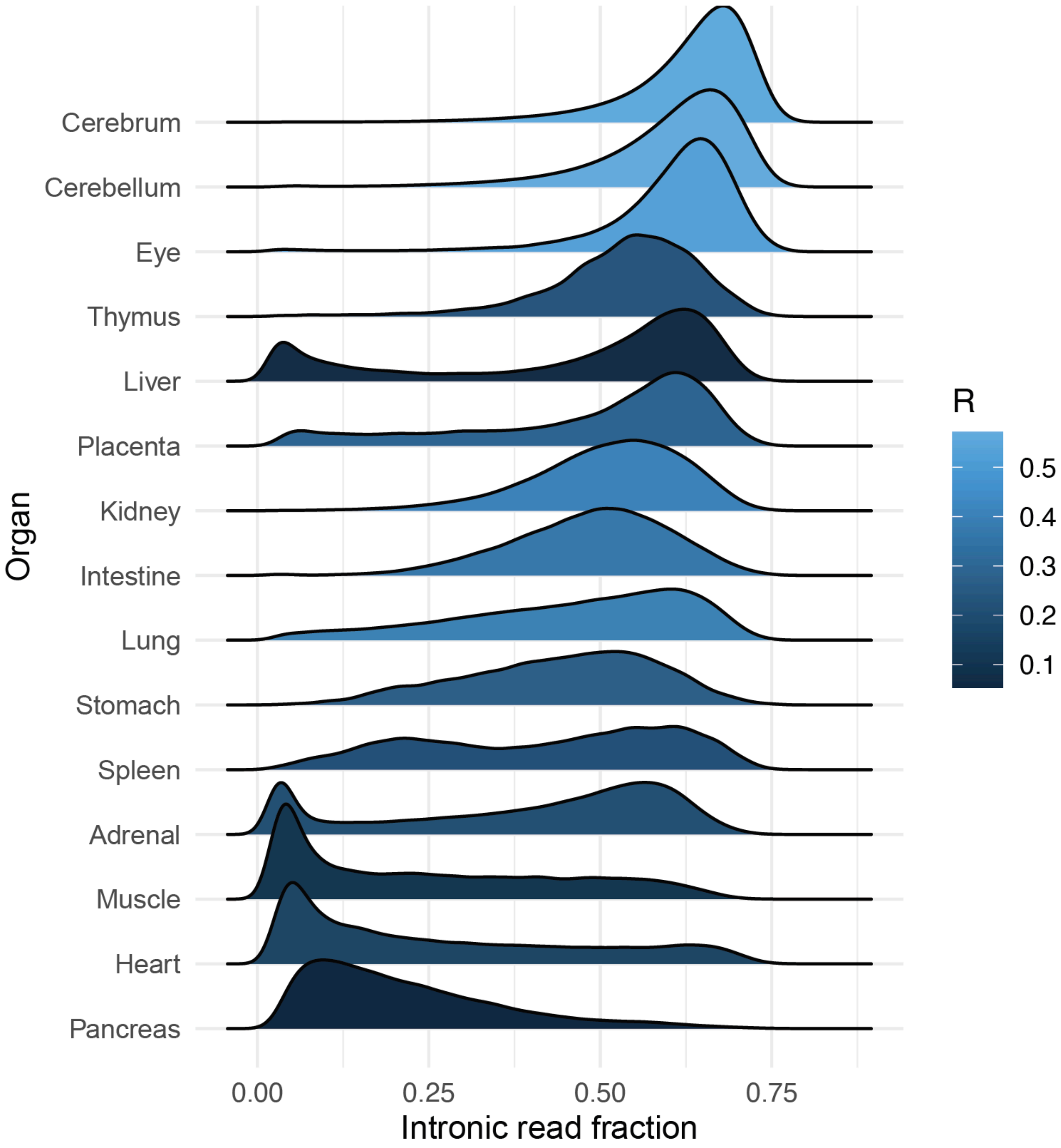


Figure S1: Variation in intronic read fraction reported as metadata by the Human Fetal Atlas(Cao et al. 2020). Tissues are ordered by median fraction; nervous tissues exhibit the strongest correlations between average *intron&exon* abundance in the tissue (log of average CPM) and gene length. Liver appears as an outlier due to bimodality in the intron fraction, which

is driven almost entirely by the 'Hepatocyte' cell type (data not shown). These data were generated with 'single-cell combinatorial indexing' (sci-RNA-seq) on nuclei, which uses a different cell barcoding strategy than 10x Genomics but is otherwise similar with regard to the use of poly(dT) primers and 3' read coverage.

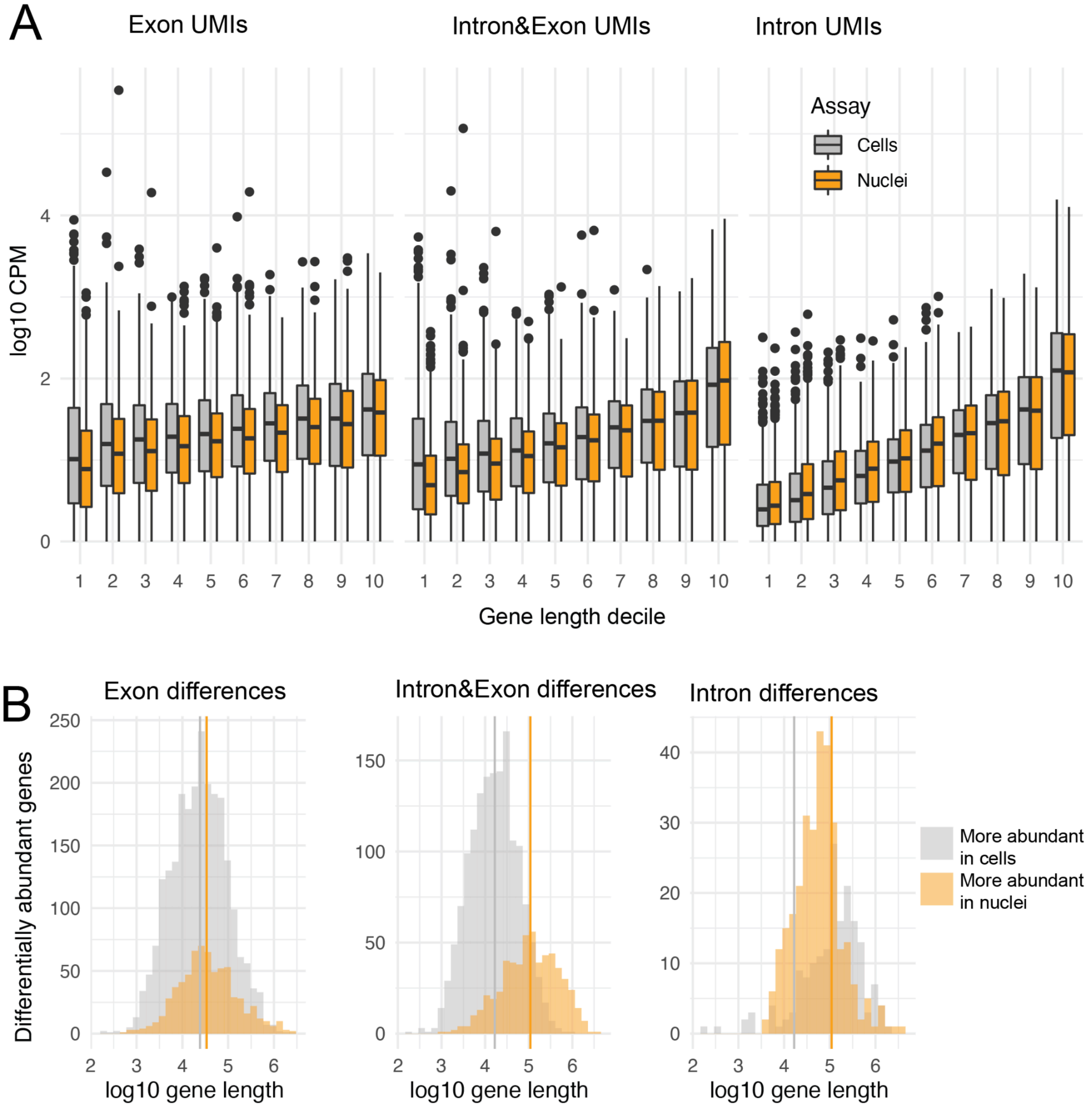


Figure S2. Gene length bias in L5 IT neurons across quantification strategies (A) Gene length vs abundance. Correlation coefficients between mean Log10 abundance and gene length in cells and nuclei: Exon: R= 0.13 and 0.21, Intron&Exon: R=0.29 and 0.44; Intron: R=0.47 and 0.47. The apparent reduction in nuclei is due to the very high abundance of Malat1, a nuclear lncRNA. (B) Gene length distribution of differentially-abundant genes between L5 IT cells and nuclei. Mean length of significantly different genes (FC > 1.5) in cells and nuclei (Log10 scale): Exon: 4.35 vs 4.6 (p = 4.6e-11), 2051 genes; Intron&Exon: 4.15 vs 5.10 (p < 2e-16), 1230 genes; Intron: 4.75 vs 4.77 (p = 0.88), 169 genes

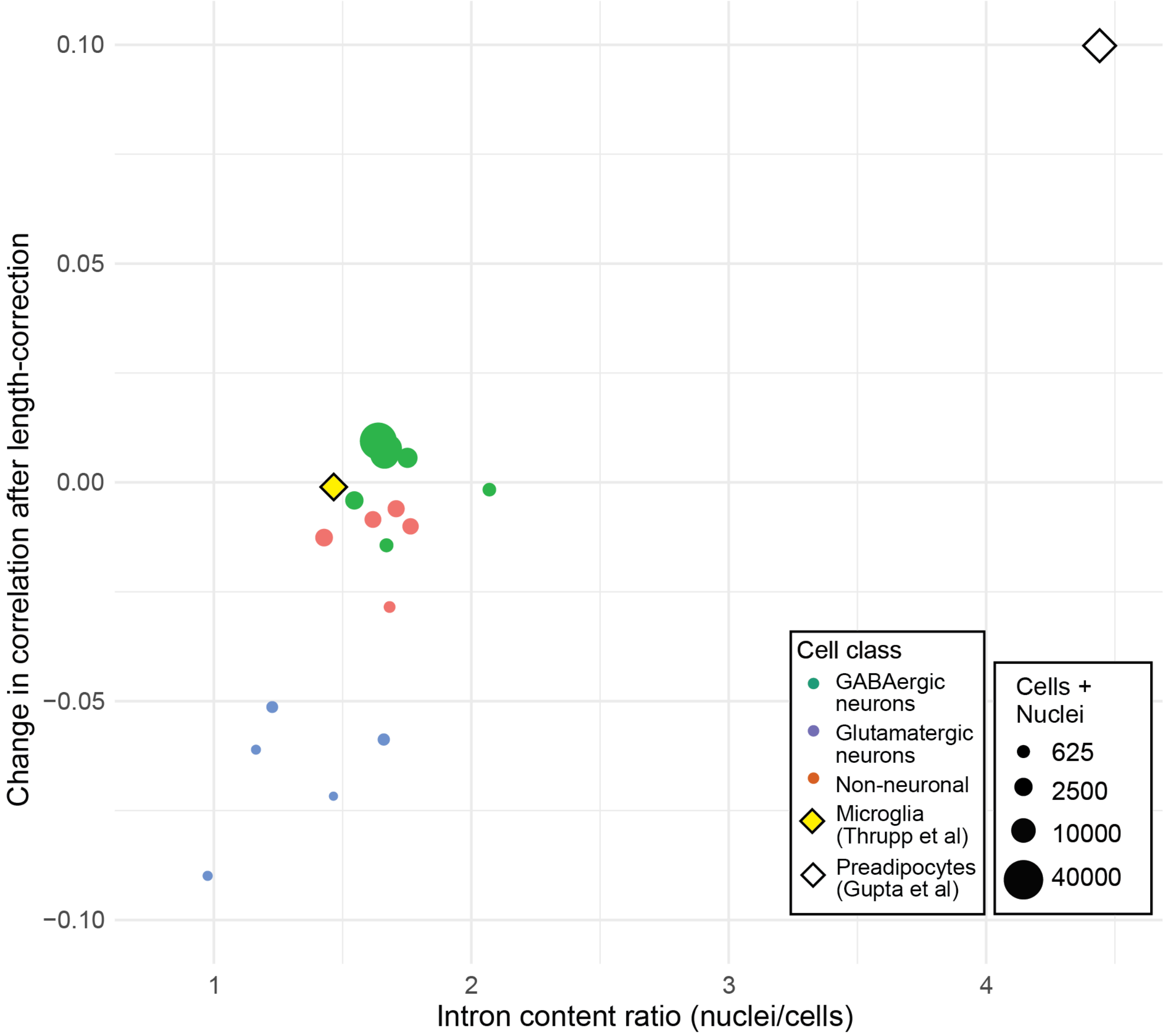


Figure S3: Effect of length-correction on Pearson correlation between average Log10 cell and nuclear abundances (genes with CPM > 1) in mouse motor cortex cell types. Points are colored by cell category and sized by the total number in both assays. The result generated from Thrupp et al data is shown as a yellow diamond and the result reported by Gupta et al is shown as a white diamond.

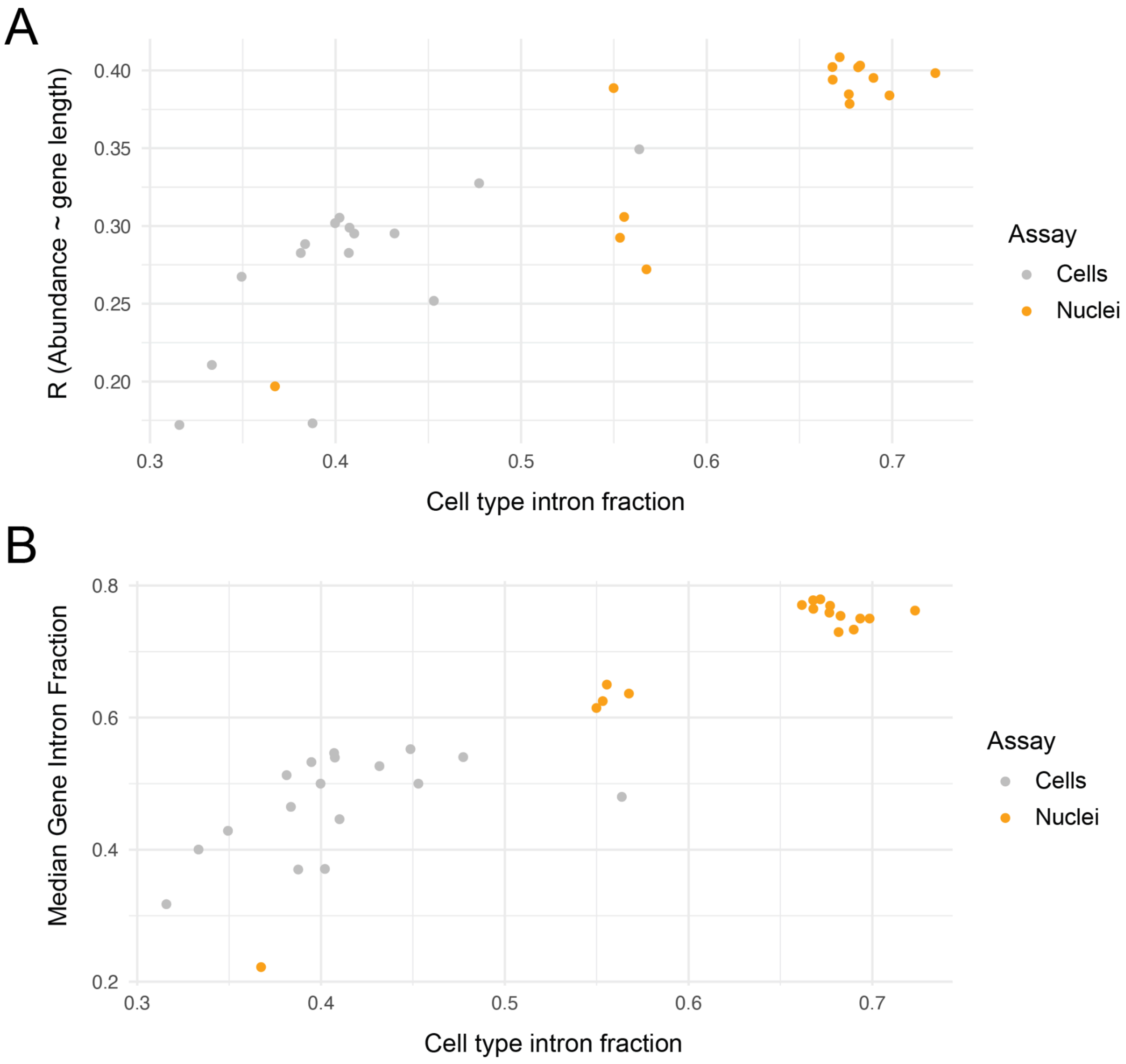
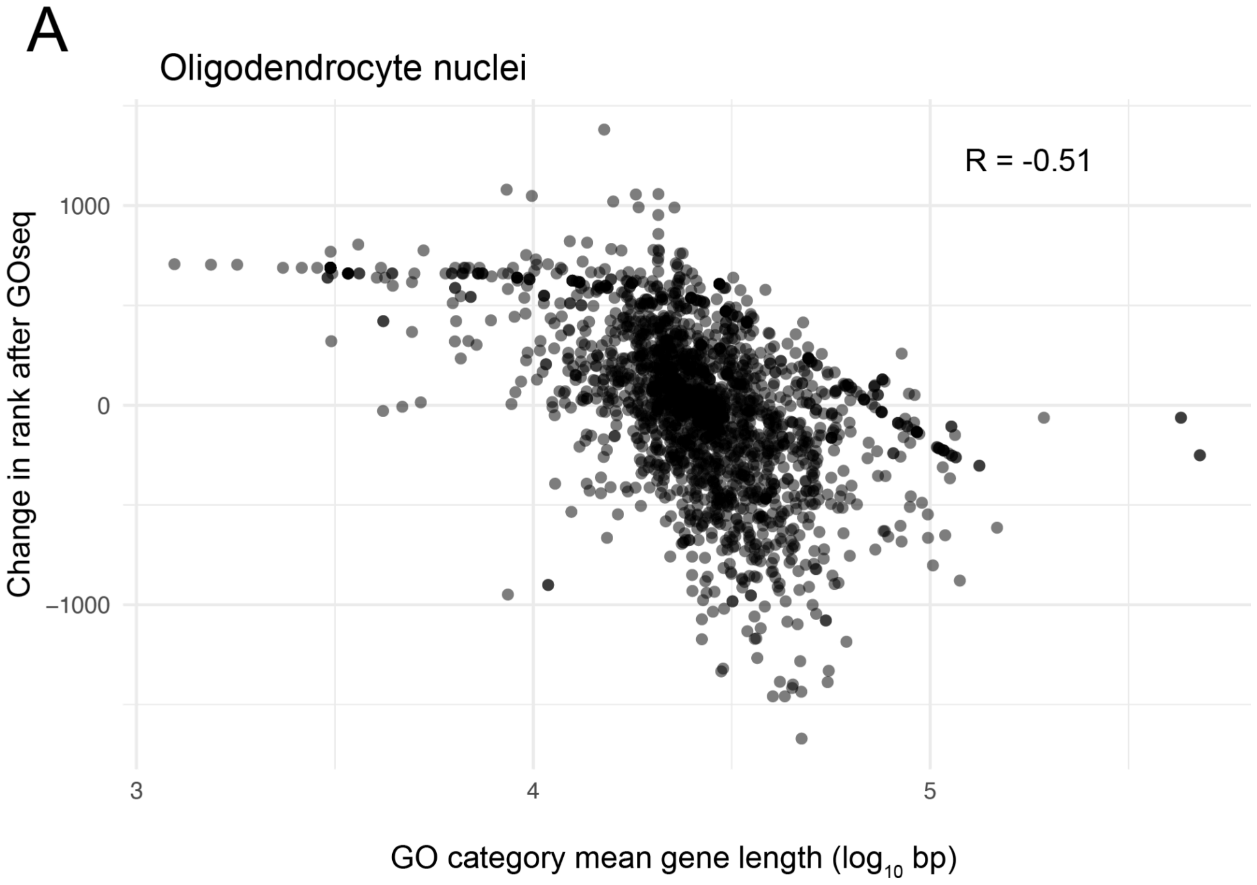


Figure S4. Variation in gene length bias is due to increased pre-mRNA sampling in some cell types. (A) Cell types with higher intron count fraction show a stronger correlation between average abundance (log10 mean CPM) and gene length (Log10 scale). (B) Cell types with a higher intronic UMI count fraction also show a higher fraction of intronic counts at the gene level, which indicates that the apparent increase in pre-mRNA content is due to increased sampling of pre-mRNA rather than a biological preference for longer genes. Y-axis shows the median gene-level intron count fraction for each cell type in cells and nuclei from the mouse cortex dataset.



B

Without GOseq

1. cell development
2. cell periphery
3. myelin sheath
4. neurogenesis
5. generation of neurons

With GOseq

1. myelin sheath
2. cell development
3. myelinination
4. neurogenesis
5. ensheathment of neurons

Figure S5. GOseq analysis of oligodendrocyte nuclei. (A) GOseq was applied to marker genes (significantly more expressed in oligodendrocytes compared to other cell types in the cortex nuclei dataset, fold change > 1.5). GO categories with nominal p values less than .1 are depicted. Categories with longer than average genes tend to decrease in prominence, while

categories with short genes become more significant. (B) Top five highest ranked GO terms in the oligodendrocyte nuclei marker genes with and without adjusting for gene length with GOseq.

Supplementary Table Captions:

Table S1. GOseq results for nucleus-enriched genes in the comparison of L5 IT neurons (fold change > 1.5). The quantification (Exon, Intron&Exon, Length-corrected) and statistical method (uncorrected Hypergeoemtric or GOseq) are stored in columns ‘mode’ and ‘quant’.

Table S2. GOseq results for marker genes (Intron&Exon, fold change > 1.5) in cortex cells and nuclei. Cell type is found in column ‘subclass’, assay in column ‘tech’ and statistical test in column ‘mode’.

Table S3. GOseq results for human microglia (Intron&Exon, fold change > 1.5). A description of the comparison (between assays and/or preprocessing) and statistical test are stored in column ‘mode’.