Incorporating single-cell RNA-sequencing data from the developing heart into the UMLS Knowledge Graph (UMLS-KG) to identify biomarkers of congenital heart disease

Shubha Vasisht (1); Ben Stear (1), MS; Taha Mohseni Ahooyi (1), PhD; Deanne Taylor (1,2), PhD.

1. Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia

2. Perelman School of Medicine, University of Pennsylvania

Children's Hospital of Philadelphia® Department of Biomedical & Health Informatics

Introduction

- Congenital heart disease (CHD): typically recognized as abnormalities in heart structure that occur within the developing human heart (Sun et al., 2015)
- Cell types of the developing heart can be classified using single-cell gene expression profiles (Asp, 2019) from single-cell RNA-sequencing data (scRNA-seq)
- Gene markers of developmental cell types in heart can be used for genetic studies of CHDs
- Importance of different gene markers can be informed by integrating different ontologies and knowledge repositories grouped together under UMLS Concepts (Bodenreider, 2004).
- Recently developed UMLS-KG is a property graph implementation which allows for fast, complex queries across a wide range of the biomedical terms (see Ben Stear's poster in session).

Dataset

- Single cell gene expression data was collected from second embryonic heart sample at 6.5-7 post-conception weeks (PCW) in the Asp 2019 study as it contained anatomical features from earlier and later time points utilized in paper
- Cellular spatial (anatomical) information was retained in order to identify 14 cell types through clustering and marker gene expression

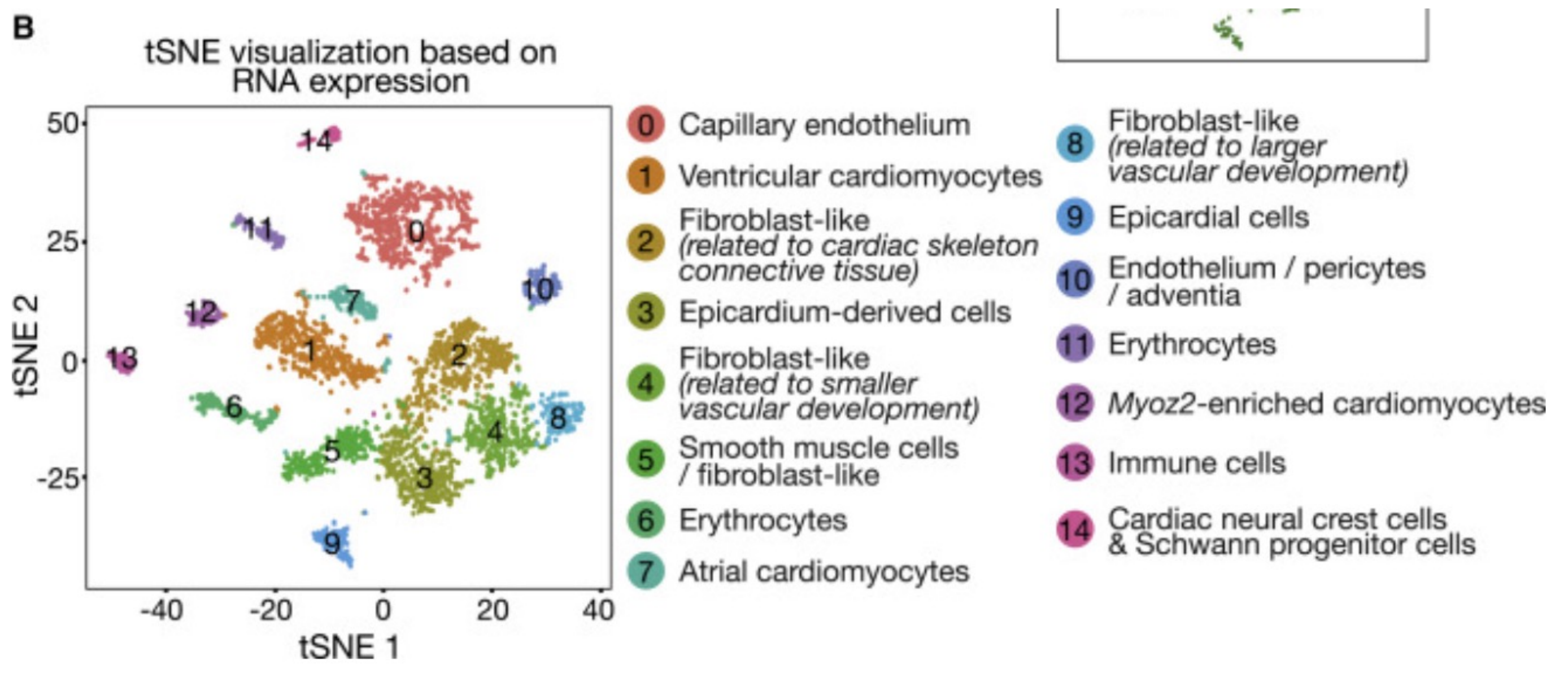


Figure 1. Figure to demonstrate the tSNE (dimensionality reduction) clustering by cell type (Asp et al. 2019)

Methods

- Seurat: package within R that aims to analyze single-cell RNA-seq data. Includes dimensionality, clustering, and differential expression analyses (Kharchenko, 2021).
- Clustered the developing heart cells by author defined cell type (Asp et al., 2019) to identify differentially expressed gene (DEG) markers (see figure 1) and other quantifiable differences (significant p-values along with the log2-fold-changes)
- Mapped 14 author-defined cell types to the Cell Ontology and UBERON for use in the UMLS-KG
- Once these biomarkers were identified for each cell type, the biomarker and cell type were used to create new Concept nodes that are uniquely identified by their relationships to the biomarker (gene) and cell type Concept nodes.
- Other quantifiable differences were used to identify the new Concept nodes

Results

- Nearly 3,000 biomarkers were identified across the 14 cell types and mapped to HGNC gene IDs
- These biomarkers by cell type are now available within an integrated KG of over 30 million Concepts representing 100s of ontologies and millions of experimental data points
- Queries of these connections are now being tested as a proof of concept for Gabriella Miller Kids First variant warehouse integration in order to identify DEGs in cell types that are associated with certain CHDs. (see Figure 4)

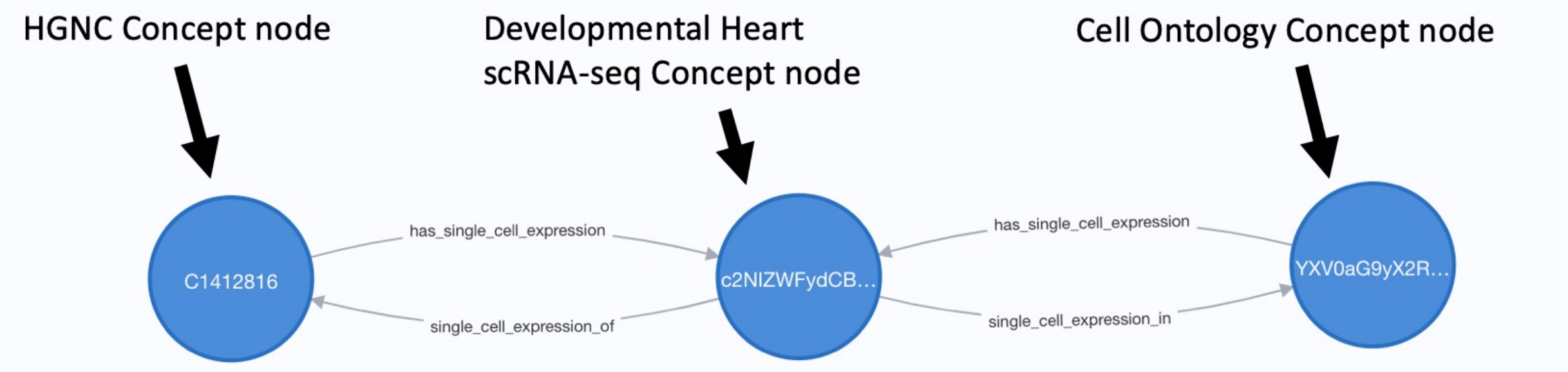
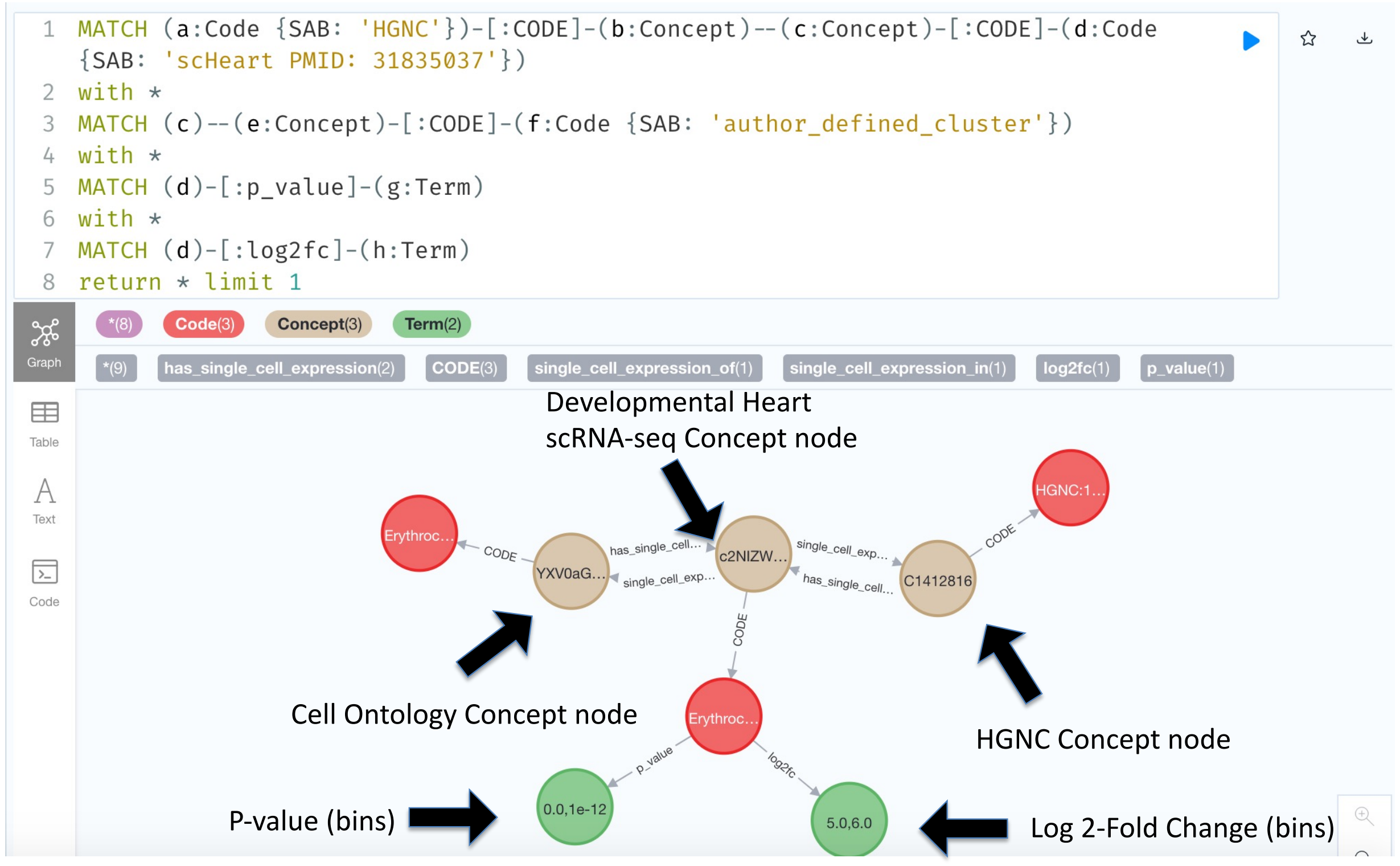

HGNC Concept node
Developmental Heart scRNA-seq Concept node
Cell Ontology Concept node
C1412816
c2NIZWFydCB...
YXV0aG9yX2R...
has_single_cell_expression
single_cell_expression_of
has_single_cell_expression
single_cell_expression_in

Figure 2. General connection between HGNC gene node to single cell Asp heart Concept node too the Cell ontology concept node (modeled in Neo4J environment)



```
MATCH (a:Code {SAB: 'HGNC'})-[:CODE]-(b:Concept)--(c:Concept)-[:CODE]-(d:Code
{SAB: 'scHeart PMID: 31835037'})
with *
MATCH (c)--(e:Concept)-[:CODE]-(f:Code {SAB: 'author_defined_cluster'})
with *
MATCH (d)-[:p_value]-(g:Term)
with *
MATCH (d)-[:log2fc]-(h:Term)
return * limit 1
```

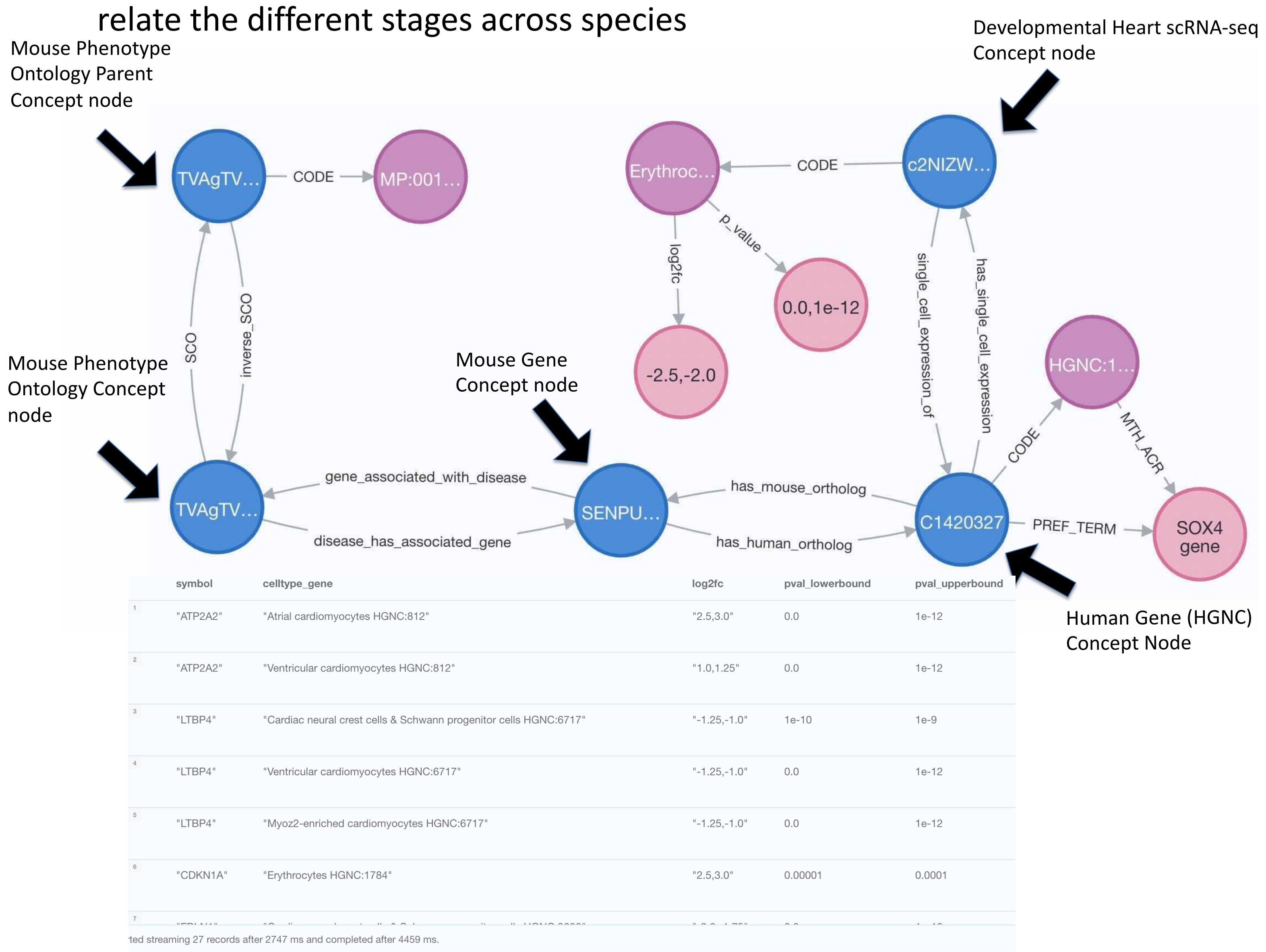
Figure 3. Example of Cypher query in Neo4J environment to produce log2fc and p-value of HGNC gene in cell type of Asp data; yields a small set from much larger UMLS-KG

Conclusions

- By integrating the developmental heart scRNAseq data into the UMLS-KG we can now query the data in much more biologically meaningful ways.
- Because the gene and cell type Concept nodes both have relationships with the other ontologies (including mouse phenotype) in UMLS, we can design complex queries between phenotype, cell type, and gene
- For example, we can take advantage of the gene nodes' relationships to other ontologies by filtering the data based on gene location, chromosome, transcript type, expression in tissue (UBERON and GTEx) or biological function (GO).
- We are able to query the UMLS-KG for DEGs that could be associated with CHDs as defined by the Kids First Data Resource Center in order to identify potential targets for further research (see Figure 4)
- Now able to leverage the rich, semantic, biomedical environment of UMLS against our scRNA-seq data in order answer more meaningful questions related to CHDs in the developing heart.

Next Steps

- Machine learning approaches to find connections between variants from Gabriella Miller Kids First variant warehouse and possible gene markers (see Ben Stear's poster in session)
- Looking to include Theiler Stages: developmental stages of mouse embryos
- Also, Carnegie Stages: developmental stages of human embryos
- Will continue to look for ways to connect within the UMLS-KG in order to easily relate the different stages across species


Mouse Phenotype Ontology Parent Concept node
Developmental Heart scRNA-seq Concept node
Mouse Phenotype Ontology Concept node
Mouse Gene Concept node
Human Gene (HGNC) Concept Node
TVAgTV...
MP:001...
Erythroc...
c2NIZW...
0.0,1e-12
-2.5,-2.0
HGNC:1...
TVAgTV...
SENPU...
C1420327
SOX4 gene
CODE
SCO
inverse_SCO
log2fc
p_value
single_cell_expression_of
has_single_cell_expression
CODE
MTH_ACR
gene_associated_with_disease
disease_has_associated_gene
has_mouse_ortholog
has_human_ortholog
PREF_TERM

	symbol	celltype_gene	log2fc	pval_lowerbound	pval_upperbound
1	"ATP2A2"	"Atrial cardiomyocytes HGNC:812"	"2.5,3.0"	0.0	1e-12
2	"ATP2A2"	"Ventricular cardiomyocytes HGNC:812"	"1.0,1.25"	0.0	1e-12
3	"LTBP4"	"Cardiac neural crest cells & Schwann progenitor cells HGNC:6717"	"-1.25,-1.0"	1e-10	1e-9
4	"LTBP4"	"Ventricular cardiomyocytes HGNC:6717"	"-1.25,-1.0"	0.0	1e-12
5	"LTBP4"	"Myoz2-enriched cardiomyocytes HGNC:6717"	"-1.25,-1.0"	0.0	1e-12
6	"CDKN1A"	"Erythrocytes HGNC:1784"	"2.5,3.0"	0.00001	0.0001
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

ted streaming 27 records after 2747 ms and completed after 4459 ms.

Figure 4. Results of Query for HGNC genes that are biomarkers for Atrial Septal Defect, a CHD that is characterized by a hole between the left and right atria. Query yields **27 cell type/ gene pairs with significant p-values** that can be further studied

Acknowledgements

- Thank you to Dr. Deanne Taylor for the guidance and mentorship during the project
- Thank you to Ben Stear and Taha Mohseni Ahooyi for the expertise and the support
- Thank you CHOP for the use of resources and support during this project

References

Olivier Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research, Volume 32, Issue suppl_1, 1 January 2004, Pages D267–D270

Asp M. A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. Cell. 2019 Dec 12;179(7):1647-1660.e19. doi: 10.1016/j.cell.2019.11.025. PMID: 31835037.

Kharchenko, P.V. The triumphs and limitations of computational methods for scRNA-seq. Nat Methods 18, 723–732 (2021). https://doi.org/10.1038/s41592-021-01171-x

Sun R, Liu M, Lu L, Zheng Y, Zhang P. Congenital Heart Disease: Causes, Diagnosis, Symptoms, and Treatments. Cell Biochem Biophys. 2015 Jul;72(3):857-60. doi: 10.1007/s12013-015-0551-6. PMID: 25638345.