Geometry of gene expression in retina of mouse and macaque

Markus Meister[1]
[1]Division of Biology and Biological Engineering
California Institute of Technology
{meister}@caltech.edu

July 22, 2023

1 Summary

This is an analysis of gene expression patterns among retinal ganglion cells of the mouse[3] and the macaque[2]. The immediate goal is to test the correspondence of certain ganglion cell types in the mouse with the canonical midget and parasol types in the macaque. The analysis builds on the report about evolution of retinal cell types by Hahn et al. [1]. That report proposes that the mouse Alpha retinal ganglion cells are orthologous to the macaque midget and parasol types, based on their gene expression patterns. Here I present an alternative analysis method that ultimately reaches the same conclusion with a high degree of confidence. The analysis also offers interesting windows on the geometry of gene expression among retinal ganglion cells.

2 Methods summary

2.1 Source of data

The starting materials are the gene expression matrices for retinal ganglion cells published by Tran et al. [3] (mouse, 35699 cells) and Peng et al. [2] (macaque, 28849 cells in fovea, 10333 cells in periphery). I merged the two data sets using a list of orthologous genes between mouse and macaque from Ensembl, which yielded $N_{\mathrm{genes}} = 10416$ common genes. The cells fall into 53 different types: the 45 mouse RGC types identified in Tran et al. [3] and the 8 midget and parasol types in the foveal and peripheral region identified in Peng et al. [2].

2.2 Selection of genes

For each gene and each type, I computed the mean and SEM of the expression level across cells of that type, and combined those into the signal-to-noise ratio (SNR): this is the variance of expression across the 53 types divided by the variance within those types. Genes were selected for analysis in order of decreasing SNR. About 50% of the genes had SNR>13, and 10% had SNR>100. Results reported here are based on genes with SNR>200, which yields about 400 genes. Owing to the high SNR, the cluster center of each cell type is very well defined: The standard error of the mean is a small fraction of the separation between the centers. In subsequent analysis I therefore ignore the variation of gene expression within clusters, and focus entirely on the cluster centers.

2.3 Dimension reduction

The N_{types} cluster centers live in an $(N_{\mathrm{types}} - 1)$-dimensional space. Dimensions outside of that subspace are uninformative, because the cluster centers don't vary along those dimensions. So one can project the data into this subspace without loss of information. This results in a $N_{\mathrm{types}} \times N_{\mathrm{features}}$ data matrix, where $N_{\mathrm{types}} = 53$ and $N_{\mathrm{features}} = N_{\mathrm{types}} - 1 = 52$. To reduce the co-variation between these features, I further reduced the dimensionality with a PCA step to N_{dims} dimensions.

Preprint. July 22, 2023.

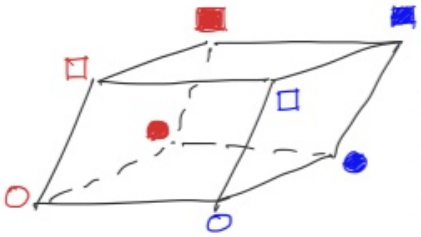

Figure 1: Sketch of a $2 \times 2 \times 2$ Cartesian model for the geometry in gene expression space among RGC types of the macaque. Symbols: red=fovea, blue=periphery; square=parasol, circle=midget; open=On, closed=Off.

2.4 Geometric model

Suppose we want to evaluate the correspondence between 4 mouse RGC types and the 4 macaque midget and parasol types. The cluster centers for the macaque types form some 4-cornered shape in the space of gene expression values. Similarly the 4 mouse types form a 4-cornered shape. I ask how well the mouse shape matches the macaque shape. I fit the data with a model that assumes the same shape, except for a translation in gene expression space. Then I score the fit by how much of the data variance the model explains within each species. Variants of the model include additional constraints, for example that the shape should be a parallelogram, or that the same shape should match RGCs in the macaque fovea and in the macaque periphery. For mathematical details, see 6.1.

2.5 Investigator's degrees of freedom

There are only two choices to make in this analysis: (1) how many genes to include, N_{genes}; (2) how many principal components to use for the reduced space to which the geometric model is applied, N_{dims}. Neither of these choices introduces a bias favoring any particular geometric arrangement among the cell types. Still, one needs to check that any interesting claims are robust to these parameters. In most cases I will show analysis results averaged across a sweep of parameters.

3 Geometric analysis of macaque RGCs

Let's start with the 8 primate RGC types. They form a $2 \times 2 \times 2$ Cartesian product of traits: On vs Off, P vs M, fovea vs periphery. [1] The 8 types make eight points in the feature space, and we want to interpret their geometric arrangement, i.e. the shape that they form. We will propose a model for that shape, fit the model to the actual points, and ask how much of the shape the model captures.

Perhaps the simplest model one could propose in this case is that each of the 3 phenotypic features maps onto a specific displacement vector in feature space (Fig 1). For example, the 4 Off cells are displaced from the corresponding 4 On cells by the same vector; the peripheral cells are displaced from the foveal cells by another vector; and the midget cells are displaced from the parasol cells by another vector. That means the 8 points should lie at the corners of a parallelepiped. This is a strong constraint on the shape in gene space, and one might think this outcome unlikely a priori, but it turns out to work pretty well.

Figure 2A shows a result from this fit, projected down into the 3-dimensional space of the best-fit parallelepiped so we can plot and examine it. The wire frame is the optimal parallelepiped to fit the 8 data points. Each data point is tied (dotted line) to its corresponding corner of the model. Note that the points are nicely arranged in line with the prediction: the foveal cells are opposite the peripheral ones, the on cells are opposite the off cells, and the sustained cells are opposite the transient cells. So the model comes reasonably close to the actual geometry of the cell types in gene space and captures about 86% of the variance in the data.

Nonetheless, the model isn't perfect, and it is instructive to look at where the data deviate from the fit. Figure 2C looks at the same shape along the direction of the On-Off axis. Here we see that the

[1] I found that the peripheral and foveal cells are far apart in gene space, separated by a greater distance than On vs Off cells, or parasols vs midgets. So we can really treat them as separate cell types.

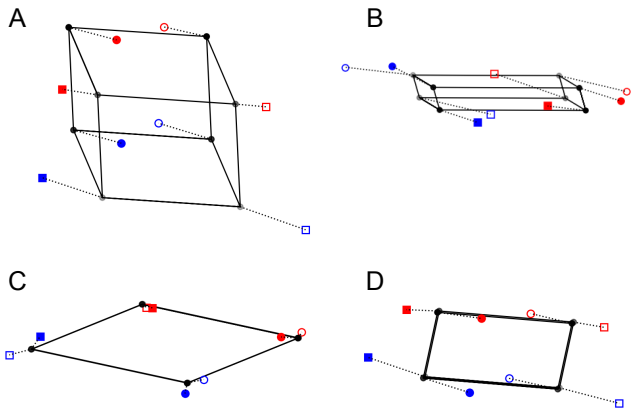

Figure 2: **A:** A $2 \times 2 \times 2$ Cartesian shape fitted to the actual geometry among RGC types of the macaque. Symbols: red=fovea, blue=periphery; square=parasol, circle=midget; open=On, closed=Off. Parameters: $N_{\text{genes}} = 200, N_{\text{dims}} = 12$. **B:** The result of the fit if the parasol and midget cell labels are mistakenly swapped among peripheral cells. Note the model shape is now much thinner along the parasol-vs-midget axis, and explains less of the variance in the data. **C:** A view of the shape in panel A along the On-vs-Off axis. Note that the fovea-periphery and P-M traits follow the Cartesian model well: in those 2 dimensions the data form a nice parallelogram. **D:** A view of the shape in panel A along the P-vs-M axis. Note the On-Off trait does not fit as nicely into the Cartesian model: The midget On and Offs are much closer together than the parasol On and Offs.

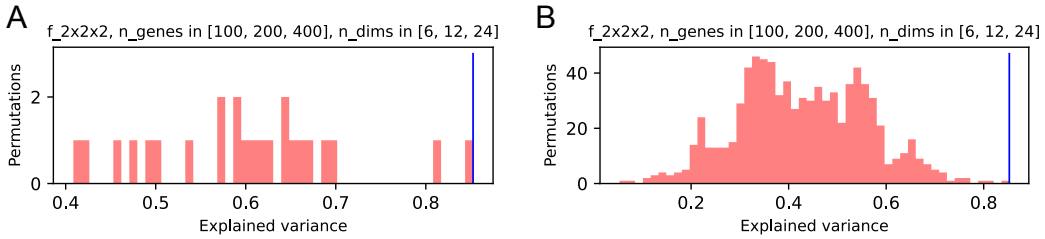

Figure 3: A $2 \times 2 \times 2$ Cartesian model for the geometry among RGC types in the macaque serves to identify the correct labeling of cell types. **A**: Here the labels on the four peripheral types were permuted in 24 different ways, and the model fit applied to each case. Histogram shows the resulting explained variance from each of these fits. The value for the correct labeling is marked in blue. **B**: As in (A), but the labels on all 8 RGC types were permuted in 840 different ways. Again the correct labeling is marked in blue. Average scores over a sweep of the parameters $N_{\text{genes}} \in \{100, 200, 400\}, N_{\text{dims}} \in \{6, 12, 24\}$.

traits for fovea-periphery (red-blue) and for parasol-midget (square-circle) form a beautiful Cartesian shape, namely a parallelogram. The parasol-midget separation vector in the fovea is almost exactly the same as in the periphery. Similarly the fovea-periphery separation vector among midgets is almost exactly the same as among parasol cells.

Figure 2D looks at the same shape along the parasol-midget axis. Now we see that the On-Off trait doesn't fit into the Cartesian prediction as nicely: The On-Off separation vector among midgets is very different from that among parasols. In particular, midget Ons and Offs are much closer in gene expression than parasol Ons and Offs.

More things could be learned from inspecting the arrangement of macaque types, but let's move closer to an inter-species comparison. We want to see how effective the Cartesian model is in identifying the correct labeling among macaque cells. First, suppose we had inadvertently swapped the parasol and midget labels on the peripheral cells. Now the model can't match the data very well (Fig 2B). The optimal parallelepiped is squashed along the parasol-midget axis, and it captures only 54% of the variance. So a bad correspondence of the type labels leads to a lower score of explained variance.

Let's pretend now that the 4 peripheral types came from a different species, and we didn't know anything about the correspondence of those types with the 4 macaque fovea types. Could we identify the correct correspondence uniquely just from the gene expression patterns? There are $4! = 24$ ways of labeling the peripheral types. Figure 3A shows the explained variance for each of these arrangements. The correct one does indeed have the highest score, so the analysis can reliably identify the 4 RGC types in the periphery that correspond to those in the fovea.

One can go further and inspect all possible labelings of all 8 cell types. There are 840 permutations among the corners of a cube. Figure 3B shows the resulting histogram of explained variance scores. The correct labeling has the largest score and can be reliably recognized among all 840. These results are robust to the choice of parameters in the analysis, and the figures quote the average over many parameter settings.

Further analysis showed that the correct assignment of the 8 macaque types can be found using just the $N_{\mathrm{genes}} = 6$ genes with highest SNR, out of more than 10000! This may be worth further attention at some other time...

For now, the above results derived from the 8 macaque types provide an important positive control. In this situation, we know the ground truth for a correct match between foveal and peripheral RGC types. And the analysis method can identify that correct match among 840 alternatives based purely on the geometry of gene expression patterns. This gives some confidence that the method will work in a situation where the ground truth is not known.

4 Correspondence between mouse and macaque RGCs

The next step is to replace the macaque peripheral RGCs in the above analysis with mouse RGCs. Then we can try different combinations of mouse cell types and score each match by explained variance. Hahn et al. [1] compiled four groups of mouse cell types that could plausibly correspond to the four macaque types, based on their visual response properties (Fig 5). By selecting one from each group one can form 432 candidate combinations of four mouse RGC types. I will pair each of these mouse combinations with the 4 macaque types and apply the $2 \times 2 \times 2$ Cartesian model to the geometry in gene expression space. To evaluate the fit, I ask how well the model explains the variance of the data within each species (see 6.1).

Fig 4A shows the result. The histogram of explained variance scores covers all combinations of cells tested. The blue ticks indicate combinations that include Alpha cells, and the size of the tick (from 1 to 4) says how many there are. The top 5 scores all go to combinations that include either 3 or 4 Alpha cells. In general, one sees a trend that more Alpha types in the group of four mouse types leads to higher scores (tick marks get taller left to right). The results are robust to the analysis parameters; in fact, the figure shows the average over a sweep of conditions.

In the following step I relax the model somewhat from the 2x2x2 Cartesian structure to a 4x2 structure. Now we allow the 4 types within a species to take on any desired shape, rather than being restricted to a parallelogram. But we ask that the corresponding types in the other species make the same shape. The two shapes are simply shifted versions of each other (Fig 4B Left). Because our focus here is on the correspondence across species, we should not impose a strong expectation on the geometry within a species. Now the explained variance is considerably higher (Fig 4B Right), because the model can capture more of the geometry within a species. The top score belongs to the combination of 4 Alpha cells. Note this score is separated from the rest of the distribution by a wide margin. Several groups with 3 Alpha cells also lie above the bulk of the distribution. Again there is a systematic trend where more Alpha cells in the group lead to higher scores.

One can further extend this model to a 4x3 structure that matches the mouse with both the fovea and the periphery in the macaque. Again, we allow the 4 types in each group (mouse, fovea, and periphery) to adopt any desired geometry, and shifted versions of that same shape should account for all 12 types (Fig 4C Left). As before we test the 432 curated combinations of mouse types (Fig 4C Right). The resulting scores are somewhat lower, because the model now needs to fit an additional 4 data points. However, as before, the all-Alpha combination of mouse types wins by a large margin.

Finally, we return to the 4x2 model structure, but go beyond the 432 combinations of mouse types that were curated ahead of time based on visual response properties. Instead, we consider all possible combinations of 4 mouse types, of which there are $45 \times 44 \times 43 \times 42 = 3575880$. This yields a

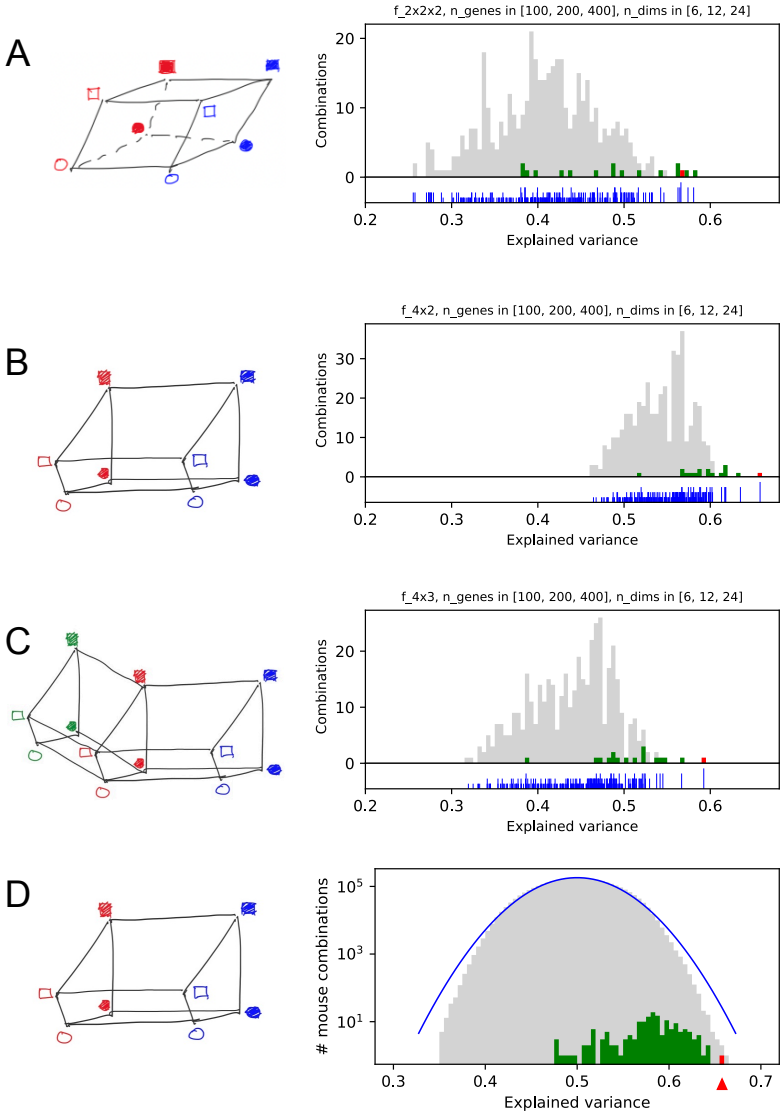

Figure 4: **A, Left:** a $2 \times 2 \times 2$ Cartesian model for the geometry among RGC types in the mouse and the macaque fovea. Symbols: red=macaque fovea, blue=mouse; square=transient, circle=sustained; open=On, closed=Off. **A, Right:** Histogram of the intra-species variance explained by this model for each of 432 combinations among 4 mouse types, curated to have plausible visual response phenotypes; green: combinations that include 3 Alpha cell types; red: the lone combination with all 4 Alpha types. Bottom: blue ticks mark combinations that include Alpha cells, tick length (1-4) indicates number of Alphas. **B:** As in (A) but for a 4×2 Cartesian model. **C:** As in (A) but for a 4×3 Cartesian model linking RGC types in the mouse, macaque fovea and macaque periphery. **D:** As in (B) but considering all 3575880 4-type combinations drawn from the 45 mouse types. Note logarithmic y-axis. Blue line: Gaussian distribution with the same mean and variance. Red arrowhead marks the combination with all 4 Alpha types. All panels show the average result over a sweep of the parameters $N_{\text{genes}} \in \{100, 200, 400\}, N_{\text{dims}} \in \{6, 12, 24\}$.

broad distribution of scores with approximately Gaussian shape (Fig 4D). The combination with 4 Alpha types in the correct order lies at the far right edge of this distribution. In fact only one other combination achieves a slightly higher score.

5 Interpretation

Based on the geometry of gene expression patterns, this analysis suggests that the Alpha cells in the mouse retina are a unique match with the macaque parasol and midget cells (Fig 4). The match extends to both fovea and periphery regions of the macaque retina (Fig 4C). Among the 432 combinations of mouse RGC types that have plausible visual response phenotypes, no other combination comes even close to matching the score of the Alpha cells (Fig 4B-C). When considering all possible combinations of mouse types, the Alpha cells beat all others with one exception. The null hypothesis – namely that the all-Alpha combination gets this high score by accident – has a *p*-value of $< 10^{-6}$.

6 Methods

6.1 How to fit a Cartesian model to the data

Here is the formalism for a 3D model, e.g. $2 \times 2 \times 2$.

Gene expression vectors:

$$\vec{x}_{i,j,k}, \quad i \in \{1, \ldots, L\}, j \in \{1, \ldots, M\}, k \in \{1, \ldots, N\}$$

where i, j, k index the 3 phenotypic dimensions, for example "transient/sustained", "on/off", "macaque/mouse".

Fit:

$$\vec{y}_{i,j,k} = \vec{a} + \vec{b}_i + \vec{c}_j + \vec{d}_k$$

The model parameters are the vectors $\vec{a}, \{\vec{b}_i\}, \{\vec{c}_j\}, \{\vec{d}_k\}$. For uniqueness we require that

$$\sum_i \vec{b}_i = \sum_j \vec{c}_j = \sum_k \vec{d}_k = 0$$

The squared residual

$$\chi^2 = \sum_{i,j,k} (\vec{x}_{i,j,k} - \vec{y}_{i,j,k})^2$$

is minimized by this solution:

$$\begin{aligned}
\vec{a} &= \langle \vec{x}_{i,j,k} \rangle_{i,j,k} = \frac{1}{LMN} \sum_{i,j,k} \vec{x}_{i,j,k} \\
\vec{b}_i &= \langle \vec{x}_{i,j,k} - \vec{a} \rangle_{j,k} = \frac{1}{MN} \sum_{j,k} (\vec{x}_{i,j,k} - \vec{a}) \\
\vec{c}_j &= \langle \vec{x}_{i,j,k} - \vec{a} \rangle_{i,k} = \frac{1}{LN} \sum_{i,k} (\vec{x}_{i,j,k} - \vec{a}) \\
\vec{d}_k &= \langle \vec{x}_{i,j,k} - \vec{a} \rangle_{i,j} = \frac{1}{LM} \sum_{i,j} (\vec{x}_{i,j,k} - \vec{a})
\end{aligned}$$

The total variance in the data is

$$V_{\text{tot}} = \frac{1}{LMN} \sum_{i,j,k} \left(\vec{x}_{i,j,k} - \vec{a}\right)^2$$

and the variance explained by the model is

$$V_{\text{exp}} = \left\langle \left(\vec{y}_{i,j,k} - \vec{a}\right)^2 \right\rangle_{i,j,k} = \frac{1}{LMN} \sum_{i,j,k} \left(\vec{b}_i + \vec{c}_j + \vec{d}_k\right)^2$$

I score the fit by the fraction of variance explained, V_{exp}/V.

For comparisons across species, we are interested only in the variance explained within each species, i.e. the degree to which a common shape explains gene expression patterns in both species. Suppose that the last index $k \in \{1, 2\}$ denotes macaque vs mouse, then the mean expression vector within each species is

$$\vec{a}_k = \langle \vec{x}_{i,j,k} \rangle_{i,j} = \frac{1}{LM} \sum_{i,j} \vec{x}_{i,j,k}$$

and the intra-species variance of the data is

$$V^{(\text{s})} = \frac{1}{LM} \sum_{i,j,k} \left(\vec{x}_{i,j,k} - \vec{a}_k\right)^2$$

The intraspecies variance explained by the model is

$$V^{(\text{s})}_{\text{exp}} = \left\langle \left(\vec{y}_{i,j,k} - \vec{a}_k\right)^2 \right\rangle_{i,j,k} = \frac{1}{LM} \sum_{i,j} \left(\vec{b}_i + \vec{c}_j\right)^2$$

The fit score in that case is the fraction of explained intra-species variance, $V^{(\text{s})}_{\text{exp}}/V^{(\text{s})}$.

The solution generalizes in obvious ways to lower or higher dimensions. For the model structures used in the analysis so far:

$2 \times 2 \times 2$: This is a 3D model with $L = 2, M = 2, N = 2$. It assumes that the shape within each species is a parallelogram.

4×2: This is a 2D model with $L = 4, M = 2$. There are no $\vec{d}_k$ vectors. Here the shape within each species is an arbitrary tetrahedron.

4×3: This is a 2D model with $L = 4, M = 3$. Again the shape within each species is an arbitrary tetrahedron, but it is translated across 3 "species": mouse, macaque fovea, macaque periphery.

6.2 Candidate mouse ganglion cell types

The following table shows the candidate RGC types in mouse retina considered for possible matches to the macaque parasol and midget types. They are grouped by their visual response properties. This is Supplementary Table 2 from Hahn et al. [1].

Mouse RGC candidates

ON-sustained	ON-transient	OFF-sustained	OFF-transient
C8 = PixON C10= ON DS C14= ON-delayed C18 = ON bursty C27 = ONhOS SmRF C31 and/or C22 = M2 C36 = ONhOS LgRF C40 and/or 33 = M1 C43 = ONsusAlpha	C3 = F-mini-ON C30 = ONtrSmRF C38 = F-midi-ON C41 = ONtrAlpha	C5 = OFFvOS/JAMB C9 = OFFhOS C28 = F-midi-OFF C42 = OFFsusAlpha	C4 = F-mini-OFF C21 = OFFtrSmRF C45 = OFFtrAlpha

Figure 5: Candidate ganglion cell types in mouse retina considered for possible matches to the macaque parasol and midget types. From Hahn et al. [1], Supplementary Table 2.

References

[1] Hahn, J., Monavarfeshani, A., Qiao, M., Kao, A., Kölsch, Y., Kumar, A., Kunze, V. P., Rasys, A. M., Richardson, R., Baier, H., Lucas, R. J., Li, W., Meister, M., Trachtenberg, J. T., Yan, W., Peng, Y.-R., Sanes, J. R., and Shekhar, K. (2023). Evolution of neuronal cell classes and types in the vertebrate retina.

[2] Peng, Y.-R., Shekhar, K., Yan, W., Herrmann, D., Sappington, A., Bryman, G. S., van Zyl, T., Do, M. T. H., Regev, A., and Sanes, J. R. (2019). Molecular Classification and Comparative Taxonomics of Foveal and Peripheral Cells in Primate Retina. *Cell*.

[3] Tran, N. M., Shekhar, K., Whitney, I. E., Jacobi, A., Benhar, I., Hong, G., Yan, W., Adiconis, X., Arnold, M. E., Lee, J. M., Levin, J. Z., Lin, D., Wang, C., Lieber, C. M., Regev, A., He, Z., and Sanes, J. R. (2019). Single-Cell Profiles of Retinal Ganglion Cells Differing in Resilience to Injury Reveal Neuroprotective Genes. *Neuron*, 104(6):1039–1055.e12.