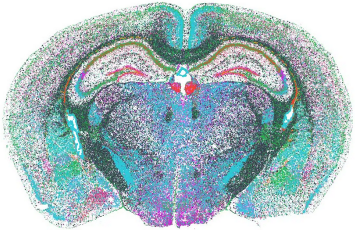
Statistics and data analysis

Dave Zhao
Department of Statistics
Carl R. Woese Institute for Genomic Biology

Statistics bridges the gap.

“Imagine a flashy spaceship lands in your backyard. The door opens and you are invited to investigate everything to see what you can learn. The technology is clearly millions of years beyond what we can make.

This is biology. That’s why it fascinates me so much. Life is billions of years old technology, and we can explore it to see ***how it works***!”



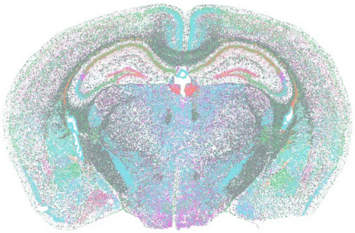
berthub.eu, Vizgen

There is a disconnect between biological question and data.

> “Imagine a flashy spaceship lands in your backyard. The door opens and you are invited to investigate everything to see what you can learn. The technology is clearly millions of years beyond what we can make.
>
> This is biology. That’s why it fascinates me so much. Life is billions of years old technology, and we can explore it to see ***how it works***!”

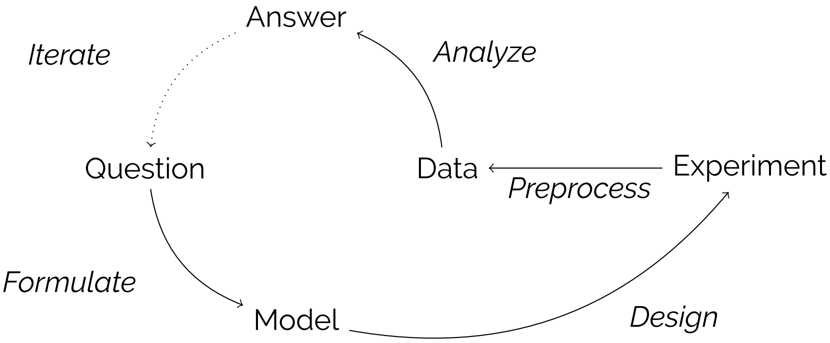
Statistical concepts help formulate scientific questions as mathematical questions.

Statistical techniques help answer mathematical questions using data.



berthub.eu, Vizgen

Conceptual framework for connecting biological question to data.



Answer

Iterate

Analyze

Question

Data

Experiment

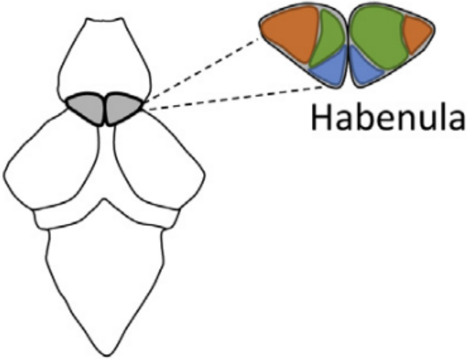
Preprocess

Formulate

Model

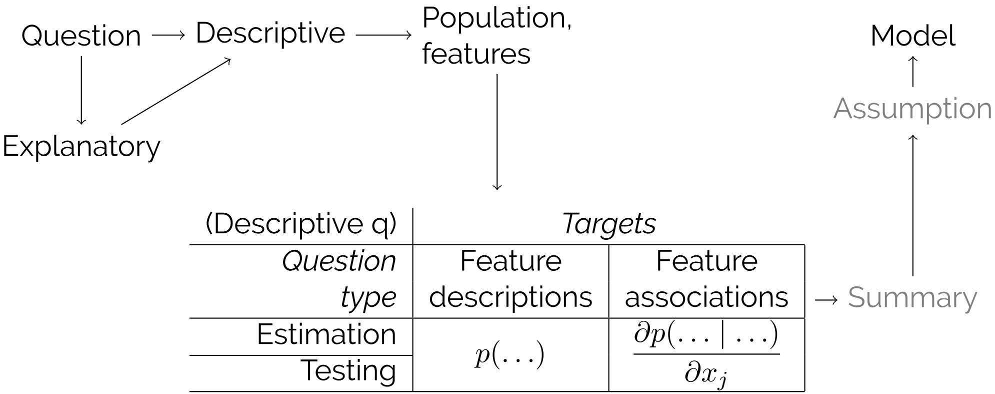
Design

How do cells in the larval zebrafish habenula coordinate their functions?



Pandey et al. (2018)

Formulate descriptive question.



(Descriptive q)	*Targets*	
Question type	Feature descriptions	Feature associations
Estimation	$p(\ldots)$	$\dfrac{\partial p(\ldots \mid \ldots)}{\partial x_j}$
Testing		

Explanatory:
How do cells in the larval zebrafish habenula coordinate their functions?

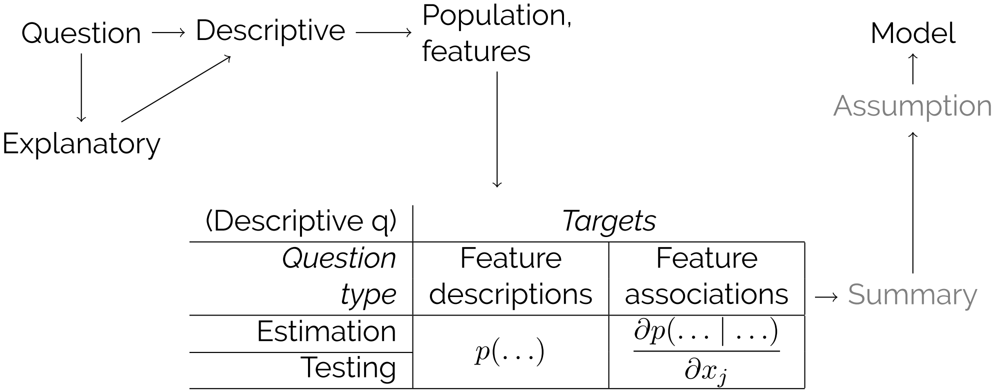
↓

Descriptive:

1. What cell types are in the habenula?
2. What are marker genes for each cell type?

Pandey et al. (2018)

Identify population and features.



(Descriptive q)	*Targets*	
Question type	Feature descriptions	Feature associations
Estimation	$p(\ldots)$	$\frac{\partial p(\ldots \mid \ldots)}{\partial x_j}$
Testing		

Descriptive question:
What cell types are in the habenula?

Population: habenula cells

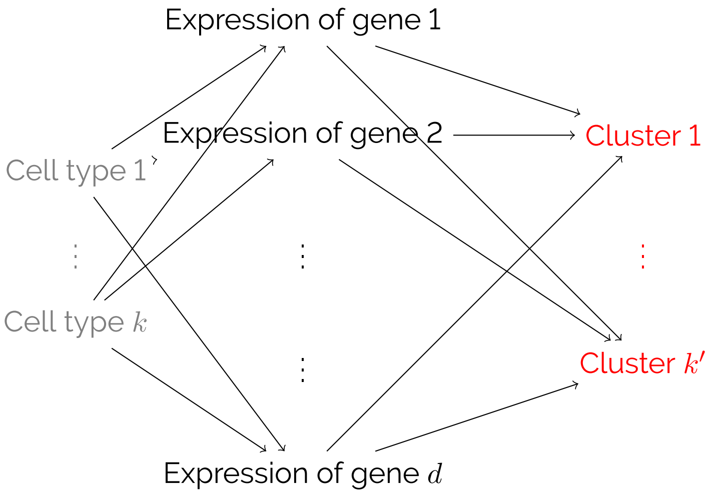
- Things you want to understand
- Things you want to sample
- "X% of the population ..."

Features: cell type

- Attributes you want to measure on each element of your population
- Can be directly observable or latent

Pandey et al. (2018)

Proxies for latent features need to be defined.



Descriptive question:
What are marker genes for each cell type?

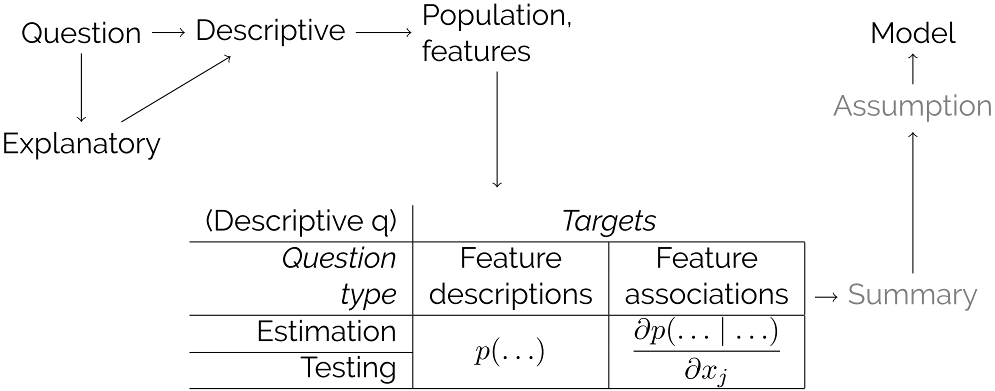
Population: habenula cells

Features: cell type and expression of gene $j = 1, \ldots, d$

- Attributes you want to measure on each element of your population
- Can be directly observable or latent

Pandey et al. (2018)

Identify target and type of question.



Question

Descriptive

Population, features

Explanatory

Model

Assumption

Summary

(Descriptive q)	*Targets*	
Question type	Feature descriptions	Feature associations
Estimation	$p(\ldots)$	$\frac{\partial p(\ldots \mid \ldots)}{\partial x_j}$
Testing		

Running example.

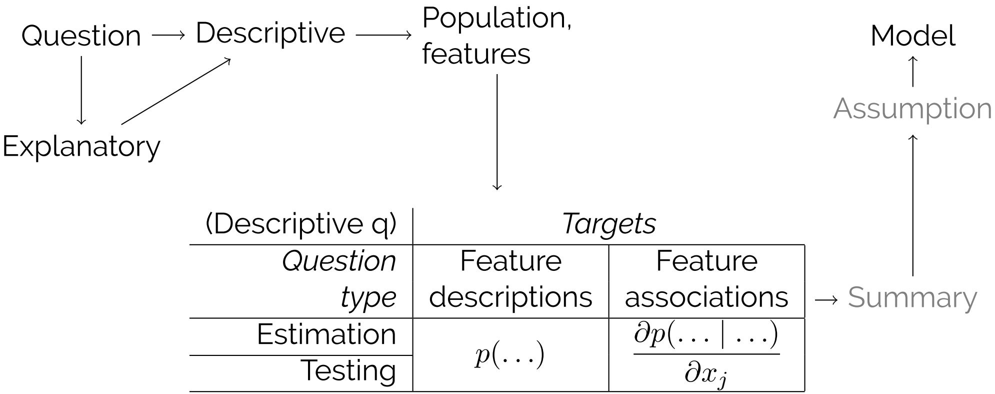
Descriptive questions:

1. What cell types are in the habenula?
2. What are marker genes for each cell type?

(Descriptive q)	*Targets*	
Question type	Feature descriptions	Feature associations
Estimation	1	
Testing		2

Pandey et al. (2018)

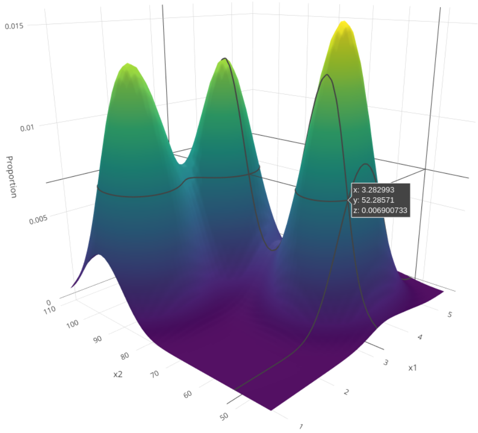
Identify mathematical quantities of interest.

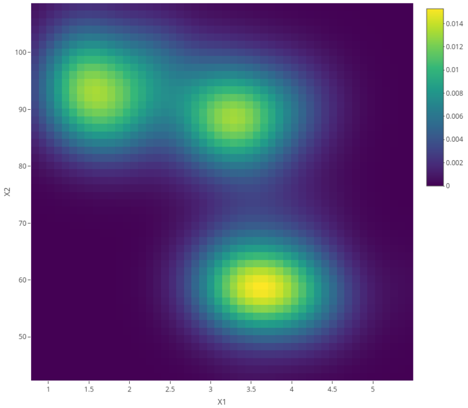


(Descriptive q)	*Targets*	
Question type	Feature descriptions	Feature associations
Estimation	$p(\ldots)$	$\dfrac{\partial p(\ldots \mid \ldots)}{\partial x_j}$
Testing		

Interesting mathematical quantities are functions of the population distribution function.

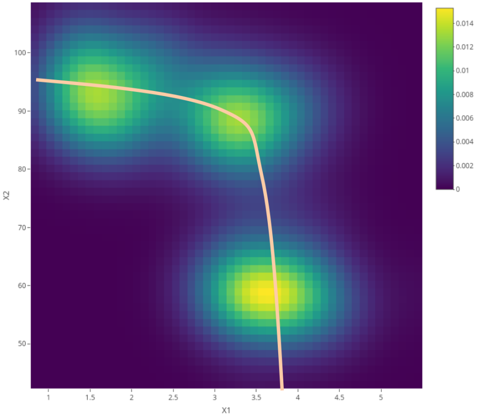
Population distribution function: $p(x_1, \ldots, x_d) = P(X_1 = x_1, \ldots, X_d = x_d)$



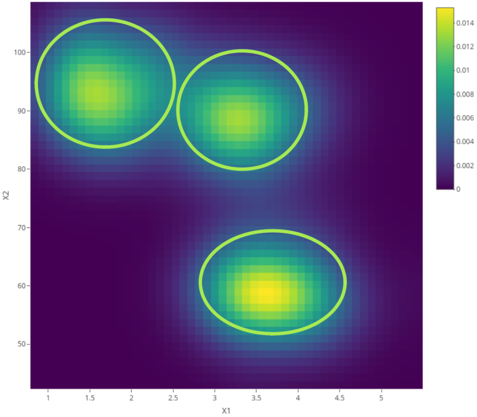


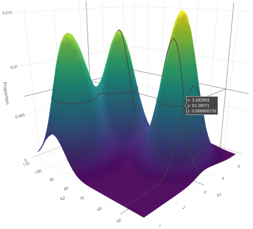
Canonical mathematical quantities of interest.

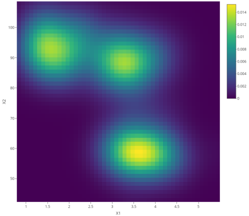
Latent numerical feature



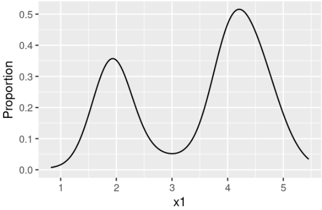
Latent categorical feature

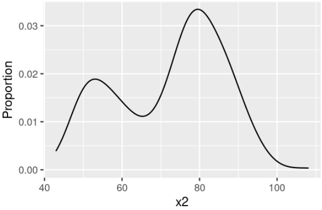




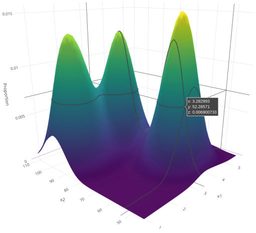


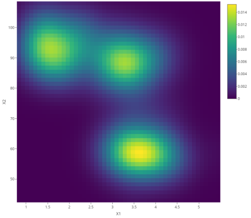
Marginal distributions: distributions of a subset of features



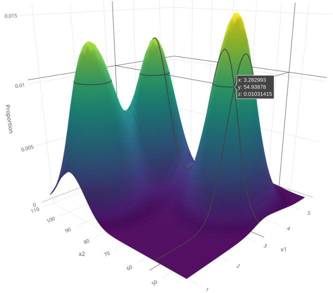


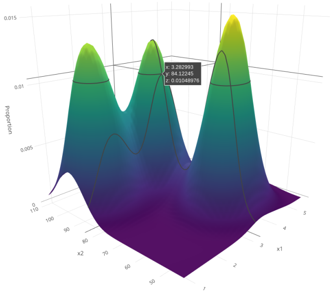
Canonical mathematical quantities of interest.





Conditional distributions: distributions in subpopulations





L: $P(X_1 = x_1 \mid X_2 = 54.94)$

R: $P(X_1 = x_1 \mid X_2 = 84.12)$

Canonical mathematical formulations of descriptive questions.

(Descriptive q)	*Targets*	
Question type	Feature descriptions	Feature associations
Estimation	What is the dist?	How does a conditional dist change?
Testing	Is the dist …?	Does a conditional dist change?

Descriptive questions:

1. What cell types are in the habenula?
2. What are marker genes for each cell type?

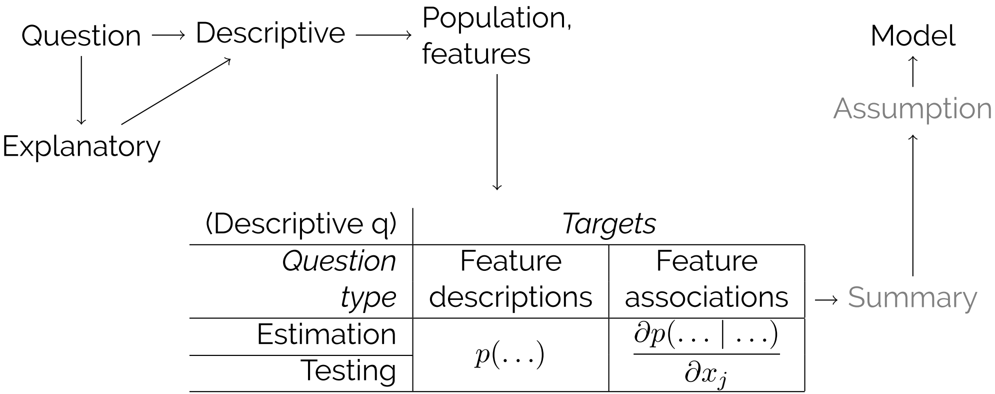
Let Z_i be the (proxy constructed for the latent categorical) cell type of cell i and G_{ij} be the (directly observed) expression of gene j in cell i.

Mathematical models of questions:

1. What is $P(Z_i = z)$?
2. Does $P(G_{ij} = g \mid Z_i = z)$ change as z changes, for each gene j?

Pandey et al. (2018)

Simplify.



Question → Descriptive → Population, features

Question → Explanatory → Descriptive

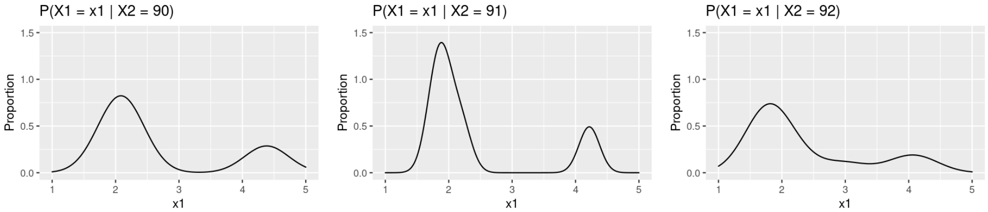
Summary → Assumption → Model

(Descriptive q)	*Targets*	
Question type	Feature descriptions	Feature associations
Estimation	$p(\ldots)$	$\dfrac{\partial p(\ldots \mid \ldots)}{\partial x_j}$
Testing		

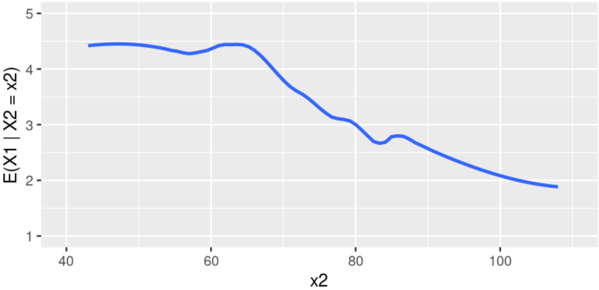
Association between X_1 and X_2

$\downarrow$

How does $P(X_1 = x_1 \mid X_2 = x_2)$ change as x_2 changes??

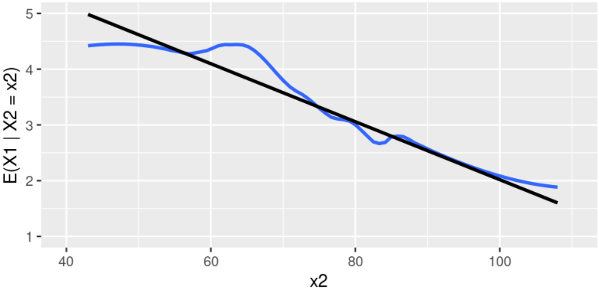


How does $E(X_1 \mid X_2 = x_2)$ change as x_2 changes?



Simplify using assumptions.

Pretend that $E(X_1 \mid X_2 = x_2) = \beta_0 + \beta_1 x_2$. There are β_0 and β_1 that make this assumption closest to the truth. What is that β_1?



Descriptive question:
What are marker genes for each cell type?

Let Z_i be the (proxy constructed for the latent categorical) cell type of cell i and G_{ij} be the (directly observed) expression of gene j in cell i.

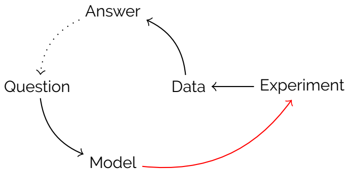
For the kth cell type, let $\tilde{Z}_{ik}$ = 1 if $Z_i = k$ and 0 if $Z_i \neq k$.

Mathematical model of question:

- Pretend that
$$\log E(G_{ij} \mid \tilde{Z}_{ik} = z) = \beta_{0jk} + \beta_{1jk} z.$$
- Is $\beta_{1jk} = 0$ for each gene j and each cell type k?

Pandey et al. (2018)

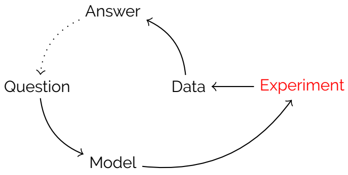
Design



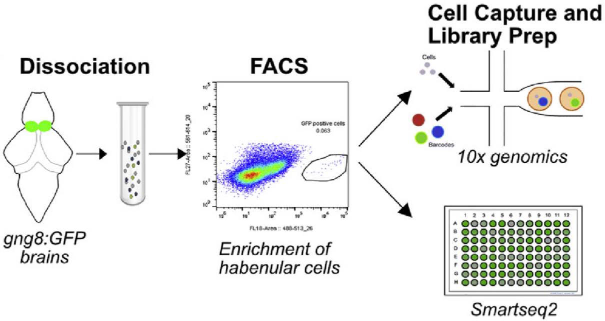
What cell types are in the habenula? What are marker genes for each cell type?

How many zebrafish, how to select zebrafish, which regions to dissect, at what times, etc.?

Experiment

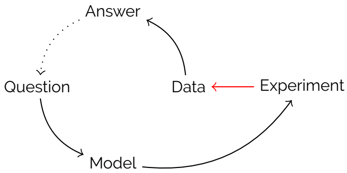


What cell types are in the habenula? What are marker genes for each cell type?

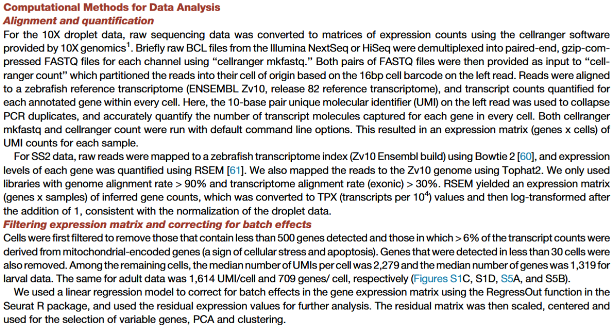


Pandey et al. (2018)

Preprocess



What cell types are in the habenula? What are marker genes for each cell type?



Computational Methods for Data Analysis

Alignment and quantification

For the 10X droplet data, raw sequencing data was converted to matrices of expression counts using the cellranger software provided by 10X genomics[1]. Briefly raw BCL files from the Illumina NextSeq or HiSeq were demultiplexed into paired-end, gzip-compressed FASTQ files for each channel using "cellranger mkfastq." Both pairs of FASTQ files were then provided as input to "cellranger count" which partitioned the reads into their cell of origin based on the 16bp cell barcode on the left read. Reads were aligned to a zebrafish reference transcriptome (ENSEMBL Zv10, release 82 reference transcriptome), and transcript counts quantified for each annotated gene within every cell. Here, the 10-base pair unique molecular identifier (UMI) on the left read was used to collapse PCR duplicates, and accurately quantify the number of transcript molecules captured for each gene in every cell. Both cellranger mkfastq and cellranger count were run with default command line options. This resulted in an expression matrix (genes x cells) of UMI counts for each sample.

For SS2 data, raw reads were mapped to a zebrafish transcriptome index (Zv10 Ensembl build) using Bowtie 2 [60], and expression levels of each gene was quantified using RSEM [61]. We also mapped the reads to the Zv10 genome using Tophat2. We only used libraries with genome alignment rate > 90% and transcriptome alignment rate (exonic) > 30%. RSEM yielded an expression matrix (genes x samples) of inferred gene counts, which was converted to TPX (transcripts per 10^4) values and then log-transformed after the addition of 1, consistent with the normalization of the droplet data.

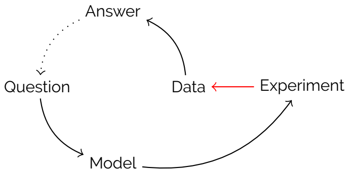
Filtering expression matrix and correcting for batch effects

Cells were first filtered to remove those that contain less than 500 genes detected and those in which > 6% of the transcript counts were derived from mitochondrial-encoded genes (a sign of cellular stress and apoptosis). Genes that were detected in less than 30 cells were also removed. Among the remaining cells, the median number of UMIs per cell was 2,279 and the median number of genes was 1,319 for larval data. The same for adult data was 1,614 UMI/cell and 709 genes/ cell, respectively (Figures S1C, S1D, S5A, and S5B).

We used a linear regression model to correct for batch effects in the gene expression matrix using the RegressOut function in the Seurat R package, and used the residual expression values for further analysis. The residual matrix was then scaled, centered and used for the selection of variable genes, PCA and clustering.

Pandey et al. (2018)

Preprocess



Answer

Question

Data

Experiment

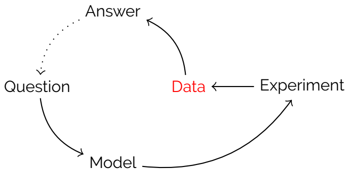
Model

What cell types are in the habenula? What are marker genes for each cell type?

```
file = "GSM2818521_larva_counts_matrix.txt"
larval = read.table(file, header = TRUE)

library(Seurat)
set.seed(1)

obj = CreateSeuratObject(counts = larval,
                         min.cells = 30,
                         min.features = 500)
obj[["percent.mt"]] = PercentageFeatureSet(obj,
                                           pattern = "^MT-")

obj = NormalizeData(obj)
obj = FindVariableFeatures(obj)
obj = ScaleData(obj, vars.to.regress = "percent.mt")
```

Pandey et al. (2018)

Data



Answer

Question

Data

Experiment

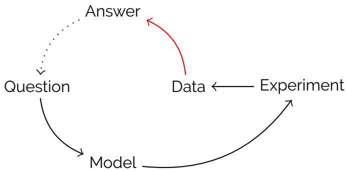
Model

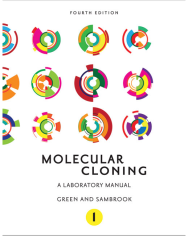
What cell types are in the habenula? What are marker genes for each cell type?

```
View(GetAssayData(obj, slot = "scale.data"))
```

Cell	Gene 1	…	Gene d
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

Analyze



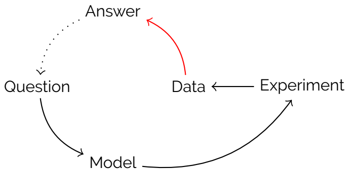


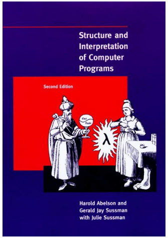
What cell types are in the habenula? What are marker genes for each cell type?

Analysis “kits” implement statistical techniques.

“Since the last edition of Molecular Cloning, there has been a relentless and continuing proliferation of commercial “kits,” which is both a blessing and a curse. On the one hand, kits offer tremendous convenience, particularly for procedures that are not routinely used in an individual laboratory. On the other hand, kits can often be too convenient, enabling users to perform procedures without understanding the underlying principles of the method.”

Analyze



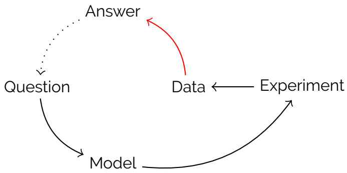

Structure and Interpretation of Computer Programs
Second Edition
Harold Abelson and Gerald Jay Sussman with Julie Sussman

What cell types are in the habenula? What are marker genes for each cell type?

“The language in which you’ll spend most of your working life hasn’t been invented yet, so we can’t teach it to you. Instead we have to give you the skills you need to learn new languages as they appear.”

Brian Harvey
University of California, Berkeley

SICP, Why *Structure and Interpretation of Computer Programs* matters

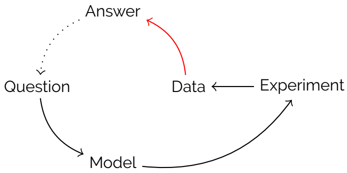


What cell types are in the habenula?

Use shared nearest neighbor clustering to construct cell type proxies Z_i.

```
## dimension reduction
obj = RunPCA(obj)

## clustering
obj = FindNeighbors(obj)
obj = FindClusters(obj,
                   resolution = 0.5)

## dimension reduction
obj = RunUMAP(obj, dims = 1:20)

## visualization
DimPlot(obj)
```

Analyze



What are marker genes for each cell type?

Use Wilcoxon tests to test the null hypothesis that $\beta_{1jk} = 0$ for each gene j and each cell type k.

```
markers = FindAllMarkers(obj)

## view top markers for cluster 0
head(markers[markers$cluster == 0,])

## view top markers for cluster 5
head(markers[markers$cluster == 5,])

## visualize markers
FeaturePlot(obj,
            features = c("G0S2",
                         "LRRTM1"))
```

Thank you

sdzhao@illinois.edu