Self-learning algorithm for denoising and advancing the integration of scRNA-seq datasets improves the identification of resilient and susceptible 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

The clustering of single cell RNA-sequencing (scRNA-seq) data enables the classification of cell types, and the development of integration algorithms has enabled the tracing of altered-from-baseline transcriptomes to their respective cell type origins. Here, we developed an algorithm that removes sources of noise from scRNA-seq datasets and improves the accuracy of their integration. We found that this improvement of integration is further increased by an algorithm we developed for auto-determination of optimal weight-assignment to anchors. As a proof-of-concept, we show that our algorithm determined the type-origin of the 17% of injured retinal ganglion cells (RGCs) that an existing integration algorithm could not identify. As we found that the majority of the originally unassigned cells belonged to only several out of 45 RGC types, a consequent change in the proportions of the surviving types resulted in a moderately different ranking of resiliency to injury. This updated standard for RGC type resilience ranking improved the understanding of RGC type biology and the ranking-dependent prediction of genes for therapeutic neuroprotective targeting. Additional bioinformatic analyses revealed single-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. We also show, for the first time, that specific global properties of the transcriptome can predict the resilience to injury of certain cell types. Furthermore, we developed a website for cluster-by-cluster comparison of gene expression between uninjured and injured RGC types. The publicly available R library package, CellTools, for denoising and improving the integration of scRNA-seq datasets, as well as for associating cell types with the global properties of the transcriptome, will assist scRNA-seq studies of developing, injured, or diseased cell types across biological fields.

INTRODUCTION

The existing integration algorithms, widely used for tracing altered transcriptomes of matured and/or injured cells to their respective scRNA-seq cluster-defined cell type origin, are evolving as new engineering and computational approaches emerge in the field. However, the limitations of comprehensiveness and accuracy of cell-to-cluster assignment by the existing integration algorithms requires new advancements[1-3]. Here, we developed a self-learning algorithm, which first denoises the reference dataset, and then determines the optimal parameters for reference-driven integration of scRNA-seq datasets. Our advancement significantly improved the quality of the reference dataset, by identifying poor quality cells, and integration of the query dataset, by increasing the percent of correct cell-to-cluster assignments. As a proof-of-concept, we show that its implementation improved the identification of scRNA-seq-derived resilient and susceptible retinal ganglion cell (RGC) types, which were originally classified using the integration algorithm iGraphBoost[4].

Changes in gene expression, whether during cell type maturation, after injury, or because of disease, require integration algorithms for tracing respective cell type origin. Mouse RGCs are comprised of 40-46 types[4-8] (sometimes referred to as subtypes[5,9-11]), with some types surviving longer after optic nerve crush (ONC) injury and/or being more capable of regenerating axons[4,12,13]. The relative proportion of surviving RGC types after injury compared to the relative proportion of uninjured atlas RGC types was used to rank the RGC types on a resilient to susceptible continuum, based on the iGraphBoost integration algorithm-determined RGC type origin of the injured cells[4]. However, the cell type origin of 17.2% of the injured RGCs (at 2 weeks after ONC, which is a standard time-point in this assay[4,14]) was not determined. Therefore, the prevalence analysis of the identified injured RGC types relied on the assumption that specific types were not over- or under-represented amongst those injured RGCs which were unassigned to types. Here, using the algorithm we developed, we determined the types of all of the injured RGCs (at 2 weeks after ONC). As we found that several types were overrepresented amongst the RGCs that were unassigned to types by the original integration analysis, consequently, a re-analysis of survivability (after assigning all the RGCs to types) resulted in a moderately different determination of the resilient and susceptible types.

The updated standard for RGC type resilience ranking, presented herein, will assist studies that rely on RGC type identification as resilient or susceptible in order to investigate the differences in their underlying biology and to identify neuroprotective and risk genes for therapeutic targeting. We also introduced an algorithm for analyzing global properties of the transcriptome, and show for the first time that it can predict biological phenotypes, such as the resiliency to injury of certain RGC types. Furthermore, we identified single gene cluster markers for RGC types, contributed new insights into how the extent of variance between transcriptomes of RGC types increases differentially during maturation and after injury, and developed a website for cluster-by-cluster comparison of gene expression between uninjured and injured RGC types.

The new findings in RGC type biology revealed herein were enabled by computational advancements, for which we made the publicly available R software library, CellTools. This suite of tools includes: DeNoise, for removing poor quality cells from reference dataset, the self-learning algorithm MapTo, for improving the accuracy of cell-to-cluster assignment in the integration of scRNA-seq datasets, and GlobalProbe, for associating cell types with global properties of the transcriptome. These tools will assist scRNA-seq studies across all biological fields.

RESULTS

To improve the integration of scRNA-seq datasets, we developed a two-step self-learning algorithm, which first denoises reference dataset, and then optimizes the *k*-weight parameter of the K-nearest neighbor (KNN) algorithm for a more accurate cell-to-cluster assignment (**Figure 1A-K** and **Supplemental Figure 1**; see in detail below). We show significant improvement by each step, using subsampling of the atlas of adult uninjured RGC types as a proof-of-concept (**Figure 1M**), which cumulatively reduced incorrect integration (i.e., mapping to reference) of query cells by 73% (on average).

Classification of the uninjured adult mouse RGC scRNA-seq-derived transcriptome profiles (adult RGC atlas, which we used for developing our algorithm) by the Louvain-Jaccard algorithm[16,17] predicted 45 clusters, one of which was manually sub-clustered into 2, resulting in a total of 46 types[4], that closely matched the 40 neonatal RGC types we identified previously[5] (see supplementary Fig. 1F in Tran et al., 2019[4]). However, the integration algorithm (iGraphBoost) traced the 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[4]). 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[4]). 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 selected this dataset for validating our algorithm by testing whether it could improve cell-to-cluster assignment in the integration and thereby enable testing this assumption.

Our algorithm applies a wrapper for the Seurat functions (FindTransferAnchors, TransferData, IntegrateEmbeddings, and ProjectUMAP)[1,15], which use high-dimensional data for determining the integration anchors used for mapping the query cells, and also uses uniform manifold approximation and projection (UMAP)[18,19] for transferring the reference cluster labels of nearest neighbors, that are determined, in part, by the *k*-weight parameter of the KNN algorithm. UMAP yielded a similar distribution of the Louvain Jaccard-determined clusters (**Figure 1B**) as did the scalable implementation of t-distributed stochastic neighbor embedding (tSNE)[20,21] that was used originally (see Fig. 1C in Tran et al., 2019[4]).

In the first step, the DeNoise function we developed, identifies and removes poor quality cells from the reference dataset, and then generates a denoised reference UMAP (**Figure 1C**; see Methods) for integration in the next step. The DeNoise function reduces ambiguity in cluster segregation (inter-cluster heterogeneity; **Figure 1D-E**) and increases intra-cluster homogeneity (**Figure 1F-G**), both of which impact weight-assignment of the anchors used for integration. The DeNoise algorithm identified ~3% of the total reference dataset cells as noise, and its validity was demonstrated by subsequent analyses, which revealed abnormalities in 83.2% of the noise cells (**Figure 1H-L**). For example, 53.7% of the noise cells had an abnormally large transcriptome size (**Figure 1H-J**), which could result from a partially damaged cell binding to the same bead as an intact cell (and pass the doublet QC because the damaged cell did not preserve its entire transcriptome), and 29.5% of the other noise cells had an abnormally high proportion of mitochondria-related genes (**Figure 1I-K**), which is typically higher in cells progressing towards apoptosis[1,22,23] (these cells could have been damaged more than other cells during processing). However, noise cells were not overrepresented in any clusters, and, accordingly, the proportions of

cluster sizes did not change after exclusion of the noise cells (no significant differences found by EdgeR analysis), further supporting our finding that these cells are a random consequence of the technical limitations detected by our DeNoise algorithm-improved Quality Control (QC). Further validation of the DeNoise algorithm was shown by that it has significantly improved the integration of the query cells, by merely eliminating sources of noise from the reference dataset, even before advancing the integration algorithm itself in the next step (**Figure 1M**).

In the second step, the MapTo function improves the accuracy of cell-to-cluster assignment in the integration by optimizing the *k*-weight parameter of the KNN algorithm. The *k*-weight parameter is the number of neighbors (i.e., cells) the KNN algorithm considers when assigning weights to anchors (i.e., neighboring pairs of cells from the reference and query datasets) in determining cell-to-cluster assignment. Our self-learning algorithm optimizes the *k*-weight parameter for the reference dataset, by randomly setting aside 20% (a default that can be changed by user) of the total cells from the reference dataset before clustering it, and determining at which *k*-weight parameter the maximum percent of these cells are accurately traced by the integration algorithm to the type-origin clusters (i.e., clusters to which the set-aside cells belonged originally) (**Supplemental Figure 1A**). To determine the optimal *k*-weight parameter, which yields the highest percent of accurately traced set-aside cells, the self-learning algorithm uses a loop-script (see Methods) that iterates through the Seurat functions' *k*-weight parameters and outputs the percent of the accurately traced set-aside cells. The algorithm then ranks the percent of the accurately traced cells from all iterations, and selects the *k*-weight parameter which yields the highest percent of correctly-mapped cells (**Supplemental Figure 1B-C**). We repeated this process 3 times, using different groups of randomly subsampled set-aside cells, and found that the optimal *k*-weight parameter significantly improved their integration (**Figure 1M**).

We then re-analyzed the original data using the algorithm we developed, in order to test whether the failure to trace the cluster origin of 17.2% of the injured RGCs resulted from computational bioinformatic limitations, or if it was due to biological changes after injury that dedifferentiated certain RGCs beyond bioinformatic traceability to types. We found that our algorithm enabled the tracing of all the RGCs (at 2 weeks after ONC) to their type-origin (**Figure 1N-O**), thereby demonstrating that computational limitations of the original approach resulted in the failure to trace the cluster origin of 17.2% of the injured RGCs, and that our algorithm overcame these limitations. We also found that our algorithm significantly increased the correlation between the uninjured and the injured RGC clusters (traced to respective type-origin), for the analyzed post-injury time-points, compared to using the original analyses (**Supplemental Figure 2A**). Moreover, injured RGCs clusters (after assignment; **Supplemental Figure 2A**), as well as all individual injured RGCs (pre-assignment; **Supplemental Figure 2B**), correlated with the uninjured RGC clusters or all individual RGCs, respectively, non-linearly relative to the analyzed post-injury time-points. For example, RGCs at 7 days after ONC were more similar to uninjured RGCs than the RGCs from a preceding time-point at 4 days after ONC (**Supplemental Figure 2A-B**). Therefore, using the sequential (i.e., iterative integration) of iGraphBoost tracing is not advantageous for the 2-week post-ONC time-point. Accordingly, there is a peak in the percentage of injured RGCs unassigned to clusters by the original algorithm at 4 and 7 days after ONC, compared to a smaller percent of RGCs unassigned at earlier time-points (at 1 and 2 days after ONC) and a later time-point (at 2 weeks after ONC; **Supplemental Figure 2C**).

Next, we analyzed whether certain RGC types were overrepresented amongst the originally unassigned injured RGCs. We found that the majority (57%) of the originally unassigned RGCs belonged to types C1, C3, C6, and C19, and that types C1, C19, C20, C26, C32, C35, C37, C42, and C45 were substantially (>25% each) overrepresented (**Figure 1P-S**). 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[4]). For example, the originally susceptible types C45, C19, and C3 now rank as resilient, the originally strongly susceptible type, C1, now ranks as mildly resilient, and the originally mildly resilient type, C39, now ranks as strongly susceptible (**Figure 2A-C**). We also developed the Subtypes Gene Browser web application (in the same format we previously presented for neonatal RGC subtypes from the left and right eyes[5]), which provides a platform for cluster-by-cluster differential gene expression analysis between the atlas and 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 which were unassigned to types by the original integration algorithm (iGraphBoost) were substantially different from the injured RGCs that were assigned to types by our algorithm. We found that, the transcriptomes of the originally unassigned overrepresented injured RGC types, not only did not belong to the types most or least changed by injury (**Figure 2D**), but when grouped separately based on our type-assignment were highly correlated ($r = 0.97$, $p < 0.0001$)

to their originally assigned counterpart types (**Figure 2E**), thereby further supporting the accuracy of their assignment to clusters by our algorithm. We also did not find the relative proportion of mitochondria-related genes (see above) 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 original algorithm (iGraphBoost) were not substantially different from other injured RGCs, but rather the limitations of the original 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[5] 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[5]. However, out of the 45 adult RGC clusters, 32 required combinations of enriched/unenriched genes to mark them uniquely in the original analysis, as single gene cluster markers were identified for only 13 clusters (see Fig. 1F and Supplementary Fig. 1F in Tran et al., 2019[4]). Of the 13 single gene cluster markers, 9 were amongst the previously predicted neonatal RGC single gene markers (see Fig. 4 and Supplementary Data/Table 1 in Rheaume et al., 2018[5]; **Supplemental Table 1**). Of the remaining 4 genes, 3 of them, 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, the neonatal clusters closely matched the adult clusters, with the exception of 2 neonatal clusters converging into 1 adult cluster in 3 instances, and the 5 clusters in the adult (see Supplementary Fig. 1F in Tran et al., 2019[4]). Therefore, we applied the algorithm we used for identifying single gene markers of neonatal RGC clusters[5], 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[4,5] (**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[5] 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 transcriptomic dynamics of RGC types, we analyzed whether the extent of variability between RGC types changes during maturation and after injury. We found that the variability between the transcriptomes of adult atlas RGC types is within an average *r* range of 0.91-0.97 (Pearson; **Figure 4A**), which is twice the variability we found for pre-eye-opening neonatal atlas RGC types (0.96-0.99 average *r* range[5]). Furthermore, the extent of variability between the transcriptomes of RGC types also increased after injury to 0.58-0.95 average *r* range (Pearson; **Figure 4B**). However, there was no significant correlation between survivability and the mean transcriptome size of uninjured ($r = 0.26$, $p = 0.07$) or injured ($r = 0.05$, $p = 0.74$) RGC types (**Supplemental Figure 4B-C**), nor the relative proportion of mitochondria-related genes (see above) of injured types ($r = 0.17$, $p = 0.25$; **Supplemental Figure 4A**). We also showed that the varying proportion of highly expressed >0.5 *NE*) genes[5,24] remained significantly correlated between uninjured and injured RGC types ($r = 0.47$, $p < 0.01$; **Figure 4C-F**), but it was not correlated to RGC survivability ($r = -0.05$, $p = 0.73$), suggesting that this varying global transcriptomic property of RGC types is largely unaffected by injury, regardless of susceptibility or resilience of an RGC type. Finally, we found that the variability between the transcriptomes of uninjured RGC types (average cluster *r* values in **Figure 4A**) is significantly correlated to the survivability of injured RGC types ($r = -0.34$, $p = 0.02$; **Figure 4G**), and that the types with the most influential regression residuals accounting for correlation are the resilient atlas types (C42, C43, and C45) belonging to the alpha/intrinsically-photosensitive RGCs (ipRGCs) class (see Table S2 in Tran et al., 2019[4]; **Figure 4G-H**).

DISCUSSION

The QC algorithm we developed for denoising the reference dataset, along with the self-learning algorithm we developed for improving the integration of scRNA-seq datasets, addressed the need in the field for new computational advancements of integration algorithms[1-3]. Retrospect analyses of the noise cells detected by our QC algorithm showed that they had either abnormally large transcriptomes (typical when a piece of a broken cell is attached to a bead with an intact cell) or an abnormally high proportion of mitochondrial genes (typical for cells

progressing towards apoptosis), were not overrepresented in any clusters, and their removal improved the accuracy of integration. Further improvement in the accuracy of cell-to-cluster assignment in the integration was achieved by the self-learning algorithm we developed for optimizing the *k*-weight parameter of the KNN algorithm. Cumulatively, our two-step algorithm reduced incorrect integration of the query cells with the reference dataset by 73% (on average), which represents a substantial advancement. As changes in gene expression during various biological and pathological processes require integration algorithms for tracing respective cell type origin, our algorithm will assist studies across biological fields aimed at tracing cell type-origin of injured, diseased, developing, matured, or otherwise broadly-changed transcriptomes.

As a proof-of-concept, we show that our algorithm improved the identification of scRNA-seq-derived resilient and susceptible RGC types, that were originally classified using the integration algorithm iGraphBoost[4]. We found that the originally unassigned injured RGCs were not substantially different from other injured RGCs in any of the global transcriptomic properties, and that RGC types C1, C19, C20, C26, C32, C35, C37, C42, and C45 were substantially overrepresented amongst the originally unassigned injured RGCs. Consequently, the re-analysis of the RGC types' survivability led to a moderately different identification of the resilient and susceptible RGC types than was originally reported[4]. Thus, the updated standard for RGC type resilience ranking presented herein, will assist the studies that rely on RGC type identification as resilient or susceptible in order to investigate the differences in their underlying biology and to identify neuroprotective and risk genes for therapeutic targeting.

We also used an algorithm we developed previously[5], to identify single gene markers uniquely enriched in all but one of the adult atlas RGC clusters[4], and found that many of these markers were also uniquely enriched in the developing neonatal atlas RGC clusters[5]. Importantly, 18.1% of these marker genes have been previously validated (e.g., Pde1a, Zic1, Kcnip2) as RGC type-specific[4,5]. 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[5]. These single gene markers will assist studies of RGC type biology.

Additionally, we provided insights into the global characteristics of RGC types[5,24] and how axonal injury affects them. We showed that the variance between RGC types' transcriptomes increases over two-fold during developmental maturation, which may be related to the fact that there are more types of RGCs in adult (46) than there are in neonatal (40)[5], as the cells continued to diverge further towards their final state during maturation into adulthood[4]. We then found that the most divergent atlas RGC types are resilient and belong to the alpha/ipRGC class, while after injury the variability between the transcriptomes of RGC types increased overall. However, the proportion of highly expressed genes[24], another varying global property of the RGC types transcriptome[5], 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[1,22,23] were significantly correlated 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 increases 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 the 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 rely 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, insights into the global characteristics of RGC types, and the Subtypes Gene Browser for cluster-by-cluster differential gene expression analysis between the atlas and injured RGC clusters (https://health.uconn.edu/neuroregeneration-lab/subtypes-gene-browser), will prove valuable for the field. We also made the publicly available R software library, CellTools, which includes the algorithm for removing poor quality cells from reference dataset (DeNoise), the self-learning algorithm for improving the accuracy of cell-to-cluster assignment in the integration of scRNA-seq datasets (MapTo), and the algorithm for associating cell types with global properties of the transcriptome (GlobalProbe). The CellTools software package will assist scRNA-seq studies across biological fields.

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. BAM files, raw counts, normalized matrices, and cell metadata (e.g., type assignment) for the mouse adult RGC atlas and the injured RGCs were obtained from the Gene Expression Omnibus (GEO) accession number GSE137400[4]. A total of 10 atlas RGC BAM files and 22 injured RGC BAM files were converted to FASTQ files using CellRanger's bamtofastq software. FASTQ files from both the atlas and injured data sets were aggregated where appropriate using CellRanger and then mapped to the CellRanger mm10-1.2.0 transcriptome, as the original analysis had mapped the atlas and injured data to different transcriptomes, making comparisons between them more difficult and, in some cases, impossible.

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[1,15]. Normalization of the raw counts was performed using the Seurat function, NormalizeData, which divides the feature counts by the number of counts per each cell and then applies natural log transformation[1,15]. The same cells that passed the initial quality check and thresholding, and the same type assignments for the atlas, from the original analysis[4] were used in the downstream characterization and analysis by the present study.

Dimensionality reduction. Dimensionality reduction, visualization, and model generation for the subsequent reference mapping was performed using the uniform manifold approximation and projection (UMAP) implementation[18,19] in Seurat v. 4.0.3 using default parameters[1,15]. Feature selection for the UMAP algorithm was determined using the top 50 principal components (PCs), which were all determined to be significant sources of variation by Seurat's JackStraw implementation[15,25]. UMAP of atlas RGCs were color-coded with their original cluster designations[4] for determining the cluster-latent-space relationship.

Denoising and reference-mapping of injured to atlas RGCs using CellTools' DeNoise and MapTo. First, the DeNoise algorithm removes the sources of noise from the reference UMAP by identifying outlier cells in the cluster-UMAP relationship. To do this, each cluster's quartiles and interquartile range (IQR) is calculated for the x and y dimensions of the UMAP. Any cell *i* belonging to cluster *c* (*c*,*i*) whose x or y dimensions are more than 1.5 times the IQR, either greater than the upper quartile (Q3) or less than the lower quartile (Q1), is considered to be *noise* within the UMAP structure and is subsequently removed for downstream analyses. The following equations summarize the discovery of *noise* and *signal* cells:

$$noise = Q1_{c,i} - \left(1.5 * IQR_{c,i}\right) > i > Q3_{c,i} + (1.5 * IQR_{c,i})$$

$$signal = Q1_{c,i} - \left(1.5 * IQR_{c,i}\right) \leq i \leq Q3_{c,i} + (1.5 * IQR_{c,i})$$

After *noise* cells are removed from the original UMAP, a new UMAP is generated with only the *signal* cells using Seurat's RunPCA and RunUMAP (with return.model = T) functions. Injured RGC transcriptomes are then mapped to the denoised reference UMAP model using the MapTo function from CellTools (which we developed) using the top 50 PCs. MapTo is a an integration algorithm, which utilizes a wrapper/iterative-looper for Seurat's FindTransferAnchors, TransferData, IntegrateEmbeddings, and ProjectUMAP functions[15]. The first step identifies the optimal *k*-weight parameter for the KNN latent search-space by iteratively mapping a test-query dataset to a test-reference dataset with known cell types. To accomplish this, the algorithm first randomly subsamples the reference dataset into a test-query dataset. By default, 20% of the cells are subsampled, but this parameter can be changed by user (for this study we used 23.7%, or 8,456 cells, which matched the size of the query dataset). The remaining cells from the reference dataset are designated as the test-reference dataset, and the test-query dataset is iteratively mapped to the test-reference dataset using every *k*-weight parameter within the user-defined range (default is between 10-100; see **Supplemental Figure 1B**). Iterations are then ranked based on the percent of correctly mapped cells, and the iteration with the highest percent of accurately mapped cells is selected as the optimal *k*-weight parameter (if more than one *k*-parameter yields same percent of accurately mapped cells, then the smaller *k*-parameter is selected; for this study the optimal $k = 23$). Next, the query dataset is mapped to the denoised reference dataset using the optimal *k*-weight parameter. The predicted cell IDs (i.e., the reference dataset IDs transferred to the query data; RGC types here) are outputted in the returned query Seurat object, along with a reference-based UMAP of the query cells.

CellTools' GlobalProbe. We designed the function GlobalProbe to explore the relationship between the global transcriptomic properties and cell types, and to test whether these properties can predict (using a linear model) cell type-specific phenotypes (e.g., resilience to injury, which is a dependent variable here). The function requires a Seurat object as input and the designation of the cluster label slot by which the user intends to make comparisons. The function will output the mean and SEM of correlation coefficients (Pearson's) for the comparisons between each cluster's mean gene expression and every other cluster's mean gene expression. The outputted graph is formatted as in **Figure 4A-B**. If a dependent variable for a cell type-specific phenotype is passed to the function (by default this option is off), the function will output the graph formatted as in **Figure 4G**, which will use a linear model to show whether the global transcriptomic properties can predict any of the cell type-specific phenotypes, that would be indicated by a significant ($p < 0.05$) coefficient of determination (R^2).

Design of the website and online tools. The Subtypes Gene Browser was designed in the same format as we previously did for RGC subtypes[5], using R and ShinyApps with R-markdown language[26]. Boxplots, violin plots, and bar plots were adapted from ggplot2 R software package for data visualization[27]. CellTools is a suite of single-cell analytical tools that was developed for this study and others, and is available via GitHub at https://github.com/BruceRheaume/CellTools.

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 3B**, 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 3A**, analysis of significance was performed using ANOVA with posthoc LSD (SPSS). Fold change significance in Fig. 2E was determined using the EdgeR algorithm[28].

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[5], 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[5].

Heatmaps, Violin plots, Density plots, and Boxplot. The heatmap for **Figure 3** was generated using the R package Superheat[29]. 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[4] and GSE115404[5]. 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. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 06 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031
2. Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet*. 05 2019;20(5):257-272. doi:10.1038/s41576-019-0093-7
3. Kharchenko PV. The triumphs and limitations of computational methods for scRNA-seq. *Nat Methods*. 07 2021;18(7):723-732. doi:10.1038/s41592-021-01171-x
4. Tran NM, Shekhar K, Whitney IE, et al. Single-Cell Profiles of Retinal Ganglion Cells Differing in Resilience to Injury Reveal Neuroprotective Genes. *Neuron*. 12 2019;104(6):1039-1055.e12. doi:10.1016/j.neuron.2019.11.006
5. Rheaume BA, Jereen A, Bolisetty M, et al. Single cell transcriptome profiling of retinal ganglion cells identifies cellular subtypes. *Nat Commun*. Jul 2018;9(1):2759. doi:10.1038/s41467-018-05134-3
6. Baden T, Berens P, Franke K, Román Rosón M, Bethge M, Euler T. The functional diversity of retinal ganglion cells in the mouse. *Nature*. Jan 2016;529(7586):345-50. doi:10.1038/nature16468
7. Bae JA, Mu S, Kim JS, et al. Digital Museum of Retinal Ganglion Cells with Dense Anatomy and Physiology. *Cell*. 05 2018;173(5):1293-1306.e19. doi:10.1016/j.cell.2018.04.040
8. Shekhar K, Sanes JR. Generating and Using Transcriptomically Based Retinal Cell Atlases. *Annu Rev Vis Sci*. Jul 06 2021;doi:10.1146/annurev-vision-032621-075200
9. Sanes JR, Masland RH. The types of retinal ganglion cells: current status and implications for neuronal classification. *Annu Rev Neurosci*. Jul 08 2015;38:221-46. doi:10.1146/annurev-neuro-071714-034120
10. Tasic B. Single cell transcriptomics in neuroscience: cell classification and beyond. *Curr Opin Neurobiol*. 06 2018;50:242-249. doi:10.1016/j.conb.2018.04.021
11. Langer KB, Ohlemacher SK, Phillips MJ, et al. Retinal Ganglion Cell Diversity and Subtype Specification from Human Pluripotent Stem Cells. *Stem Cell Reports*. Mar 2018;doi:10.1016/j.stemcr.2018.02.010
12. Bray ER, Yungher BJ, Levay K, et al. Thrombospondin-1 Mediates Axon Regeneration in Retinal Ganglion Cells. *Neuron*. 08 2019;103(4):642-657.e7. doi:10.1016/j.neuron.2019.05.044
13. Duan X, Qiao M, Bei F, Kim IJ, He Z, Sanes JR. Subtype-Specific Regeneration of Retinal Ganglion Cells following Axotomy: Effects of Osteopontin and mTOR Signaling. *Neuron*. Mar 2015;85(6):1244-56. doi:10.1016/j.neuron.2015.02.017
14. Kim J, Sajid MS, Trakhtenberg EF. 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*. Jun 2018;8(1):9839. doi:10.1038/s41598-018-28209-z
15. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *bioRxiv*. 2020:2020.10.12.335331. doi:10.1101/2020.10.12.335331
16. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*. 2008/10/09 2008;2008(10):P10008. doi:10.1088/1742-5468/2008/10/p10008
17. Peng YR, Shekhar K, Yan W, et al. Molecular Classification and Comparative Taxonomics of Foveal and Peripheral Cells in Primate Retina. *Cell*. 02 2019;176(5):1222-1237.e22. doi:10.1016/j.cell.2019.01.004
18. McInnes L, Healy J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *CoRR*. / 2018;abs/1802.03426
19. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. Dec 2018;doi:10.1038/nbt.4314
20. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. Article. *Journal of Machine Learning Research*. 2008;9:2579-2625.
21. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods*. 03 2019;16(3):243-245. doi:10.1038/s41592-018-0308-4
22. Ilicic T, Kim JK, Kolodziejczyk AA, et al. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol*. Feb 17 2016;17:29. doi:10.1186/s13059-016-0888-1
23. Zhao Q, Wang J, Levichkin IV, Stasinopoulos S, Ryan MT, Hoogenraad NJ. A mitochondrial specific stress response in mammalian cells. *EMBO J*. Sep 02 2002;21(17):4411-9. doi:10.1093/emboj/cdf445
24. Trakhtenberg EF, Pho N, Holton KM, Chittenden TW, Goldberg JL, Dong L. Cell types differ in global coordination of splicing and proportion of highly expressed genes. *Sci Rep*. 2016;6:32249. doi:10.1038/srep32249

25. Chung NC, Storey JD. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*. 2015;31(4):545-554.
26. R-Core_Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org; 2014.
27. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer; 2009.
28. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*. May 2012;40(10):4288-97. doi:10.1093/nar/gks042
29. Barter RL, Yu B. Superheat: An R Package for Creating Beautiful and Extendable Heatmaps for Visualizing Complex Data. *Journal of Computational and Graphical Statistics*. 2018;doi:10.1080/10618600.2018.1473780

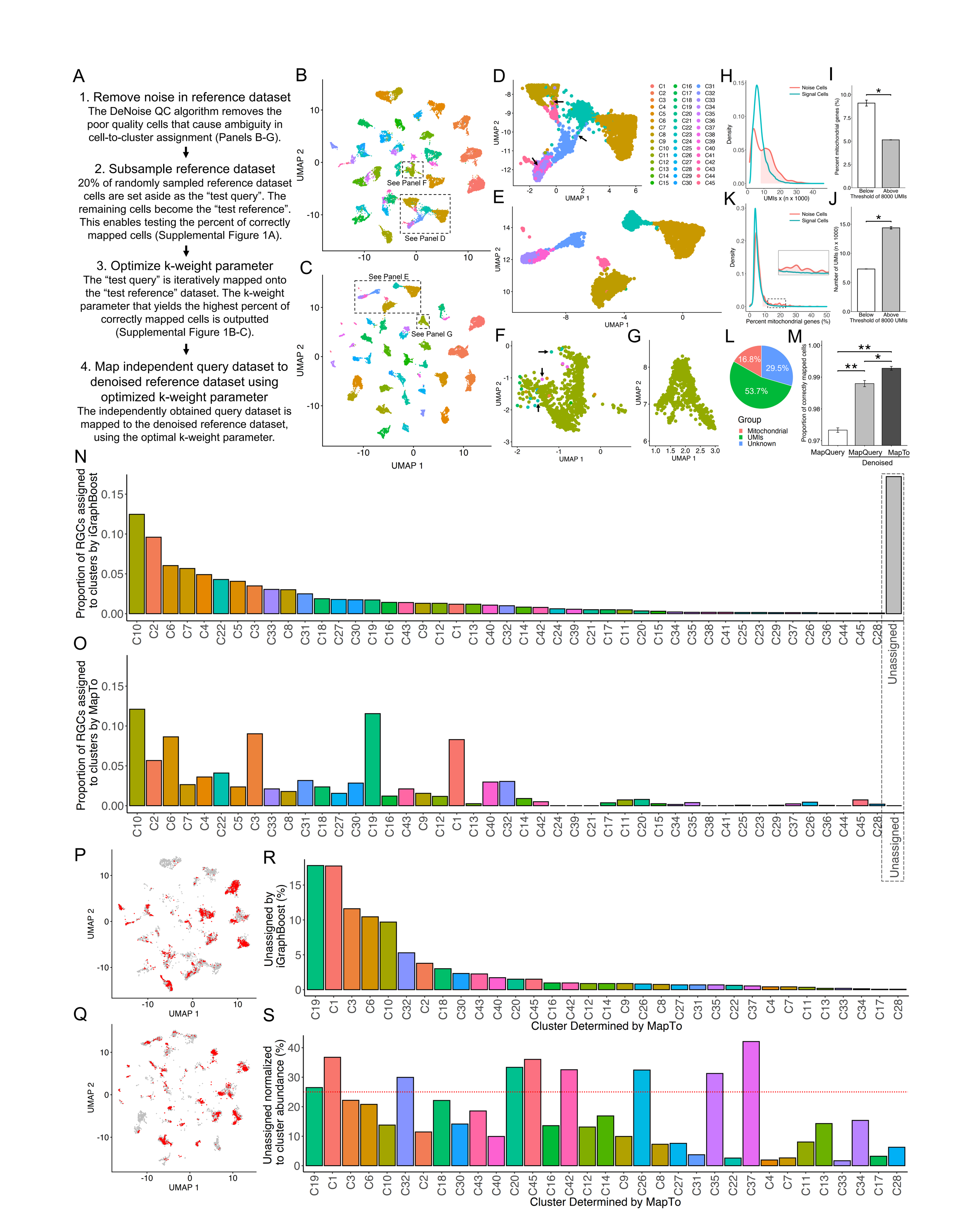

Figure 1. For figure legend see next page.

Figure 1. MapTo algorithm removes noise and improves integration, showing overrepresentation of originally unassigned injured RGCs in certain clusters

(**A**) A flowchart of the steps that the MapTo algorithm performs. First, the noise cells are removed from the reference dataset by the DeNoise function. Next, 20% of the randomly subsampled reference dataset is set aside as a "test query" and the remainder becomes the "test reference". Then, the test query is iteratively mapped over a range of *k*-weights (see methods) to the test reference, and the *k*-weight that yields the highest percent of correctly mapped cells is selected as optimal for mapping the original query dataset.

(**B-C**) The UMAP of uninjured adult atlas RGCs (***B***) 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 all UMAPs and bar plots (*B-G, N-O, and R-S*) have the same color-code, shown in the legend to the right of (*D*). A denoised reference dataset UMAP (***C***) shows improved cluster-UMAP agreement (compared to B).

(**D-G**) Insets (outlined in *B-C*) show reduction in inter-cluster heterogeneity (***D*** compared to improved ***E***), and increase in intra-cluster homogeneity (***F*** compared to improved ***G***). Arrows point to examples of ambiguity that are resolved by the DeNoise algorithm.

(**H-K**) A unimodal distribution in density of the UMIs of the "signal" (i.e., not noise) cells overlapped the first peak in the bimodal distribution of the "noise" cells (***H***), but the second peak representing an abnormally large transcriptome size was unique to the "noise" cells (***I***). However, the first (normal) peak of UMIs of the "noise" cells (in *H*, annotated "below" in *I*) was significantly enriched for mitochondria-related genes (***J***). Conversely, a second peak representing abnormally high proportion of mitochondria-related genes was also unique to "noise" cells on the mitochondria-related genes density plot, whereas "signal" cells unimodal distribution overlapped the first peak in the bimodal distribution of the "noise" cells (***K***). * $p < 0.0001$ by *t*-test, 2-tailed; error bars = SEM.

(**L**) A color-coded pie graph of the "noise" cells, showing abnormalities in 83.2% of these cells accounted for by abnormally large transcriptome size (53.7%) or abnormally high proportion of mitochondria-related genes (29.5%).

(**M**) Average (from using different groups of randomly subsampled set-aside cells, $N = 3$) percent of "test query" cells correctly mapped to the "test reference" dataset by default integration (MapQuery) is significantly improved by denoising the reference dataset (MapQuery-denoised), and further improved by optimizing the *k*-weight parameter (MapTo-denoised). * $p < 0.01$, ** $p < 0.001$ by ANOVA; error bars = SEM.

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

(**P-Q**) UMAPs of injured RGCs (at 2 weeks after ONC) assigned by the MapTo algorithm to uninjured atlas reference dataset, before (***P***; mapped to the original reference dataset shown in *B*) and after denoising (***Q***; mapped to the denoised reference dataset shown in *C*). Colored-coded (in *P-Q*) previously assigned (gray) and unassigned (red) by the iGraphBoost algorithm cells, help to visualize that originally unassigned injured RGCs (red cells) remained overrepresented in certain clusters even after denoising (*Q*).

(**R-S**) MapTo algorithm assignments of RGCs previously unassigned by the iGraphBoost algorithm (at 2 weeks after ONC) as percent of total originally unassigned RGCs (***R***), and as percent of total RGCs assigned per respective cluster (***S***), showing substantial (>25%) overrepresentation in clusters C1, C19, C20, C26, C32, C35, C37, C42, and C45. Clusters in *R* and *S* are ordered (left to right) following higher to lower percent in *F*.

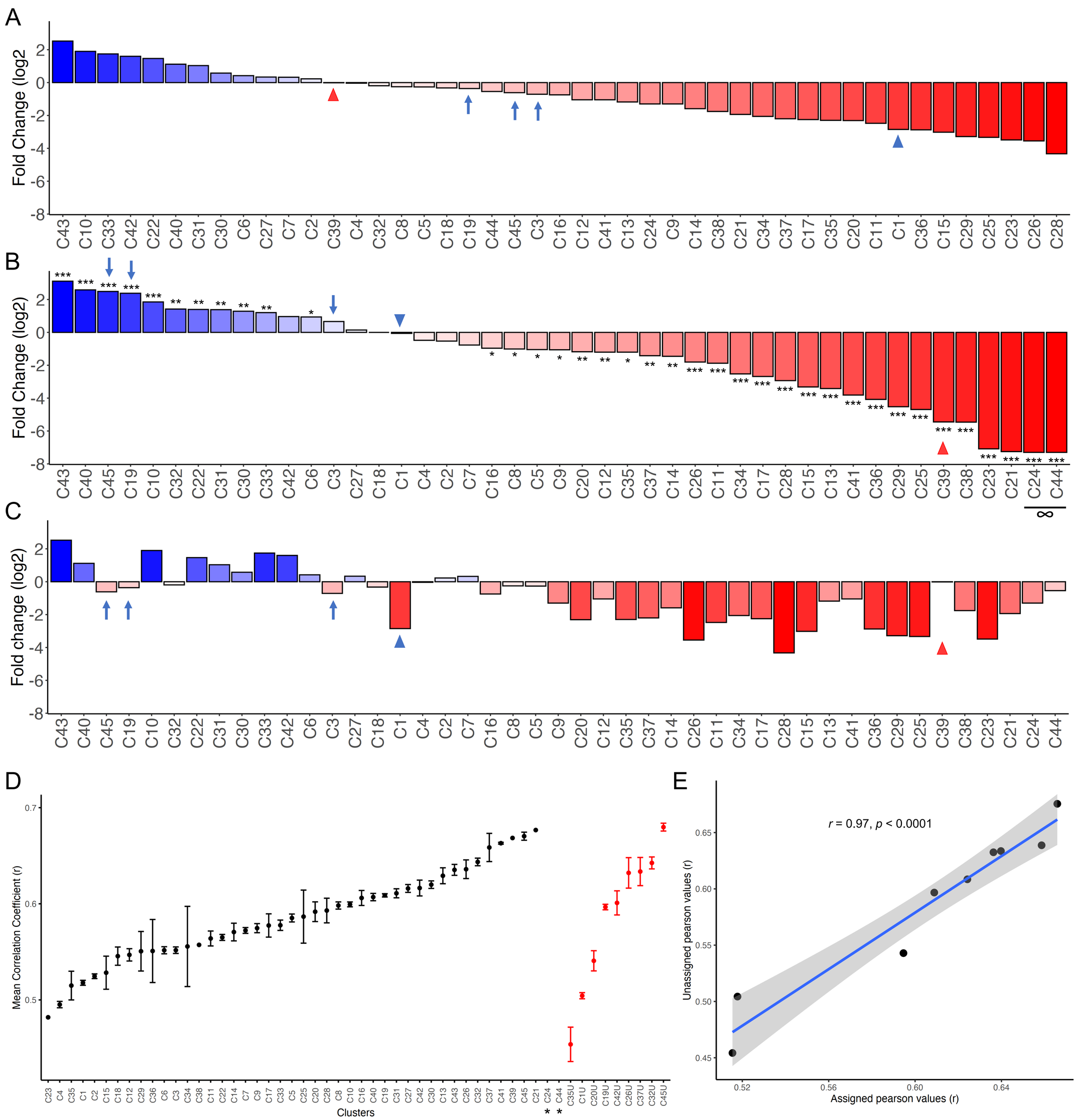

Figure 2. Re-analysis of the RGC types survivability after assignment to clusters of the previously unassigned injured RGCs

(**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 MapTo (***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* (MapTo-determined) is shown with the cluster log2 fold-change and color-code as in *A* (iGraphBoost-determined). Arrows and arrowheads 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" or to "neutral" (i.e., clusters that did not show significant change in *B*), respectively; red arrowhead indicates an example of a cluster that changed determination from "neutral" to "susceptible".

(**D**) Black datapoints: the mean correlation coefficients (*r*) of the injured (2 weeks after ONC) RGCs assigned to clusters by the MapTo algorithm (excluding those that were originally unassigned to clusters by the iGraphBoost algorithm) and all uninjured atlas RGCs from their respective clusters. Red datapoints: Mean correlation coefficients (*r*) of the injured (2 weeks after ONC) RGCs that were originally unassigned to clusters by the iGraphBoost algorithm (but assigned to clusters in retrospect by the MapTo algorithm) and uninjured atlas RGCs from their respective clusters. C24 and C44, which did not survive 2 weeks after injury are marked by asterisk (*). The range of distribution for the black and red datapoints is similar, indicating that the changes due to injury in the transcriptomes of originally unassigned RGC types are within the normal range compared to the other RGC types. U = Unassigned; Error bars = SEM.

(**E**) Linear regression of mean correlation coefficients of the C1U, C19U, C20U, C26U, C32U, C35U, C37U, C42U, and C45U (that were originally unassigned to clusters by the iGraphBoost algorithm, but assigned to clusters in retrospect by the MapTo algorithm) and uninjured atlas RGCs from their respective clusters (red datapoints in *D*), vs. mean correlation coefficients of injured RGCs assigned to counterpart clusters (excluding the RGCs in C1U, C19U, C20U, C26U, C32U, C35U, C37U, C42U, and C45U that were originally unassigned to clusters by the iGraphBoost algorithm) and uninjured atlas RGCs from their respective clusters (black datapoints in *D* for C1, C19, C20, C26, C32, C35, C37, C42, and C45). High correlation coefficient ($r = 0.97$; Pearson's) indicates that the transcriptomes of originally unassigned RGC types are highly correlated with their assigned counterpart clusters. Gray area around the fit-line indicates 95% confidence interval for the linear regression model.

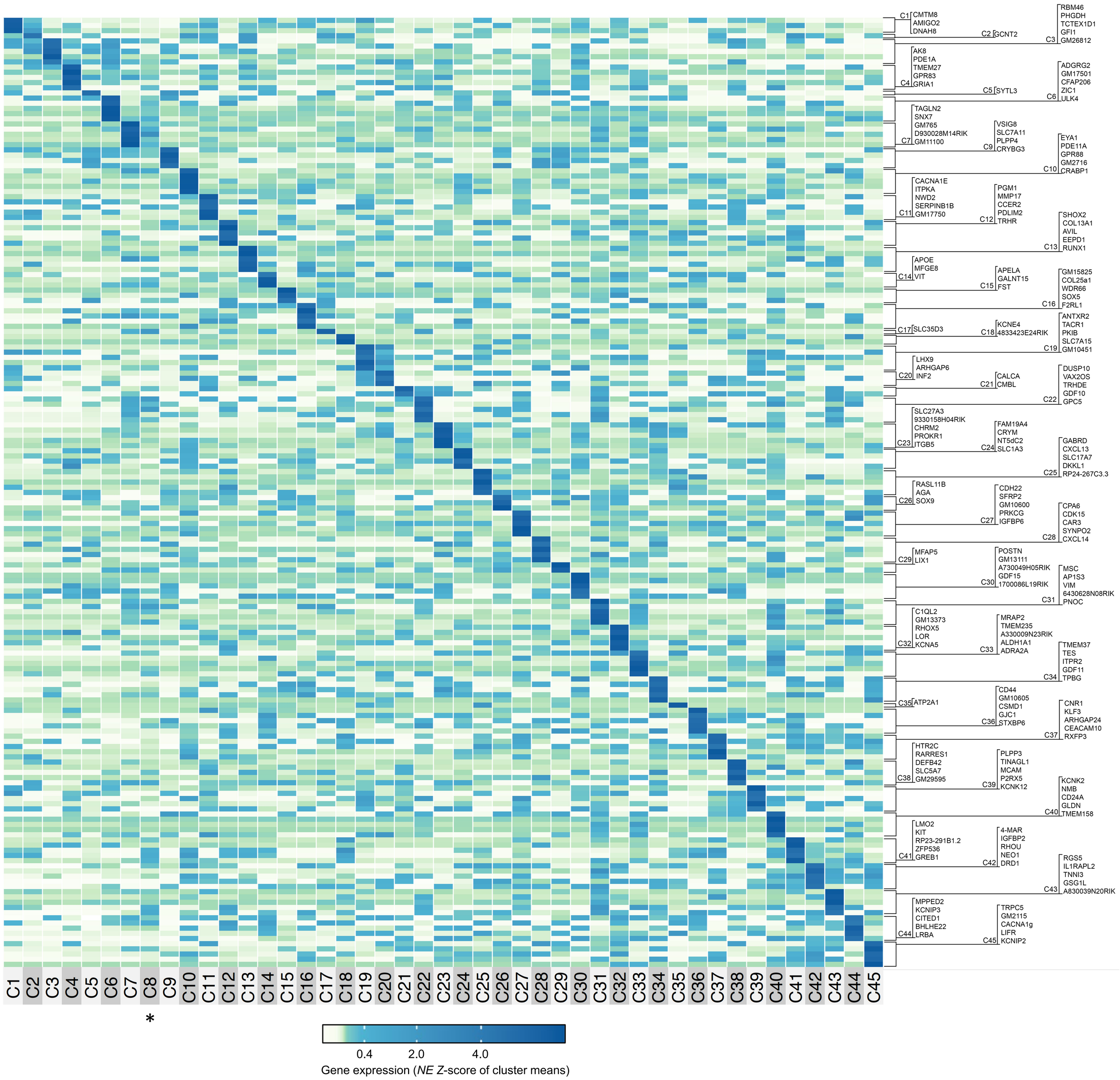

Figure 3. Heatmap signatures of cluster marker genes enriched in adult atlas RGC types

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.

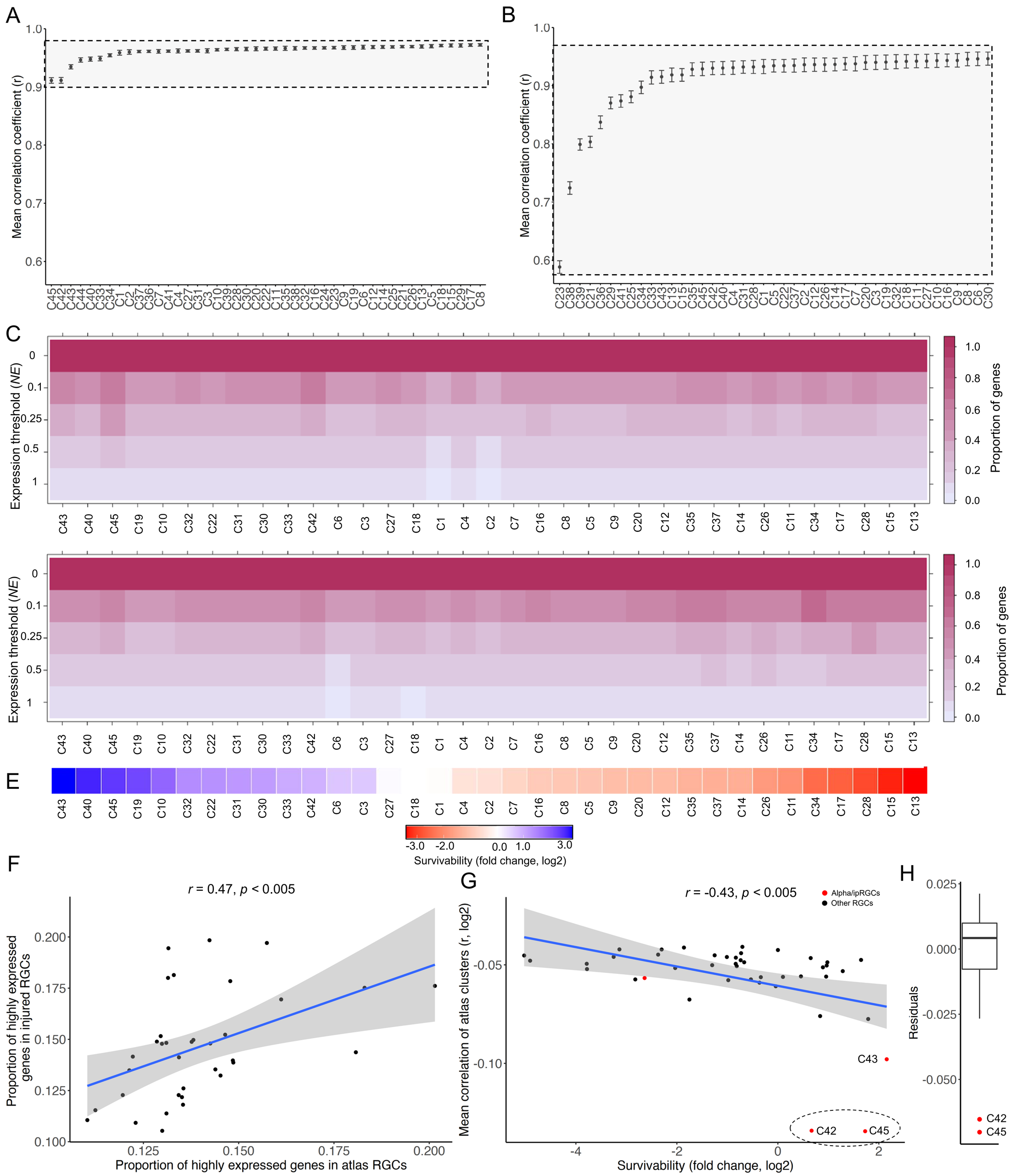

Figure 4. The effects of injury on global transcriptome parameters of RGC types

(**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. C24 and C44, which did not survive 2 weeks after injury, are not shown in *B*. 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. The order of the clusters (left to right) is the same in *C*-*E*. Clusters that contain 0 or < 5 cells at 2 weeks after ONC are excluded from analyses in *C*-*F*.

(**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 (>0.5 *NE*) in uninjured atlas RGC clusters vs. the proportion of highly expressed genes (>0.5 *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.*