Supplemental Figures

Supplemental figure legends

Figure S1 | Testing the pre-clustering heuristic. (**A**) (Left) Default, unsupervised heuristic sets a cut of 7% of the total dendrogram depth, which results in 52 pre-clusters. (Right) The numerical model calculated using the 52 pre-clusters. X_{c1} and X_{c2} represent the expression (in a binned UMIs grid) of a given gene *X* in two cells *c1* and *c2* belonging to the same pre-cluster. The cumulative distribution plot estimates the frequency, hence likelihood, of an expression change. (**B**) (Left) Forcing a cut of only 4% creates 1152 pre-clusters, more than 20-fold increase compared to the default 7% depth. Also, given the reduction of the average cluster size and the consequent reduction of possible intra-cluster pair-wise comparison, the number of data points used to fit the model decreases of more than 5-fold compared to default 7% cut (from 3.79E+9 to 6.56E+8). (Right) Despite this, the difference between the numerical model of 4% cut and 7% cut is marginal. (**C**) (Left) Forcing a cut of 20% creates only 9 pre-clusters, which is less than the number of final clusters (in this case, 11) and therefore represents a miscalculated configuration. Still the difference between the numerical model of 20% cut and 7% cut is marginal (right). (**D**) Also switching from *Pearson* to *Spearman* correlation is associated with neglectable differences in the numerical model. (**E**) (Top) **N**umber of pre-clusters associated with the different cutting depths, correlations metrics (*Pearson*, *Spearman*) or linkage metrics (*complete* or Weighted average distance, *WPGMA*, instead of default *Ward's*). *Complete* and *WPGMA* linkages were both applied with *Pearson* correlation and default 7% cutting depth. (Bottom) Rand indexes indicating nearly complete similarity between the final clustering obtained from the default pre-clustering and its variations (Rand index=1 indicates exact similarity).

Figure S2 | *BigSCale* correction of confounding signatures. (**A**) Confounding factors, such as cell cycle, gender, mitochondrial or ribosomal content can negatively condition the outcome of the

clustering. In *bigSCale*, correction of any given confounding effect is performed in two steps: 1) The set of genes creating the confounding signature is determined. To determine these genes *bigSCale* starts from the previously assigned hierarchical markers and the Z-scores which indicate, for each marker, how strongly it is expressed in each of the various populations of cells (clusters) at multiple hierarchical levels. The Z-scores are clustered to identify signatures of co-expressed marker genes via the *Jaccard* metric, in which the distance is measured as the percentage of non-zero coordinates that differ. As a result, markers expressed in the same populations (clusters) at multiple hierarchical levels will be classified as co-expressed. 2) A coefficient ranging from zero (no change) to one (complete removal) is applied to reduce or completely remove the effects of the confounding signature. In our iPSC-derived neuronal progenitor cells (NPC) dataset, one of the patients was male while the others were females. We observe that male cells tend to cluster together because of a gender signature which includes genes located on *chrY*, such as *RPS4Y1* or *PCDH11Y*. *BigSCale* determines the gender signature to consist of a set of 476 genes whose average expression is clearly clustered, same as *RPS4Y1* or *PCDH11Y* (heatmap). (**B**) Re-clustering of the dataset after correcting the gender confounding signature results in the male cells being distributed over the clusters, as shown by the individual genes plots and the average signature expression. (**C**) Next, we found two more confounding signatures persisting after the gender correction. These two signatures correspond to genes expressed in two distinct phases of the cell cycle, namely *G1/S* (93 genes) and *G2/M* (197 genes). (**D**) Again, re-clustering after correcting for both signatures efficiently prevents the cell cycle to drive the clustering. Notably, the cells lacking cell cycle genes are post-mitotic neuroblasts whose phenotypic differences compared to the remaining neuronal progenitor cells are not limited to the cell cycle (clustering separately also after the correction of cell cycle).

Figure S3 | Differential expression analysis in neuronal progenitors derived from WB (WB2, **A**) and Dup7 (Dup7.1/2, **B,C**) syndrome patients compared to a healthy control. (**A,B,C**) Differential expression analysis of genes within the disease related region using five DE tools and displaying the top 2500 genes. For the genes located in the deleted (**A**) or amplified (**B,C**) region the *p*-values

of are shown in Z-score scale (red: down-regulated; blue: up-regulated). Genes correctly assigned as down- or up-regulated are highlighted by grey. (**D**) Average number of detected down- (red) and up-regulated (blue) genes in the two WB and Dup7 patients, respectively, compared to healthy donor and using the top 2000 (left) or top 1500 (right) genes for each tool.

Figure S4 | Benchmarking *bigSCale* using simulated datasets. (**A**) Characteristics of the simulated datasets sim_10x (red) and sim_NPC (green) in terms of library sizes (left), distribution of zeros per gene (middle) and distribution of zero per cell (right). (**B-E**) Partial AUCs of ROC curves computed across tools in the simulated datasets sim_10x (**B,C**) and sim_NPC (**D,E**). The *bigSCale* method shows highest sensitivity at high specificity (>90%, grey area) at both group size conditions (1:2, **B,D** and 1:10, **C,E**).

Figure S5 | Partial AUCs of ROC curves computed across the *Seurat*'s alternative tests in the two simulated datasets (NPC; 10x Genomics) with group sizes having proportions 1:1 (1x), 1:2 (2x) and 1:10 (10x). The sensitivity at high level of specificity (>90%) is highlighted (grey area).

Figure S6 | (**A**) Comparison of *bigSCale* with default normalization (library size) or following *SCRAN* normalization. Partial AUCs of ROC curves computed in the two simulated datasets (NPC; 10x Genomics) with group sizes having proportions 1:1 (1x), 1:2 (2x) and 1:10 (10x). (**B**) Scatter plots of the normalization coefficients of *SCRAN* compared to library size. The correlation is nearly perfect with *Pearson* or *Spearman* metrics.

Figure S7 | *BigSCale* analysis of 3,005 mouse cortical and hippocampal cells (Zeisel et al. 2015). (**A**) Comparison of *bigSCale* and *BackSPIN* in the detection of markers for oligodendrocytes in the turquois cluster (high expression, yellow; low expression, blue). *BigSCale* identified 126 additional markers with high specificity for oligodendrocytes and markers uniquely identified by *BackSPIN* display a weaker specificity and achieved low scoring in *bigSCale*. (**B**) Hierarchical signature markers of *bigSCale*. Signatures of different hierarchical levels exemplified by unique

vascular, interneuronal and Pyr3 (Level 1) signatures and shared, higher level markers expressed in *Cornu Ammonis* Pyramidal neurons (Level 3) or generally in neurons (Level 6).

Figure S8 | Population and marker identification by *bigSCale*. (**A**) Deconvolution of 3,005 single cells of the adult brain using tSNE representation indicating nine main subpopulations. (**B,C**) Markers unique for the *bigSCale* analysis identified for neurons (**B**) or astrocytes (**C**). The tSNE plots highlight the expression levels of the neural markers *Stmn3* and *Snap25* and the astrocyte markers *Aqp4*, *Atp1a2*, *Mt1* and *Slc1a3* (high expression, blue; low expression, yellow).

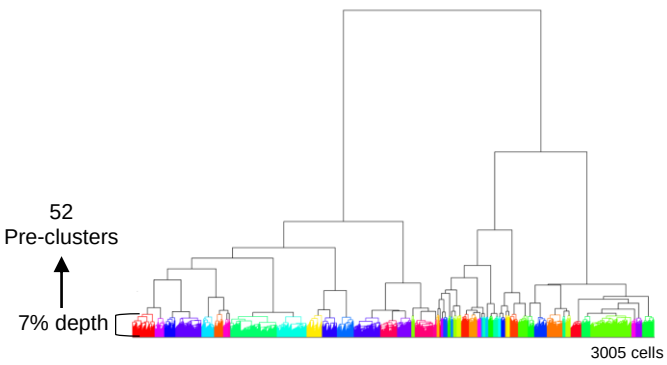
Figure S9 | Differential expression of neurotransmitter receptors in subpopulations of Cajal-Retzius cells. Heatmap of Z-scores representing the relative expression level (higher expression, red; lower expression, blue) of each receptor subunit in each cell cluster (CR1-CR8) Besides the AMPA receptors (Gria1-4), the most significant variation was detected for *Grin2b* (NMDA receptor), *Grm2* (Glutamate Metabotropic receptor) and *Gabra2* (GABA(A) receptor).

Figure S10 | The *bigSCale* analytical framework. (**A**) Surface-plot for the numerical approximation of a cumulative distribution function (computed for the adult brain dataset (Zeisel et al. 2015). (**B**) Section of the cumulative distribution function. The plot shows the *bigSCale* estimated likelihood of an expression change for a gene with expression x=10 UMIs in cell A and y UMIs in cell B. Example from adult brain dataset (Zeisel et al. 2015). (**C**) Differential expression. Example of the relation of the raw scores (y-axis) with the number of non-null comparisons (x-axis). A null-comparison results from genes with zero UMIs in both compared cells. (**D**-**G**) Steps for determining overdispersed genes (example from the NPC cell dataset). (**D**) Mean-variance relationship is fitted with a smoothing spline. The fitted line (red) is next used to normalize the relationship obtaining the plot shown in (E). (**E**) The mean-variance relationship is further normalized by diving for the smoothing spline of the standard deviation to obtain the plot shown in (F). (**F**) Genes exceed the threshold (default: Z-score=2) are selected as overdispersed genes. (**G**) Skewed genes are discarded. The plot shows for each gene (dot) the relationship

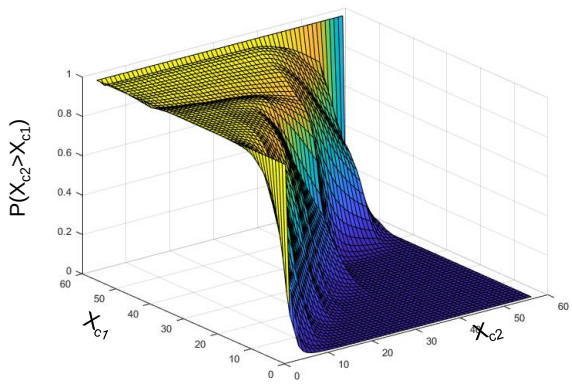
between the top 10 highest expression values and the mean-normalized standard deviation. The red circled genes show very high dispersions in their top 10 values, pointing to a high expression in a small number of cells (outliers). (**H**) Batch effect removal. Example of the procedure for the gene *EIF4H* in the NPC syndrome dataset. The dataset contains 1,920 cells divided in 10 batches of 192 cells each. The distribution of UMIs for *EIF4H* is variable between the batches, indicating the presence of batch effects (left). *BigSCale* forces the batches to follow the same distribution, calculated as the weighted average of all batches (right). The procedure is iterated gene by gene, for each condition. (**I**) Batch effect removal and changes in read counts. To remove batch effects *bigSCale* re-assigns UMIs/reads to genes, resulting in changes in library size. To ensure that re-assignment of UMIs/reads did not introduce artifacts, we compared library size before and after removing batch effects. Batch removal had minor effects on the library sizes, as shown by the close linear relationship in the scatter plot ($R^2=0.995$, left). Histogram plots of the normalized library sizes further showed minimal artifacts introduced by the procedure (right).

Supplemental Figure 1

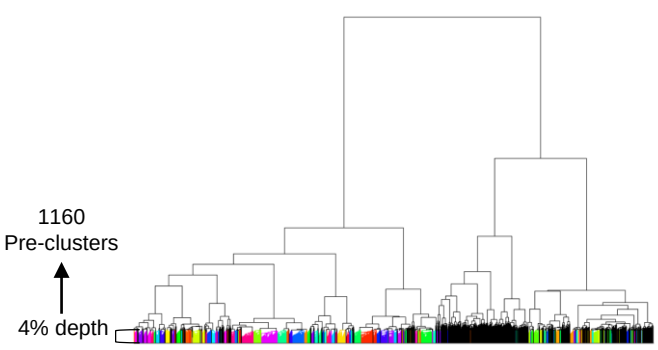
A PRE-CLUSTERING (*PEARSON*, UNSUPERVISED, 7%)



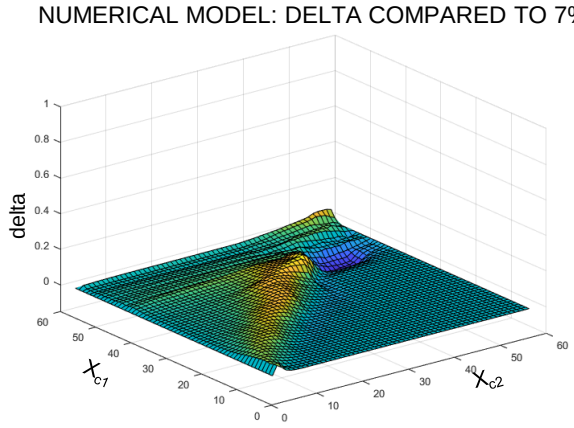
NUMERICAL MODEL



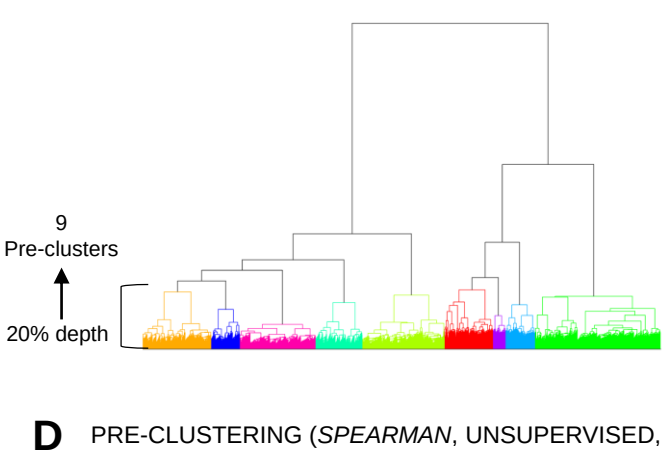
B PRE-CLUSTERING (*PEARSON*, FORCED TO 4%)



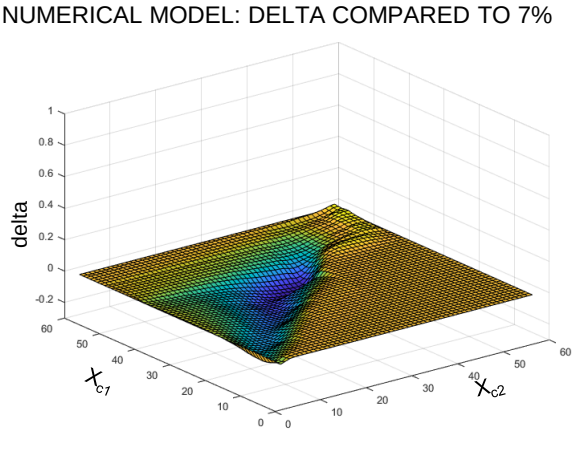
NUMERICAL MODEL: DELTA COMPARED TO 7%



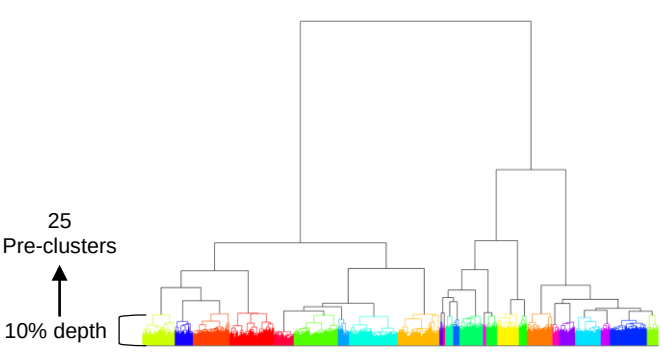
C PRE-CLUSTERING (*PEARSON*, FORCED TO 20%)



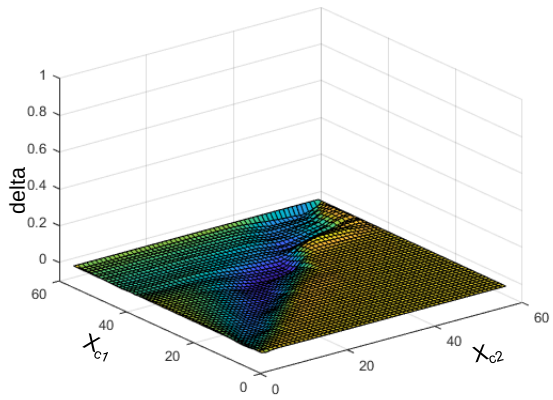
NUMERICAL MODEL: DELTA COMPARED TO 7%



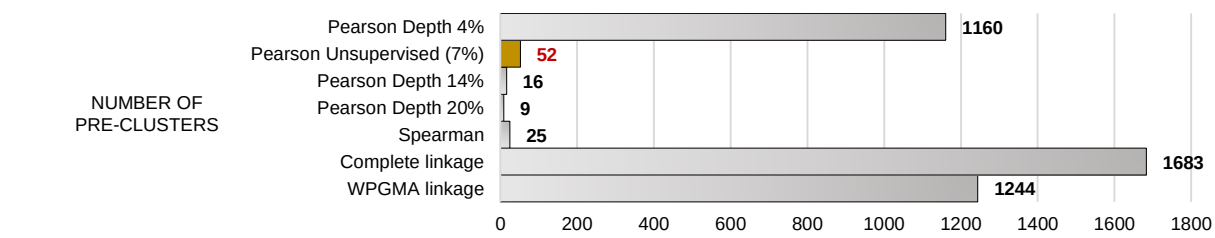
D PRE-CLUSTERING (*SPEARMAN*, UNSUPERVISED, 10%)

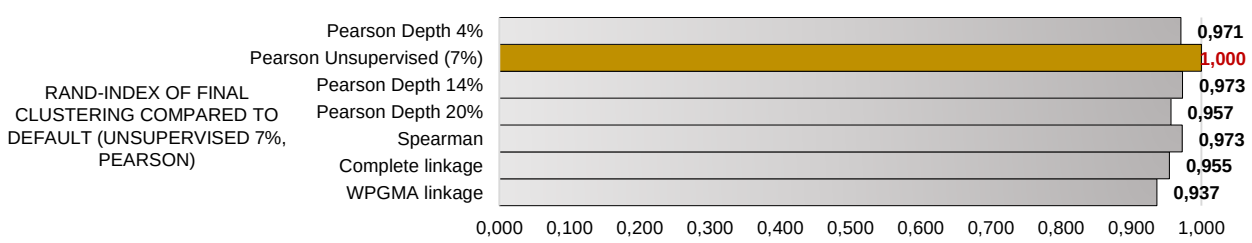


NUMERICAL MODEL: DELTA COMPARED TO 7%

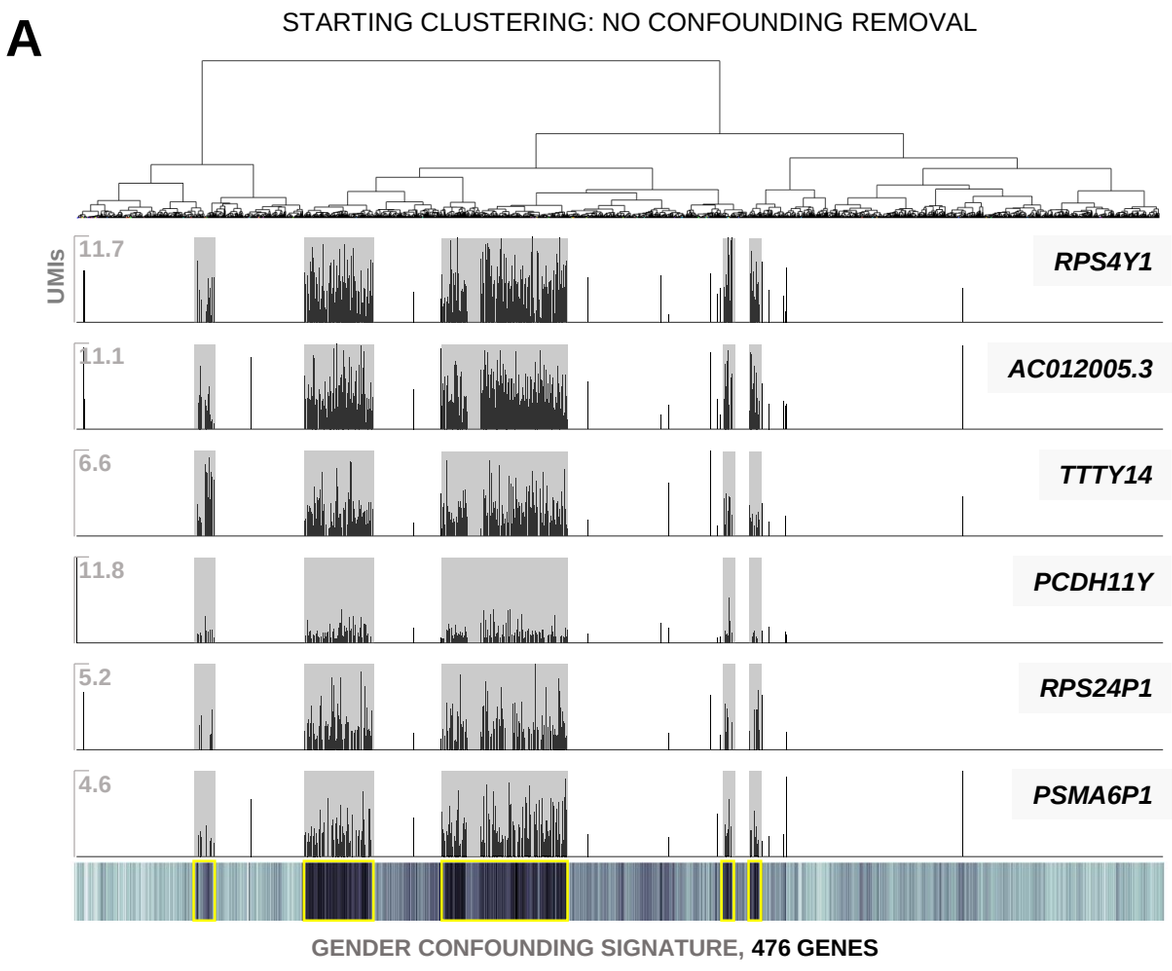


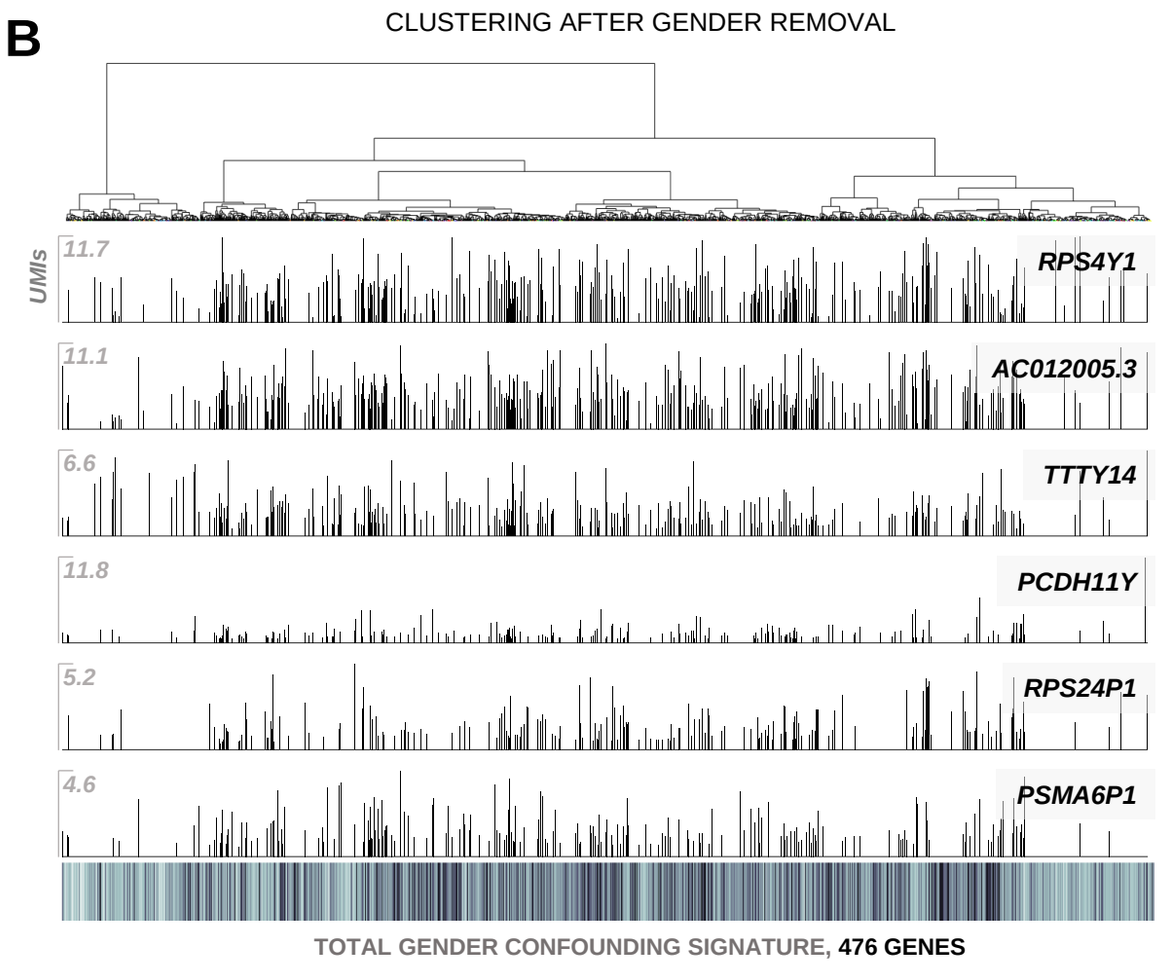
E OVERVIEWIEW OF PRE-CLUSTERING DATA AND RAND-INDEX OF FINAL CLUSTERING

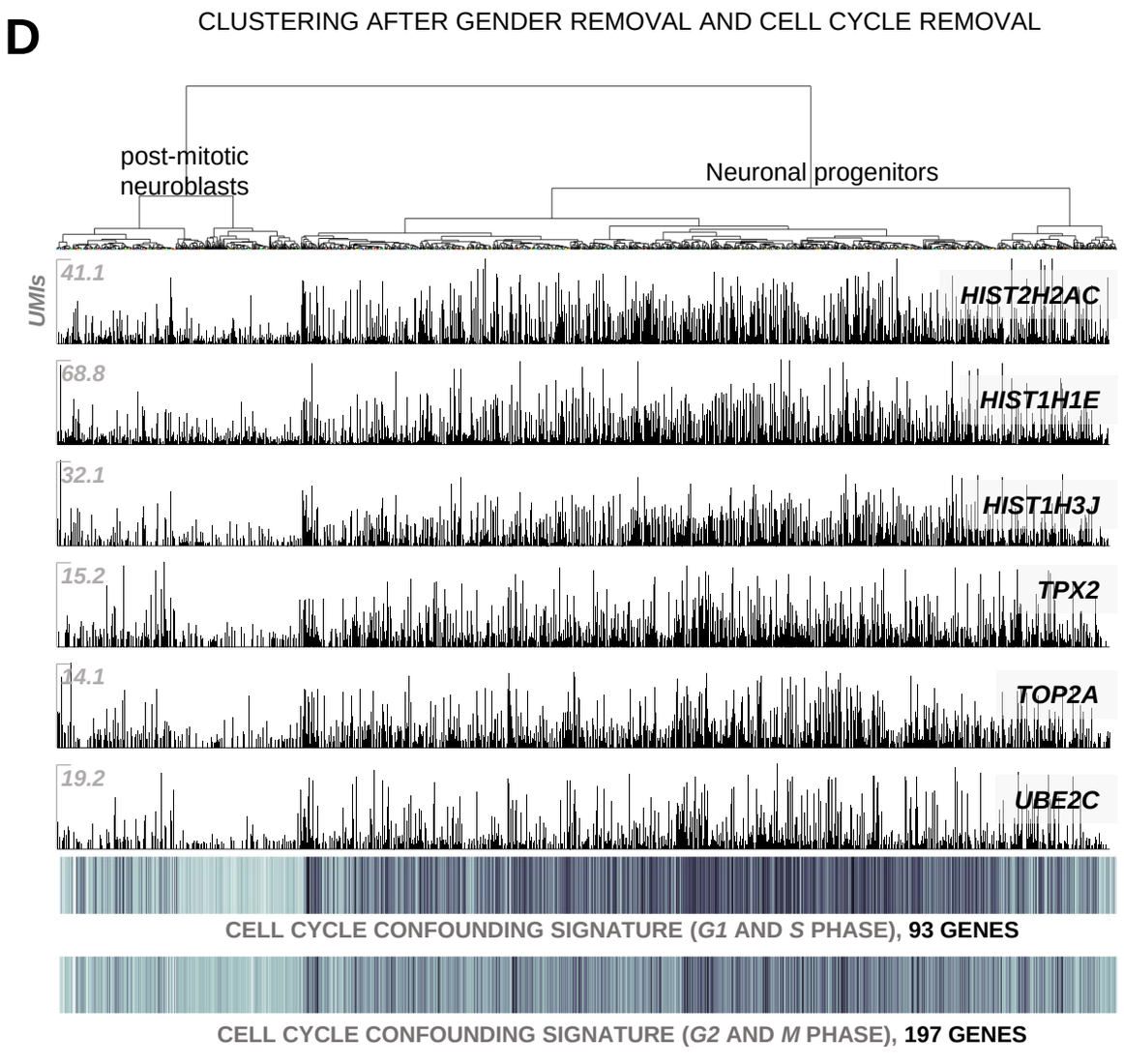




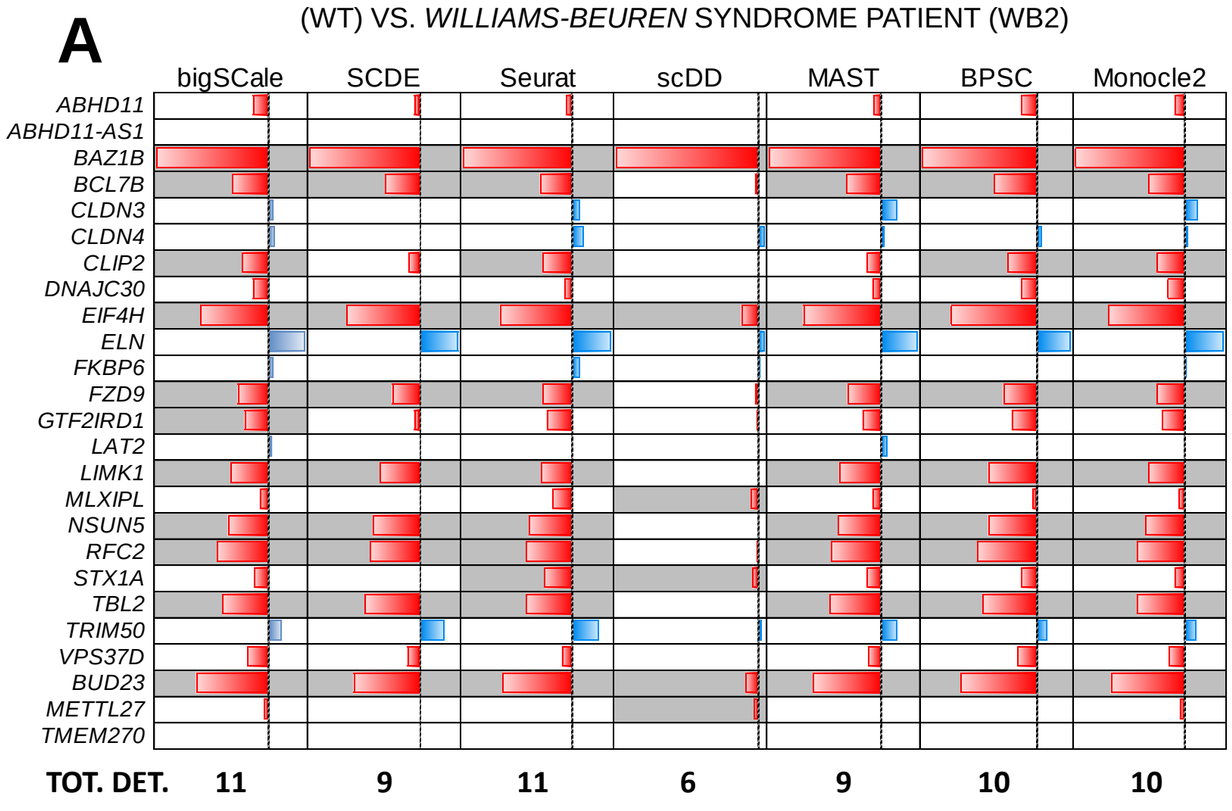
Supplemental figure 2

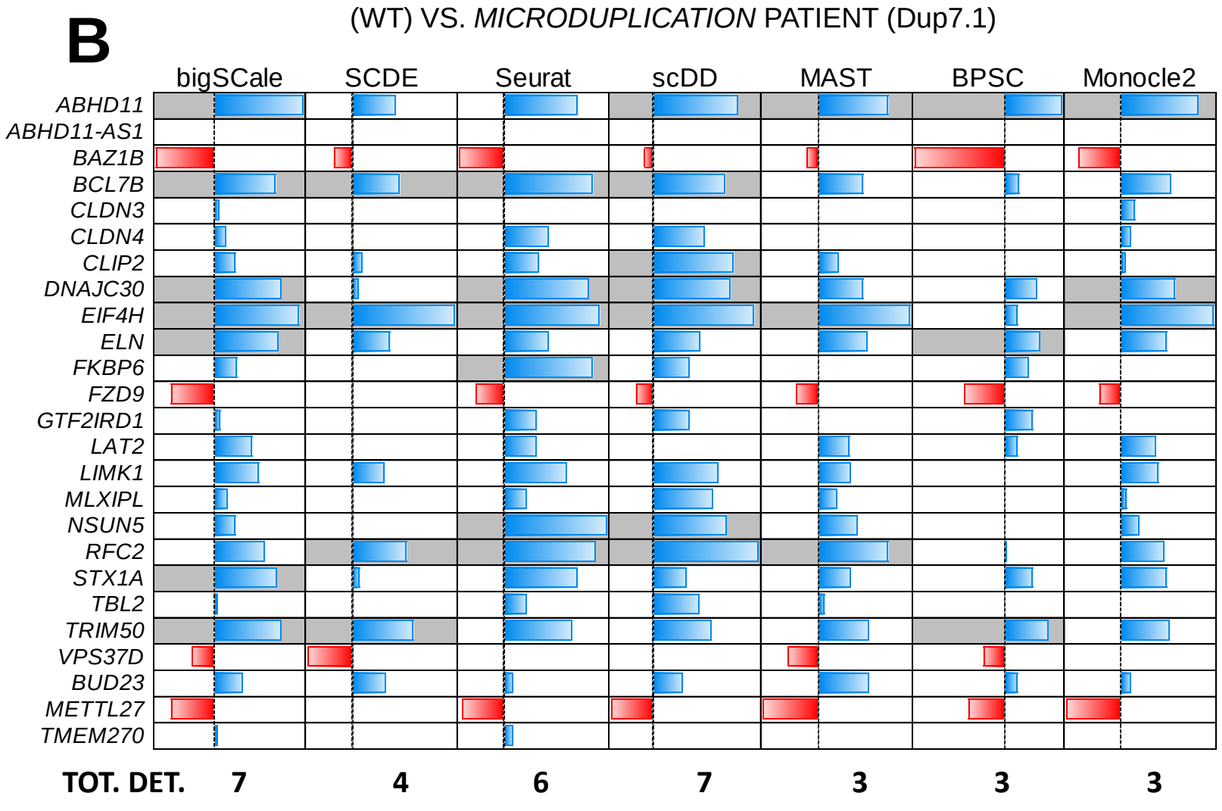


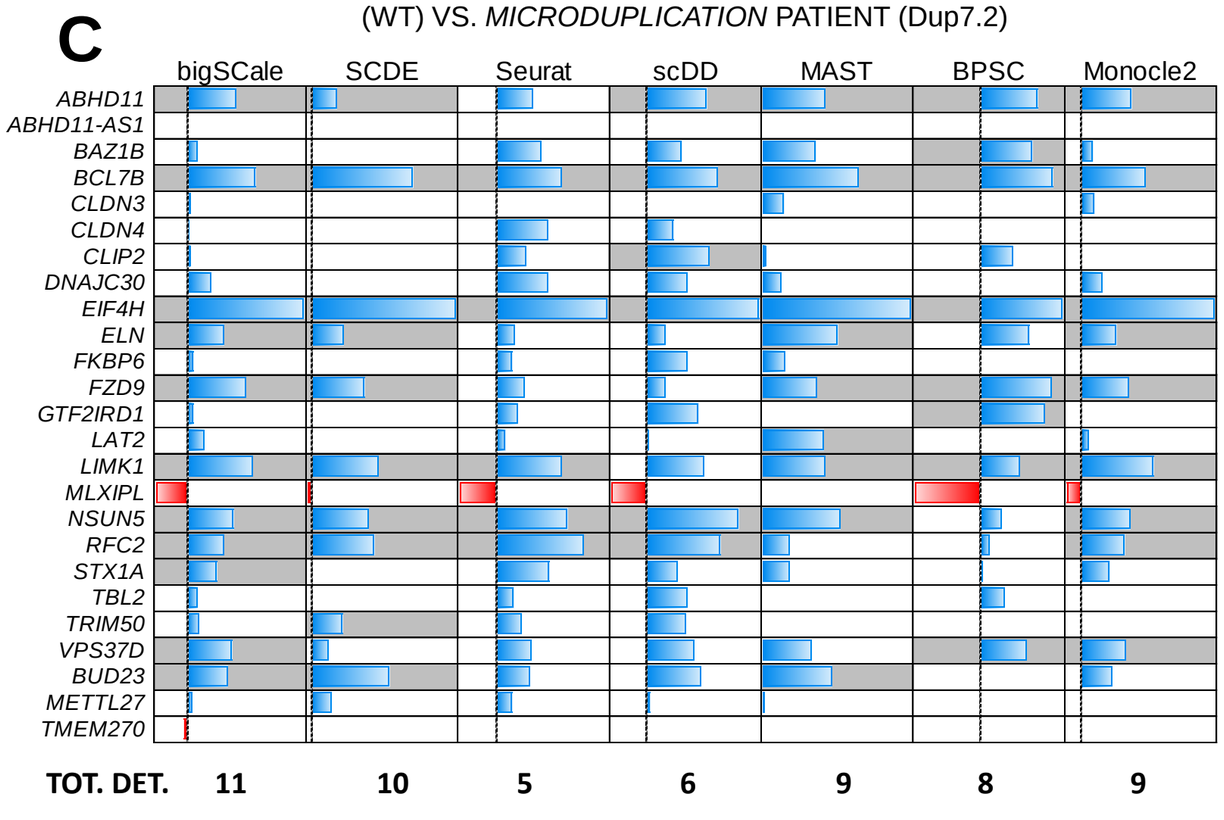


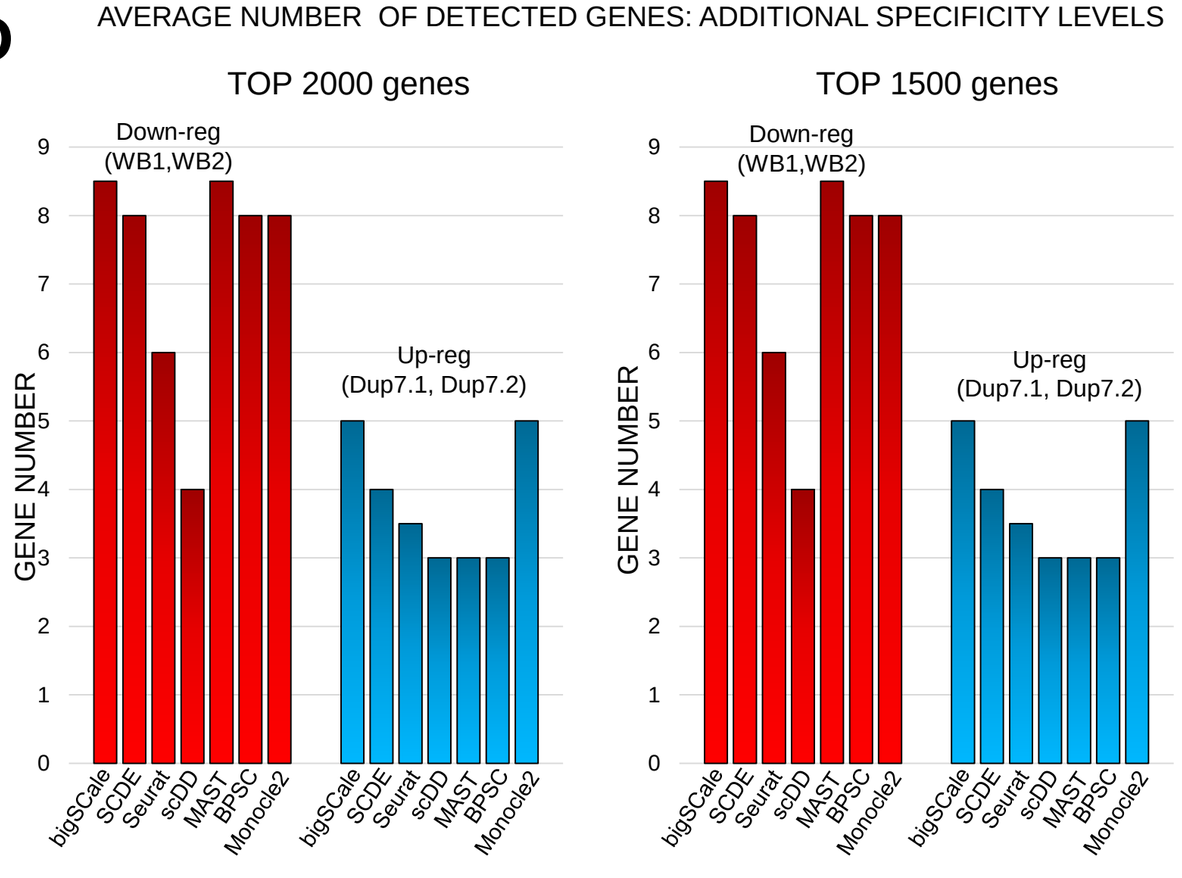


Supplemental figure 3





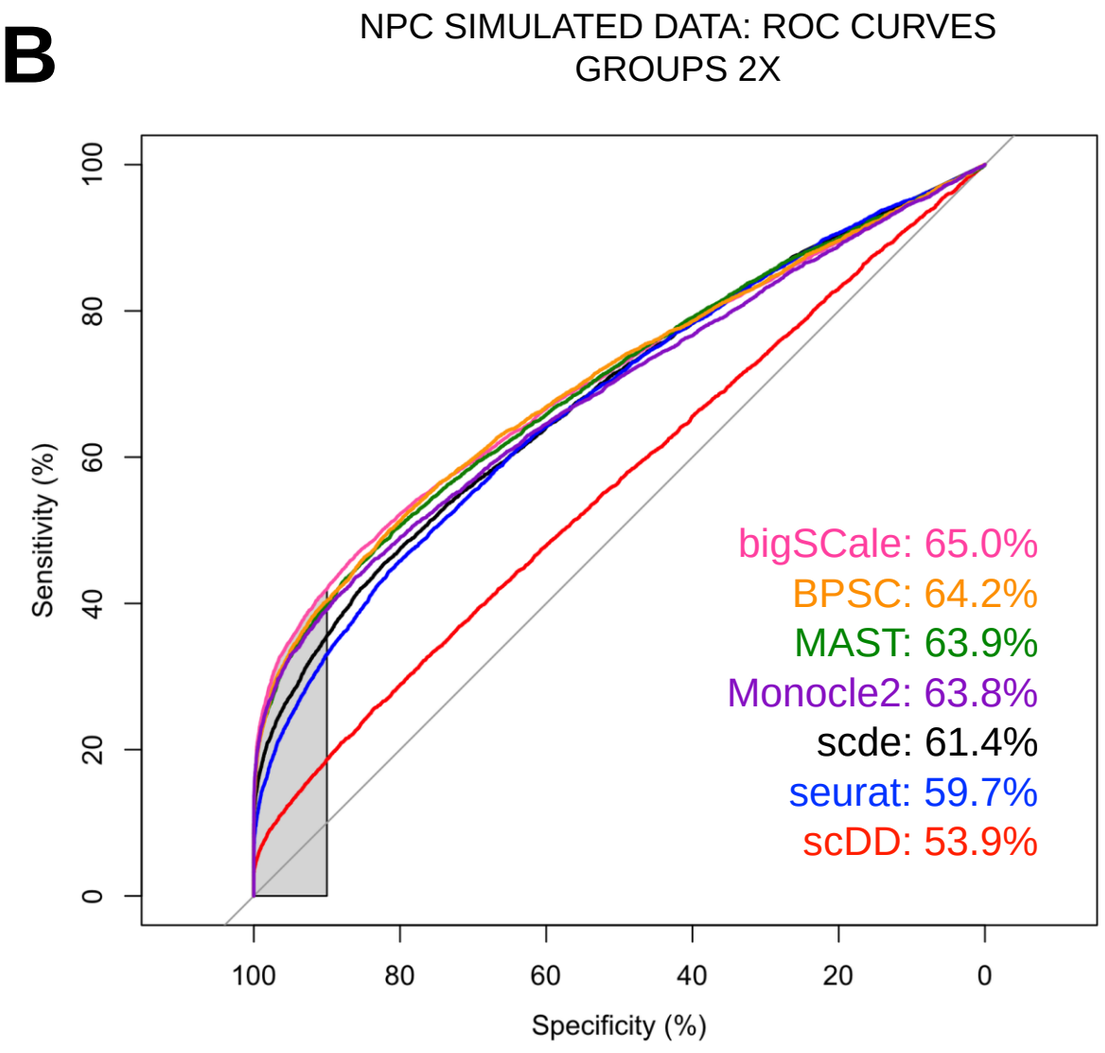




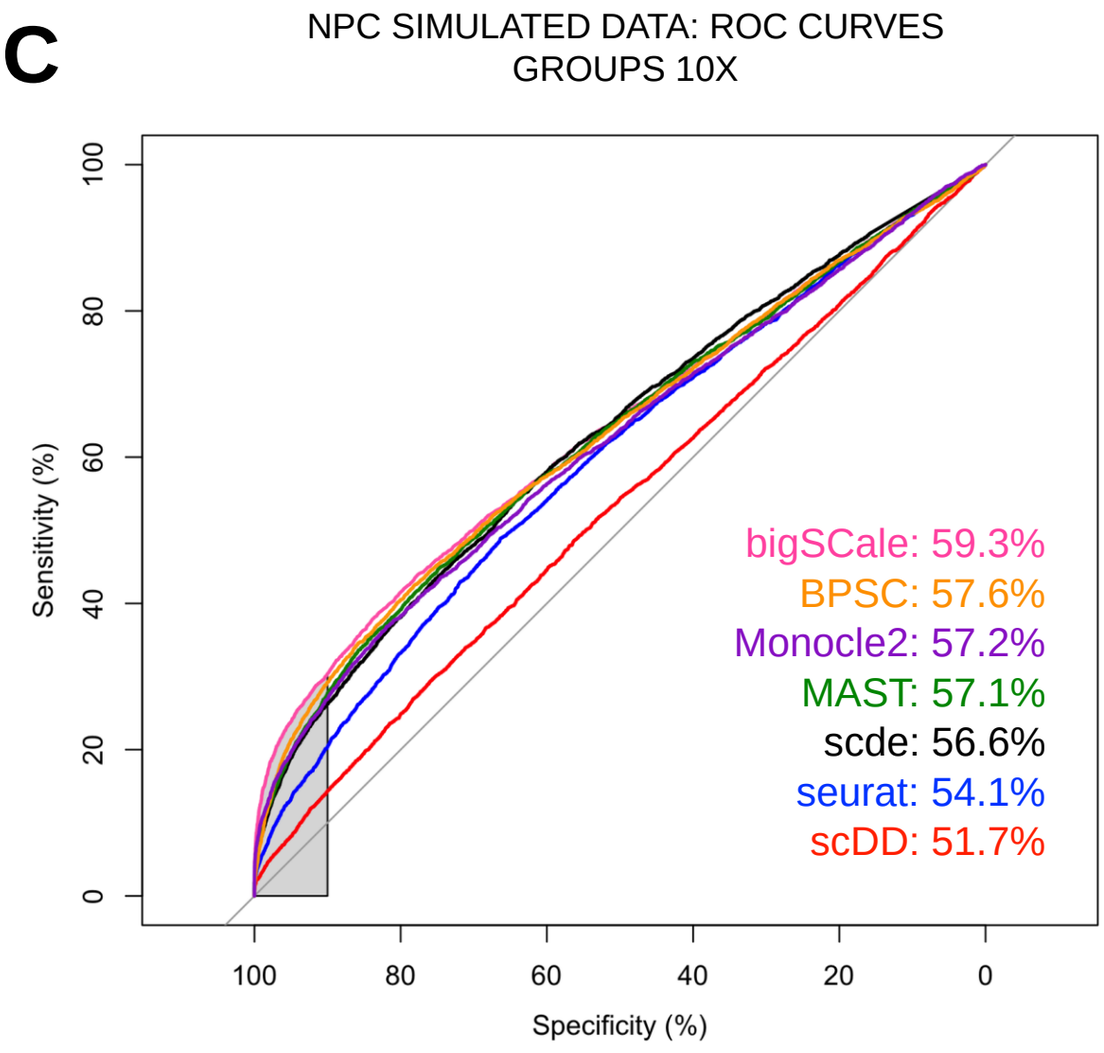
Supplemental figure 4

A

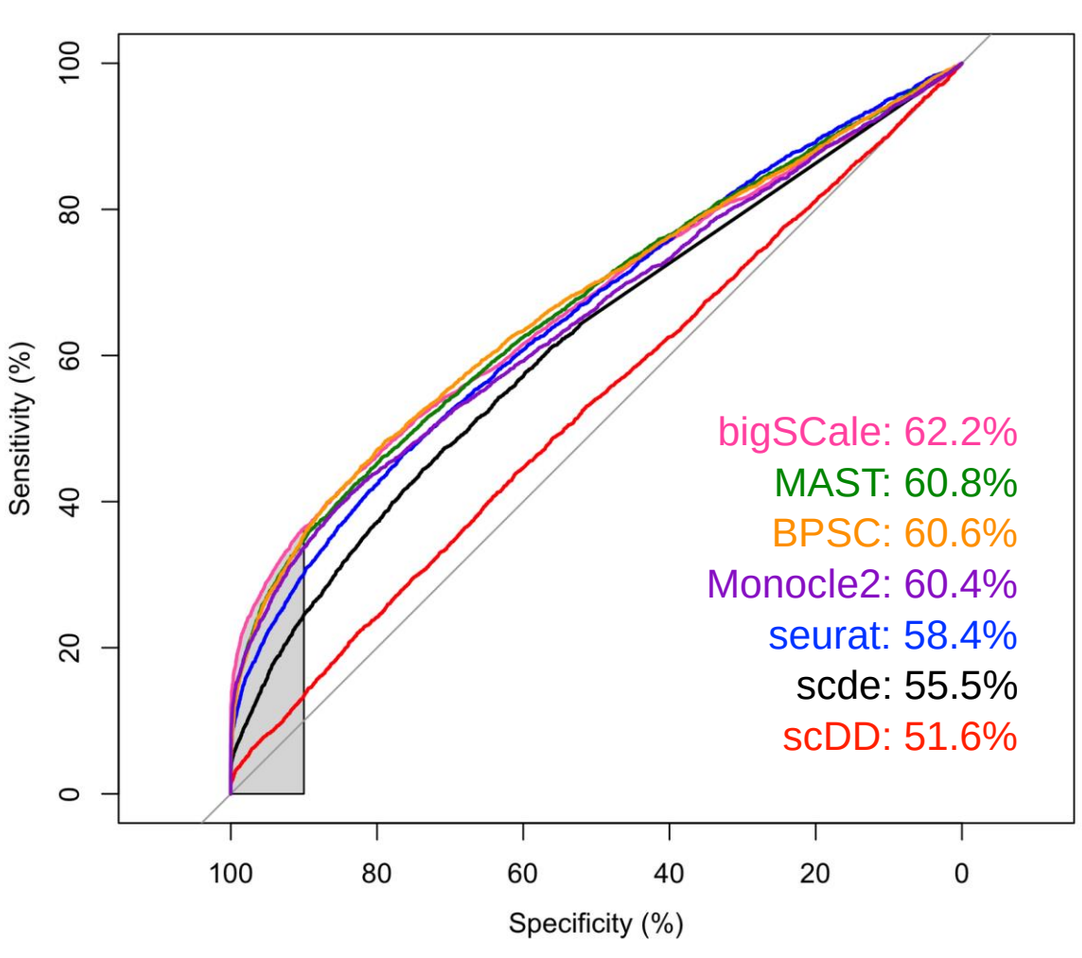
B



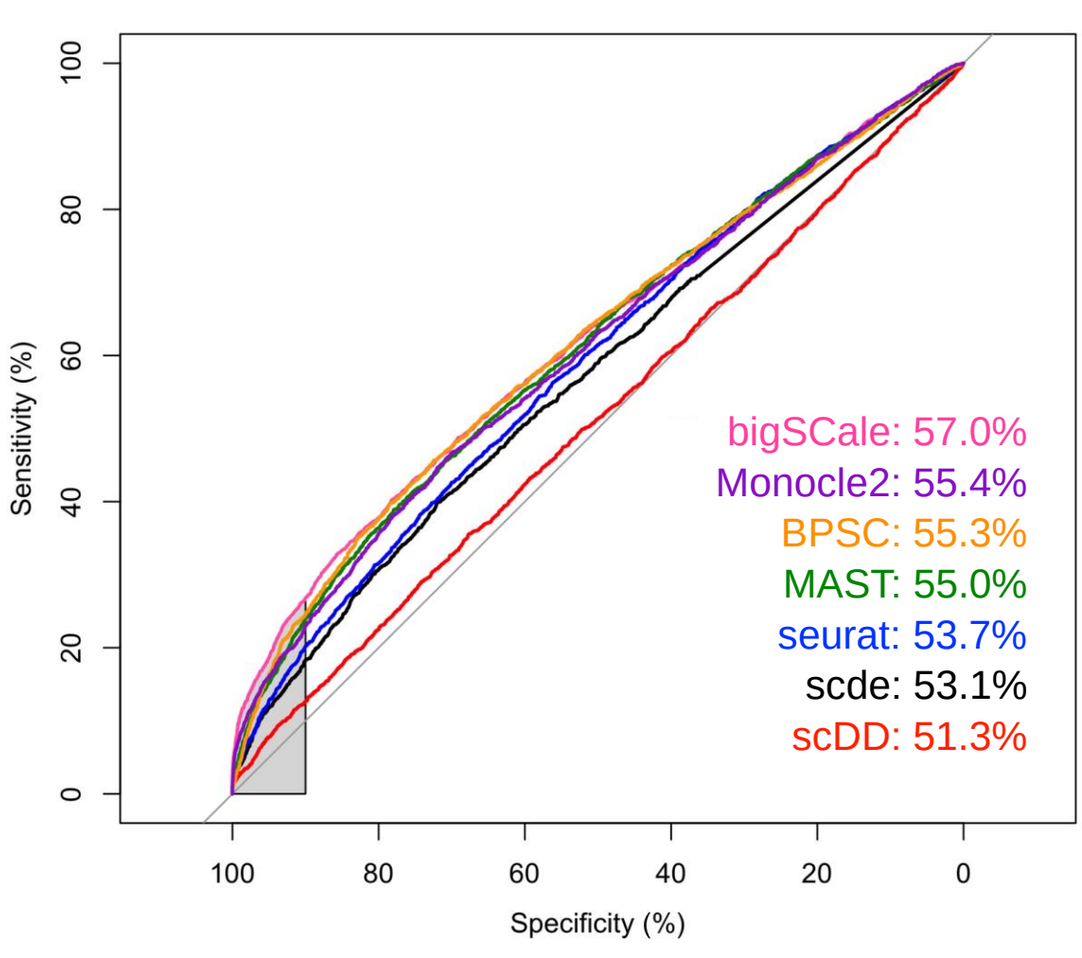
C



D

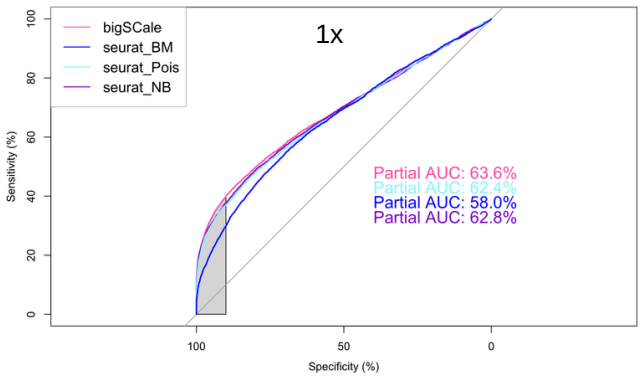


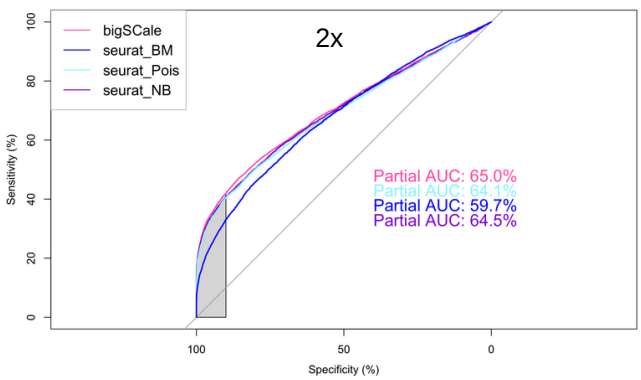
E

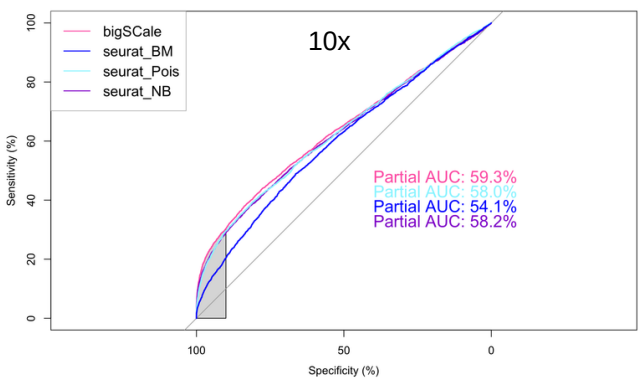


Supplemental figure 5

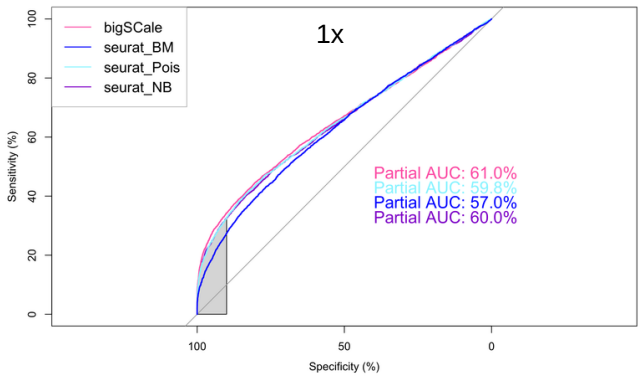
NPC

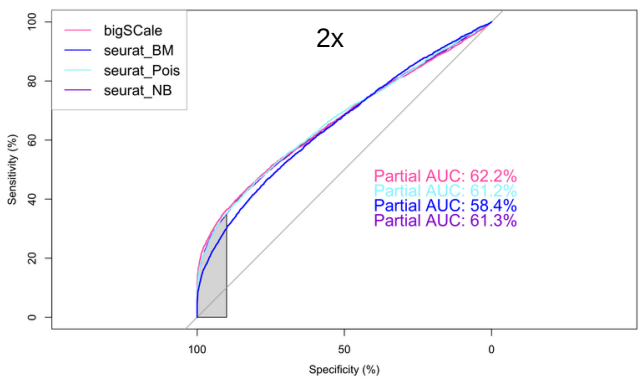


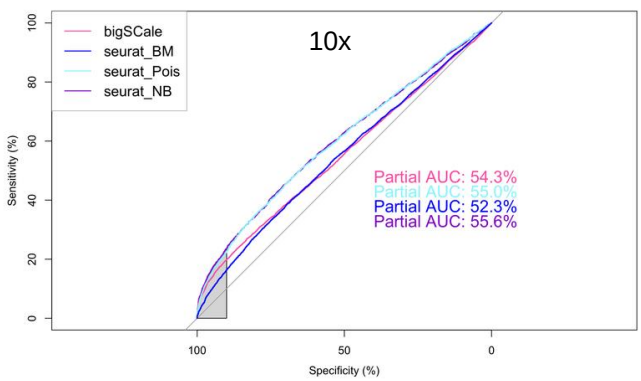




10x Genomics

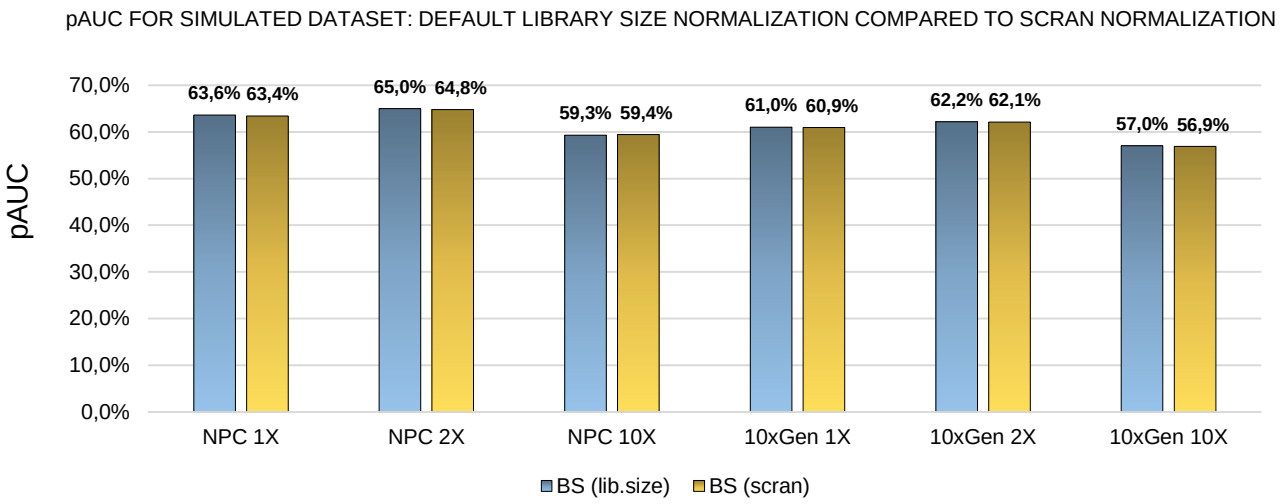




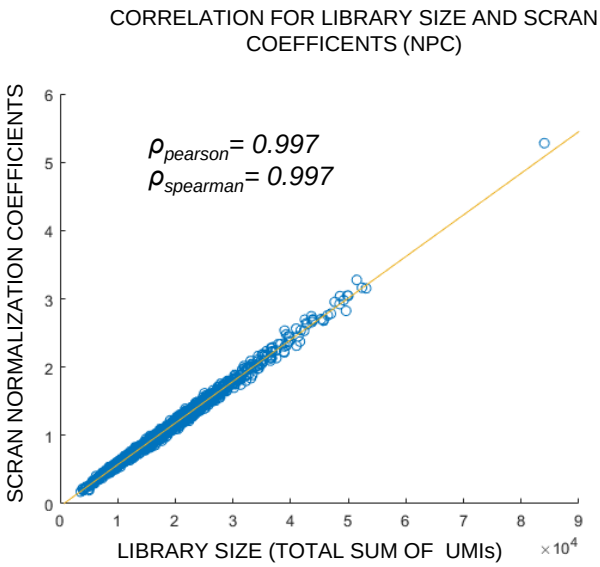


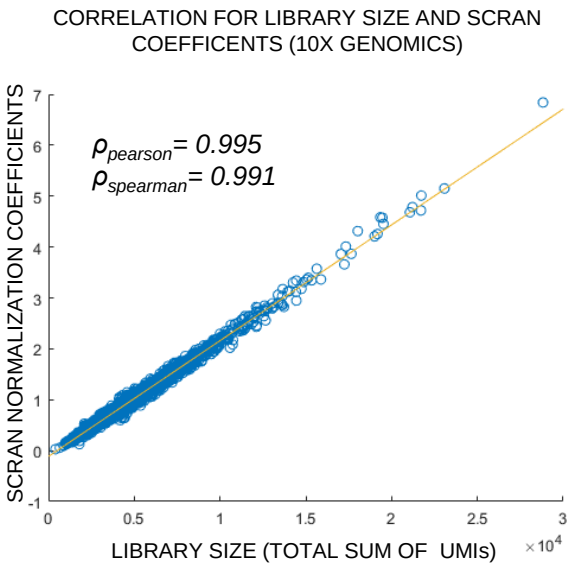
Supplemental figure 6

A



B

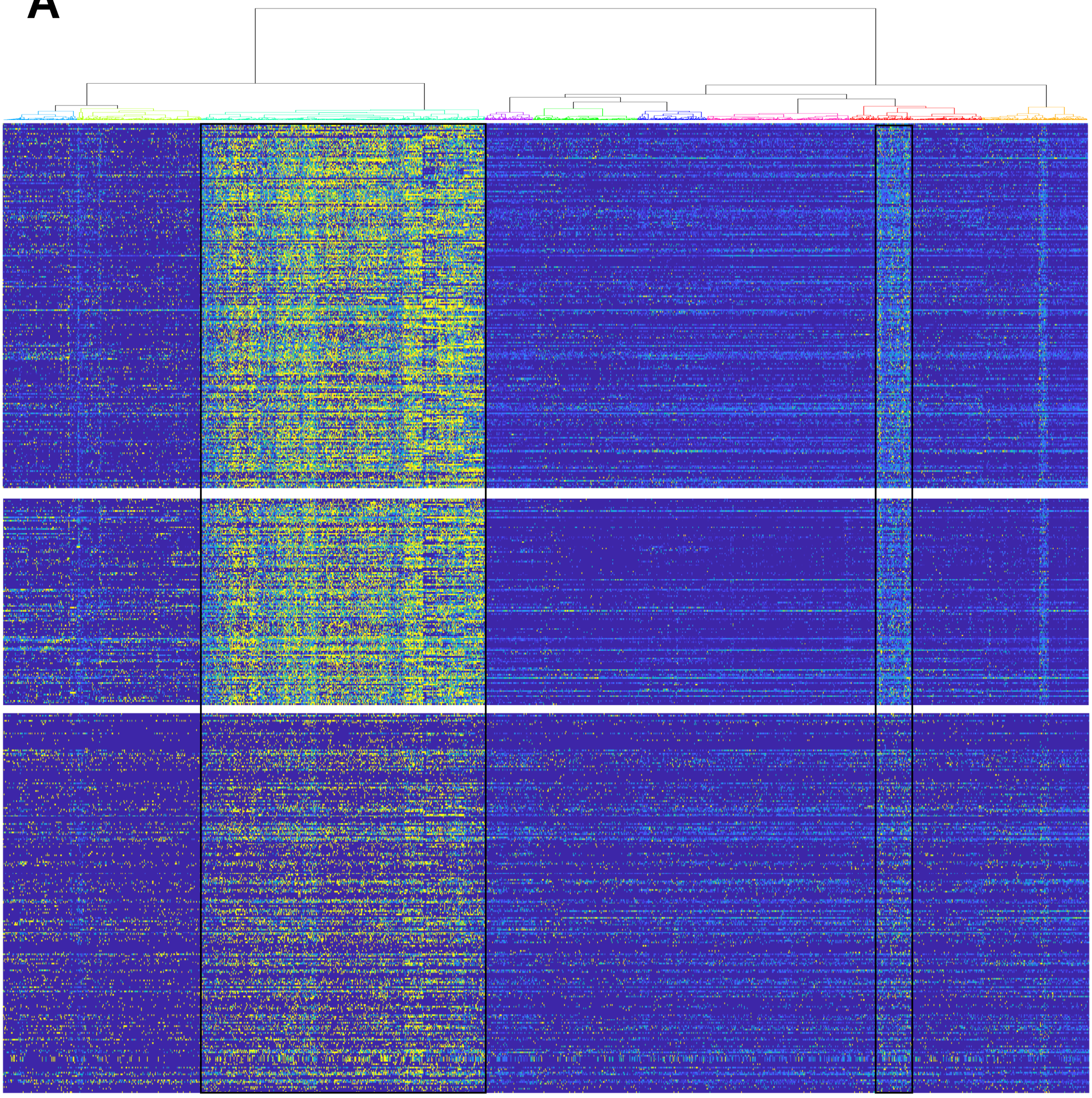




Supplemental figure 7

A

MARKERS OF OLIGODENDROCYTES: *bigSCale* VS. *BackSPIN*

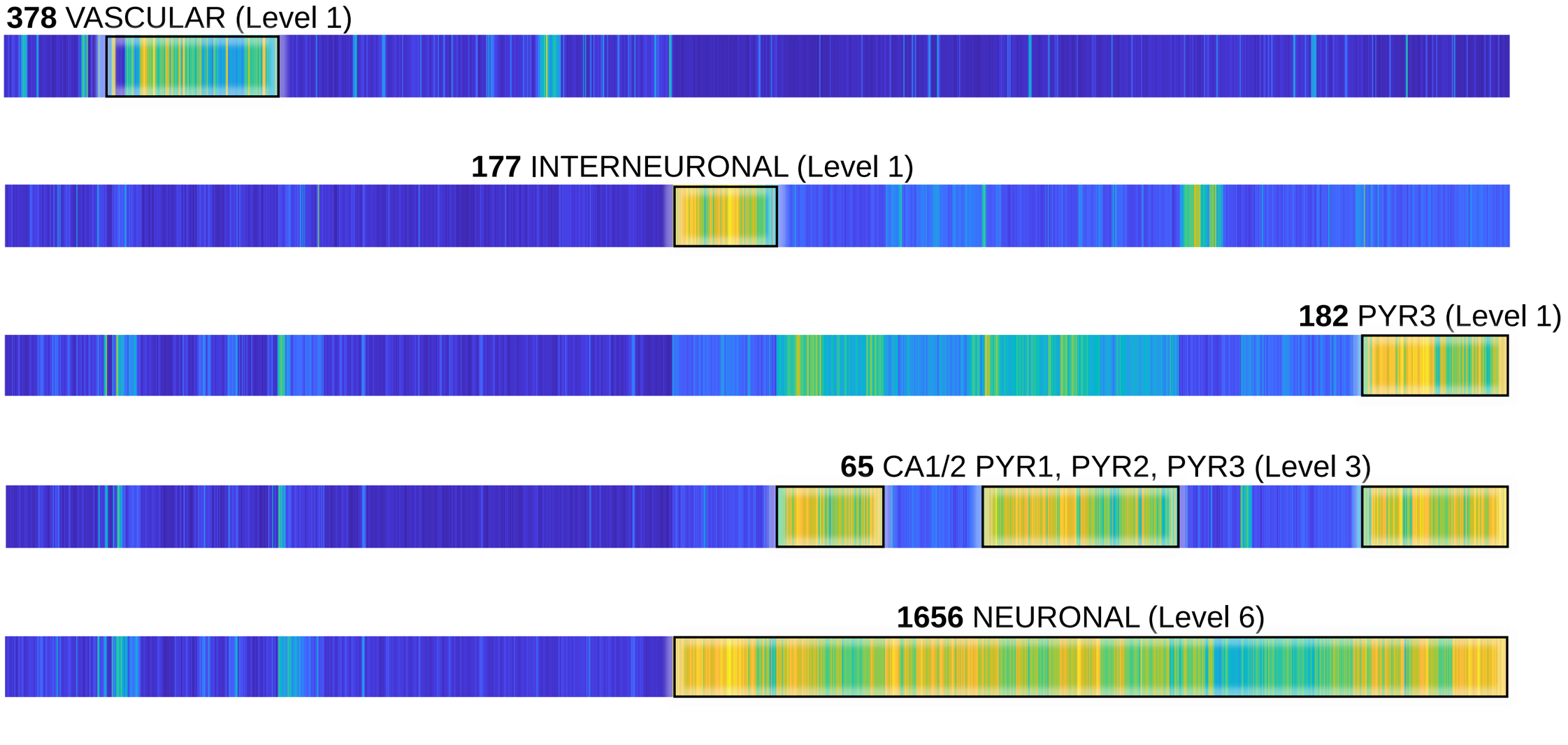


OLIGODENDROCYTES

DOUBLETS
(neurons+oligodendrocytes)

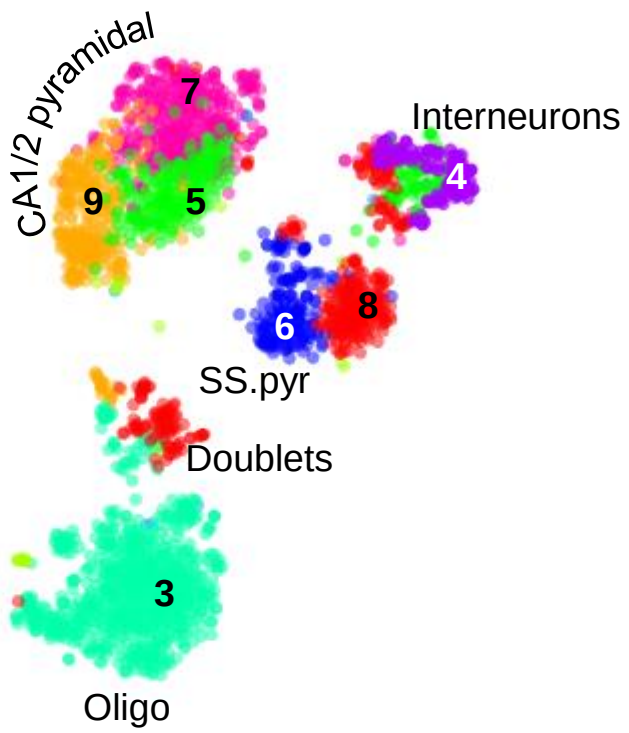
B

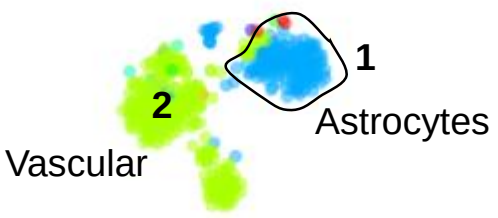
HIERARCHICAL MARKERS CALCULATED BY *bigSCale*: EXAMPLES



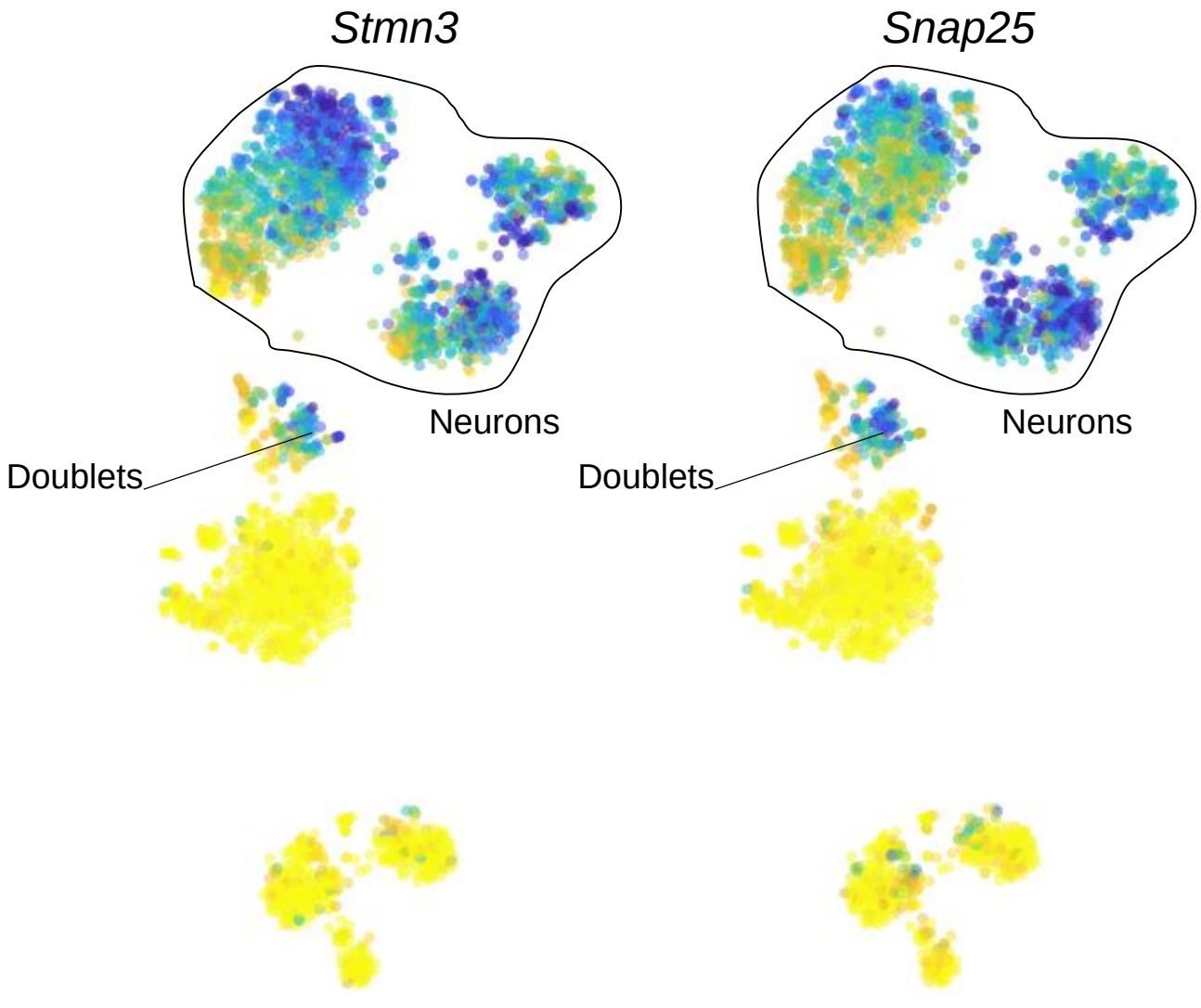
Supplemental figure 8

A tSNE OF 3,005 CELLS
ADULT BRAIN, 9 CLUSTERS

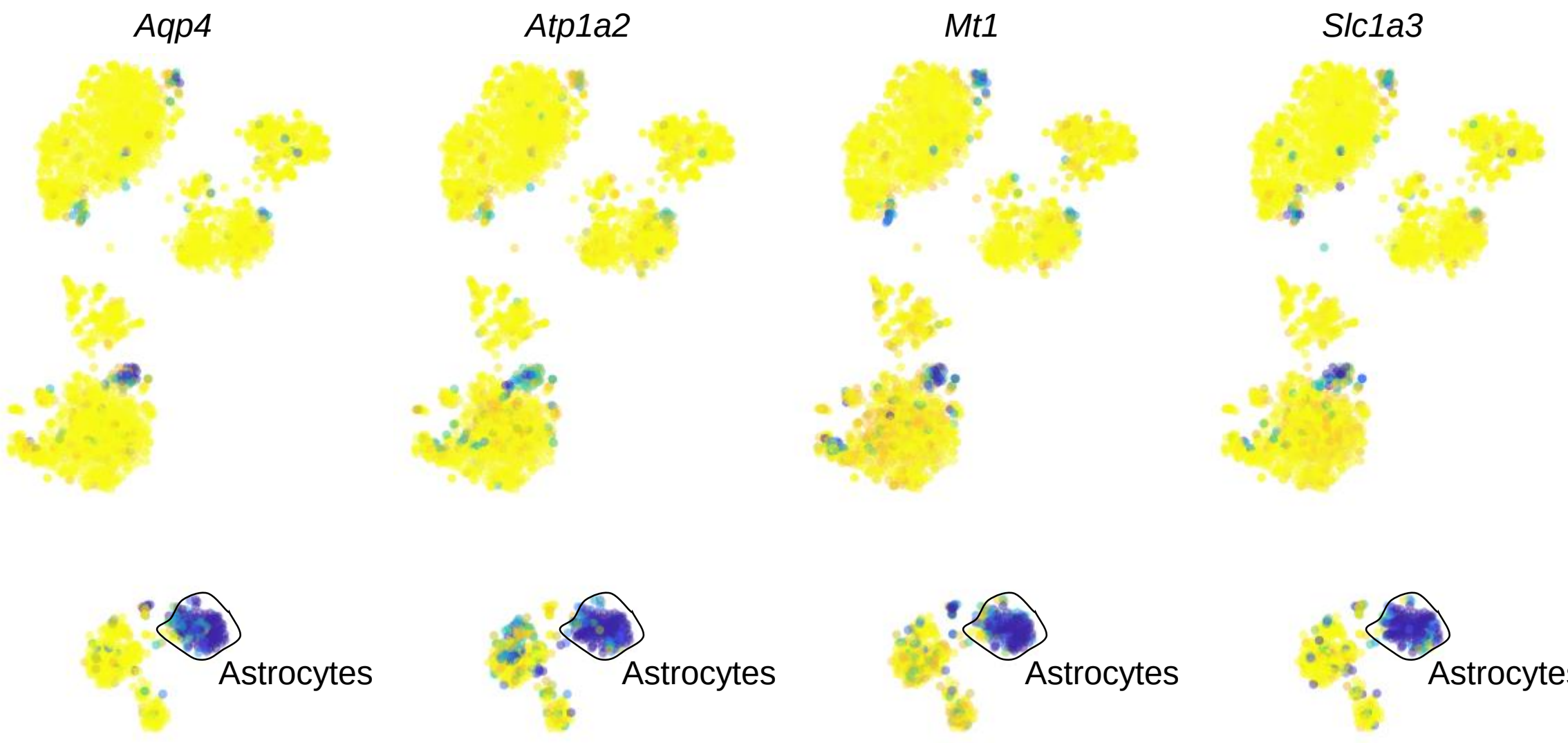




B ESTABLISHED MARKERS FOR NEURONS
PREVIOUSLY UNIDENTIFIED BY *BackSPIN*



C ESTABLISHED MARKERS FOR ASTROCYTES PREVIOUSLY UNIDENTIFIED BY *BackSPIN*



Supplemental figure 9

RELATIVE EXPRESSION (Z-SCORES) OF NEUROTRANSMITTER RECEPTORS IN THE SUBTYPES OF CAJAL-RETZIUS NEURONS

NMDA receptor

	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
Grin1	-11,9	-3,2	-0,7	2,7	4,6	-6,0	4,6	9,9
Grin2a	0,9	-0,3	1,1	0,0	-0,3	-0,2	-0,7	-0,4
Grin2b	2,1	0,2	16,4	-8,0	-5,4	-1,3	-5,1	1,0
Grin2c	0,0	0,1	0,0	0,0	-0,1	0,0	0,0	0,0
Grin2d	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0
Grin3a	-2,8	2,8	-0,2	5,8	1,5	-4,0	-2,7	-0,4
Grin3b	0,0	0,1	0,0	0,0	0,0	0,1	0,0	0,0

AMPA receptor

	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
Gria1	-10,6	-1,8	4,1	9,0	1,3	-4,8	0,2	2,5
Gria2	-13,7	-12,3	46,9	-9,1	-12,6	-5,7	-4,3	10,8
Gria3	1,2	0,1	8,2	-2,2	-2,8	-1,6	-1,2	-1,7
Gria4	-7,8	1,1	0,9	3,6	-1,7	-0,9	-1,0	5,8

Glutamate Metabotropic Receptor

	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
Grm1	0,1	0,5	1,0	0,0	-0,6	0,1	-0,5	-0,6
Grm2	14,3	3,4	2,8	-3,2	-4,6	-2,3	-4,4	-6,1
Grm3	0,0	1,0	-0,1	-0,6	-0,1	0,3	-0,1	-0,5
Grm4	0,2	0,1	1,9	-0,3	0,2	-1,0	-0,2	-0,9
Grm5	-1,8	-0,6	3,9	-0,8	-0,4	-0,1	-0,4	0,1
Grm6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Grm7	-1,9	-1,5	1,2	-0,5	1,5	-1,1	0,7	1,5
Grm8	0,0	-0,2	0,1	0,0	0,2	0,3	-0,2	-0,1

GABA(A) Receptor

	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
Gabra1	-2,6	1,7	1,2	1,6	0,9	-2,4	1,2	-1,6
Gabra2	0,6	4,7	12,3	-2,5	0,7	-5,7	-3,9	-6,3
Gabra3	-1,0	2,0	0,4	-0,1	1,7	-1,0	-0,3	-1,8
Gabra4	-0,8	1,2	2,2	-0,6	-0,2	-0,5	-0,2	-1,2
Gabra5	-0,4	0,0	2,3	0,1	-0,6	-0,3	-0,3	-0,7
Gabra6	-1,1	0,5	-0,2	0,0	0,1	-0,1	0,6	0,1
Gabrb1	-2,5	2,8	2,9	-0,7	-0,1	-1,9	-0,3	-0,3
Gabrb2	-2,8	1,1	-0,8	-1,4	-0,8	0,0	1,6	3,1
Gabrb3	1,4	2,9	6,6	4,0	-0,5	-6,5	-3,8	-4,0
Gabrd	-0,2	0,0	-0,1	0,2	0,3	-0,1	0,0	-0,1
Gabre	0,1	0,0	0,0	0,0	-0,1	0,0	-0,1	0,0
Gabrg1	1,0	1,1	-0,1	-0,2	0,3	-1,0	-0,4	-0,7
Gabrg2	-8,2	2,2	0,6	-0,1	8,9	-3,0	2,9	-3,4
Gabrg3	3,6	2,6	0,0	0,3	-1,2	-2,8	-1,1	-1,5

GABA(B) Receptor

	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
Gabbr1	-4,0	0,1	1,9	4,8	3,7	-4,5	0,5	-2,4
Gabbr2	1,3	-1,7	4,5	2,9	0,5	-2,5	-2,8	-2,1

Glycine receptor

	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
Glra1	-0,8	-0,8	0,3	-0,5	0,2	-0,3	0,2	1,7
Glra2	1,2	-0,2	5,5	-1,6	-0,7	-0,2	-1,5	-2,6
Glra3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Glra4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Glrb	-7,6	-0,1	-0,6	9,6	7,0	-5,4	0,5	-3,4

5-HT serotonin receptor ionotropic

	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
Htr3a	-4,1	-2,0	0,6	3,9	5,7	-1,4	0,4	-3,0
Htr3b	-0,6	-0,5	0,0	0,5	0,7	-0,3	0,4	-0,2

5-HT serotonin receptor metebotropic

	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
Htr1a	-0,1	-0,1	0,4	0,0	0,0	-0,1	0,0	-0,1
Htr1b	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0
Htr1d	0,1	0,0	0,0	0,1	0,1	0,0	0,0	0,0
Htr1f	-5,4	-2,8	-1,4	1,4	-0,1	-2,6	4,4	6,6
Htr2a	0,0	0,1	0,1	0,0	0,0	-0,1	0,0	-0,2
Htr2b	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Htr2c	0,2	0,4	-0,2	-0,3	-0,3	-0,1	-0,3	0,6
Htr4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Htr5a	0,0	0,0	0,1	-0,1	0,0	0,0	0,1	0,0
Htr6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Htr7	0,0	0,0	0,2	0,0	0,1	0,1	-0,1	-0,1

Adrenoceptors

	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
Adra1a	0,0	0,2	-0,1	0,0	-0,1	0,0	0,0	0,1
Adra1b	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0
Adra1d	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Adra2a	0,5	3,2	3,2	4,7	-0,2	-2,7	-5,3	-3,3
Adra2b	0,0	0,2	0,0	0,0	0,0	0,0	0,0	0,0
Adra2c	0,3	-0,9	2,0	2,8	1,4	-2,7	-0,8	-2,1
Adrb1	1,5	-0,4	1,9	1,7	-0,2	-1,4	-1,9	-1,4
Adrb2	0,1	0,2	0,1	0,1	-0,2	-0,1	-0,1	-0,1
Adrb3	0,1	0,0	0,0	0,1	0,0	0,0	0,0	0,0

Supplemental figure 10