Supplemental Figures

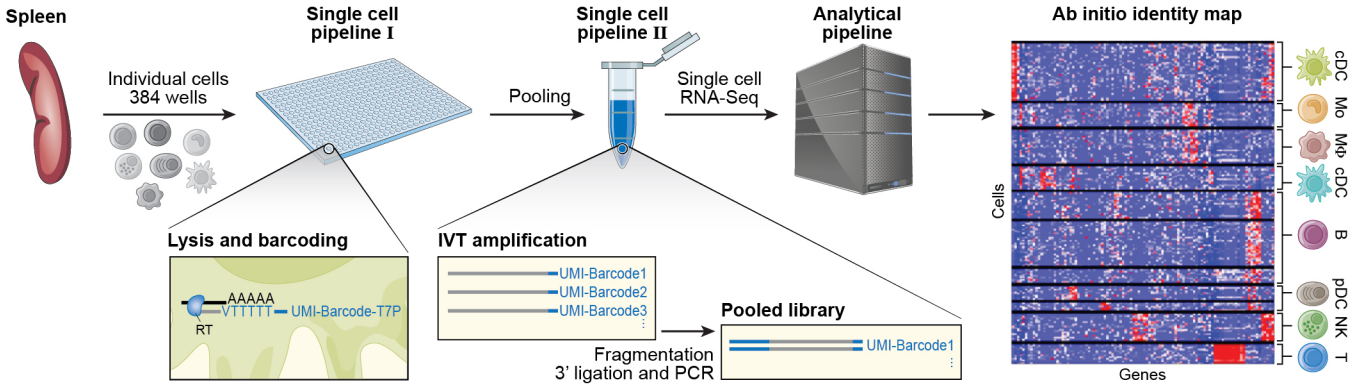

Figure S1. **Massively parallel RNA single-cell sequencing framework (MARS-Seq)**. Schematic diagram of the massively parallel approach to single-cell RNA-seq, involving the use of randomized molecular tags to initially label poly-A tailed RNA molecules, followed by pooling labeled samples and performing two rounds of amplification, generating sequencing ready material (see fig. S2 and methods for an expanded version).

Supplemental Figures

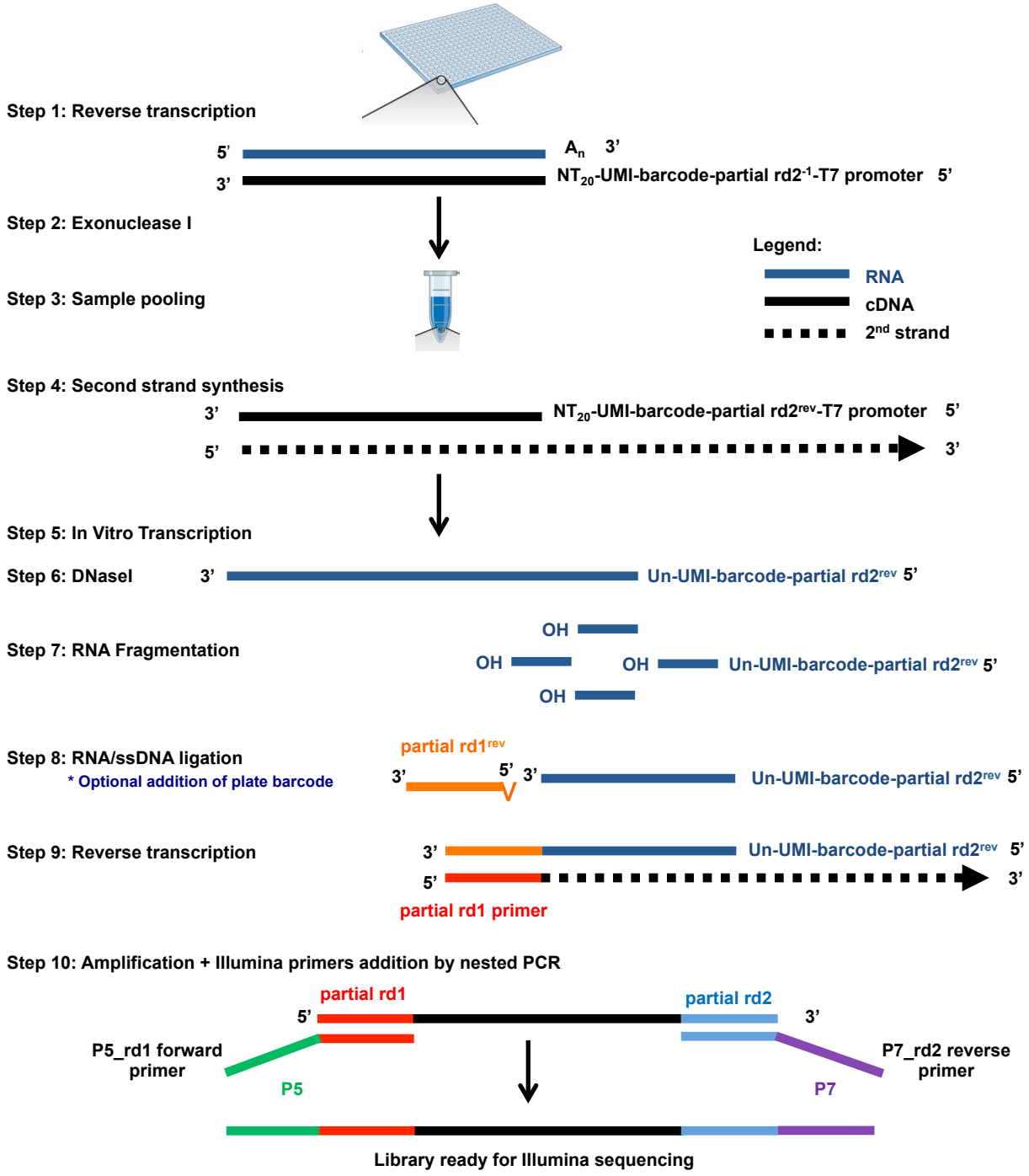

Figure S2. Experimental procedure. Schematic diagram presenting the process of converting single-cell RNA samples to sequencing-ready DNA libraries. Shown are ten experimental steps describing how RNA is tagged, pooled, amplified, fragmented, and how library construction is being performed. Colored lines represent RNA (blue) or DNA (black) molecules, or oligos and primers (see methods for a detailed description).

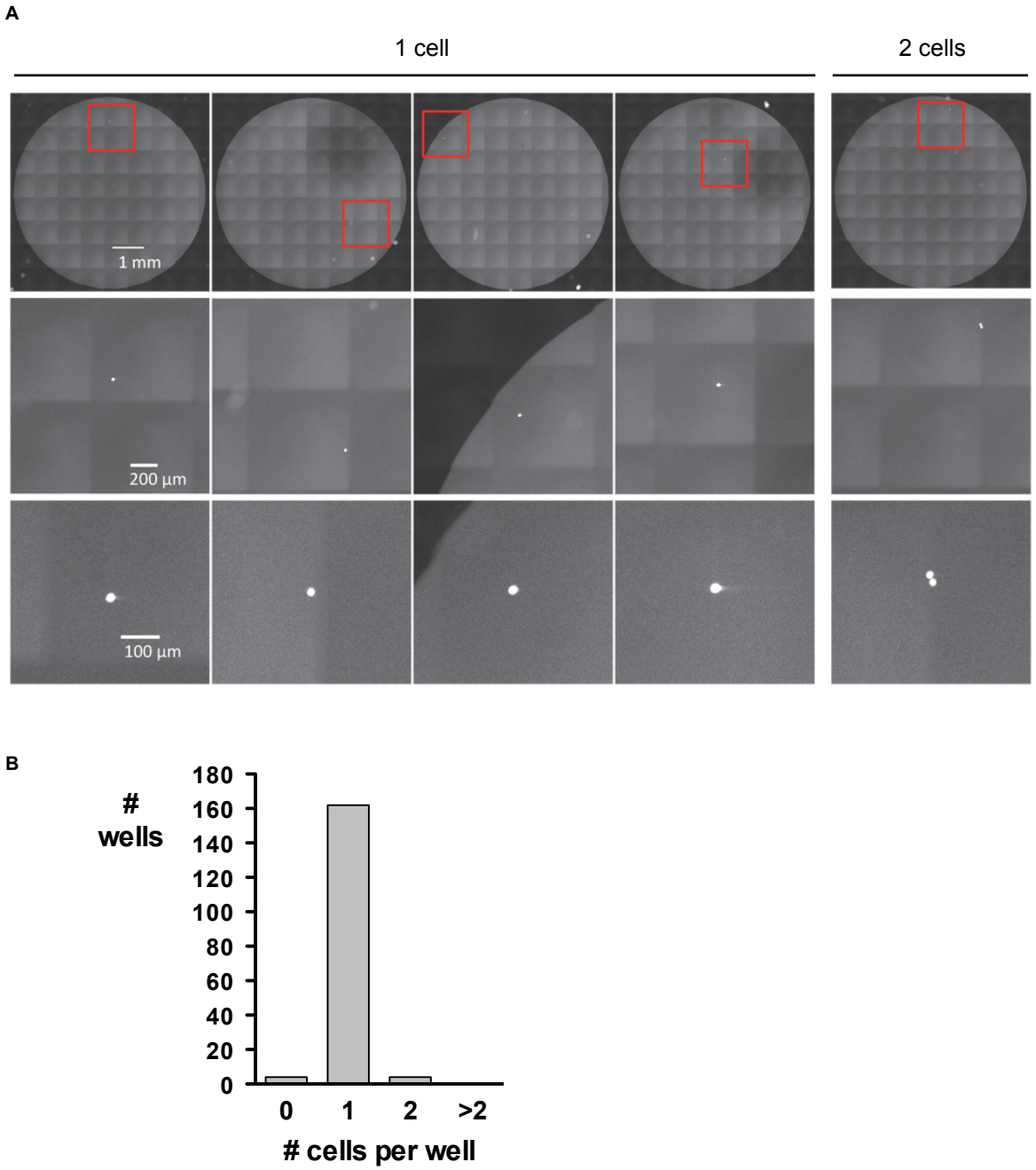

Figure S3. Single-cell sorting assessment. (A) Automated microscope scanning of CFSE-stained cells single cell-sorted into 96-well plates. Representative pictures and software-generated magnifications of wells containing one or two cells. **(B) Single-cell sorting quantification.** We sorted two 96-well plates in single-cell mode into all wells of columns 2 to 12 of each plate (column 1 was used for calibration; see Materials and Methods). The histogram shows the number of wells in which no cell (4 wells), one cell (162 wells), two cells (4 wells) or more than two cells (0) were detected (six wells were not properly scanned by the microscope and excluded from the analysis).

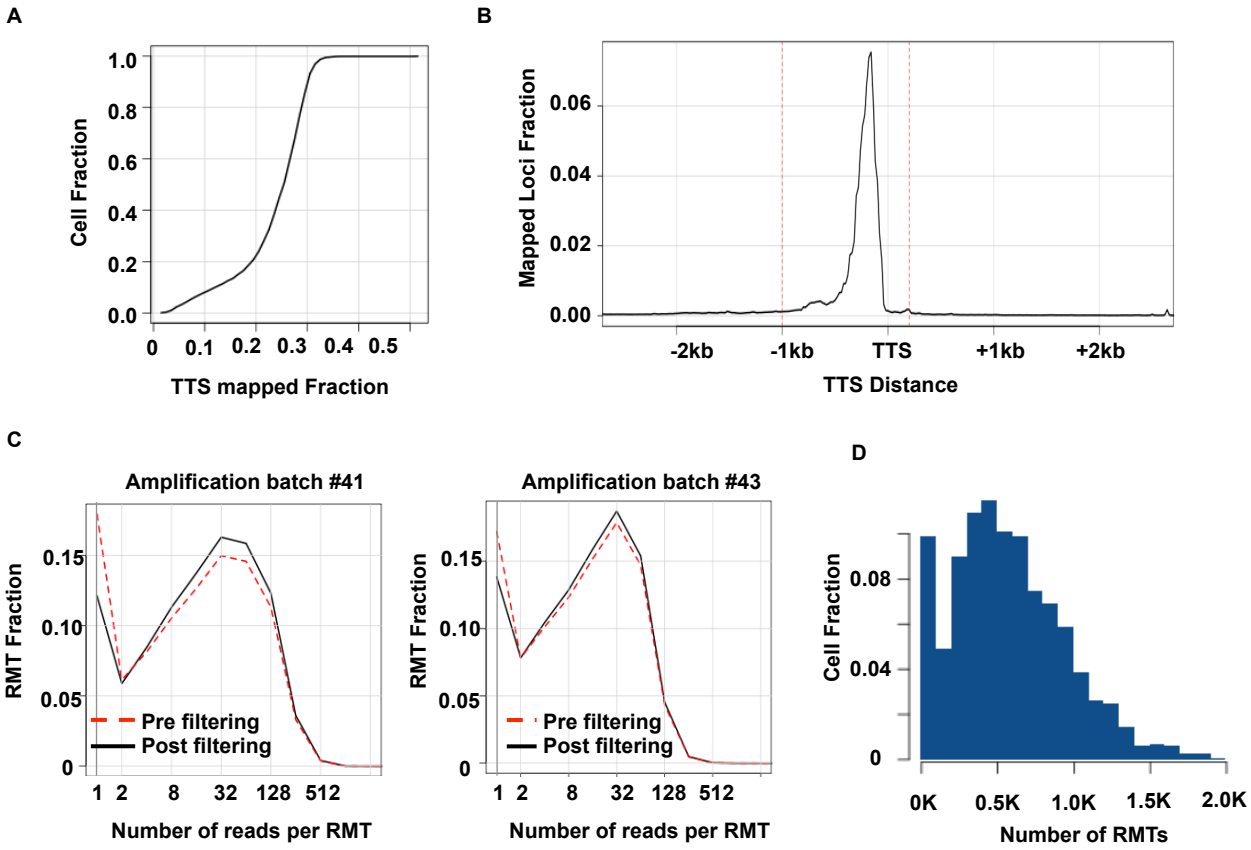

Figure S4. Sequence mapping and tag filtering. (A) TTS mapping efficiency. Cumulative distribution of the fraction of reads that were mapped to mouse TTS (x-axis, see methods) across 1528 multiplexed cells (y-axis show the cumulative cell fraction). Median mapping percentage is ~25%. **(B) Distribution of mapping loci around TTS.** Shown is the spatial distribution of mapped molecules (Cell/RMT) around TTS. Dashed red lines demarcate the (-1000, 200) range in which we consider a molecule as associated with a TTS. **(C) Read saturation.** Distribution of number of sequencing products per inferred molecule (unique and valid RMT) before (dashed red curve) and after (black solid curve) barcode and RMT filtering (see methods) for two amplification batches (i.e. batches sharing the same plate barcode). These data show that majority of the molecules were excessively sequenced. **(D) RMT yield.** Shown is a distribution of RMTs after filtering (x-axis) and the fraction of cells (y-axis).

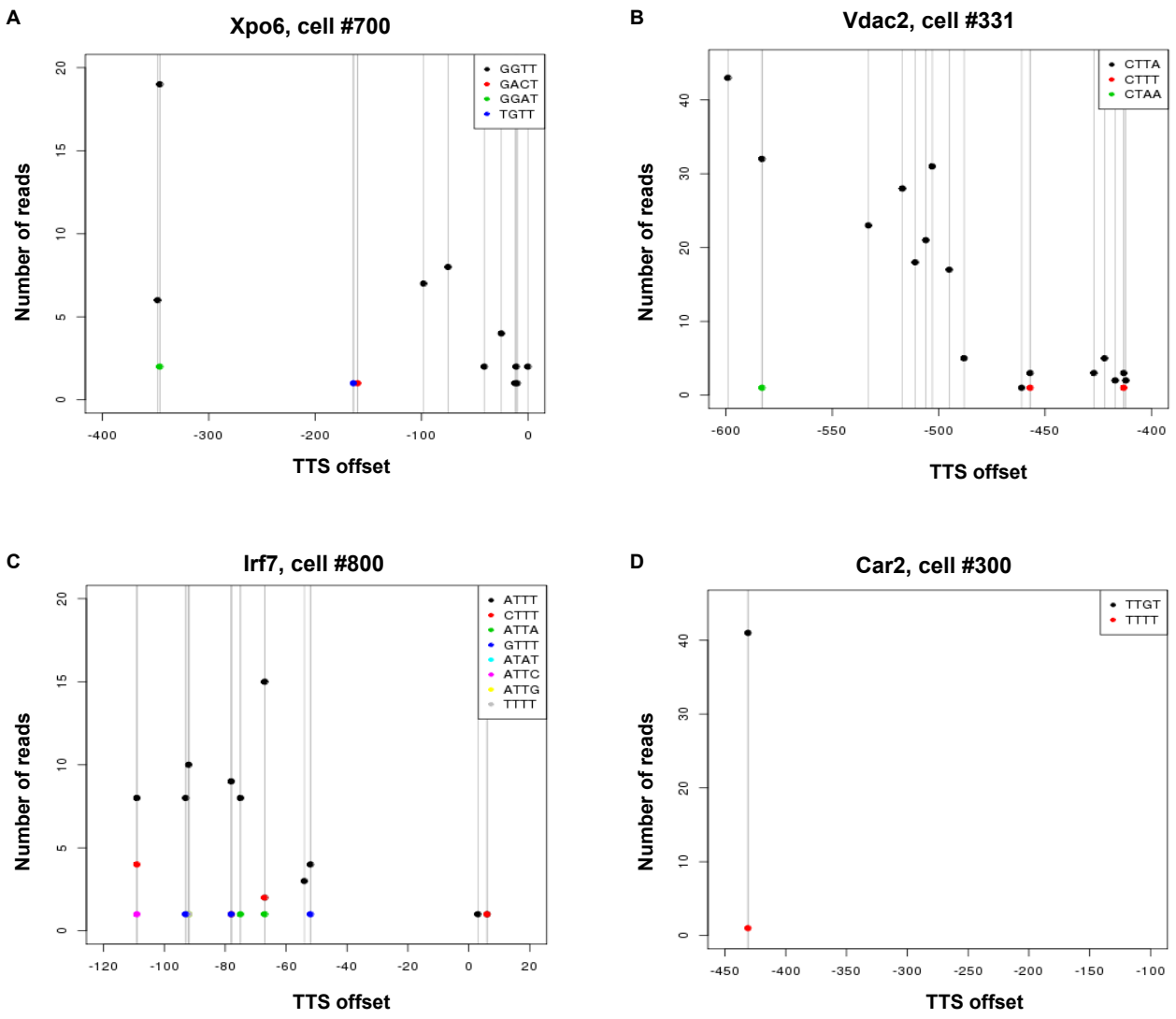

Figure S5. Filtering RMT sequencing errors. Shown are full sequencing and labeling data for 4 gene/cell examples. For each offset (x-axis) the number of reads of each color coded RMTs is shown. These profiles exemplify RMTs with multiple offsets (black dots, panels **A-C**) that undergo sequencing errors and create spurious RMT with edit distance of one (colored dots). Specifically in the Xpo6 example **(A)**, the GGTT RMT (black) is mapped to multiple positions with high read counts, but the GGAT is mapped to a single loci which is shared with GGTT. In some cases poorly mapped molecules undergo RMT sequencing errors (as shown in **(D)**). Detection and filtering of RMT sequence errors is described in the methods section.

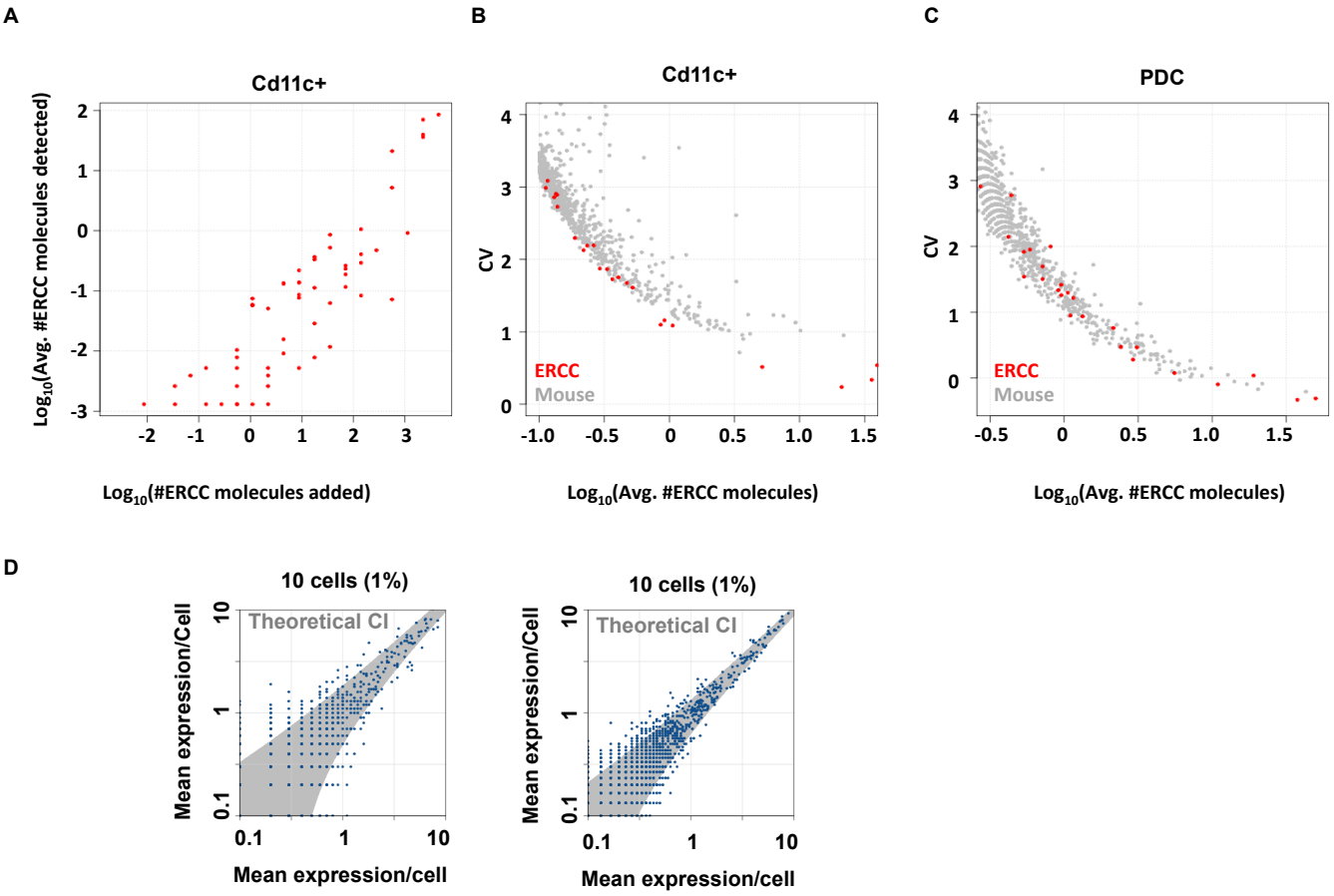

Figure S6. Technical variance. (A) ERCC RNA recovery of over four orders of magnitude. Shown is the average detection rate of ERCC spike-in molecules (y-axis) vs. the number of ERCC molecules added to each single cell (log scale) across 1536 CD11c$^+$ cells. The value for undetected molecule was set to (-3). The data reflect robust estimation of concentrations over ~4 orders of magnitudes, with some technical variability and provide bounds on the expected technical sequence specific recovery and sequencing bias in the protocol. Overall, following extensive filtering, we estimate recovery rate of 2-3% of the spiked-in molecules. **(B-C) Association between standard deviation and average of detected molecule counts.** Shown are coefficients of variance (CV, y-axis) vs. the average cellular mRNA (gray dots) and ERCC spike-in (red dots) molecule counts across 1536 Cd11c$^+$ cells (**B**) and 95 Cd8$^+$ pDCs (**C**). This analysis shows low technical variance between cellular mRNA and spike-in molecules compared to recently published methods (Wu et al. (*21*)). As expected, spike-in molecules have lower variability than cellular mRNA molecules, as their variance is only technical and does not involve a biological component. pDCs also show relatively homogeneous expression, especially compared to the heterogeneous CD11c+ dataset. Spike-in controls were processed using the same pipeline used for mouse sequences. (**D**) Similar to Fig 1C, but using 10, 30 cells.

We note that we recommend assessing technical variance in single-cell RNA-seq by plotting variance/mean against the mean, rather than the CV (as done in Fig. 1). This reflects the desire to see only technical binomial variance (that scale with the mean) in the data.

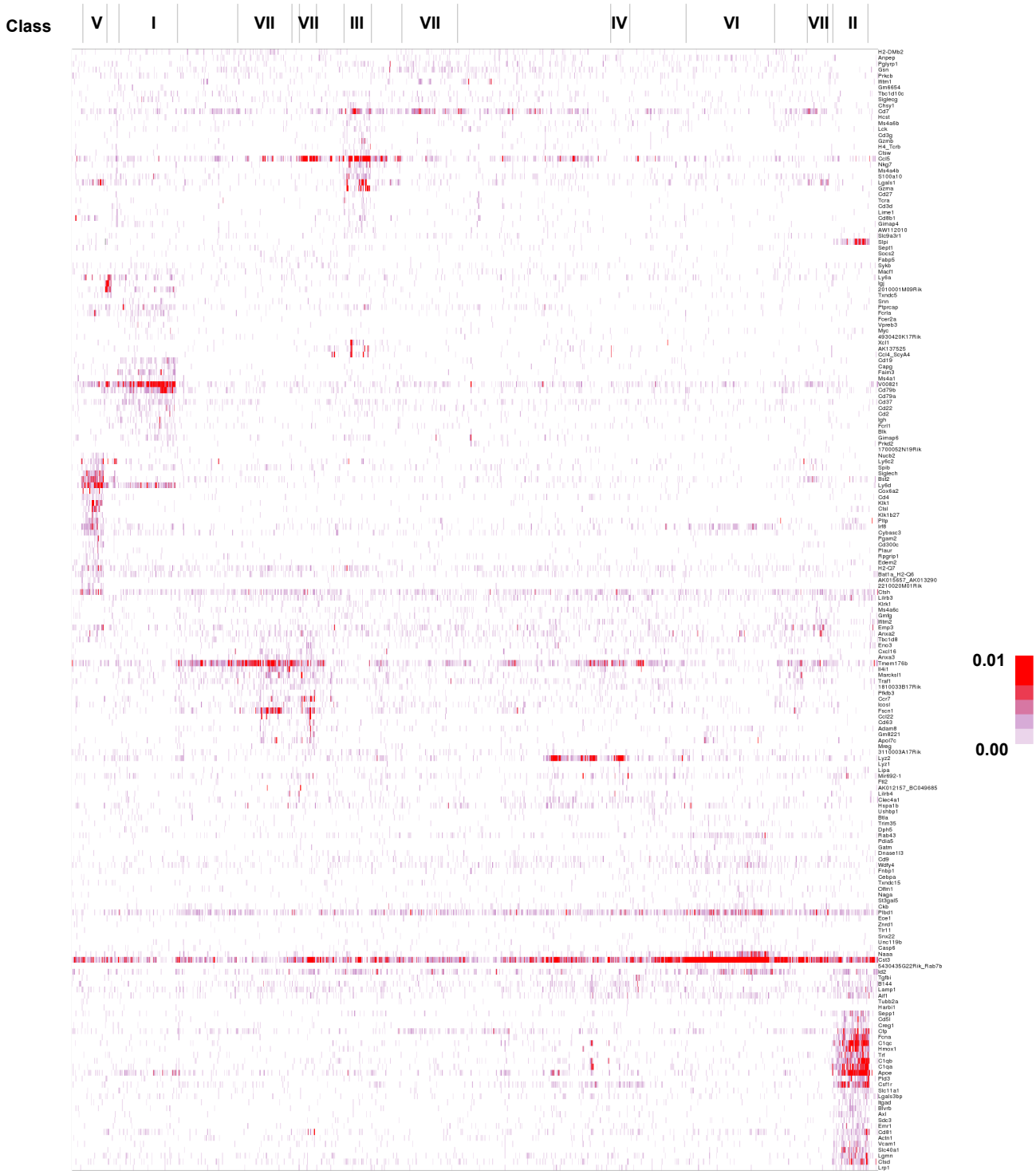

Figure S7. Gene-cell covariation structure over splenic cell population. Shown are color-coded (logarithmic scale) down-sampled molecular counts for selected genes (rows) over 1040 single cells (columns; ordering is identical to the clustered correlation map map of figure 2A). Groups of strongly correlated cells (as in figure 2A) are marked by black lines on top. This direct visualization of the dataset demonstrates how the correlation between cell type-specific gene-expression profiles combine to generate effective and usually unambiguous classification of cells into types. Once cell classes are identified, a much large number of cell-type specific genes are characterized through pooling together single cells.

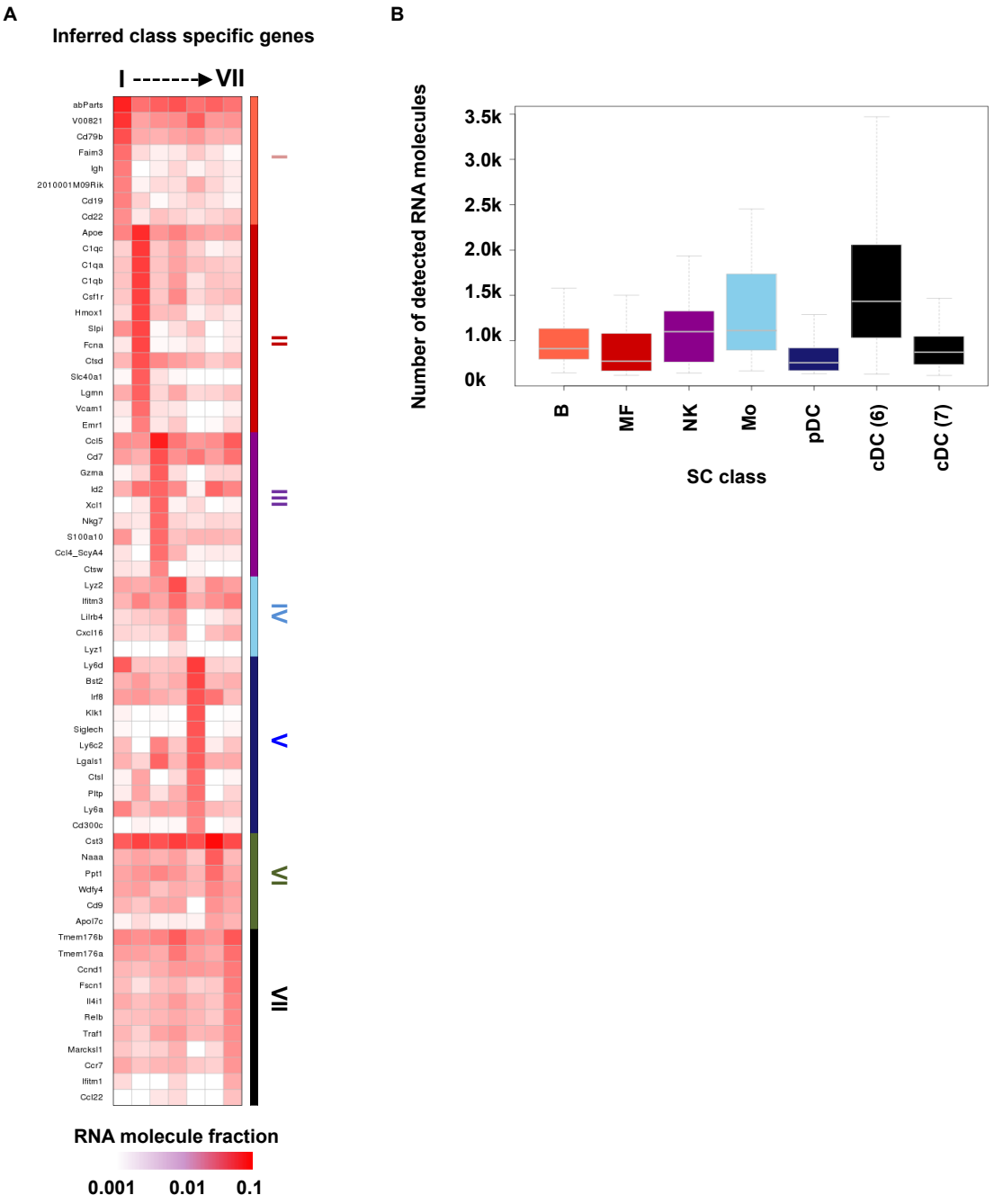

Figure S8. Mixture model for the CD11c+ enriched splenic cell population. (A) Selected cell-type specific genes for the spleen CD11c+ model. The color coded matrix depict class-specific mean expression for the CD11c^{+} model shown in Figure 2. Complete data for this model is provided in Table S2. **(B) Coverage varies among cell types.** Shown are the distributions pre-downsampled RMT coverage (Y axis) for over 2000 splenic cells, organized according to the inferred maximum *a posteriori* group (x-axis). While the model is inferred from uniform coverage data, different cell types are likely to provide markedly different numbers of recovered mRNAs.

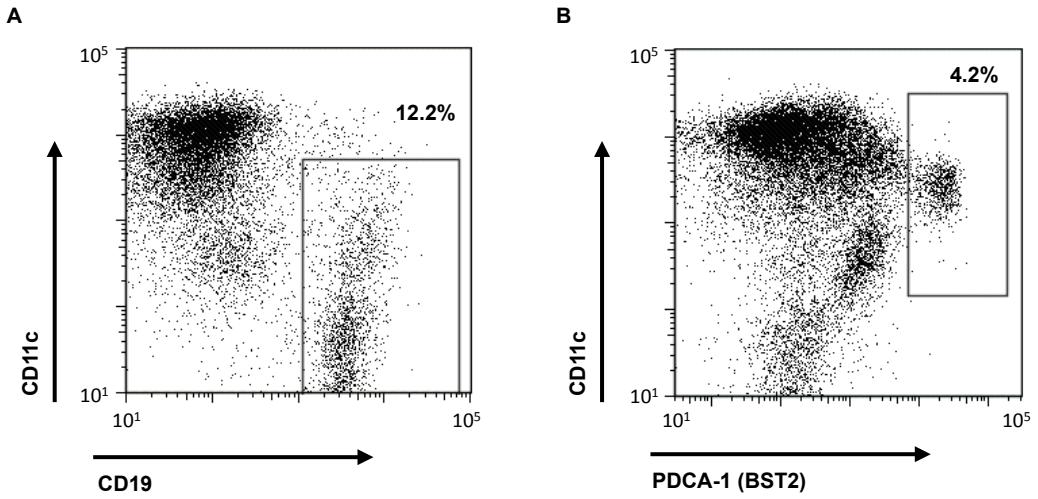

Figure S9. Non-DC frequency-estimation validation. FACS analysis was used to validate the estimated frequencies of B cells **(A)** and pDCs **(B)** in the CD11c$^+$ pool. Shown are independent experiments analyzing CD11c vs. CD19, a B cell marker, and PDCA-1 (Bst2), a pDC marker. B and pDC subpopulation frequencies are shown above the gating frame (gray).

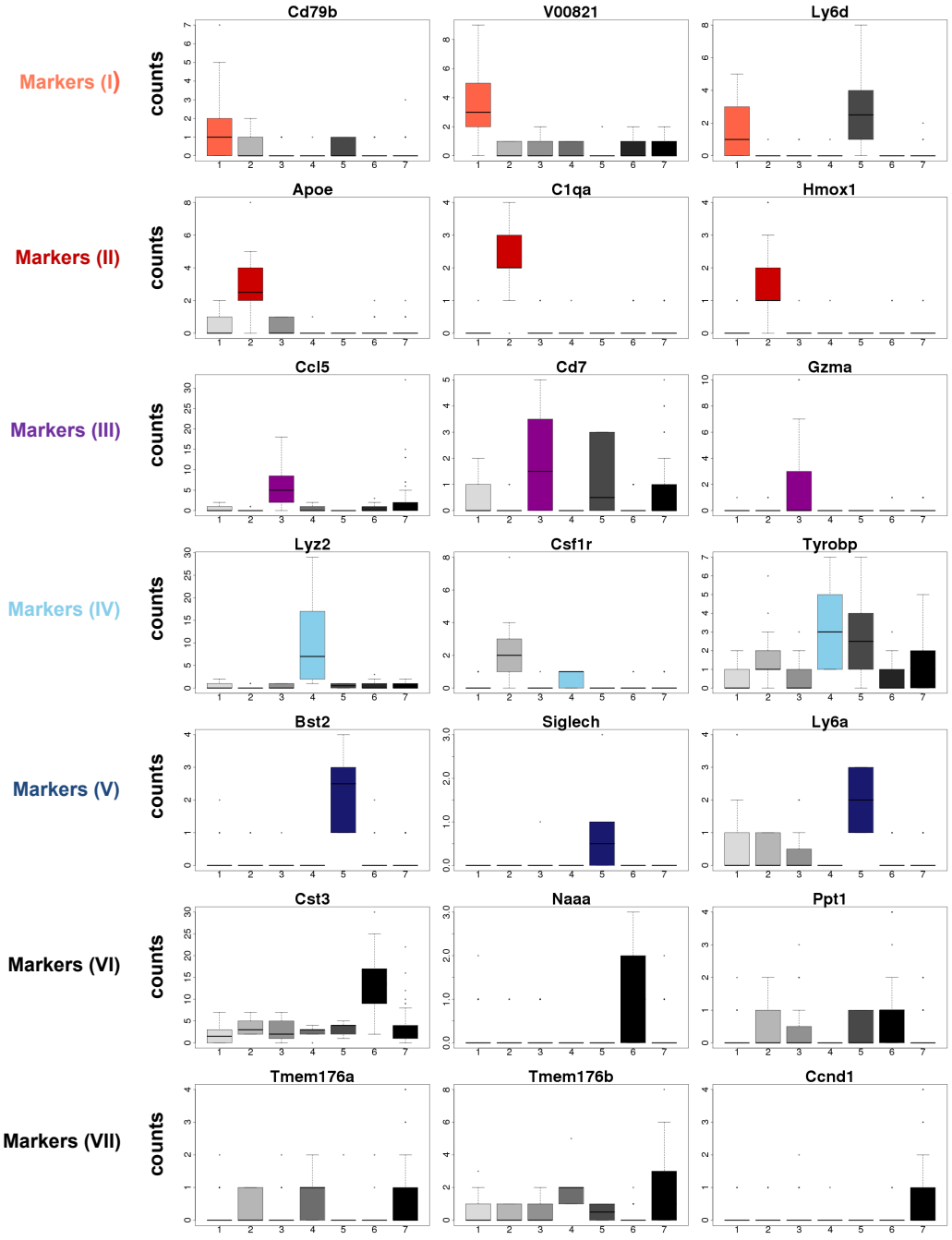

Figure S10. Expression distribution of representative genes shows variability between single-cell classes. Three strong "marker" genes were selected for each single-cell class. Shown here is their RMT coverage distribution in each of the seven classes (using down-sampled data to eliminate coverage difference between cell types).

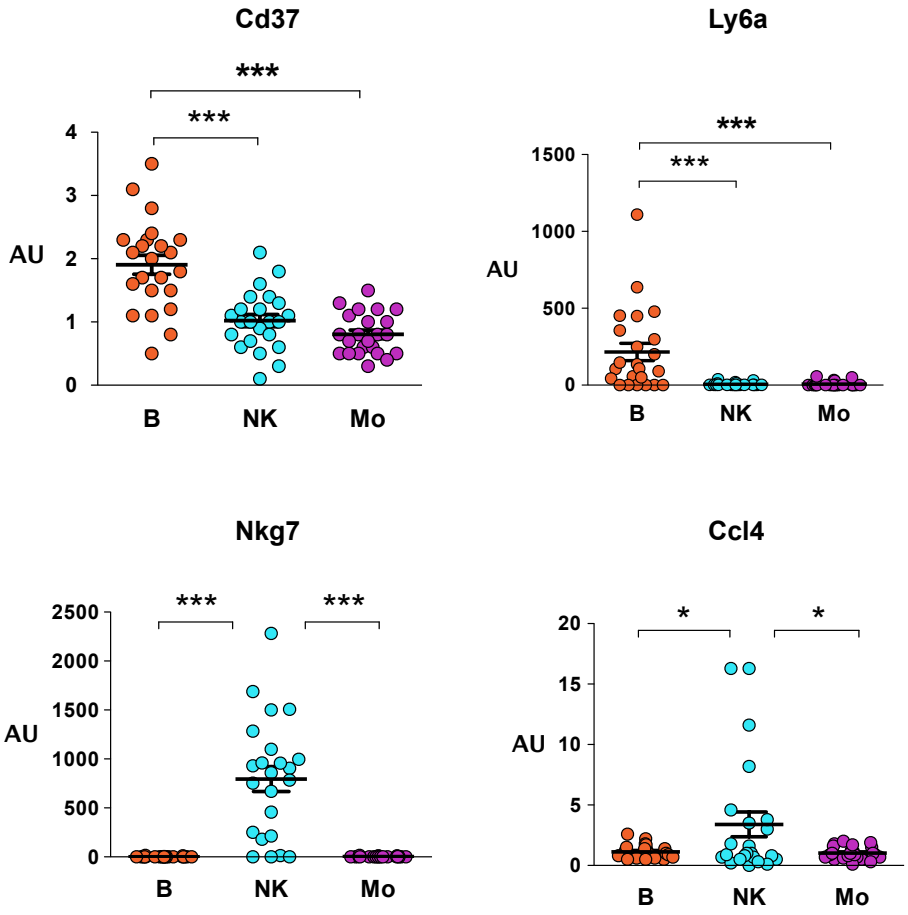

Figure S11. Single-cell gene expression validation by real-time PCR. Shown are four genes differentially expressed among three identified CD11c^{+} subpopulations (scatter plots) validated by single-cell RT pre-amplification real time qPCR. Error bars represent mean ± s.e.m. (n = 23 single cells); AU, arbitrary units; asterisks indicates ANOVA Bonferroni's Multiple Comparison test: *** *P* value < 0.001; * *P* value < 0.05

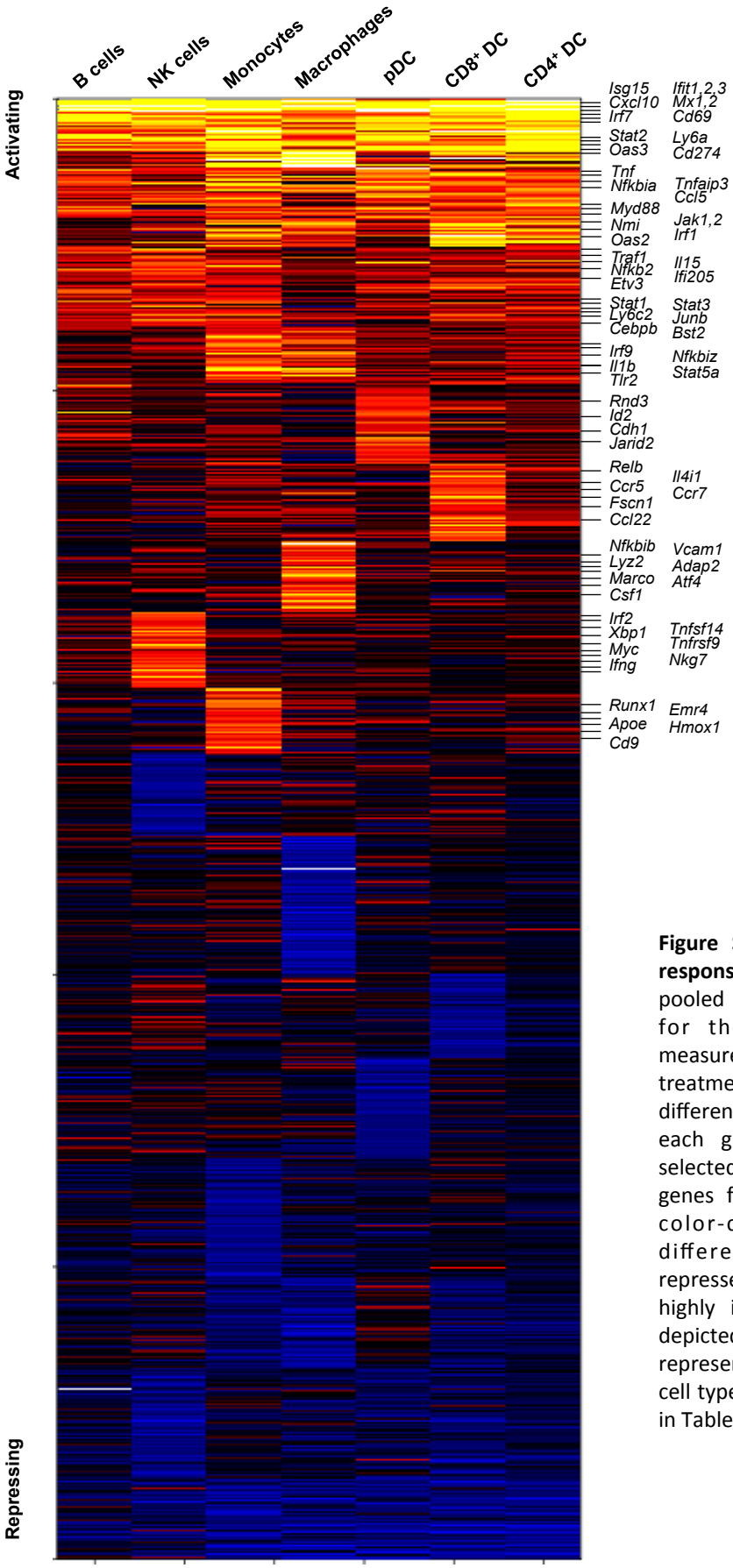

Figure S12. Clustering the spleen response to LPS by cell types. We pooled single-cell RNA-seq profiles for the seven splenic types, measured before and 2h after LPS treatment. We then computed the difference in mean expression for each gene and each class, and selected 2065 variably expressed genes for k-means clustering. The color-coded matrix indicates differential expression (blue-repressed, red – induced, yellow-highly induced) where genes are depicted in rows, and each column represent response in a different cell type. Complete data is available in Table S4.

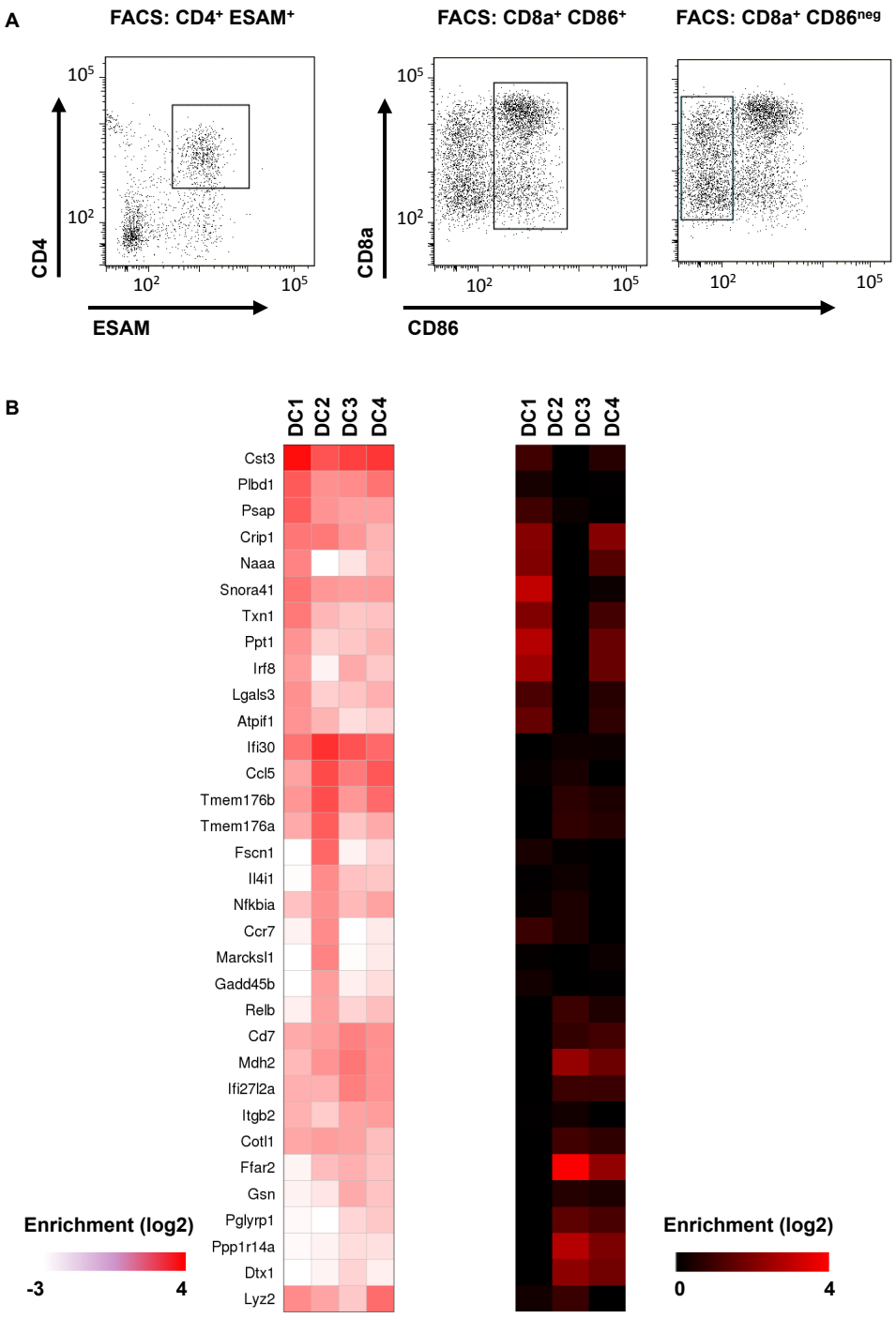

Figure S13. Marker-sorted DCs. (A) Comparison of FACS and single-cell RNA-based sorting was facilitated by FACS sorting and sequencing RNA from three DC subpopulations: $CD8^{high}$ $CD86^{+}$, $CD8^{inter}$ $CD86^{-}$ and $CD4^{+}$ $ESAM^{+}$. Gating is shown by gray boxes in the corresponding FACS plots. These sorted population were analyzed by single-cell RNA-seq and compared to the DC mixture model as shown in Fig 4B. **(B)** Shown are pooled single-cell mRNA mean counts (left) side by side with ImmGen gene expression data for three sorted DC classes (right). Genes that were specifically enriched in at least one of the three classes were selected for presentation. For the complete table of differentially expressed genes see table S5.

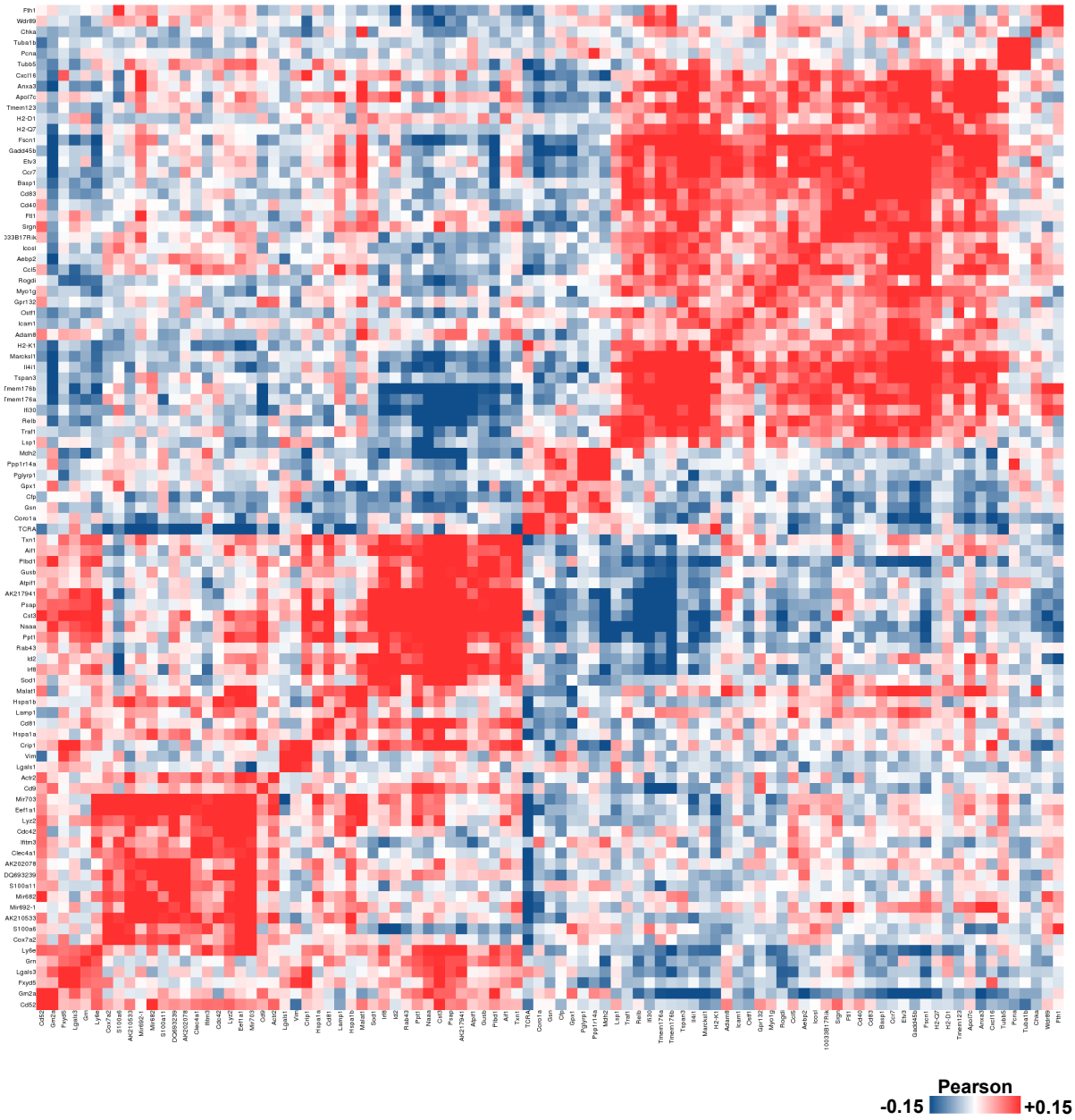

Figure S14. Shown is a gene correlation matrix depicting Pearson correlations between the single-cell RNA-seq profiles within 595 cells classified as DCs by our model. Only genes with at least 4 pairing with R > 0.14 are shown.

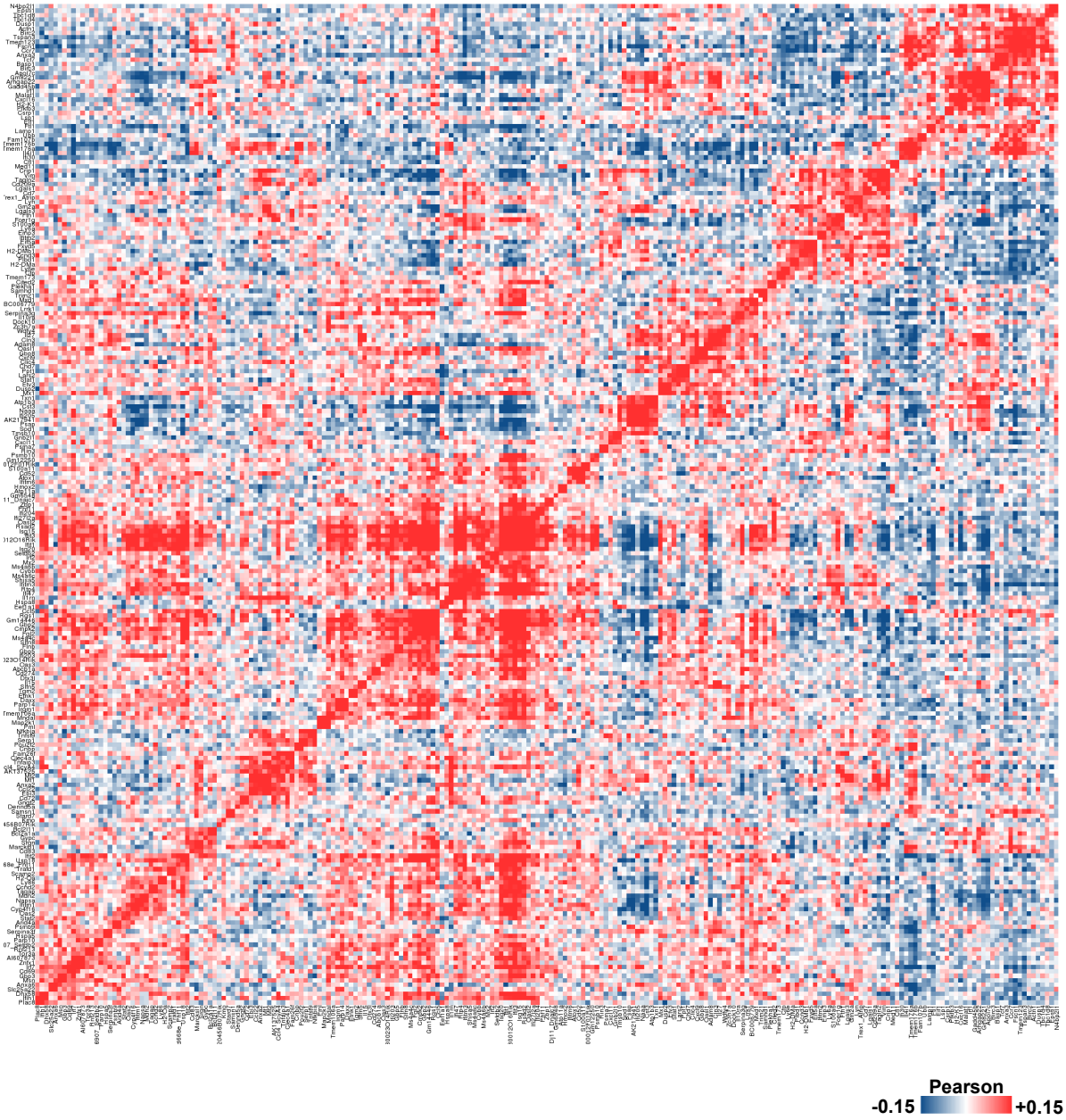

Figure S15. Shown is a gene correlation matrix depicting Pearson correlations between the *in vivo* LPS-treated single-cell RNA-seq profiles within 403 cells classified as DCs by our model. Only genes with at least 4 pairing with r>0.15 are shown.

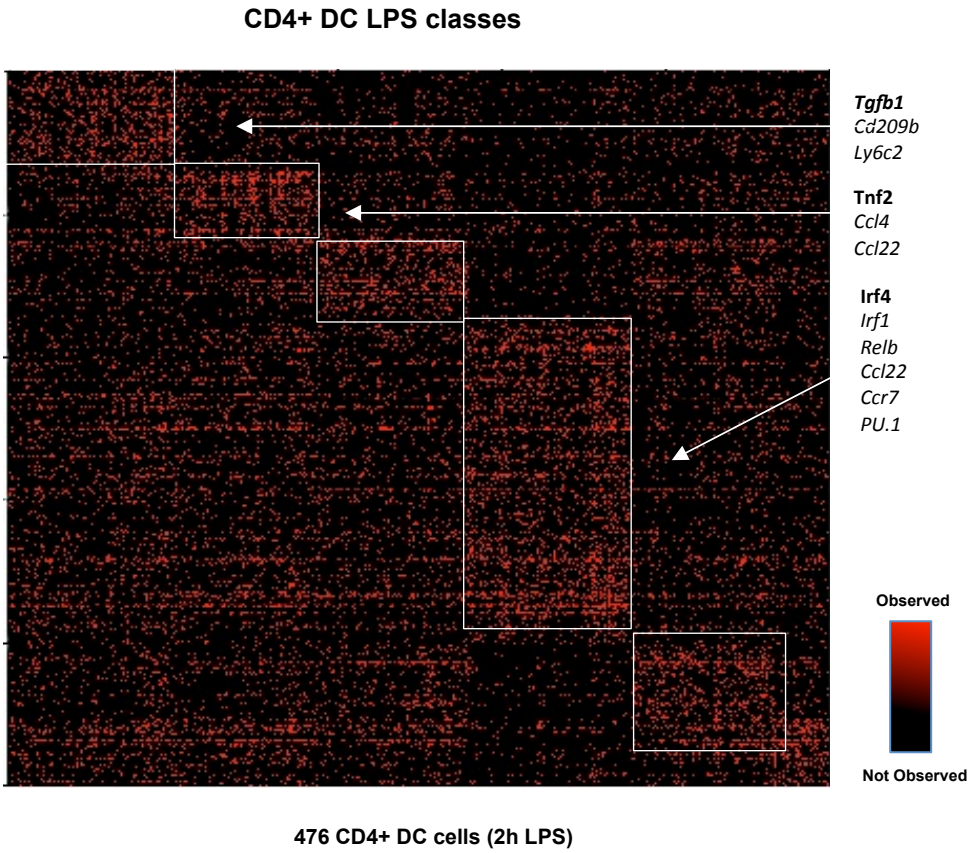

Figure S16. The heat map depicts mRNA counts in 476 single cells that were classified into class VII following *in vivo* LPS treatment. Cells were clusters using our EM-like iterative approach as described in the Methods. Genes are grouped according to the class in which they are mostly enriched. Complete data is available in Table S6.

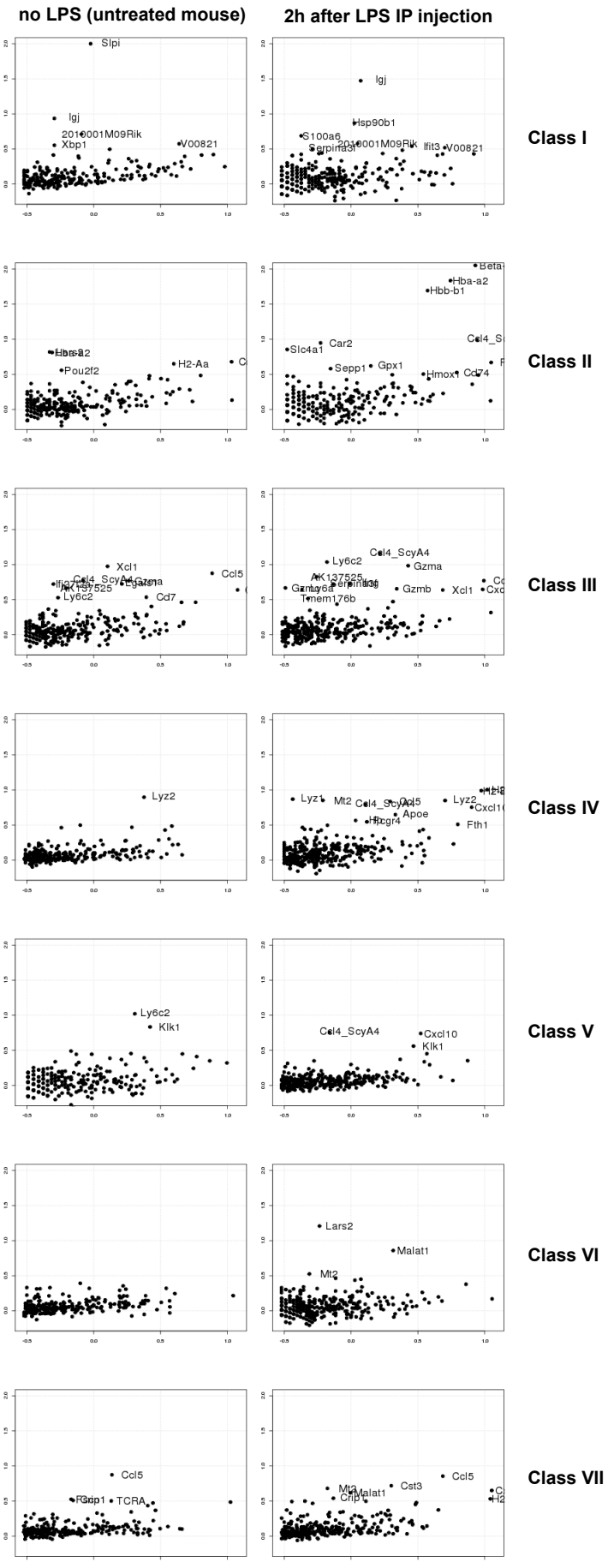

Figure S17. For each of the seven splenic cell classes, we show the mean and variance of all genes, before and after LPS treatment. Genes showing high variance indicate potential further organization of these populations into subtypes. Genes affected following LPS treatment indicate potential re-organization of the cell population into specific response classes, as outlined above in more detail for $CD4^+$ DCs (Fig S16).