Contents

1 Supplemental Methods S1: Modeling and Simulation

For a given gene q and a tissue with m cell types, we use the vector $\mathbf{c}_q$ to represent its Cell Type expression profile (CT profile), which contains the expression level of q in each of the m cell types:

$$\mathbf{c}_q = [\, t_{q1},\, t_{q2},\, ...,\, t_{qm}\,], \qquad t_{q1}, t_{q2}, ..., t_{qm} \geq 0,\, \sum_{i=1}^{m} t_{qi} > 0 \tag{1}$$

Where t_{qi} is the mean expression level of q in cell type i. We denote the variance of values t_{qi}s by $var(t_q)$, that is, the variance of the expression level of q among different cell types.

For a given dataset with n samples from the tissue, vector $\boldsymbol{\alpha}_j$ represents the Cell Type Composition (CTC) in sample j:

$$\boldsymbol{\alpha}_j = [\, \alpha_{j,1},\, \alpha_{j,2},\, ...,\, \alpha_{j,m}\,], \qquad \alpha_{j,i} \geq 0, \sum_{i=1}^{m} \alpha_{j,i} = 1 \tag{2}$$

Where m is the count of cell types in the tissue, as it was in definition 1 and $\alpha_{j,i}$ is the proportion of cell type i in sample j. Matrix $\mathbf{A}$ defined as:

$$\mathbf{A} = \left[\boldsymbol{\alpha}'_1,\, \boldsymbol{\alpha}'_2\,, ..., \boldsymbol{\alpha}'_n\right] \tag{3}$$

contains cellular composition vectors for the n samples of the dataset.

1.1 Simulation 1 - variance of a gene in bulk tissue

Having the above definitions, the expression level of q in n samples of the dataset is given as $\mathbf{e}_q$:

$$\mathbf{e}_q = \mathbf{c}_q\, \mathbf{A} \tag{4}$$

We denote the variance of the expression level of q among the n samples of dataset presented in $\mathbf{e}_q$ by $var(q)$. From the above formulation, it is apparent that for the two special cases when $t_{q1} = t_{q2} = ... = t_{qm}$, (i.e $var(t_q = 0)$ or when $\boldsymbol{\alpha}_1 = ... = \boldsymbol{\alpha}_n$, we have $var(q) = 0$. In a bulk tissue dataset, these two cases are equivalent to when q has the same expression level in all cell

types and when the cellular composition remains the same among the samples.

For simulation, vectors c (CT expression profiles) were generated for 1000 genes for $m = 10$ cell types. For each gene q, $\mathbf{c}_q$ was generated using normal distribution with mean and variance obtained from a CT expression profile of a randomly selected gene from the sNuc-seq dataset. Bulk tissue data was generated with $n = 100$ samples. Matrix A was generated using uniform distribution with the criteria that each column has sum one. The results show that among the genes, $var(t_q)$ and $var(q)$ are highly correlated (see Supplemental Fig S1).

MATLAB code for simulation

```
myVar % variance of the CT expression profiles obtained from sNuc-seq data
myMean % mean of the CT expression profiles obtained from sNuc-seq data

m = 10 % count of tissues
n = 100 % count of bulk samples

% normalizing the weight matrix
A = rand(m,  n);
for i = 1:n
    A(:, i) = A(:, i) ./(sum(A(:,i)));
end

orVar = zeros(1,1000); % variance of the CT profile
obVar = zeros(1,1000); % the observed variance in bulk tissue
for k = 1:1000 % for 1000 genes
    ind = datasample(1:length(myVar), 1);
    v = myVar(ind);
    orVar(k) = v;
    c = normrnd(myMeans(ind), sqrt(v), 1, t);
    exps = c * A;
    obVar(k) = var(exps);
end
```

1.2 Simulation 2 - correlation of two genes in bulk tissue

For given two genes p and q, with expression vectors $\mathbf{e}_p$ and $\mathbf{e}_q$ in the bulk tissue dataset, the higher the correlation of their CT profiles $\mathbf{c}_p$ and $\mathbf{c}_q$ (in any direction), the more likely that $\mathbf{e}_p$ and $\mathbf{e}_q$ are also correlated in the same direction. Having centered $\mathbf{c}_p$ and $\mathbf{c}_q$ as $\mathbf{c}_p^c = \mathbf{c}_p - \bar{\mathbf{c}}_p$ and $\mathbf{c}_q^c = \mathbf{c}_q - \bar{\mathbf{c}}_q$ and formulation (4), since $corr(\mathbf{c}_p, \mathbf{c}_q) = corr(\mathbf{c}_q^c, \mathbf{c}_p^c) = cos(\mathbf{c}_p^c, \mathbf{c}_q^c)$; correlation

between $\mathbf{e}_p$ and $\mathbf{e}_q$ can be written as:

$$\begin{aligned} corr(\mathbf{e}_p, \mathbf{e}_q) &= corr\big(\ [\ \mathbf{c}_p^c \cdot \alpha_1',\ \mathbf{c}_p^c \cdot \alpha_2', ..., \mathbf{c}_p^c \cdot \alpha_n'], [\ \mathbf{c}_q^c \cdot \alpha_1',\ \mathbf{c}_q^c \cdot \alpha_2', ..., \mathbf{c}_q^c \cdot \alpha_n']\big) \\ &= corr\big(\ [\ \|\alpha_1\| cos(\mathbf{c}_p^c, \alpha_1),\ \|\alpha_2\| cos(\mathbf{c}_p^c, \alpha_2), ...,\ \|\alpha_n\| cos(\mathbf{c}_q^c, \alpha_n)\ ] \\ &\qquad [\ \|\alpha_1\| cos(\mathbf{c}_q^c, \alpha_1),\ \|\alpha_2\| cos(\mathbf{c}_q^c, \alpha_2), ...,\ \|\alpha_n\| cos(\mathbf{c}_q^c, \alpha_n)\ ]\ \big) \end{aligned} \tag{5}$$

Where $cos(a, b)$ is the cosine of the angle between the two vectors. As the angle between $\mathbf{c}_p^c$ and $\mathbf{c}_q^c$ gets smaller (equivalent to their correlation getting higher), the difference between $cos(\mathbf{c}_p^c, \alpha_i)$ and $cos(\mathbf{c}_q^c, \alpha_i)$ gets smaller and therefore correlation between $\mathbf{e}_p$ and $\mathbf{e}_q$ gets higher.

Figure 3C shows simulation results for correlation of CT expression profiles versus the observed correlation in the bulk tissue for 1000 gene pairs.

MATLAB code for the simulation

```
n = 100
CTCorr = zeros(1, 1000); % correlation of the CT expression profiles
sampleCorr = zeros(1, 1000); % correlation of the gene pair in the bulk tissue
for k = 1:1000
    P = (randn(1,10)*3 + 4);  % generating CT profile for gene P
    Q = (randn(1,10)*3 + 4);  % generating CT profile for gene Q
    Gs = [P, Q];

    CTCorr(k) = corr(P', Q');

    A = rand(10,  n);
    % normalizing the weight matrix
    for i = 1:n
        A(:, i) = A(:, i) ./(sum(A(:,i)));
    end

    % getting the final vectors
    exps = Gs * A;

    sib = corr(exps');
    sampleCorr(k) = sib(1, 2);
end
```

2 Supplemental Methods S2: Gene variance explained by variance of the markers

We examined the results from the Principal Component Regression method from the two sets of marker genes from the same five cell types - Astrocyte, Microglia, Oligodendrocyte, Endothelial and Pyramidal. This was to study the specificity of the model in capturing cellular composition induced variance (refelected specifically in the variation of the marker genes) as oppose to variance induced by a general confounding factor shared among all the genes. The two sets of marker genes are identified independently and have a small overlap (see Supplemental Fig. S3).

3 Supplemental Figures

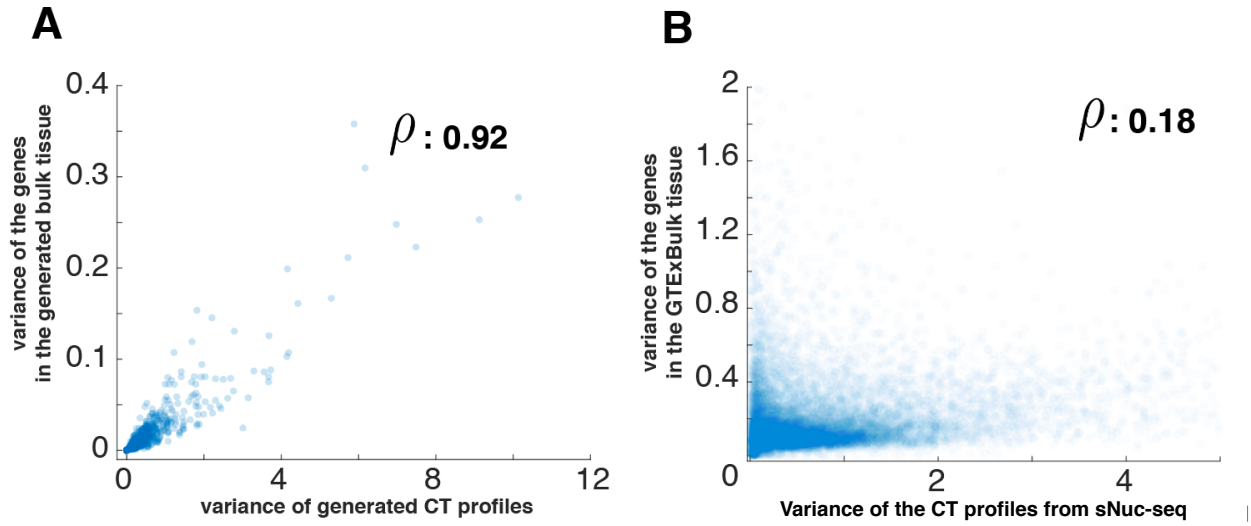

Figure S1: Variance of the CT profiles versus the variance observed in the bulk tissue. (A) Results from simulation for 1000 genes. Each point is data from a gene. **(B)** Results from data: CT profiles estimated using sNuc-seq data, compared to the observed variance in the GTExBulk tissue dataset.

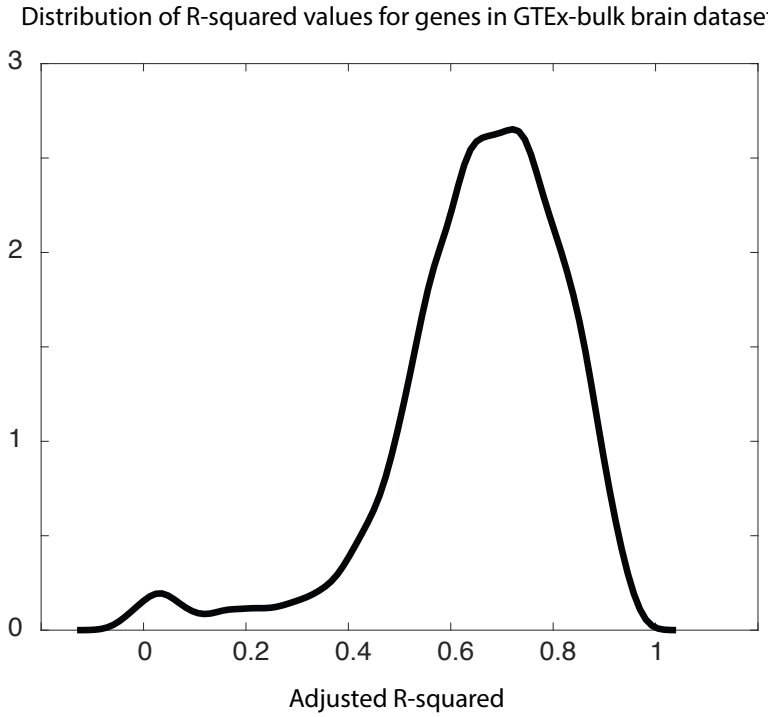

Figure S2: Distribution of the R-squared values for genes in the GTExBulk coexpression network.

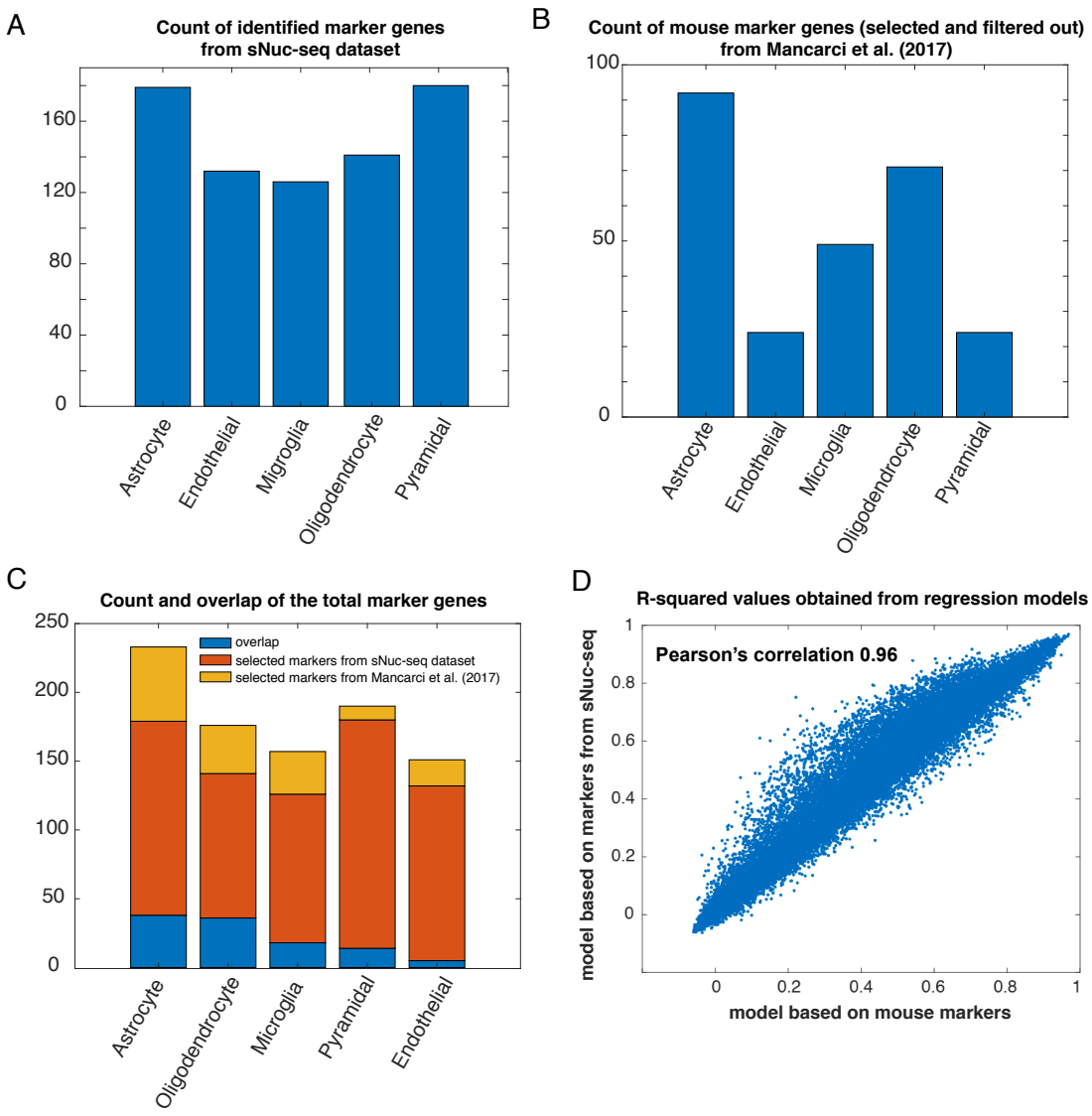

Figure S3: Comparison of two sets of marker genes. (A) Count of marker genes identified in sNuc-seq dataset. **(B)** Marker genes from Mancarci et al. (2017). **(C)** Overlap of the marker genes between the two sources. (D) R-squared values from the two sets of marker genes are highly correlated.

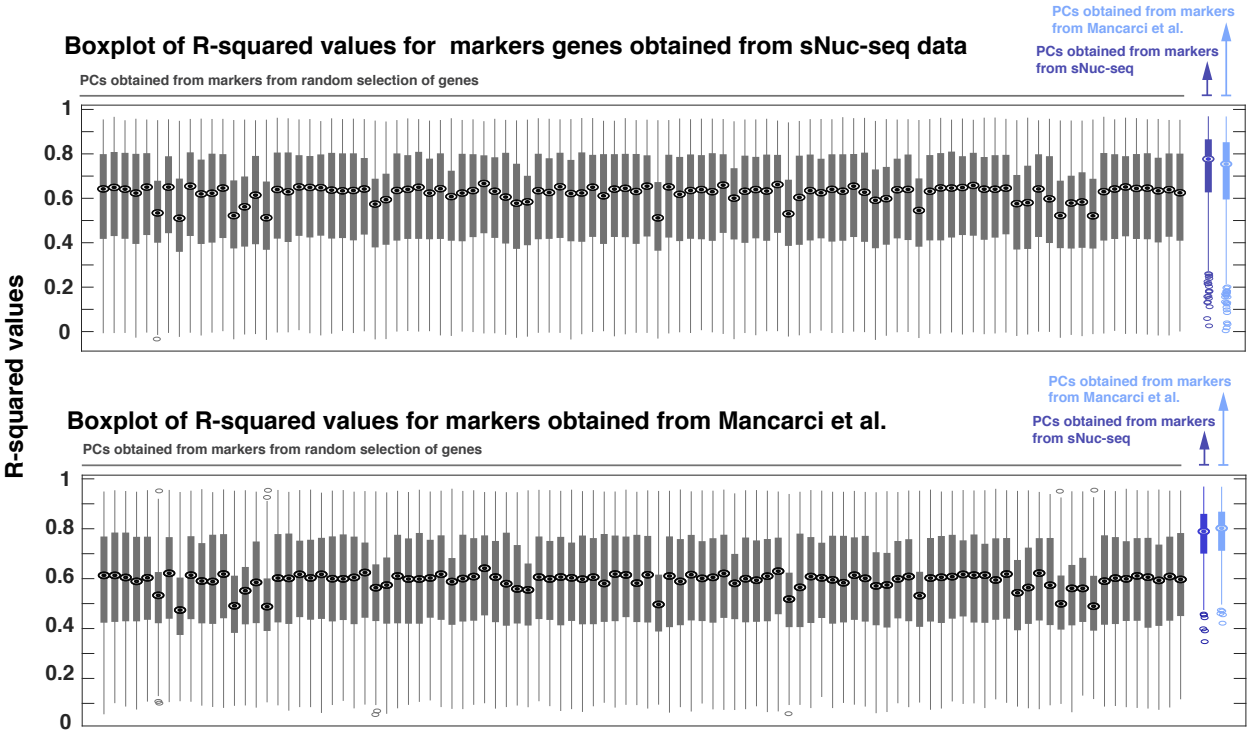

Figure S4: Comparison of the distribution of R^2 values for marker genes, when PCs are obtained from random selection of genes versus when they are obtained from marker genes.

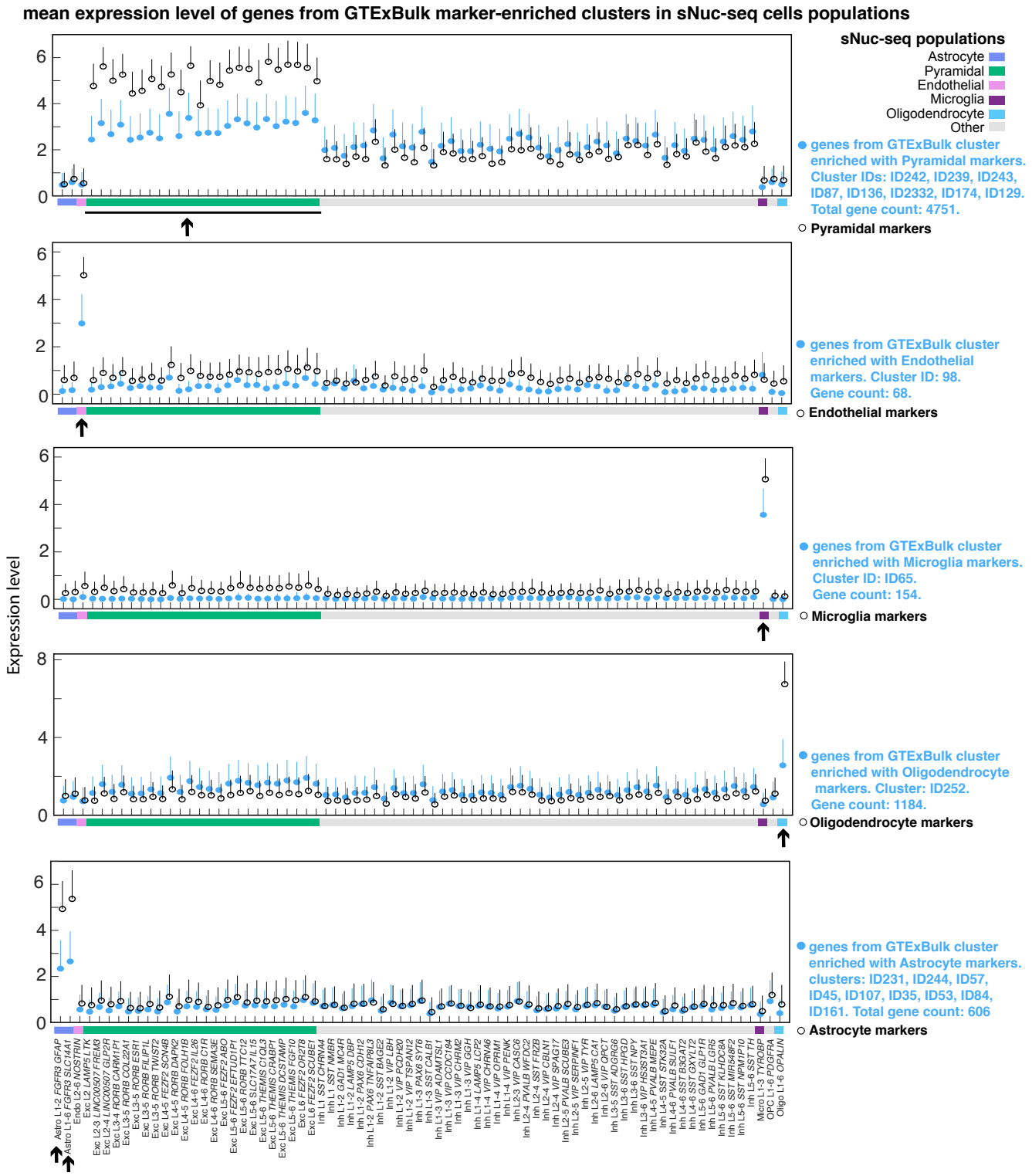

Figure S5: Mean expression level of genes from GTExBulk marker-enriched clusters in different cell populations of the sNuc-seq data. Each plot shows the expression levels for genes from GTExBulk clusters enriched with markers of specific cell types (from top to bottom: Pyramidal, Endothelial, Microglia, Oligodendrocyte and Asctrocyte). Lines show the standard deviation. Generally, genes in marker-enriched clusters have higher expression levels in the relevant cell populations -marked by an arrows- compared to other cell populations.

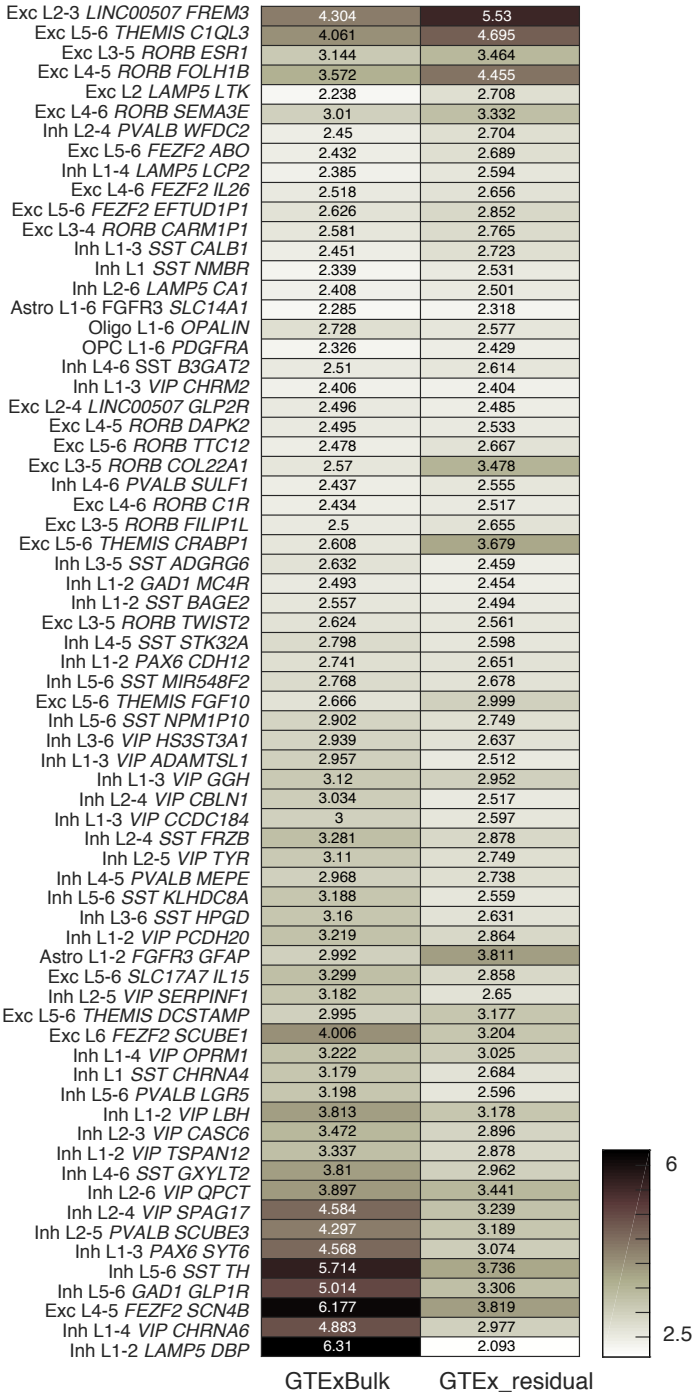

Figure S6: Ratio of the observed versus the expected link overlap of the links from sNuc-seq networks with GTExBulk and GTEx_residual networks. sNuc-seq populations are sorted based on the count of cells, with the top one having the highest count of cells. Generally, the link overlap between GTExBulk and GTEx_residual doesn't change much and there is some level of agreement between sNuc-seq population networks and the two bulk networks.

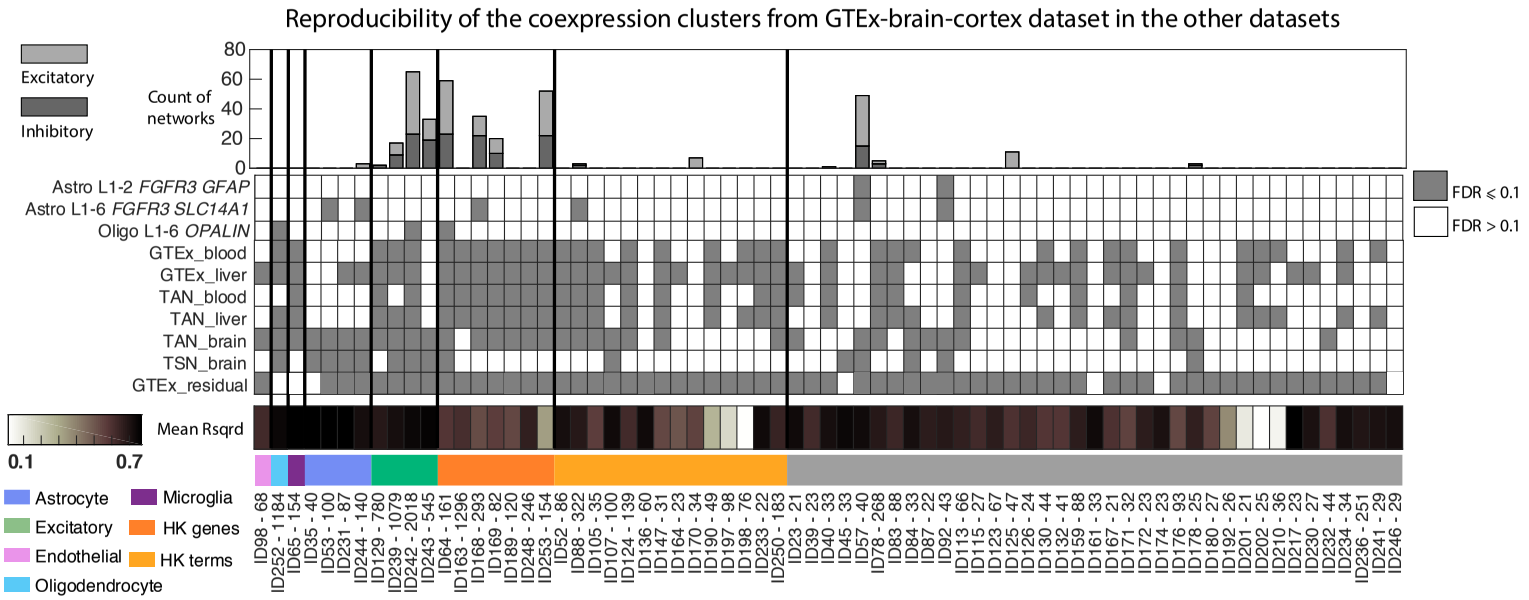

Figure S7: Reproducibility of clusters identified in GTExBulk in other networks. Rows in the heatmap show the reproducibility of clusters in different networks. Each column in the heatmap (and the same column in the bar plot) shows data for one cluster identified in GTExBulk network. The gray color in the heatmap shows that the cluster has significantly high count of links (FDR $\leq$ 0.1). The top bar plot shows, for each cluster, the count of Excitatory and Inhibitory networks (built from populations of cell types identified in sNuc-seq data) in which the cluster has significantly high count of links (FDR $\leq$ 0.1). The bottom color bar shows if the cluster has markers of specific cell types, enriched by housekeeping genes or functional terms.

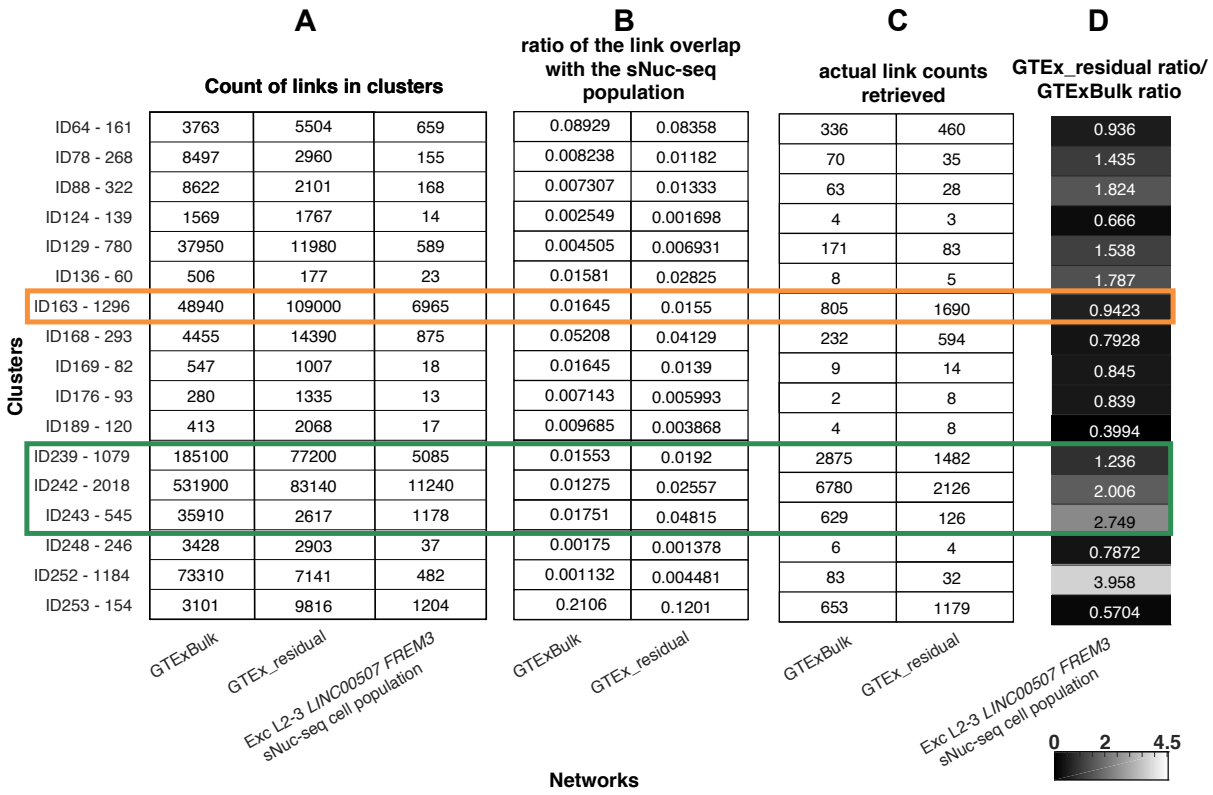

Figure S8: Representation of the links in a sNuc-seq population (Exc L2-3 *LINC00507 FREM3*). This population was selected because it has the highest count of cells among all the sNuc-seq populations and therefore a more complete network regarding the gene count. **A** shows count of links in each of the GTExBulk clusters (clusters were selected to have 10 or more links in the sNuc-seq population). **B** shows ratio of links from GTExBulk and GTEx_Residual networks that overlap with links in the sNuc-seq network. **C** shows the actual count of links from sNuc-seq network retrieved (has overlap) in GTExBulk and GTEx_residual networks. **D** shows the proportion of the ratios in **B**. For the Pyramidal clusters in the green box, although the count of links has decreased, precision has almost doubled for the GTEx_residual network. For the housekeeping cluster in the orange box, count of links retrieved is more than double in GTEx_residual network compared to GTExBulk network and the precision has not changed much.

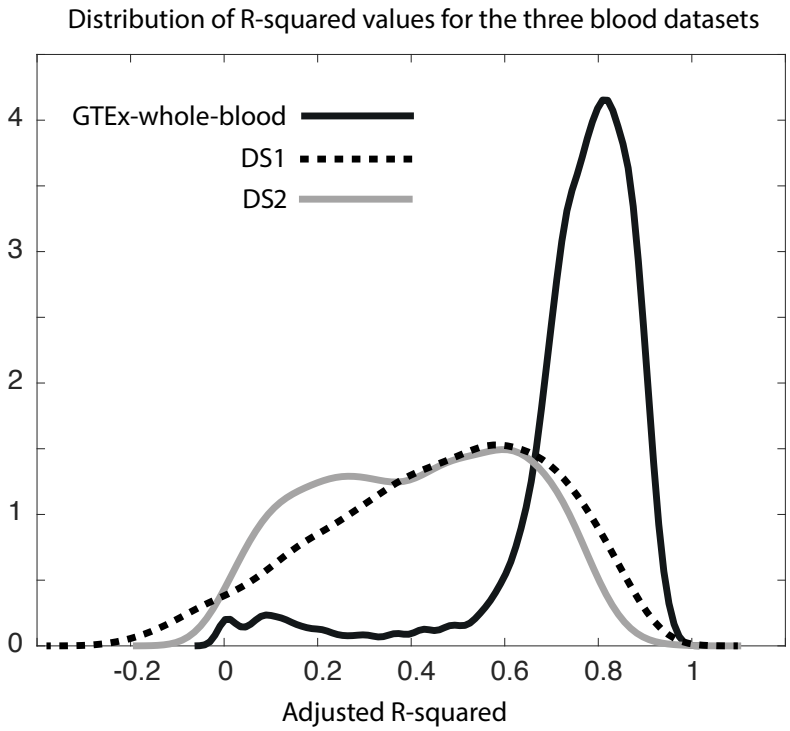

Figure S9: Distribution of the R-squared values for genes in three blood datasets.

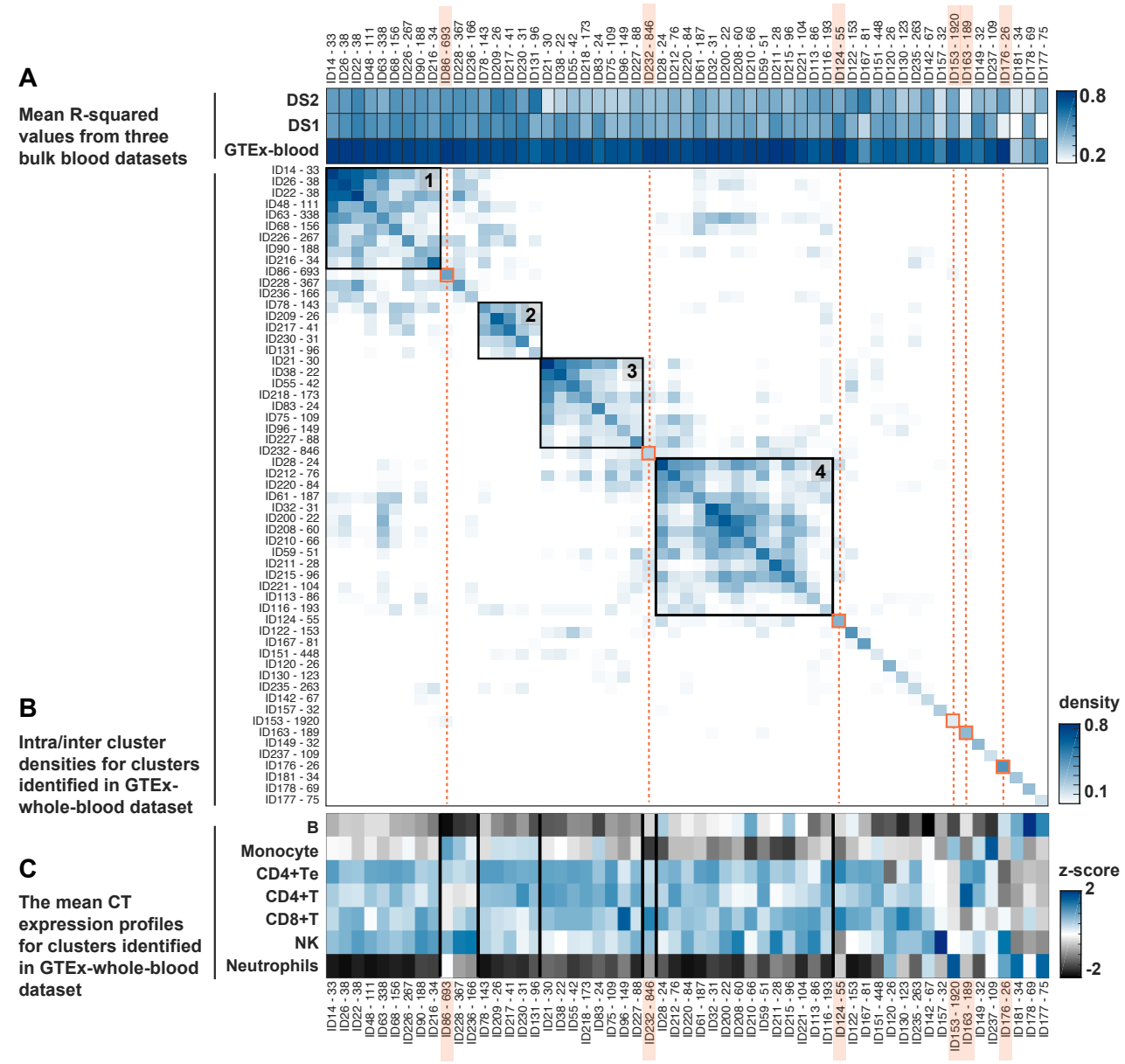

Figure S10: Cell type expression profiles of genes in whole blood dataset can explain grouping of the genes in coexpression clusters. (A) The heatmap shows the average R^2 values for coexpression clusters (identified in GTEx-whole-blood network) in GTEx-whole-blood and two microarray bulk blood datasets (DS1 and DS2). Clusters are labeled with their ID and count of genes. Clusters with > 2 genes identified as differentially expressed among multiple blood cell types are highlighted with color orange. The dotted lines trace their inter and intra cluster density in the heatmap in **B**. **(B)** The heatmap shows grouping of coexpression clusters identified in GTEx-whole-blood dataset. Four major groups of clusters are identified and labeled. **(C)** The heatmap shows zscores for average CT expression profiles for the clusters. All the four groups have low expression levels in Neutrophils, while expression level of Monocytes and B cells vary between them.

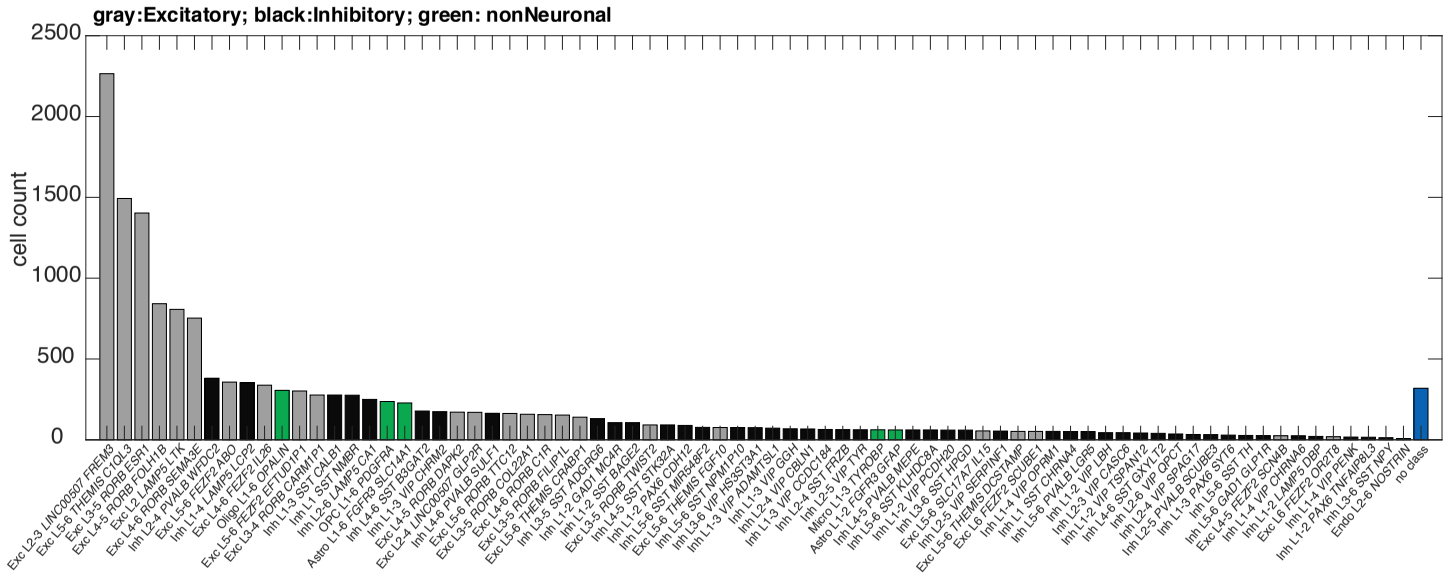

Figure S11: Count of samples (cells) in each cluster of identified cell types for the Allen institude sNuc-seq dataset.

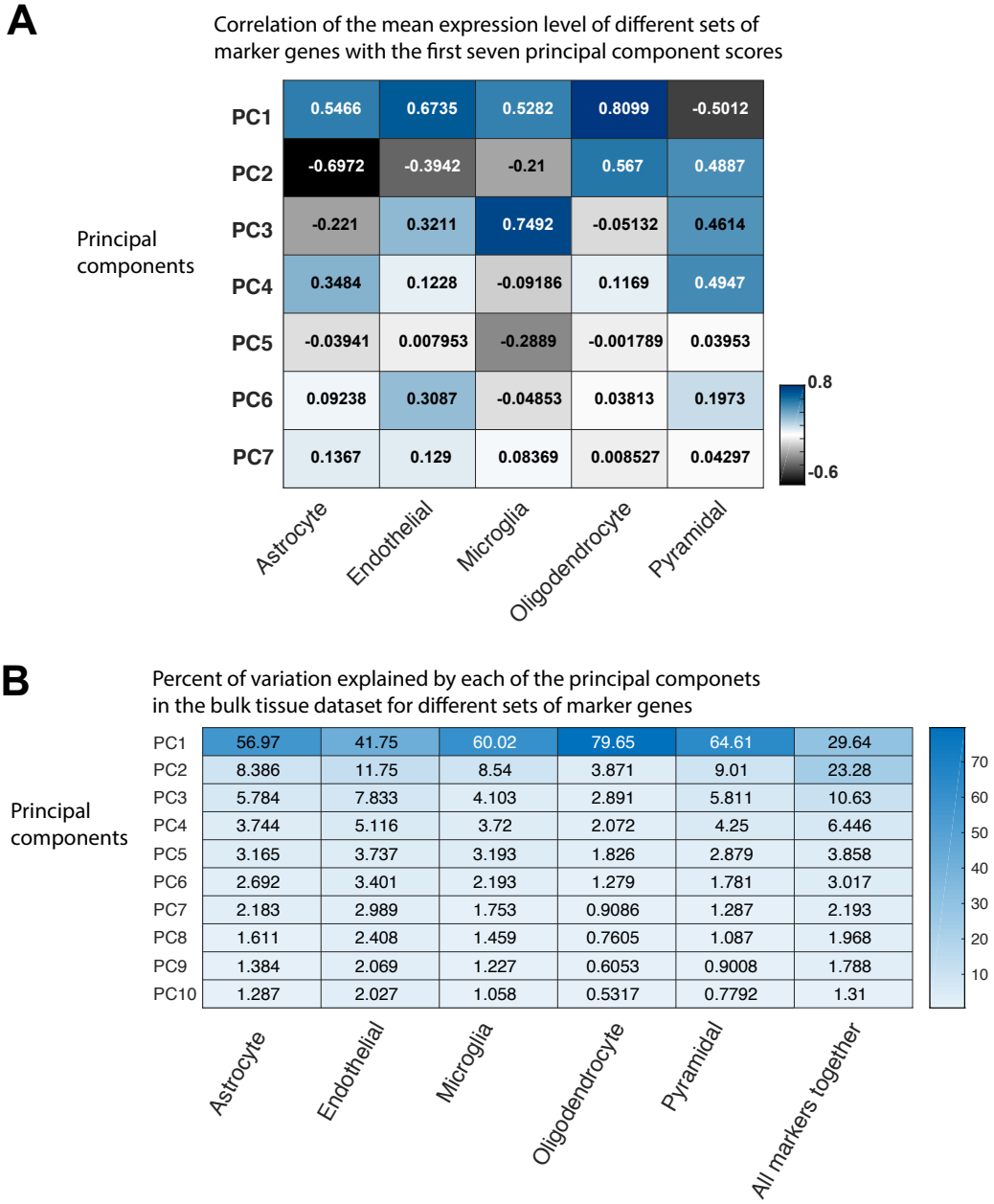

Figure S12: Cellular compositional varince captured by principal components. **A** shows correlation of the mean zscore of expression of sets of marker genes among the GTExBulk (brain) samples, with the first seven principal component scores obtained from the whole set of marker genes. **B** shows the percent of variation explained by first ten principal components of different marker sets as well as all markers together.

4 Supplemental Tables

Table S1: Count of links and genes in each of the networks

networks	gene count	link count
GTEx brain cortex	14,102	1,774,291
GTEx blood	11,348	1,312,228
GTEx liver	16,388	1,481,028
TAN brain	8,761	1,118,791
TAN blood	8,747	380,875
TAN liver	10,227	496,476
TSN brain	6,422	358,531

	Count of genes in the networks	Count of inks in the networks
Exc L2-3 *LINC00507 FREM3*	10610	563000
Exc L5-6 *THEMIS C1QL3*	10710	573300
Exc L3-5 *RORB ESR1*	8509	362000
Exc L4-5 *RORB FOLH1B*	11040	609700
Exc L2 *LAMP5 LTK*	8530	363800
Exc L4-6 *RORB SEMA3E*	9194	422600
Inh L2-4 *PVALB WFDC2*	9107	410900
Exc L5-6 *FEZF2 ABO*	10140	508900
Inh L1-4 *LAMP5 LCP2*	8169	326000
Exc L4-6 *FEZF2 IL26*	9410	434500
Exc L5-6 *FEZF2 EFTUD1P1*	10810	573000
Exc L3-4 *RORB CARM1P1*	10340	515300
Inh L1-3 *SST CALB1*	5968	157300
Inh L1 *SST NMBR*	7839	278400
Inh L2-6 *LAMP5 CA1*	8656	342100
Astro L1-6 FGFR3 *SLC14A1*	4427	70560
Oligo L1-6 *OPALIN*	3286	32340
OPC L1-6 *PDGFRA*	3281	29850
Inh L4-6 SST *B3GAT2*	7402	198500
Inh L1-3 *VIP CHRM2*	7583	201700
Exc L2-4 *LINC00507 GLP2R*	8894	317200
Exc L4-5 *RORB DAPK2*	8636	285700
Exc L5-6 *RORB TTC12*	10480	464400
Exc L3-5 *RORB COL22A1*	8420	252100
Inh L4-6 *PVALB SULF1*	8339	246800
Exc L4-6 *RORB C1R*	9216	323000
Exc L3-5 *RORB FILIP1L*	9276	321500
Exc L5-6 *THEMIS CRABP1*	10150	411200
Inh L3-5 *SST ADGRG6*	7035	130400
Inh L1-2 *GAD1 MC4R*	6328	79800
Inh L1-2 *SST BAGE2*	5777	64540
Exc L3-5 *RORB TWIST2*	7885	147100
Inh L4-5 *SST STK32A*	5401	55640
Inh L1-2 *PAX6 CDH12*	7204	118000
Inh L5-6 *SST MIR548F2*	8185	177600
Exc L5-6 *THEMIS FGF10*	9303	271200
Inh L5-6 *SST NPM1P10*	7086	116700
Inh L3-6 *VIP HS3ST3A1*	6367	84610
Inh L1-3 *VIP ADAMTSL1*	5939	73990
Inh L1-3 *VIP GGH*	4470	42500
Inh L2-4 *VIP CBLN1*	3773	29240
Inh L1-3 *VIP CCDC184*	5904	78420
Inh L2-4 *SST FRZB*	4674	48690
Inh L2-5 *VIP TYR*	4724	48510
Inh L4-5 *PVALB MEPE*	7049	127800
Inh L5-6 *SST KLHDC8A*	5764	74940
Inh L3-6 *SST HPGD*	5986	87560
Inh L1-2 *VIP PCDH20*	4353	42050
Astro L1-2 *FGFR3 GFAP*	917	2964
Exc L5-6 *SLC17A7 IL15*	6698	119400
Inh L2-5 *VIP SERPINF1*	2972	20940
Exc L5-6 *THEMIS DCSTAMP*	7535	173700
Exc L6 *FEZF2 SCUBE1*	6841	138800
Inh L1-4 *VIP OPRM1*	2960	20860
Inh L1 *SST CHRNA4*	3136	23040
Inh L5-6 *PVALB LGR5*	3522	27870
Inh L1-2 *VIP LBH*	4474	51470
Inh L2-3 *VIP CASC6*	4596	54250
Inh L1-2 *VIP TSPAN12*	2550	16730
Inh L4-6 *SST GXYLT2*	3315	29010
Inh L2-6 *VIP QPCT*	1717	9126
Inh L2-4 *VIP SPAG17*	1074	4576
Inh L2-5 *PVALB SCUBE3*	1516	7848
Inh L1-3 *PAX6 SYT6*	2202	15420
Inh L5-6 *SST TH*	1450	8536
Inh L5-6 *GAD1 GLP1R*	856	3302
Exc L4-5 *FEZF2 SCN4B*	2936	32310
Inh L1-4 *VIP CHRNA6*	409	1300
Inh L1-2 *LAMP5 DBP*	158	304

Table S2: Count of genes (with one or more links) and links in different sNuc-seq networks.