Comparison of Boolean Implication and Correlational Methods for In Silico Gene Reporting of Retinal Cell Types

Rohan Subramanian

3rd August 2020

UC San Diego
Boolean Lab

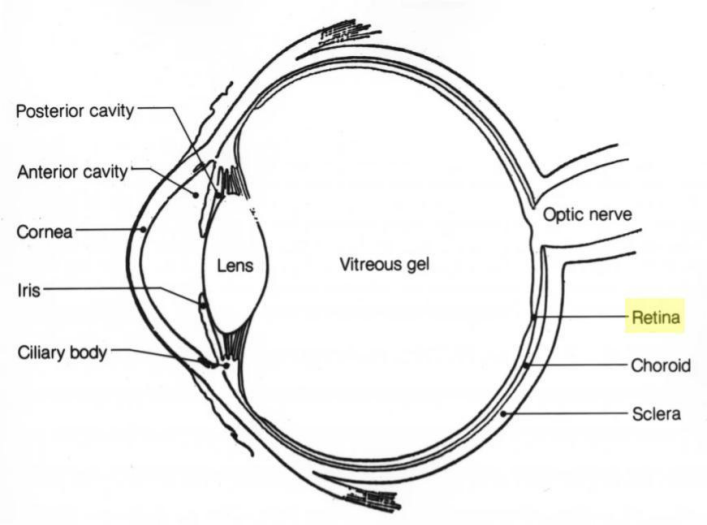
Overview

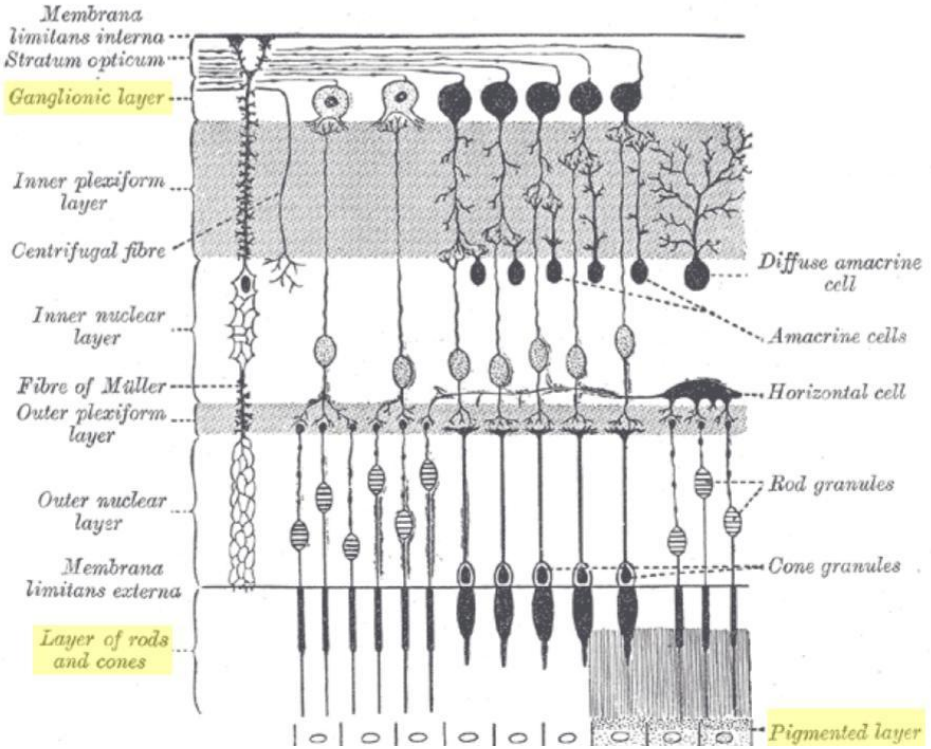
Aim: **Does boolean implication improve the prediction accuracy of in silico gene reporting of retinal cell types compared to correlational methods?**

1. Significance of research into retinal cell types.
2. Previous approaches to this problem.
3. Advantages of boolean approach and application to specific datasets used.
4. Quantification of results and comparison with correlation.
5. New discoveries from method.

UC San Diego
Boolean Lab

Structure of Retina



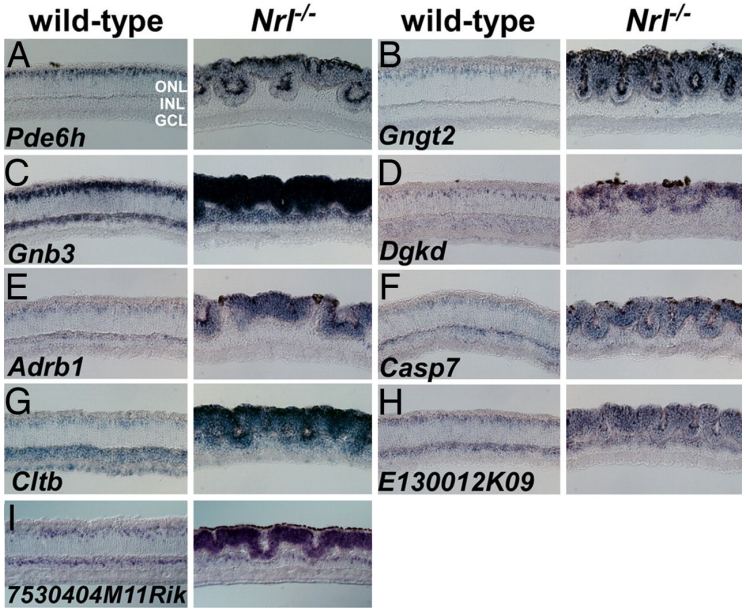


UC San Diego
Boolean Lab

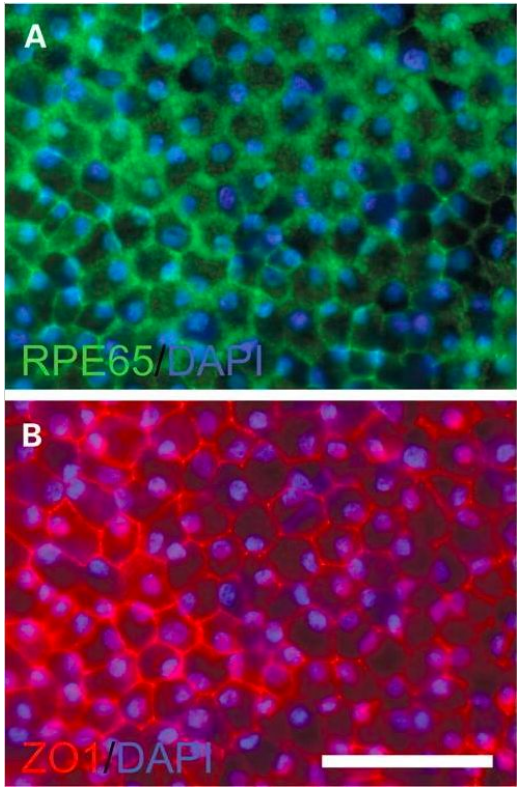
Clinical Significance of Studying Retinal Cell Types

- Generation of retinal cell types from stem cells to treat diseases such as AMD, AIR and retinitis pigmentosa that remain major causes of blindness in the developing world.
- Treatment of cancers affecting eye such as retinoblastoma.
- California Project to Cure Blindness-Retinal Pigment Epithelium 1 (CPCB-RPE 1) transplanted embryonic stem cell-derived RPE into retina of AMD-affected patients in a clinical trial, with limited success.
- Differentiation and purification of other retinal cell types more difficult.

Previous Approaches In Vivo



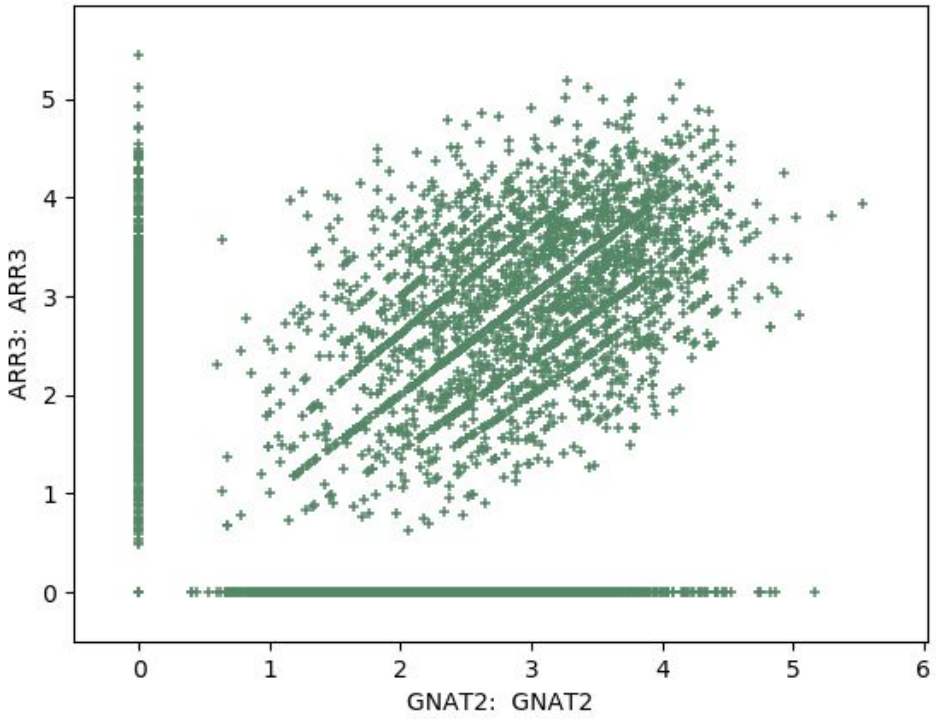
Corbo et al. 2007: Identification of differentially expressed genes from microarray data, followed by confirmation through staining of retinal tissue from genetically modified murine models.



Liao et al. 2010: Analysis of gene expression levels in purified cell type (RPE) through a genetically engineered fluorescent reporter line.

UC San Diego
Boolean Lab

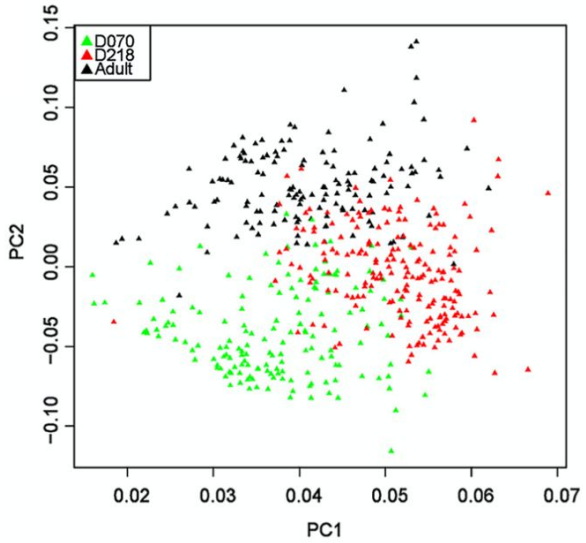
Single Cell RNA-seq data and In Silico Gene Reporting



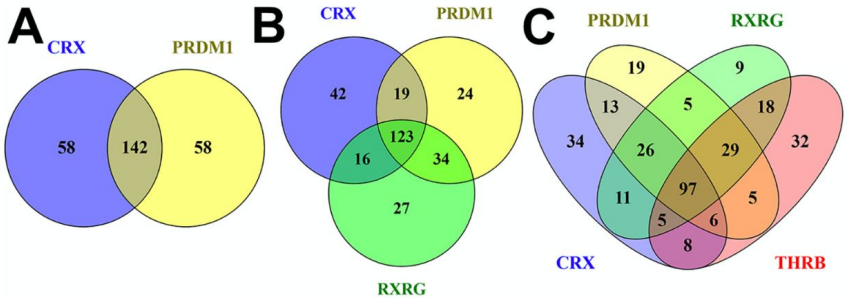
- Captures transcriptome of a single cell at a single point in time.
- Large numbers of zeroes, often “dropouts”, that may be false negatives.
- Technical artifacts due to PCR.
- Has led to the development of in silico gene reporting methods.

UC San Diego
Boolean Lab

Previous Approaches In Silico



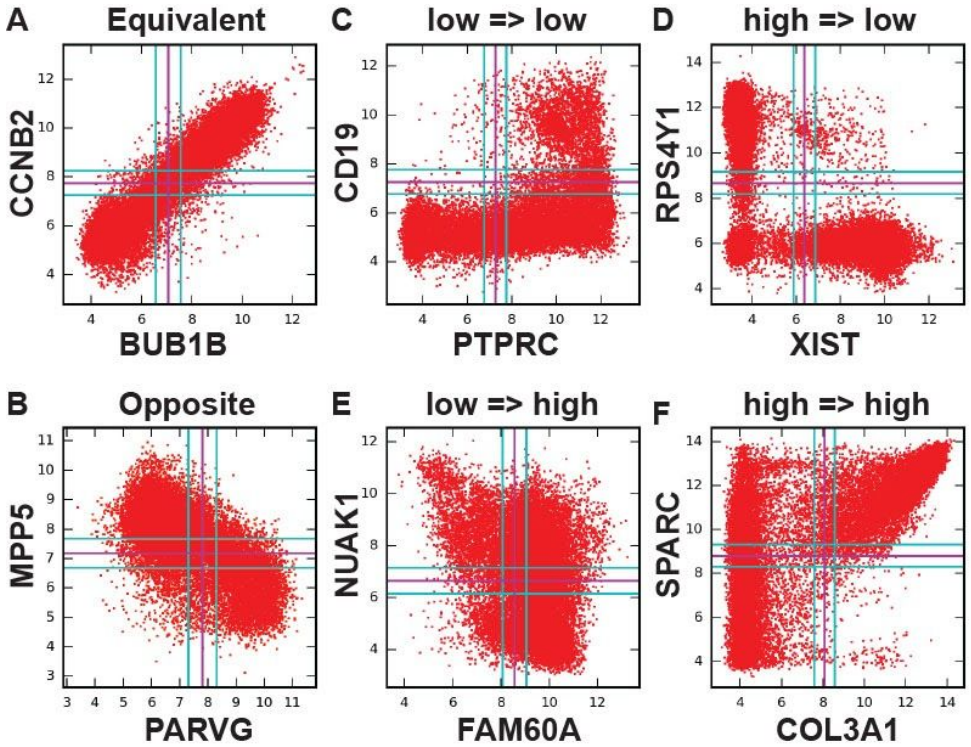
Phillips et al. 2018: Attempt to classify individual cells into clusters representing different cell types using PCA plots, not successful in their case.



Phillips et al. 2018: Novel method to find genes which are highly correlated with known markers of cones by taking the intersection of their top 200 correlating genes. (Spearman's rank correlation)

UC San Diego
Boolean Lab

A Boolean Approach

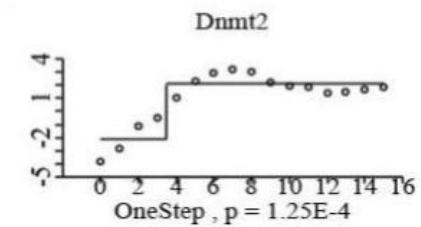


- Six types of boolean implication relationship: two symmetric, four asymmetric.
- Correlation cannot capture asymmetric relationships.
- Work at the Boolean Lab has shown that boolean implication has the ability to filter out noise better than correlational methods.
- We will attempt to test that hypothesis in this application.

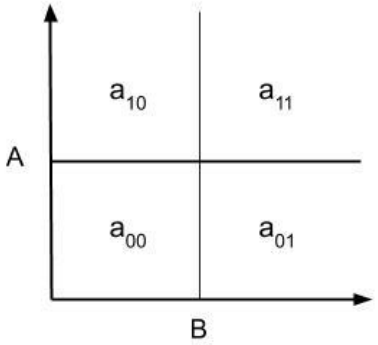
UC San Diego
Boolean Lab

StepMiner and BooleanNet

A



B



C

$$nA_{low} = (a_{00} + a_{01})$$
$$nB_{low} = (a_{00} + a_{10})$$
$$\text{total} = a_{00} + a_{01} + a_{10} + a_{11}$$

$$\hat{e} = \frac{\frac{nA_{low}}{\text{total}} \cdot \frac{nB_{low}}{\text{total}}}{\text{total}}$$

$$S_{00} = \frac{\hat{e} - n}{\sqrt{\hat{e}}}$$

$$p_{00} = \frac{1}{2}\left(\frac{a_{00}}{(a_{00} + a_{01})} + \frac{a_{00}}{(a_{00} + a_{10})}\right)$$

Boolean Implication $= S > 2.5, p < 0.35$

D

Equivalent $= S_{01} > 2.5, p_{01} < 0.35, S_{10} > 2.5, p_{10} < 0.35$

Opposite $= S_{00} > 2.5, p_{00} < 0.35, S_{11} > 2.5, p_{11} < 0.35$

A low $\Longrightarrow$ B high $= (S_{00} > 2.5, p_{00} < 0.35)$

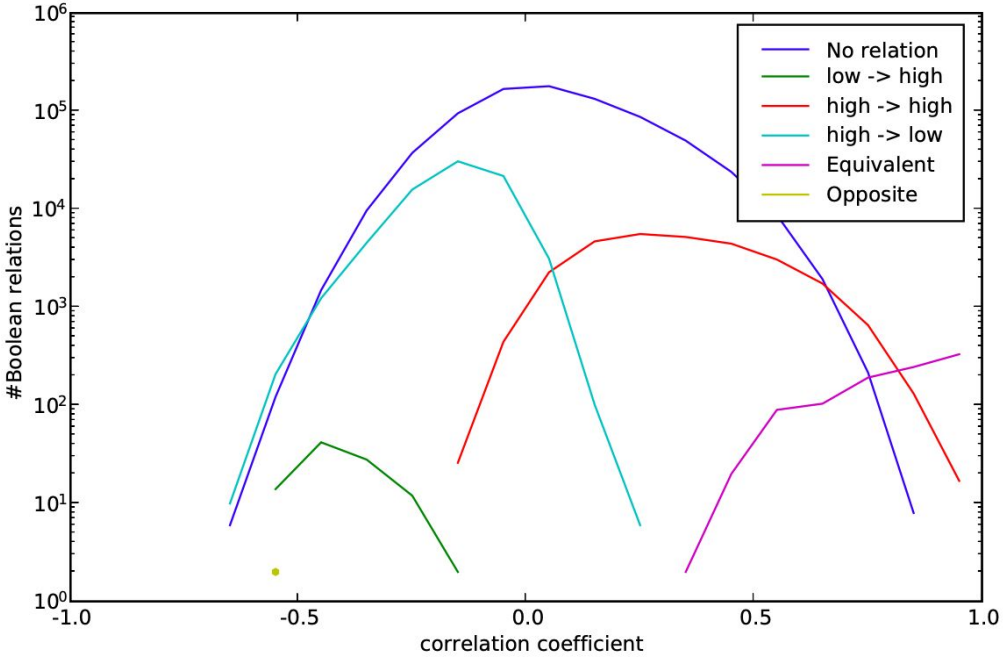
A low $\Longrightarrow$ B low $= (S_{01} > 2.5, p_{01} < 0.35)$

A high $\Longrightarrow$ B high $= (S_{10} > 2.5, p_{10} < 0.35)$

A high $\Longrightarrow$ B low $= (S_{11} > 2.5, p_{11} < 0.35)$

UC San Diego
Boolean Lab

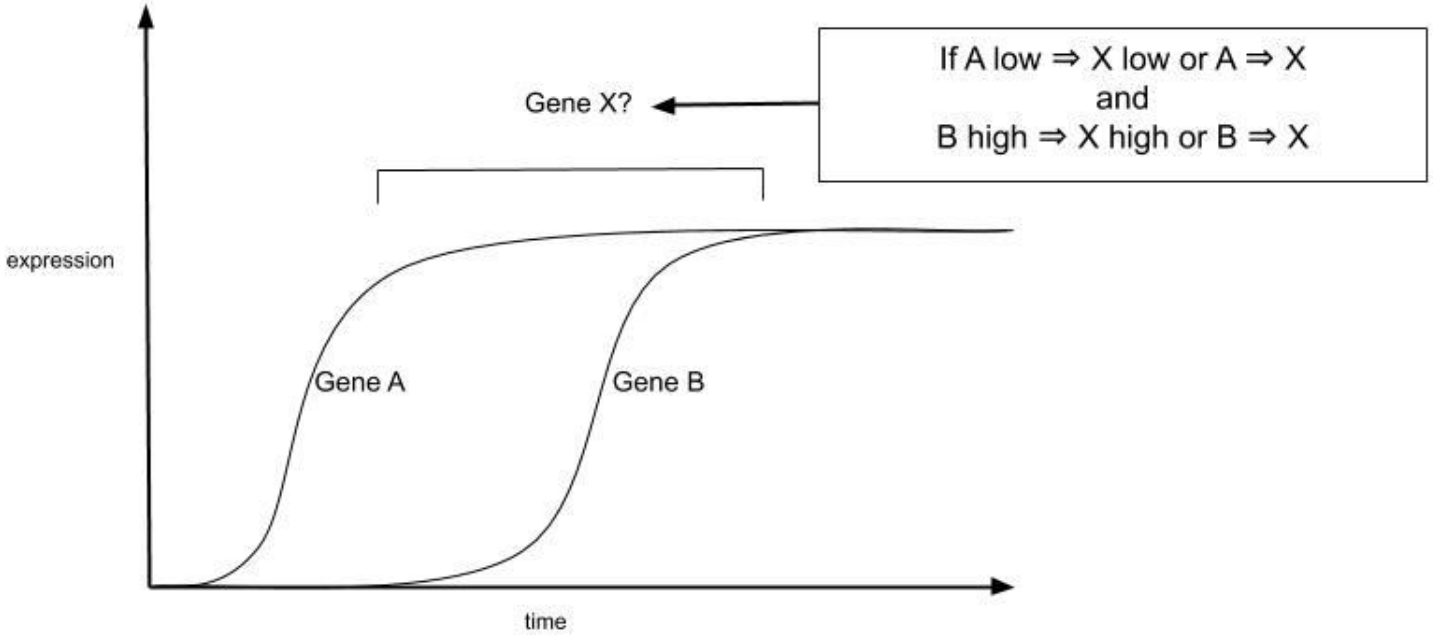
Boolean Implication and Correlation



- Boolean implication relationships can also be interpreted as subsets, allowing insight into specificity of genes that correlation does not provide.
- Correlation completely disregards asymmetric relationships.
- Correlation may only be observed in a single subset of data that the operator must choose, but boolean implication is evident across entire dataset.

UC San Diego
Boolean Lab

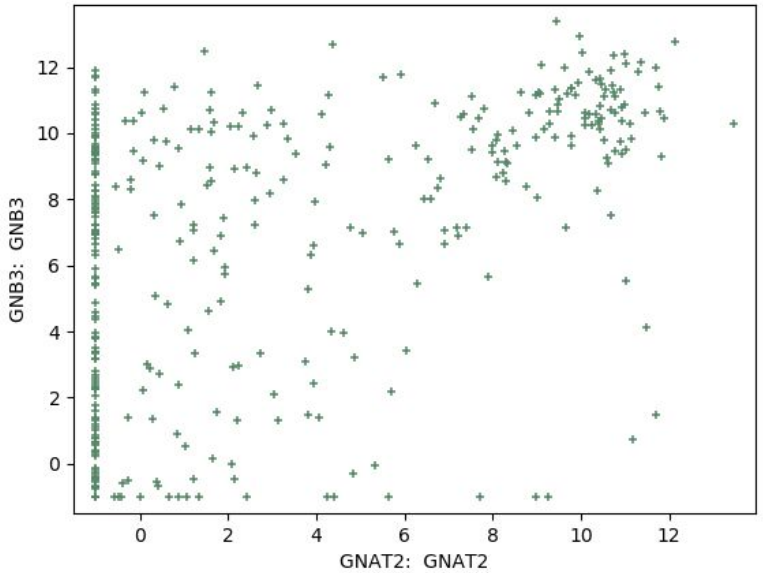
Application of Boolean Implication to Genes involved in Cell Fate



UC San Diego
Boolean Lab

Characteristics and Processing of Phillips 2018 Dataset

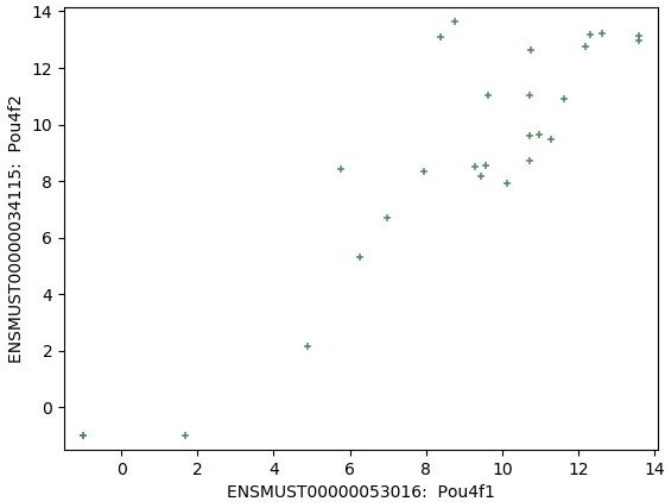
- scRNA-seq data from 546 cells of optic vesicles (organoids) developing from hPSCs.
- Mixed culture of photoreceptors, progenitors, ganglion and RPE cells from 3 time points: day 70, day 218 and adult.
- Phillips et al. 2018 developed Spearman's rank correlation coefficient analysis (SRCCA) when PCA failed to yield well-defined cell clusters.
- Phillips et al. 2018 used **median-by-ratio normalization** on TPM values, while we used **$\log_2$(v+1) transformation** as scRNA-seq data follows the Poisson distribution.



UC San Diego
Boolean Lab

Identification of Gold Standard and Bait Genes

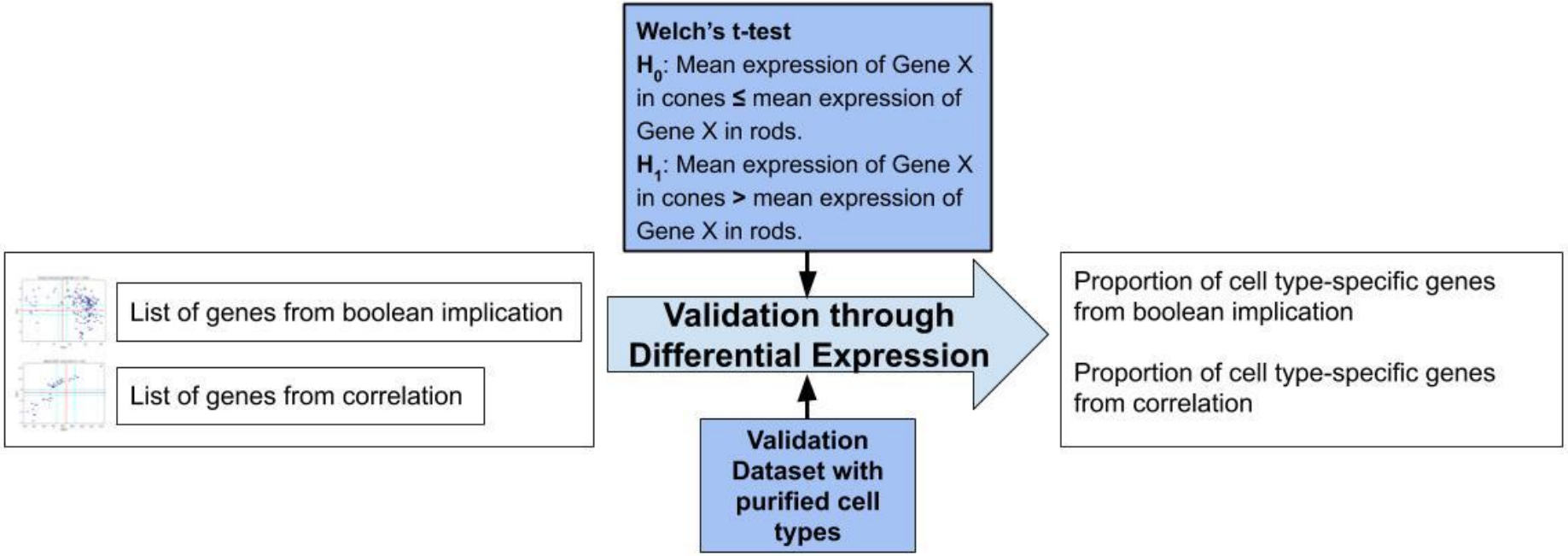
- To find “gold standard” genes, genes reported in literature were verified in datasets containing purified RGCs, RPCs and RPEs.
- We searched for bait genes which yielded a shorter list of genes with a large number of “gold standard” genes.



	Boolean Analysis	SRCCA
Cone photoreceptors	GNAT2, ARR3	CRX, PRDM1, RXRG, THRB
Rod photoreceptors	PDE6B, NR2E3	NRL, NR2E3
Retinal ganglion cells (RGC)	ISL1, POU4F2	ATOH7, POU4F2
Retinal progenitor cells (RPC)	PAX6, VSX2	VSX2, VIM
Retinal pigment epithelium (RPE)	DIXDC1, PMEL, RPE65	PMEL, TYRP1

UC San Diego
Boolean Lab

Quantification of Results



Additional level of specificity: Threshold for minimum difference in mean expression values (>0.5)

UC San Diego
Boolean Lab

Cell Type Specificity for Rod Photoreceptor Genes

	Correlation (NRL, NR2E3)	Boolean (PDE6B, NR2E3)	Correlation and Boolean	Correlation and not Boolean
Rod-specific	29	21	16	12
Specific in cones	7	5	2	5
No statistically significant difference	20	4	1	19
Total	56	30	19	36
Proportion	0.517	0.700	0.842	0.333

P-value from 2-proportion z test between cone-specific genes of 1st and 3rd column = **0.013 < 0.05**

UC San Diego
Boolean Lab

Cell Class Specificity for Rod Photoreceptor Genes

We verified whether the improvement in prediction accuracy for cell type also held true for cell class (photoreceptor).

	Correlation (NRL, NR2E3)	Boolean (PDE6B, NR2E3)	Correlation and Boolean	Correlation and not Boolean
PR-specific	42	34	19	22
Specific to non-PR cell types	2	1	0	2
No statistically significant difference	12	5	0	12
Total	56	40	19	36

P-value from 2-proportion z test between rod-specific genes of 1st and 3rd column = **0.016 < 0.05**

UC San Diego
Boolean Lab

Cell Type Specificity for Cone Photoreceptor Genes

	Correlation (CRX, PRDM1, THRB, RXRG)	Boolean (ARR3 and GNAT2)	Correlation and Boolean	Correlation and not Boolean
Cone-specific	32	15	9	23
Specific in rods	26	4	3	23
No statistically significant difference	36	11	6	30
Total	94	30	18	76
Proportion	0.340	0.500	0.500	0.303

(Hartl 2017 dataset)

UC San Diego
Boolean Lab

Quantifying Reproducibility using common Bait Genes

Direct repetition of analysis in Voigt 2020 human scRNA seq dataset (≈21,000 cells) using CRX, GNAT2 and GNB3 as bait genes for **cone photoreceptors** in both correlation and boolean implication.

	Correlation	Correlation and Boolean
Reproduced	11	7
Not Reproduced	24	0
Total	35	7

P-value from 2-proportion z test = 0.00082 < **0.05**

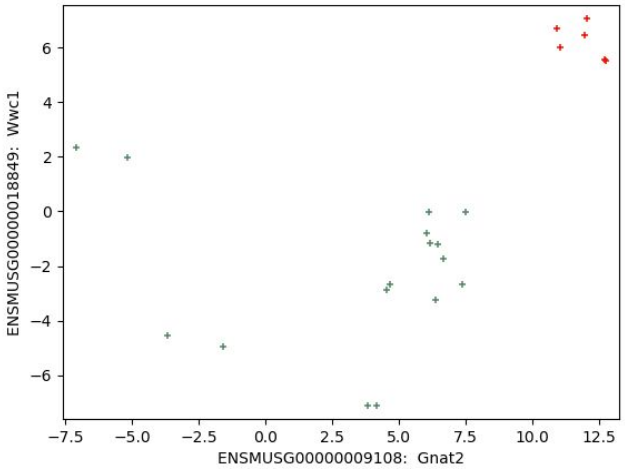
UC San Diego
Boolean Lab

WWC1: Novel Cone Photoreceptor Marker Gene

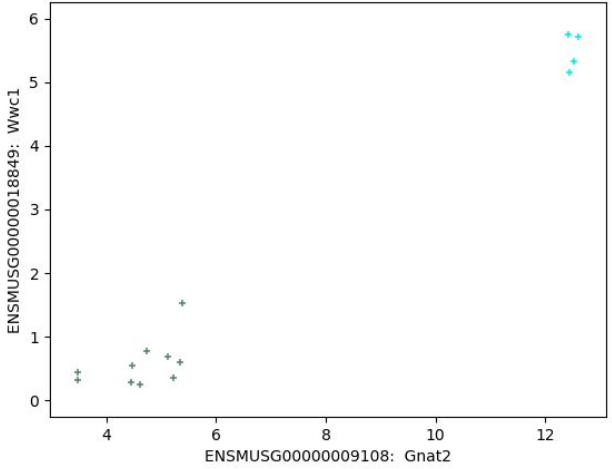
- WWC1 encodes WW domain-containing protein 1, which has broadly been described to have a function in the nervous system.

- Boolean implication identified WWC1 as a cone-specific gene, which was confirmed in the Sarin and Hartl purified cone datasets.

Sarin 2018



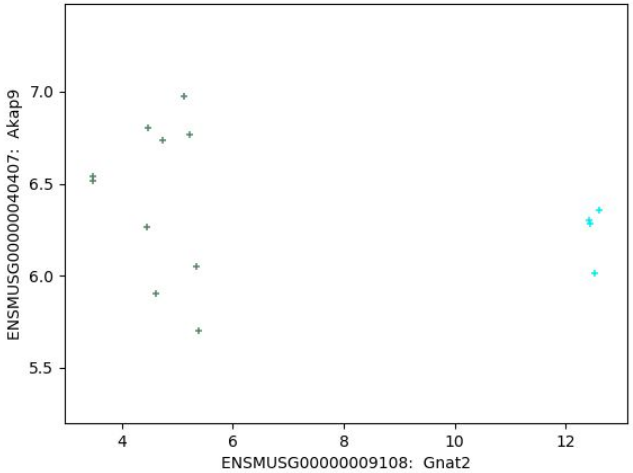
Hartl 2017



UC San Diego
Boolean Lab

Boolean Implication Refutes High Confidence SRCCA Gene AKAP9

- AKAP9, a novel candidate cone gene, was identified by Phillips et al., but not by boolean implication.
- Validation in Hartl 2017 dataset shows that it is not cone-specific.
- While identification of high confidence genes from SRCCA may be arbitrary, boolean implication can lend insight into the importance of the gene in determining cell fate.

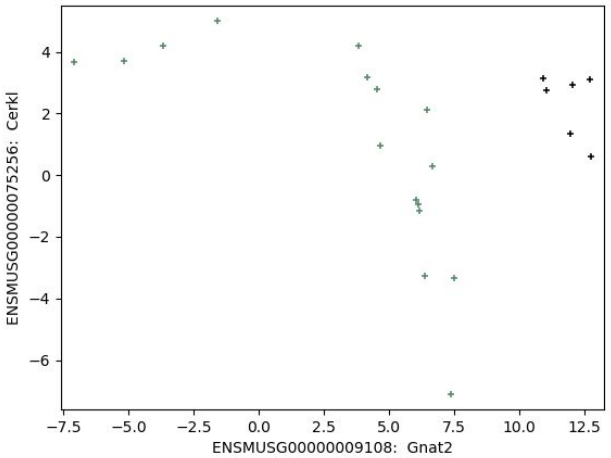


UC San Diego
Boolean Lab

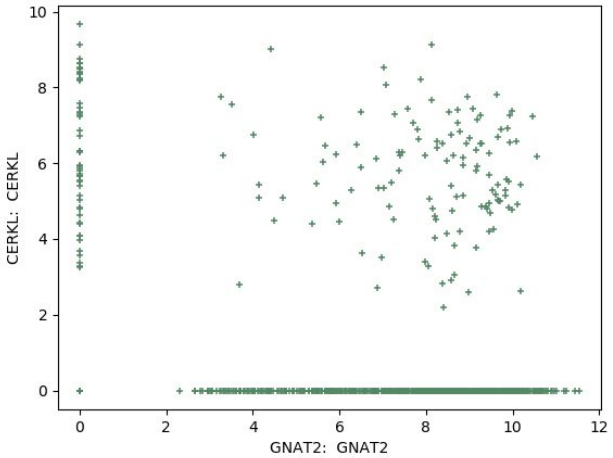
Differences in Human and Mouse Retina

- This method of quantification using purified retinal cell types from Mus musculus may have errors due to differences between species.
- In depth analysis allowed identification of CERKL, a gene specific to cones in humans but more general in mouse retina.

Mouse (Sarin 2018)



Human (Kim 2016)



UC San Diego
Boolean Lab

Conclusion

- Boolean implication improved upon correlational methods by filtering out noise and identifying asymmetric relationships that lend insight into the specificity of genes.
- Filtering correlating genes using boolean implication led to a statistically significant improvement in cell type-specificity for rod photoreceptor genes, and reproducibility for cone photoreceptor genes.
- Difference between results in rods and cones suggests that the Phillips dataset may not be comprehensive enough to provide accurate distinction between rods and cones.
- Boolean implication can be used to give a more accurate insight into “high confidence” genes, and lead to identification of novel markers of retinal cell types such as WWC1.
- Boolean implication offers all the advantages presented by Phillips et al. 2018 for SRCCA, including efficiency, ability to combine multiple bait genes, and improved prediction accuracy.

UC San Diego
Boolean Lab

Further research

- We have not completed the quantification for the remaining 3 cell types.
- However, it is less likely that there is an improvement due to smaller number of cells present, hence more noise.
- Application of similar methods in other larger retina datasets may lend greater insight into novel markers of retinal cell types.

UC San Diego
Boolean Lab

Thank You!

Questions?

UC San Diego
Boolean Lab