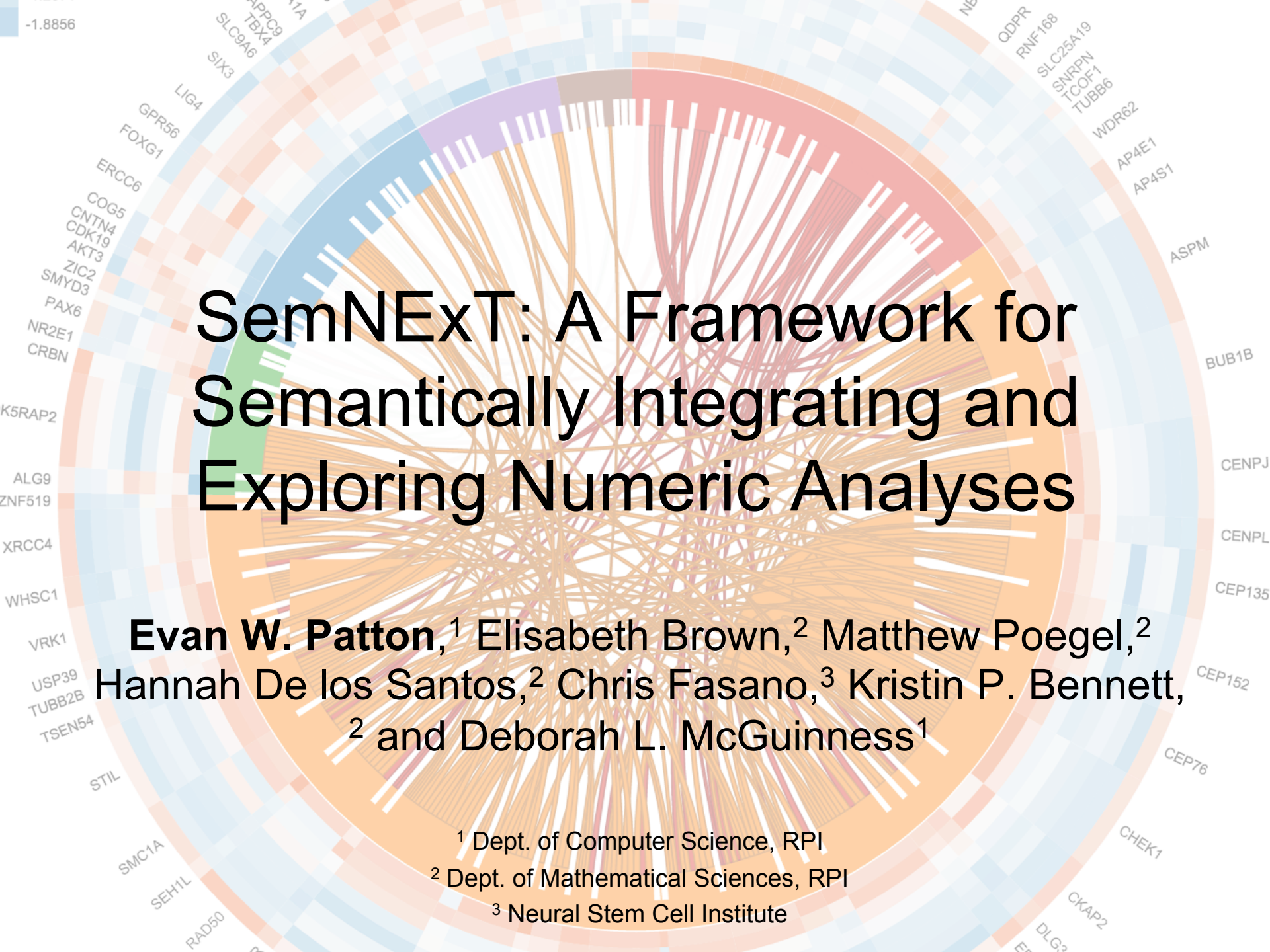


SemNExT: A Framework for Semantically Integrating and Exploring Numeric Analyses

Evan W. Patton,[1] Elisabeth Brown,[2] Matthew Poegel,[2] Hannah De los Santos,[2] Chris Fasano,[3] Kristin P. Bennett,[2] and Deborah L. McGuinness[1]

[1] Dept. of Computer Science, RPI

[2] Dept. of Mathematical Sciences, RPI

[3] Neural Stem Cell Institute

Overview

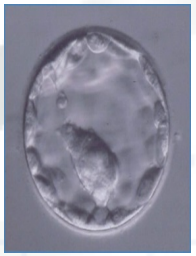
- Motivation & Domain
- Example
- Architecture
  - Datasets
  - Statistics
  - Linking
  - Provenance
- Conclusions

Motivation

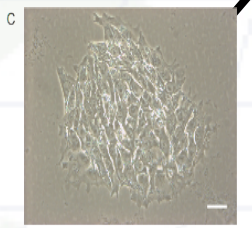
- Neural Stem Cell Institute collects significant amounts of data on the state of brain development in the form of *gene reads*
- Questions:
  - How to *cluster genes* together based on *activity over time*?
  - How to associate those gene clusters with *diseases of interest/gene ontology annotations*?
  - Are there other relationships that may not be obvious from the underlying data alone?

Understanding origins of disease in human cerebral cortex development

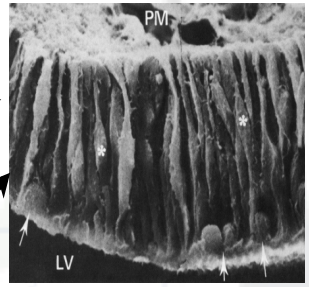
Fertilization



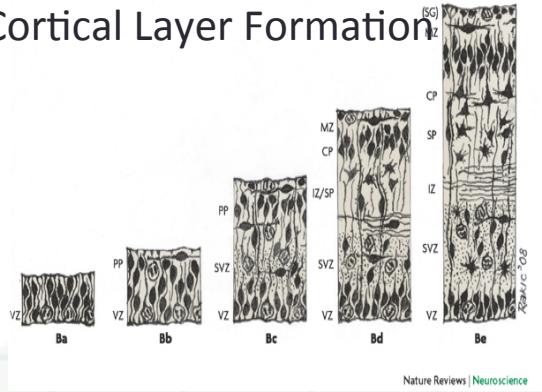
Human Embryonic Stem Cells



Neural Differentiation



Cortical Layer Formation



- Create molecular signature of normal cortical development *in a dish*
- Understand stages of development and when layers form
- Analyze mutated genes associated with disease to understand developmental origins
- Understand which pathways are associated with stages and diseases.

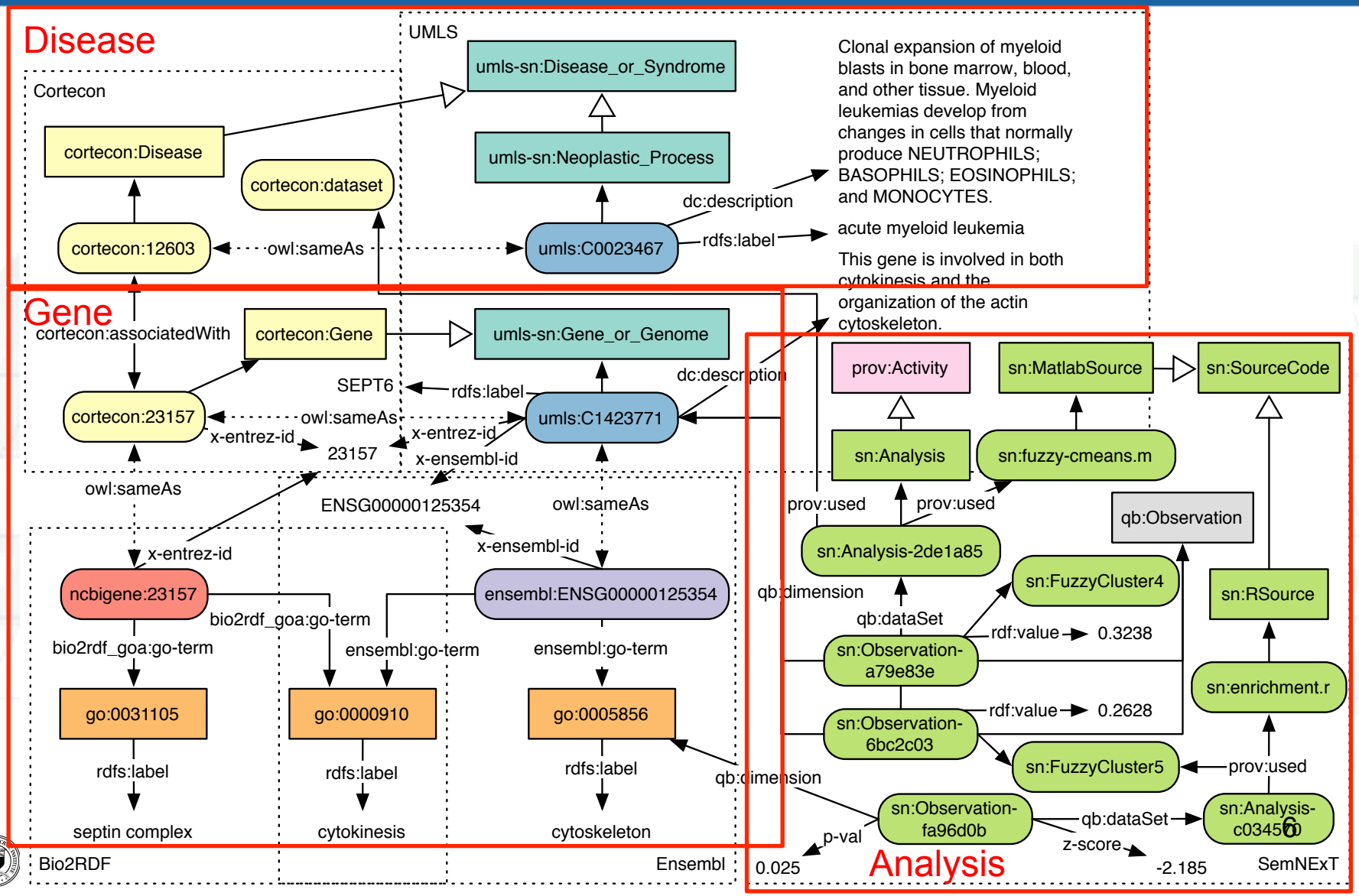
Joint work on CORTECON Data with Dr. Chris Fasano and Sally Temple at Neural Stem Cell Institute

Domain

- Brain development data, primarily focused on ***RNA-seq counts*** from RNA sequencing
- ***Genes*** encode ***proteins*** and ***enzymes***, which perform various biological functions
- Proteins often act as part of ***pathways*** and ***interact*** with one another, may cause certain ***diseases***, and be affected by ***compounds***, such as ***drugs***

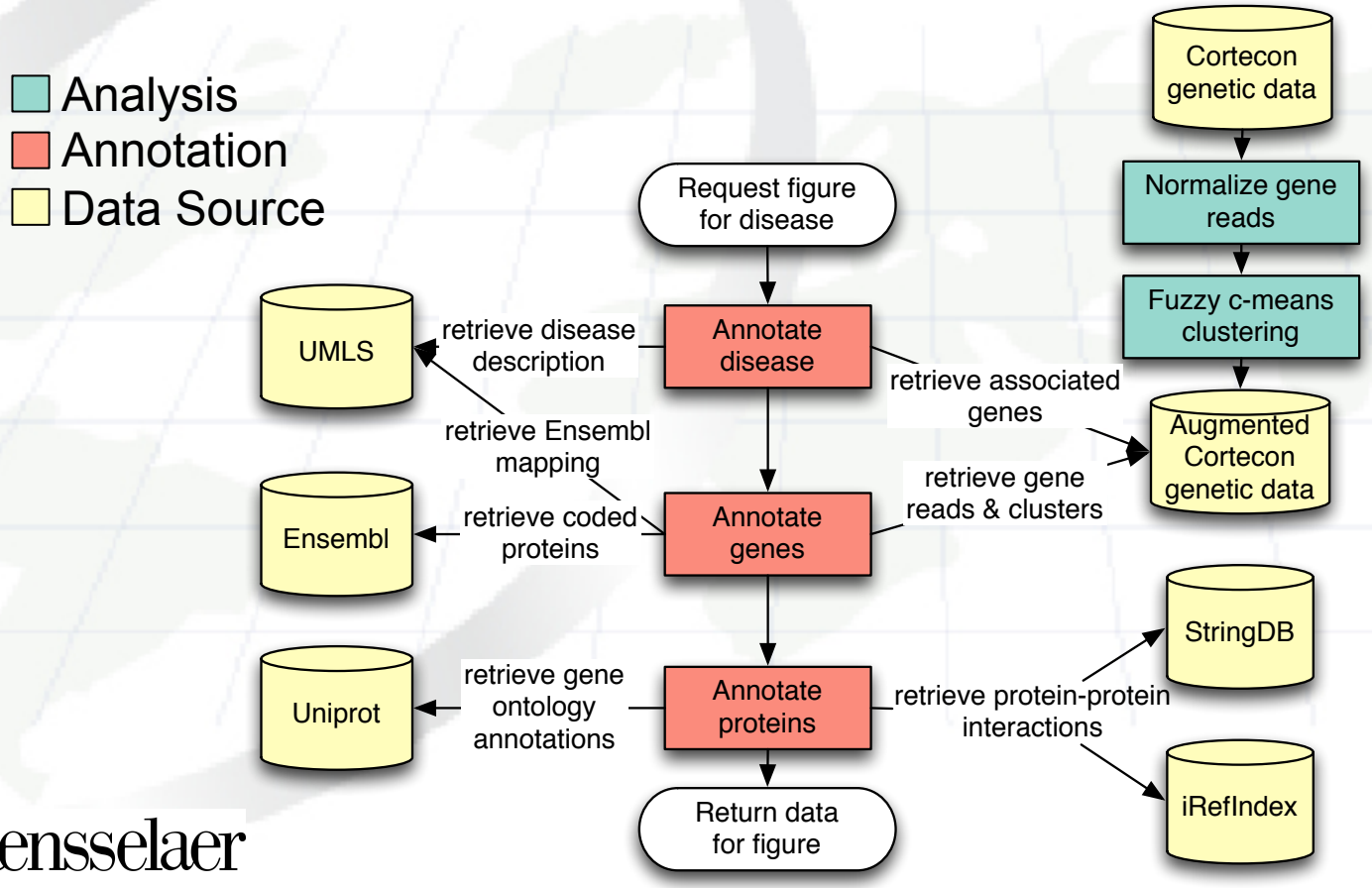
For more info, see J. van de Leemput el al. “CORTECON: a temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells.” *Neuron*: 83.1 (2014): 51-68.

Example


Disease
UMLS
Cortecon
umls-sn:Disease_or_Syndrome
cortecon:Disease
umls-sn:Neoplastic_Process
cortecon:dataset
dc:description
Clonal expansion of myeloid blasts in bone marrow, blood, and other tissue. Myeloid leukemias develop from changes in cells that normally produce NEUTROPHILS; BASOPHILS; EOSINOPHILS; and MONOCYTES.
cortecon:12603
owl:sameAs
umls:C0023467
rdfs:label
acute myeloid leukemia
This gene is involved in both cytokinesis and the organization of the actin cytoskeleton.
Gene
cortecon:associatedWith
cortecon:Gene
umls-sn:Gene_or_Genome
dc:description
SEPT6
rdfs:label
cortecon:23157
owl:sameAs
umls:C1423771
x-entrez-id
23157
x-entrez-id
x-ensembl-id
owl:sameAs
ENSG00000125354
owl:sameAs
x-entrez-id
x-ensembl-id
ncbigene:23157
ensembl:ENSG00000125354
bio2rdf_goa:go-term
bio2rdf_goa:go-term
ensembl:go-term
ensembl:go-term
go:0031105
go:0000910
go:0005856
rdfs:label
rdfs:label
rdfs:label
septin complex
cytokinesis
cytoskeleton
Bio2RDF
Ensembl
prov:Activity
sn:MatlabSource
sn:SourceCode
sn:Analysis
sn:fuzzy-cmeans.m
prov:used
prov:used
qb:Observation
sn:Analysis-2de1a85
qb:dimension
sn:FuzzyCluster4
sn:RSource
qb:dataSet
rdf:value
0.3238
sn:Observation-a79e83e
sn:enrichment.r
rdf:value
0.2628
sn:Observation-6bc2c03
sn:FuzzyCluster5
prov:used
qb:dimension
sn:Observation-fa96d0b
qb:dataSet
sn:Analysis-c03450
p-val
z-score
0.025
-2.185
Analysis
SemNExT

Architecture

SemNeXT annotates ***entities*** from ***data sources***, performs ***analyses***, and visualizes the results as ***Chord and Heat Map (CHeM)*** diagrams.



Datasets

Relational Databases

1. NSCI Cortecon
   - Genes, diseases, RNAseq read data
2. KEGG
   - Genes, pathways, proteins
3. StringDB
   - Protein interactions
4. Ensembl
   - Genes, proteins, gene ontology annotations
5. Unified Medical Language System
   - Genes, diseases

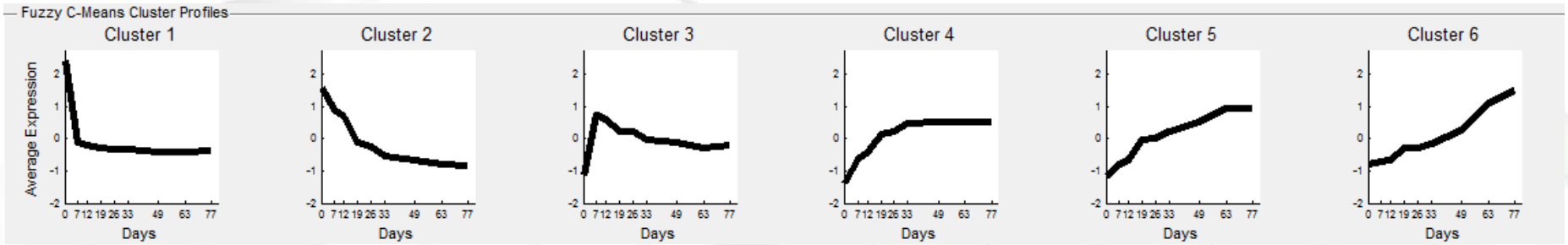
RDF Quad Stores

6. Bio2RDF
7. ReDrugS (RPI)
   - IRefIndex
     - Proteins, protein interactions
   - Drugbank
     - Proteins, drugs
   - Online Mandellian Inheritance in Man
     - Diseases, genes
8. Uniprot
   - Genes, proteins, protein interactions, gene ontology annotations

Statistics

- We used ***fuzzy c-means clustering*** and ***singular value decomposition*** to cluster genes and order them based on activation
  - The fuzzy c-means clustering technique identified one more gene cluster than the CORTECON analysis
  - What is significant about this cluster?
- ***Enrichment analysis*** was used to determine when a disease/pathway is enriched or depleted for a particular set of genes
- Statistical analyses are made available at dereferenceable, content-negotiable URIs

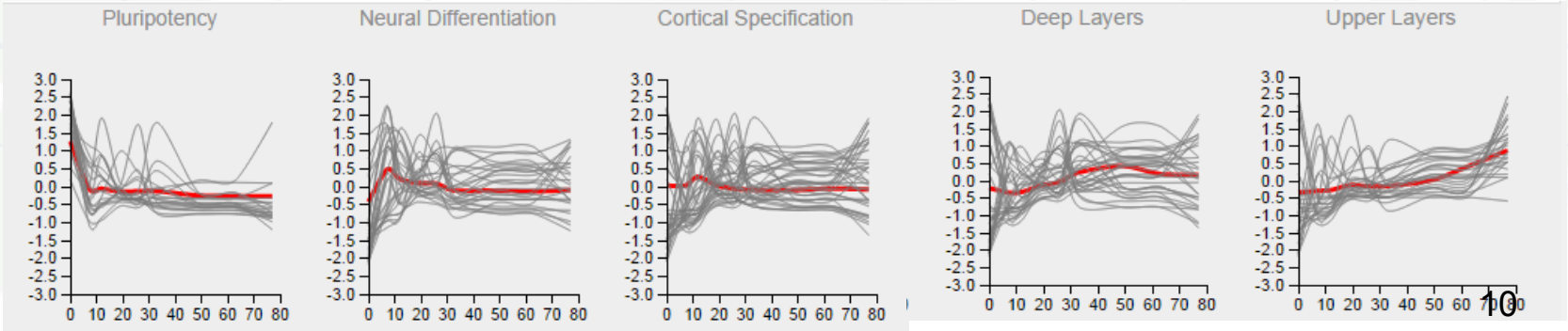
Extracted Brain Development Clusters



Placenta Epithelial Brain Brain Brain Brain

Tissue Enrichment

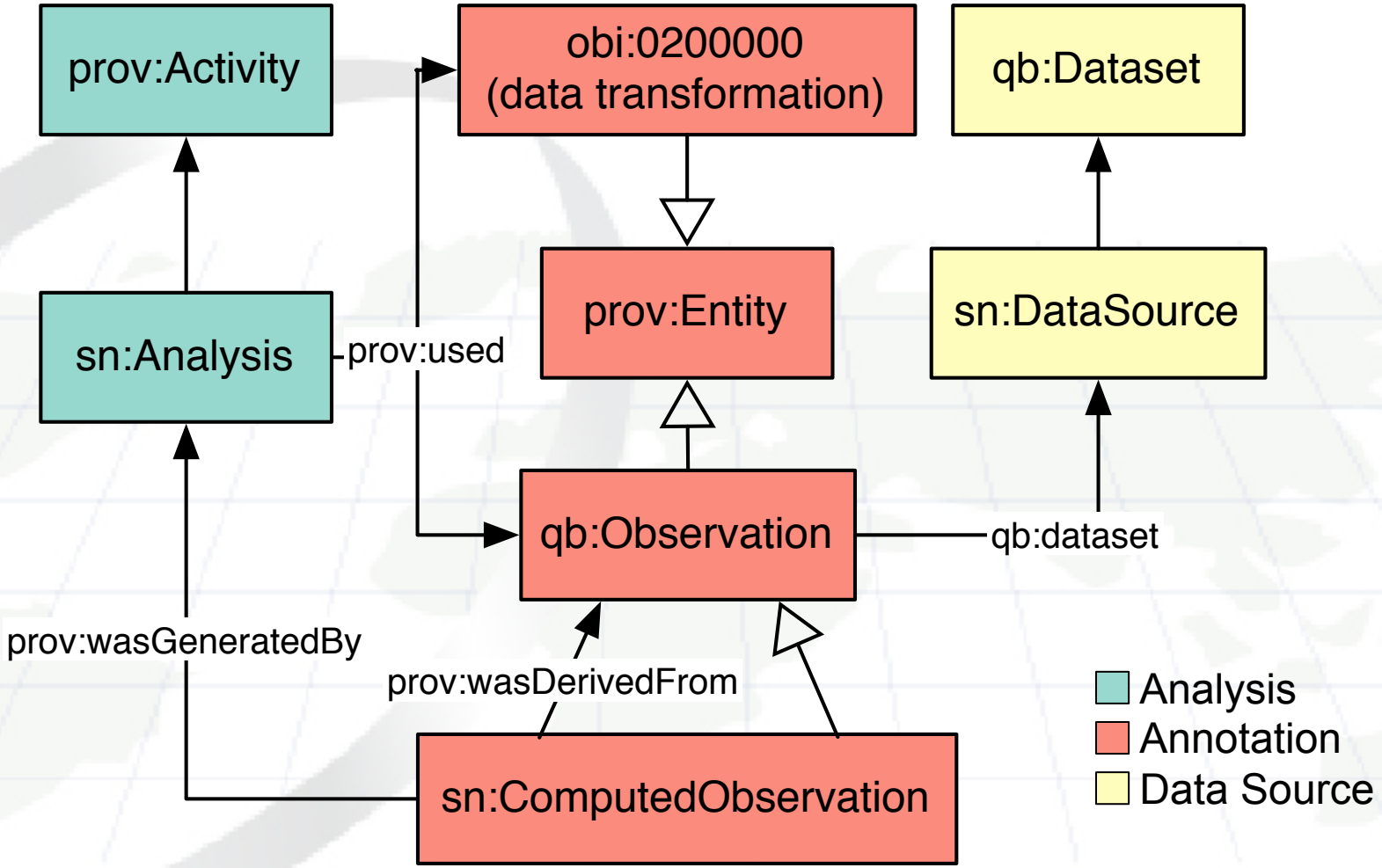
Stages from Cortecon http://cortecon.neuralsci.org/



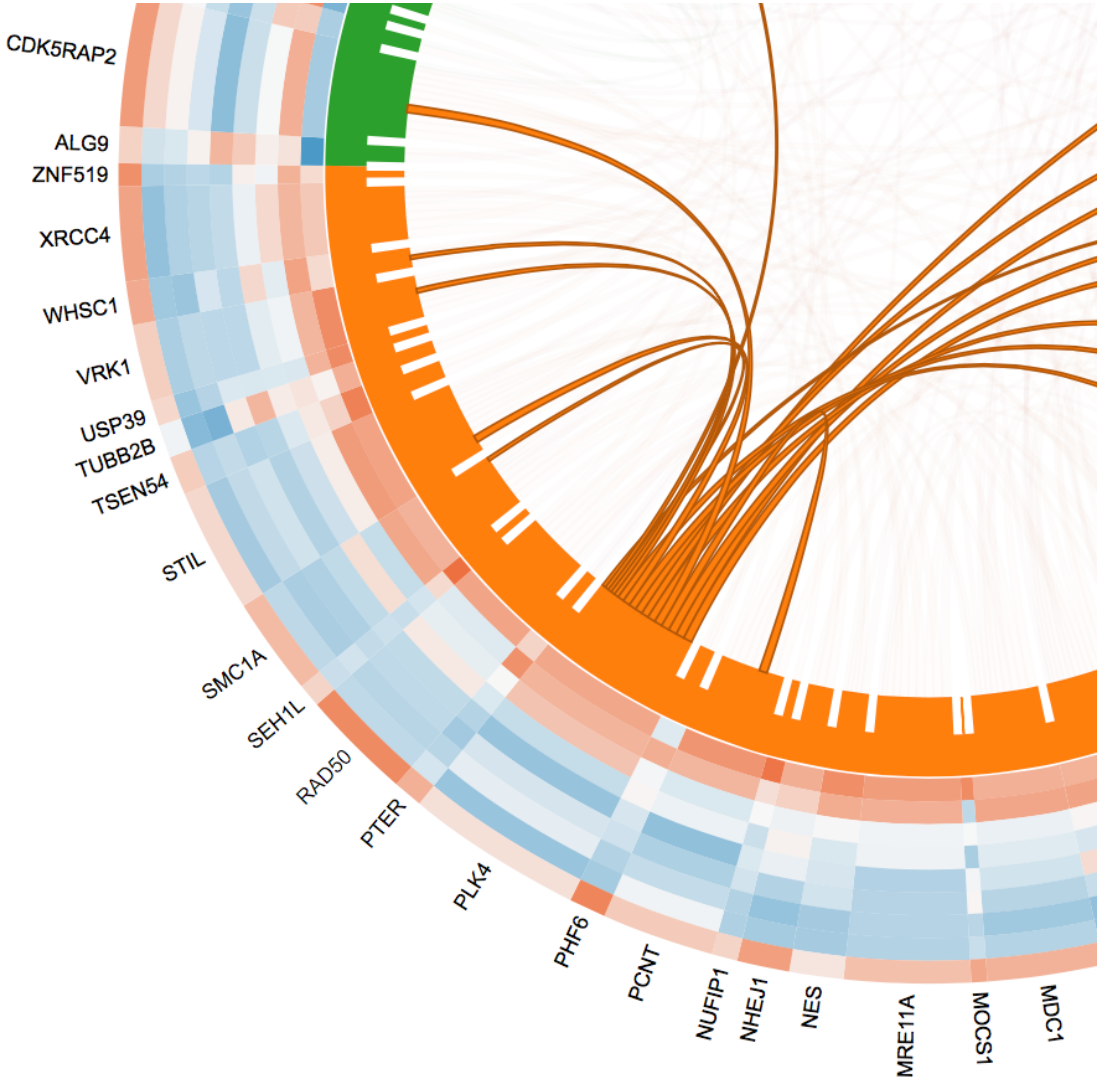
Linking

- Composed of two techniques:
  - Cross-database identifier linking:
    - Typical identifiers include Gene Ontology, NCBI, KEGG, Ensembl identifiers
  - Textual matching in the absence of unique identifiers:
    - “Alzheimer’s disease, familial, 1”, “Alzheimer’s disease, familial, 2”, etc. => “Alzheimer’s disease”
    - Exploit broader and type relations to collapse results into a “summary” entity

Provenance

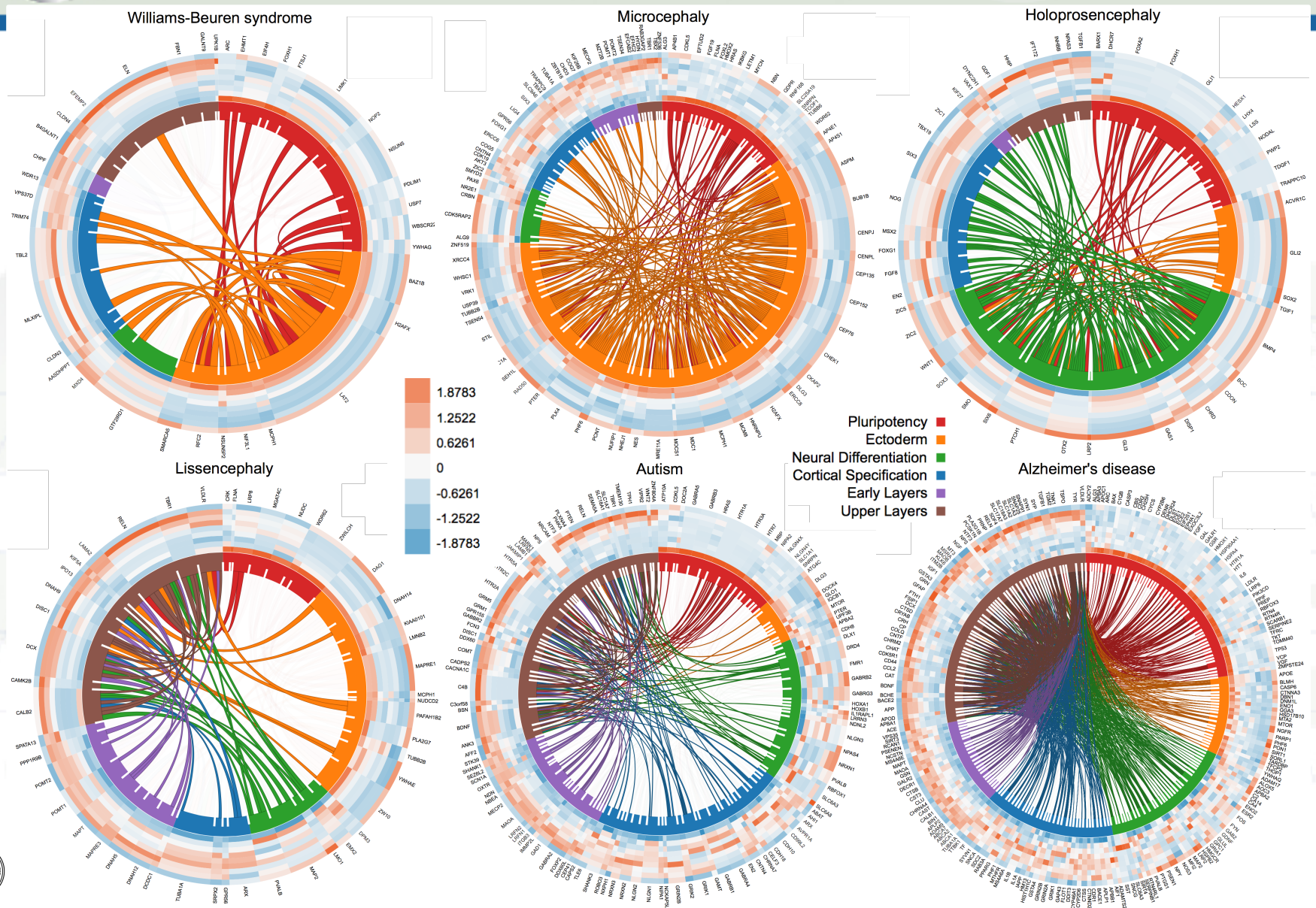

prov:Activity
obi:0200000
(data transformation)
qb:Dataset
sn:Analysis
prov:used
prov:Entity
sn:DataSource
qb:Observation
qb:dataset
prov:wasGeneratedBy
prov:wasDerivedFrom
sn:ComputedObservation
Analysis
Annotation
Data Source

Visualization



https://semnext.tw.rpi.edu/chem/

Visualization



Results

- Through the application of fuzzy c-means clustering, we discovered a new cluster of genes (Ectoderm) in the early stages of brain development
- Our combination of statistical analysis, linking to structured data sources, and visualization allow us to provide domain scientists with deeper insights into relationships within and between clusters

Conclusions

- We modeled statistical analyses of genomic data using best-in-class ontologies
- Analysis results were linked with additional structured data to provide literature support
- SemNExT combines statistical analysis and linked data in a generalizable way by incorporating new analyses, data sources, and ontologies
- We are using SemNExT with structured knowledge to make sense of a newly identified cluster of genes in neural development

Future Directions

- Applying SemNExT to two related domains of plant microbiome and child development
- Looking for users/collaborators for feedback
  - Contact dlm@cs.rpi.edu or pattoe@rpi.edu

Acknowledgements

- NSF Graduate Research Fellowship: Mr. Patton
- NSF Grant 1331023: Ms. Brown, Ms. De los Santos, and Dr. Bennett
- RPI Internal Funding
- Dr. John Erickson for valuable feedback on the project
- SemStats 2015 workshop chairs and reviewers

References

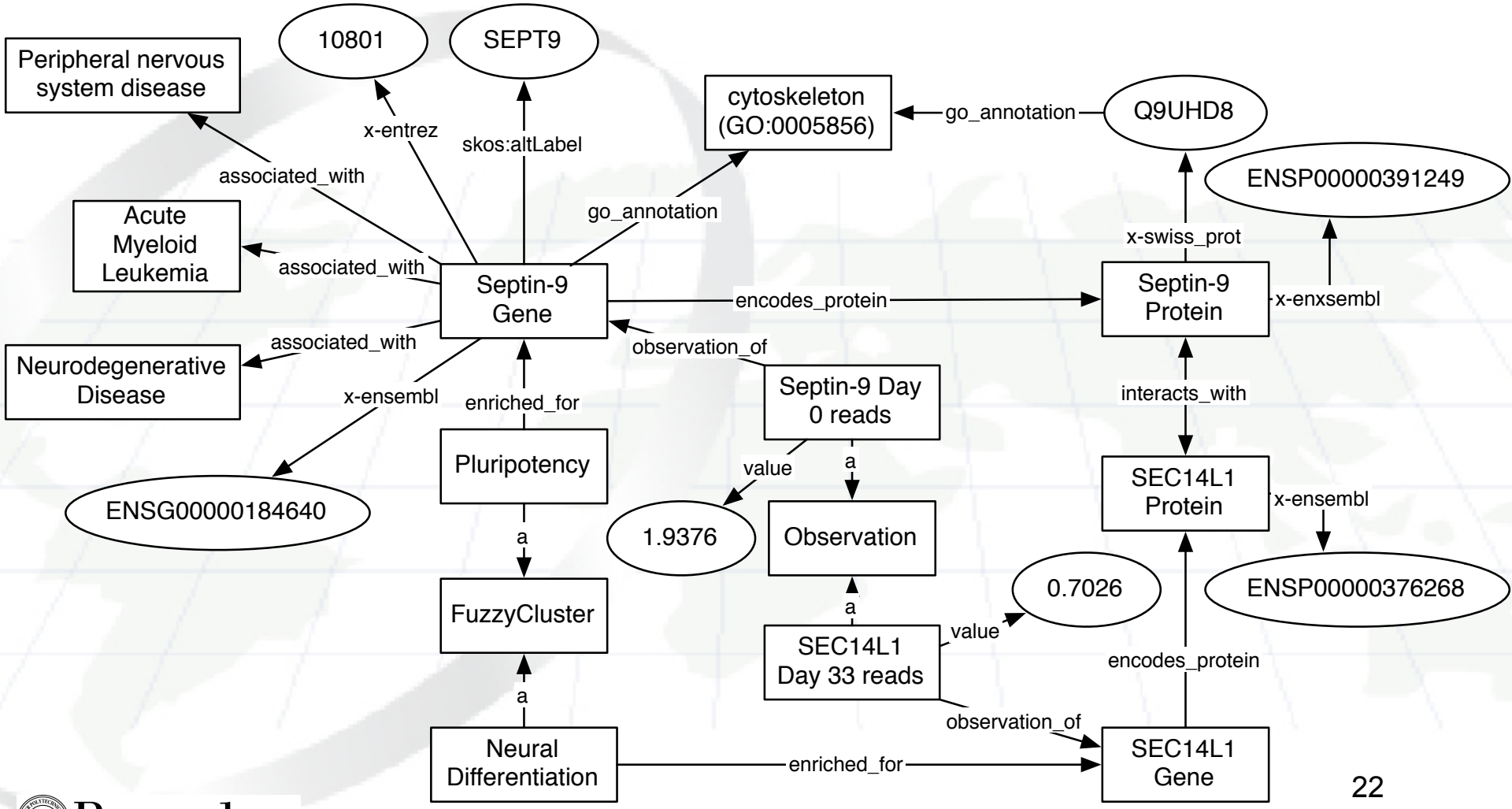
- Neural Stem Cell Institute – http://www.neuralsci.org/
- NSCI Cortecon – http://cortecon.neuralsci.org/
- KEGG – http://www.genome.jp/kegg/
- StringDB – http://string-db.org/
- Ensembl – http://www.ensembl.org/
- UMLS – https://www.nlm.nih.gov/research/umls/
- Bio2RDF – http://bio2rdf.org/
- ReDrugS – http://redrugs.tw.rpi.edu/
- Uniprot – http://www.uniprot.org/
- SemNExT Demo – https://semnext.tw.rpi.edu/chem/

QUESTIONS?

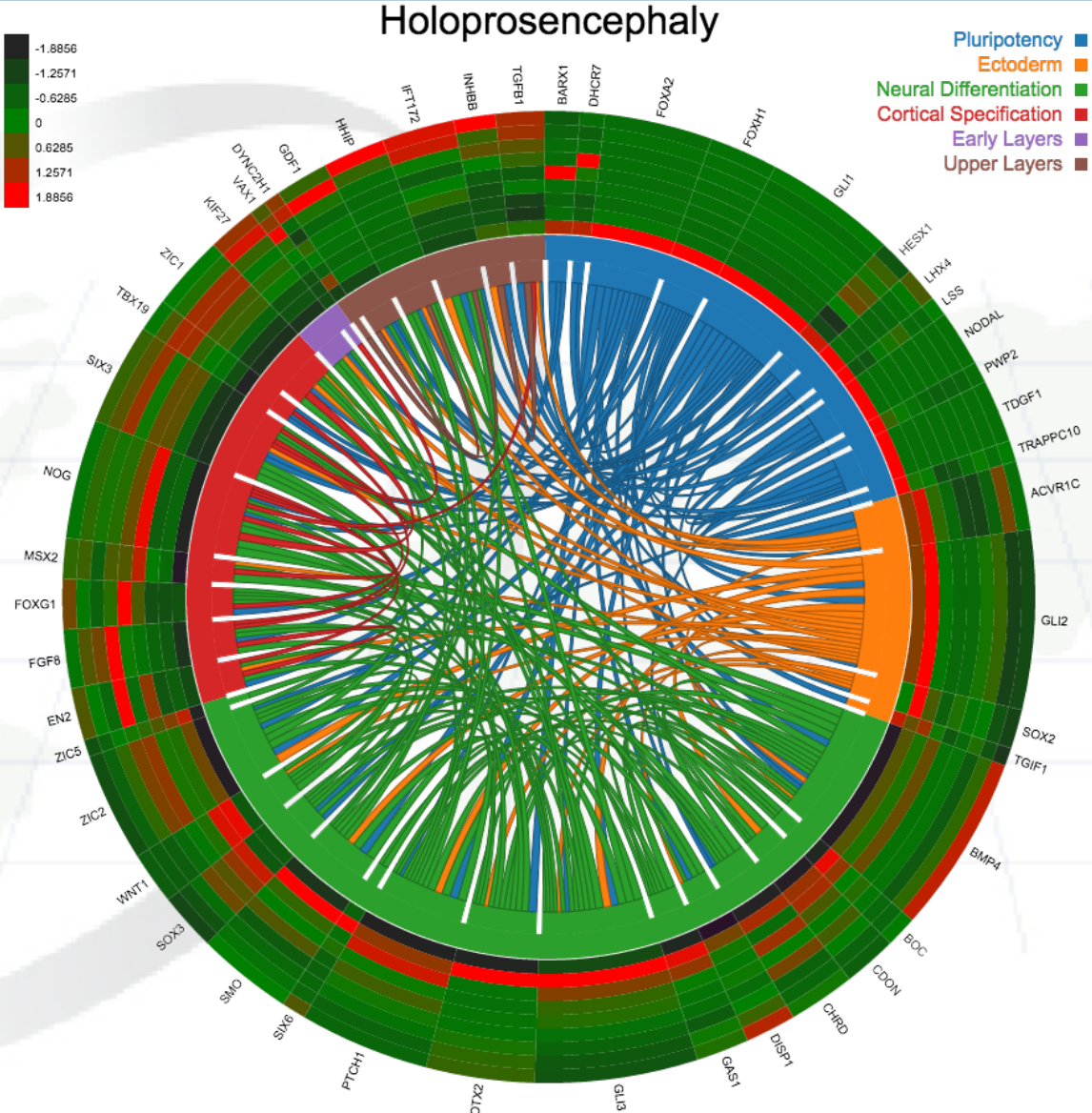
dlm@cs.rpi.edu | pattoe@rpi.edu

Extras

Example


10801
SEPT9
Peripheral nervous system disease
x-entrez
skos:altLabel
cytoskeleton (GO:0005856)
go_annotation
Q9UHD8
ENSP00000391249
associated_with
Acute Myeloid Leukemia
go_annotation
x-swiss_prot
associated_with
Septin-9 Gene
encodes_protein
Septin-9 Protein
x-enxsembl
associated_with
observation_of
Neurodegenerative Disease
Septin-9 Day 0 reads
interacts_with
x-ensembl
enriched_for
Pluripotency
value
a
SEC14L1 Protein
x-ensembl
ENSG00000184640
1.9376
Observation
a
0.7026
ENSP00000376268
a
FuzzyCluster
value
SEC14L1 Day 33 reads
encodes_protein
a
observation_of
Neural Differentiation
enriched_for
SEC14L1 Gene

Visualization



Rensselaer

Stage Enrichment for Selected Diseases