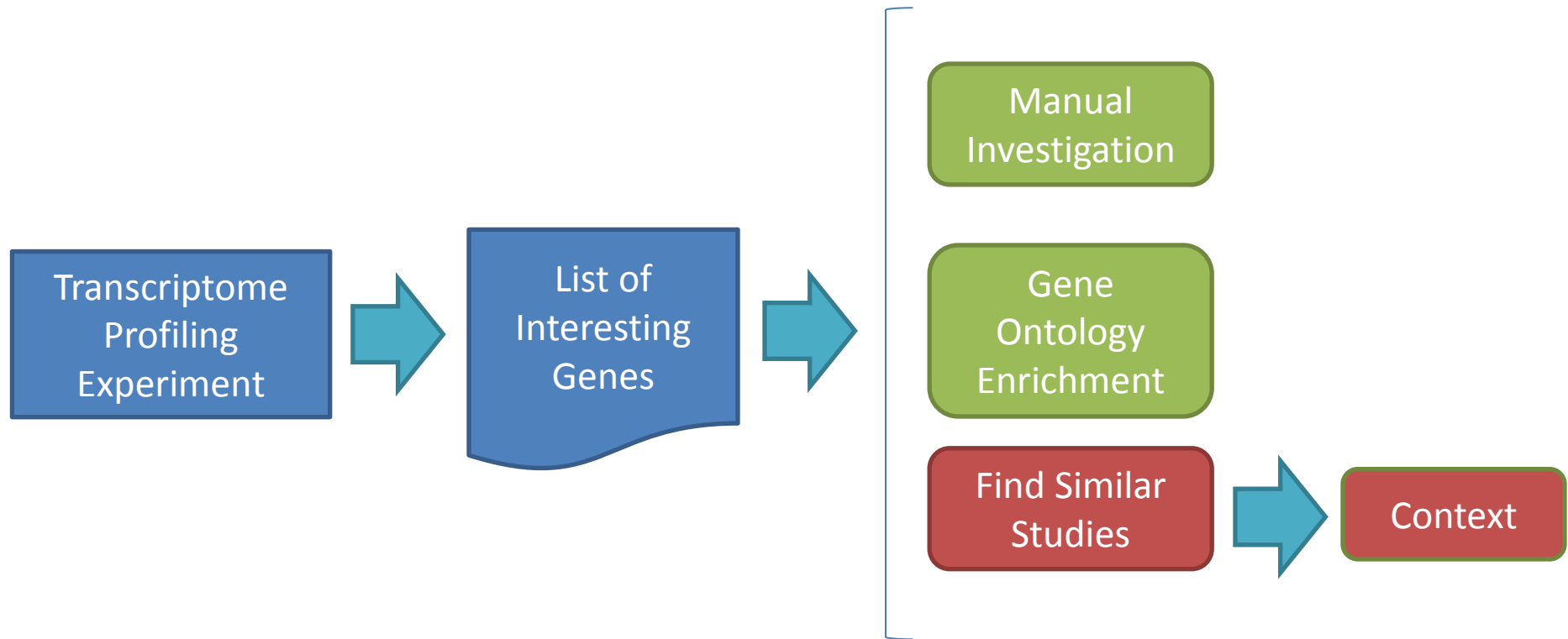


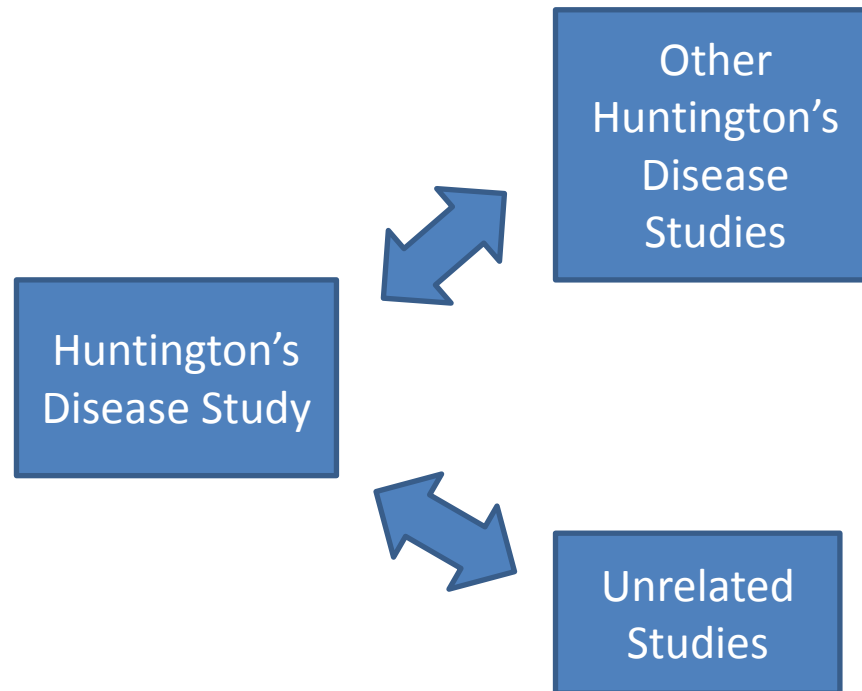
# Wide-scale Comparison of Transcriptome Data

Raymond Lim, Pavlidis Lab

# Motivation



# Contextualize Studies



# Background: Microarray Study

- Profile expression of thousands of genes
- Differential expression between groups
  - e.g. control vs. treatment
  - Potentially complicated design
- Different platforms
  - Versions
  - Manufacturers, e.g. Affymetrix, Illumina

# Overview

Gene	Rank
Pzp	11586
Aanat	12671
Aatk	1022
...	...

Gene	Dataset A Rank	Dataset B Rank	...
Pzp	334	6721	...
Aanat	752	384	...
Aatk	76	103	...
...	...	...	...

Dataset	Similarity Score
A	0.3
B	0.4
...	...



# Challenges

- Measuring similarity between gene lists
  - Noisy data
- Validation of results
  - Gold standard
- Confounds
  - Platform effects
  - Difference in statistical power

# Data Overview

- Gemma
  - Framework for meta-analysis of gene expression data
  - Automated differential expression analyses
    - Not published results
- Annotations

Taxon	NumSamples	Platforms	Factors
human:224	Min. : 4.00	GPL1261:137	Treatment : 44
mouse:349	1st Qu.: 12.00	GPL570 : 79	time : 30
	Median : 18.00	GPL81 : 70	DiseaseState : 29
	Mean : 26.46	GPL96 : 46	Treatment SamplingTimePoint: 25
	3rd Qu.: 29.00	GPL339 : 32	Genotype : 22
	Max. :280.00	GPL260 : 29	Genotype Treatment : 22
		(Other):180	(Other) :401

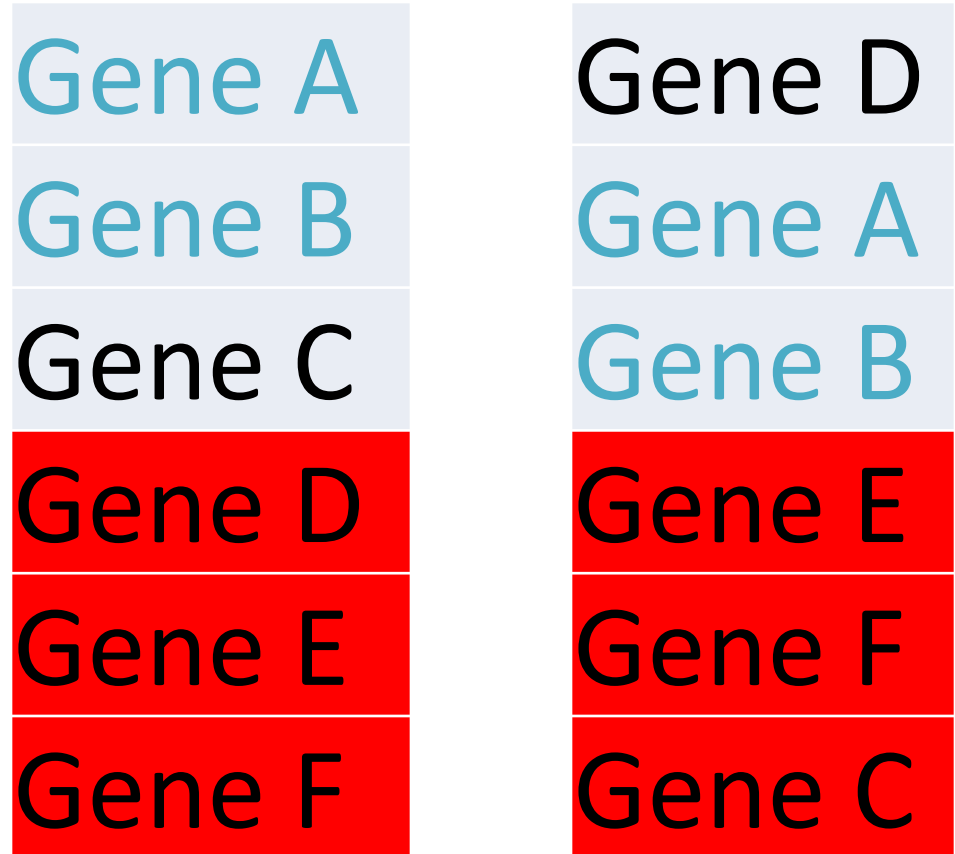
# Dataset Annotations

- Automated and manual
- Manually clustered subset
- Measure similarity
- Semi-gold standard

	Brain	Gene X	Disease
Dataset A	1	1	0
Dataset B	0	1	1
Dataset C	0	0	1

# Measuring Gene Signature Similarity

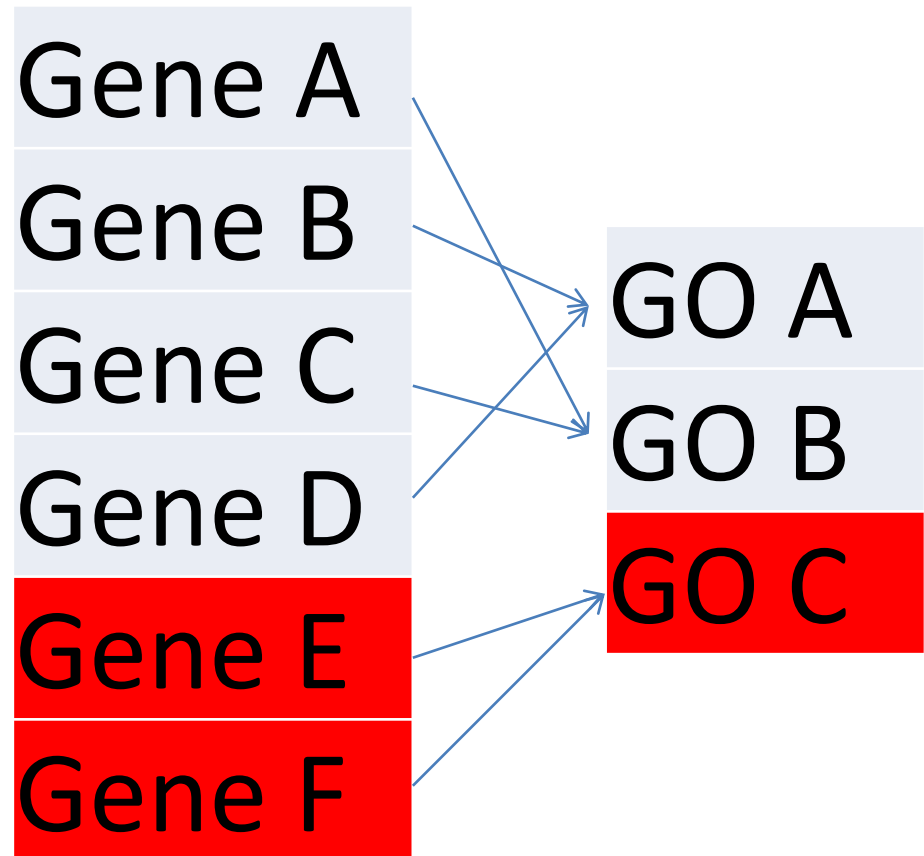
- Rank based
  - Spearman, Kendall's tau
  - Bottom of list noisy
- Threshold based
  - Gene overlap
    - Fisher's exact test
- Threshold/rank based
  - ROC method
  - Top-k Kendall
- Ontology based
  - Low Coverage
  - Redundancy in GO
    - Filtering
  - Low resolution
  - Performs best



Red: Below threshold

# Measuring Gene Signature Similarity

- Rank based
  - Spearman, Kendall's tau
  - Bottom of list noisy
- Threshold based
  - Gene overlap
    - Fisher's exact test
- Threshold/rank based
  - ROC method
  - Top-k Kendall
- Ontology based
  - Low coverage
  - Redundancy in GO
    - Filtering
  - Low resolution



Red: Below threshold

# Validation

Gene Expression Signatures



Pair-wise Comparison



Similarity Scores



Comparison



	A	B	C
A	1	0.5	0.4
B		1	0.3
C			1

	A	B	C
A	1	0.4	0.3
B		1	0.2
C			1



Pair-wise Comparison



Similarity Scores



Comparison

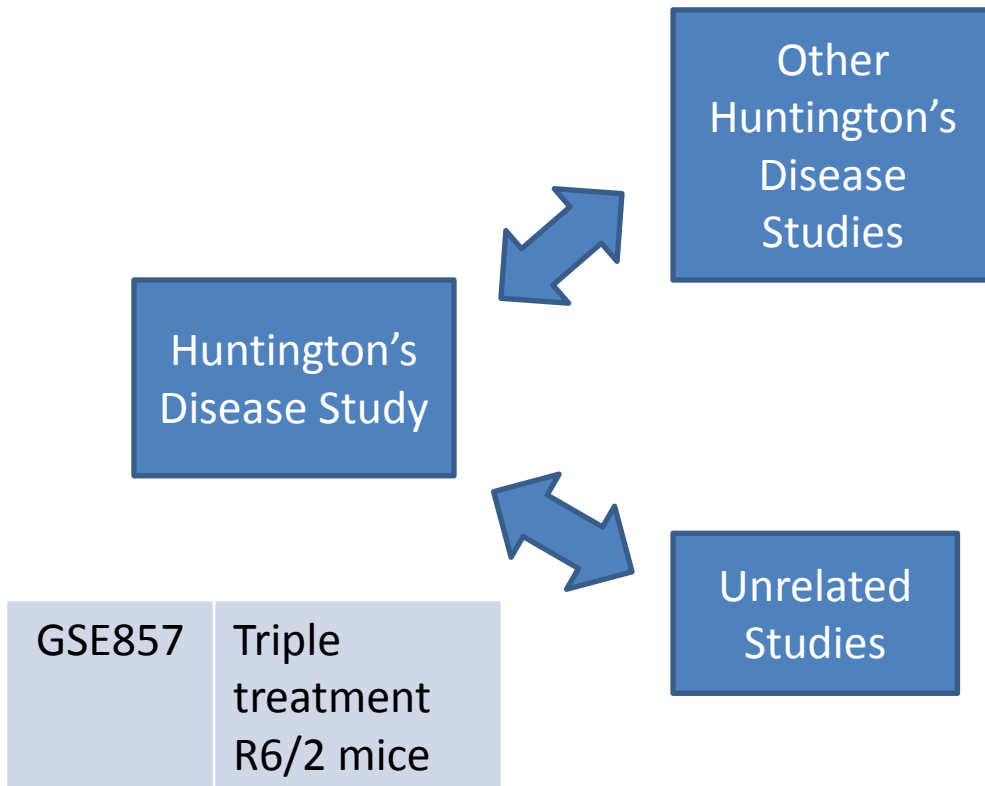


	A	B	C
A	1	0.5	0.4
B		1	0.3
C			1

	A	B	C
A	1	0.4	0.3
B		1	0.2
C			1

Annotations  
e.g. brain,  
kidney, cancer

# Example

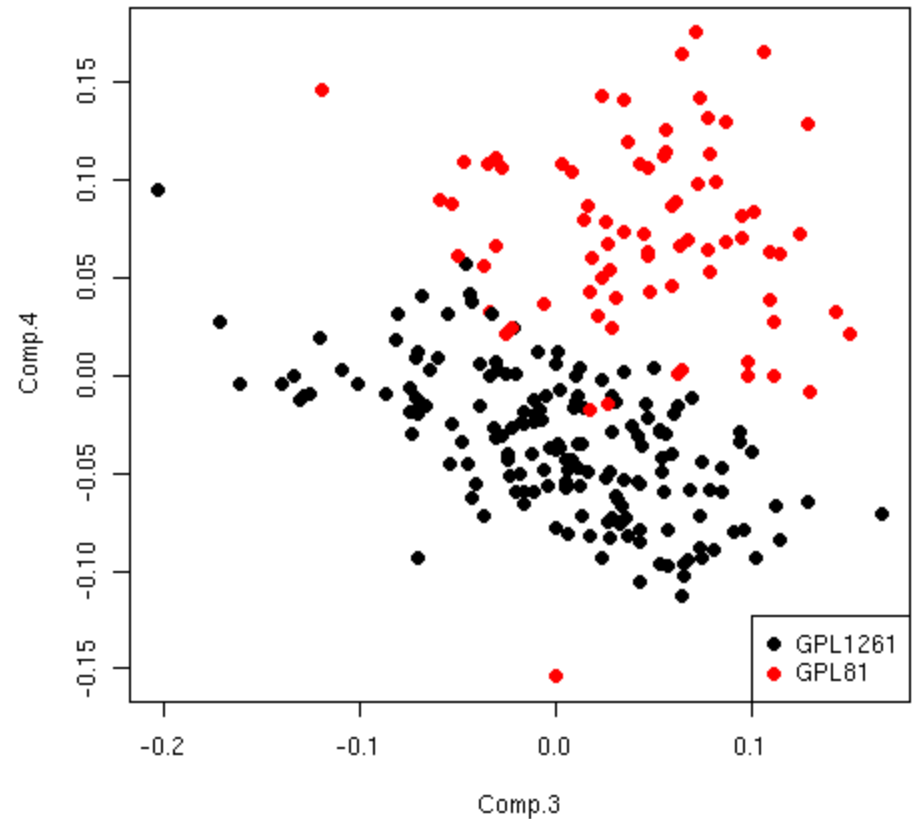


Dataset	Description
GSE3634	Molecular pathways involved in neuronal degeneration of polyglutamine mouse models
GSE3621	R6/1 brain hemisphere time series gene expression

Dataset	Description
GSE2161	Identification of genes that are dysregulated in the telencephalon of Dlx1/2 mutants.
GSE6678	Palmitoyl protein thioesterase-1 knockout mice

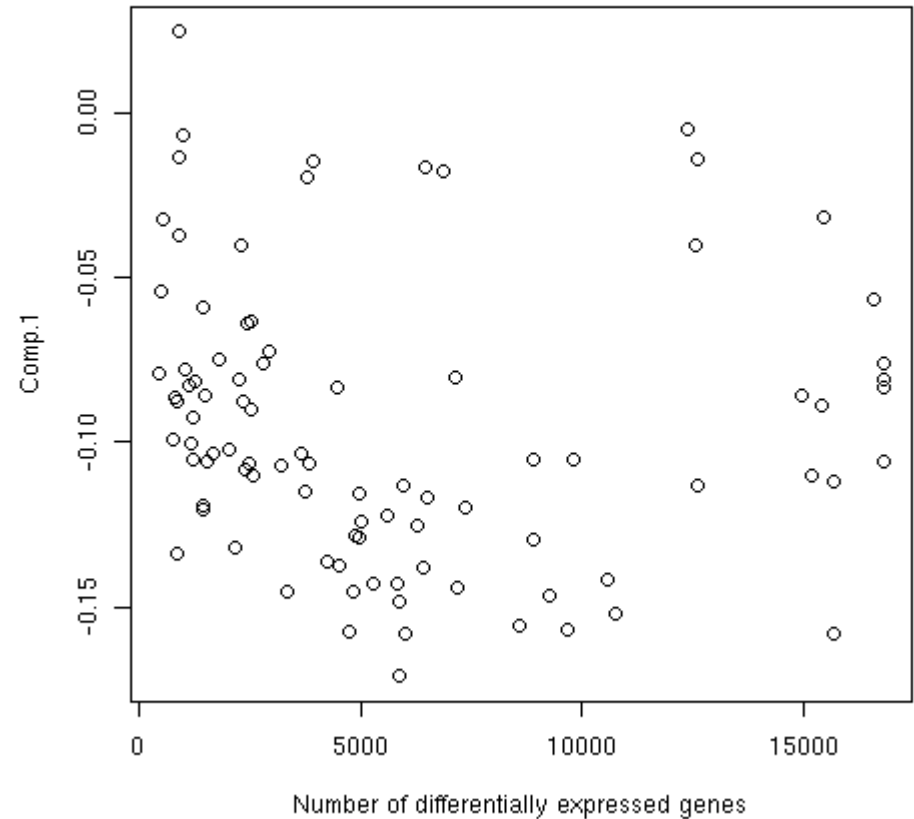
# Platforms Cluster

- Intersection of genes (~7000)
- GPL1261: Affymetrix Mouse 430 2.0
- GPL81: Affymetrix Mouse U74 2.0



# Studies Differ in Statistical Power

- Widely varying numbers of differentially expressed genes
- Datasets with similar power tend to cluster together



# Concluding Remarks and Future Direction

- Provide context for gene lists
- Develop better gold standard
- Manual quality control
- Investigate platform effect
- Gene dynamics
- Identify modules of differential expression

# Acknowledgements

- Paul Pavlidis
- Leon French
- Artemis Lai
- Members of Pavlidis Lab