

Title: A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA.

Authors: Blue B. Lake^{1†}, Simone Codeluppi^{2,3†}, Yun C. Yung^{4†}, Derek Gao¹, Jerold Chun^{4*}, Peter V. Kharchenko^{5*}, Sten Linnarsson^{2*}, and Kun Zhang^{1*}

Supplementary Material:

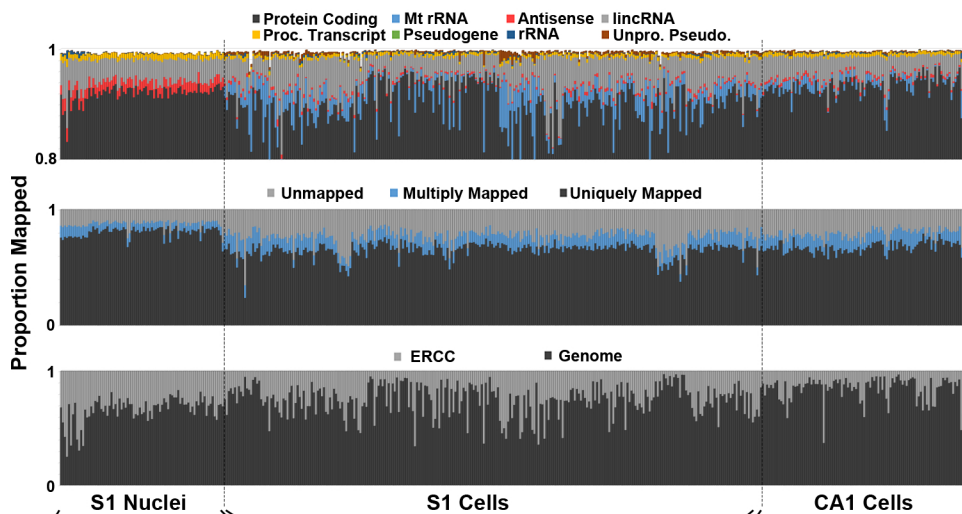
- Supplementary Tables S1-S5 (Excel file, .xlsx)
 - **Table S1.** Single cell or nucleus data sets used for analyses.
 - **Table S2.** Cell type-specific gene sets.
 - **Table S3.** Genes identified as differentially detected between S1 excitatory neuronal nuclei or cells.
 - **Table S4.** Go annotations for all genes over-represented in cellular over nuclear data sets for S1 excitatory neurons.
 - **Table S5.** Go annotations for all genes over-represented in nuclear over cellular data sets for S1 excitatory neurons
- Supplementary Figures S1-S6 (.pdf file)
 - **Fig. S1.** SNS is comparable to whole cell RNA-seq. **a.** Top panel: proportion of reads mapped to different gene types. Middle panel: relative proportion of all reads that were unmapped, multiply mapped or uniquely mapped to ERCC transcripts and reference genes. Bottom panel: the proportion of reads that mapped to either genome or ERCC. Results are shown for S1 nuclei, S1 cortical cells and CA1 hippocampal cells. **b.** Plots showing the frequency distribution of total number of reads sequenced (counts ≥ 4) and ERCC Pearson correlation r values [$\log(\text{counts}+0.1)$ versus $\log(\text{concentration})$] for all single nuclei libraries. **c.** Sample pairwise correlation plots using expression values ($\log(\text{TPM}+1)$) from the top 20 genes from each significant PC loading used to cluster and classify cells in Fig. 1e. Associated sample clusters are indicated. **d.** Sample pairwise correlation plots based on ERCC TPM values ($\log(\text{TPM}+1)$) with cluster identities indicated in (c). Hierarchical clustering was based on the “complete” method and euclidian distance.

- **Fig. S2.** Samples cluster based on cell type expression. **a.** Sample pairwise correlation plots using cell type-specific gene sets ($\log(\text{TPM}+1)$) as shown in Fig. 2a (Supplementary Table S1-S2). Associated sample clusters are indicated. **b.** Hierarchical clustering of samples (based on “ward.D” method and euclidian distance) for heatmap shown in Fig. 2a, with cluster identities indicated in (a), and showing co-clustering of neuronal data sets (nuclear and cellular).
- **Fig. S3.** Nuclear data sets represent S1 excitatory neurons. Scatter plots showing averaged expression values for nuclei compared the indicated cell types. Left column shows plots using all genes, right column shows plots using only genes associated with the cell type being compared (Supplementary Table S2). Pearson correlation values (r) are indicated.
- **Fig. S4.** Correction of systematic gene length and intronic fraction bias. **a.** Pronounced length bias in comparison of whole cell and nuclei expression estimates. The scatter plot shows mean expression ($\log_{10}$ FPM) levels calculated across the measured nuclei (x-axis) and cells. The genes with mean $\log_2$ expression fold ratio (M value) over 3 are shown in red (higher in whole cells) and green (higher in nuclei). **b.** Correction based on the derived model (Fig. 4d) eliminates systematic length bias between whole-cell and nuclei estimates. The plot shows analysis of corrected values in the same way as shown in (a). **c.** Heatmap of S1 excitatory neuronal marker gene expression showing improved consistency across single cell/nuclei data sets after gene length bias correction.
- **Fig. S5.** Retained cell type classification after gene length bias correction. **a.** t-SNE plots showing expected identity of cluster groupings using corrected expression data and based on markers used in Fig. 1d. **b.** Heatmap of expression for clusters identified in (a) using cell type enriched marker genes Supplementary Table S2).
- **Fig. S6.** Differentially detected transcripts are more exclusively detected in cellular data. **a.** Heatmap of expression ($\log(\text{TPM}+1)$) for top ($p < 1 \times 10^{-20}$) differentially detected genes between nuclei and whole cells. **b.** Heatmap of expression ($\log(\text{TPM}+1)$) for bottom ($p \geq 1 \times 10^{-20}$) differentially detected genes between nuclei and whole cells. **c.** Scatter plot for nuclear and cellular data using using genes associated with the mouse genome-scale metabolic reconstruction (iMM1415, 1088 of 1415 genes detected). Pearson correlation

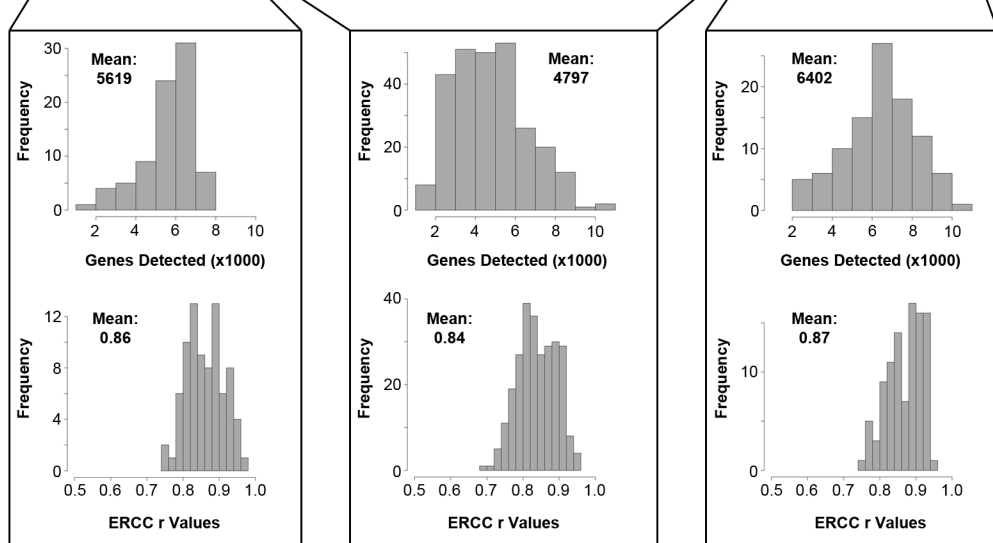
62 coefficient (r) is indicated. **d.** Heatmap of expression ($\log(\text{TPM}+1)$) for detected IMM1415
63 genes.

Fig. S1

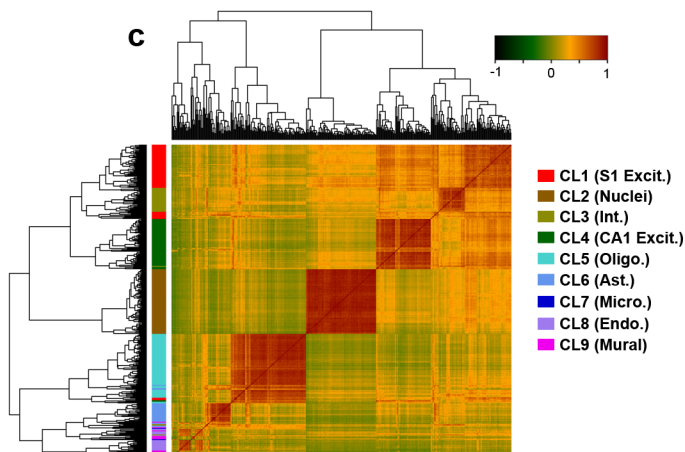
a



b



c



d

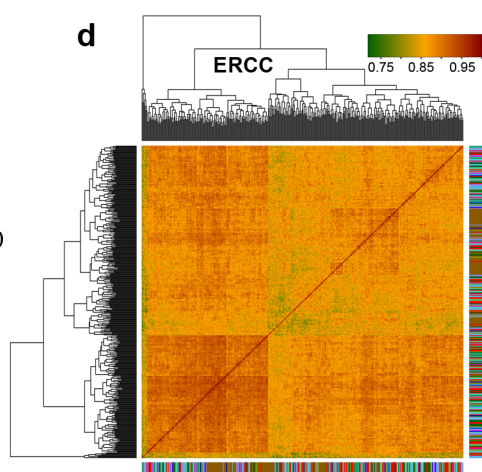


Fig. S2

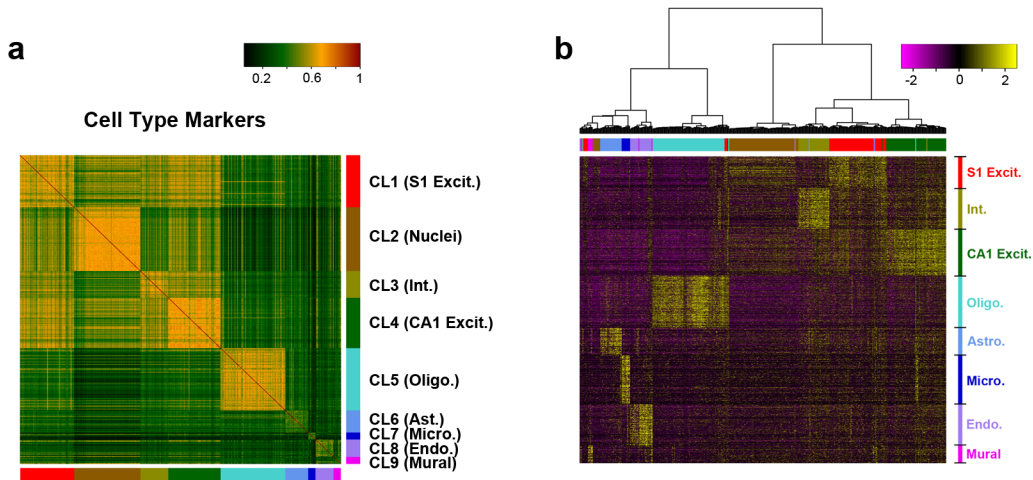


Fig. S3

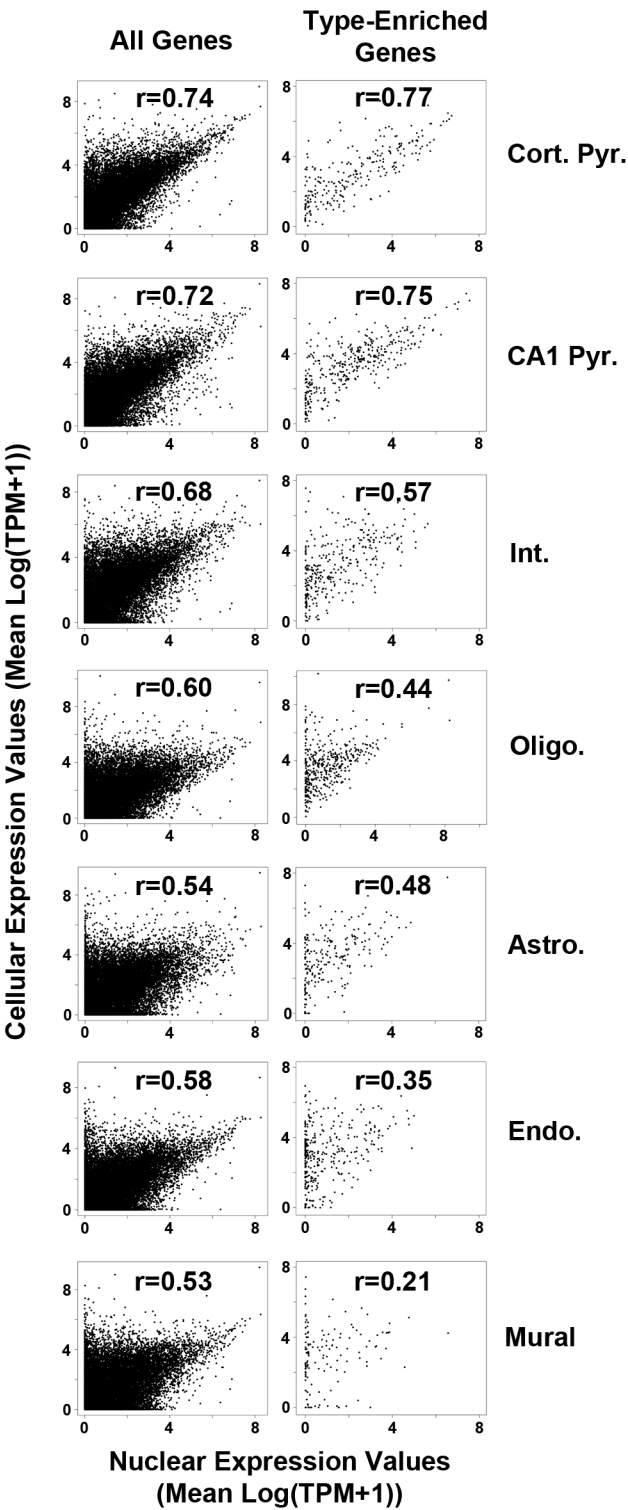


Fig. S4

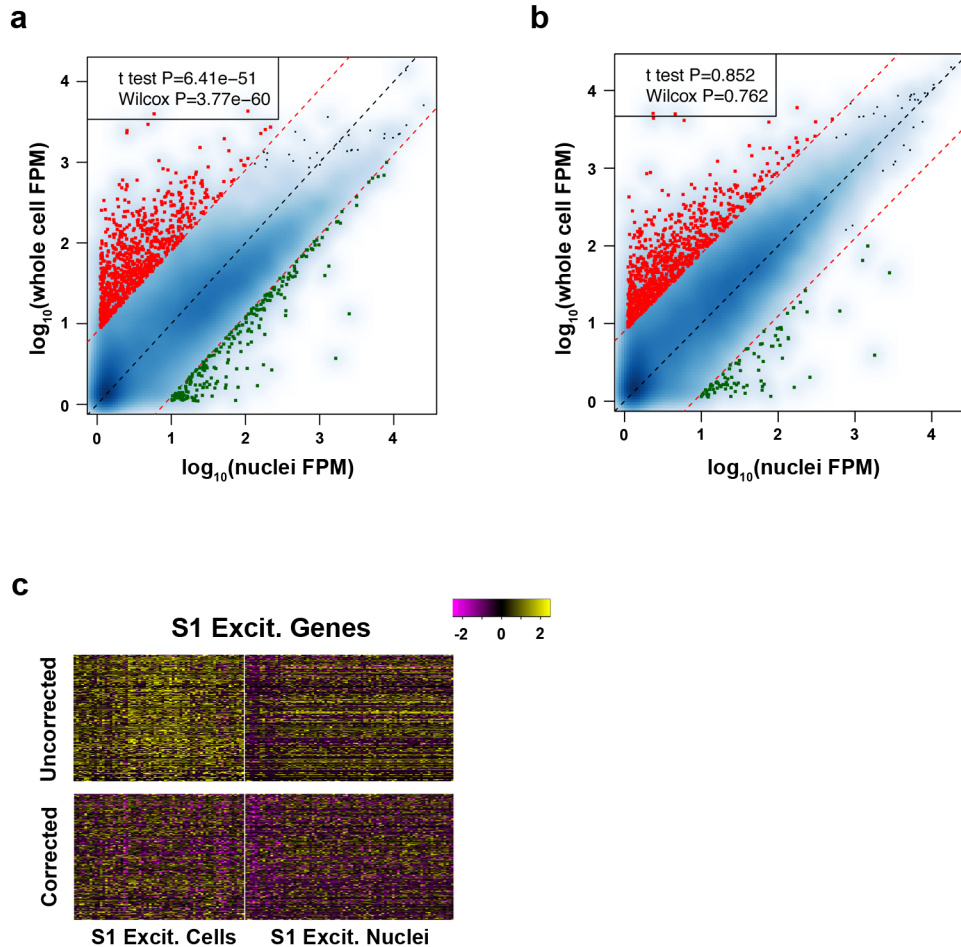
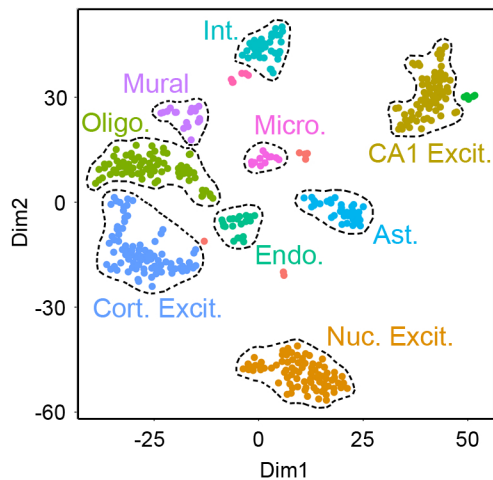


Fig. S5

a



b

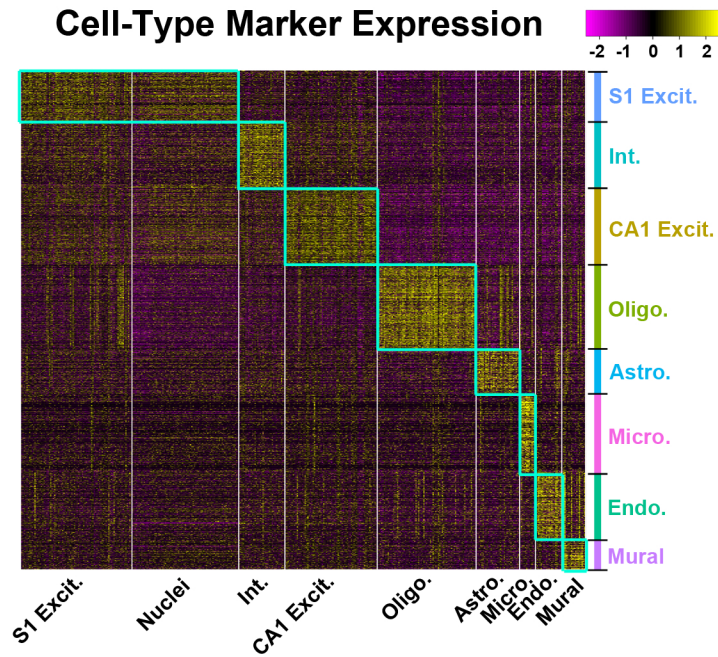


Fig. S6

