Supplemental Data for

Early adaptive immune activation detected in monozygotic twins with prodromal multiple sclerosis

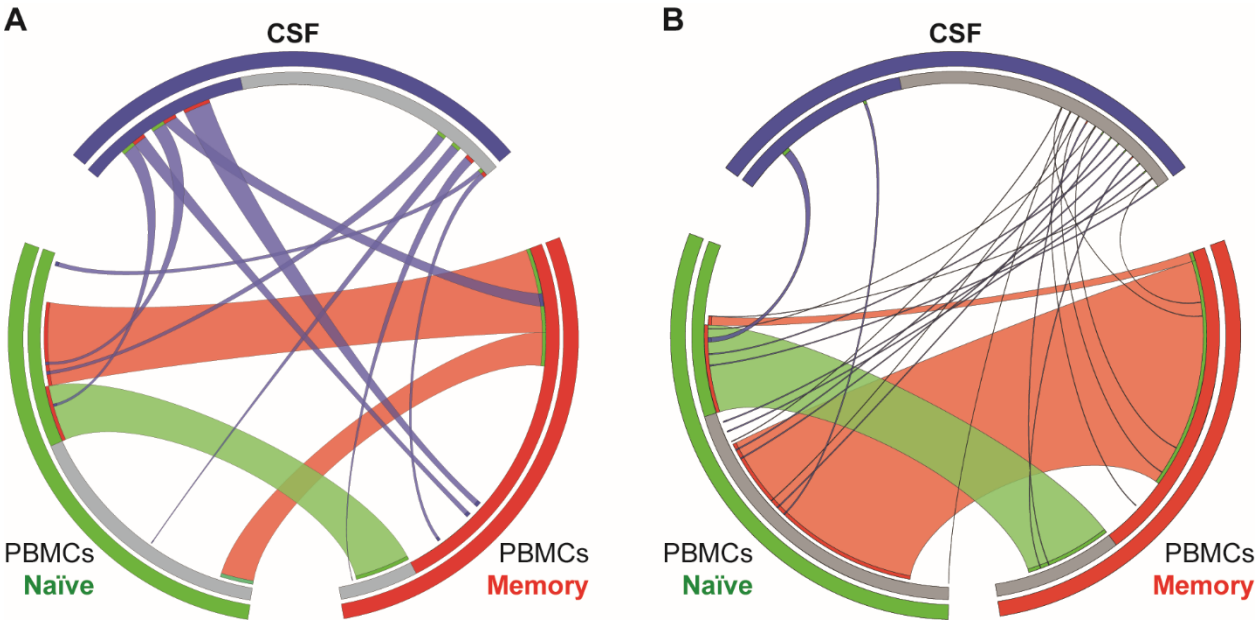
Eduardo Beltrán, Lisa Ann Gerdes, Julia Hansen, Andrea Flierl-Hecht, Stefan Krebs, Helmut Blum, Birgit Ertl-Wagner, Frederik Barkhof, Tania Kümpfel, Reinhard Hohlfeld, and Klaus Dornmair

* Corresponding authors:

Klaus.Dornmair@med.uni-muenchen.de and Eduardo.Beltran@med.uni-muenchen.de

The file includes:

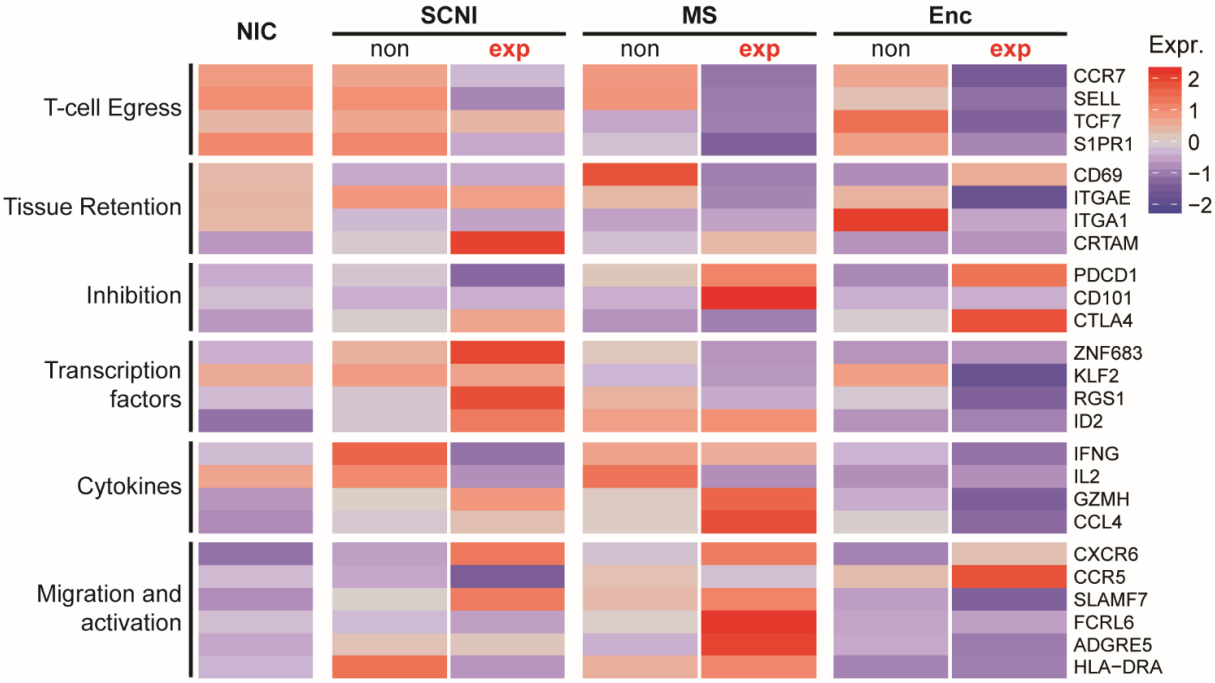
- Supplemental Figures S1-5
- Supplemental Table 1
- Supplemental Material



Supplemental Figure S1

TCR-ß repertoire of single CD8 T cells in CSF, and memory and naïve CD8 T cells in blood.

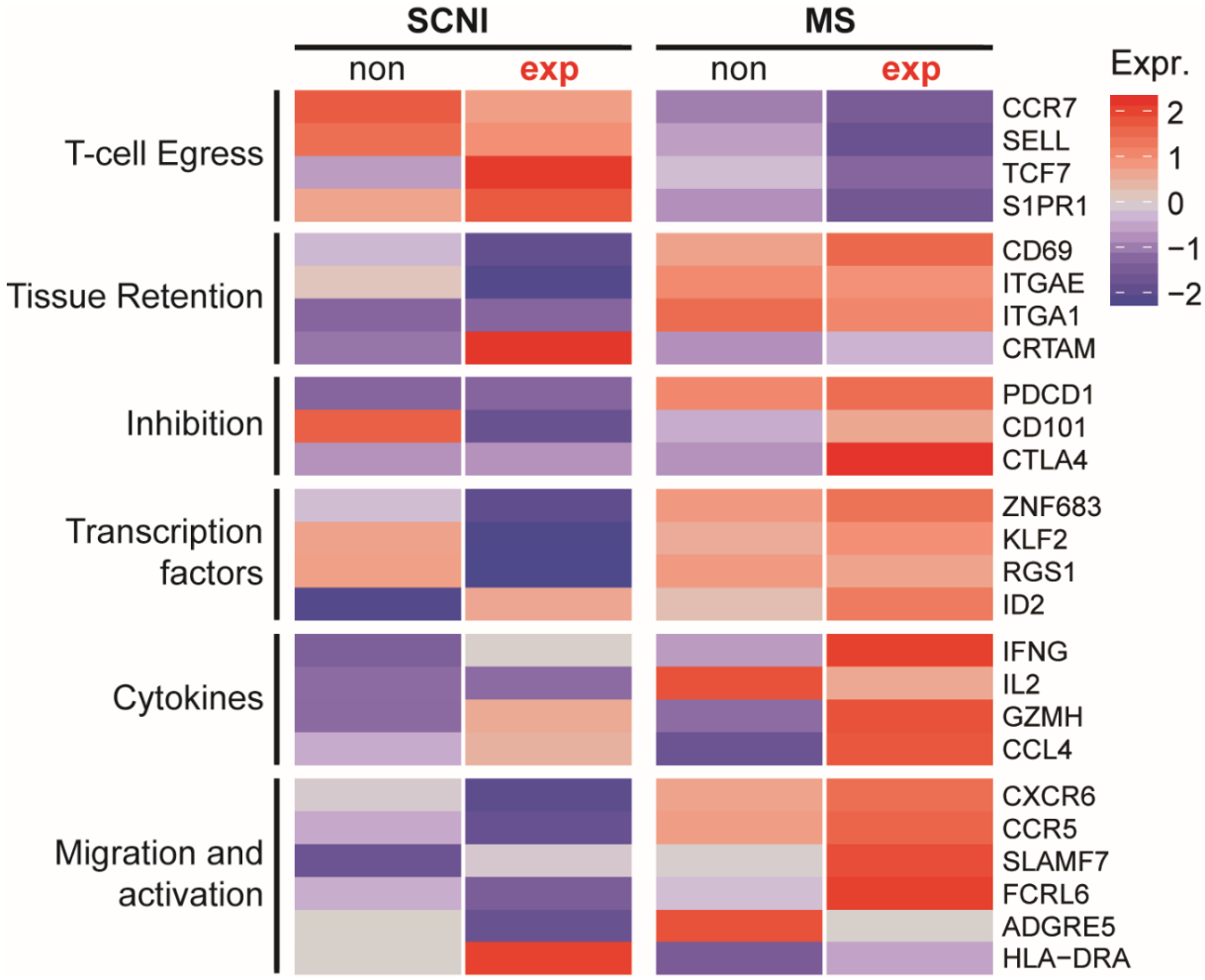
(A) The TCR β-chain repertoire of two SCNI co-twins (AV-H, BJ-H) is shown. **(B)** The TCR β-chain repertoire of the two corresponding MS twins (AV-MS, BJ-MS) is shown. The TCR repertoires of CSF derived CD8+ T cells is highlighted in blue at the top of each circle. The TCR repertoires of memory and naïve CD8 T cells from peripheral blood are shown in red at the right, and in green on the left of each circle. In the inner circles, the colored segments represent expanded clones, whereas the polyclonal background is depicted in gray. The widths of the segments in the inner circles indicate the relative abundance of clones. T cell clones shared between CSF and blood or between blood compartments are visualized as semicircular connection. Overlaps between the memory and naïve repertoires in blood are shown by semicircular connections. Images were generated using CIRCOS software.



Supplemental Figure S2

Heat map of gene expression levels of selected function-associated genes of CD4⁺ T cells.

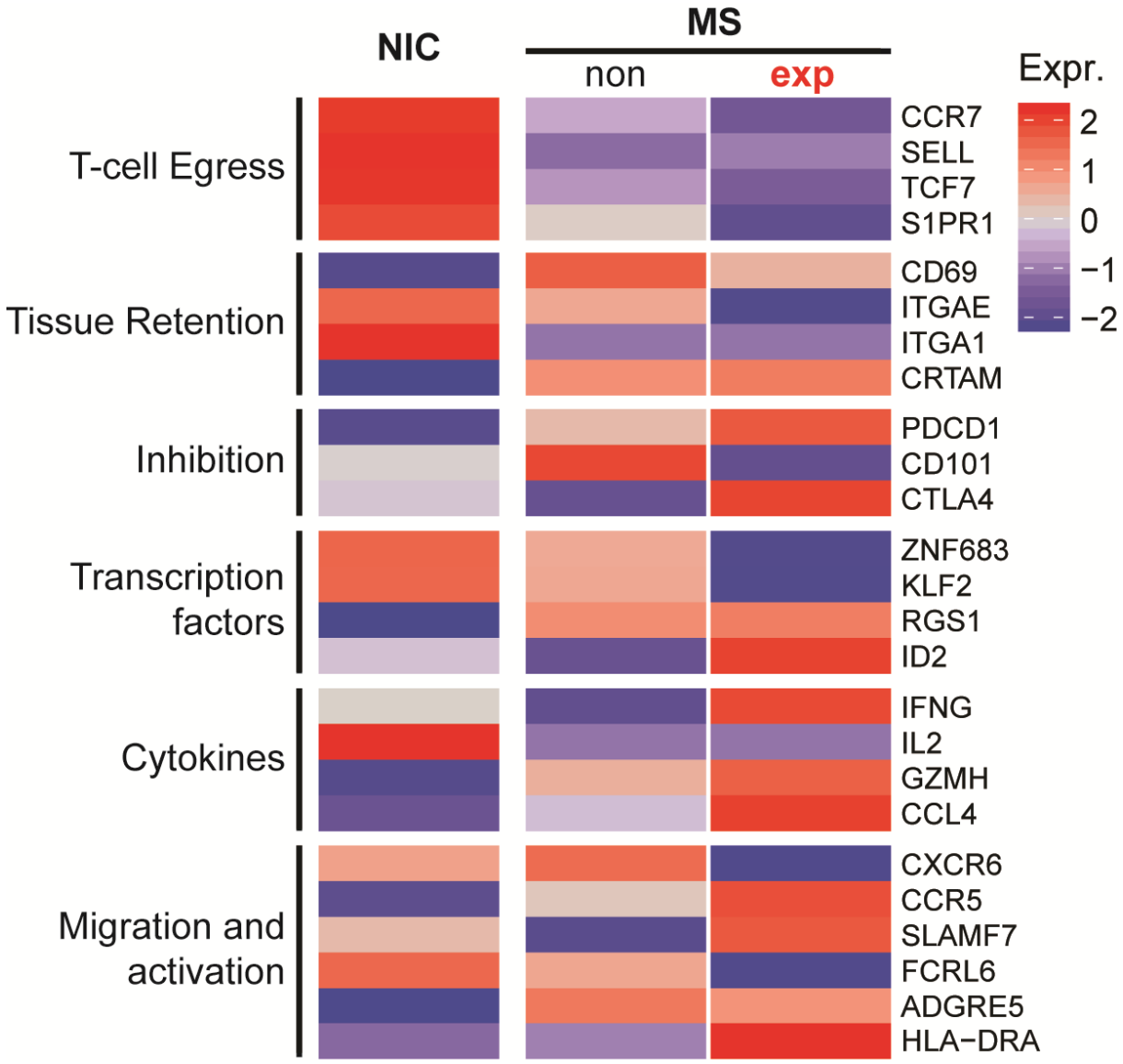
CD4⁺ T cells were selected based on CD4 surface marker expression and presence of at least one TCR chain as determined by NGS. Gene expression of 25 marker genes for homing, migration, and activation are shown for non-expanded CD4⁺ T cells and expanded CD4⁺ T cell clones from SCNI, MS, and Enc subjects. No distinction is made for NIC as the number of expanded clones is too low. Color scheme is based on z-score distribution from –2.5 (blue) to 2.5 (red).



Supplemental Figure S3

Heat map of gene expression levels of non-expanded and expanded CD8⁺ T cells from twin pair AU-SCNI and AU-MS.

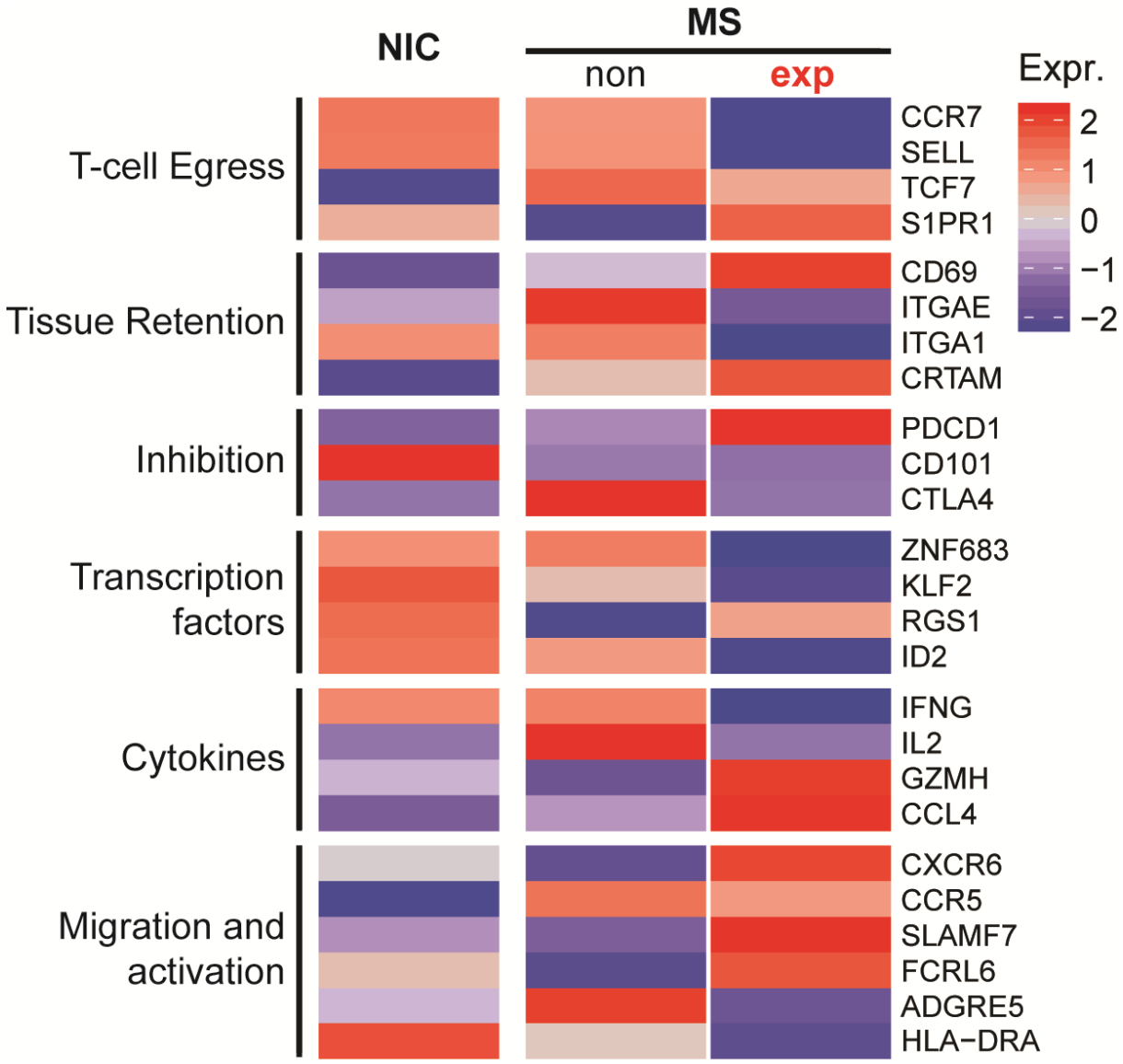
CD8⁺ T cells were selected based on flow cytometry staining for CD8 and presence of at least one TCR chain as determined by NGS. Gene expression of 25 marker genes for homing, migration, and activation are shown for non-expanded CD8⁺ T cells and expanded CD8⁺ T cell clones from SCNI, MS, and Enc subjects. No distinction is made for NIC as the number of expanded clones is too low. Color scheme is based on z-score distribution from –2.5 (blue) to 2.5 (red).



Supplemental Figure S4

Heat map of gene expression levels of non-expanded and expanded CD8$^+$ T cells from twin pair AV-H and AV-MS.

See legend to **Supplemental Figure S2** for details.



Supplemental Figure S5

Heat map of gene expression levels of non-expanded and expanded CD8⁺ T cells from twin pair BJ-H and BJ-MS.

See legend to **Supplemental Figure S2** for details. In the healthy co-twin #BJ-H, no expanded clones were detected.

Supplemental Table 1: List of the top 20 discriminative genes per cluster identified through unbiased clustering in Figure 1

gene	avg_logFC	pct.1	pct.2	p_val	p_val_adj	cluster ID
CD79A	2.18	0.70	0.04	1.79E-238	6.73E-234	Bcell
TNFRSF13C	1.86	0.74	0.05	5.90E-234	2.22E-229	Bcell
BANK1	2.14	0.70	0.05	4.87E-220	1.83E-215	Bcell
FCRL1	1.46	0.45	0.01	6.62E-214	2.49E-209	Bcell
IGHG3	1.78	0.67	0.04	3.50E-209	1.31E-204	Bcell
IGHA2	1.83	0.58	0.03	4.52E-201	1.70E-196	Bcell
VPREB3	1.95	0.38	0.01	3.61E-195	1.36E-190	Bcell
IGHG2	1.74	0.70	0.06	5.75E-193	2.16E-188	Bcell
LINC00926	1.49	0.41	0.01	8.46E-189	3.18E-184	Bcell
IGHA1	2.88	0.70	0.07	7.11E-181	2.67E-176	Bcell
IGHG1	1.97	0.78	0.09	1.05E-173	3.95E-169	Bcell
BLK	1.50	0.57	0.04	1.86E-173	6.99E-169	Bcell
FCRLA	1.68	0.43	0.02	1.30E-166	4.90E-162	Bcell
LINC01781	1.97	0.35	0.01	7.94E-158	2.98E-153	Bcell
IGHM	2.53	0.68	0.08	1.12E-151	4.20E-147	Bcell
IGHG4	1.22	0.63	0.06	4.67E-147	1.76E-142	Bcell
MS4A1	2.61	0.83	0.15	3.07E-143	1.16E-138	Bcell
IGKC	3.12	0.88	0.18	1.04E-140	3.89E-136	Bcell
CD22	1.02	0.34	0.01	1.31E-140	4.91E-136	Bcell
IGHGP	1.03	0.51	0.04	3.45E-135	1.30E-130	Bcell
CD4	0.86	0.59	0.21	1.33E-93	5.00E-89	CD4inTcellClusterI
MAL	1.01	0.40	0.09	3.27E-87	1.23E-82	CD4inTcellClusterI
CD40LG	1.02	0.39	0.10	8.17E-76	3.07E-71	CD4inTcellClusterI
IL7R	0.70	0.89	0.63	8.83E-72	3.32E-67	CD4inTcellClusterI
SERINC5	0.61	0.61	0.27	7.50E-67	2.82E-62	CD4inTcellClusterI
LDHB	0.67	0.86	0.63	1.97E-66	7.41E-62	CD4inTcellClusterI
CCR7	0.81	0.63	0.33	1.01E-55	3.81E-51	CD4inTcellClusterI
CD52	0.37	0.96	0.81	9.63E-54	3.62E-49	CD4inTcellClusterI
PRKCQ-AS1	0.78	0.42	0.17	9.36E-51	3.52E-46	CD4inTcellClusterI
TRAC	0.46	0.97	0.80	1.20E-46	4.50E-42	CD4inTcellClusterI
TNFRSF25	0.62	0.34	0.11	2.36E-46	8.85E-42	CD4inTcellClusterI
INPP4B	0.47	0.60	0.33	8.52E-42	3.20E-37	CD4inTcellClusterI
LEF1	0.77	0.42	0.19	1.21E-40	4.55E-36	CD4inTcellClusterI
AQP3	0.68	0.31	0.12	1.95E-37	7.32E-33	CD4inTcellClusterI
RCAN3	0.48	0.43	0.20	8.42E-37	3.17E-32	CD4inTcellClusterI
S1PR1	0.44	0.73	0.48	2.34E-35	8.78E-31	CD4inTcellClusterI
TESPA1	0.49	0.43	0.20	6.92E-35	2.60E-30	CD4inTcellClusterI
LTB	0.38	0.53	0.28	1.06E-33	4.00E-29	CD4inTcellClusterI

MYC	0.66	0.47	0.25	8.28E-33	3.11E-28	CD4inTcellClusterI
CD3G	0.37	0.97	0.80	1.33E-32	4.99E-28	CD4inTcellClusterI
CD4	0.66	0.63	0.28	4.40E-20	1.65E-15	CD4inTcellClusterII
GZMK	0.60	0.75	0.42	2.65E-16	9.94E-12	CD4inTcellClusterII
GZMA	0.50	0.94	0.64	1.09E-14	4.09E-10	CD4inTcellClusterII
GZMH	0.84	0.44	0.20	2.39E-13	8.99E-09	CD4inTcellClusterII
CD5	0.66	0.46	0.23	1.28E-11	4.80E-07	CD4inTcellClusterII
IKZF3	0.59	0.67	0.45	5.26E-11	1.97E-06	CD4inTcellClusterII
CST7	0.45	0.83	0.54	4.25E-10	1.60E-05	CD4inTcellClusterII
CD6	0.37	0.40	0.20	1.47E-09	5.51E-05	CD4inTcellClusterII
CXCR6	0.69	0.44	0.24	5.48E-09	0.00020581	CD4inTcellClusterII
SLAMF6	0.62	0.47	0.27	9.26E-09	0.00034791	CD4inTcellClusterII
F2R	0.42	0.34	0.16	1.50E-08	0.00056465	CD4inTcellClusterII
CD40LG	0.49	0.33	0.16	1.46E-07	0.00550472	CD4inTcellClusterII
EOMES	0.51	0.37	0.19	1.51E-07	0.00565932	CD4inTcellClusterII
ITM2A	0.44	0.70	0.51	2.81E-07	0.01054598	CD4inTcellClusterII
GIMAP4	0.35	0.81	0.61	1.18E-06	0.04448057	CD4inTcellClusterII
TC2N	0.39	0.72	0.56	3.79E-06	0.14248714	CD4inTcellClusterII
CNN2	0.37	0.64	0.49	4.20E-06	0.15766943	CD4inTcellClusterII
HMOX2	0.56	0.42	0.27	7.45E-06	0.27996581	CD4inTcellClusterII
FYN	0.50	0.71	0.59	7.80E-06	0.29327835	CD4inTcellClusterII
ARF1	0.38	0.74	0.58	1.30E-05	0.48675212	CD4inTcellClusterII
IL7R	0.49	0.91	0.67	1.42E-22	5.34E-18	CD8inTcellClusterI
ENSG00000239470	0.34	0.96	0.88	2.60E-17	9.76E-13	CD8inTcellClusterI
CCR7	0.62	0.60	0.38	1.92E-15	7.22E-11	CD8inTcellClusterI
LDHB	0.43	0.85	0.67	3.81E-14	1.43E-09	CD8inTcellClusterI
LINC00861	0.37	0.80	0.62	4.35E-13	1.64E-08	CD8inTcellClusterI
LINC02273	0.66	0.28	0.13	5.12E-12	1.93E-07	CD8inTcellClusterI
LEF1	0.51	0.41	0.23	2.75E-11	1.03E-06	CD8inTcellClusterI
SELL	0.32	0.81	0.63	1.46E-10	5.48E-06	CD8inTcellClusterI
S1PR1	0.30	0.72	0.52	4.13E-10	1.55E-05	CD8inTcellClusterI
TOB1	0.37	0.80	0.65	1.44E-09	5.42E-05	CD8inTcellClusterI
BCL11B	0.44	0.59	0.41	3.70E-09	0.0001389	CD8inTcellClusterI
TESPA1	0.32	0.41	0.24	7.72E-09	0.00029023	CD8inTcellClusterI
PRKCQ-AS1	0.34	0.37	0.21	1.68E-08	0.00062985	CD8inTcellClusterI
DGKA	0.40	0.60	0.44	1.99E-08	0.00074601	CD8inTcellClusterI
ARHGAP15	0.34	0.78	0.64	3.16E-08	0.00118719	CD8inTcellClusterI
RPL9	0.45	0.85	0.78	9.09E-08	0.0034161	CD8inTcellClusterI
PASK	0.33	0.27	0.15	2.95E-07	0.01108464	CD8inTcellClusterI
SERINC5	0.37	0.48	0.34	5.85E-07	0.02196802	CD8inTcellClusterI
TRAF3IP3	0.30	0.75	0.63	7.14E-07	0.02681642	CD8inTcellClusterI
ENSG00000170089	0.42	0.28	0.16	1.64E-06	0.06171542	CD8inTcellClusterI

CD8A	1.62	0.81	0.24	2.45E-217	9.19E-213	CD8inTcellClusterII
CD8B	1.53	0.68	0.16	2.22E-186	8.33E-182	CD8inTcellClusterII
CCL5	0.98	1.00	0.70	9.33E-176	3.51E-171	CD8inTcellClusterII
CD8B2	0.52	0.47	0.06	4.12E-153	1.55E-148	CD8inTcellClusterII
NKG7	1.03	0.86	0.36	3.06E-140	1.15E-135	CD8inTcellClusterII
CCL4	1.12	0.70	0.23	2.11E-133	7.92E-129	CD8inTcellClusterII
CST7	0.91	0.88	0.45	3.13E-126	1.18E-121	CD8inTcellClusterII
GZMA	0.77	0.92	0.57	2.64E-109	9.93E-105	CD8inTcellClusterII
KLRK1	0.88	0.63	0.21	3.59E-106	1.35E-101	CD8inTcellClusterII
GZMK	0.90	0.75	0.34	5.42E-100	2.04E-95	CD8inTcellClusterII
CTSW	0.74	0.82	0.45	2.08E-81	7.80E-77	CD8inTcellClusterII
TRAC	0.49	0.97	0.80	1.08E-66	4.07E-62	CD8inTcellClusterII
IL32	0.42	0.98	0.78	4.21E-59	1.58E-54	CD8inTcellClusterII
GZMH	1.07	0.41	0.15	5.13E-58	1.93E-53	CD8inTcellClusterII
CCL4L2	0.52	0.31	0.08	1.14E-56	4.27E-52	CD8inTcellClusterII
PRF1	0.69	0.64	0.36	8.45E-52	3.17E-47	CD8inTcellClusterII
SH2D1A	0.79	0.66	0.38	6.65E-51	2.50E-46	CD8inTcellClusterII
LINC01871	0.57	0.54	0.24	9.53E-51	3.58E-46	CD8inTcellClusterII
TRGC2	0.80	0.58	0.31	5.82E-49	2.19E-44	CD8inTcellClusterII
PYHIN1	0.61	0.74	0.47	8.72E-49	3.28E-44	CD8inTcellClusterII
CLEC10A	2.05	0.70	0.03	4.77E-203	1.79E-198	DC
CD1C	3.05	0.79	0.05	5.89E-198	2.21E-193	DC
FCER1A	3.60	0.91	0.09	2.31E-184	8.66E-180	DC
CD1E	2.56	0.65	0.04	1.32E-157	4.97E-153	DC
AIF1	2.08	0.87	0.08	1.96E-155	7.37E-151	DC
LYZ	1.48	0.76	0.06	3.15E-145	1.18E-140	DC
RNA5SP151	2.79	0.64	0.04	1.07E-144	4.04E-140	DC
LGALS2	1.55	0.29	0.00	3.13E-143	1.18E-138	DC
MNDA	2.45	0.75	0.07	1.80E-135	6.76E-131	DC
LST1	1.75	0.84	0.10	2.61E-131	9.80E-127	DC
IGSF6	1.96	0.71	0.07	2.55E-127	9.57E-123	DC
CPVL	2.03	0.65	0.06	5.70E-127	2.14E-122	DC
CST3	2.16	0.95	0.16	8.29E-123	3.11E-118	DC
HLA-DQA1	2.34	0.89	0.14	2.05E-121	7.70E-117	DC
IL13RA1	1.78	0.63	0.06	3.23E-111	1.21E-106	DC
CSF2RA	1.45	0.70	0.08	1.10E-106	4.13E-102	DC
PKIB	1.62	0.41	0.02	1.34E-104	5.04E-100	DC
HLA-DQB1	2.13	0.92	0.19	1.77E-103	6.64E-99	DC
FGL2	1.95	0.73	0.10	2.94E-98	1.11E-93	DC
MPEG1	1.57	0.65	0.08	4.16E-96	1.56E-91	DC
C1QA	2.26	0.78	0.03	1.04E-289	3.91E-285	Monocyte
C1QC	2.41	0.81	0.03	1.18E-286	4.42E-282	Monocyte

CD163	1.46	0.68	0.02	9.38E-281	3.53E-276	Monocyte
TREM2	1.84	0.61	0.01	2.18E-270	8.18E-266	Monocyte
FCGR1A	1.49	0.62	0.02	2.42E-252	9.10E-248	Monocyte
AIF1	2.01	0.96	0.07	1.05E-237	3.95E-233	Monocyte
C1QB	2.32	0.77	0.04	2.24E-231	8.42E-227	Monocyte
CD14	2.61	0.96	0.09	7.35E-231	2.76E-226	Monocyte
CSF1R	1.96	0.92	0.07	4.11E-228	1.55E-223	Monocyte
MS4A7	1.97	0.81	0.05	1.68E-226	6.32E-222	Monocyte
FCGR2A	1.87	0.98	0.09	2.23E-219	8.38E-215	Monocyte
CLEC7A	2.22	0.99	0.10	1.89E-214	7.10E-210	Monocyte
MSR1	1.69	0.88	0.07	3.58E-214	1.35E-209	Monocyte
SERPINA1	1.59	0.81	0.06	3.27E-210	1.23E-205	Monocyte
TLR2	1.71	0.64	0.03	3.61E-206	1.36E-201	Monocyte
MARCKS	1.71	0.90	0.08	7.89E-203	2.97E-198	Monocyte
LYZ	2.07	0.78	0.06	6.75E-201	2.54E-196	Monocyte
VMO1	1.67	0.50	0.02	2.85E-197	1.07E-192	Monocyte
CSF2RA	1.54	0.84	0.07	1.65E-191	6.21E-187	Monocyte
MS4A4A	1.61	0.57	0.02	2.54E-191	9.53E-187	Monocyte
TRDJ1	2.44	0.51	0.01	9.83E-246	3.69E-241	NK
SH2D1B	2.44	0.72	0.05	2.17E-234	8.17E-230	NK
TRGJP1	2.15	0.42	0.01	8.65E-198	3.25E-193	NK
TRDC	2.15	0.91	0.15	1.55E-172	5.83E-168	NK
KLRF1	2.29	0.65	0.06	3.37E-165	1.27E-160	NK
TYROBP	1.64	0.85	0.15	2.84E-138	1.07E-133	NK
FCER1G	1.66	0.89	0.20	4.55E-126	1.71E-121	NK
KLRD1	1.60	0.94	0.28	1.03E-119	3.87E-115	NK
KLRC2	1.39	0.58	0.07	4.39E-116	1.65E-111	NK
SPTSSB	1.77	0.37	0.02	1.98E-111	7.44E-107	NK
GNLY	2.18	0.85	0.30	1.87E-98	7.02E-94	NK
XCL2	1.37	0.68	0.13	3.15E-94	1.19E-89	NK
KLRC1	1.87	0.62	0.12	3.62E-90	1.36E-85	NK
XCL1	1.26	0.63	0.12	2.36E-88	8.85E-84	NK
KLRC3	0.66	0.37	0.04	2.20E-84	8.26E-80	NK
IL18RAP	1.70	0.57	0.11	3.38E-84	1.27E-79	NK
IL2RB	1.35	0.75	0.21	3.08E-80	1.16E-75	NK
ITGAX	1.11	0.60	0.13	6.10E-77	2.29E-72	NK
ENSG00000259188	1.32	0.28	0.02	9.22E-73	3.47E-68	NK
CTSW	1.08	0.94	0.51	2.91E-62	1.10E-57	NK
LILRA4	1.92	0.81	0.01	4.79E-293	1.80E-288	pDC
CLEC4C	2.64	0.81	0.01	3.99E-271	1.50E-266	pDC
ENSG00000248991	2.04	0.27	0.00	1.42E-123	5.35E-119	pDC
SERPINF1	2.89	0.88	0.05	1.86E-119	7.01E-115	pDC

LAMP5	1.60	0.34	0.01	6.86E-114	2.58E-109	pDC
TSPAN13	2.44	0.68	0.03	2.65E-112	9.95E-108	pDC
C1orf186	2.27	0.85	0.05	6.77E-111	2.54E-106	pDC
DERL3	1.96	0.81	0.04	2.42E-110	9.10E-106	pDC
FAM129C	1.57	0.78	0.04	7.86E-109	2.95E-104	pDC
TCL1A	2.29	0.56	0.02	3.35E-106	1.26E-101	pDC
MYBL2	1.46	0.44	0.01	1.32E-98	4.96E-94	pDC
IGKJ1	2.14	0.46	0.01	3.37E-96	1.27E-91	pDC
IL3RA	2.13	0.76	0.05	3.75E-89	1.41E-84	pDC
MPEG1	2.49	0.93	0.09	1.01E-81	3.80E-77	pDC
ENSG00000230138	2.37	0.51	0.02	2.79E-79	1.05E-74	pDC
SCN9A	1.54	0.44	0.02	1.72E-77	6.46E-73	pDC
LINC00996	2.27	0.51	0.03	1.99E-72	7.48E-68	pDC
GZMB	2.89	0.98	0.12	1.79E-71	6.73E-67	pDC
BCL11A	2.24	0.95	0.11	1.62E-70	6.08E-66	pDC
TPM2	1.53	0.51	0.03	1.26E-67	4.72E-63	pDC
IGLL5	1.44	0.64	0.00	0	0	Plasmablast
TNFRSF17	2.47	0.92	0.02	2.64E-302	9.93E-298	Plasmablast
DERL3	1.83	0.98	0.04	2.41E-199	9.04E-195	Plasmablast
IGHJ3P	1.80	0.46	0.01	1.22E-183	4.59E-179	Plasmablast
HRASLS2	1.29	0.40	0.00	2.23E-165	8.37E-161	Plasmablast
MZB1	2.56	1.00	0.06	2.34E-155	8.78E-151	Plasmablast
IGHJ6	1.67	0.42	0.01	3.55E-145	1.33E-140	Plasmablast
CD38	1.55	0.90	0.05	2.22E-140	8.35E-136	Plasmablast
IGLV6-57	1.60	0.46	0.01	5.11E-140	1.92E-135	Plasmablast
CD79A	2.03	0.98	0.06	1.10E-131	4.15E-127	Plasmablast
IGKJ5	1.87	0.48	0.01	3.25E-126	1.22E-121	Plasmablast
EAF2	2.00	0.96	0.07	5.36E-123	2.01E-118	Plasmablast
IGHA2	2.19	0.82	0.05	2.70E-119	1.02E-114	Plasmablast
IGHV3-30	1.77	0.44	0.01	7.41E-117	2.79E-112	Plasmablast
CLIC4	1.27	0.74	0.04	1.02E-108	3.83E-104	Plasmablast
IGHG3	1.99	0.88	0.07	3.32E-105	1.25E-100	Plasmablast
MAN1A1	1.41	0.92	0.07	1.76E-103	6.62E-99	Plasmablast
SEMA4A	1.39	0.86	0.07	5.53E-98	2.08E-93	Plasmablast
IGHG2	2.02	0.92	0.08	2.72E-96	1.02E-91	Plasmablast
IGKV4-1	1.76	0.48	0.02	2.16E-95	8.12E-91	Plasmablast

Footnotes:

avg_logFC - Value refers to average differential expression within one subset (log fold change)

pct.1 - Percentage of cells, within the cluster ID for which the gene is a marker, that detect the gene

pct.2 - Percentage of all the other cells, excluding the cluster ID for which the gene is a marker, that detect the gene

Supplemental Material:

R script utilizing Seurat package to generate t-SNE plots for the single-cell data.

```
#########################################################################
###  Single-cell analysis using Seurat – Beltran et al.  ################################
#########################################################################

library(devtools)
library(Seurat)
library(dplyr)
library(Matrix)

# Load the single-cell dataset
csf.data=read.table('counts_TPM_ALL.csv',header=T,row.names=1,sep='\t')
csf.data=log(csf.data+1)
celltypes=unlist(lapply(colnames(csf.data), ExtractField, 1))
table(celltypes)

# Create Seurat object. Keep all genes expressed in >= 3 cells (~0.1% of the data). Keep all cells with at least 200 detected genes
csf <- CreateSeuratObject(raw.data = csf.data, min.cells = 3, min.genes = 200, project = 'CSF_2018')

# AddMetaData adds columns to object@meta.data:
sc_cell_info <- read.table('sc_cell_info.txt', header = T,row.names=1)
csf <- AddMetaData(object = csf, metadata = sc_cell_info, col.name = c('Twin', 'Case', 'Sample',
'index.sort', 'Clones'))

# QC and selecting cells for further analysis. We calculate the percentage of mitochondrial genes here and store it in percent.mito using AddMetaData. We use object@raw.data since this represents non-transformed and non-log-normalized counts. The % of counts mapping to MT-genes is a common scRNAseq QC metric.
Sys.setlocale('LC_ALL','C')
mito.genes <- grep(pattern = "^MT-", x = rownames(x = csf@data), value = TRUE)
percent.mito <- Matrix::colSums(csf@raw.data[mito.genes, ])/Matrix::colSums(csf@raw.data)
csf <- AddMetaData(object = csf, metadata = percent.mito, col.name = 'percent.mito')

VlnPlot(object = csf, features.plot = c('nGene', 'percent.mito'), nCol = 2, x.lab.rot = TRUE)
csf <- FilterCells(object = csf, subset.names = c('nGene', 'percent.mito'), low.thresholds = c(200, -Inf),
high.thresholds = c(6000, 0.05))

# Normalizing the data
## Further normalization is performed since the dataset used to create the Seurat object is a merge of normalized datasets into one file. That means that, in the final merged file, the data from different dataset are not normalized to the same sequencing depth. Therefore, a further normalization is required.
csf <- NormalizeData(object = csf, normalization.method = 'LogNormalize', scale.factor = 10000)
```

```
# Detection of variable genes across the single cells
csf <- FindVariableGenes(object = csf, mean.function = ExpMean, dispersion.function = LogVMR,
x.low.cutoff = 0.0125, x.high.cutoff = 3, y.cutoff = 0.5)
length(x = csf@var.genes)

# Scaling the data and removing unwanted sources of variation
csf <- ScaleData(object = csf, vars.to.regress = 'percent.mito')

# Perform linear dimensional reduction
csf <- RunPCA(object = csf, pc.genes = csf@var.genes, do.print = TRUE, pcs.print = 1:5, genes.print = 5)
VizPCA(object = csf, pcs.use = 1:2)
PCAPlot(object = csf, dim.1 = 1, dim.2 = 2)

# ProjectPCA scores each gene in the dataset (including genes not included in the PCA) based on their
correlation with the calculated components
csf <- ProjectPCA(object = csf, do.print = FALSE)

# Heatmaps based on the PCA
PCHeatmap(object = csf, pc.use = 1, cells.use = 500, do.balanced = TRUE, label.columns = FALSE)
PCHeatmap(object = csf, pc.use = 1:20, cells.use = 500, do.balanced = TRUE, label.columns = FALSE,
use.full = FALSE)
PrintPCA(object = csf, pcs.print = 1:20, genes.print = 5, use.full = FALSE)

# Determine statistically significant principal components
csf <- JackStraw(object = csf, num.replicate = 100)
JackStrawPlot(object = csf, PCs = 1:20)
PCElbowPlot(object = csf)

# Cluster the cells
csf <- FindClusters(object = csf, reduction.type = 'pca', dims.use = 1:10, resolution = 0.6, print.output = 0,
save.SNN = TRUE)
PrintFindClustersParams(object = csf)

# Run Non-linear dimensional reduction (tSNE)
csf <- RunTSNE(object = csf, dims.use = 1:10, do.fast = TRUE, check_duplicates = FALSE)
TSNEPlot(object = csf, do.label=T)

# QC of each cluster
VlnPlot(object = csf, features.plot = c('nGene', 'percent.mito'), nCol = 2,
x.lab.rot = TRUE)

## Cells in Cluster #3 that were characterized by low number of detected genes and low mitochondrial
gene transcripts (indicating low-quality cells) were removed
new.subset <- SubsetData(object = csf, ident.remove = "3")
csf_v1 <- csf
save(csf_v1, file = "./csf_v1_2018.Rda")
csf <- new.subset
csf <- FilterCells(object = csf, subset.names = c('nGene', 'percent.mito'), low.thresholds = c(200, -Inf),
high.thresholds = c(6000, 0.025))
```

```
## Additional round of clustering after SubsetData
# re-running FindVariableGenes() and ScaleData()
csf <- FindVariableGenes(object = csf, mean.function = ExpMean,
dispersion.function = LogVMR, x.low.cutoff = 0.0125, x.high.cutoff = 3, y.cutoff = 0.5)
length(x = csf@var.genes)
csf <- ScaleData(object = csf, vars.to.regress = 'percent.mito')
csf <- RunPCA(object = csf, pc.genes = csf@var.genes, do.print = TRUE, pcs.print = 1:5, genes.print = 5)
VizPCA(object = csf, pcs.use = 1:2)
PCAPlot(object = csf, dim.1 = 1, dim.2 = 2)
csf <- ProjectPCA(object = csf, do.print = FALSE)
PCHeatmap(object = csf, pc.use = 1, cells.use = 500, do.balanced = TRUE, label.columns = FALSE)
PCHeatmap(object = csf, pc.use = 1:20, cells.use = 500, do.balanced = TRUE,label.columns = FALSE,
use.full = FALSE)
PrintPCA(object = csf, pcs.print = 1:20, genes.print = 5, use.full = FALSE)
csf <- JackStraw(object = csf, num.replicate = 100)
JackStrawPlot(object = csf, PCs = 1:20)
PCElbowPlot(object = csf)

csf <- FindClusters(object = csf, reduction.type = 'pca', dims.use = 1:11,
resolution = 0.6, print.output = 0, save.SNN = TRUE, force.recalc = TRUE)
PrintFindClustersParams(object = csf)
csf <- RunTSNE(object = csf, dims.use = 1:11, do.fast = TRUE, check_duplicates = FALSE)
TSNEPlot(object = csf, do.label=T)

# QC of each cluster
VlnPlot(object = csf, features.plot = c('nGene', 'percent.mito'), nCol = 2, x.lab.rot = TRUE)

## Cells in Cluster #6 that were characterized primarily by mitochondrial and ribosomal gene transcripts
(indicating low-quality cells) were removed
new.subset <- SubsetData(object = csf, ident.remove = "6")
csf_v2 <- csf
csf_v2
save(csf_v2, file = "./csf_v2_2018.Rda")
csf <- new.subset

## Additional round of clustering after SubsetData
# re-running FindVariableGenes() and ScaleData()
csf <- FindVariableGenes(object = csf, mean.function = ExpMean,
dispersion.function = LogVMR, x.low.cutoff = 0.0125, x.high.cutoff = 3, y.cutoff = 0.5)
length(x = csf@var.genes)
csf <- ScaleData(object = csf, vars.to.regress = 'percent.mito')
csf <- RunPCA(object = csf, pc.genes = csf@var.genes, do.print = TRUE, pcs.print = 1:5, genes.print = 5)
VizPCA(object = csf, pcs.use = 1:2)
PCAPlot(object = csf, dim.1 = 1, dim.2 = 2)
csf <- ProjectPCA(object = csf, do.print = FALSE)
PCHeatmap(object = csf, pc.use = 1, cells.use = 500, do.balanced = TRUE, label.columns = FALSE)
PCHeatmap(object = csf, pc.use = 1:20, cells.use = 500, do.balanced = TRUE, label.columns = FALSE,
use.full = FALSE)
```

```
PrintPCA(object = csf, pcs.print = 1:20, genes.print = 5, use.full = FALSE)
csf <- JackStraw(object = csf, num.replicate = 100)
JackStrawPlot(object = csf, PCs = 1:20)
PCElbowPlot(object = csf)

csf <- FindClusters(object = csf, reduction.type = 'pca', dims.use = 1:11, resolution = 0.6, print.output = 0,
save.SNN = TRUE, force.recalc = TRUE)
PrintFindClustersParams(object = csf)
csf <- RunTSNE(object = csf, dims.use = 1:11, do.fast = TRUE, check_duplicates = FALSE)

TSNEPlot(object = csf, do.label=T)

csf.markers <- FindAllMarkers(object = csf, only.pos = TRUE, min.pct = 0.25, thresh.use = 0.25)
csf.markers %>% group_by(cluster) %>% top_n(2, avg_logFC)
write.table(csf.markers %>% group_by(cluster) %>% top_n(10, avg_logFC), 'csf.markers.tsv', sep='\t')

top10 <- csf.markers %>% group_by(cluster) %>% top_n(10, avg_logFC)
# setting slim.col.label to TRUE will print just the cluster IDS instead of every cell name
DoHeatmap(object = csf, genes.use = top10$gene, slim.col.label = TRUE, remove.key = TRUE)

## Heatmap shows that cells in clusters #0 and #3 share same top markers, and therefore were clustered
together. And the same holds for clusters #1 and #2. Tiny cluster between clusters #5 and #9 which
contains platelet-like cells coming from the blood samples, was removed for the figure in the publication.

TSNEPlot(object = csf, do.return = TRUE, group.by = 'index.sort', do.label = TRUE)
TSNEPlot(object = csf, do.return = TRUE, group.by = 'Clones', do.label = TRUE)
VlnPlot(object = csf, features.plot = c('nGene', 'percent.mito'), nCol = 2, x.lab.rot = TRUE)

write.table(csf@meta.data, 'meta.data.tsv', sep='\t')

save(csf, file = "./csf_2018.Rda")


###################################

sessionInfo()

# R version 3.4.4 (2018-03-15)
# Platform: x86_64-pc-linux-gnu (64-bit)
# Running under: Ubuntu 18.04.1 LTS

# Matrix products: default
# BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
# LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1

# locale:
# [1] LC_CTYPE=C LC_NUMERIC=C
# [3] LC_TIME=C LC_COLLATE=C
# [5] LC_MONETARY=C LC_MESSAGES=en_US.UTF-8
```

```
# [7] LC_PAPER=de_DE.UTF-8 LC_NAME=C
# [9] LC_ADDRESS=C LC_TELEPHONE=C
# [11] LC_MEASUREMENT=de_DE.UTF-8 LC_IDENTIFICATION=C

# attached base packages:
# [1] stats graphics grDevices utils datasets methods # base

# other attached packages:
# [1] bindrcpp_0.2.2 dplyr_0.7.8 Seurat_2.3.4 Matrix_1.2-15 # cowplot_0.9.3
# [6] ggplot2_3.1.0 usethis_1.4.0 devtools_2.0.1
```