Data-driven identification of total RNA expression genes for estimation of RNA abundance in heterogeneous cell types highlighted in brain tissue

Louise A. Huuki-Myers[1], Kelsey D. Montgomery[1], Sang Ho Kwon[1,2], Stephanie C. Page[1], Stephanie C. Hicks[3], Kristen R. Maynard[1,4*] and Leonardo Collado-Torres[1,3*]

*Correspondence:
kristen.maynard@libd.org;
lcolladotor@gmail.com

[1] Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, USA
[2] The Solomon H. Snyder Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD, USA
[3] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[4] Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD, USA

Abstract

We define and identify a new class of control genes for next-generation sequencing called total RNA expression genes (TREGs), which correlate with total RNA abundance in cell types of different sizes and transcriptional activity. We provide a data-driven method to identify TREGs from single-cell RNA sequencing data, allowing the estimation of total amount of RNA when restricted to quantifying a limited number of genes. We demonstrate our method in postmortem human brain using multiplex single-molecule fluorescent in situ hybridization and compare candidate TREGs against classic housekeeping genes. We identify AKT3 as a top TREG across five brain regions.

Keywords: RNA abundance, RNAscope, snRNA-seq, TREG, Deconvolution, Bioconductor

Background

In genomic analyses, researchers frequently face the decision of whether to use a list of genes identified a priori for an analysis or to identify new genes in a data-driven manner that have specific desirable qualities to answer a biological question. This duality reflects the nature of how our knowledge evolves as experimental assays generate more data and provide further insight into our understanding of biological systems. This expansion of knowledge is reflected in approaches such as single cell or nucleus RNA sequencing (sc/snRNA-seq) where known cell-type marker genes are used to annotate cells, and the annotations are used to find new cell-type marker genes [1–3]. Similarly, in spatially resolved transcriptomics, previous knowledge of genes with distinct spatial expression can be used to annotate cells in situ but also identify anatomical domains leading to the identification of new spatially variable gene sets [4, 5].

Huuki-Myers *et al. Genome Biology*   (2023) 24:233

Page 2 of 21

Methods for gene selection, either data-driven or based on previous knowledge from the literature [6], are not only relevant to genes with high variability but also to identify "control" genes with stable levels of expression used, for example, in normalization, such as microarray channel [7] or quantitative PCR normalization [8]. One data-driven approach to identify control genes for these assays when samples contain different amounts of RNA is to rely on a rank-invariant approach [9].

Different cell types contain variable amounts of RNA due to differences in cell size and transcriptional activity. In brain tissue, this variation in cell size and RNA abundance can negatively impact the accuracy of bulk RNA-seq deconvolution methods, which aim to identify cell type proportions in homogenate tissue by using sc/snRNA-seq reference profiles [10]. For example, neurons are larger and more transcriptionally active than glia and therefore have more RNA content and more genes detected per nucleus in snRNA-seq data [11]. With the exception of two methods [12, 13], the majority of existing bulk RNA-seq deconvolution methods [10] fail to incorporate this variation and hence report potentially biased estimates of the *relative* fraction of RNA attributable to each cell type rather than the true proportion of cell types [14]. This is a challenge for bulk RNA-seq deconvolution algorithms that have to properly take into account heterogeneous cell types when estimating the proportions of cell types [15]. However, methods to robustly estimate cell or nuclear size and total RNA abundance in the same assay are limited as approaches that capture global RNA expression, such as snRNA-seq, require tissue homogenization preventing the acquisition of cell size measurements. Here, we introduce total RNA expression genes (TREGs), which can be used in combination with histological approaches such as single molecule fluorescence in situ hybridization (smFISH) to link estimates of cell size and total RNA abundance in the same assay.

One approach to measure nuclear size and relative RNA abundance is to use RNAscope smFISH technology [16], which allows quantification of both cell morphology and gene expression for a small number of target genes. Specifically, RNAscope fluorescently labels individual RNA transcripts, which are represented as "dots" or puncta in the image that can be segmented and used to quantify gene expression per nucleus [17]. In parallel, these images can be used to estimate spatially resolved nuclear size across heterogenous cell types in situ. A TREG can be used in combination with RNAscope to estimate total RNA expression in differently sized cell types by utilizing one channel of the multiplex assay. However, there are no rigorous and data-driven approaches to identify candidate target genes to estimate total RNA abundance compatible with smFISH assays, such as RNAscope.

Here, we propose a data-driven approach using sc/snRNA-seq data to identify TREGs to estimate total RNA abundance in heterogeneous cell types. These genes should ideally be highly correlated with total RNA abundance and predictive of transcriptional activity (Fig. 1a). In the postmortem human brain, single-unit measurements are limited to the nucleus, but it has been established that nuclear RNA content is representative of the whole cell [18]. In other research settings, single-unit measurements could encompass the whole cell using scRNA-seq. When TREGs are applied in smFISH using RNAscope, they can be used to link spatially resolved size and total RNA expression in different cells.
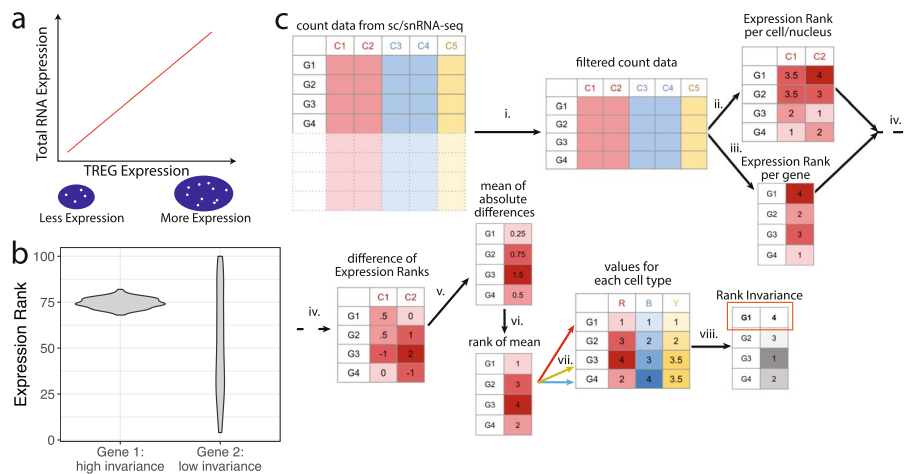
Fig. 1 Overview of TREG motivation and methodology. **a** Illustration of the relationship between the expression of a TREG and the total RNA expression of a nucleus. TREG expression can be quantified with puncta (white dots) in a nucleus (blue area), where the nucleus is identified with DAPI. **b** Illustration of the distribution of expression rank, which is the rank of the expression of a given gene among all genes, computed individually for each cell/nucleus, depending on the measurement technology used: sc or snRNA-seq. Two theoretical genes are shown: gene 1 with high rank invariance and gene 2 with low rank invariance across cells/nuclei. **c** Rank invariance workflow to identify a TREG (the "Rank invariance calculation" section), with a gene expression matrix with genes on the rows and cells/nuclei on the columns. (i) Filter for low-expressed genes (the "Expression and proportion zero filtering" section). Onward working with one cell type at a time; (ii) compute expression rank of each cell/nucleus for each gene (example distribution in *b*); (iii) calculate the mean gene expression across all cells/nuclei for one cell type and then its rank expression; (iv) per gene, find the difference of the rank expression against the mean rank expression for each cell/nucleus in a given cell type; (v) calculate the mean of the absolute expression rank differences for each gene; (vi) rank the mean absolute expression rank differences; (vii) repeat steps ii–vi for each cell type; (viii) per gene, compute the sum of the previous ranks across all cell types and then rank these sums across genes such that the highest rank is given to the gene with the smallest sum. This is the final rank invariance value

Using sc/snRNA-seq data, we define a candidate TREG as a gene that (1) has non-zero expression in most cells/nuclei across groups of interest, such as tissue-specific cell types, and (2) is expressed at a constant level with respect to other genes across different cell types of a given tissue. To be compatible with RNAscope, candidate TREGs also meet the following criteria: (1) expressed in the top 50% of genes for easy detection, (2) have a dynamic range of puncta to provide a continuous metric, and (3) expressed at a level that individual puncta can be accurately counted.

While TREGs theoretically share some similarities with classical housekeeping (HK) genes, such as being expressed in every cell, they have other distinct properties. By definition, TREGs are tissue-specific and are associated with total RNA expression. In other words, TREGs are identified in one reference dataset specific to an experimental condition; therefore, TREGs are not necessarily generalizable to other experimental conditions. Furthermore, they are not defined by the function of the protein they encode. In contrast, classic HK genes are associated with cell maintenance, tissue agnostic, and expressed at a constant level regardless of cell type and condition [19].

While TREG is a general method, our research focus is motivated by understanding the transcriptional landscape in the human brain and identifying changes associated with psychiatric disorders [20]. We are interested in identifying a TREG that could

be used in multiple cortical and subcortical brain regions linked to psychiatric disorders [20]. We focused on broad cell type categories that are diverse across size and expression levels and are frequently present in these brain regions [20, 21]. With this in mind, we demonstrated the use of TREGs by applying our approach to snRNA-seq data from five brain regions, with focused RNAscope analyses in the dorsolateral prefrontal cortex (DLPFC). We compared candidate TREGs against classic HK genes and identified *AKT3* as the best-performing TREG in the DLPFC. To identify candidate TREGs in other tissues, we provide open-source software available as an R/Bioconductor package at https://bioconductor.org/packages/TREG.

Results

Overview of method to identify TREGs

Our approach to identify total RNA expression genes (TREGs) was inspired by rank-invariance methods originally developed for microarrays that were used to identify stably expressed genes within normalization methods applied to unbalanced transcriptome data (or containing different amounts of RNA) [7–9]. Briefly, after applying a filter to remove lowly expressed genes in a given sc/snRNA-seq reference dataset, our approach compares the ranks of expression across cells/nuclei (rather than comparing the gene expression values themselves across cells of different sizes) and identifies genes that are consistently ranked (or high "rank invariance") (Fig. 1b). In our algorithm, to identify a TREG, we compared the stability of each gene's expression rank within and across cell types to identify high rank invariant genes (Fig. 1c, the "Rank invariance calculation" section). Genes consistently expressed in all cells/nuclei across all cell types were identified by high rank invariance values and were considered TREG candidates. We implemented our data-driven method in an open-source R/Bioconductor package (https://bioconductor.org/packages/TREG) [22] to identify candidate TREGs in any sc/snRNA-seq dataset. The package includes functionality for both gene filtering and rank invariance methods.

Datasets and TREG experiment overview

We applied our method to identify TREGs in a publicly available snRNA-seq dataset from the human postmortem brain. Specifically, the dataset included 70,527 nuclei from eight donors across five brain regions [20]. We identified candidate TREGs among 10 broad cell types across these brain regions: amygdala (AMY), dorsolateral prefrontal cortex (DLPFC), hippocampus (HPC), nucleus accumbens (NAc), and subgenual anterior cingulate cortex (sACC) (the "snRNA-seq reference data" section, Additional file 1: Table S1). Gene expression from top candidate TREGs was measured with smFISH using RNAscope technology and compared to a classic housekeeping gene, *POLR2A* [23], to evaluate TREG predictiveness of total RNA expression.

Filtering genes from the snRNA-seq data in the postmortem human brain

To maximize detection compatibility with the RNAscope assay, the expression data was filtered for highly expressed genes, specifically the top 50% of the 23,038 genes in the snRNA-seq dataset, retaining 11,519 genes. Genes were also filtered to remove those with a high maximum proportion zero (ranges between 0 and 1) expression across all cell type and brain region combinations (the "Expression and proportion zero filtering"

section). The proportion zero filtering process avoids rank ties in the downstream steps due to the high number of genes with no expression. A high proportion zero also suggested that a gene may not be observable in most nuclei in that population using RNAscope. Frequently, nuclei from a specific cell type and brain region combination had a high frequency of genes whose proportion zero exceeded 0.75 for common cell types including astrocytes, microglia, oligodendrocytes, oligodendrocyte precursor cells, excitatory and inhibitory neurons (Fig. 2a), and for more rare cell types including endothelial, macrophages, mural, and T-cells (Additional file 2: Fig. S1). After filtering the genes for a maximum proportion zero of less than 0.75 across all cell types and region combinations, 877 genes remained (3.8% out of the initial 23,038 genes). The classic housekeeping gene *POLR2A* showed a high proportion zero in many cell types across brain regions and did not pass this filtering step unlike *AKT3*, *ARID1B*, and *MALAT1* (Fig. 2b, Additional file 2: Fig. S2a).

Identification of TREG candidates in the postmortem human brain

After applying the filtering steps, the rank invariance workflow (Fig. 1c) was applied to the five brain regions in the postmortem human brain to identify candidate TREGs (the "Rank invariance calculation" section). From the top ten rank invariance values, we
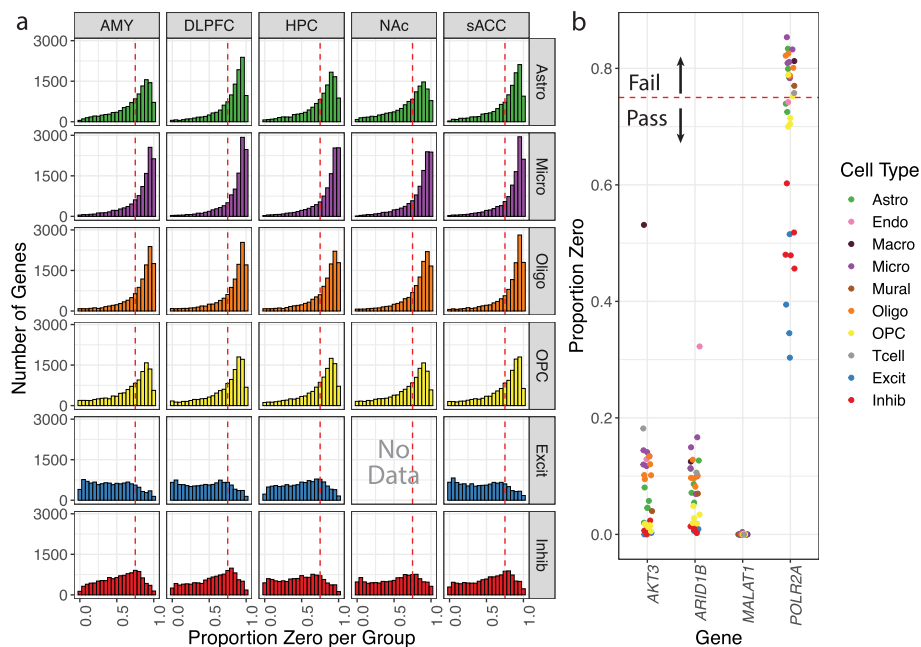

Fig. 2 Overview of the proportion zero filtering process. **a** Histogram frequency of proportion zeros for each nuclei population for a given cell type and brain region combination. These combinations are arranged with cell types along the rows [astrocytes (Astro), microglia (Micro), oligodendrocytes (Oligo), oligodendrocyte precursor cells (OPC), excitatory (Excit), and inhibitory neurons (Inhib)] and by brain region along the columns [amygdala (AMY), dorsolateral prefrontal cortex (DLPFC), hippocampus (HPC), nucleus accumbens (NAc), and subgenual anterior cingulate cortex (sACC)]. Consistent with the inhibitory neuron-rich cell type composition of the NAc, there were no excitatory neurons found in this region and therefore no data to report. The red dashed line represents the 0.75 cutoff for filtering. **b** Proportion zero filtering process detailed for *AKT3*, *ARID1B*, and *MALAT1* compared to the classic HK gene *POLR2A*. If any cell type and brain region combination (individual colored points) has a proportion zero > 0.75, then the gene fails the filtering step. Unlike *AKT3*, *ARID1B*, and *MALAT1*, *POLR2A* fails proportion zero filtering

selected three candidate TREGs (*AKT3*, *ARID1B*, and *MALAT1*) for further evaluation based on the commercial availability of RNAscope probes. *MALAT1* was the top rank invariance gene and also the gene with the highest mean expression. The expression rank of these TREGs has a small variance across 70 k nuclei (Fig. 3a), as well as within different cell types (Fig. 3b). This is in contrast to the HK gene *POLR2A*, which shows a more variable expression rank distribution (Fig. 3b). We note that this same relationship holds if we compare the distribution of log-transformed gene expression across cell types, which is more variable than using the expression rank distribution (Additional file 2: Fig. S3).

By definition, the expression of a TREG should be predictive of the total expression of RNA in a cell (or nuclear expression when limited to snRNA-seq data). We compared the relationship between the gene expression of a TREG and the total expression of RNA in a nucleus (estimated by the log2 sum of all counts) and quantified the strength of the association by fitting a linear model for all nuclei within each cell type. We found a strong linear relationship for *MALAT1* (Fig. 3c), *AKT3*, and *ARID1B* (Fig. 3, Additional file 2: Fig. S4, Additional file 1: Table S3; the "Total RNA linear regression" section). Among the genes passing the proportion zero filter, the strength of their association with total RNA expression generally increased as their rank invariance increased (Additional file 2: Fig. S2b). Furthermore, Gene Ontology enrichment analysis with the top 20 or 50 candidate TREGs showed that these genes are enriched for key biological processes such as chromatin binding and transcription regulator activity (Additional file 2: Fig. S5; the "Gene Ontology and KEGG pathway enrichment analysis" section). Among the GO-enriched terms, *ARID1B* was a very frequent contributor, whereas for the KEGG-enriched pathways, *AKT3* was the principal contributor (Additional file 1: Table S2).

We ran the filtering process and TREG candidate identification independently for each of the five brain regions and identified the top 50 rank invariance genes (Additional file 1: Table S3). We identified 13 TREGs common across all five brain regions; therefore, for the main analysis, we used the combined dataset (Additional file 2: Fig. S6). The top 13 TREGs across the brain regions included *AKT3*, *ARID1B*, and *MALAT1*.

Identification of TREGs in a case–control dataset

To examine the stability of the method using data collected from a disease state and control donors, we performed the rank invariance workflow on a snRNA-seq dataset containing samples from neurotypical donors and donors with autism spectrum disorder from Velmeshev et al. [24]. We identified a list of the top 50 candidate TREGs from this dataset as a whole, then separately for the case and control samples. We found that 34 genes (50.7% of 67 unique genes between all sets) were present in all three sets of candidate TREGs (Additional file 2: Fig. S7a). This shows that the rank invariance workflow is able to identify the same TREG candidates between disease states. *MALAT1* was the top rank invariance gene in all three conditions. We also noted the gene *CADM2* was identified in the top 15 candidate TREGs across our evaluations of the Velmeshev et al. dataset, as well as in the top ten TREGs identified in our primary analysis in the Tran-Maynard et al. data. Like other highlighted TREG candidates, CADM2 shows expression rank stability across most cell types and between diagnosis (Additional file 2: Fig. S7b).
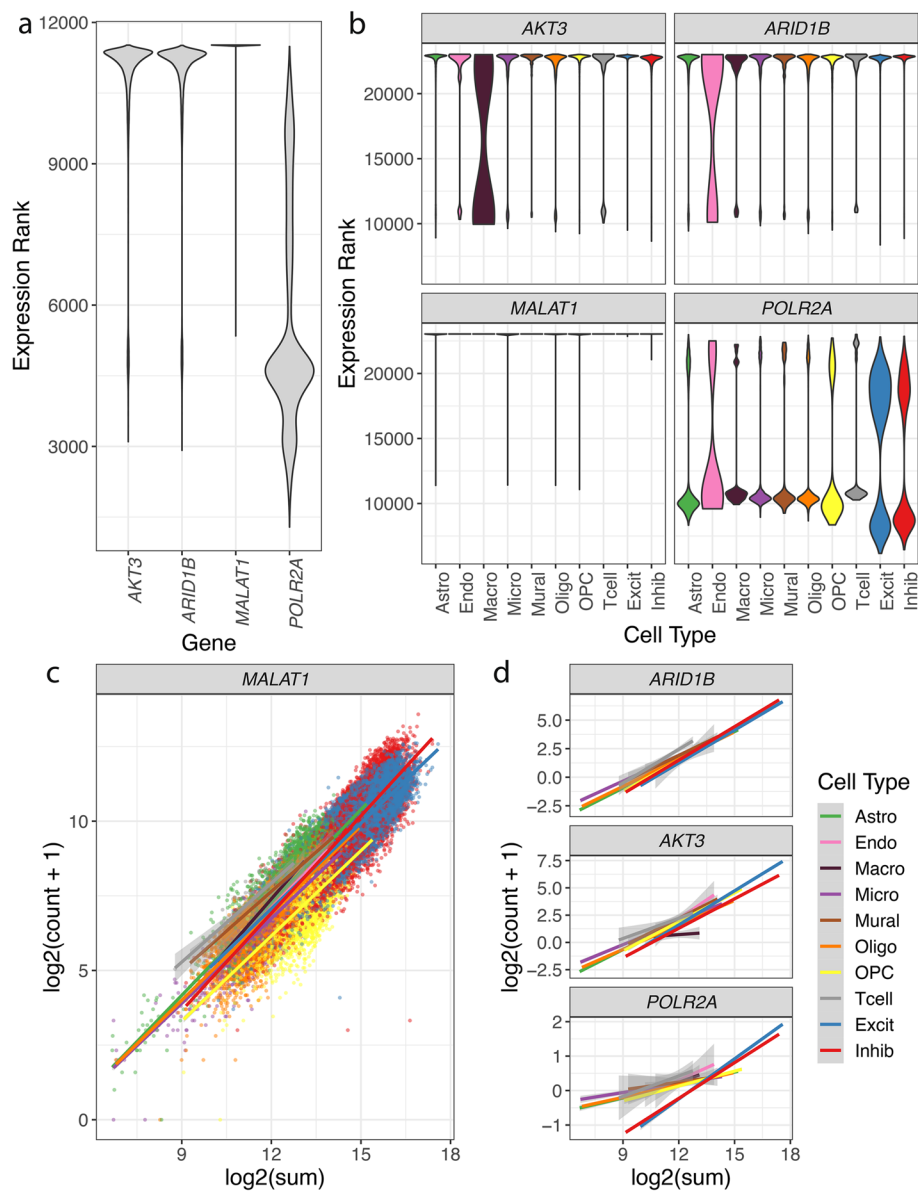
Fig. 3 Distribution of ranks and relationship between total nuclear expression and expression of candidate TREGs. **a** Distribution of the expression rank (*y*-axis) over all nuclei for genes *AKT3*, *ARID1B*, and *MALAT1* (three candidate TREGs) and *POLR2A* (a known HK gene). The candidate TREGs show higher rank invariance compared to *POLR2A* (related to Fig. 1b). **b** The distribution of the expression ranks (*y*-axis) over all cell types (*x*-axis) for the three candidate TREGs shows less expression rank variability across most cell types compared to *POLR2A*. **c** Scatter plot of the total RNA expression (estimated by the nuclei log2 sum of all counts) against the nuclei gene expression (log2 of the count plus one) for *MALAT1*, overlaid with the linear fit for each cell type and colored by cell type. **d** Linear fits of total nuclear RNA expression against the gene expression in the nuclei, similar to **c** for *POLR2A*, *AKT3*, and *AR1D1B*. The expression of candidate TREGs show consistent positive linear relationships with total RNA expression in each nucleus across all cell types, unlike *POLR2A* where the neurons have a different pattern than other cell types

Validation of TREGs with smFISH using RNAscope technology

Next, we chose to further evaluate TREGs in DLPFC tissue given its implication in several psychiatric disorders. We used multiplex fluorescent smFISH with RNAscope technology to label candidate TREGs *AKT3*, *ARID1B*, and *MALAT1* as well as HK gene

POLR2A in different cell types in DLPFC tissue sections from an independent brain donor ($n = 9$ tissue sections with 3 tissue sections per gene, Fig. 4a; the "Postmortem human tissue" section). We performed RNAscope with three probe combinations (Additional file 1: Tables S4 and S6; the "RNAscope multiplex single molecule fluorescent in situ hybridization (smFISH)" section). Each combination probed a TREG with a panel of cell type marker genes including *SLC17A7*, *GAD1*, and *MBP* (labeling excitatory neurons, inhibitory neurons, and oligodendrocytes, respectively). *POLR2A* and *MALAT1* were hybridized in the same experiment, and due to limitations in multiplexing, *GAD1* was omitted. Following high-magnification imaging, *AKT3*, *ARID1B*, and *POLR2A* transcripts were detected as discrete puncta (white dots) within individual nuclei (Fig. 5a–c; the "Image acquisition" section). However, due to high expression, individual puncta could not be observed for *MALAT1* and fluorescent signals were too saturated for quantification (Fig. 5d).

For TREGs showing discrete puncta, image segmentation and transcript quantification were performed using the HALO software (the "Image analysis with HALO" section). HALO identified 1,099,931 individual nuclei across the nine DLPFC tissue sections, with 80 k–109 k nuclei segmented per tissue section (Additional file 1: Table S7). After quality control for poorly segmented regions (the "Quality control and spatial quantification of HALO data" section, Additional file 2: Fig. S8), the number of nuclei per section was reduced to 68 k–106 k (Additional file 1: Table S7). We show accurate segmentation of
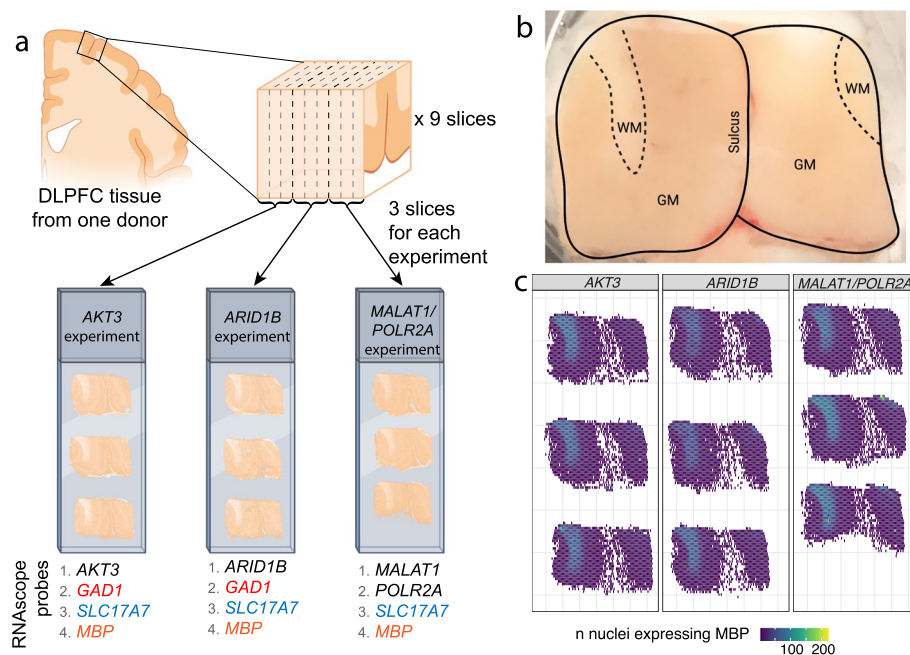

Fig. 4 Overview of the smFISH RNAscope experiment and DLPFC anatomy. **a** Illustration of RNAscope experimental design where a single DLPFC tissue block was used to generate 9 spatially adjacent slices. These 9 slices were hybridized with 3 RNAscope probe combinations noted as the *AKT3*, *ARID1B*, and *MALAT1/POLR2A* experiments (related to Additional file 1: Tables S4-S5). Candidate TREGs and *POLR2A* are shown in black, while *GAD1*, *SLC17A7*, and *MBP* are cell-type marker genes for inhibitory neurons (red), excitatory neurons (blue), and oligodendrocytes (orange), respectively. **b** Annotated image of DLPFC tissue, noting the location of gray matter (GM), white matter (WM), and sulcus. **c** Spatial distribution of cells expressing *MBP* for each sample. *MBP* is an oligodendrocyte cell type marker gene enriched in white matter
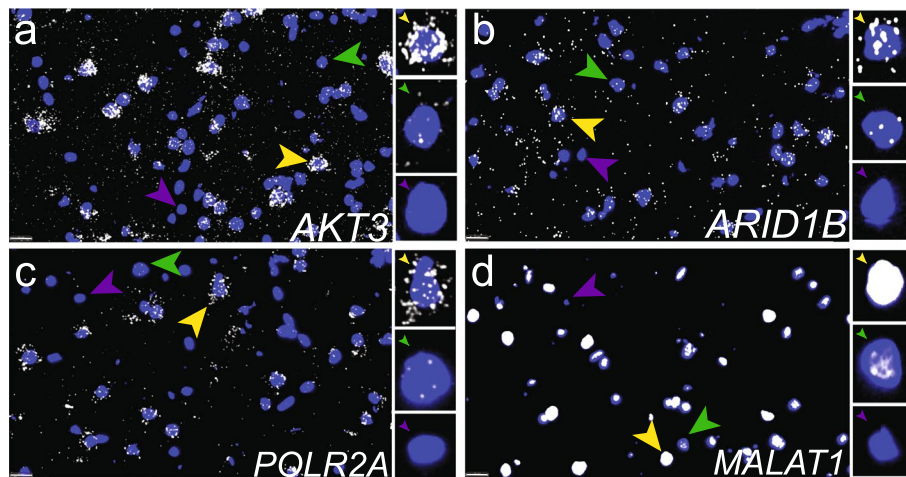
Fig. 5 Expression of TREGs in individual nuclei using smFISH with RNAscope. Representative high-magnification images showing the expression of candidate TREGs **a** *AKT3*, **b** *ARID1B*, **c** HK gene *POLR2A*, and **d** *MALAT1* and in human DLPFC. Insets show individual nuclei with high expression (yellow arrow), low expression (green arrow), or in rare cases (≤ 14% for candidate TREGs and 22% for *POLR2A*, Table 1) no expression (purple arrow). Each puncta represents a single transcript, as illustrated in Fig. 1a. *MALAT1* shows extremely high expression in the majority of nuclei such that individual puncta cannot be quantified (yellow arrow). Scale bar is 20 um

fluorescent signals in a representative DLPFC tissue section including neuron-enriched gray matter and glial-enriched white matter (Figs. 4b, c and 6). As expected, quantification of nuclear area based on DAPI signal confirmed that neurons located in the gray matter have a larger nuclear area than glia located in the white matter (Fig. 6a, Additional file 2: Fig. S8). Quantification of *AKT3* puncta within nuclei confirmed higher expression of AKT3 in neuron-enriched gray matter, which is consistent with higher transcriptional activity in neurons compared to glia (Fig. 6b). *ARID1B* and *POLR2A* also showed elevated expression in neurons located in the gray matter across the 3 different tissue sections (Additional file 2: Fig. S9). Segmentation of *SLC17A7* (excitatory neurons), *GAD1* (inhibitory neurons), and *MBP* (oligodendrocytes) fluorescent signals revealed the expected spatial distribution of these cell types within the gray and white matter (Fig. 6c).

Image segmentation and transcript quantification revealed that candidate TREGs were consistently expressed in the majority of nuclei across cell types. Specifically, TREG expression was recorded in over 86% of nuclei (Table 1). The HK gene *POLR2A* had puncta that could be quantified in 78% of nuclei with RNAscope, which was unexpected given that only 30% of nuclei had non-zero expression values in snRNA-seq. Additionally, *AKT3* and *ARID1B* provided a larger dynamic measurement range given that the mean puncta per nucleus is 4.09 and 2.08, respectively, compared to 2.75 for *POLR2A* (Table 1).

Next, we evaluated how the number of measured puncta (for each TREG) in a nucleus related to total nuclear RNA expression measured by snRNA-seq for excitatory neurons, inhibitory neurons, and oligodendrocytes in the DLPFC. DLPFC snRNA-seq data showed that excitatory neurons have the highest total nuclear RNA expression (estimated with the sum of total UMI counts per nucleus), followed by inhibitory neurons, and then oligodendrocytes (Fig. 7a). We quantified this pattern using the standardized
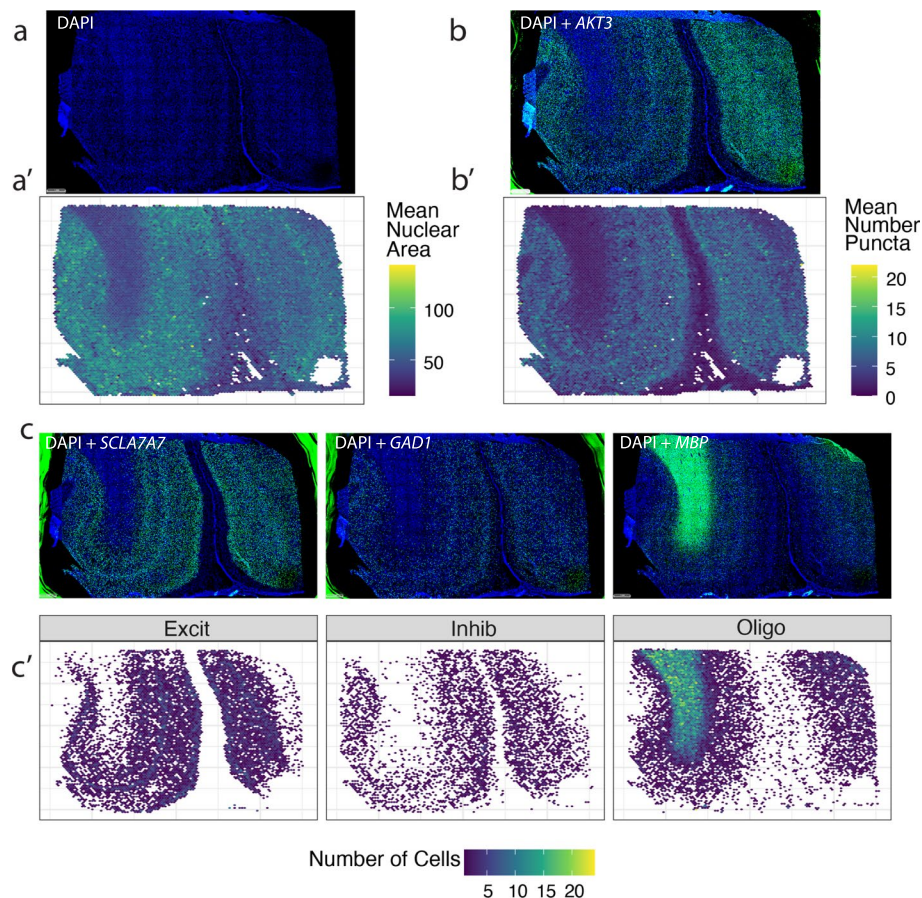
Fig. 6 Quantification of candidate TREG *AKT3* in differently sized cell types in human DLPFC. Representative tissue section showing **a** raw fluorescence for nuclear DAPI signals and **a'** corresponding mean nuclear area size. The nuclear area based on the DAPI signal shows larger excitatory neuron nuclei in gray matter from smaller glial nuclei in the white matter, related to Fig. 4b. **b** Raw fluorescence for DAPI and *AKT3* and **b'** corresponding quantification of the mean number of *AKT3* puncta per nucleus. **c** Raw fluorescence for *DAPI* and one of the following (*SLC17A7*, *GAD1*, and *MBP*) compared to **c'** the quantified distribution of the number of *SLC17A7* + excitatory neurons (Excit), *GAD1* + inhibitory neurons (Inhib), and *MBP* + oligodendrocytes (Oligo), respectively. Scale bar in a for **a**–**c** is 1 mm

regression coefficient for total nuclear RNA expression vs. cell types, which is -1.33 (95% CI $= -1.35, -1.31$, Table 1; the "Linear regression of puncta across cell types" section). For candidate TREGS *AKT3* and *ARID1B*, we measured the number of puncta per nucleus across cell types and found that *AKT3* tracks the closest with the pattern of total RNA expression measured by snRNA-seq and has a more symmetric distribution than *POLR2A*, particularly for oligodendrocytes (Fig. 7b). Excitatory neurons contain the most puncta, followed by inhibitory neurons, and then oligodendrocytes. *ARID1B* breaks from the expression pattern shown in the reference snRNA-seq data, given that inhibitory neurons are higher than excitatory neurons, although neurons still show more expression than the oligodendrocytes. *POLR2A* was only measured in two cell types, excitatory neurons, and oligodendrocytes but also follows this pattern with higher expression in neurons compared to glia. The standardized regression coefficient for number of puncta vs. cell types for *AKT3* is -1.07 (95% CI $= -1.07, -1.06$) and is the

Table 1 Proportion of nuclei that displayed any TREG candidate or *POLR2A* puncta. Proportion of nuclei with a non-zero count in the DLPFC snRNA-seq data compared against the mean proportion of non-zero puncta in the nucleus and mean number of puncta observed in the RNAscope data for the candidate TREGs and *POLR2A*. Beta values are the slope of the linear fit of the number of puncta over ordered cell types and the 95% confidence interval. The standardized beta is the slope of the linear fit of the number of puncta divided by the standard deviation of the number of puncta for each gene. Standardized betas enable the comparison between snRNA-seq and RNAscope data. The standardized beta in snRNA-seq is -1.33 (-1.35, -1.31). With RNAscope, *AKT3* is the TREG that most similarly follows the trend across all genes in snRNA-seq (see also Fig. 7). Due to the inability to resolve individual puncta for *MALAT1*, the observed trend (Additional file 2: Fig. S10) is unreliable

| Gene	Prop. non-zero in DLPFC snRNA	Mean prop. non-zero puncta in the nucleus	Mean *n* puncta	β (95% CI)	Standard deviation	Standardized β (95% CI)
AKT3	0.92	0.88	4.09	-5.52 (-5.55, -5.49)	5.18	-1.07 (-1.07, -1.06)
ARID1B	0.94	0.86	3.08	-2.63 (-2.65, -2.6)	3.42	-0.77 (-0.77, -0.76)
MALAT1	1.00	0.98	2.07	-1.22 (-1.24, -1.21)	1.53	-0.8 (-0.81, -0.79)
POLR2A	0.30	0.78	2.75	-3.49 (-3.51, -3.47)	3.34	-1.05 (-1.05, -1.04)
All genes in snRNA-seq	NA	NA	NA	$-21,844.07$ ($-22,172.45$, $-21,515.68$)	15,560.76	-1.33 (-1.35, -1.31)
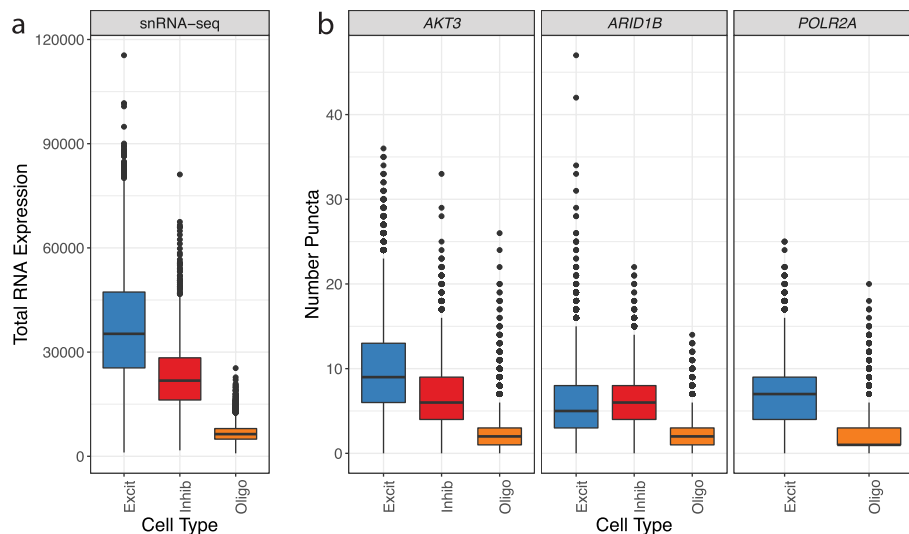
Fig. 7 Boxplots of total RNA nuclear expression in the nucleus across cell types. **a** Distribution of total nuclear RNA expression (estimated with the sum of total UMIs per nucleus) in DLPFC snRNA-seq data across excitatory neurons (Excit), inhibitory neurons (Inhib), and oligodendrocytes (Oligo). **b** Distribution of the number of puncta quantified by RNAscope for each observed gene across the same cell types as in **a**. The number of puncta by RNAscope estimates the total RNA expression by snRNA-seq (Fig. 1a). *POLR2A* was only evaluated in excitatory neurons and oligodendrocytes as it was multiplexed with *MALAT1* and *GAD1* was omitted

closest to the snRNA-seq coefficient (-1.33, 95% CI $= -1.35$, -1.31) of the experimentally validated genes (Table 1). This pattern is also consistent when considering other cell types (Additional file 2: Fig. S10). By measuring total nuclear RNA abundance and

nuclear area across cell types in the same experiment, we demonstrate that the relationship between RNA content and nuclear area changes across cell types (Additional file 2: Fig. S11).

Discussion

In this work, we showed that our new data-driven rank invariance method successfully determines candidate total RNA expression genes (TREGs) from snRNA-seq data that can be used in combination with smFISH to accurately estimate relative RNA abundance in distinct cell types of varying sizes and transcriptional activity. We investigated the properties of TREGs by evaluating the consistency of the expression ranks and their relationship with total RNA expression in snRNA-seq data. We further validated TREG candidates by quantifying puncta with smFISH using RNAscope, which found that *ATK3* best reconstructed the pattern between cell type and total RNA expression observed in human DLPFC snRNA-seq data. While the rank invariance method was successful, it cannot guarantee that candidate TREGs will be experimentally compatible for downstream needs. For example, *MALAT1* was the top candidate TREG in our study, but *MALAT1* was incompatible with resolving individual puncta by RNAscope because of its extremely high expression. 10x Genomics advises users that independent of protocol *MALAT1* is frequently observed in poly-A captured RNA-seq data [25], which is consistent with comparisons between polyA selection versus rRNA depletion protocols [26] and could be due to its structure [27]. Furthermore, *MALAT1* has been used as a proxy for nuclear expression linked to damaged cells in scRNA-seq data [28]. We thus recommend that TREGs be evaluated in the assay or experimental setting of choice before implementing experiments using rare and valuable samples.

The rank invariance method demonstrated stability over biological states by returning a similar set of candidate TREGs across disease states as evaluated in the Velmeshev et al. dataset [24]. The gene *CADM2* was a highly ranked TREG candidate when evaluating either case or control data separately, or the full dataset (which we recommend), as well as in the control-only Tran, Maynard et al. dataset [20]. *CADM2* may be another TREG of interest for the human cortex, which could be validated in future studies.

To identify relevant TREGs for a particular study, it is important to use sc/snRNA-seq data that is compatible with the experimental design. That is, sc/snRNA-seq data should contain the tissue and cell types combinations of interest, as well as match the experimental conditions for which the TREG will be used, such as age, sex, diagnosis, and experimental exposure. Otherwise, candidate TREGs might be less reliable for quantifying total RNA as they could be specific to an organism, tissue, cell type, or experimental condition. As more snRNA-seq datasets come online across tissues and experimental organisms, our rank invariance methodology can be used to identify TREGs in different experimental settings. Furthermore, the mean absolute differences (Fig. 1c, arrow v.) will be more stable when larger numbers of cells/nuclei are present per cell type. Thus, the rank invariance process might be less reliable for datasets with rare cell types, for which it might be best to perform a sensitivity analysis without the rare cell types to compare resulting TREGs and ultimately identify reliable TREGs.

While TREGs share some similarities with housekeeping (HK) genes, they are by definition different. A HK gene typically has a defined central molecular function and is expected to be expressed at a constant level across multiple tissues [19]; for example, the GTEx Portal [29] shows less variable expression across tissues for *POLR2A* than *AKT3* (https://gtexportal.org/home/gene/POLR2A   vs   https://gtexportal.org/home/gene/AKT3). In contrast, the RNA level of a TREG should be quantifiable in most cells among all cell types in a particular experimental setting, and most importantly, it should be highly predictive of the total RNA expression of those cells or nuclei. In our snRNA-seq data, *POLR2A* and other HK genes had high proportion zero and were not as strongly predictive of total nuclear RNA expression as other candidate TREGs. Interestingly, by RNAscope, *POLR2A* could be measured in a larger percent of nuclei compared to snRNA-seq (78% vs 30%, Table 1). We note that TREGs were detected in the majority of nuclei by RNAscope as expected, but we did observe some nuclei lacking expression. This could be due to only a fraction of the nucleus being present in the 10 μm tissue section plane or technical limitations in multiplex fluorescent slide scanning with spectral imaging, including low resolution and image acquisition in only the *x* and *y* dimensions, with no *z*-axis component [30]. However, *AKT3* was present in 88% of nuclei by RNAscope and had a higher mean number of puncta compared to *POLR2A* (4.09 vs 2.75). *AKT3* best recapitulated the observed trend in snRNA-seq data (Fig. 7b). Thus, while *POLR2A* performed better than expected on RNAscope, *ATK3* still outperformed *POLR2A* across different metrics.

The protein encoded by *AKT3* is a member of the AKT/protein kinase B family of serine/threonine protein kinases. AKT plays a key role in the phosphatidylinositol 3-kinase (PI3K)-Akt-mTOR signaling cascade, which regulates numerous biological processes such as cell proliferation, growth, apoptosis, and metabolism [31]. *AKT3* is one isoform of AKT that is predominantly expressed in the human and mouse brain and plays a significant role in brain development [32]. Dysfunction of AKT3 is implicated in a variety of neurodevelopmental and neurodegenerative brain disorders and tumors, such as glioma [32, 33]. The *AKT3* gene has also been associated with risk for neuropsychiatric disorders [34]. AKT3 is an important enzyme whose function needs to be carefully regulated at the protein level. Thus, the *AKT3* gene expression is likely highly regulated across cell types, which is consistent with its dynamic expression in neurons and glia (Additional file 2: Fig. S3). While candidate TREGs, such as *AKT3* and *ARID1B*, are clinically relevant [30, 32–37], the snRNA-seq data used in this study is from neurotypical control subjects. As more snRNA-seq datasets are generated from subjects with neuropsychiatric and neurological disorders, it will be important to assess candidate TREGs in the context of brain disorders. More generally, other candidate TREGs identified in our neurotypical control tissue are functionally implicated in chromatin binding and transcription (Additional file 2: Fig. S5, Additional file 1: Table S2), which is consistent with the notion that a TREG should be predictive of total RNA expression. However, we do not expect that candidate TREGs in other biological contexts will be enriched for the same biological functions or pathways.

In contrast to rank-invariant methods [7–9], the method we developed here is not a normalization method, but a method for gene selection that is specific to the desired properties of a TREG, namely correlation with total RNA expression. Our

implementation is thoroughly tested with 100% code coverage [38] and available via Bioconductor at https://bioconductor.org/packages/TREG [22]. While our list of candidate TREGs could be valid for other DLPFC studies (Additional file 1: Table S3), we recommend that you use our R package with your own sc/snRNA-seq data. TREGs could be useful for other research purposes and other contexts than the ones envisioned here. For example, emerging spatial transcriptomics technologies such as MERFISH [39] and Xenium [40] (where a limited number of probe targets must be selected) can employ TREGs as a complementary class of genes to traditional housekeeping genes to aid in cell type characterization.

We used smFISH with RNAscope technology to validate candidate TREGs across three major cell types in the human DLPFC: oligodendrocytes, excitatory neurons, and inhibitory neurons. We selected only 3 cell types due to the limitations in multiplexing with RNAscope (the V2 assay supports a maximum of 4 targets). In the future, we plan to expand our experimental design to include other major cell types captured in snRNA-seq data, such as astrocytes, microglia, and oligodendrocyte progenitor cells. Another limitation of our study is that we focused only on TREG expression in the nucleus, but many TREGs are also expressed in the cytoplasm. Currently, the HALO FISH-IF module will only quantify puncta within a nucleus or a dilated area around the nucleus, which is a limitation when working with cell types that are not oval in shape, such as neurons and glia. While our analysis was focused only in the nucleus, previous work suggests that gene-level expression between the nucleus and cytoplasmic compartments are comparable [18, 41]. In future studies, we aim to improve cell segmentation to be able to estimate cell size and RNA content instead of restricting analyses to the nucleus.

While sc/snRNA-seq and bulk RNA-seq data is commonly derived from pulverized tissue, our work suggests considering a different experimental design. In particular, if you are designing a paired sc/snRNA-seq and bulk RNA-seq study where you will use the snRNA-seq data as a reference for deconvolution of bulk RNA-seq, generating spatially adjacent dissections in order to use them for RNAscope experiments could be useful to "future-proof" your datasets [15]. With this experimental design, you could identify cell types in the sc/snRNA-seq data, then identify candidate TREGs based on those cell types, and use these candidate TREGs with smFISH to quantify the size and total RNA content for the cell types of interest (Additional file 2: Fig. S11). These cell metrics could improve the precision of deconvolution algorithms and generate a potential gold standard dataset to evaluate the performance of the deconvolution methods.

Conclusion

Through the data-driven rank invariance process, we have identified several candidate genes as total RNA expression genes (TREGs) in five postmortem human brain regions. RNAscope validation experiments revealed that *AKT3* is a strong proxy for total nuclear RNA expression in the DLPFC. Future work will use individual TREGs to estimate total RNA abundance in differently sized cell types of the DLPFC to bolster deconvolution approaches. As more snRNA-seq data comes online, this rank invariance methodology could facilitate the identification of TREGs in other experimental settings with differences in cell types, donor demographics, brain regions, tissues, or species. Similar to highly variable genes or housekeeping genes, TREGs represent an important class of

genes that could be used for a variety of assays and downstream analyses. Our method for identifying TREGs is accessible, integrated with the Bioconductor ecosystem [42], and available at https://bioconductor.org/packages/TREG [22].

Methods

R [43] and HALO (version 3.3.2541.383, Indica Labs) were used for the data analysis, and plotting was done with the ggplot2 [44] and UpSetR [45] packages.

snRNA-seq reference data

The single-nucleus RNA sequencing (snRNA-seq) reference data used for the rank invariance process is a publicly available dataset ($10 \times$ Genomics, single-cell 3′ gene expression) from postmortem human brain, which includes tissue from eight donors and five brain regions, including the amygdala (AMY), dorsolateral prefrontal cortex (DLPFC), hippocampus (HPC), nucleus accumbens (NAc), and subgenual anterior cingulate cortex (sACC) [20]. The original study classifies many region-specific subtypes of inhibitory and excitatory neurons; however, for the purpose of this study, a lower resolution of broad cell types was used: astrocytes (Astro), endothelial cells (Endo), microglia, mural cells, oligodendrocytes (Oligo), oligodendrocyte precursor cells (OPC), T-cells, excitatory (Excit), and inhibitory neurons (Inhib). Specialized medium spiny neurons found exclusively in the NAc were classified as Inhib (Additional file 1: Table S1).

Expression and proportion zero filtering

In R [43], by default, the rank() function returns high ranks for high values, where equivalent values or ties are given an average value. To reduce the occurrence of ties, we removed genes that would introduce many low-value ties. The data was first filtered to the top 50% of expressed genes. The nuclei were grouped by cell type and brain region, the proportion zero counts were calculated for each gene in each group and is defined as $PZ_{i,j,k} = \sum\limits_{z=1}^{n_{j,k}} I\left(C_{i,j,k} > 0\right)/n_{j,k}$ where $c_{i,j,k,z}$ is the number of snRNA-seq counts for cell/nucleus z for gene i, cell type j, and brain region k, and $n_{j,k}$ is the number of cells/nuclei for cell type j and brain region k. In our dataset, if the cell type was rare (less than or equal to 100 total nuclei, as was the case for Endo, Macro, Mural, and T-cells), the nuclei from different regions were combined into one group, effectively ignoring the brain region from which the cell type was derived (Additional file 2: Fig. S1). The distribution of proportion zeros for each group was visualized and used to select a cutoff value of 0.75, which included the peak of the proportion zero distributions (Fig. 2a). Then, for each gene, the maximum proportion zero across groups was required to be less than the cutoff (i.e., < 0.75) to pass the filtering step (Fig. 2b).

Rank invariance calculation

After proportion zero filtering, the remaining genes were evaluated for rank invariance jointly across all five brain regions; thus, the nuclei were grouped only by cell type. The normalized expression (logcounts) of each gene was ranked for each nucleus, and the result was a matrix of expression rank values (the number of nuclei * number of genes). Within each cell type, the mean expression for each gene was ranked, and the result was

a vector called mean expression rank (length is the number of genes). Then, the absolute difference between the expression rank of each nucleus and the mean expression rank was found. From here, the mean of the differences for each gene was calculated and then ranked. These steps were repeated for each cell type, and the result was a matrix of ranks, (number of cell types * number of genes). From here, the sum of the ranks for each gene were reverse-ranked such that low values were given a high rank. This process resulted in the final value for each gene called the rank invariance value (Fig. 1b). The genes with the highest rank invariance were considered as candidate TREGs. Classic housekeeping (HK) genes [46, 47] and brain data-driven HK genes [6] that fail these filtering steps are shown in Additional file 2: Fig. S2.

Total RNA linear regression

We tested for an association between the expression of each gene and the overall RNA expression of each nucleus using a linear regression model $log2(counts+1) \sim log2(sum)+cellType$ with limma package version 3.48 and "voom" [48, 49] (Fig. 3c, d). The *t*-statistics from this analysis are plotted in Additional file 2: Fig. S2. The combination of the rank invariance values and the rank of the *t*-statistics from this linear model were used to help identify the best candidate TREGs.

Gene ontology and KEGG pathway enrichment analysis

Of the 877 genes evaluated for rank invariance, the top 20 or top 50 were evaluated for Gene Ontology and KEGG pathway enrichment. Of the 11,519 genes expressed in the top 50% in the snRNA-seq dataset, 10,875 (94.41%) have ENTREZ ids and were used as the universe. The enrichment analysis was performed with the compareCluster() function from clusterProfiler package version 4.8.1 [50, 51]. Ontologies biological processes (BP), cellular components (CC), molecular function (MF), and KEGG pathways were all tested.

Postmortem human tissue

The human postmortem brain used in this study for RNAscope was obtained by autopsy from the Offices of the Chief Medical Examiner of the District of Columbia and of the Commonwealth of Virginia, Northern District, with informed consent from the legal next of kin (protocol 90-M-0142 approved by the NIMH/NIH Institutional Review Board). Details regarding curation, diagnosis, tissue handling, processing, and quality control measures have been described previously [52]. The study included a single neurotypical control adult donor (Br1531). A small piece of frozen DLPFC was dissected under visual guidance with a handheld dental drill on dry ice by a neuroanatomist. Gray and white matter tissue from the crown of the middle frontal gyrus was obtained from the coronal slab corresponding to the middle one-third of the DLPFC (along its rostral-caudal axis) immediately anterior to the genu of the corpus callosum. Microdissected DLPFC tissue was stored at $-80\,°C$ until cryosectioning.

RNAscope multiplex single molecule fluorescent in situ hybridization (smFISH)

DLPFC tissue was cryo-sectioned at 10 μm on a Leica cryostat. Three tissue sections were collected per slide. Prior to use, the slides were stored at $-80\,°C$. Using RNAscope

technology (RNAscope Multiplex Kit V2 and 4-plex Ancillary Kit: Cat # 323,100, 323,120, Advanced Cell Diagnostics, Newark, CA), probe hybridization and labeling were completed following the manufacturer's instructions. Briefly, the protocol includes fixing the tissue sections in 10% neutral buffered formalin solution (Cat # HT501128-4L, Sigma-Aldrich, St. Louis, MO), dehydration in a series of ethanol solutions, pretreatment with hydrogen peroxide, and permeabilizing with proteases. Each slide was then incubated with one of the following probe combinations: *AKT3*/*GAD1*/*SLC17A7*/*MBP*; *ARID1B*/*GAD1*/*SLC17A7*/*MBP*; *POLR2A*/*MALAT1*/*SLC17A7*/*MBP* (Additional file 1: Table S4). These combinations were named according to the candidate TREGs or housekeeping (HK) gene included (i.e., *AKT3*, *ARID1B*, or *POLR2A/MALAT1*) (Cat # 434211, 404031-C2, 415611-C3, 411051-C4, 551721, 310451, 400811-C2, Advanced Cell Diagnostics, Newark, CA). After washing briefly, the slides were stored in $4 \times$ saline-sodium citrate (Cat # 351–003-101, Quality Biological, Gaithersburg, MD) overnight at 4 °C. Probes were then fluorescently labeled using opal dyes (Opal 520, Opal 570, Opal 620, and Opal 690, Perkin Elmer, Waltham, MA). Dyes were assigned to probes and diluted in concentration as described in Additional file 1: Table S4 and Additional file 1: Table S5. The nuclei were labeled with DAPI (4′,6-diamidino-2-phenylindole) and coverslipped with fluoromount-G mounting media.

Image acquisition

Slides were imaged at $\times 20$ magnification using a Vectra Polaris Automated Quantitative Pathology Imaging System (Akoya Biosciences), which performs multi-spectral imaging. For each probe combination, a scanning protocol was created. Each protocol optimized the exposure time for a given opal dye in each probe combination as listed in Additional file 1: Table S6. Scanning generated a large.qptiff image file, which was then pre-processed in Phenochart (Akoya Biosciences). Briefly, the boundary of each slide (including the 3 tissue sections) was traced, and the individual.tiff tiles making up the scan area (1141–1489 tiles per slide) were extracted. These tiles were then subjected to linear unmixing in InForm (Akoya Biosciences). Unmixed.tiff tiles were then fused in HALO (version 3.3.2541.383, Indica Labs).

Image analysis with HALO

Fused images from each scanned slide were annotated in HALO by drawing a boundary around each tissue section. The annotated areas across tissue sections ranged from 156,716,789.46 to 162,640,367.42 μm^2 and annotations were consistent among the tissue sections on each slide. The FISH-IF module (version 2.1.5) was then used to segment cells and assign phenotype (i.e. cell type). Briefly, we first assigned dyes to either FISH or immunofluorescence (IF). While these experiments were exclusively FISH, the distinction between FISH and IF dyes allows for visualization and segmentation of diffuse staining vs. individual puncta. DAPI, *GAD1*, *SLC17A7*, and *MBP* were assigned IF values given their strong signals resembling diffuse labeling. The FISH dye assignment changed for each experiment (*AKT3*, *ARID1B*, *POLR2A/MALAT1*). Segmentation was optimized for each dye for each tissue section by adjusting several values with reference to the manufacturer's guidelines: HALO 3.3 FISH-IF Step-by-Step guide (Indica Labs, version 2.1.4 July 2021) and Quantitative RNAscope Image Analysis Guide (Indica Labs).

Size thresholds for nuclei, cytoplasm, cells, and FISH probe puncta were held constant across all tissue sections. Once the puncta counting was completed, the object data and settings were exported as.csv files (available via Globus at "jhpce#TREG_paper") and.txt files (available on GitHub), respectively.

Quality control and spatial quantification of HALO data

Visual inspection of the images revealed some technical artifacts related to cryosectioning and slide scanning (Additional file 2: Fig. S8), including tissue tears/shredding, small bubbles, and out-of-focus fields. Out-of-focus fields caused the nuclei to appear bigger and blurred together multiple puncta so they were not clearly resolved. The nuclei from these regions were excluded from the data analysis during these quality control steps.

The nuclei with only *MBP* expression were classified as oligodendrocytes (Oligo), with only *SLC17A7* expression as excitatory neurons (Excit), and with only *GAD1* expression as inhibitory neurons (Inhib). The nuclei with the expression of multiple marker genes that could not definitively be assigned a cell type were classified as "Multi." The nuclei with no markers were classified as "Other" and likely represent other non-neuronal cell types in the brain that were not labeled, including astrocytes and microglia. As the *MALAT1*/*POLAR2A* samples were not labeled with *GAD1* due to technical limitations in multiplexing (Additional file 1: Table S4), the number of Inhib nuclei could not be determined for these samples (Additional file 1: Table S7).

Linear regression of puncta across cell types

TREG candidates were evaluated by the proportion of cells where any puncta were recorded in the HALO-segmented nuclear area, as well as the mean number of puncta recorded (Table 1). The pattern of expression across cell types was compared to the sum of total counts of that cell type in the reference snRNA-seq data (Fig. 7). To quantify this relationship, we estimated the regression coefficient of total RNA over the three cell types that were sampled (Excit, Inhib, Oligo), for the snRNA-seq total RNA of a nucleus was estimated by the sum of UMIs. For the RNAscope data, total nuclear RNA is estimated by the number of segmented puncta. To compare these different data types, the standardized regression coefficient was calculated by dividing by the standard deviation of the total UMIs and number of puncta respectively.

External data replication in case–control data

A second snRNA-seq dataset from Velmeshev et al.'s case (autism spectrum disorder) and control samples from the prefrontal cortex and anterior cingulate cortex was used to validate this process in a more diverse biological context [24]. Cell types were evaluated at a similar seven broad cell type level as used in the Tran et al. data for the main analysis. The data was evaluated in all samples, then case samples only, and control samples only. In the *expression and proportion zero filtering* step, the proportion zero limit was set to 0.85 for all three groups instead of 0.75 as this dataset was more sparse than the Tran et al. dataset. The *rank invariance calculation* was performed on the resulting set of genes for each group. The fifty genes with the highest rank invariance were considered TREG candidates, and compared to evaluate the stability of the analysis between the case and control conditions.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-03066-w.

> **Additional file 1.** This additional file contains all supplementary tables and their captions.
>
> **Additional file 2.** This additional file contains all supplementary figures and their captions.
>
> **Additional file 3.** This additional file contains the review history.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 3.

Authors' contributions

LAHM: conceptualization, formal analysis, methodology, software, visualization, writing—original draft, and writing—review and editing. KDM: investigation, validation, visualization, and writing—original draft. SHK: formal analysis, investigation, validation, visualization, and writing—original draft. SCP: writing—review and editing. SCH: conceptualization, funding acquisition, and writing—review and editing. KRM: conceptualization, funding acquisition, project administration, supervision, writing—original draft, and writing—review and editing. LCT: conceptualization, methodology, software, supervision, writing—original draft, and writing—review and editing.

Authors' Twitter handles

Twitter handles: @lahuuki (Louise A. Huuki-Myers), @sanghokwon17 (Sang Ho Kwon), @cerceopage (Stephanie C. Page), @stephaniehicks (Stephanie C. Hicks), @kr_maynard (Kristen R. Maynard), @lcolladotor (Leonardo Collado-Torres).

Availability of data and materials

The snRNA-seq data [20] is available from https://github.com/LieberInstitute/10xPilot_snRNAseq-human#processed-data [53]. The code for the analysis is available from https://github.com/LieberInstitute/TREG_paper [54] and the software from https://github.com/LieberInstitute/TREG [22]. The raw data are publicly available from the Globus endpoint "jhpce#TREG_paper" that is also listed at http://research.libd.org/globus/jhpce_TREG_paper/. The raw data provided through Globus include all the raw image files and the HALO.csv segmentation results.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

References

1. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9:75.
2. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating single-cell analysis with Bioconductor. Nat Methods. 2020;17:137–45.

Huuki-Myers *et al. Genome Biology*   (2023) 24:233

Page 20 of 21

3. Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol. 2013;14:R31.

4. Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat Neurosci. 2021;24:425–36.

5. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. Nat Methods. 2018;15:343–6.

6. Caracausi M, Piovesan A, Antonaros F, Strippoli P, Vitale L, Pelleri MC. Systematic identification of human housekeeping genes possibly useful as references in gene expression studies. Mol Med Report. 2017;16:2397–410.

7. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res. 2001;29:2549–57.

8. Mar JC, Kimura Y, Schroder K, Irvine KM, Hayashizaki Y, Suzuki H, et al. Data-driven normalization strategies for high-throughput quantitative RT-PCR. BMC Bioinformatics. 2009;10:110.

9. Liu X, Li N, Liu S, Wang J, Zhang N, Zheng X, et al. Normalization methods for the analysis of unbalanced transcriptome data: a review. Front Bioeng Biotechnol. 2019;7:358.

10. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat Commun. 2020;11:5650.

11. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods. 2017;14:955–8.

12. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10:380.

13. Fan J, Lyu Y, Zhang Q, Wang X, Li M, Xiao R. MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. Brief Bioinformatics. 2022;23(6):bbac430. https://doi.org/10.1093/bib/bbac430.

14. Sosina OA, Tran MN, Maynard KR, Tao R, Taub MA, Martinowich K, et al. Strategies for cellular deconvolution in human brain RNA sequencing data. F1000Res. 2021;10:750.

15. Maden SK, Kwon SH, Huuki-Myers LA, Collado-Torres L, Hicks SC, Maynard KR. Challenges and opportunities to computationally deconvolve heterogeneous tissue with varying cell sizes using single cell RNA-sequencing datasets. arXiv. 2023. https://doi.org/10.48550/arXiv.2305.06501.

16. Wang F, Flanagan J, Su N, Wang L-C, Bui S, Nielson A, et al. RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. J Mol Diagn. 2012;14:22–9.

17. Maynard KR, Tippani M, Takahashi Y, Phan BN, Hyde TM, Jaffe AE, et al. dotdotdot: an automated approach to quantify multiplex single molecule fluorescent in situ hybridization (smFISH) images in complex tissues. Nucleic Acids Res. 2020;48(11):e66. https://doi.org/10.1093/nar/gkaa312.

18. Price AJ, Hwang T, Tao R, Burke EE, Rajpurohit A, Shin JH, et al. Characterizing the nuclear and cytoplasmic transcriptomes in developing and mature human cortex uncovers new insight into psychiatric disease gene regulation. Genome Res. 2020;30:1–11.

19. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. 2013;29:569–74.

20. Tran MN, Maynard KR, Spangler A, Huuki LA, Montgomery KD, Sadashivaiah V, et al. Single-nucleus transcriptome analysis reveals cell-type-specific molecular signatures across reward circuitry in the human brain. Neuron. 2021;109:3088-3103.e5.

21. Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science. 2016;352:1586–90.

22. Huuki-Myers LA, Collado-Torres L. LieberInstitute/TREG. Zenodo. 2022.

23. Touchberry CD, Wacker MJ, Richmond SR, Whitman SA, Godard MP. Age-related changes in relative expression of real-time PCR housekeeping genes in human skeletal muscle. J Biomol Tech. 2006;17:157–62.

24. Velmeshev D, Schirmer L, Jung D, Haeussler M, Perez Y, Mayer S, et al. Single-cell genomics identifies cell type-specific molecular changes in autism. Science. 2019;364:685–9.

25. 10x Genomics. Why do I see high levels of Malat1 in my gene expression data? https://kb.10xgenomics.com/hc/en-us/articles/360004729092-Why-do-I-see-high-levels-of-Malat1-in-my-gene-expression-data-. Accessed 27 Apr 2022.

26. Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. Sci Rep. 2018;8:4781.

27. Wilusz JE, Spector DL. An unexpected ending: noncanonical 3'end processing mechanisms. RNA. 2010;16:259–66.

28. Muskovic W, Powell JE. DropletQC: improved identification of empty droplets and damaged cells in single-cell RNA-seq data. Genome Biol. 2021;22:329.

29. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–5.

30. Halgren C, Kjaergaard S, Bak M, Hansen C, El-Schich Z, Anderson CM, et al. Corpus callosum abnormalities, intellectual disability, speech impairment, and autism in patients with haploinsufficiency of ARID1B. Clin Genet. 2012;82:248–55.

31. Xie Y, Shi X, Sheng K, Han G, Li W, Zhao Q, et al. PI3K/Akt signaling transduction pathway, erythropoiesis and glycolysis in hypoxia (Review). Mol Med Report. 2019;19:783–91.

32. Wang L, Zhou K, Fu Z, Yu D, Huang H, Zang X, et al. Brain development and Akt signaling: the crossroads of signaling pathway and neurodevelopmental diseases. J Mol Neurosci. 2017;61:379–84.

33. Long H-Z, Cheng Y, Zhou Z-W, Luo H-Y, Wen D-D, Gao L-C. PI3K/AKT signal pathway: a target of natural products in the prevention and treatment of Alzheimer's disease and Parkinson's disease. Front Pharmacol. 2021;12:648636.

34. Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli T, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature. 2022;604:502-8. https://doi.org/10.1038/s41586-022-04434-5.

35. D'Gama AM, Pochareddy S, Li M, Jamuar SS, Reiff RE, Lam A-TN, et al. Targeted DNA sequencing from autism spectrum disorder brains implicates multiple genetic mechanisms. Neuron. 2015;88:910–7.

36. Shibutani M, Horii T, Shoji H, Morita S, Kimura M, Terawaki N, et al. Arid1b Haploinsufficiency Causes Abnormal Brain Gene Expression and Autism-Related Behaviors in Mice. Int J Mol Sci. 2017;18(9):1872. https://doi.org/10.3390/ijms18091872.

Huuki-Myers *et al. Genome Biology*   (2023) 24:233

Page 21 of 21

37. Alessi DR, Andjelkovic M, Caudwell B, Cron P, Morrice N, Cohen P, et al. Mechanism of activation of protein kinase B by insulin and IGF-1. EMBO J. 1996;15:6541–51.
38. Hester J. covr: test coverage for packages. 2020.
39. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015;348:aaa6090.
40. Janesick A, Shelansky R, Gottscho A, Wagner F, Rouault M, Beliakoff G, et al. High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. BioRxiv. 2022. https://doi.org/10.1101/2022.10.06.510405.
41. Lake BB, Codeluppi S, Yung YC, Gao D, Chun J, Kharchenko PV, et al. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. Sci Rep. 2017;7:6031.
42. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12:115–21.
43. R: a language and environment for statistical computing. https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing. Accessed 2 Mar 2022.
44. Wickham H. ggplot2: elegant graphics for data analysis (use R!). 2nd ed. Cham: Springer; 2016.
45. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33:2938–40.
46. Advanced Cell Diagnostics, Inc. Control Probe- Manual RNAscope HiPlex. https://acdbio.com/product-type2/control-probe-manual-rnascope-hiplex. Accessed 26 Apr 2022.
47. Hounkpe BW, Chenou F, de Lima F, De Paula EV. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. Nucleic Acids Res. 2021;49:D947-55.
48. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29.
49. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
50. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2:100141.
51. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.
52. Lipska BK, Deep-Soboslay A, Weickert CS, Hyde TM, Martin CE, Herman MM, et al. Critical factors in gene expression in postmortem human brain: focus on studies in schizophrenia. Biol Psychiatry. 2006;60:650–8.
53. Tran MN, Collado-Torres L, Seyedian A, Huuki L. LieberInstitute/10xPilot_snRNAseq-human: Revision. Zenodo. 2021. https://doi.org/10.5281/zenodo.5149046.
54. Huuki-Myers LA, Collado-Torres L. LieberInstitute/TREG_paper. Zenodo. 2022.

Publisher's Note