Human iPSC glial mouse chimeras reveal glial contributions to schizophrenia

Martha S. Windrem[1], Mikhail Osipovitch[2], Su Wang[1], Janna Bates[1], Lisa Zou[1], Zhengshan Liu[1], Jared Munir[1], Devin Chandler-Militello[1], Robert Miller[5], Maiken Nedergaard[1,2], Robert Findling[3], Paul J. Tesar[5], Steven A. Goldman[1,2]

[1]Center for Translational Neuromedicine, University of Rochester Medical Center, Rochester, NY 14642; [2]Center for Translational Neuroscience, University of Copenhagen Faculty of Health Sciences, 2200 Copenhagen N, Denmark; [3]Dept. of Psychiatry, Johns Hopkins Univ. School of Washington University Medical Center, Washington, DC.

Materials and Methods for SCZ RNA-Seq Analysis

mRNA was isolated by polyA-selection protocol from FACS-sorted PDGFRa-positive glial progenitor cell (GPC) lines produced from iPS cells derived from 4 schizophrenic patients (designated to SCZ lines 8 [N = 4 independent cell preparations], 29 [N = 3], 51 [N = 7], and 164 [N = 8]) and 3 healthy controls (designated to CTR lines 22 [N = 3], 37 [N = 4], and 205 [N = 7]). Sequencing libraries were prepared using TruSeq RNA v2 kit and sequenced on Illumina HiSeq 2500 platform for approximately 45 million of 100-bp singe-end reads per sample. The sequencing reads were then pre-processed by trimming off adapter and low-quality sequences from 3' end using Trimmomatic [3]. The quality of reads before and after pre-processing was assessed with FastQC [1]. The pre-processed reads were then aligned to the RefSeq NCBI reference [10] human genome version GRCh38 with Subread read aligner [7] using Hamming distance to break ties between more than one optimal mapping locations. Raw gene counts were obtained from BAM alignment files with featureCounts [8]. After eliminating lowly expressed transcripts with a count of fewer than 5 reads in more than 5 samples across the gene by sample matrix, the count data were normalized using RUVSeq [12] R Bioconductor [4] package to account for variance. As described in RUVSeq manual, the normalization was accomplished in the following three-step procedure: *(1)* negative *in silico* control genes were determined by first-pass differential expression analysis by edgeR [13] and DESeq2 [9] R Bioconductor packages taking genes with FDR-adjusted P Values of above 0.75 as calculated by both methods (approximately 7000 genes unaffected by the condition of interest); *(2)* the negative control genes were then used in RUVs function of RUVSeq package for calculation of variance factors; and *(3)* the second-pass differential expression analysis (5% FDR and log2 fold change > 1) for determination of disease-dysregulated genes was performed using the original raw counts with adjusting for RUVs-calculated variance factors by multi-factor GLM models implemented in edgeR and DESeq2 packages. The filtering for lowly expressed transcripts and the three-step analysis procedure were employed for comparison of each SCZ-derived GPC cell line to pooled CTR-derived GPCs. The intersection of the resulting four lists of differentially expressed genes was taken as the conserved representative list of SCZ-dysregulated genes. In the normalization procedure for each comparison, the number of variance factors was limited to 1 for line 29, 3 for lines 8 and 164, and 7 for line 51, as determined by principal component and hierarchical clustering analyses performed with native R functions [11].

To obtain average fold changes and P Values for dysregulated genes in all four SCZ-derived GPC lines, a differential expression comparison of pooled SCZ to pooled CTR lines was performed by the same workflow with the number of variance factors limited to 9. For all differential expression comparisons, only the significant results that agreed between edgeR and DESeq2 methods were used in downstream analysis. Fold changes and FDR-adjusted P Values reported in the results section were calculated by edgeR. Functional annotation of the conserved set of SCZ-dysregulated genes was performed in ToppCluster [6] and Ingenuity

Pathway Analysis (IPA) software [5]. Network visualization and analysis of the results of functional annotation were performed in Gephi [2] graph visualization software. R code and data files to reproduce the normalization and differential expression workflow are available from https://github.com/cbtncph/GoldmanetalSCZ2016.

Citations in Materials and Methods for SCZ RNA-Seq Analysis

[1] Andrews S. (2010). FastQC: A quality control tool for high throughput sequence data. *Reference Source*.

[2] Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM 8*, 361-362.

[3] Bolger A.M., Lohse M., & Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170.

[4] Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., ... & Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), R80.

[5] Ingenuity Pathway Analysis, QIAGEN Redwood City, www.qiagen.com/ingenuity

[6] Kaimal V. et al. (2010). ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Research*, gkq418.

[7] Liao Y., Smyth G.K., Shi W. (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10):e108.

[8] Liao Y., Smyth G.K., Shi W. (2014). featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923-30.

[9] Love M.I., Huber H., & Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome biology*, 15(12), 1-21.

[10] Pruitt K.D., Tatusova T., & Maglott D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(suppl 1), D61-D65.

[11] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[12] Risso D., Ngai J., Speed T., Dudoit S. (2014). "Normalization of RNA-seq data using factor analysis of control genes or samples." *Nature Biotechnology*, 32(9), pp. 896–902.

[13] Robinson M.D., McCarthy D.J., & Smyth G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139-140.