Gene Mutation Web Application (MUTABLE)

Data Science Institute
Columbia University

**Authors: Junyi Yao, Yi Duan,
Yiquan Li, Zhe Hou, Zining Chen
Faculty Mentor: Yufeng Shen**

Data Science Capstone Project with Professor Yufeng Shen, Department of Systems Biology and Biomedical Informatics

Intuition and Project Scope

From a micro perspective, human contains enormous amounts of genes within cells of different type, and mutations of genes are highly related to human diseases. With the large dataset of cells and mutations, we build a website that aim to curate and visualize genomic data to help exploratory analysis of genetic mutations in human diseases.

- **Visualization of gene expression is important in genetic analysis. Disease relevant mutations are likely to have specific patterns in their location, therefore we use lollipop plots to visualize the mutations.**
- **Main learning objectives contain machine learning methods applied in genomics and genetics including data engineering, hierarchical clustering, statistical analysis, etc.**

Single Cell Data Analysis

- **Curated single cell expression data of *AnnData* type (Table 1) are provided containing info about tissue, gene, cell type, expression level, etc.**
- **Based on the data, single cell expression profile (Figure 4) for each gene are generated including two parts:**
  - **A table showing gene expression with respect to cell type, in which we display average expression level, percentage of positive expression, and number of cells**
  - **A bar plot (Figure 1) showing positive expression percentage of each gene in each cell type. For x-axis, the cell types are ordered according to hierarchical clustering (Figure 2) so that similar cell types are put together.**

Citation	tissue	location on md22	number of cells	number of cell types
La Manno, et al. 2016	Fetal Midbrain	/fisher/Projects/SingleCellData/LaManno_2016_midbrain	1,977	26
Zhong, et al. 2018	Prefrontal cortex	/fisher/Projects/SingleCellData/Zhong_2018_PrefrontalContex	8,686	6
Cao, et al. 2020	Cerebellum	/fisher/Projects/SingleCellData/Cao_2020	1,080,771	9
same as above	Cerebrum	same as above	1,711,950	9
same as above	Heart	same as above	96,622	16
same as above	Lung	same as above	214,387	13

Table 1. Main data sources of single cell data

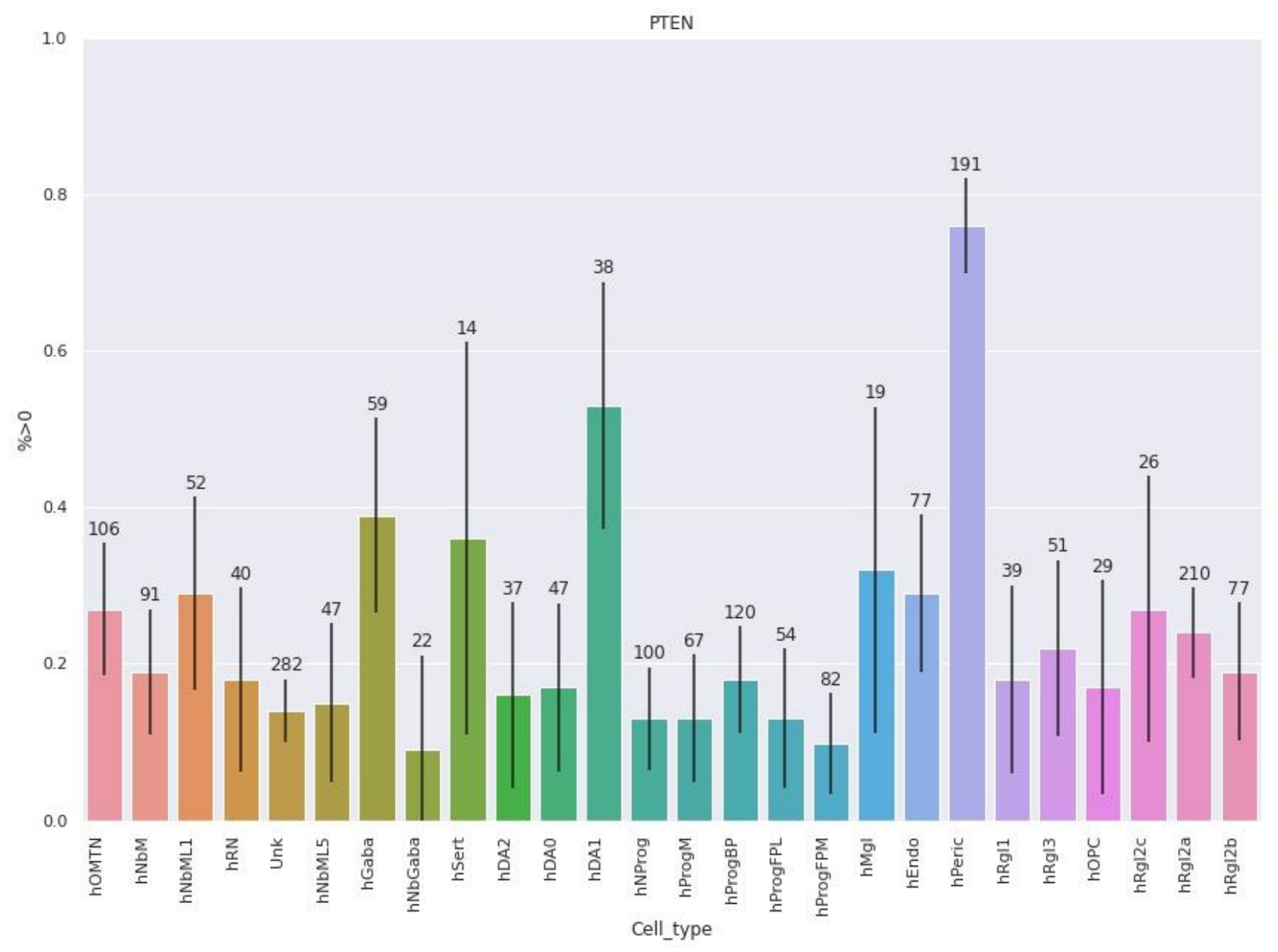


Figure 1. Bar Plot of Gene PTEN in Fetal Midbrain

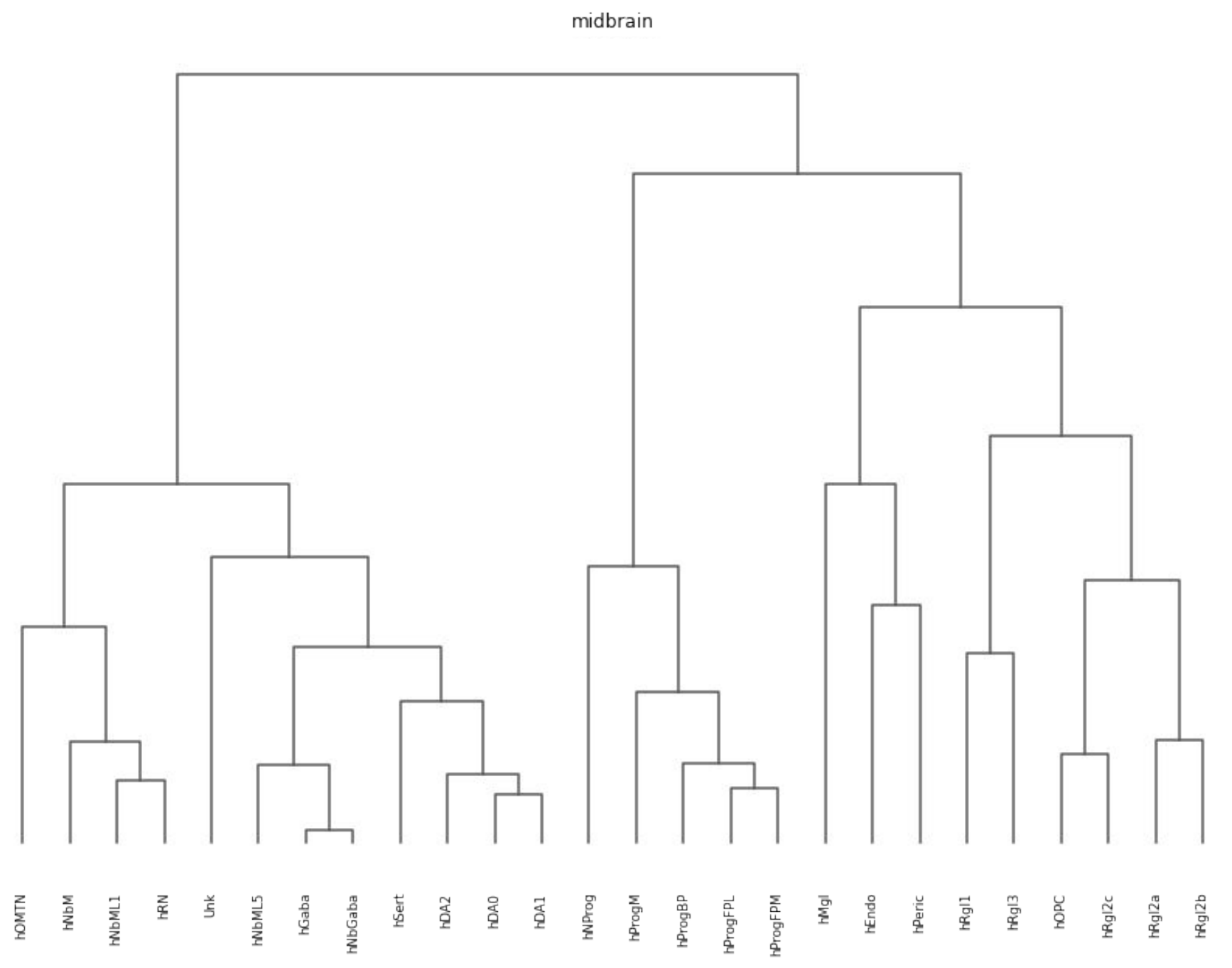


Figure 2. Dendrogram of Fetal Midbrain

Acknowledgments

Appreciation and thanks to the guidance and support of Professor Yufeng Shen for this project. Also grateful to the assistance of Jake Hagen and Lu Qiao from Shen Lab and Katie Jooyoung Kim.

Lollipop Plots for Gene Mutations

To visualize the mutations of each gene, lollipop plots are generated. Each mutation is presented as a lollipop plotted along the protein sequence. The stacked number of lollipops indicates the number of mutations on the same sequence. And the x-axis represents the protein sequence with *PFAM* protein blocks colored. Mutations are categorized into different groups based on impact (Figure 3). All mutation data and attributes are retrieved from de novo mutations (Table 2).

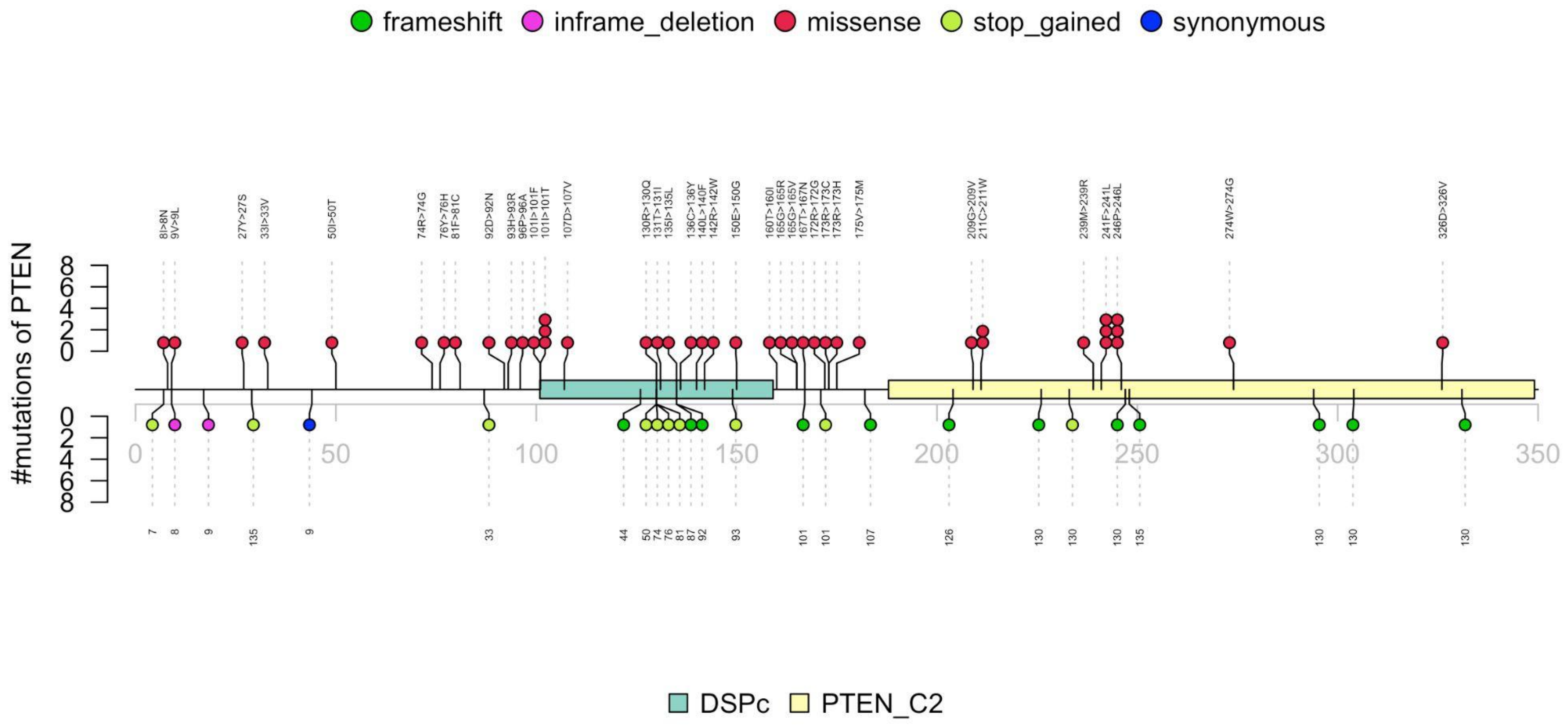


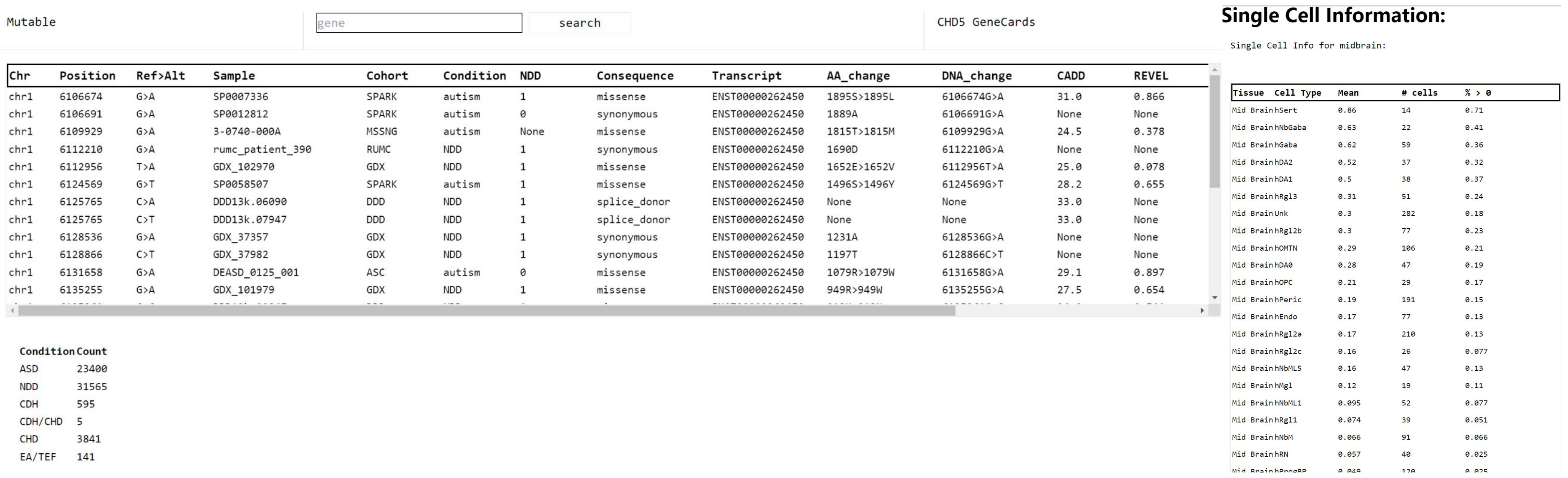
Figure 3. Lollipop Plot for Gene PTEN

Condition	Sample Size
Autism	23400
Neurodevelopmental Disorders	31565
Congenital Diaphragmatic hernia	595
Congenital Heart Disease	3841
Esophageal Atresia	141

Table 2. Data source for mutations

Website Layout (plots excluded)

The website supports searching a specific gene in different tissues for corresponding mutation data, single cell info, bar plot & lollipop plot (plots excluded in figures below).



Chr	Position	Ref>Alt	Sample	Cohort	Condition	NDD	Consequence	Transcript	AA_change	DNA_change	CADD	REVEL
chr1	6106674	G>A	SP0007336	SPARK	autism	1	missense	ENST00000262450	1895S>1895L	6106674G>A	31.0	0.866
chr1	6106691	G>A	SP0012812	SPARK	autism	0	synonymous	ENST00000262450	1889A	6106691G>A	None	None
chr1	6109929	G>A	3-0740-000A	MSSNG	autism	None	missense	ENST00000262450	1815T>1815M	6109929G>A	24.5	0.378
chr1	6112210	G>A	rumc_patient_390	RUMC	NDD	1	synonymous	ENST00000262450	1690D	6112210G>A	None	None
chr1	6112956	T>A	GDX_102970	GDX	NDD	1	missense	ENST00000262450	1652E>1652V	6112956T>A	25.0	0.078
chr1	6124569	G>T	SP0058507	SPARK	autism	1	missense	ENST00000262450	1496S>1496Y	6124569G>T	28.2	0.655
chr1	6125765	C>A	DDD13k.06090	DDD	NDD	1	splice_donor	ENST00000262450	None	None	33.0	None
chr1	6125765	C>T	DDD13k.07947	DDD	NDD	1	splice_donor	ENST00000262450	None	None	33.0	None
chr1	6128536	G>A	GDX_37357	GDX	NDD	1	synonymous	ENST00000262450	1231A	6128536G>A	None	None
chr1	6128866	C>T	GDX_37982	GDX	NDD	1	synonymous	ENST00000262450	1197T	6128866C>T	None	None
chr1	6131658	G>A	DEASD_0125_001	ASC	autism	0	missense	ENST00000262450	1079R>1079W	6131658G>A	29.1	0.897
chr1	6135255	G>A	GDX_101979	GDX	NDD	1	missense	ENST00000262450	949R>949W	6135255G>A	27.5	0.654

Condition	Count
ASD	23400
NDD	31565
CDH	595
CDH/CHD	5
CHD	3841
EA/TEF	141

Single Cell Information:

Single Cell Info for midbrain:

Tissue	Cell Type	Mean	# cells	% > 0
Mid Brain	hSert	0.86	14	0.71
Mid Brain	hNbGaba	0.63	22	0.41
Mid Brain	hGaba	0.62	59	0.36
Mid Brain	hDA2	0.52	37	0.32
Mid Brain	hDA1	0.5	38	0.37
Mid Brain	hRgl3	0.31	51	0.24
Mid Brain	Unk	0.3	282	0.18
Mid Brain	hRgl2b	0.3	77	0.23
Mid Brain	hOMTN	0.29	106	0.21
Mid Brain	hDA0	0.28	47	0.19
Mid Brain	hOPC	0.21	29	0.17
Mid Brain	hPeric	0.19	191	0.15
Mid Brain	hEndo	0.17	77	0.13
Mid Brain	hRgl2a	0.17	210	0.13
Mid Brain	hRgl2c	0.16	26	0.077
Mid Brain	hNbML5	0.16	47	0.13
Mid Brain	hMgl	0.12	19	0.11
Mid Brain	hNbML1	0.095	52	0.077
Mid Brain	hRgl1	0.074	39	0.051
Mid Brain	hNbM	0.066	91	0.066
Mid Brain	hRN	0.057	40	0.025
Mid Brain	hProgBP	0.049	120	0.025

Figure 4. Website Layout when Searching for Gene CHD5

Conclusion

Genes highly expressed in cell types relevant to a condition or diseases are more likely to be a risk gene. On the flip side, cell types in which known risk genes are highly and specifically expressed are likely to be relevant to a disease. We create a gene centric view in a web engine to support the relevant exploratory analysis of genomic data in human disease studies. The website is built reliable and sustainable for future research in Shen Lab at Columbia University.

References

The Tabula Muris Consortium., Overall coordination., Logistical coordination. et al. Nature 562, 367–372 (2018). https://doi.org/10.1038/s41586-018-0590-4