Retroactive analysis of single cell transcriptome profiles using next-generation algorithms revised the identification of several resilient retinal ganglion cell types

Bruce A. Rheaume[1] and Ephraim F. Trakhtenberg[1]*.

[1]Department of Neuroscience, University of Connecticut School of Medicine, 263 Farmington Ave., Farmington, CT, 06030, USA.

*Correspondence to: trakhtenberg@uchc.edu

ABSTRACT

Mammalian central nervous system (CNS) projection neurons do not regenerate axons damaged by an injury or disease, and certain CNS neurons, such as retinal ganglion cells (RGCs), eventually die by retrograde degeneration. Recently, RGC types that are more resilient or susceptible to death after axonal injury were identified. Investigating the differences between these RGC types could help identify neuroprotective and risk genes for therapeutic targeting. Therefore, accurate identification of the resilient or susceptible RGC types is critical, not only for advancing fundamental neurobiological knowledge, but also for guiding the studies which rely on this knowledge. Here, using a next-generation integration algorithm for analyzing single cell transcriptome profiles, we were able to identify the types of injured RGCs that the original algorithm could not identify. Unexpectedly, the majority of these cells originated from only a few RGC types, and consequently the re-analysis resulted in a substantially different determination of the resilient and susceptible RGC types. We also provide a website for comparing the gene expression in atlas RGC clusters to the re-analyzed injured RGC clusters. Additional bioinformatics analyses also revealed singe gene cluster markers for RGC types and contributed new insights into the global characteristics of RGC types and how axonal injury affects them, showing how the extent of variance between transcriptomes of RGC types increases differentially during maturation and after injury, and that specific global properties of the transcriptome are associated with several resilient RGC types.

INTRODUCTION

The failure of mammalian CNS projection neurons to spontaneously regenerate injured axons limits the ability to restore functions lost due to injury or neurodegenerative diseases that damage the axons in CNS white matter tracts and nerves. Optic nerve crush (ONC) injury of the mouse optic nerve (CNS white matter) models human traumatic optic neuropathy and is used to study the failure of RGCs (a type of CNS projection neuron) to regenerate damaged axons through the optic nerve[1-5]. Mouse RGCs are comprised of 40-46 types[6-10] (sometimes referred to as subtypes[7, 11-13]), with some types surviving longer after ONC injury or being more prone to regenerating axons[6, 14, 15]. The most comprehensive study of these resilient RGC types used single cell RNA sequencing (scRNA-seq) to determine the prevalence of the surviving RGC types[6]. Because injury changes gene expression in the RGCs, the authors developed the iGraphBoost algorithm for tracing injured RGC transcriptomes to their respective RGC types[6]. However, the algorithm's limitations precluded it from determining the type of 17.2% of the injured RGCs 2 weeks after ONC (a standard time-point in this assay). Therefore, the prevalence analysis of the identified injured RGC types relied on the assumption that specific types were not overrepresented amongst those injured RGCs which were unassigned to types. Here, using a recently developed next-generation integration algorithm[16, 17] in retroactive analysis, we determined the types of all the injured RGCs (at 2 weeks after ONC) and found that some types were significantly overrepresented amongst the RGCs which were unassigned to types by the original iGraphBoost algorithm. Consequently, re-analysis of survivability after assigning all the RGCs to types resulted in a substantially different determination of the resilient and susceptible types, reported herein. A number of studies rely on the accuracy of RGC type identification as resilient or susceptible to investigate the differences in their underlying biology and to identify neuroprotective and risk genes. Therefore, an update of the relevant conclusions is important, particularly in the context of new insights into how axonal injury affects the global characteristics of RGC types and improved identification of RGC type markers, revealed here by further analysis.

RESULTS

Classification of the uninjured adult mouse RGC scRNA-seq-derived transcriptome profiles (adult RGC atlas) by the Louvain-Jaccard algorithm[18, 19] predicted 45 clusters, one of which was manually sub-clustered into 2, resulting in a total of 46 types[6], that closely matched the 40 neonatal RGC types we identified previously[7] (see supplementary Fig. 1F in Tran et al., 2019[6]). However, the iGraphBoost algorithm was able to trace cluster origin of only 82.8% of the injured RGCs (that survived 2 weeks after ONC) to the uninjured RGC types, despite the data from earlier time-points (i.e., before 2 weeks after ONC) enabling sequential tracing of RGC types, as their

transcriptomes were fluctuating over time after injury (see Fig. 3C in Tran et al., 2019[6]). The resilience or susceptibility of RGC types was, in turn, determined by the prevalence of the types remaining 2 weeks after injury compared to the prevalence of their uninjured counterpart types (see Fig. 3E in Tran et al., 2019[6]). The implicit assumption of such an analysis is that specific RGC types were not overrepresented amongst the unaccounted 17.2% of injured RGCs; otherwise, such types would have been deemed susceptible when in fact they are resilient. Because accurate determination of the resilient and susceptible RGC types depends on this assumption, we used next-generation algorithms to probe it.

To test whether the failure to trace the cluster origin of 17.2% of the injured RGCs resulted from the limitations of the first-generation integration algorithm, rather than due to biological changes that de-differentiated certain RGCs after injury beyond bioinformatic traceability to types, we re-analyzed the data using the next-generation Seurat algorithm for tracing cell type-origin[16, 17]. The Seurat integration algorithm uses high-dimensional data for calculating the integration anchors used for mapping the query cells, but it also uses uniform manifold approximation and projection (UMAP)[20, 21] for transferring the reference cluster labels of nearest neighbors (that are determined, in part, by the *k*-weight parameter; see below). UMAP yielded a similar distribution of the Louvain Jaccard-determined clusters (**Figure 1A**) as did the scalable implementation of t-distributed stochastic neighbor embedding (tSNE)[22, 23] that was used originally (see Fig. 1C in Tran et al., 2019[6]). Next, applying the Seurat integration algorithm using UMAP and an optimized k-weight parameter, we found that all the RGCs (at 2 weeks after ONC) were successfully traced to their type-origin (**Figure 1B-C**). The k-weight parameter is the weight assigned to clusters based on size (number of cells predicted to comprise it) and distance in high-dimensional space from other clusters, in determining cell-to-cluster assignment. To determine an optimal k-weight parameter, we used a loop-script (see Methods) that iterates through the Seurat functions' k-weight parameters and outputs the average confidence score (ranging 0 to 1, with 1 being the highest) of all cell-to-cluster assignments. We then ranked average confidence scores from all the runs, and found that the highest mean confidence score (with the k-weight parameter set to 6) was 1.44-fold higher compared to the mean score using the default k-weight parameter (**Figure 1D**). We also found that using the Seurat algorithm, the correlation between the uninjured and the injured RGC clusters (traced to respective type-origin) was higher, for the analyzed post-injury time-points, compared to using iGraphBoost (**Supplemental Figure 1A**). Moreover, the RGCs from the latest time-point (2 weeks after ONC) correlated with the uninjured RGCs more than the RGCs from two preceding time-points (4 and 7 days after ONC; **Supplemental Figure 1B**), suggesting that the transcriptome of the surviving RGCs reverts, over time, towards an uninjured state. Therefore, using the sequential (i.e., iterative integration) of iGraphBoost tracing was not advantageous for the 2-week post-ONC time-point. Accordingly, during this period (4 and 7 days after ONC) there is a peak in the percentage of injured RGCs unassigned to clusters by the iGraphBoost algorithm, compared to a smaller percent of RGCs unassigned at 2 weeks after ONC (**Supplemental Figure 1C**).

Next, we analyzed whether certain RGC types were overrepresented amongst the originally unaccounted injured RGCs, and found that types C3, C1, C18, C42, C21, and C45 were substantially (>25% each) overrepresented, accounting for the majority (60.4%) of the unassigned cells (**Figure 1E-G**). Then, we analyzed the prevalence of the surviving RGC types, with all injured RGCs now assigned to types, and found that determination of several resilient and susceptible RGC types changed compared to their original ranking (see Fig. 3E in Tran et al., 2019[6]), including the susceptible types C45 and C3 now ranking as resilient, the resilient types C39 and C7 now ranking as susceptible, and the most abundant type, C1, which was previously considered highly susceptible, now ranking as mildly susceptible (**Figure 2A-C**). We also developed the Subtypes Gene Browser web application (in the same format we previously presented for neonatal RGC subtypes[7]), which provides a platform for comparing gene expression in atlas RGC clusters to the re-analyzed injured RGC clusters (https://health.uconn.edu/neuroregeneration-lab/subtypes-gene-browser; see Methods for details).

Then, we analyzed whether, at 2 weeks after ONC, the injured RGCs originally unassigned to types by the iGraphBoost algorithm were substantially different from the injured RGCs that were assigned to types. We did not find the cell-to-cluster confidence scores for the originally unassigned injured RGCs (mean = 0.74) to be substantially different from the assigned RGCs (mean = 0.81). The transcriptomes of the originally unassigned overrepresented injured RGC types, not only were not amongst the types most or least changed by injury (**Supplemental Figure 2A**), but were highly correlated ($r = 0.99$, $p < 0.0001$) to their assigned counterpart types (**Supplemental Figure 2B**), highlighting the accuracy of their assignment to clusters in retrospect by the Seurat algorithm. We also did not find the relative proportion of mitochondria-related genes (which is typically higher in cells progressing towards apoptosis[16, 24, 25]) in the overrepresented unassigned RGC types to be amongst the highest or lowest relative to the assigned types (**Supplemental Figure 3A**), nor was the total number of genes

detected in the originally unassigned injured RGC types amongst the highest or lowest relative to the assigned types (**Supplemental Figure 3B-C**). The overrepresented unassigned RGC types also were not amongst the least or most abundant types in the RGC atlas (**Supplemental Figure 3D**). These data suggest that the injured RGCs unassigned to types by the iGraphBoost algorithm were not substantially different from other injured RGCs, but rather the limitations of the first-generation integration algorithm precluded their tracing to respective type-origin.

Next, we asked whether applying the algorithm we developed for determining unique single gene markers for individual RGC clusters[7] would identify more single gene markers enriched in the adult RGC clusters. In our classification of neonatal RGC types, we identified several enriched single gene markers for each of the 40 clusters[7]. However, out of the 45 adult RGC clusters, 32 required combinations of enriched/unenriched genes to mark them uniquely, as single gene cluster markers were identified for only 13 clusters (see Fig. 1F and Supplementary Fig. 1F in Tran et al., 2019[6]). Of the 13 single gene cluster markers, 9 were amongst the predicted neonatal RGC single gene markers (see Fig. 4 and Supplementary Data/Table 1 in Rheaume et al., 2018[7]; **Supplemental Table 1**). Of the remaining 4 genes, Ceacam10, Nmb, and Slc17a7 were developmentally upregulated from neonatal to adult RGCs (thus not sufficiently enriched in neonatal RGC clusters), and Rhox5 was enriched in several clusters of neonatal RGCs (and thus did not qualify as a neonatal cluster marker; **Supplemental Figure 4**) but was downregulated during maturation in these clusters except for in cluster C32 (for which Rhox5 became a marker in adult RGCs). Overall, neonatal clusters closely matched adult clusters, with an exception of 2 neonatal clusters converging into 1 adult cluster in 3 instances, and the 5 clusters added in adult (see Supplementary Fig. 1F in Tran et al., 2019[6]). Therefore, we applied the algorithm we used for identifying single gene markers of neonatal RGC clusters[7], and found single gene markers uniquely enriched in all but one (C8) adult atlas RGC clusters (**Figure 3** and **Supplemental Data 1**). At least one single gene marker from 37 adult atlas RGC clusters was also amongst single gene markers from 31 neonatal RGC clusters, and 18.1% of these marker genes have been previously validated, including Pde1a, Zic1, and Kcnip2[6, 7] (**Supplemental Table 2**; **Supplemental Figure 5A-C**). Because we did not find a single gene marker for cluster C8, we applied an algorithm we used[7] for identifying a combination of genes uniquely co-enriched in a cluster, and found that Prkcq and Tac1 are co-enriched only in cluster C8 (**Supplemental Figure 5D-E**); however, as Tac1 is also modestly co-expressed with Prkcq in cluster C43, including Syt6 provides a more robust 3-way combination to mark cluster C8 (**Supplemental Figure 5F**).

To gain insight into the global transcriptome dynamics of RGC types, we analyzed whether the extent of variability between RGC types changes during maturation and after injury. We found that variability between the transcriptomes of adult RGC types is within the 0.92-0.98 average *r* range (Pearson; **Figure 4A**), which is twice the variability we found for pre-eye-opening neonatal RGC types (0.96-0.99 average *r* range[7]). Furthermore, the extent of variability between the transcriptomes of RGC types more than doubled after injury to 0.73-0.90 average *r* range (Pearson; **Figure 4B**). However, there was no significant correlation between survivability and the mean transcriptome size of uninjured ($r = 0.05$, $p = 0.74$) or injured ($r = 0.22$, $p = 0.16$) RGC types (**Supplemental Figure 3B-C**), nor the relative proportion of mitochondria-related genes (see above) of injured types ($r = -0.21$, $p = 0.17$; **Supplemental Figure 3A**). We also showed that the varying proportion of highly expressed (>1 *NE*) genes[7, 26] remained significantly correlated between uninjured and injured RGC types ($r = 0.65$, $p < 0.001$; **Figure 4C-F**), but it was not correlated to RGC survivability ($r = -0.05$, $p = 0.73$), suggesting that this varying global property of RGC type transcriptome is largely unaffected by injury regardless of susceptibility or resilience of an RGC type. Finally, we found that the variability between the transcriptomes of atlas adult RGC types (average cluster *r* values in **Figure 4A**) significantly correlated to survivability of injured RGC types ($r = -0.37$, $p < 0.02$; **Figure 4G**), and that the types with the most influential regression residuals accounting for correlation are the resilient types C42, C43, and C45 belonging to the alpha/intrinsically-photosensitive RGCs (ipRGCs) class (see Table S2 in Tran et al., 2019[6]; **Figure 4G-H**).

DISCUSSION

Our retroactive analysis of the scRNA-seq-derived transcriptome profiles of the injured RGCs using a next-generation integration algorithm[16, 17] accounted for those injured RGCs which were unassigned to types by the original algorithm. We found that the unassigned RGCs were not substantially different from other injured RGCs on any of the global transcriptome properties, and belonged primarily to types clusters C3, C1, C18, C42, C21, and C45. Consequently, the re-analysis of the RGC types survivability led to a different identification of the resilient and susceptible RGC types than was originally reported[6].

We also interpret the observation that the transcriptomes of uninjured atlas RGCs are more similar to RGCs at 2 weeks after ONC than at 4-7 days after ONC, as evidence that by 2 weeks after ONC, the transcriptomes of the surviving RGCs attempt to recover towards an uninjured state. Furthermore, the retrograde RGC degeneration and death that occurs weeks/months later would likely be associated with another peak of changes in the RGC transcriptome.

Next, using an algorithm we developed previously[7], we identified single gene markers uniquely enriched in all but one of the adult atlas RGC clusters[6], and found that many of these markers were also uniquely enriched in the developing neonatal atlas RGC clusters[7]. Importantly, 18.1% of these marker genes have been previously validated (e.g., Pde1a, Zic1, Kcnip2) as RGC type-specific[6, 7]. Some markers were cluster-specific only in adult RGCs, as during maturation they were either upregulated (e.g., Ceacam10 for C37, Nmb for C40, Slc17a7 for C25) or downregulated (e.g., Rhox5 for C32) in all but one cluster. For the adult RGC cluster C8, which did not have a uniquely enriched single gene, we identified a uniquely co-enriched 3-way combination of Prkcq, Tac1, and Syt6, using an algorithm we developed previously[7].

We also provided insights into the global characteristics of RGC types[7, 26] and how axonal injury affects them. We showed that the variance between RGC types' transcriptomes increases during developmental maturation, which may be related to the fact that there are more types in the adult (46) compared to the neonatal RGCs (40)[7], as the cells continued to diverge further towards their final state during maturation into adulthood[6]. We then found that the most divergent RGC types are resilient and belong to the alpha/ipRGC class, while the overall variability between the transcriptomes of RGC types substantially increases after injury. However, the proportion of highly expressed genes[26], another varying global property of the RGC types transcriptome[7], was largely unchanged by injury regardless of susceptibility or resilience of an RGC type. We also showed that neither the cluster size (average number of cells), cluster transcriptome size (average number of expressed genes), nor the proportion of cellular stress genes[16, 24, 25] were significantly related to the differential survivability between RGC types. Thus, while the factors unique to susceptible or resilient RGC types are primary determinants of survivability and may be associated with differential increase in variability between the transcriptomes of RGC types after injury, the differences in variability between global properties of atlas RGC transcriptomes are associated with the survivability of alpha/ipRGC types. This is the first time a global parameter of an uninjured cell type transcriptome has been linked to a neuronal response to injury, suggesting that transcriptomic global parameters are important to consider in future studies of neuronal survivability and regeneration.

As a number of studies are relying on the accuracy of RGC type identification as resilient or suspectable in investigating the differences in their underlying biology and identifying neuroprotective and risk genes, this timely update, along with new cluster-markers and insights into the global characteristics of RGC types, will prove valuable for the field.

SUPPLEMENTAL INFORMATION

Supplemental Information includes 5 Figures, 2 Tables, and 1 Data file.

ACKNOWLEDGMENTS

This work was supported by grants from The University of Connecticut School of Medicine, Start-Up Funds (to E.F.T.), the BrightFocus Foundation (Grant G2017204, to E.F.T.), and the National Institutes of Health (NIH) (Grant R01-EY029739, to E.F.T.). Portions of this research were conducted at the High Performance Computing Facility, University of Connecticut. We thank Nicholas Tran, Karthik Shekhar, and Joshua Sanes (Center for Brain Science and Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, Broad Institute of Harvard and MIT, Cambridge, MA, University of California, Berkeley CA, and Baylor College of Medicine, Houston TX) for providing us with raw data from which they generated the figures and tables in Tran et al. (2019). We also thank Sophan Iv and Vijender Singh (Research IT Services, University of Connecticut), and Stephen King (High Performance Computing Facility, University of Connecticut), for assistance with computational resources and the development of the website. Finally, we thank Ashiti Damania, Mahit Gupta, Anthony Antony, and Samantha DeRosa (undergraduate and graduate students, University of Connecticut) for technical assistance.

AUTHOR CONTRIBUTIONS

B.A.R. performed the analysis and contributed to conceptualization of the study, and E.F.T. conceptualized the study and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

METHODS

Single cell RNA-seq data procurement and initialization. Raw counts, normalized matrices, and cell metadata (e.g., cluster assignment) for the mouse adult RGC atlas and the injured RGCs were obtained from the Gene Expression Omnibus (GEO) accession number GSE137400[6]. Because the atlas RGCs were sequenced in 3 separate batches, we first performed batch correction using the FindIntegrationAnchors and IntegrateData functions from Seurat v. 4.0.3[16, 17]. Normalization of the raw counts was also performed using Seurat function, which divides the feature counts by the number of counts per each cell and then applying natural log transformation[16, 17].

Dimensionality reduction. Dimensionality reduction, visualization, and model generation for the subsequent reference mapping was performed using the uniform manifold approximation and projection (UMAP) implementation[20, 21] in Seurat v. 4.0.3 using default parameters[16, 17]. Feature selection for the UMAP algorithm was determined using the top 30 principal components (PCs). UMAP of atlas RGCs were color-coded with their original cluster designations[6] for determining cluster-latent-space relationship.

Reference mapping of injured to atlas RGCs. Injured RGC transcriptomes were mapped to the reference UMAP model using a custom deconstructed version of Seurat's MapQuery function that allows for the control of various parameters for optimizations. We used the FindTransferAnchors function from the Seurat package to discover shared anchors between reference and query datasets before transferring cluster labels and integrating and projecting embeddings onto the existing reference UMAP model. Projecting these embeddings requires the TransferData, IntegrateEmbeddings, and ProjectUMAP Seurat functions. The final algorithm outputs predicted ID scores (i.e., confidence scores) alongside the cluster designations. These confidence scores range from 0-1, with 1 denoting the highest confidence in the assignment of query cells to their respective reference clusters. In order to optimize the confidence of cell assignments, we created an R script which cycles through all valid k-weight parameters used in transferring, integrating, and projecting the query dataset, and outputted the average predicted confidence score of each fit for each k-weight. Subsequently, we used the optimal (e.g., resulting in the highest average confidence score) k-weight parameter (k = 6) in the TransferData, IntegrateEmbeddings, and ProjectUMAP Seurat functions, which generated the integrated dataset used for all the analyses of injured RGCs assigned to clusters. All of the injured RGCs at 2 weeks post-ONC were successfully mapped to their type-origin on the UMAP latent-space and assigned their respective cluster designations.

Design of the website and online tools. The Subtypes Gene Browser was designed in the same format as we previously did for RGC subtypes[7], using R and ShinyApps with R-markdown language[27]. Boxplots, violin plots, and bar plots were adapted from ggplot2 R software package for data visualization[28].

Statistical analyses. Pearson (2-tailed) correlation analyses were used for the transcriptomes of cells and clusters, and the mean correlation coefficient r was computed as reported, with mean ± SEM shown and significance cutoff at $p \leq 0.05$. In **Figure 4F-G** and **Supplemental Figure 2B**, correlations were visualized using linear regressions. For **Figure 4F**, clusters C24, C38, and C39 were excluded from the correlation and regression analyses, because they had no cells that survived 2 weeks after ONC. For **Supplemental Figure 1A**, analysis of significance was performed using ANOVA with Repeated Measures and posthoc LSD (SPSS). Fold change significance in Fig. 2E was determined using the EdgeR algorithm[29].

Thresholds for signature genes enriched in adult atlas RGC types. For determining unique single gene markers enriched in individual adult atlas RGC clusters (shown in **Figure 3**), we applied an algorithm we described previously in the classification of neonatal RGC types[7], using the following thresholds. Genes uniquely enriched per cluster were expressed (>0.05 *NE*) at least 1.8-fold more compared to every other cluster, respectively, at p-value ≤ 0.05. The p-value had to pass ≤ 0.05 on two separate tests: independent samples t-test (2-tailed) and a nonparametric independent samples Mann–Whitney U test (using R software). Enriched gene markers that didn't pass one or both statistical tests but met the first two criteria are listed in **Supplemental Data 1**. The threshold of 1.8-fold was selected because it predicted at least one uniquely enriched gene marker for all clusters except cluster 8. To identify a combination of genes to mark cluster C8 (**Supplemental Figure 5D-F**), we applied an algorithm we described previously for identifying a combination of transcriptional regulators uniquely co-enriched in a cluster[7].

Heatmaps, Violin plots, Density plots, and Boxplot. The heatmap for **Figure 3** was generated using the R package Superheat[30]. Expression values for each gene across different clusters were normalized with z-scores for each row using the centered Scale R function prior to plotting, and the columns were then specified in an increasing order from cluster 1 through 45. Violin plots were generated using Seurat's VlnPlot function. Base R

statistical software functions were used for generating the density plots (density function from the stats package) and boxplot (boxplot function from the graphics package).

Data and code availability. The raw and processed data used in this study are available through the NCBI GEO under accession numbers GSE137400[6] and GSE115404[7]. Data processed in this study are available through the NCBI GEO under accession number *Pending*. Access to the dataset is also available through a user-friendly Subtypes Gene Browser web application, https://health.uconn.edu/neuroregeneration-lab/subtypes-gene-browser. All R codes and scripts designed for this study will be shared upon request.

REFERENCES

1. Benowitz, L. I.; He, Z.; Goldberg, J. L., Reaching the brain: Advances in optic nerve regeneration. *Exp Neurol* **2017,** *287* (Pt 3), 365-373.
2. Chun, B. Y.; Cestari, D. M., Advances in experimental optic nerve regeneration. *Curr Opin Ophthalmol* **2017,** *28* (6), 558-563.
3. Ghaffarieh, A.; Levin, L. A., Optic nerve disease and axon pathophysiology. *Int Rev Neurobiol* **2012,** *105*, 1-17.
4. Williams, P. R.; Benowitz, L. I.; Goldberg, J. L.; He, Z., Axon Regeneration in the Mammalian Optic Nerve. *Annu Rev Vis Sci* **2020,** *6*, 195-213.
5. Kim, J.; Sajid, M. S.; Trakhtenberg, E. F., The extent of extra-axonal tissue damage determines the levels of CSPG upregulation and the success of experimental axon regeneration in the CNS. *Sci Rep* **2018,** *8* (1), 9839.
6. 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.; Sanes, J. R., Single-Cell Profiles of Retinal Ganglion Cells Differing in Resilience to Injury Reveal Neuroprotective Genes. *Neuron* **2019,** *104* (6), 1039-1055.e12.
7. Rheaume, B. A.; Jereen, A.; Bolisetty, M.; Sajid, M. S.; Yang, Y.; Renna, K.; Sun, L.; Robson, P.; Trakhtenberg, E. F., Single cell transcriptome profiling of retinal ganglion cells identifies cellular subtypes. *Nat Commun* **2018,** *9* (1), 2759.
8. Baden, T.; Berens, P.; Franke, K.; Román Rosón, M.; Bethge, M.; Euler, T., The functional diversity of retinal ganglion cells in the mouse. *Nature* **2016,** *529* (7586), 345-50.
9. Bae, J. A.; Mu, S.; Kim, J. S.; Turner, N. L.; Tartavull, I.; Kemnitz, N.; Jordan, C. S.; Norton, A. D.; Silversmith, W. M.; Prentki, R.; Sorek, M.; David, C.; Jones, D. L.; Bland, D.; Sterling, A. L. R.; Park, J.; Briggman, K. L.; Seung, H. S.; Eyewirers, Digital Museum of Retinal Ganglion Cells with Dense Anatomy and Physiology. *Cell* **2018,** *173* (5), 1293-1306.e19.
10. Shekhar, K.; Sanes, J. R., Generating and Using Transcriptomically Based Retinal Cell Atlases. *Annu Rev Vis Sci* **2021**.
11. Sanes, J. R.; Masland, R. H., The types of retinal ganglion cells: current status and implications for neuronal classification. *Annu Rev Neurosci* **2015,** *38*, 221-46.
12. Tasic, B., Single cell transcriptomics in neuroscience: cell classification and beyond. *Curr Opin Neurobiol* **2018,** *50*, 242-249.
13. Langer, K. B.; Ohlemacher, S. K.; Phillips, M. J.; Fligor, C. M.; Jiang, P.; Gamm, D. M.; Meyer, J. S., Retinal Ganglion Cell Diversity and Subtype Specification from Human Pluripotent Stem Cells. *Stem Cell Reports* **2018**.
14. Bray, E. R.; Yungher, B. J.; Levay, K.; Ribeiro, M.; Dvoryanchikov, G.; Ayupe, A. C.; Thakor, K.; Marks, V.; Randolph, M.; Danzi, M. C.; Schmidt, T. M.; Chaudhari, N.; Lemmon, V. P.; Hattar, S.; Park, K. K., Thrombospondin-1 Mediates Axon Regeneration in Retinal Ganglion Cells. *Neuron* **2019,** *103* (4), 642-657.e7.
15. Duan, X.; Qiao, M.; Bei, F.; Kim, I. J.; He, Z.; Sanes, J. R., Subtype-Specific Regeneration of Retinal Ganglion Cells following Axotomy: Effects of Osteopontin and mTOR Signaling. *Neuron* **2015,** *85* (6), 1244-56.
16. Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck, W. M.; Hao, Y.; Stoeckius, M.; Smibert, P.; Satija, R., Comprehensive Integration of Single-Cell Data. *Cell* **2019,** *177* (7), 1888-1902.e21.
17. Hao, Y.; Hao, S.; Andersen-Nissen, E.; Mauck, W. M.; Zheng, S.; Butler, A.; Lee, M. J.; Wilk, A. J.; Darby, C.; Zagar, M.; Hoffman, P.; Stoeckius, M.; Papalexi, E.; Mimitou, E. P.; Jain, J.; Srivastava, A.; Stuart, T.; Fleming, L. B.; Yeung, B.; Rogers, A. J.; McElrath, J. M.; Blish, C. A.; Gottardo, R.; Smibert, P.; Satija, R., Integrated analysis of multimodal single-cell data. *bioRxiv* **2020**, 2020.10.12.335331.
18. Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E., Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** *2008* (10), P10008.
19. Peng, Y. R.; Shekhar, K.; Yan, W.; Herrmann, D.; Sappington, A.; Bryman, G. S.; van Zyl, T.; Do, M. T. H.; Regev, A.; Sanes, J. R., Molecular Classification and Comparative Taxonomics of Foveal and Peripheral Cells in Primate Retina. *Cell* **2019,** *176* (5), 1222-1237.e22.
20. McInnes, L.; Healy, J., UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *CoRR* **2018,** *abs/1802.03426*.
21. Becht, E.; McInnes, L.; Healy, J.; Dutertre, C. A.; Kwok, I. W. H.; Ng, L. G.; Ginhoux, F.; Newell, E. W., Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* **2018**.

22. Van Der Maaten, L.; Hinton, G., Visualizing data using t-SNE. *Journal of Machine Learning Research* **2008,** *9*, 2579-2625.
23. Linderman, G. C.; Rachh, M.; Hoskins, J. G.; Steinerberger, S.; Kluger, Y., Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods* **2019,** *16* (3), 243-245.
24. Ilicic, T.; Kim, J. K.; Kolodziejczyk, A. A.; Bagger, F. O.; McCarthy, D. J.; Marioni, J. C.; Teichmann, S. A., Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **2016,** *17*, 29.
25. Zhao, Q.; Wang, J.; Levichkin, I. V.; Stasinopoulos, S.; Ryan, M. T.; Hoogenraad, N. J., A mitochondrial specific stress response in mammalian cells. *EMBO J* **2002,** *21* (17), 4411-9.
26. Trakhtenberg, E. F.; Pho, N.; Holton, K. M.; Chittenden, T. W.; Goldberg, J. L.; Dong, L., Cell types differ in global coordination of splicing and proportion of highly expressed genes. *Sci Rep* **2016,** *6*, 32249.
27. R-Core_Team., *R: A language and environment for statistical computing*. http://www.R-project.org: Vienna, Austria, 2014.
28. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. Springer: New York, NY, 2009.
29. McCarthy, D. J.; Chen, Y.; Smyth, G. K., Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **2012,** *40* (10), 4288-97.
30. Barter, R. L.; Yu, B., Superheat: An R Package for Creating Beautiful and Extendable Heatmaps for Visualizing Complex Data. *Journal of Computational and Graphical Statistics* **2018**.

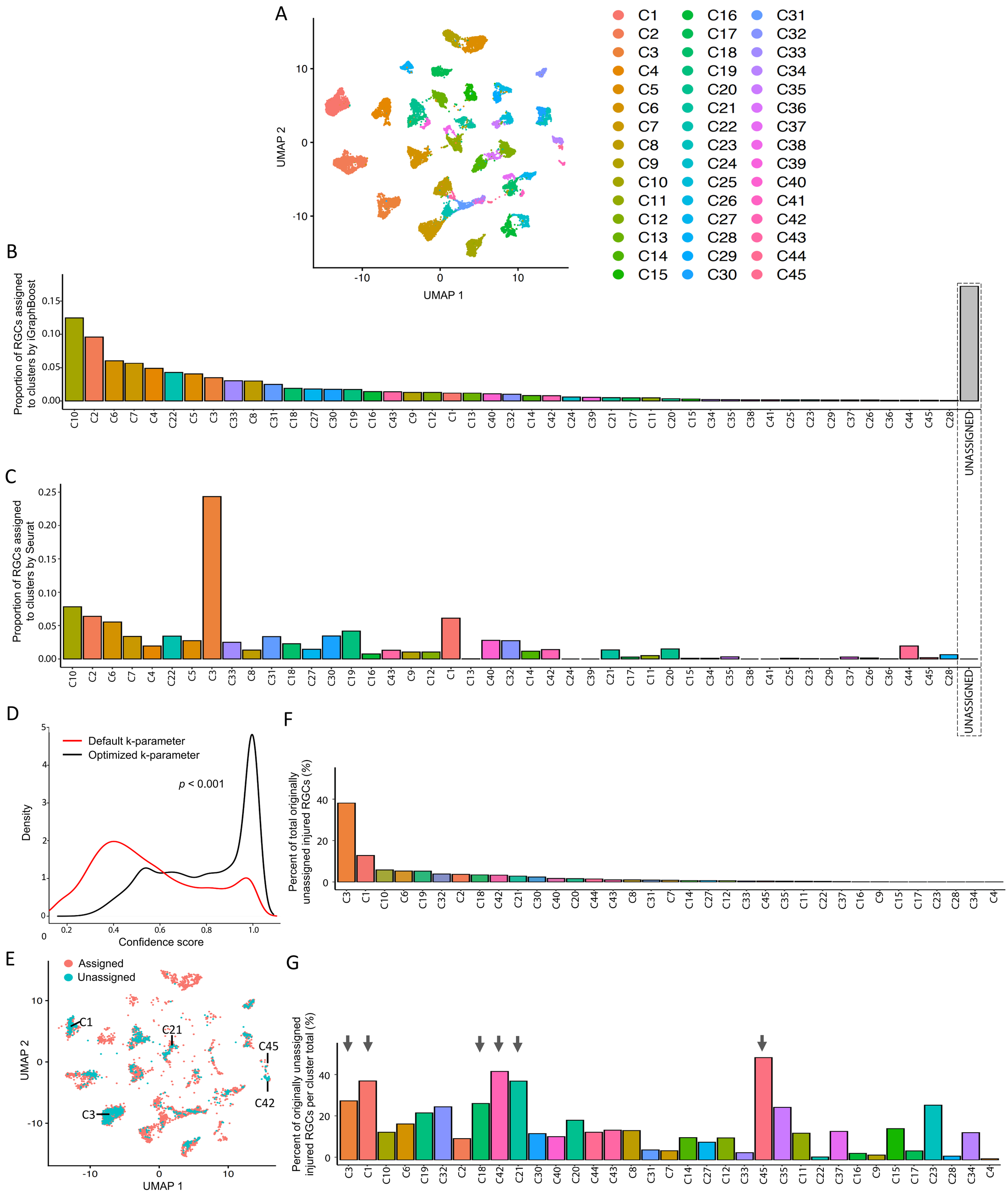

Figure 1. Cluster assignment by Seurat algorithm reveals overrepresented clusters within originally unassigned injured RGCs

(**A**) UMAP of uninjured adult atlas RGCs yielded a similar distribution of the clusters as the tSNE (colored by Louvain Jaccard-determined clusters) that was used originally (see Fig. 1C in Tran et al., 2019[6]). Clusters in UMAP and bar plots (*B*, *C*, *F*, *G*) all have the same color-code.

(**B-C**) Proportion of RGCs at 2 weeks after ONC assigned to clusters by the iGraphBoost (*B*) and Seurat (*C*) algorithms, showing a substantial percent of the RGCs unassigned to clusters by the iGraphBoost algorithm and the absence of unassigned RGCs by the Seurat algorithm (bar representing the unassigned RGCs outlined by dashed line). Clusters in *B* and *C* are ordered (left to right) following higher to lower proportion in *B*.

(**D**) Density plot comparing the confidence scores of cell-to-cluster assignments before (red line) and after (black line) k-weight parameter optimization ($p < 0.001$ by Wilcoxon Rank Sum test).

(**E**) UMAP of injured RGCs (at 2 weeks after ONC) assigned by Seurat algorithm to uninjured atlas clusters (shown in *A*). Red cells are those previously assigned by the iGraphBoost algorithm and green cells were previously unassigned. The UMAP demonstrates overrepresentation of unassigned injured RGCs in clusters C3, C1, C18, C42, C21, and C45 (labeled).

(**F-G**) Seurat algorithm assignments of RGCs previously unassigned by the iGraphBoost algorithm (at 2 weeks after ONC) as percent of total originally unassigned RGCs (*F*), and as percent of total RGCs assigned per respective cluster (*G*), showing substantial (>25%) overrepresentation in clusters C3, C1, C18, C42, C21, and C45 (indicated by arrows). Clusters in *F* and *G* are ordered (left to right) following higher to lower percent in *F*.

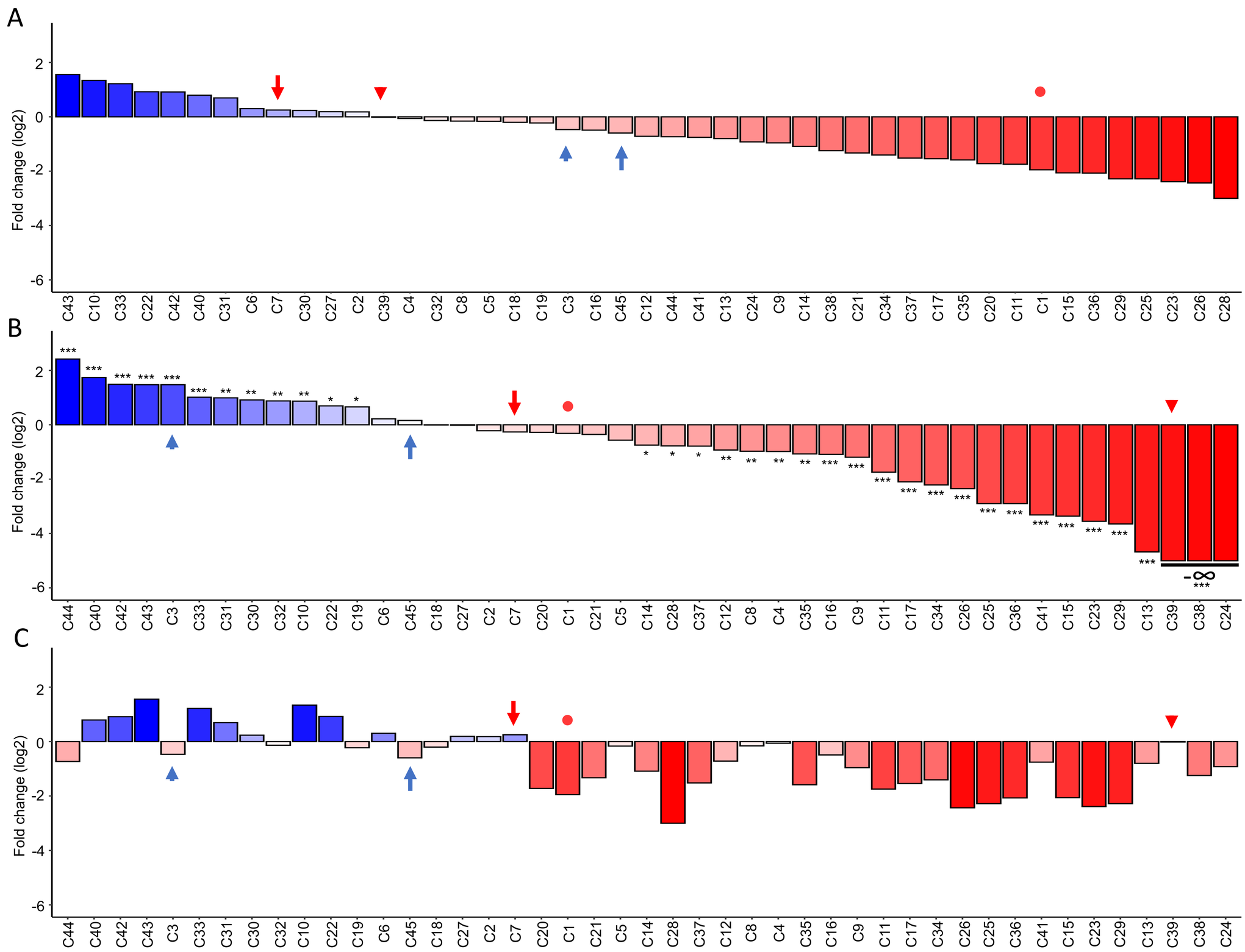

Figure 2. Re-analysis of the RGC types survivability after assignment to clusters of the previously unassigned injured RGCs
(**A**-**B**) RGC types' resilience/susceptibility determined by log2 fold-change of uninjured atlas RGCs to injured RGCs (at 2 weeks after ONC) based on cell-to-cluster assignment using iGraphBoost (replicated from Fig. 3E in Tran et al., 2019[6]) (*A*) or Seurat (*B*) algorithm. Fold change significance in (*B*) determined by EdgeR analysis, as indicated (***$p < 0.001$, **$p < 0.01$, and *$p < 0.05$; see Methods).
(**C**) To better visualize the shifts in type resilience/susceptibility determination (i.e., the differences between *A* and *B*), the same cluster order as in *B (*Seurat-determined) is shown with the cluster log2 fold-change and color-code as in *A* (iGraphBoost-determined). Arrows, arrowheads, and circles point to the same clusters at different positions in the panels (*A*-*C*): Blue arrow and arrowhead indicate examples of clusters that changed determination from "susceptible" to "resilient", red arrow and arrowhead indicate examples of clusters that changed determination from "resilient" to "susceptible", and a red circle indicates an example of a cluster that changed determination from highly "susceptible" to mildly "susceptible".

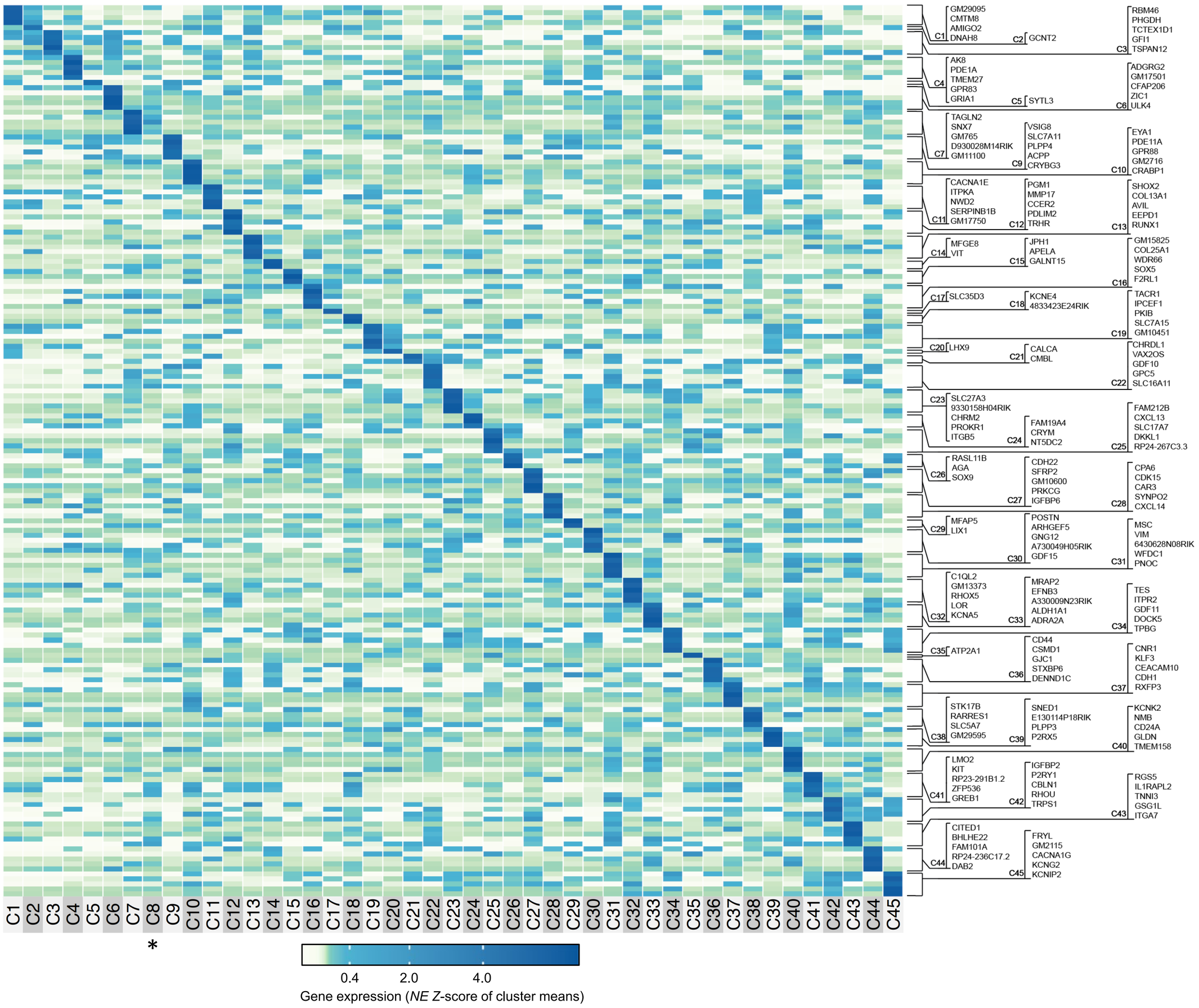

Figure 3. Heatmap signatures of cluster marker genes enriched in adult atlas RGC types

Heatmap of adult atlas RGC clusters, with up to 5 enriched single gene cluster markers indicated on the right side. The enriched genes required the following selection criteria: Gene expression in a cluster >0.05 *NE* and ≥1.8-fold relative to every other cluster, with enrichment *p*-value ≤ 0.05 by both independent samples *t*-test (2-tailed) and the nonparametric independent samples Mann–Whitney *U* test. Color-coded scale bar of gene expression indicates *z*-scores after normalizing to all the clusters (see Methods). Other genes enriched in clusters that did not pass one or both statistical tests but met the first two criteria are listed in Supplemental Data 1.

*Only cluster C8 did not have a single gene marker. Thus, a combination of genes to mark cluster C8 is shown in Supplemental Figure 5D-F.

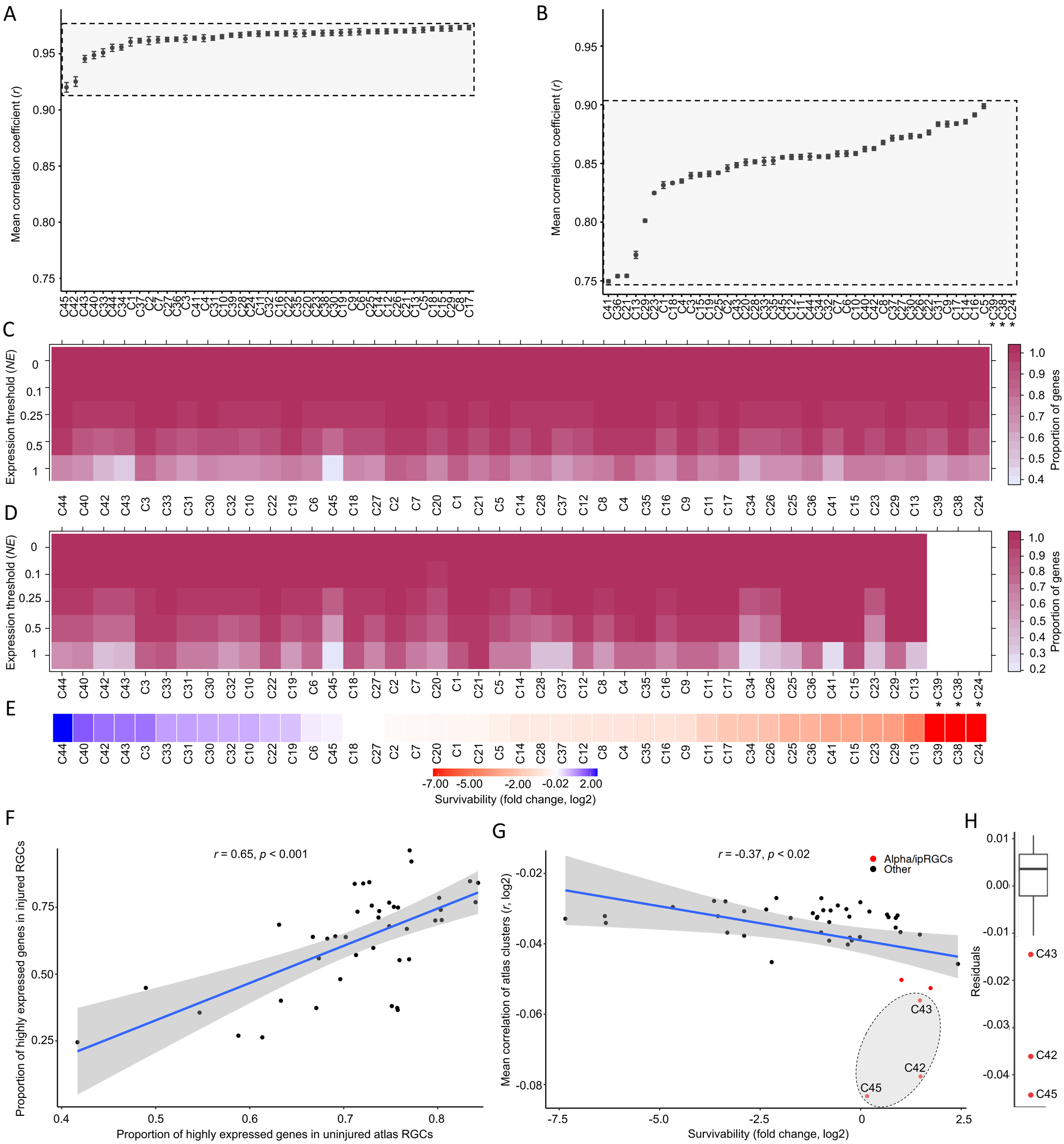

Figure 4. The effects of injury on global transcriptome parameters of RGC types

(**A**-**B**) Mean correlation coefficient (*r*) of atlas RGC clusters compared to every other uninjured atlas RGC cluster (*A*) and of injured RGC clusters compared to every other injured RGC cluster (*B*; at 2 weeks after ONC). The grey area (in *A* and *B*) represent the inter-cluster variance of Pearson correlation coefficients, highlighting >2-fold increase in variability of RGC transcriptomes after injury. C39, C38, and C24, which did not survive 2 weeks after injury are marked by asterisk (*) in *B* and *D*. Error bars = SEM.

(**C**-**D**) Proportions of low and highly expressed genes at different *NE* ranges, as marked, in uninjured atlas (*C*) and injured (*D*) RGC clusters (at 2 weeks after ONC). The order of the clusters (left to right) is the same in *C*-*E*.

(**E**) Survivability of RGC clusters (at 2 weeks after ONC), based on fold change (log2) of uninjured atlas to injured RGC clusters.

(**F**) Linear regression of the proportion of highly expressed genes (>1 *NE*) in uninjured atlas RGC clusters vs. the proportion of highly expressed genes (>1 *NE*) in injured (2 weeks after ONC) RGC clusters. Gray area around the fit-line (in *F* and *G*) indicates 95% confidence interval for the linear regression model.

(**G**) Linear regression of mean correlation coefficients of uninjured atlas RGC clusters compared to every other cluster (shown in *A*) vs. survivability of injured RGC clusters (ranked in *E*; at 2 weeks after ONC). The red clusters belong to the ip/alpha RGC class, as marked.

(**H**) A boxplot of the residuals for the regression in (*G*), with the most influential datapoints (i.e., outlier residuals shown in red) corresponding to the ip/alpha RGC clusters that are shown in dashed shaded oval in *G.*