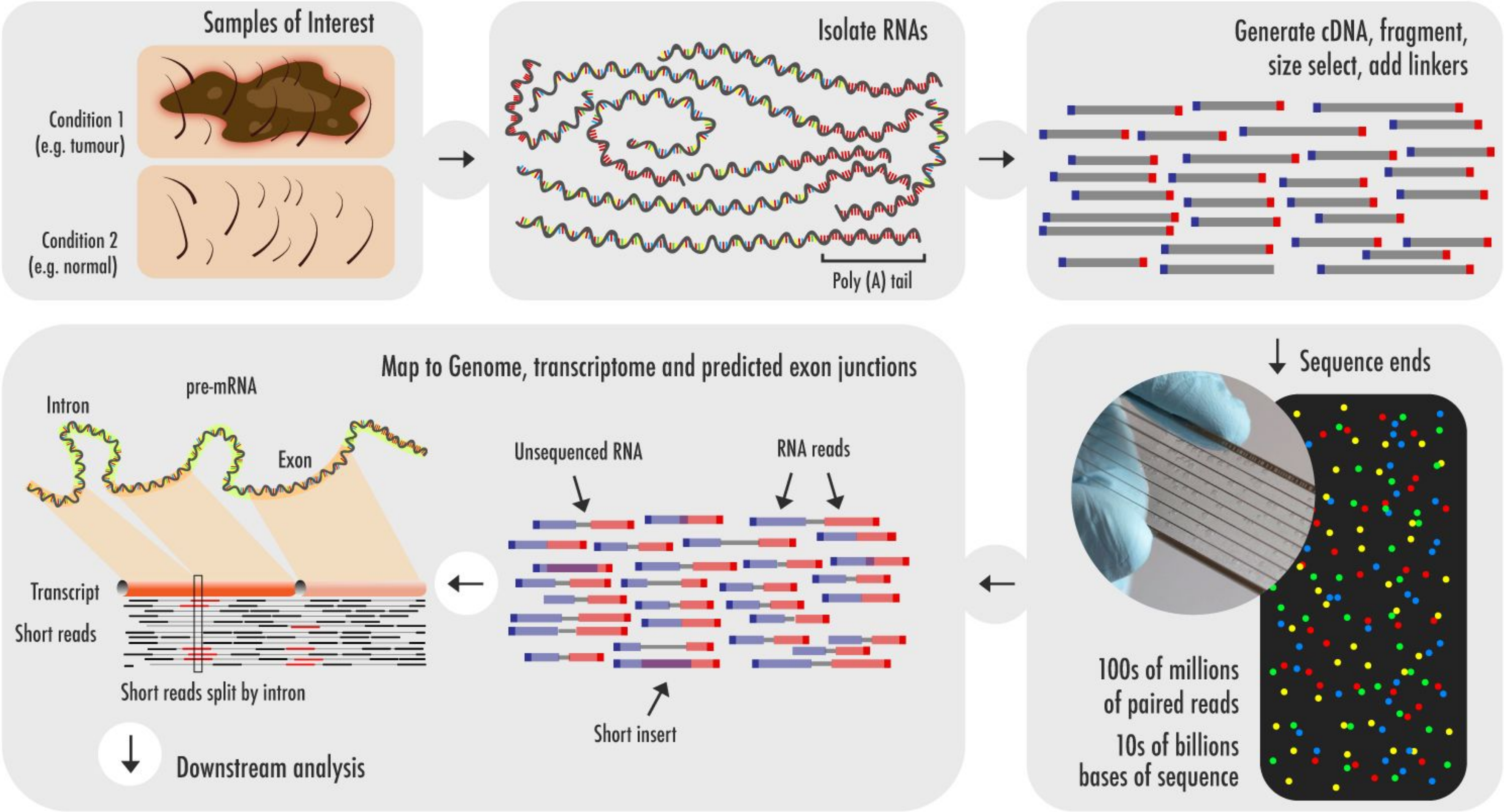
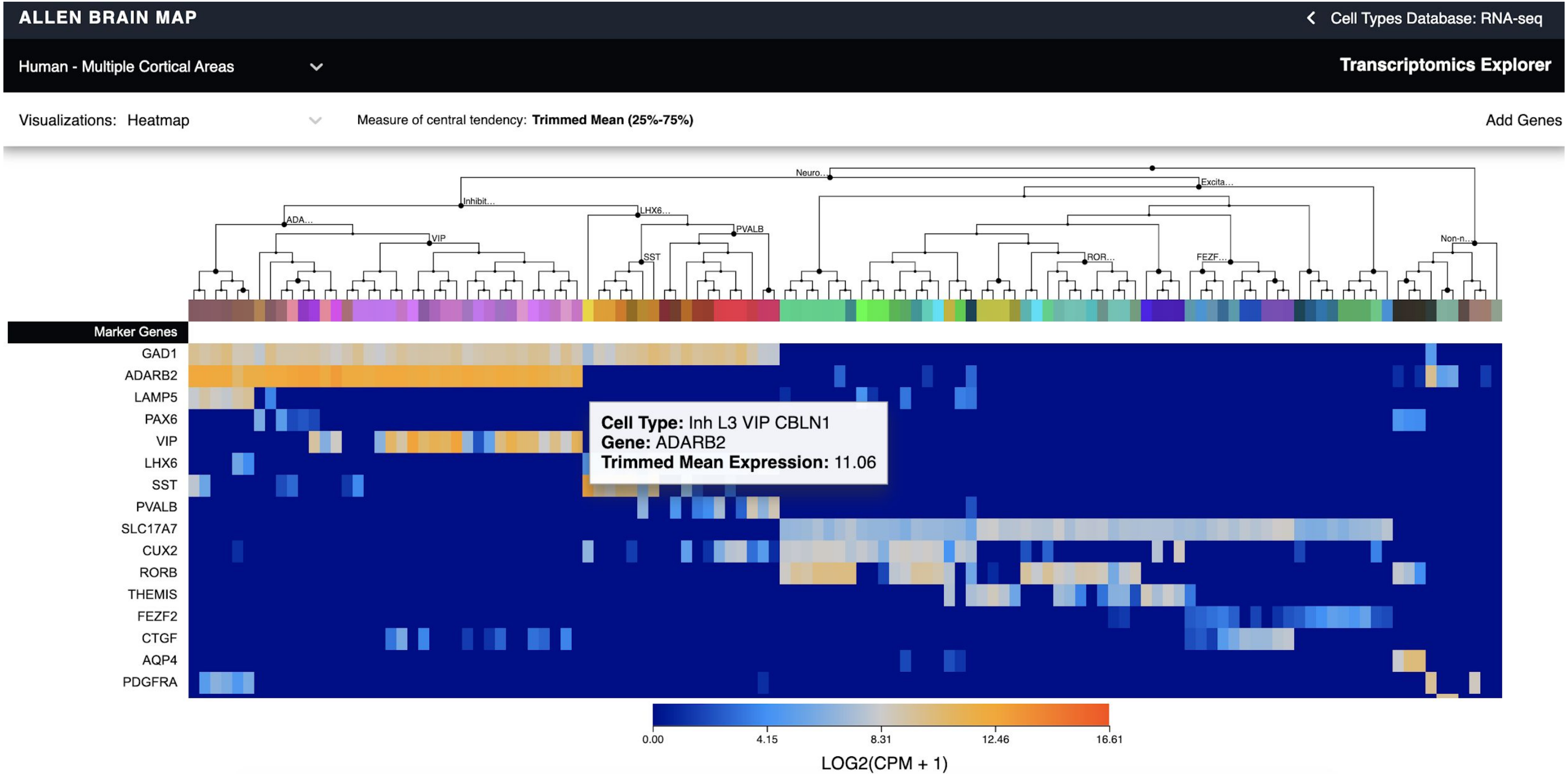


Image: https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461


Samples of Interest
Condition 1
(e.g. tumour)
Condition 2
(e.g. normal)
Isolate RNAs
Poly (A) tail
Generate cDNA, fragment,
size select, add linkers
Sequence ends
100s of millions
of paired reads
10s of billions
bases of sequence
Map to Genome, transcriptome and predicted exon junctions
Unsequenced RNA
RNA reads
Short insert
Intron
pre-mRNA
Exon
Transcript
Short reads
Short reads split by intron
Downstream analysis



ALLEN BRAIN MAP

Cell Types Database: RNA-seq

Human - Multiple Cortical Areas

Transcriptomics Explorer

Visualizations: Heatmap

Measure of central tendency: **Trimmed Mean (25%-75%)**

Add Genes

Cell Type: Inh L3 VIP CBLN1
Gene: ADARB2
Trimmed Mean Expression: 11.06

https://celltypes.brain-map.org/rnaseq/human/cortex

Normalization of RNAseq data

- **Normalization** is the process of rescaling values to be able to make comparisons (e.g., across reads, cells, brain areas, subjects, etc.)
  - Typically, we divide by the thing we're controlling for!

- Because the number of reads per cell in RNAseq can vary, it is commonplace to normalize by the total # of reads
  - This normalizes by **depth** but not gene length (shorter genes are sequenced more often!)

- One common way to normalize raw read counts is to calculate **counts per million (CPM)**, obtained by dividing read counts for genes by the sum of raw read counts in that cell (**library size**), and multiplying the results by a million

Comparing microarrays to RNA-seq

	Microarray	**RNA-seq**
Principle	Hybridization	Cloning & sequencing
Required amount of RNA	High	Low
Resolution	Several to 100 bp	Single base
Distinguish allelic expression	Limited	Yes
Discover new genes	No (requires species-specific probes)	Yes
Dynamic range	Few hundred-fold	> 8000-fold
Cost	Medium	High (computational, but getting cheaper)

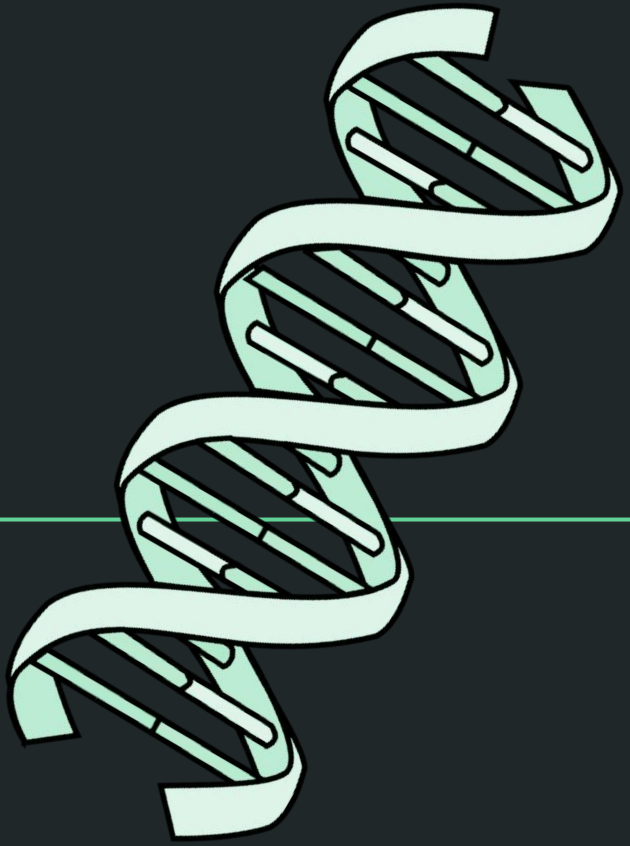
Slide: Jeremy Miller, Allen Brain Institute (Talk)

Important points to remember about measuring gene expression via microarrays or RNAseq

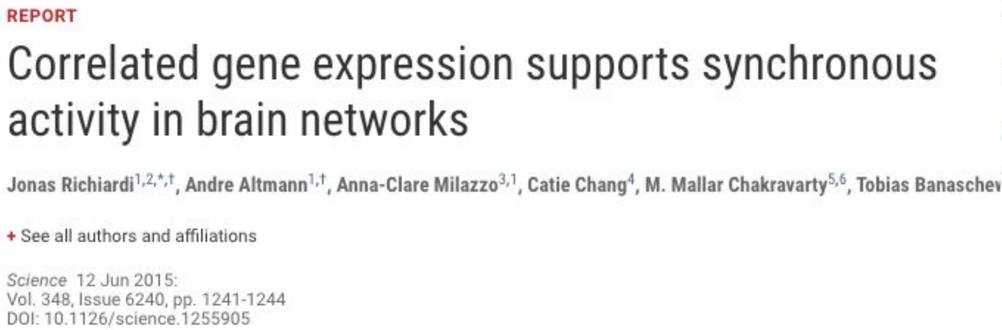
- *Neither* measure actual protein output — which may vary from genes or transcript abundance
- *Both* provide relative measurements that need to be normalized (and the mechanism of normalization can impact results)
- RNA quality also impacts results, particularly for post mortem human studies
- Findings should always be validated (e.g., with Western blotting or immunostaining for proteins, or RT-PCR for better quantification)

Kiel et al. (2018)

Bringing this back to data science...



Sample publications using the Allen Institute for Brain Science microarray dataset



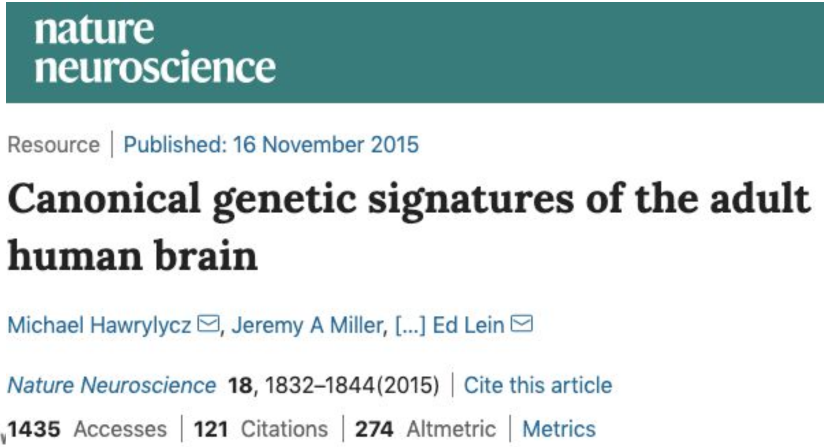
REPORT

Correlated gene expression supports synchronous activity in brain networks

Jonas Richiardi[1,2,*,†], Andre Altmann[1,†], Anna-Clare Milazzo[3,1], Catie Chang[4], M. Mallar Chakravarty[5,6], Tobias Banaschev

+ See all authors and affiliations

Science 12 Jun 2015:
Vol. 348, Issue 6240, pp. 1241-1244
DOI: 10.1126/science.1255905



nature neuroscience

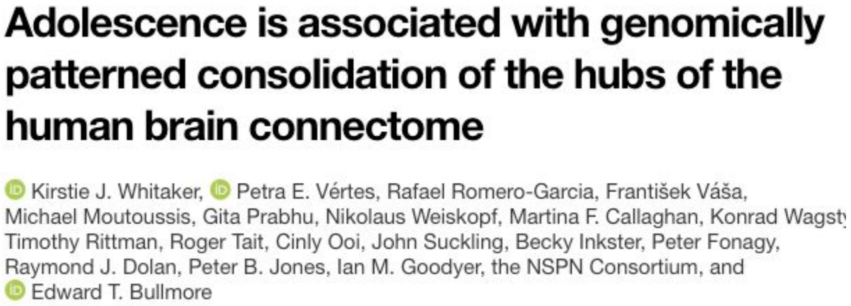
Resource | Published: 16 November 2015

Canonical genetic signatures of the adult human brain

Michael Hawrylycz ✉, Jeremy A Miller, [...] Ed Lein ✉

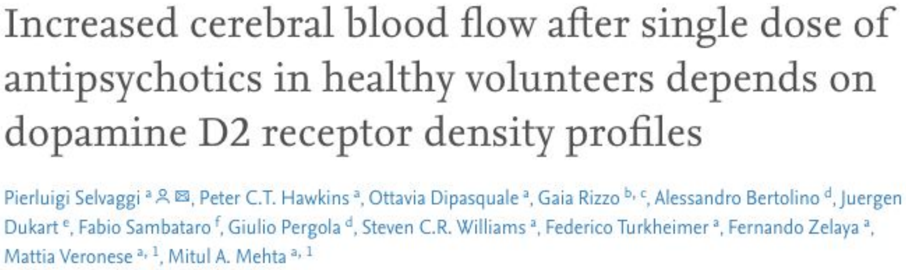
Nature Neuroscience **18**, 1832–1844(2015) | Cite this article

1435 Accesses | **121** Citations | **274** Altmetric | Metrics



Adolescence is associated with genomically patterned consolidation of the hubs of the human brain connectome

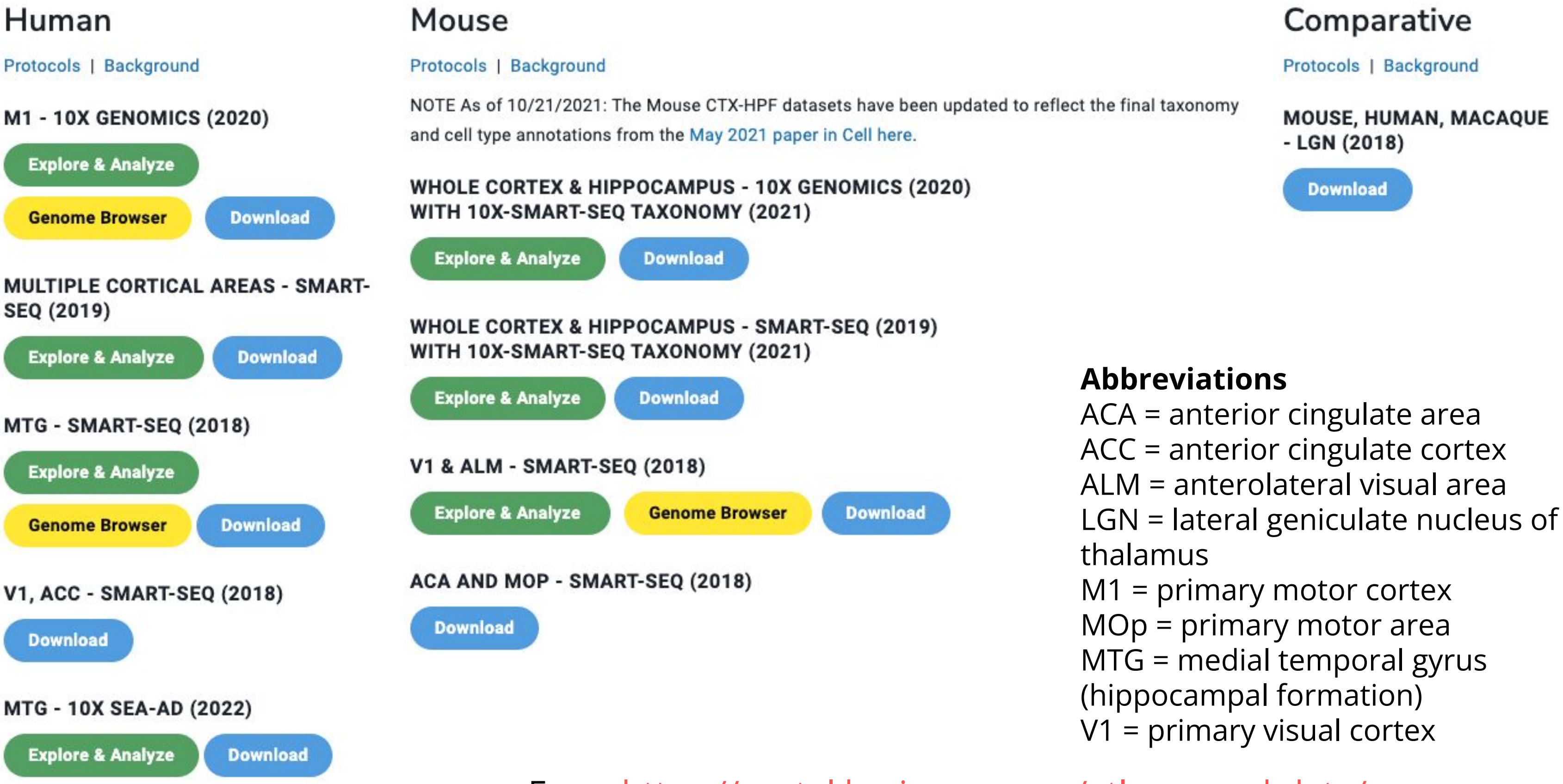
Kirstie J. Whitaker, Petra E. Vértes, Rafael Romero-Garcia, František Váša, Michael Moutoussis, Gita Prabhu, Nikolaus Weiskopf, Martina F. Callaghan, Konrad Wagsty Timothy Rittman, Roger Tait, Cinly Ooi, John Suckling, Becky Inkster, Peter Fonagy, Raymond J. Dolan, Peter B. Jones, Ian M. Goodyer, the NSPN Consortium, and Edward T. Bullmore



Increased cerebral blood flow after single dose of antipsychotics in healthy volunteers depends on dopamine D2 receptor density profiles

Pierluigi Selvaggi [a] ✉, Peter C.T. Hawkins [a], Ottavia Dipasquale [a], Gaia Rizzo [b, c], Alessandro Bertolino [d], Juergen Dukart [e], Fabio Sambataro [f], Giulio Pergola [d], Steven C.R. Williams [a], Federico Turkheimer [a], Fernando Zelaya [a], Mattia Veronese [a, 1], Mitul A. Mehta [a, 1]

Overview of Allen RNA-seq datasets



Human

Protocols | Background

M1 - 10X GENOMICS (2020)

Explore & Analyze

Genome Browser Download

MULTIPLE CORTICAL AREAS - SMART-SEQ (2019)

Explore & Analyze Download

MTG - SMART-SEQ (2018)

Explore & Analyze

Genome Browser Download

V1, ACC - SMART-SEQ (2018)

Download

MTG - 10X SEA-AD (2022)

Explore & Analyze Download

Mouse

Protocols | Background

NOTE As of 10/21/2021: The Mouse CTX-HPF datasets have been updated to reflect the final taxonomy and cell type annotations from the May 2021 paper in Cell here.

WHOLE CORTEX & HIPPOCAMPUS - 10X GENOMICS (2020) WITH 10X-SMART-SEQ TAXONOMY (2021)

Explore & Analyze Download

WHOLE CORTEX & HIPPOCAMPUS - SMART-SEQ (2019) WITH 10X-SMART-SEQ TAXONOMY (2021)

Explore & Analyze Download

V1 & ALM - SMART-SEQ (2018)

Explore & Analyze Genome Browser Download

ACA AND MOP - SMART-SEQ (2018)

Download

Comparative

Protocols | Background

MOUSE, HUMAN, MACAQUE - LGN (2018)

Download

Abbreviations

ACA = anterior cingulate area

ACC = anterior cingulate cortex

ALM = anterolateral visual area

LGN = lateral geniculate nucleus of the thalamus

M1 = primary motor cortex

MOp = primary motor area

MTG = medial temporal gyrus (hippocampal formation)

V1 = primary visual cortex

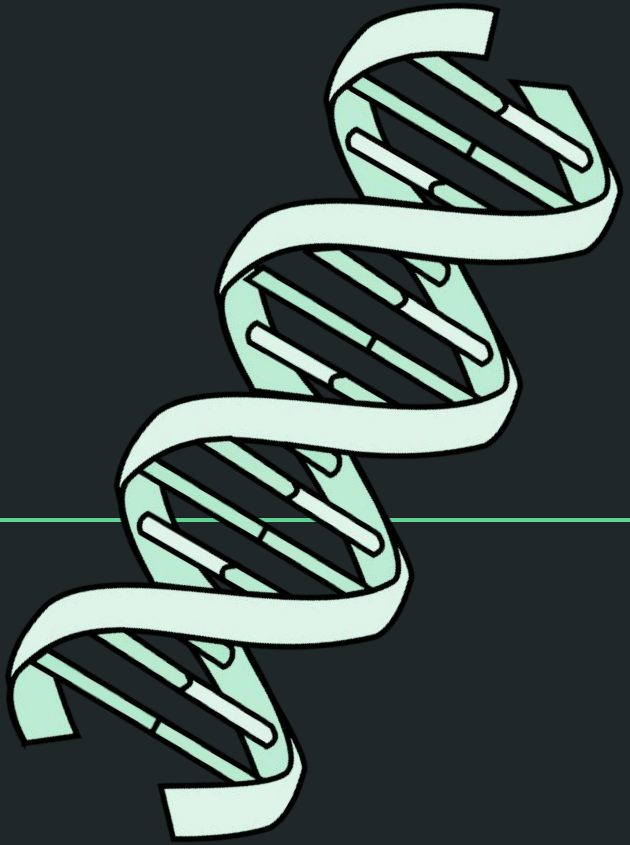
From https://portal.brain-map.org/atlases-and-data/rnaseq

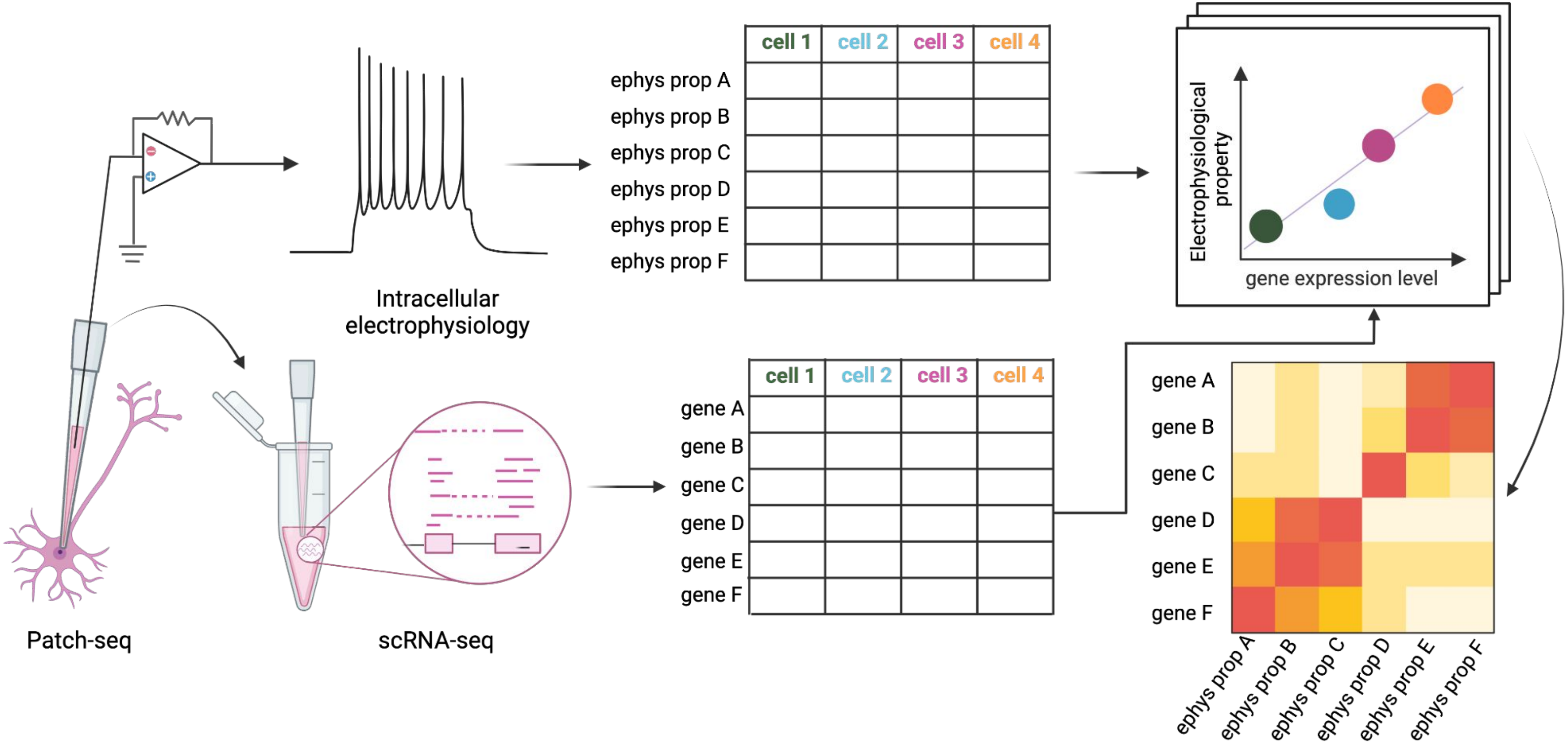
(Somewhat comprehensive) overview of Allen gene expression datasets

	DNA Microarray http://human.brain-map.org/microarray/search	**RNA-seq** https://portal.brain-map.org/atlases-and-data/rnaseq	***In situ* hybridization** https://human.brain-map.org/ish/search https://mouse.brain-map.org/search/index Across development: http://brainspan.org/ish
Whole brain	**Human**	--	**Mouse**
Cortex	**Human**	**Mouse, human**	**Mouse, human**
Motor Cortex		**Human**	
Visual Cortex	**Human**	**Mouse (V1 & ALM), human (V1)**	**Mouse, human**
Subcortex (e.g., caudate putamen)	**Human**	--	**Mouse, human**
Lateral geniculate nucleus	**Human**	**Mouse, human**	**Mouse**
Hippocampus	**Human**	**Mouse**	**Mouse**
Medial temporal gyrus	**Human**	**Human**	
Anterior cingulate cortex/area	**Human**	**Mouse, human**	**Mouse**

What if we want information about a cell's gene expression ***and*** its electrophysiology?

PATCH-SEQ!



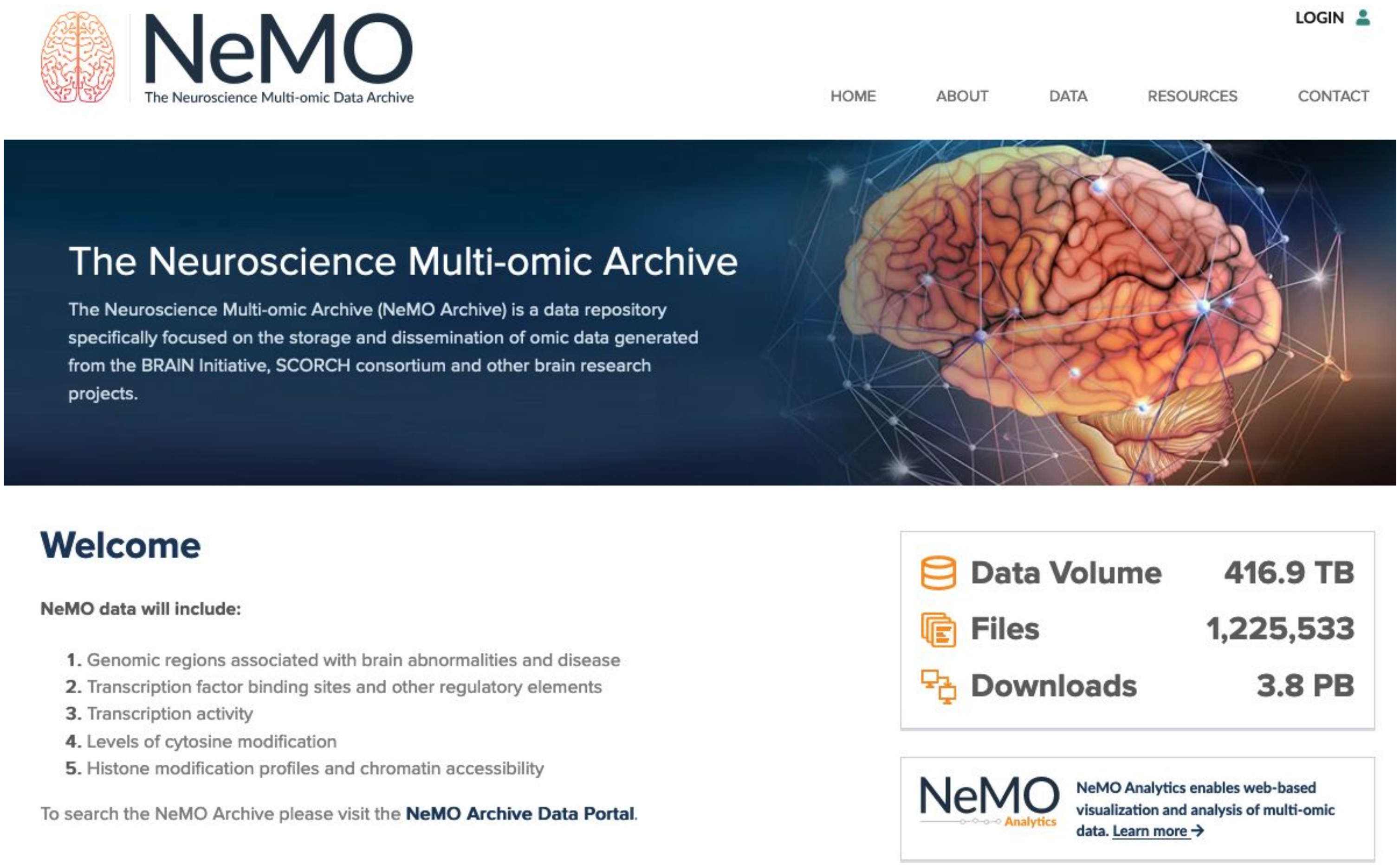


	cell 1	cell 2	cell 3	cell 4
ephys prop A				
ephys prop B				
ephys prop C				
ephys prop D				
ephys prop E				
ephys prop F				

	cell 1	cell 2	cell 3	cell 4
gene A				
gene B				
gene C				
gene D				
gene E				
gene F				

Patch-seq overview (see Gouwens et al. 2020 for details)

Image: Nuo Xu



The Neuroscience Multi-omic Archive

The Neuroscience Multi-omic Archive (NeMO Archive) is a data repository specifically focused on the storage and dissemination of omic data generated from the BRAIN Initiative, SCORCH consortium and other brain research projects.

Welcome

NeMO data will include:

1. Genomic regions associated with brain abnormalities and disease
2. Transcription factor binding sites and other regulatory elements
3. Transcription activity
4. Levels of cytosine modification
5. Histone modification profiles and chromatin accessibility

To search the NeMO Archive please visit the **NeMO Archive Data Portal**.

Data Volume	416.9 TB
Files	1,225,533
Downloads	3.8 PB

NeMO Analytics enables web-based visualization and analysis of multi-omic data. Learn more →

Where the https://nemoarchive.org/

Common file types in data science

not compressed

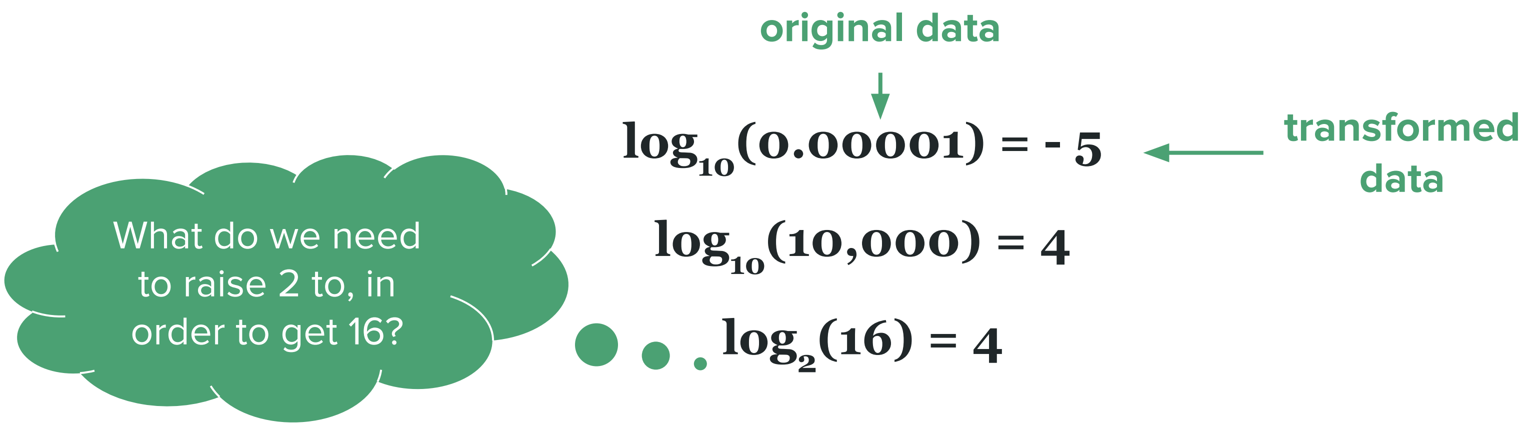
- **Comma-delimited file** (.csv): each data point separated by a comma: `.42,84,24`
- **Tab-delimited file** (.tsv): each data point separated by a tab: `.42	84	24`
- **JSON, HTML, XML** (info)
- **Tar file**
  - Tape Archive
  - Combines multiple files into one (it’s **compressed**)

Log transformations

Why do we transform data?

- Purely for visualization -- *less problematic*
- To run statistics that make assumptions about distribution of data (e.g., require normality) -- *more problematic and not always useful; see discussion here*

A **log transform** is just one way of transforming data.



original data

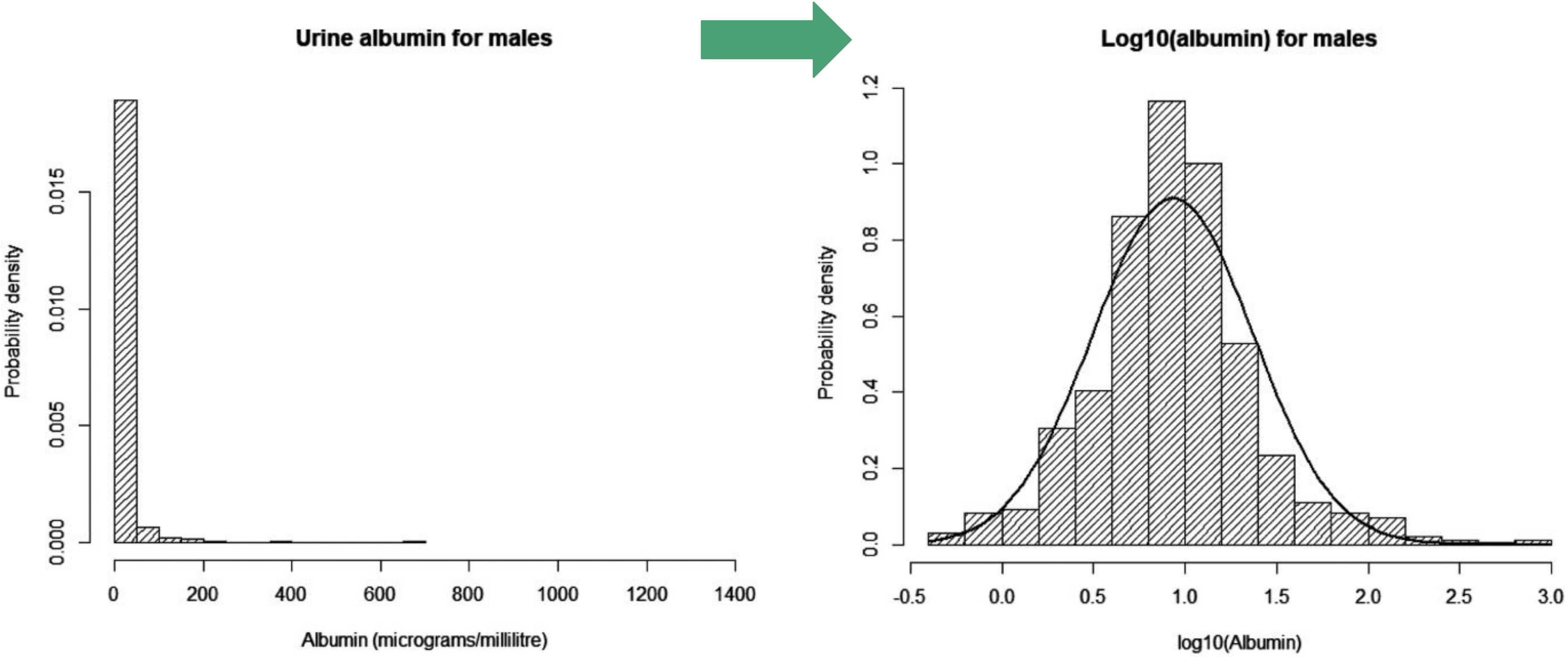
$$\log_{10}(0.00001) = -5$$

transformed data

$$\log_{10}(10{,}000) = 4$$

$$\log_2(16) = 4$$

What do we need to raise 2 to, in order to get 16?



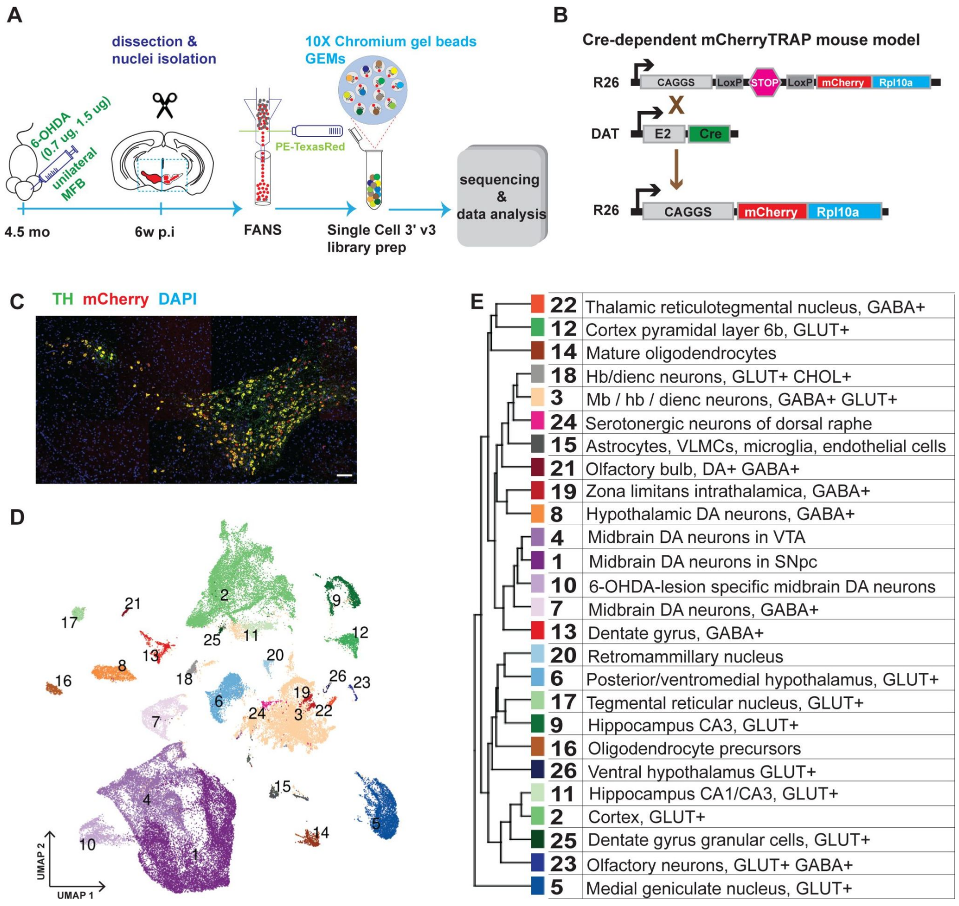
Example of log transform (from this paper)
“The main use of the urinary albumin/creatinine ratio is to provide early evidence of microvascular renal disease in patients with diabetes; values much above 3 mg/mmol are considered to be clinically significant”

(Brief) Introduction to dimensionality reduction & UMAP

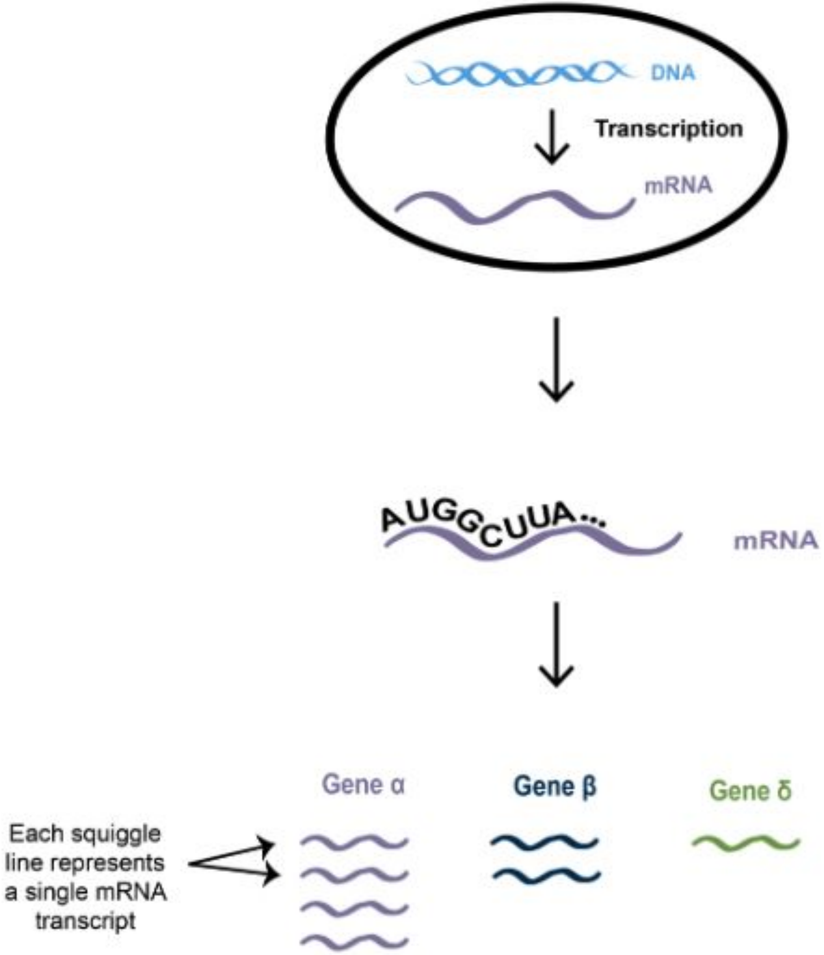
Dimensionality reduction is a technique that helps represent many-dimensional data in fewer dimensions -- *we'll spend a whole week on this later!*

Uniform Manifold Approximation and Projection (UMAP) is one way of transforming and visualizing high dimensional data so that it is more interpretable, commonly transcriptomics data.

https://alleninstitute.org/resource/what-is-a-umap/
https://umap-learn.readthedocs.io/en/latest/how_umap_works.html


A
dissection & nuclei isolation
10X Chromium gel beads GEMs
6-OHDA (0.7 ug, 1.5 ug) unilateral MFB
PE-TexasRed
4.5 mo
6w p.i
FANS
Single Cell 3' v3 library prep
sequencing & data analysis
B
Cre-dependent mCherryTRAP mouse model
R26
CAGGS
LoxP
STOP
LoxP
mCherry
Rpl10a
X
DAT
E2
Cre
R26
CAGGS
mCherry
Rpl10a
C
TH mCherry DAPI
D
UMAP 1
UMAP 2
2
9
21
17
12
25
11
13
20
8
16
18
26
23
6
19
24
3
22
7
15
4
5
14
10
1
E
22 Thalamic reticulotegmental nucleus, GABA+
12 Cortex pyramidal layer 6b, GLUT+
14 Mature oligodendrocytes
18 Hb/dienc neurons, GLUT+ CHOL+
3 Mb / hb / dienc neurons, GABA+ GLUT+
24 Serotonergic neurons of dorsal raphe
15 Astrocytes, VLMCs, microglia, endothelial cells
21 Olfactory bulb, DA+ GABA+
19 Zona limitans intrathalamica, GABA+
8 Hypothalamic DA neurons, GABA+
4 Midbrain DA neurons in VTA
1 Midbrain DA neurons in SNpc
10 6-OHDA-lesion specific midbrain DA neurons
7 Midbrain DA neurons, GABA+
13 Dentate gyrus, GABA+
20 Retromammillary nucleus
6 Posterior/ventromedial hypothalamus, GLUT+
17 Tegmental reticular nucleus, GLUT+
9 Hippocampus CA3, GLUT+
16 Oligodendrocyte precursors
26 Ventral hypothalamus GLUT+
11 Hippocampus CA1/CA3, GLUT+
2 Cortex, GLUT+
25 Dentate gyrus granular cells, GLUT+
23 Olfactory neurons, GLUT+ GABA+
5 Medial geniculate nucleus, GLUT+

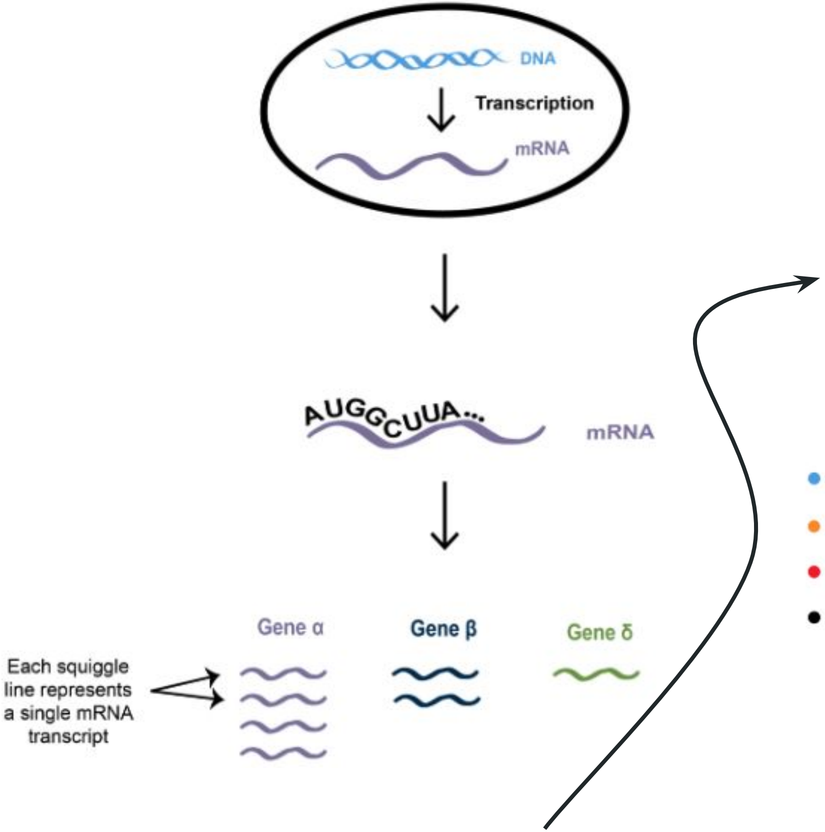
scRNA-seq in mouse midbrain
https://elifesciences.org/reviewed-preprints/89482v2



Isolate the nuclei from the cells in the sample of brain tissue and extract the RNA found in each nucelus.

Sequence the mRNA transcripts found in each cell's nucleus in order to determine which genes each brain cell was expressing.

Count the number of mRNA transcripts found for each gene. This allows us to quantify how much each cell was expressing each gene.



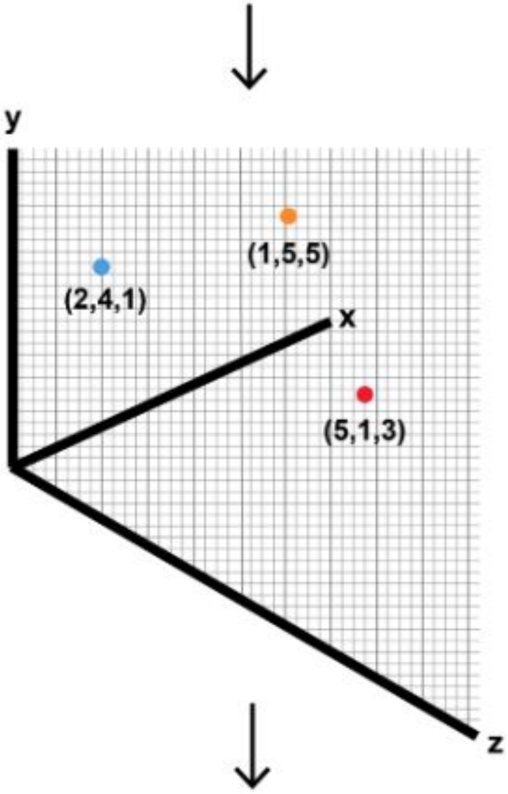
	Gene α	Gene β	Gene δ
# of mRNA transcripts found in Cell 1	4	2	1

Create a table that shows how much each cell was expressing each gene.

		Gene α	Gene β	Gene δ
●	# of mRNA transcripts found in Cell 1	2	4	1
●	# of mRNA transcripts found in Cell 2	1	5	5
●	# of mRNA transcripts found in Cell 3	5	1	3
●	repeat count for thousands of cells...	...	...	...

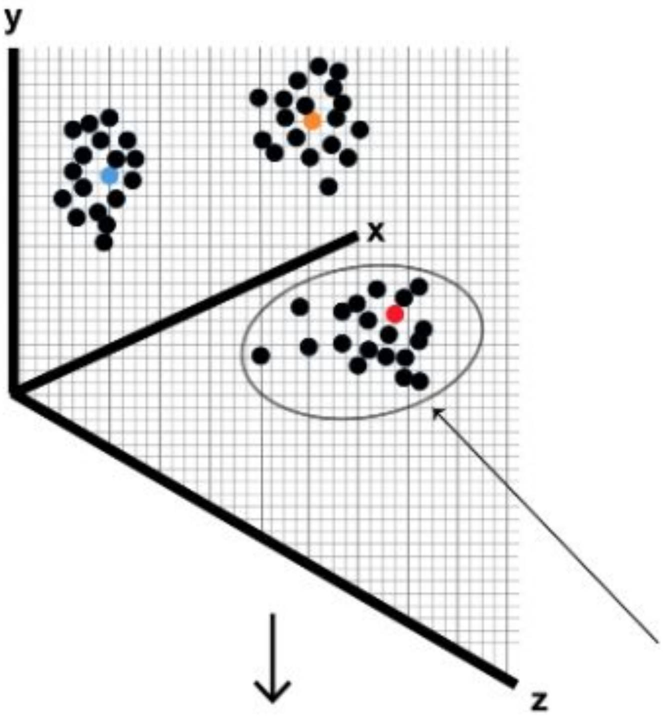
Repeat this process for THOUSANDS of cells. Remember, this means we are counting how much EACH cell was expressing EACH gene. If we wanted to create a table that listed the data in full, this data table would have thousands of rows.

x = gene α value
y = gene β value
z = gene δ value



If we wanted to create a graph that plotted the inital data for cell 1, cell 2, and cell 3 and their relative amount of expression of gene alpha, gene beta, and gene delta, we would need a 3D graph like the one on the left.

x = gene α value
y = gene β value
z = gene δ value

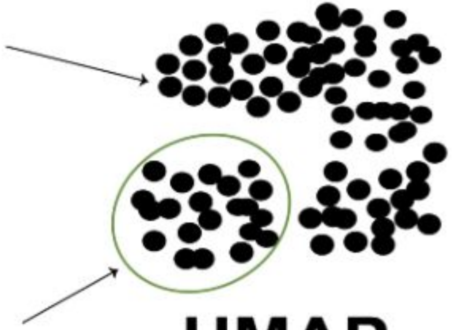


We can repeat this process for the thousands of cells that were collected from the brain tissue sample. Notice that the cells begin to cluster based on how similar their gene expression for gene alpha, gene beta, and gene delta is to one another. These clusters help us identify which cells may be more similar and/or dissimilar to one another!

when we plot the gene expression data for more cells, we notice that cell 3 (red dot) clusters next to these other cells from the sample

		Gene α	Gene β	Gene δ	repeat for thousands of genes...
●	# of mRNA transcripts found in Cell 1	2	4	1	•••
●	# of mRNA transcripts found in Cell 2	1	5	5	•••
●	# of mRNA transcripts found in Cell 3	5	1	3	•••
●	repeat count for thousands of cells...	•••	•••	•••	•••

In addition to collecting data on gene expression for thousands of cells, scientists will add another layer of complexity by measuring the gene expression of these thousands of cells for THOUSANDS of genes. A table displaying this data would have thousands of rows and thousands of columns. Since the graph would now have much more than just 3 dimensions, we will need a special type of tool to graphically represent this data in a way that humans can visualize.



Each dot represents a single nucleus isolated from a single brain cell

Identify clusters in the data--these clusters represent cells that are more like each other than they are like any other cells

UMAP

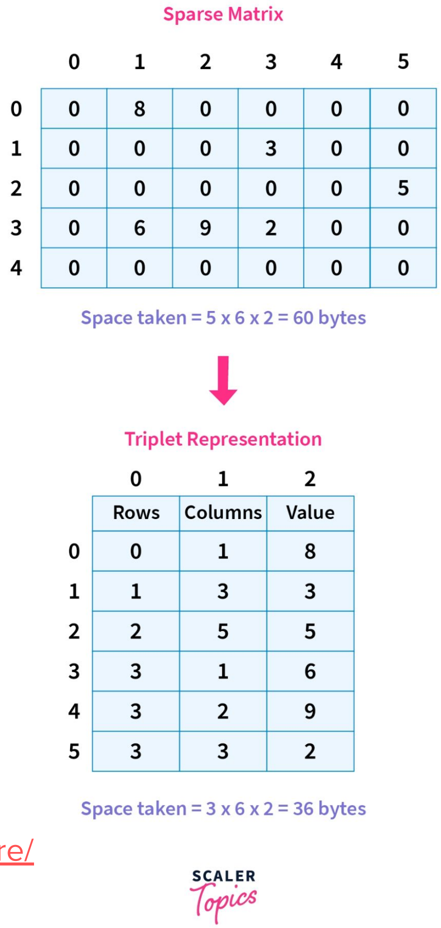
In order to plot this many-dimensional graph in a way humans can visualize, we use a dimensionality reduction tool, such as a UMAP, to plot it in a 2D space. Dimensionality reduction is a technique that helps represent many-dimentional data in just two or three dimensions.

Before running the UMAP, we need to change the way our array is represented

Sparse matrices (or arrays) contain many zeros.

We can change the representation of these to use less memory and require less computation time.

We'll use **`scipy.sparse`** to represent our data as compressed sparse row (CSR)



Sparse Matrix

	0	1	2	3	4	5
0	0	8	0	0	0	0
1	0	0	0	3	0	0
2	0	0	0	0	0	5
3	0	6	9	2	0	0
4	0	0	0	0	0	0

Space taken = 5 x 6 x 2 = 60 bytes

Triplet Representation

	0	1	2
	Rows	Columns	Value
0	0	1	8
1	1	3	3
2	2	5	5
3	3	1	6
4	3	2	9
5	3	3	2

Space taken = 3 x 6 x 2 = 36 bytes

Image from https://www.scaler.com/topics/data-structures/sparse-matrix-in-data-structure/
See also https://en.wikipedia.org/wiki/Sparse_matrix and
https://www.geeksforgeeks.org/what-is-meant-by-sparse-array/

About the author of the Patch-seq notebook: Nuo Xu! https://triplab.org/

Resources

Unraveling the Complexity of the Mammalian Brain - Allen Institute

UMAP Zoo

How UMAP Works

Log-transformation and its implications for data analysis - PMC