MAPPING NEURONAL CELL TYPES USING INTEGRATIVE MULTI-SPECIES MODELING OF HUMAN AND MOUSE SINGLE CELL RNA SEQUENCING[*]

TRAVIS JOHNSON MS,
Dept. Biomedical Informatics, Ohio State University,
250 Lincoln Tower, 1800 Cannon Dr. Columbus, Ohio, 43210
Travis.Johnson@osumc.edu

ZACHARY ABRAMS PhD,
Dept. Biomedical Informatics, Ohio State University,
250 Lincoln Tower, 1800 Cannon Dr. Columbus, Ohio, 43210
Zachary.Abrams@osumc.edu

YAN ZHANG PhD,
Dept. Biomedical Informatics, Ohio State University,
250 Lincoln Tower, 1800 Cannon Dr. Columbus, Ohio, 43210
Yan.Zhang@osumc.edu

KUN HUANG PhD,
Dept. Biomedical Informatics, Ohio State University,
250 Lincoln Tower, 1800 Cannon Dr. Columbus, Ohio, 43210
Kun.Huang@osumc.edu

Mouse brain transcriptomic studies are important in the understanding of the structural heterogeneity in the brain. However, it is not well understood how cell types in the mouse brain relate to human brain cell types on a cellular level. We propose that it is possible with single cell granularity to find concordant genes between mouse and human and that these genes can be used to separate cell types across species. We show that a set of concordant genes can be algorithmically derived from a combination of human and mouse single cell sequencing data. Using this gene set, we show that similar cell types shared between mouse and human cluster together. Furthermore we find that previously unclassified human cells can be mapped to the glial/vascular cell type by integrating mouse cell type expression profiles.

[*] This work is partially supported by RGP0053 of the Human Frontier Science Program

1. Introduction

Mouse models are an important part of biomedical research and are routinely used as a stepping-stone towards treatments for humans – gleaning knowledge from high-throughput low risk experiments. Translating this knowledge requires a firm understanding of similarities between these two species [1-2]. Homologous genes exist between these species and these genes often play similar roles in the brain [3]. However, the biochemical pathways within each species have subtle to extreme differences leading to subsets of homologous genes without exact mechanistic overlap in the brain [4]. To address the issue of identifying functionally similar homologous genes we propose the concept of concordant genes defined as gene homologs that mechanistically behave similarly between two species [5]. Specifically, we hypothesize that concordant genes between mouse and human exist and that those genes can be algorithmically derived from combined mouse-human data. We also hypothesize that based off of these concordant genes we can determine cell type matching between mouse and human. Specifically in this study we focus on the comparison of brain cell gene expression profiles between mouse and human to identify concordant gene expression patterns in the brain tissue associated with different cell types taking advantages of recent development in single cell transcriptomics for brain cells. We hope that the single cell granularity of these comparisons will augment the tissue level comparisons of the human and mouse brain transcriptome [6].

RNA sequencing (RNA-Seq) in the past has been used to study brain structure, development, and disease [7]. Recently RNA-Seq has become more granular in the form of single cell RNA sequencing (scRNA-Seq) which is an important tool in the study of tissue heterogeneity due to its unique ability to characterize transcriptomes at the cellular level [8]. Recent advances in single cell transcriptomics in the brain have provided researchers with an influx of new data spanning different brain regions, diseases, and species [9]. Specifically, the Linnarsson group amassed a large single cell dataset from the mouse cortex and hippocampus which was clustered into multiple cell types based expression profiles [10]. Subsequent to the mouse single cell transcriptomic study, the Zhang group created a large human brain scRNA-Seq dataset from postmortem brain tissue and clustered the cells into unique cell types based on expression profiles [11]. Because of the availability of both datasets we believe that in-depth comparative analyses of these two datasets is fundamental to our understanding of neuronal cell types, the distribution of these cell types, and the evolution of brain anatomy in these two species. Furthermore a clear understanding of concordant genes in both human and mouse provides valuable information on how mouse studies can be translated to human research. We provide a methodology and gene set that can be used for these comparative studies and hopefully for future translational research. We demonstrate the method by not only identifying concordant cell types between mouse and human brains with the same set of concordant feature genes, but also matching un-categorized cells in the human brain to a salient cell type based on mouse brain information.

2. Methods

2.1. *Data normalization and cleaning*

The mouse scRNA-Seq unique molecular identifier (UMI) counts [12] were downloaded from the Data section of the Linnarsson lab website (http://linnarssonlab.org/) and human scRNA-Seq transcripts per million (TPM) data was downloaded from the Links section of the SCAP-T website (scap-t.org). Since these data files contain various numbers of genes with different order, we preprocessed the files by scanning matching gene symbols between files then sorting the gene symbols so that the orders were consistent. While this process may not be able to identify all homologous genes, it provides a large list for us to extract concordant genes. The shared gene symbols in the human and mouse datasets were retained for further study (Figure 1). Within the human dataset there were genes that were originally left out of analysis by the original authors due to low expression, resulting in some cells with low number of expressed genes. Because of this, such human cells as well as human cells without annotation in the metadata were also removed from further analysis, resulting in 3,086 human cells each containing 13,355 genes. The mouse dataset resulted in 3,005 cells each containing expression values from 13,355 genes. Both human and mouse data then were transformed into comparable units. Each dataset was log2 transformed and the expression values converted into the within cell z-scores.

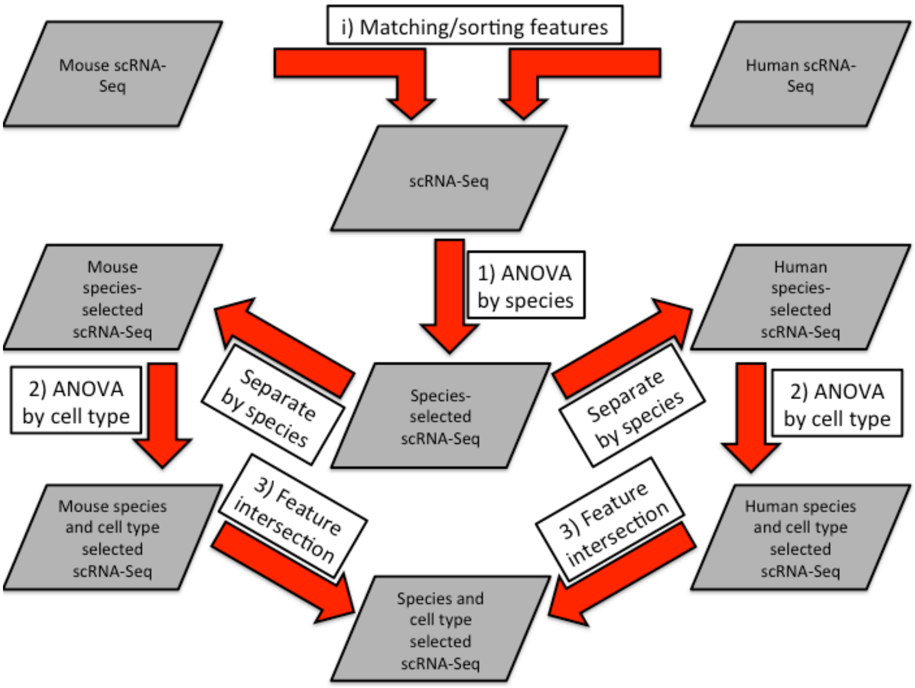
Mouse scRNA-Seq
i) Matching/sorting features
Human scRNA-Seq
scRNA-Seq
1) ANOVA by species
Mouse species-selected scRNA-Seq
Human species-selected scRNA-Seq
Separate by species
Separate by species
Species-selected scRNA-Seq
2) ANOVA by cell type
2) ANOVA by cell type
Mouse species and cell type selected scRNA-Seq
Human species and cell type selected scRNA-Seq
3) Feature intersection
3) Feature intersection
Species and cell type selected scRNA-Seq

Figure 1. Workflow of data normalization (i) and three step feature selection method (1-3).

2.2. *Feature selection*

We developed a three-step approach to find concordant genes between mouse and human based on gene expression profiles (Figure 1). This feature selection was performed to identify genes that were informative at separating cell type but uninformative at separating mouse from human cells. Genes that meet this criterion would be more useful at identifying similar cell types across species.

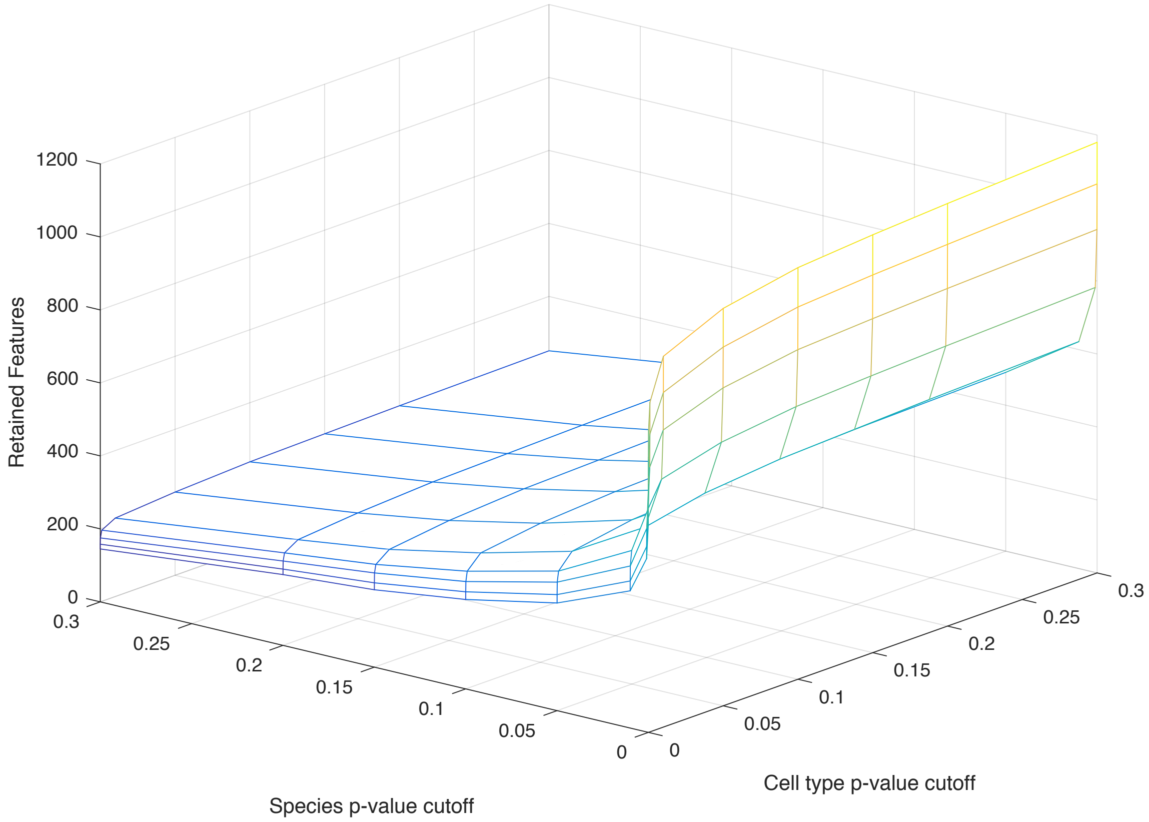

Figure 2. Number of retained features as a function of p-value cutoffs.

First, the human and mouse data matrices were concatenated such that the first 3086 columns consisted of human cells and the last 3005 columns consisted of mouse cells. For each gene in the data matrix, a one-way ANOVA was performed grouped by species to detect genes with significantly different expression level between human and mouse. Only genes with p-values larger than 0.1 were kept. This was done to remove genes that would separate cells by species. Because we are removing the significant genes from our gene set in Step 1, a greater threshold makes our criterion for retaining genes more strict than using a standard significance level. Second, the human and mouse matrices were separated and in each separate matrix a one-way ANOVA was performed on the remaining genes grouped by cell type label and using a threshold p-value of 0.01 – any genes found with a p-value of 0.01 or less were retained. The 0.01 threshold was used to provide stricter criteria for retained genes that were informative about cell type. The 0.1 and 0.01 p-value cutoffs used in the feature selection method are near the inflection point of retained features as a function of cutoff p-value (Figure 2). Third, the intersection of retained genes from human and mouse were retained in the final dataset such that genes that existed in both human and mouse gene sets after Step 2 were retained in the final combined mouse-human gene set.

To compare the differences between cell types and in concordance with previous single cell studies [13], principal component analysis (PCA) was applied to the human and mouse datasets prior to feature selection. The first 2 principal components were then plotted to visually show the differences in cell types and species (Figure 3). After feature selection, principal cross-species cell-type clusters can be viewed in the PCA of the first two principal components colored by species (left) and cell type (right) (Figure 4).

2.3. *Functional annotation of retained concordant genes*

When selecting features, it is important to study the relation of these feature/gene sets to the functional, anatomic, and phenotypic relationships that are being selected for. If there are functional relationships related to a phenotype, then the feature selection method targeting that phenotype is likely more robust. The retained genes from the feature selection step were used as input for the DAVID functional annotation software [13-14]. The functional annotation clusters were reviewed for over represented terms that can be attributed to neural pathways and cell types. We display the three most highly enriched terms within the three most highly enriched clusters from the DAVID functional annotation clustering (Table 1).

2.4. *Clustering cells using Gaussian mixture models*

Gaussian mixture models are effective in clustering microarray expression profiles [16]. We apply Gaussian mixed models (GMMs) in the mouse and human scRNA-Seq data to cluster the cells into principal cell types and to compare the relative proportions of human and mouse cells within each cluster. To perform the GMM we used the first two principal components, the same components used in the PCA plot of cell types. Four GMMs were fit to the data with two, three, four and five components respectively. The cells were clustered into three major cluster using the three component GMM fit in concordance with the three major cell types present in the human dataset. The remaining GMM fits were used in comparison against the three-component GMM fit.

Principal cell types of the mouse and human labels were compared in the PCA space to determine the most similar cell types between both species. To quantitate the mouse-human overlap the mouse and human data were split into three groups from the three major cell types in the original publications. Human cells were split into 3 major groups from their original labels [11]. All “Int” labeled cells were considered Interneuron. All “Ex” labeled cells were considered pyramidal. All “NoN” (No Nomenclature) labeled cells from a C1 Fluidigm chip with reduced mapping rates were without a biologically derived label but were considered a singular group. Similarly, mouse cells were also split based on cell type label mapping to GMM clusters [10]. All cells labeled Interneurons were still considered Interneurons. All S1 Pyramidal and CA1 Pyramidal were considered Pyramidal. All Oligodendrocytes, Microglia, Endothelial, Astrocytes, Ependymal and Mural were considered Glial/Vascular cells. All human and mouse cells that were contained within each GMM cluster were compared by the their original cell type labels to the labels of the GMM cluster. For each cluster a fisher exact test was conducted to calculate the odds ratios and confidence intervals between published cell type labels and GMM predicted cell types.

The VennX package in MATLAB was used to convert the cell type labels into Venn Diagrams to show overlap with both three component GMM predicted cell types and original mouse/human cell type labels from their original publications.

3. Results

3.1. *Feature selection*

Prior to feature selection the human and mouse cells created two clusters separated by species. The mouse cells formed sub-clusters within the major mouse clustering of cells. The human cells formed one main cluster with little differentiation (Figure 3).

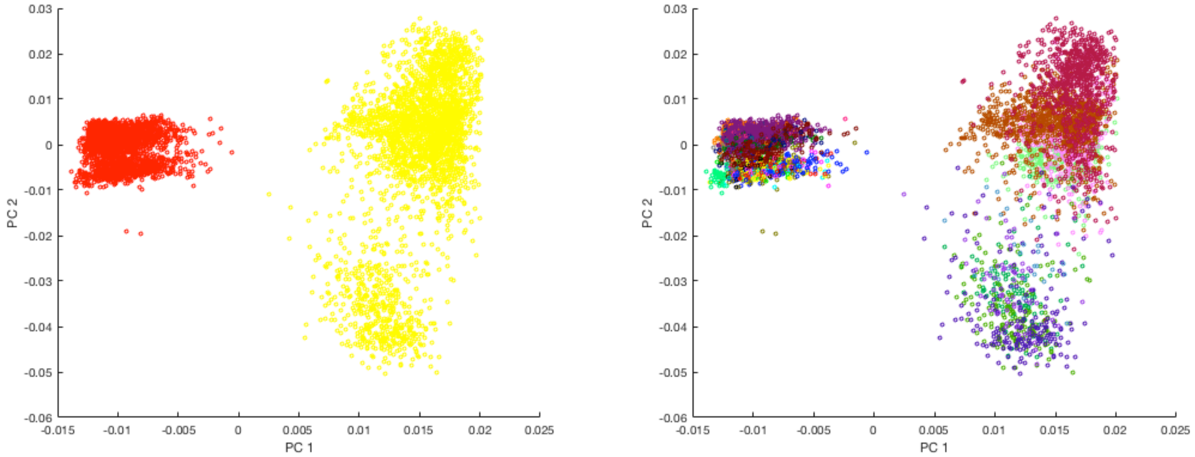

Figure 3. PCA of all human and mouse cells after normalization/cleaning. Left is colored by species, mouse (yellow) and human (red). Right is colored by cell type (36 cell types).

After feature selection, 358 concordant genes were retained, which are informative in terms of distinguishing cell types and uninformative in terms of separating species. As a result, human and mouse cells were no longer completely separate from each other. The mouse cell types still have more variability than the human cell types in the PCA space but cells from both species are contained within the same major clusters of cells (Figure 4).

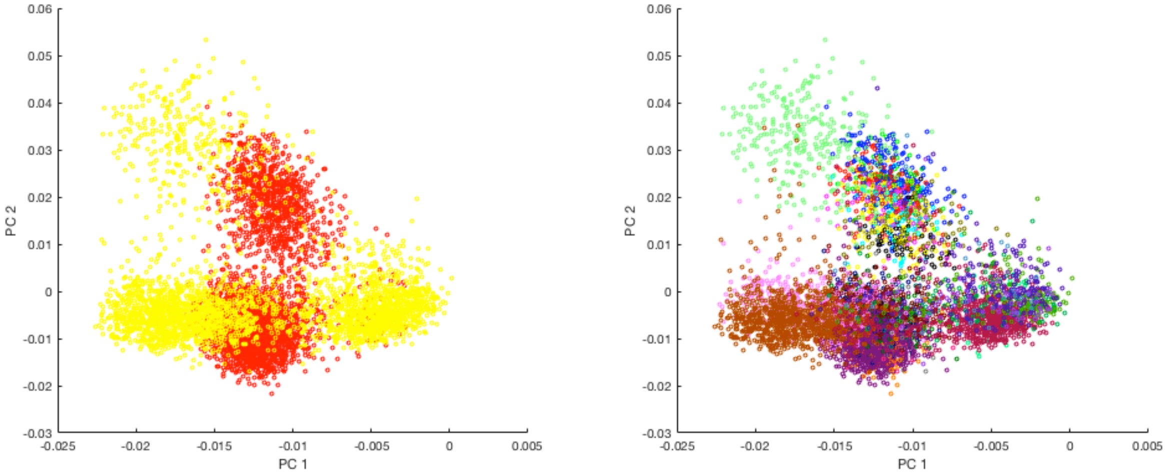

Figure 4. PCA of all human and mouse cells after normalization/cleaning and feature selection. Left is colored by species, mouse (yellow) and human (red). Right is colored by cell type (36 cell types).

3.2. *Functional annotation of concordant genes*

Functional annotation analysis of the concordant gene set revealed GO terms related to binding, ion transport and neural cells. The third most highly enriched annotation cluster was that of the GO terms axon, cell projection and neuron projection with an enrichment score of 1.57 (Table 1). Cluster 7 (not displayed) also contained many neuron related ontology terms.

Table 1. Functional annotation clustering using DAVID. Shown below are the three most highly enriched clusters and three most highly enriched terms within each cluster.

Category	Term	PValue	Fold Enrichment	Bonferroni
Annotation Cluster 1	Enrichment Score: 1.670			
SP_PIR_KEYWORDS	atp-binding	0.008	1.573	0.939
SP_PIR_KEYWORDS	nucleotide-binding	0.010	1.478	0.969
GOTERM_MF_FAT	GO:0032559~adenyl ribonucleotide binding	0.012	1.463	0.997
Annotation Cluster 2	Enrichment Score: 1.594			
GOTERM_BP_FAT	GO:0006826~iron ion transport	0.002	8.868	0.979
SP_PIR_KEYWORDS	iron transport	0.007	10.076	0.919
GOTERM_BP_FAT	GO:0000041~transition metal ion transport	0.012	4.347	1.000
Annotation Cluster 3	Enrichment Score: 1.568			
GOTERM_CC_FAT	GO:0030424~axon	0.010	3.027	0.946
GOTERM_CC_FAT	GO:0042995~cell projection	0.036	1.611	1.000
GOTERM_CC_FAT	GO:0043005~neuron projection	0.056	1.877	1.000

3.3. *Clustering cells using gaussian mixture models*

Gaussian mixture models showed major patterns within the cell profiles. Interneurons from both human and mouse (red and yellow respectively)(Figure 5) clustered in the same GMM. Whereas human pyramidal/projection neurons clustered (green) clustered with the remaining 2 cell types in mouse (S1 pyramidal, CA1 pyramidal). It is also worth consideration that the non-biologically labeled “NoN” human cell types in purple are mapped to a third cluster that begins to appear at 3 GMM components that contains the remaining 6 mouse cell types (mural, endothelial, microglia, ependymal, astrocytes, oligodendrocytes) (Figure 5).

The GMM clustering using three components (BIC = -9.08×10^4) split the cells into three groups that can be roughly defined as Interneurons (red), Pyramidal cells (green) and Glial/Vascular cell types (blue) (Figure 6: Top left). After identifying these three groups and comparing the mouse and human labels the GMM labels it was found that these three groups, Interneurons, Pryamidal cells, and Glial/Vascular cells are very closely mapped between both mouse and human. Also the “NoN” cell type cluster found in the human scRNA-Seq paper were clearly and uniquely clustered with the mouse Glial/Vascular cells (Figure 6 bottom right) with no significant difference between Glial/Vascular mouse cells and “NoN” human cells on PC 1 p-value = 0.41.

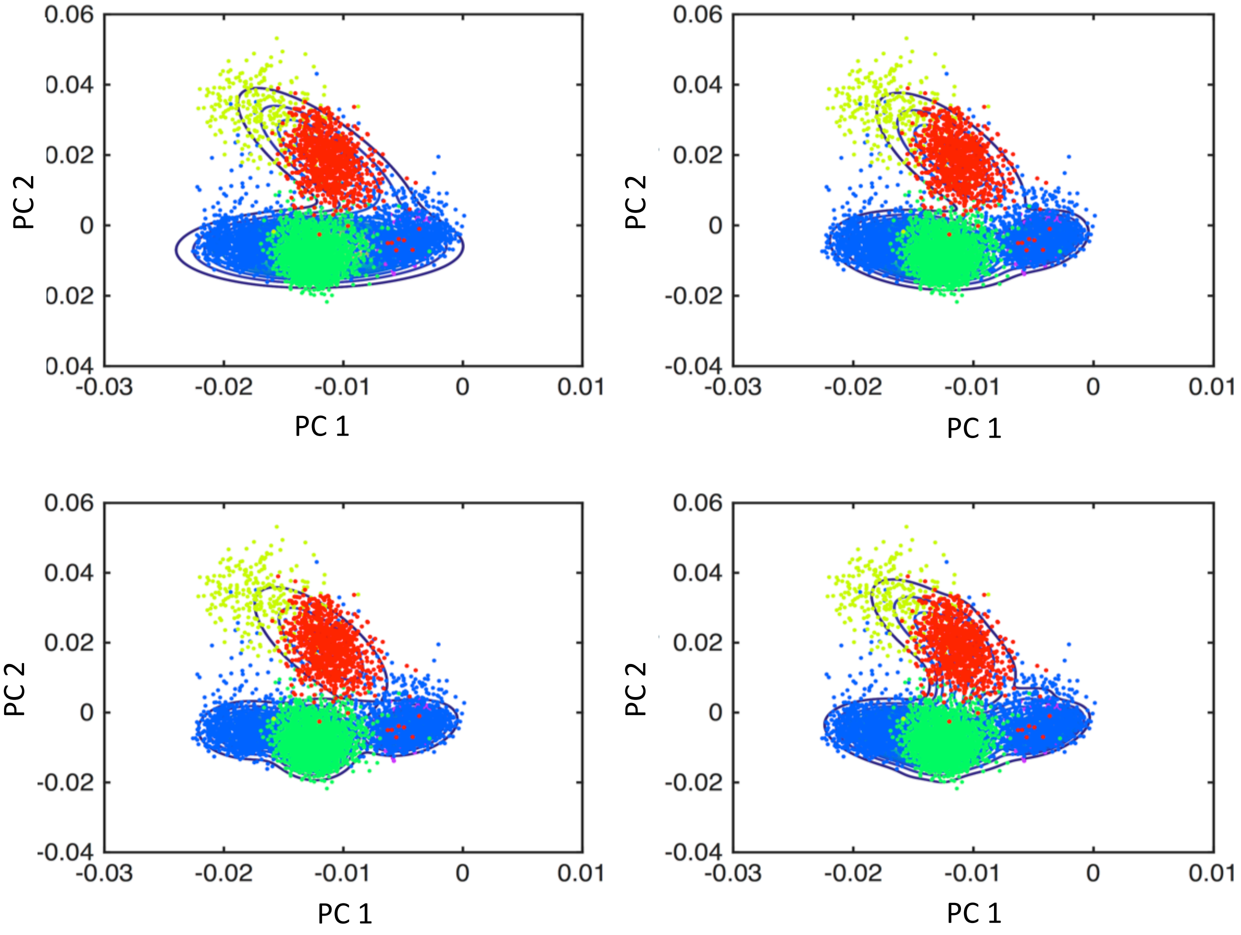

Figure 5. Gaussian mixture model clustering of human and mouse cell types where top left: two components, top right: three components, bottom left: four components and bottom right: five components.

The cell types predicted by the three component GMM were representative of the original cell type labels. The interneuron GMM had an odds ratio of 2.00×10^3 and confidence interval of (1.16×10^3,3.46×10^3), the pyramidal GMM had an odds ratio of 9.93×10^2 and a confidence interval of (6.84×10^2, 1.44×10^3), and the glial/vascular GMM had an odds ratio of 1.15×10^2 and a confidence interval of (91.34, 1.44×10^2) (Figure 6). The GMM cluster for glial/vascular cells had a higher false negative rate than the other GMM clusters due to incorrect clustering of glial/vascular labeled mouse cells.

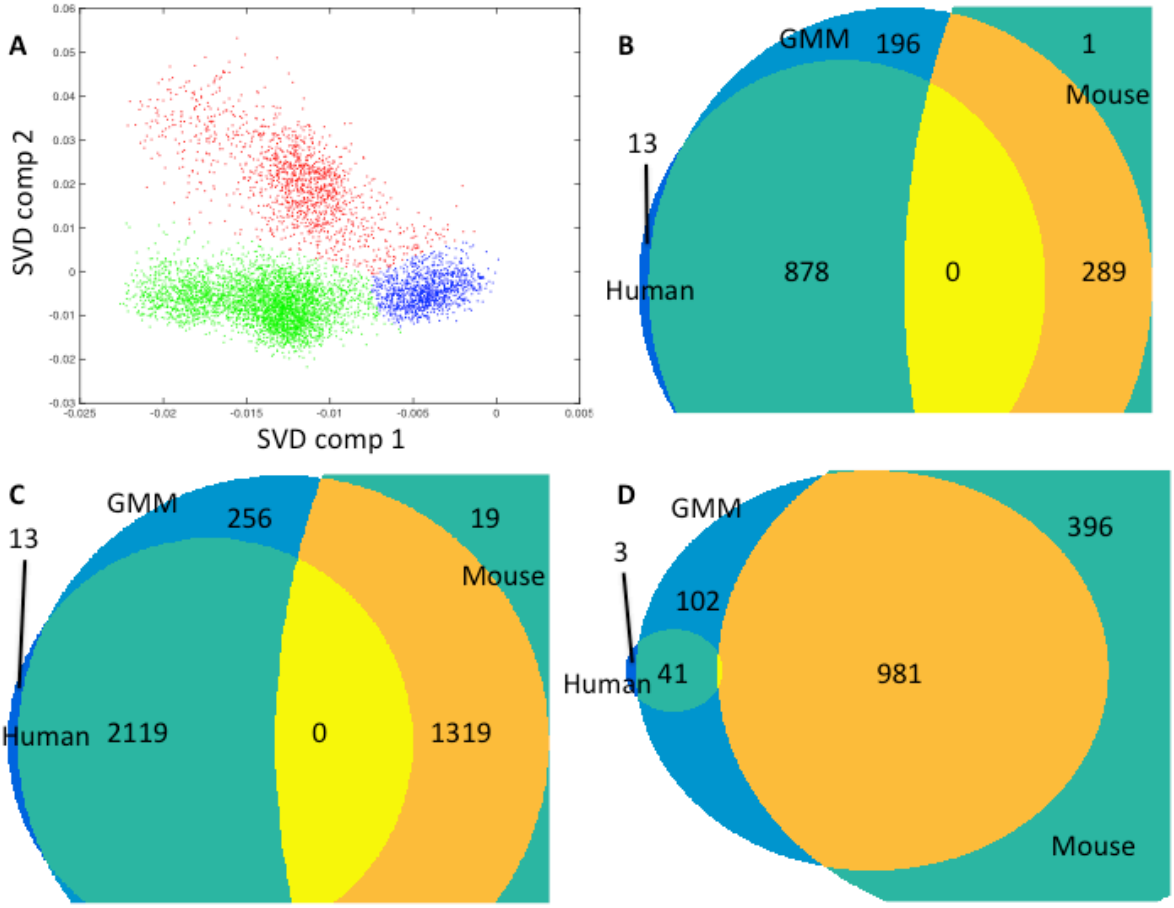

Figure 6. Comparing GMM clustering of human and mouse cells versus reported cell types. A) SVD components colored by GMM predicted clusters red (interneurons), green (pyramidal) and blue (mural/vascular). B-D are Venn diagrams comparing reported human and mouse cell types with GMM predicted cell types. The following superscripts represent if the point was included + or excluded – from species and GMM cluster. The colors from left to right consist of Human^{+}-GMM^{-} (blue), Human^{+}-GMM^{+} (green), Human^{-}-Mouse^{-}-GMM^{+} (blue), Null set (yellow), Mouse^{+}-GMM^{+} (orange), Mouse^{+}-GMM^{-} (green). B) GMM Interneurons cluster (red in panel A) with mouse and human interneuron labeled cells. C) GMM Pyramidal (green in panel A) with mouse and human pyramidal labeled cells ("CA1, S1" and "Ex" respectively). D) GMM Glial/Vascular (blue in panel A) with mouse glial/vascular labeled cells and human "NoN" labeled cells.

4. Discussion

4.1. *Insights*

In this study we found that through feature selection it is possible to find informative gene sets that can be used across species. This feature selection of "concordant gene sets" is an important application of single cell data that has multiple downstream applications in relation to cross species modeling, especially in translation of preclinical studies. It is important to note that the data used to find the concordant genes cannot be paired by sample which makes correlation matrices impossible to generate. Without correlation matrices to discover concordant genes, the gene sets must be derived from ulterior methods such as minimizing redundant gene sets through machine learning [17] or grouped statistical tests like ANOVA.

4.1.1. *Scalability*

The feature selection method is based on ANOVA which is calculated across multiple groups. Unlike t-tests, this facet of ANOVA makes the feature selection method scalable in relation to number of species and cell types being studied. Because of this, finding concordant gene sets between many organisms and cell types simultaneously is possible and should be pursued.

4.1.2. *Functional relevance*

The annotated concordant gene set had a clear relationship to the brain through gene ontology which is an important control due to the tissue origin [18]. It is important to note that gene sets with no functional overlap to the phenotype being selected for could potentially be selecting for unknown associated phenotypes. The functional ontology analyses of this concordant gene set shows that there is selection of genes with direct relation to neuronal phenotypes. Because of the enrichment of phenotypically similar ontology terms, a case can be made that seemingly phenotypically dissimilar ontology terms are more likely to have an unknown but direct relationship to our concordant gene set.

4.1.3. *Evolutionary potential*

Concordant gene sets also contain unique evolutionary information. Gene homologs which express differently between two species (Discordant genes) potentially do not share exactly the same functionality. Discordant genes may have the same down-stream effects but the biological mechanism may have changed [6] such that the same quantity of mRNA is not produced across species. Concordant genes are informative because they could represent pathways that are relatively conserved between through the evolution of species.

4.1.4. *Medical and research potential*

In the medical realm concordant gene sets could be of use in translational research. Much of research is conducted in model organisms and using concordant gene sets gives the user an ability to distinguish between transcriptional changes that likely cause similar phenotypes or likely do not between the model and human. Though we do not immediately condone the clinical use of concordant genes at the present these concordant gene sets could help to quickly and efficiently integrate cross-species knowledge to improve translational research.

4.1.5. *Future work*

The scalability of cell type and species number should be tested upon the arrival of comparable data in other species. Aside from the direct feature selection of concordant genes multiple comparisons could be carried out to create hierarchical concordant gene sets for higher granularity. Another option to improve granularity would be to test models that include interaction variables between species, brain location, and cell type. With the generation of concordant gene sets cross-species deconvolution could become more accurate than with more heuristic approaches. Also concordant gene sets can be used in classification of cell types across species. With further refinement of the procedure human cell types could be classified using mouse expression profiles which would require refinement of feature selection and of classification algorithms and validation of such methods on another dataset.

4.1.6. *Importance of single cell granularity*

Single cell technologies in the form of fluorescence-activated cell sorting (FACS) and flow cytometry have been effectively used to model cell heterogeneity [19] before the advent of single cell transcriptomics. Through FACS sorting [20] and flow cytometry [21] deriving the transcriptome of a single cell is much higher throughput than original methodologies that required manual isolation of single cells [22]. Without the single cell granularity of these techniques, it would be impossible to study concordant genes effectively at the cellular level and acquire the sample sizes large enough to properly study concordant gene sets, especially when many species and phenotypes are involved. Only through these recent advances in scRNA-Seq is it possible to properly glean enough information about cell types to model across species.

4.2. *Limitations*

There are some limitations to this study, which included the use of zscores as the measurement of expression. This measurement makes the assumption that the data has a normal distribution. Because of the nature of scRNA-Seq data the distribution is negative binomial. It was important to use zscores because other normalization techniques would not be effective. Quantile normalization introduced artificats in the data that made it unrepresentative. Conversion of UMI counts to TPM alos posed a problem because TPM is based on aligned reads opposed to tag counts from UMIs.
Aside from normalization, the diversity of cell types in each dataset also potentially introduced bias. The human dataset consisted of fewer major cell types than the mouse dataset. The mouse dataset contained more glial cell types while the human dataset had higher granularity within interneurons and pyramidal cells.

5. Conclusion

We were able to find a concordant gene set between mouse and human brain cells that had direct functional ontology relationships to the brain. The concordant gene set allowed us to reduce the distance between cell types of different species allowing separation of cell type regardless of each cell's species. Through the study of these aggregate cell types the biologically unresolved human cell type "NoN" (No Nomenclature) was able to be categorized as Glial/Vascular. Furthermore we show that our methodology is scalable to multiple species and cell types to find concordant gene sets between multiple species and these concordant genes sets are important stepping stones toward evolutionary and translational research goals.

6. Acknowledgements

This work is partially supported by the Human Frontier Science Program (RPG0053/2014) and the NLM T15 training grant. The Ohio Supercomputer Center provided computing support.

References

[1] S. Lin, Y. Lin, J. R. Nery, M. A. Urich, A. Breschi, C. A. Davis, A. Dobin, C. Zaleski, M. A. Beer, W. C. Chapman, T. R. Gingeras, J. R. Ecker, and M. P. Snyder, *Proc. Natl. Acad. Sci.*, **111**, 17224 (2014).

[2] P. P. C. Tan, L. French, and P. Pavlidis, *Front. Neurosci.*, **7**, 1 (2013).

[3] K. Taeho, G. S. Vidal, M. Djurisic, C. M. William, M. E. Birnbaum, C. K. Garcia, B. T.

Hyman, and C. J. Shatz, **341**, 1399 (2013).

[4] S. Matsuda, M. Katane, K. Maeda, Y. Kaneko, Y. Saitoh, T. Miyamoto, M. Sekine, and H. Homma, *Amino Acids* **47**, 975 (2015)

[5] J. W. Rowley, A. J. Oler, N. D. Tolley, B. N. Hunter, E. N. Low, D. a Nix, C. C. Yost, G. a Zimmerman, and A. S. Weyrich, *Blood* **118**, 101 (2011).

[6] J. a Miller, S. Horvath, and D. H. Geschwind, *Proc. Natl. Acad. Sci.* **107**, 12698 (2010).

[7] S. a. Fietz, R. Lachmann, H. Brandl, M. Kircher, N. Samusik, R. Schroder, N. Lakshmanaperumal, I. Henry, J. Vogt, a. Riehn, W. Distler, R. Nitsch, W. Enard, S. Paabo, and W. B. Huttner, *Proc. Natl. Acad. Sci.* **109**, 11836

[8] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll, *Cell* **161**,1202 (2015).

[9] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suvà, A. Regev, and B. E. Bernstein, *Science* **344**, 1396 (2014).

[10] A. Zeisel, A. B. M. Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, and S. Marques, *Science* **347**, 1138 (2015).

[11] B. B. Lake, R. Ai, G. E. Kaeser, N. S. Salathia, Y. C. Yung, R. Liu, A. Wildberg, D. Gao, H.-L. Fung, S. Chen, R. Vijayaraghavan, J. Wong, A. Chen, X. Sheng, F. Kaper, R. Shen, M. Ronaghi, J.-B. Fan, W. Wang, J. Chun, and K. Zhang, *Science* **352**, 1586 (2016).

[12] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson, *Nature methods* **11**, 163 (2014).

[13] D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. a Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, and R. Sandberg, *Nat. Biotechnol.* **30**, 777 (2012).

[14] D. W. Huang, R. a Lempicki, and B. T. Sherman, *Nat. Protoc.* **4**, 44 (2009).

[15] D. W. Huang, B. T. Sherman, and R. A. Lempicki, *Nucleic Acids Res.* **37**, 1 (2009).

[16] P. D. McNicholas and T. B. Murphy, *Bioinformatics* **26**, 2705 (2010).

[17] C. Ding and H. Peng, *Journal of Bioinformatics and Computational Biology* **3**, 185 (2003).

[18] N. A. Twine, K. Janitz, M. R. Wilkins, and M. Janitz, *PLoS One* **6,** e16266.

[19] P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs, R. V Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis, *Nat. Biotechnol.* **29**, 886–891 (2011).

[20] N. K. Wilson, D. G. Kent, F. Buettner, M. Shehata, I. C. Macaulay, F. J. Calero-Nieto, M. Sanchez Castillo, C. A. Oedekoven, E. Diamanti, R. Schulte, C. P. Ponting, T. Voet, C. Caldas, J. Stingl, A. R. Green, F. J. Theis, and B. Gottgens, *Cell Stem Cell* **16**, 712 (2015).

[21] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit, *Science* **343**, 776 (2014).

[22] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani, *Nat. Methods* **6**, 377 (2009).