Gene regulatory network of human GM-CSF secreting T helper cells

Szabolcs Éliás[1,*], Angelika Schmidt[1], David Gomez-Cabrero[1,2,3] and Jesper Tegnér[1,4,*]

[1] Unit of Computational Medicine, Center for Molecular Medicine, Department of Medicine Solna, Karolinska Institutet, Karolinska University Hospital & Science for Life Laboratory, 17176 Solna, Stockholm, Sweden

[2] Mucosal & Salivary Biology Division, King's College London Dental Institute, London, SE1 9RT, UK

[3] Navarrabiomed, Complejo Hospitalario de Navarra (CHN), Universidad Pública de Navarra (UPNA), IdiSNA, 31008 Pamplona, Spain

[4] Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955–6900, Saudi Arabia

* To whom correspondence should be addressed. Email: sz.e@outlook.com or jesper.tegner@ki.se

ABSTRACT

GM-CSF produced by autoreactive CD4 positive T helper cells is involved in the pathogenesis of autoimmune diseases, such as Multiple Sclerosis. However, the molecular regulators that establish and maintain the features of GM-CSF positive CD4 T cells are unknown. In order to identify these regulators, we isolated human GM-CSF producing CD4 T cells from human peripheral blood by using a cytokine capture assay. We compared these cells to the corresponding GM-CSF negative fraction, and furthermore, we studied naïve CD4 T cells, memory CD4 T cells and bulk CD4 T cells from the same individuals as additional control cell populations. As a result, we provide a rich resource of integrated chromatin accessibility (ATAC-seq) and transcriptome (RNA-seq) data from these primary human CD4 T cell subsets, and we show that the identified signatures are associated with human autoimmune disease, especially Multiple Sclerosis. By combining information about mRNA expression, DNA accessibility and predicted transcription factor binding, we reconstructed directed gene regulatory networks connecting transcription factors to their targets, which comprise putative key regulators of human GM-CSF positive CD4 T cells as well as memory CD4 T cells. Our results suggest potential therapeutic targets to be investigated in the future in human autoimmune disease.

INTRODUCTION

CD4 positive T helper cells (Th) are crucial players in the immune system which exert their effects mainly by producing cytokines. CD4 T cell subsets are usually classified based on expression of "lineage-defining" transcription factors (TFs) as well as the signature cytokines they secrete [1]. However, the distinction is not clear-cut, since different signature cytokines can be expressed simultaneously and plasticity between subsets occurs [2,3]. In view of the "classical" distinction of CD4 T cell subsets, particularly Th1 and Th17 subsets are involved in the establishment of autoimmune diseases such as Multiple Sclerosis (MS) and the corresponding rodent model Experimental Autoimmune Encephalomyelitis (EAE), which are thought to be driven by the T cell-released cytokines Interleukin-17 (IL-17), Interferon-γ (IFN-γ), IL-22 and granulocyte-macrophage colony-stimulating factor (GM-CSF). Of these, GM-CSF was determined as the key cytokine in EAE pathogenesis, since only knocking out GM-CSF (but neither IFN-γ, IL-17A nor IL-17F) could completely protect the animals from induced EAE [4–6]. Furthermore, it has been demonstrated that specifically the GM-CSF produced by autoreactive T cells was necessary for EAE induction, while T cell-produced IFN-γ and IL-17 were dispensable [7–9]. In line with these results, GM-CSF expression by Th cells was required for neuroinflammation in EAE, and even in the presence of IFN-γ and IL-17A producing Th cells, pathogenicity was lost upon loss of GM-CSF [10]. Importantly, in humans the fraction of GM-CSF positive (and IFN-γ positive) cells within CD4 T cells was elevated in MS patients' cerebrospinal fluid compared to controls, while IL-17A positive cell fractions were not strikingly different in these reports [11,12]. The fraction of GM-CSF positive and IFN-γ positive cells was also increased in peripheral blood of MS patients in one report [13], but not in another [11]. Similarly, enhanced fractions of GM-CSF producing CD4 T cells have been observed in synovial fluid of patients with juvenile arthritis along with the well-known enhanced GM-CSF levels in synovial fluid [14,15]. Interestingly, an expanded Th subset producing GM-CSF was found in blood and CNS of MS patients and could be diminished by disease-modifying therapy [16], suggesting GM-CSF producing Th cells to be an attractive therapeutic target. Of note, targeting GM-CSF in MS or arthritis is subject to several ongoing clinical studies, highlighting the importance of this cytokine in these diseases [15]. Based on this cumulative evidence of the significance of GM-CSF producing CD4 T cells in human autoimmune disease, understanding the factors driving and defining GM-CSF positive T cells would be of utmost importance for targeting them therapeutically.

Beyond their established pathogenic role in autoimmune diseases, GM-CSF producing Th cells have also been implicated in other, inflammatory diseases. In sepsis, enhanced fractions of GM-CSF producing T cells were associated with a poor outcome [17]. Notably, GM-CSF producing Th cells have also been implicated in SARS-CoV-2 infection, especially in patients with severe course of coronavirus disease 2019 (COVID-19) [18–20]. Due to the pleiotropic roles of GM-CSF in immune disease and lung inflammation, GM-CSF-targeting therapeutic approaches are currently explored in clinical trials to treat COVID-19 [21]. Murine CD4 T cell populations expressing IL-17A and GM-CSF have been observed and termed "pathogenic Th17" cells because they have the potential to induce EAE [22–24]. However, only one of these studies showed co-expression of both cytokines on the

single cell level [24]. Although a T cell can express GM-CSF simultaneously with other cytokines such as IFN-γ, a “GM-CSF only” producing murine T cell subset was also proposed and associated with enhanced encephalitogenic activity over IL-17 and IFN-γ producing T cells in EAE [9]. The existence of a corresponding separate “GM-CSF-only” human T cell subset has also been proposed [25], because a substantial subset of human GM-CSF positive CD4 T cells produces GM-CSF in the absence of any other classical Th lineage-defining cytokines, transcription factors or surface markers [11,26]. Furthermore, GM-CSF producing CD4 T cells are induced by different sets of cytokines compared to other Th cell subsets [25,27]. In fact, GM-CSF and IL-17A expression by human CD4 T cells has been found to be mutually exclusive on single cell level [11] or at least less frequent than co-expression of IFN-γ and GM-CSF [26,27].

Despite the importance of GM-CSF producing T cells, there is no specific marker to distinguish such cells from others to date. Although combinations of presence and absence of non-exclusive surface markers has been useful to delineate “GM-CSF-only” cells [11], the fraction of GM-CSF producing cells that also produces other cytokines such as IFN-γ is excluded by this approach. Furthermore, the driving molecules for GM-CSF production remain unclear. Together, these observations suggest that the characterization of human GM-CSF positive CD4 T cells isolated based on their functional profile (GM-CSF production) rather than by distinction of the “classical” Th1 and Th17-like phenotypic markers may enable the identification of factors regulating GM-CSF production in CD4 T cells. Hence, in order to understand the regulators and molecular patterns defining human GM-CSF positive CD4 T cells, we here isolated those cells actively secreting GM-CSF from human peripheral blood *ex vivo* by cytokine “capture” assay, starting with bulk CD4 T cells. We then studied their transcriptome by RNA-sequencing (RNA-seq). Studying a single data type like mRNA expression separately may not be sufficient for identification of all regulatory factors, since TFs themselves are often regulated by post-transcriptional modifications, intracellular translocation or co-binding with other TFs, rather than by changes in their gene expression. Thus, we assessed in parallel the DNA accessibility of the same samples in order to gain a global picture of putative TF binding patterns and enabling integration with the expression of regulated target genes on RNA level. Due to the limited number of primary, *ex vivo* isolated GM-CSF positive human cells, we employed a recently described highly sensitive method, assay for transposase-accessible chromatin using sequencing (ATAC-seq) [28] to study DNA accessibility from 50,000 cells per sample. As a control, we used the respective GM-CSF depleted (“GM-CSF negative”) fraction derived from the capture assay procedure. Since GM-CSF positive cells may differ from GM-CSF negative CD4 T cells simply by containing largely reduced fractions of naïve cells, we furthermore undertook RNA-seq and ATAC-seq profiling of several control cell populations that is, naïve CD4 T cells and memory CD4 T cells. As an additional control, we studied bulk CD4 T cells without capture assay procedure.

This study hence reveals molecular patterns specific for GM-CSF positive CD4 T cells or shared with memory, naïve or bulk CD4 cells. To our knowledge, this is the first study of global molecular signatures of GM-CSF positive CD4 T cells derived *ex vivo* without restimulation. Besides serving as a control we furthermore provide a novel resource of ATAC-seq and RNA-seq data of human primary

naïve, memory and bulk CD4 T cells from several human healthy donors. A large body of knowledge exists on molecular signatures and regulation of human naïve and memory T cells [29]. This encompasses large consortium efforts to map human memory and naïve CD4 T cell subsets' transcriptomes and epigenomes including chromatin accessibility [30], albeit these authors did not use the ATAC-seq method. Recently, few reports using ATAC-seq for T cells have been published and the impactful results support the power of the methodology. Many of these studies focus on murine CD8 T cell differentiation and exhaustion [31–35], and several recent studies also comprise ATAC-seq on CD4 T cells [10,28,36–40] . However, none of these works studied all the types of CD4 T cell subsets we analyzed here.

Through interpreted, integrative analysis of mRNA expression and DNA accessibility data from primary human CD4 T cell subsets we provide novel gene regulatory networks underlying GM-CSF production as well as the memory phenotype in human CD4 T cells, and we propose novel key TFs regulating these cells. The enrichment of the identified genes for human immune system diseases and specifically MS for GM-CSF positive cells underlines the clinical relevance of our data, which may be exploited in a multitude of basic and applied immunology studies in the future.

MATERIALS AND METHODS

Ethics Statement

Peripheral blood mononuclear cells were freshly isolated from anonymized healthy donor buffy coats purchased from the Karolinska University Hospital (Karolinska Universitetssjukhuset, Huddinge), Sweden. Research was performed according to the national Swedish ethical regulations (ethical review act, SFS number 2003:460).

Experimental methods

PBMC and T cell isolation. Human peripheral blood mononuclear cells (PBMCs) were isolated using Ficoll-Paque gradient centrifugation from buffy coats according to standard procedures. In brief, buffy coats were diluted in PBS, layered on Ficoll-Paque (GE healthcare) and centrifuged at 1200xg for 20 min without break. Subsequently, the PBMC ring was collected. PBMCs were washed with PBS (450xg, 10 min) and monocytes were depleted by plastic adherence in RPMI 1640 medium containing GlutaMAX (LifeTechnologies, Thermo Fisher Scientific) and 10% (v/v) heat inactivated fetal bovine serum (FBS; Gibco Performance Plus certified; Thermo Fisher Scientific) for 60-80 minutes. Platelets were removed by centrifugation (200xg, 5-10 min, 20°C; 4 times). Subsequently, human naïve,

146 memory, and total CD4 T cells were isolated negatively ("untouched") in parallel by magnetic
147 activated cell sorting (MACS) from each donor. The following MACS kits were used according to the
148 instructions from the manufacturer (Miltenyi Biotec): human naïve $CD4^+$ T cell isolation kit II, human
149 memory $CD4^+$ T cell isolation kit II, and human $CD4^+$ T cell isolation kit II. The purity of naïve,
150 memory, and total CD4 T cells was controlled by flow cytometry (see below). Cells were counted with
151 the Countess Automated Cell Counter (Invitrogen) and viability (determined by Trypan blue stain) was
152 96.5±1.5% (mean±SD). T cells were cultured at 37°C and 5% CO_2 in serum-free X-Vivo 15 medium
153 (Lonza) supplemented with GlutaMAX (Gibco), unless otherwise stated. T cells were rested overnight
154 before sample preparation for RNA-seq, ATAC-seq, and GM-CSF secretion assay (see below). 6
155 healthy male anonymized donors (age 35.7±8.5 years, mean±SD; range 22-44 years) were used for
156 molecular profiling.

157

158 *GM-CSF secretion assay ("capture assay").* Total CD4 T cells were isolated and rested as above,
159 before "capturing" GM-CSF producing cells using the GM-CSF Secretion Assay - Cell Enrichment and
160 Detection Kit (PE), human (Miltenyi Biotec) according to the manufacturer's instructions with the
161 following modifications and details. 65-150x10^6 (98.3x10^6±33.7x10^6, mean±SD) unstimulated purified
162 CD4 T cells were centrifuged (300xg, 10 min), X-Vivo 15 medium was removed completely, and cells
163 were washed with 15 ml MACS buffer (0.5 % human serum albumin (HSA) and 2 mM EDTA in PBS,
164 4°C). Cells were resuspended in ice-cold RPMI 1640 medium including GlutaMAX (LifeTechnologies,
165 Thermo Fisher Scientific) and 10% (v/v) FBS (Gibco Performance Plus certified, heat inactivated;
166 Thermo Fisher Scientific). GM-CSF catch reagent was added, mixed, and incubated for 5 minutes on
167 ice. The GM-CSF secretion period was performed by adding pre-warmed (37°C) RPMI 1640 medium
168 (containing GlutaMAX and 10% FBS) to a cell density of 1x10^6 cells/ml, under continuous rotation (10
169 rpm orbital mixing) of the cells for 45 minutes at 37°C/5% CO_2. Labeling cells with GM-CSF Detection
170 Antibody (Biotin) and Anti-Biotin-PE, and magnetic labeling with Anti-PE MicroBeads UltraPure, were
171 performed as per instructions and cells were washed in 50 ml MACS buffer. Cells were resuspended
172 in 3 ml MACS buffer, cell suspensions were filtered with 30 µm Filcon strainer (BD Biosciences) and
173 magnetic separation was performed on LS columns (Miltenyi Biotec) according to standard protocols.
174 For Donor A, the GM-CSF+ eluate was passed over a second MS column following the LS column
175 procedure, but since a second column did not increase purity but led to loss of cells (data not shown),
176 for all other donors the GM-CSF+ eluate from the first LS column was used for subsequent analyses.
177 The GM-CSF- fraction from the flow-through was passed over a second column (except for Donor A)
178 to increase purity of the negative fraction. Yield of GM-CSF+ cells was 1.2±0.6% of CD4 T cells
179 (mean±SD) and purity was controlled by flow cytometry (see below).

180

181 *Flow cytometry.* Purity of MACS-isolated naïve CD4 T cells, memory CD4 T cells and total CD4 T
182 cells was verified by surface staining using the following antibodies (all against the human proteins):
183 anti-CD4-PerCP (clone SK3, BD Biosciences), CD45RA-FITC (clone T6D11, Miltenyi Biotec),

184 CD45RO-PE (clone UCHL1, BD Biosciences), CD19-APC (Clone HIB19, BD Biosciences), CD8-
185 eFlour450 (clone OKT8, eBiosciences). Staining was performed in the dark with antibody dilutions in
186 FACS buffer (PBS with 0.5% HSA) for 15 min at 20°C. Single stained PBMC or T cell samples were
187 used as compensation controls. Cells were washed once with PBS, resuspended in FACS buffer and
188 acquired on a CyAn ADP 9-Color Analyzer (Beckman Coulter). Compensation was performed with the
189 CyAn software (Summit) tool. Purity of GM-CSF+ and GM-CSF- samples after GM-CSF secretion
190 assay from bulk CD4 T cells was performed by measuring the fraction of PE-labeled (GM-CSF
191 positive) cells in a CyAn ADP 9-Color Analyzer (Beckman Coulter). Cell purities and gating strategy
192 are shown in Figure 1 and Supplementary Figure 1. Flow cytometry data analysis and visualization
193 was performed using the FlowJo software v7.6.5 (Tree Star), and exported percentage values were
194 plotted in GraphPad Prism v7.02 (GraphPad Software, Inc.).

195

196 *ATAC-seq sample preparation.* Nuclear isolation, tagmentation and PCR amplification was carried out
197 according to Buenrostro *et al.* [28]. In brief, 50000 cells per sample were transferred to 0.2 ml tubes,
198 centrifuged (500xg, 6 min, 4°C) and the supernatant was removed. Cells were washed with PBS
199 (500xg, 6 min, 4°C) and lysed in lysis buffer (10 mM Tris-HCl, pH 7.4; 10 mM NaCl; 3 mM $MgCl_2$;
200 0.1% IGEPAL CA-630) to isolate nuclei. 3 technical replicates (50000 cells each) were processed in
201 parallel for each sample (except donor A, 2 technical replicates). Nuclei were washed in PBS (500xg,
202 6 min, 4°C), and resuspended in Transposase Reaction mix. Transposition was carried out at 37 °C
203 for 30 minutes, followed by clean up using Qiagen Minelute Reaction Clean Up kit according to the
204 manufacturer's instructions (Qiagen). PCR amplification using reagents from Nextera DNA Sample
205 Preparation Kit (Illumina) and barcoding of replicates was performed with reaction conditions and
206 index primers as described in [28]. 20 different index primers were used and distributed across
207 replicate samples and donors in a balanced way to control for potential batch effects. The PCR
208 product was cleaned up using Qiagen Minelute Reaction Clean Up kit. Subsequently, gel size
209 selection was performed by gel electrophoresis [1.8% (w/v) certified low-melt agarose (Bio-Rad) in 1x
210 UltraPure Tris Acetate-EDTA (TAE) buffer (Invitrogen, Thermo Fisher Scientific), with a free well
211 between each of the samples] and DNA in the size range of 150-230 bp was excised using surgical
212 blades. Replicate samples were allocated to the gels in a balanced way regarding donor and
213 experimental condition to control for potential batch effects. Resulting DNA libraries were purified
214 using Qiagen Minelute Gel Extraction Kit according to the manufacturer's instructions (Qiagen).

215 Size distribution of ATAC-seq sequencing libraries was determined on a 2100 Bioanalyzer instrument
216 (Agilent Technologies) using the Agilent DNA High Sensitivity Kit according to the manufacturer's
217 instructions. Libraries were quantified by real-time PCR on a StepOne plus detector system (Applied
218 Biosystems) using the KAPA Library Quantification Kit (KAPA Biosystems). Sequencing was
219 performed on an Illumina HiSeq 2500 instrument (Illumina) with single-read setting and read length
220 42 bp.

221

222 *RNA-seq sample preparation.* Cells were centrifuged (1000xg, 5 min, 20°C), washed with PBS, and
223 lysed in QIAzol Lysis Reagent (Qiagen) by vortexing and incubating for 5 min at 20°C. Lysates were
224 stored at -80°C until RNA extraction. RNA was extracted with the miRNeasy Micro Kit (Qiagen)
225 according to the manufacturer's instructions. RNA concentration was determined on a Nanodrop 2000
226 spectrophotometer (Thermo Fisher Scientific) and RNA quality was controlled on a 2100 Bioanalyzer
227 instrument (Agilent Technologies) using an Agilent RNA 6000 Pico Kit. RNA integrity numbers (RIN)
228 were 8.5±0.3 (mean±SD). Libraries were prepared in one batch with the TruSeq Stranded mRNA HT
229 kit (Illumina) with Dual Indexing adapters and Ambion ERCC Spike-In Control (Thermo Fisher
230 Scientific). Samples were allocated to sequencing indexes and -lanes in a balanced fashion to control
231 for potential batch effects (7-8 samples/lane). Library concentration was determined on a Qubit 2.0
232 Fluorometer (Thermo Fisher Scientific). Library size and quality were measured on a 2100
233 Bioanalyzer instrument (Agilent Technologies) using an Agilent High Sensitivity DNA Kit. Libraries
234 were quantified with the KAPA Library Quantification Kit (KAPA Biosystems). Sequencing was
235 performed on a HiSeq 2500 Sequencing Platform (Illumina; High Output Run) with 76 nt paired-end
236 reads.

237

238 **Computational Methods**

239

240 *Preprocessing of sequencing data.* BCL base-call files were demultiplexed and converted to FASTQ
241 files using bcl2fastq version 2.17.1.14. For quality control FastQC version 0.11.5 was used.

242 FASTQ files from the ATAC-seq experiments were trimmed from adapters and low quality bases
243 using scythe version 0.991. and sickle version 1.33. FASTQ files from the ATAC-seq experiments
244 were aligned to the human genome version hg38 using bowtie version 2.3.0 with the '--very-sensitive'
245 option. After alignment, BAM files of ATAC-seq experiments were filtered to eliminate: duplicates
246 (samtools rmdup), alignments with a mapping score below 10, and alignments that are not mapped to
247 chromosome 1-22, chromosome X or Y. ATAC-seq peaks were called using the 'findPeaks' script
248 from the HOMER suite (version 4.9.1) with the '-style factor' option [41]. ATAC-seq peaks were
249 assigned to a gene using the 'annotatePeaks' script from the HOMER suite.

250 FASTQ files from the RNA-seq experiments were aligned using STAR version 2.5.2b. Indexes for
251 RNA-seq alignment were created using the gencode version 25 annotation file. RNA-seq alignment
252 was run with STAR's built-in adapter trimming option ('--clip3pAdapterSeq
253 AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
254 AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA') and its built-in counting option ('--
255 quantMode'). Only genes with more than 1 count per million in at least 3 samples were included in the
256 downstream RNA-seq analysis.

257 *Generation of consensus peak-set.* In order to generate a set of comparable features (genomic
258 regions) for read-counting and quantifying differential accessibility from the ATAC-seq data, a set of
259 consensus peaks was generated in two subsequent steps. (1) Generation of consensus peaks on the
260 technical replicate level: First, the peaks that appeared in at least two technical replicates (out of a
261 total of three, except one donor with a total of two technical replicates) with at least 75% reciprocal
262 overlap were selected. Then these selected regions were partitioned into disjoint non-empty subsets
263 so that each element is contained in precisely one subset. Only the partitions appearing in at least two
264 replicates were retained, and afterwards adjacent regions were merged. A bed file resulting from
265 these steps is herein referred to as a sample. (2) Next, all bed files containing each set of technical
266 replicate level consensus regions were concatenated and the presence of each region was counted
267 within each experimental sample; one occurrence corresponds to one donor (biological replicate) and
268 one cell type (experimental condition). Only the regions appearing in at least four samples were kept
269 (n=5 is the number of biological replicates in the smallest group regarding experimental condition).
270 Regions having a distance of 42 bases or less between them were subsequently merged (the number
271 42 corresponds to the sequencing read length in bases). Afterwards, reads were counted using the
272 featureCounts tool [42] with the criterion that at least half of the read had to overlap with a feature to
273 be assigned.

274 *Calculation of differential expression and differential accessibility.* Differential expression and
275 accessibility was calculated using the edgeR [43] library version 3.18.1 from Bioconductor. Donors
276 (biological replicates) and cell types (experimental conditions) were used as explanatory variables in
277 the generalized linear models. ATAC-seq data were normalized to length and GC content by
278 conditional quantile normalization (CQN) [44]. Comparisons were made between GM-CSF positive
279 CD4 T cells *versus* GM-CSF negative CD4 T cells, or between memory *versus* naïve CD4 T cells (see
280 Figure 1A). The cutoff to call differentially expressed genes (DEGs) or differentially accessible regions
281 (DARs) was FDR<0.05 and >25% fold change (in the direction of either up- or down-regulation that is,
282 either 1.25 or 0.75 fold change).

283

284 *Footprinting.* Footprinting was carried out using the Wellington algorithm [45], i.e. the
285 wellington_footprints.py script from the pyDNase library version 0.2.5 with the following settings: -fp
286 6,41,1 -sh 7,36,1 -fdr 0.01 -fdriter 100 -fdrlimit -30 -A. The footprint occupancy score (FOS) for each
287 footprint was calculated using the pyDNase library as described in [46]. For subsequent network
288 reconstruction, we considered only footprints with a FOS smaller than the following threshold:
289 median(FOS) + [median(FOS) – minimum(FOS)].

290

291 *Network reconstruction.* A directed network was reconstructed by combining information from the
292 ATAC-seq and RNA-seq data in two subsequent steps. (1) Identifying source nodes: Peaks were
293 ranked based on their combined measure of significance and direction of differential accessibility [-

294 $\log_{10}$(FDR) x sign($\log_2$(fold change))]. Peaks containing footprints were scanned for TF binding motifs
295 using the TRANSFAC database [47]. An enrichment score was calculated to identify TFs with binding
296 sites enriched in differentially accessible peaks, using tools similar to gene set enrichment analysis
297 (GSEA) [48]. In detail, random sampling was performed on the ranked list of peaks to assess how
298 probable it is to observe at least the same enrichment by chance (P value) [49]. After multiple test
299 correction, TFs with FDR<0.05 were selected and the normalized enrichment score (NES) was
300 obtained. Only those source nodes (TFs) that were detectably expressed on RNA level (according to
301 a minimal RNA-seq filtering rule) were considered. Of these, most fell into the class of highly
302 expressed genes (HEGs) according to [50]. (2) Identifying target nodes: Target nodes are defined as
303 peaks with an assigned gene. The selection criteria were that (i) the peak contains a footprint with a
304 binding motif of the source node (TF) and (ii) the peak and/or the assigned gene has to be
305 differentially accessible or differentially expressed respectively.

306 To assign an importance measure to the source nodes in networks generated as above, the
307 PageRank [51] of the network nodes was calculated after inverting the directionality of all edges in the
308 network (only for the purpose of this computation). After this computation, the nodes with high
309 PageRank values (higher than the 99th percentile of all node values within a given network) were
310 selected from both the GM-CSF and memory network, and afterwards their values were investigated
311 in each of the two networks.

312

313

314 **RESULTS AND DISCUSSION**

315

316 GM-CSF positive CD4 T cells are enriched in MS patients and play a crucial role in EAE; nevertheless
317 the factors driving and markers defining those cells are largely unknown. To better understand the
318 features and regulatory networks of GM-CSF positive CD4 T cells, we therefore studied the
319 transcriptional profiles and chromatin accessibility of these cells. *In vitro* differentiated GM-CSF
320 producing cells comprise several subsets [27] and are likely to differ from those generated *in vivo*.
321 Further, *ex vivo* re-stimulation with strong artificial stimuli such as PMA and ionomycin - which is
322 usually necessary to reach sufficient signal strength for detection by intracellular cytokine staining -
323 drastically alters the transcriptome of T cells. Hence, we aimed to isolate GM-CSF producing cells *ex*
324 *vivo* in an as much as possible unmanipulated state by GM-CSF secretion assay, “capturing” and
325 isolating those cells that actively secrete GM-CSF (experimental setup, see Figure 1A), here defined
326 as GM-CSF positive cells. The capture assay was performed starting from highly purified CD4 T cells
327 derived from human peripheral blood (purity 97.3±0.6 %, mean±SEM, Supplementary Figure S1A). As
328 controls, we used the respective GM-CSF negative fraction from the isolation procedure, as well as
329 the bulk CD4 T cells before any capture assay procedure. The latter should, given the low fraction of
330 GM-CSF positive cells, be very similar to the GM-CSF negative fraction and hence allows for

331 estimation of the effects arising from the capture assay procedure. The purity of GM-CSF positive and
332 GM-CSF negative fractions was assessed by flow cytometry (Figure 1B) and the yield of isolated GM-
333 CSF positive cells was 1.2±0.6 % (mean±SD) of CD4 T cells. To measure the transcriptome and DNA
334 accessibility from limited cell numbers, we employed highly sensitive next-generation sequencing
335 (NGS) methods (RNA-seq and ATAC-seq respectively). Since cytokine secreting cells may differ from
336 naïve cells due to a memory-like phenotype, and a large fraction of CD4 T cells are naïve
337 (mean±SEM 42.3±5.8 % in the donors here; see Supplementary Figure S1A), we further profiled
338 highly purified naïve and memory CD4 T cells from the same donors (Figure 1A and Supplementary
339 Figure S1B-D).

340 Altogether, we obtained DNA accessibility and transcriptome data from highly purified *ex vivo* derived
341 human naïve CD4 T cells, memory CD4 T cells, bulk CD4 T cells and well as GM-CSF positive and
342 corresponding GM-CSF negative CD4 T cells. To enable paired analysis within a donor, these cell
343 populations were isolated in parallel within a donor; for 6 donors in total. DNA accessibility and mRNA
344 data were obtained in parallel from the same samples allowing for matched integration of the data.

345

346 **Unique and shared DNA accessibility and gene expression signatures of GM-CSF positive and**
347 **memory CD4 T cells.**

348 We studied the above-described five different CD4 T cell populations by RNA-seq and ATAC-seq. To
349 minimize potential batch effects due to technical factors, the library preparations and sequencing runs
350 were designed in such a way that donors, cell populations, and (for ATAC-seq) technical triplicates
351 were distributed in a balanced fashion. It is also worth noting that only donors of the same gender
352 were studied here (male, age 35.7±8.5 years, mean±SD), which may be important since it was
353 recently shown that gender was the largest source of variation explaining chromatin accessibility in
354 primary human CD4 T cells measured by ATAC-seq [36]. That study further discovered novel
355 elements escaping X chromosome inactivation and affecting immune genes [36]. To assess which
356 factors explained most of the variability between the samples under study here, we performed
357 principle component analysis (PCA). Indeed, for both data types there was a grouping of the samples
358 based on the cell subset, outweighing donor or experimental variation (Figure 2A, B) and confirming
359 the quality of our samples and data. Notably, for RNA data, the cell populations were generally more
360 distinct from each other than for DNA accessibility data. However, the GM-CSF positive and
361 corresponding GM-CSF negative fraction appeared relatively similar to each other in the PCA
362 performed on RNA data, while PCA results from ATAC-seq data were closer to the expected pattern
363 that is, bulk CD4 T cells appearing “between” GM-CSF positive and GM-CSF negative populations
364 (Figure 2A, B). The difference between RNA-seq and ATAC-seq data with respect to separation of
365 GM-CSF positive and negative cells may indicate that the capture assay procedure imposes distinct
366 changes on the transcriptome, highlighting the importance of using correspondingly treated controls to
367 determine differential expression. In contrast, changes in DNA accessibility appeared more robust
368 towards changes due to the experimental procedure at least within the experimental time frame under

369 study, although the distinction of the other groups was generally less apparent with ATAC-seq data.
370 Because the first two PCs only explained about 50% of the variation in the data, we also used another
371 dimensionality reduction method to explore the sample-to-sample relationships, namely t-distributed
372 stochastic neighbor embedding (t-SNE) [52]. t-SNE, in contrast to PCA, is a non-linear dimensionality
373 reduction algorithm, and it is suited for capturing local and global relationships at the same time. The
374 t-SNE results (Figure 2C, D) generally confirmed the results of the PCA analysis (Figure 2A, B) that is,
375 the groups (cell types) being more distinct in RNA data than ATAC-seq data, with the exception of
376 GM-CSF positive and GM-CSF negative cells.

377 We defined significantly differentially accessible DNA regions (DARs) and significantly differentially
378 expressed genes (DEGs) in GM-CSF positive CD4 T cells or in memory CD4 T cells. To do so, we
379 specifically analyzed the signatures of GM-CSF positive *versus* GM-CSF negative CD4 T cells, as
380 well as the profiles of memory *versus* naïve CD4 T cells (Figure 1A and Figure 3A, B). We used
381 generalized linear models based on the negative binomial distribution (edgeR) to determine
382 differential expression and accessibility respectively, and we called DEGs and DARs based on
383 combined FDR and fold change cutoffs. We called 16571 DARs in GM-CSF positive CD4 T cells
384 (compared to corresponding GM-CSF negative cells; Figure 3A) and 13705 DARs in memory CD4 T
385 cells (compared to naïve CD4 T cells; Figure 3A). On the transcriptome level, we called 124 DEGs in
386 GM-CSF positive and 5383 DEGs in memory CD4 T cells (Figure 3B). The relatively low number of
387 DEGs in GM-CSF positive cells is in agreement with the PCA and t-SNE data (Figure 2B, D) and may
388 suggest that combination with ATAC-seq data drastically improves the possibility to define molecular
389 signatures specific to GM-CSF positive *ex vivo* captured cells.

390 Next, we studied the DARs and DEGs in more detail. We first focused on the signatures of memory
391 CD4 T cells, which are well studied in the literature [29] and hence enabled to assess the biological
392 quality of our data, besides providing a new NGS data set of human primary memory and naïve CD4
393 T cells. DARs and DEGs defined in memory cells are shown in Supplementary Figure S2, along with
394 their molecular patterns in the other cell types under study. We next extracted those memory-specific
395 DARs that were assigned to a gene from a list of genes known to be involved in T cell memory as
396 compiled by Polansky and colleagues [30]. Several of the DARs in memory cells were assigned to
397 such known memory-associated genes, most of those in the promoter region (Figure 3C).
398 Furthermore, we assessed a selected subset of memory related genes that were shown to be up- or
399 down-regulated on RNA level in memory T cells [30]. The majority of these genes were DEG in
400 memory cells in our data, notably up- or down-regulated almost exclusively (36 of 37 studied genes;
401 97%) in the expected direction (Figure 3D), validating our data.

402

403 **The molecular signature of GM-CSF positive CD4 T cells.**

404 Next, we focused on the signatures of GM-CSF positive cells by studying the respective DARs and
405 DEGs in more detail. All the DARs in GM-CSF positive cells are displayed in Figure 4A with the color

406 scale representing accessibility. A substantial fraction of DARs in GM-CSF positive cells displayed a
407 similar pattern in memory cells, while naïve cells and GM-CSF negative cells were most distinct from
408 GM-CSF positive cells (Figure 4A). In bulk CD4 T cells, DARs defined in GM-CSF positive cells
409 showed heterogeneity between donors (Figure 4A), which may reflect the variability in the fraction of
410 memory and naïve cells within bulk CD4 T cells depending on the donor (Supplementary Figure S1A).
411 Importantly, several DARs displayed a unique accessibility pattern in GM-CSF positive cells distinct
412 from other cell populations under study (Figure 4A). These data show that GM-CSF positive cells can
413 be assigned a specific pattern of accessible DNA regions that distinguish them from other CD4 T cell
414 subsets, and that may contribute important information about regulation of GM-CSF positive cells.
415 Different DNA accessibility can functionally affect the status of a cell by, for example, modifying
416 expression of genes regulated by these regions.

417 Thus, we then studied the DEGs defined in GM-CSF positive cells by analyzing their expression in
418 GM-CSF positive cells along with the other CD4 T cell subsets under study. Like observed for the
419 DARs, DEGs in GM-CSF positive cells shared a large part of the RNA signature with memory cells,
420 but also displayed distinct patterns, and differed largely from naïve cells, GM-CSF negative and bulk
421 CD4 T cells (Figure 4B). Despite the general similarity of cells treated with the capture assay
422 regarding the global transcriptome (Figure 2B, D), sub-setting on the DEGs defined between GM-CSF
423 positive and negative cells with stringent statistical cutoffs visualized clearly the differences in these
424 "signature genes" for those cell types (Figure 4B), albeit this is not surprising due to the respective
425 pre-selection of the genes. However and importantly, sub-setting on these genes that were defined as
426 DEGs of GM-CSF positive *versus* corresponding negative cells (that is, without considering the bulk
427 CD4 T cell samples) also clearly showed the expected similarity of bulk CD4 T cells and GM-CSF
428 negative CD4 T cells (Figure 4B) that was not apparent in the RNA-seq PCA on all genes (see above;
429 Figure 2B, D), confirming that the selected GM-CSF positive cell "signature" DEGs are likely not
430 affected by the capture assay and column procedure. DEGs in GM-CSF positive cells comprised
431 multiple genes with a well-known role in T cells such as *EGR2, IFNG, CTLA4, CXCL8* and *CXCR5*
432 along with multiple genes with an unknown role in T cells and potentially regulating GM-CSF positive
433 cells (Figure 4B). *CSF2RB* (encoding for the high affinity receptor subunit for IL-3, IL-5 and GM-CSF)
434 was among the DEGs in GM-CSF positive cells, suggesting a possible feedback loop; *CSF2RB*
435 expression being lower in GM-CSF positive than negative cells excludes a technical artifact of
436 isolation of cells binding GM-CSF through this receptor (Figure 4B). Albeit not passing the
437 significance threshold for being called a DEG in GM-CSF+ cells, relatively high expression of *IFNG* in
438 GM-CSF positive cells (data not displayed) is in accordance with our and other's previous findings
439 from flow cytometry of T cells from healthy donors as well as MS patients demonstrating preferential
440 co-expression of IFN-γ and GM-CSF in some, but not all GM-CSF producing human CD4 single T
441 cells or clonal populations thereof [11,14,26,27,53]. Also in line with the majority of these studies of
442 human T cells, *IL17A* and *IL17F* expression was below detection limit in GM-CSF positive cells (as
443 well as any other cell population under study; data not displayed). It should be noted that *CSF2*
444 mRNA, which encodes for GM-CSF, was lowly expressed in all samples (RNA-seq counts close to
445 detection limit), which may be explained by rapid mRNA decay conferred by the adenine and uridine-

446 rich elements (ARE) in the GM-CSF promoter – AREs in fact have been discovered in the *CSF2* gene
447 which codes for a particularly unstable transcript [54–56]. Importantly, considering the instability of
448 *CSF2* mRNA and relatively low expression levels, our approach of isolating GM-CSF protein secreting
449 cells is likely to be more suitable to define signatures of *ex vivo* derived GM-CSF positive cells, as
450 opposed to for example single cell RNA-seq of mixed T cell populations.

451 Since to our knowledge, there is no other data set available that studied the signatures of purified
452 GM-CSF secreting cells, the RNA expression signatures defined in the GM-CSF positive cells in this
453 study could not be validated externally in an independently published transcriptome data set.
454 Nevertheless, we strived to confirm the expression signature from GM-CSF positive captured cells by
455 comparing with a completely independent data set and experimental setup. To this end, we asked
456 whether the RNA expression pattern of captured GM-CSF positive T cells generally agreed with the
457 respective proteins in GM-CSF positive T cells defined by intracellular cytokine staining. To explore
458 overlap of as many as possible markers, we studied a mass cytometry (CyTOF) data set [57] which
459 comprises staining of human PBMCs with CD4 and GM-CSF along with other markers measured on
460 protein level. In pre-gated CD4 positive T cells of this CyTOF data set, we gated on GM-CSF+ cells
461 and GM-CSF- cells and determined the relative expression of other available protein markers in these
462 populations. We compared the up- or down-regulation of these protein markers in gated GM-CSF+
463 *versus* GM-CSF- cells [$\log_2$(fold change) of the median signal intensity of the two cell populations]
464 with the up- or down-regulation of the corresponding protein-coding transcripts of these markers in
465 isolated GM-CSF positive *versus* GM-CSF negative cells from our data [$\log_2$(fold change) of mRNA
466 expression]. As a result, 16 of the 24 markers (two thirds) had a fold change with identical directions
467 in the CyTOF and RNA-seq data (Figure 4C). According to a binomial distribution, the probability of
468 observing this or greater concordance between the two datasets by chance is 7.6%. This needs to be
469 acknowledged considering that those markers that do not concur might be regulated by protein
470 internalization from the surface, such as well-known for CD3 [58] that was down-regulated in CyTOF
471 data but barely affected on RNA level. Also, the total abundance of certain proteins may be regulated
472 on post-transcriptional level, as transcript levels cannot always predict protein abundance [59,60].
473 Therefore, we concluded that the expression profile of the GM-CSF positive captured T cells matches
474 well with the profile of independently characterized human GM-CSF positive T cells.

475

476 **GM-CSF positive CD4 T cell transcript signatures and chromatin accessibility are associated**
477 **with autoimmune diseases, especially MS.**

478 Next, we asked whether the gene expression pattern of GM-CSF positive cells was enriched for any
479 diseases by exploring the Open Targets platform [61]. We ranked the detected genes based on their
480 differential expression in GM-CSF positive *versus* negative cells using the $-\log_{10}(\text{FDR}) \times \text{sign}(\log_2(\text{fold}$
481 $\text{change}))$ function for ranking, and calculated enrichment scores for diseases. Indeed, among the few
482 significantly enriched diseases (FDR<0.05), the autoimmune diseases MS and rheumatoid arthritis
483 were represented (Figure 5A). Notably, for both diseases GM-CSF targeting is in clinical trials [15],

484 supporting the relevance of our data. The data also showed enrichment for several other diseases
485 related to the immune system, infection or metabolism, although it should be noted that some of these
486 contained only few elements (genes) and may thus be less relevant than MS or rheumatoid arthritis,
487 which comprised a large number of genes (Figure 5A and Supplementary Table S1A). When
488 performing the same enrichment analysis for the ranked gene list from memory *versus* naïve CD4 T
489 cells, a large number of diseases were significantly enriched including many diseases involving the
490 immune system, as expected (Supplementary Table S1B).

491 Due to the relevance of GM-CSF positive T cells in MS, we studied in more detail whether DEGs
492 identified in GM-CSF positive cells displayed altered expression in MS patients' T cell samples. T cells
493 recognizing peptides derived from myelin proteins as autoantigen and being activated and migrating
494 to the CNS are thought to be crucial mediators of inflammation in MS [62]. Hafler and colleagues have
495 performed RNA-seq of expanded CD4 T cells derived from MS patients and healthy controls, with two
496 subsets of samples each containing those auto-reactive to antigenic peptides derived from myelin
497 (tetramer positive) *versus* tetramer negative T cells [26]. The authors discovered that myelin reactive
498 T cells from patients with MS displayed strongly enhanced production of IFN-γ, IL-17 and GM-CSF
499 compared to those isolated from healthy controls, which instead secreted more anti-inflammatory IL-
500 10 [26]. We studied whether these cells would also display altered expression of the genes we
501 defined as signature genes of GM-CSF positive cells. Indeed, a subset of these genes was down- and
502 another subset was up-regulated in myelin-reactive T cells from MS patients (Figure 5B), potentially
503 identifying genes relevant to the disease pathogenesis. Here, the T cells displayed detectable *CSF2*
504 mRNA encoding for GM-CSF, perhaps due to the *in vitro* stimulation and expansion of these T cells.
505 Remarkably, only myelin reactive cells from MS patients expressed high levels of *CSF2* (Figure 5B).
506 Interestingly, this *CSF2* expression pattern strongly resembled the expression patterns of several of
507 the GM-CSF positive T cell signature DEGs defined here, namely *DUSP5*, *IL1R1*, *KDM6B*, *EGR1* and
508 *EGR2* (Figure 5B) which may be interesting candidates to explore in the future regarding their role in
509 GM-CSF positive T cells and MS. While this analysis confirmed the relevance of the GM-CSF positive
510 T cells' RNA-seq data in MS, we next analyzed whether also the ATAC-seq data could reveal
511 information on those DNA regions most relevant for GM-CSF positive T cells in the context of MS. To
512 do so, we asked whether the peaks called from our ATAC-seq analysis contained any SNPs
513 associated with MS. Interrogating the current best list of non-MHC MS susceptibility variants
514 comprising 110 established risk variants from the International Multiple Sclerosis Genetics
515 Consortium [63,64], we identified two SNPs mapping to the accessible DNA peaks (considering all
516 consensus peaks) from our study. These SNPs were assigned to the protein-coding genes *Regulator*
517 *Of G Protein Signaling 1* (*RGS1*) and *Engulfment And Cell Motility 1* (*ELMO1*) genes respectively
518 (Figure 5C). Importantly, the regions containing these SNPs were significantly differentially accessible
519 (FDR<0.05) in GM-CSF positive *versus* GM-CSF negative cells (Figure 5C), suggesting a putative
520 role of these regions in T cell mediated MS pathogenesis.

521

522 **Relationship of differential gene expression and chromatin accessibility.**

523 Since chromatin accessibility can directly affect gene expression, we next combined the ATAC-seq
524 and RNA-seq data, aiming to identify key TFs that may bind to open chromatin regions and hence
525 affect the expression of their target genes. The methods for calling consensus peaks are not trivial.
526 Therefore, we first confirmed that the genes assigned to the consensus peaks defined in this study
527 were enriched for several pathways involved in immune regulation including T cell receptor signaling
528 and Th subset differentiation, as well as for immune-related diseases including MS, RA and other
529 autoimmune disease (Supplementary Table S2A, B). Furthermore, the distribution of the consensus
530 peaks (open chromatin regions) defined in our data showed an enrichment for being located in CpG
531 islands, promoters, 5' UTRs, exons and protein-coding regions (Figure 6A), suggesting that the
532 ATAC-seq data generated here should be well suited to identify TF binding and the expression of
533 corresponding target genes. We first studied the relationship of RNA and chromatin data on a global
534 level: We considered the genes that were both detected on RNA level and also had an ATAC-seq
535 peak assigned to them. We ranked these genes using the $-\log_{10}(\text{FDR}) \times \text{sign}(\log(\text{fold change}))$
536 function separately for the RNA-seq and ATAC-seq data. Next, we visualized the correlation between
537 these two ranks for each contrast (Figure 6B). Considering only the genes significantly changing
538 (FDR<0.05) in the given contrast and assigning them to up- and down-regulated categories, we
539 detected more than random coincidence in the direction of the change between the RNA-seq and
540 ATAC-seq data (using Fisher's exact test with Monte Carlo simulation; Figure 6B). Increased
541 openness of the chromatin was associated with increased expression of the corresponding gene's
542 RNA for the majority of genes. Less accessible chromatin also coincided with low expression of the
543 corresponding gene in many cases, although a substantial fraction of lowly accessible regions also
544 displayed high expression of the corresponding gene (Figure 6C).

545

546 **Identification of key TFs linked to the signatures of GM-CSF positive and memory CD4 T cells.**

547 To identify potential TFs that may establish the gene signatures of GM-CSF positive (*versus* GM-CSF
548 negative) and memory (*versus* naïve) CD4 T cells, we scanned the consensus peaks for footprints
549 and subsequently we scanned the identified footprints for TF binding motifs. For motif scanning we
550 used the TRANSFAC database [47] that contains experimentally validated binding sites, consensus
551 binding sequences (positional weight matrices) and regulated genes of eukaryotic TFs. Confirming
552 the methodology used to identify footprints, they were enriched in differentially accessible regions as
553 expected for the corresponding cell type comparisons (Supplementary Figure S3). Next, we defined
554 the TFs whose binding sites were most enriched in peaks of GM-CSF positive or memory cells
555 respectively (ranked based on differentially accessibility), and identified about 20 TFs each that
556 passed the significance threshold (FDR<0.05) for enrichment (Figure 7A). These lists contained
557 several TFs with a well-known role in T cells, such as SATB1, YY1, ETS- and EGR-family TFs,
558 among other factors with a less defined role. Notably there was little overlap between the key TFs in
559 GM-CSF positive and memory cells, suggesting that our strategy may have identified key factors to
560 specifically define the GM-CSF positive T cell phenotype. The majority of these TFs were highly

561 expressed on RNA level (Figure 7B), falling within the class of highly expressed genes (HEGs) that
562 were suggested to be more functional than lowly expressed genes [50]. Interestingly however, most of
563 these TFs were not differentially expressed themselves in the cell population comparisons under
564 study (Figure 7C), suggesting that they may be regulated on post-transcriptional levels such as
565 protein phosphorylation and intracellular localization which is well-known for many TFs. Hence, with a
566 strategy exploring solely the transcriptome (or even proteome) without integrating ATAC-seq data,
567 several key TFs would likely be missed, while our integrative strategy combining RNA-seq and ATAC-
568 seq data successfully identified such TFs from limited amounts of primary human T cells.

569

570 **Integration of RNA-seq and ATAC-seq data identifies gene regulatory networks of GM-CSF**
571 **positive and memory CD4 T cells.**

572 Having identified key TFs in GM-CSF positive and memory CD4 T cells, we were interested whether
573 these factors regulated certain groups of target genes and whether several TFs may act together in a
574 concerted fashion. Some TFs regulated a large number of target genes, and clusters of TFs were
575 grouping together (Figure 8A, B). Exploring the co-binding between TFs in more detail showed that
576 certain groups of TFs bound together to the same targets (Figure 8C, D). These included the
577 TCF3:LEF1 pair that was always co-binding the same regions, as evident from the Memory/Naïve
578 CD4 T cell contrast (Figure 8B, D). TCF/LEF family proteins act downstream of the Wnt pathway and
579 it is well-known that they often display overlapping expression patterns and functional redundancy
580 [65]. In accordance with our data on human memory and naïve CD4 T cells, binding motifs for TCF
581 family members were recently also determined to be depleted in murine memory CD8 T cells as well
582 as human memory CD4 T cells using ATAC-seq [35,39].

583 Finally, by connecting the TFs to the genes assigned to their target regions, we generated a directed
584 gene regulatory network representing GM-CSF positive (*versus* negative) and memory (*versus* naïve)
585 CD4 T cells respectively (Figure 9). The source nodes were represented by the key TFs as selected
586 above (key TFs from Figure 7A, B) and target nodes were selected when the region was either
587 differentially accessible and/or the mRNA of the assigned gene was differentially expressed in the
588 respective cell population comparisons. Both the GM-CSF as well as the memory comparisons led to
589 similarly sized networks however, only some of the target and source nodes were shared between
590 both networks (Figure 9A, B).

591 To quantify the importance of the source nodes in each individual network as well as in the
592 comparison of both networks, we calculated scores based on the PageRank algorithm [51]. To do
593 such a calculation on a network where the outgoing edges from a TF reflect its importance, rather
594 than the incoming edges like in the original application of the algorithm, the edges in our networks
595 were inverted (solely for the purpose of calculating the PageRank values). The obtained PageRank
596 values give information on the importance of a node based on the number of other nodes it influences
597 (directly or indirectly). Some of the key TFs that were shared between both networks also had a

598 relatively high PageRank in both networks, such as the TFs E2F3, SPIB and GABPA (Figure 9A, B),
599 and may be general regulators of T cell activation/differentiation. Importantly, we also identified key
600 TFs with a relatively high PageRank in the GM-CSF T cell regulatory network whose PageRank value
601 in the memory/naïve network was 0. These TFs, namely ZNF35, SP2, SATB1, YY1, TEF and ZNF333
602 may, thus, be novel regulators specifically controlling the molecular signatures of GM-CSF positive
603 CD4 T cells (Figure 9A, B).

604

605

606 **CONCLUSIONS AND OUTLOOK**

607 In summary, we here provide interpreted high-quality novel RNA-seq and ATAC-seq data from
608 primary human CD4 T cells. We provide a network of key TFs and their targets representing GM-CSF
609 positive cells and memory CD4 T cells respectively. To our knowledge, this is the first study providing
610 signatures of GM-CSF secreting cells isolated by capture assay without restimulation, which are likely
611 as similar as possible to the state *in vivo* in humans. Very recently, others have performed a similar
612 approach, capturing IFN-γ and IL-17 producing cells from human donors, notably also from MS
613 patients [66], and the authors obtained important results on the transcriptional signatures of these
614 cells, including genes distinguishing clinically stable *versus* active MS patients. This raises interesting
615 prospects for studying GM-CSF positive cells specifically from MS patients as well by the methods
616 and comparing to the data resource provided here. This will have important clinical implications given
617 the likely important contribution of GM-CSF positive CD4 T cells to MS. Hu *et al.* [66] studied only a
618 limited set of 418 genes in the captured IFN-γ and IL-17 secreting cells, while the methods employed
619 in the present work enable genome-wide studies of DNA accessibility and RNA expression patterns.
620 Notably, the authors [66] stated that their approach may "miss other disease related" genes and that
621 "follow-up studies, such as RNA-sequencing analysis on TH17 and TH1/17 subsets isolated from MS
622 patients, may help to identify more disease-related genes" which we propose should be extended to
623 ATAC-seq and the GM-CSF producing Th subset as well. Future work should also address what
624 distinguishes patterns and contributions of GM-CSF *versus* IFN-γ and IL-17 producing T cells
625 particularly in MS patients. Although not from MS patients, another recent report [67] provides RNA-
626 seq data of such captured cells from healthy donors however, PMA and ionomycin restimulation
627 (which by itself drastically alters the T cell transcriptome) render it difficult to directly compare these
628 data to ours. Interestingly, recent ATAC-seq data from murine GM-CSF *versus* IFN-γ and IL-17
629 producing T cells upon *in vitro* stimulation suggests that while all subsets showed distinct epigenetic
630 profiles, GM-CSF producing cells were more related to IFN-γ than IL-17 producing T cells [10]. With
631 an elegant fate-mapping system, that study [10] could also demonstrate that murine GM-CSF
632 producing murine T cells display a stable epigenetic imprint. While such fate mapping systems are not
633 possible with human cells, and also, capturing multiple cytokine-producing Th subsets in parallel from
634 the same donor without artificially raising their fraction by PMA and ionomycin was not feasible in our
635 study due to the limited cell numbers, our study provides data from unmanipulated *ex vivo* isolated

human cells that may be most relevant for human diseases. Together, our data on naïve, memory, and GM-CSF positive CD4 T cells (and corresponding controls) can be exploited for a multitude of future studies for basic and translational immunology concerning autoimmune diseases. Beyond that, the results may have important implications for other diseases with involvement of GM-CSF producing cells, such as sepsis and COVID-19. Further, our data can be a test-bed for bioinformatics method development for the integration of RNA-seq and ATAC-seq data from the same samples, which may help understanding basic principles of gene regulation in primary eukaryotic cells.

DATA AVAILABILITY

The data sets are available in the GEO repository under the SuperSeries accession number GSE119734 (with the SubSeries GSE119731 and GSE119732).

CONFLICT OF INTEREST

The authors have no conflict of interest.

FUNDING STATEMENT

This work was supported by Karolinska Institute's faculty funds for doctoral education [KID-funding to S.E.]; the Center of Excellence for Research on Inflammation and Cardiovascular disease [CERIC, to A.S. and J.T.]; the 7th European Community Framework Programme [FP7-PEOPLE project 326930 to A.S.; FP7-IDEAS-ERC project 617393 to S.E., A.S., J.T. and D.G.C.; FP7-HEALTH project 306000 to J.T. and D.G.C.]; Vetenskapsrådet Medicine and Health [2011-3264 to J.T.]; and Torsten Söderberg Foundation [to J.T.].

AUTHOR CONTRIBUTIONS

S.E. and A.S. designed the project, designed and carried out all the experiments, interpreted results, prepared figures and wrote the paper; S.E. designed the computational data analysis pipeline and

666 performed all the data analysis including visualization; D.G.C provided advice for data analysis and
667 feedback on the manuscript; J.T. supervised research and provided feedback on the manuscript.

668

669

670 **ACKNOWLEDGEMENTS**

671 The authors thank Matilda Eriksson and Peri Noori for performing RNA extractions, library preparation
672 and quality control for RNA-seq, and next-generation sequencing, as well as for excellent general lab
673 management; Sunjay Jude Fernandes for reagents and advice for the ATAC-seq protocol; and Gilad
674 Silberberg for helpful discussions about footprinting and motif scanning (all from Computational
675 Medicine Unit, Karolinska Institute). We thank John Andersson (Translational Immunology Unit,
676 Karolinska Institute) for helpful suggestions and discussions.

677

678

679 **SUPPLEMENTARY MATERIALS**

680 Supplementary Materials contain the following Figures and Tables: Supplementary Table S1A:
681 Disease enrichment analysis based on gene lists ranked by gene expression in GM-CSF positive
682 CD4 T cells (*versus* GM-CSF negative CD4 T cells). Supplementary Table S1B: Disease enrichment
683 analysis based on gene lists ranked by gene expression in memory (*versus* naïve) CD4 T cells.
684 Supplementary Table S2A: KEGG pathway enrichment analysis of genes assigned to consensus
685 peaks. Supplementary Table S2B: GWAS catalog disease enrichment analysis of genes assigned to
686 consensus peaks. Supplementary Figure S1: Purity of cell populations used in this study.
687 Supplementary Figure S2: Expression profile and DNA accessibility signatures of human memory
688 CD4 T cells. Supplementary Figure S3. Enrichment of footprints across peaks ranked based on
689 differential accessibility.

690

691

692 **REFERENCES**

693

694 [1] Geginat J, Paroni M, Facciotti F, Gruarin P, Kastirr I, Caprioli F, et al. The CD4-centered
695 universe of human T cell subsets. Semin Immunol 2013;25:252–62.
696 doi:10.1016/j.smim.2013.10.012.

697 [2] Geginat J, Paroni M, Maglie S, Alfen JS, Kastirr I, Gruarin P, et al. Plasticity of Human CD4 T

698 Cell Subsets. Front Immunol 2014;5:630. doi:10.3389/fimmu.2014.00630.

699 [3] Caza T, Landas S. Functional and Phenotypic Plasticity of CD4+ T Cell Subsets. Biomed Res
700 Int 2015;2015:1–13. doi:10.1155/2015/521957.

701 [4] McQualter JL, Darwiche R, Ewing C, Onuki M, Kay TW, Hamilton JA, et al. Granulocyte
702 macrophage colony-stimulating factor: a new putative therapeutic target in multiple sclerosis. J
703 Exp Med 2001;194:873–82.

704 [5] Ferber IA, Brocke S, Taylor-Edwards C, Ridgway W, Dinisco C, Steinman L, et al. Mice with a
705 disrupted IFN-gamma gene are susceptible to the induction of experimental autoimmune
706 encephalomyelitis (EAE). J Immunol 1996;156:5–7.

707 [6] Haak S, Croxford AL, Kreymborg K, Heppner FL, Pouly S, Becher B, et al. IL-17A and IL-17F
708 do not contribute vitally to autoimmune neuro-inflammation in mice. J Clin Invest 2009;119:61–
709 9. doi:10.1172/JCI35997.

710 [7] Ponomarev ED, Shriver LP, Maresz K, Pedras-Vasconcelos J, Verthelyi D, Dittel BN. GM-CSF
711 Production by Autoreactive T Cells Is Required for the Activation of Microglial Cells and the
712 Onset of Experimental Autoimmune Encephalomyelitis. J Immunol 2007;178:39–48.
713 doi:178/1/39 [pii].

714 [8] Codarri L, Gyülvészi G, Tosevski V, Hesske L, Fontana A, Magnenat L, et al. RORγt drives
715 production of the cytokine GM-CSF in helper T cells, which is essential for the effector phase
716 of autoimmune neuroinflammation. Nat Immunol 2011;12:560–7. doi:10.1038/ni.2027.

717 [9] Sheng W, Yang F, Zhou Y, Yang H, Low PY, Kemeny DM, et al. STAT5 programs a distinct
718 subset of GM-CSF-producing T helper cells that is essential for autoimmune
719 neuroinflammation. Cell Res 2014;24:1387–402. doi:10.1038/cr.2014.154.

720 [10] Komuczki J, Tuzlak S, Friebel E, Hartwig T, Spath S, Rosenstiel P, et al. Fate-Mapping of GM-
721 CSF Expression Identifies a Discrete Subset of Inflammation-Driving T Helper Cells Regulated
722 by Cytokines IL-23 and IL-1β. Immunity 2019;50:1289-1304.e6.
723 doi:10.1016/j.immuni.2019.04.006.

724 [11] Noster R, Riedel R, Mashreghi M-F, Radbruch H, Harms L, Haftmann C, et al. IL-17 and GM-
725 CSF expression are antagonistically regulated by human T helper cells. Sci Transl Med
726 2014;6:241ra80. doi:10.1126/scitranslmed.3008706.

727 [12] Restorick SM, Durant L, Kalra S, Hassan-Smith G, Rathbone E, Douglas MR, et al. CCR6 +
728 Th cells in the cerebrospinal fluid of persons with multiple sclerosis are dominated by
729 pathogenic non-classic Th1 cells and GM-CSF-only-secreting Th cells. Brain Behav Immun
730 2017;64:71–9. doi:10.1016/j.bbi.2017.03.008.

731 [13] Hartmann FJ, Khademi M, Aram J, Ammann S, Kockum I, Constantinescu C, et al. Multiple

732 sclerosis-associated IL2RA polymorphism controls GM-CSF production in human TH cells. Nat
733 Commun 2014;5:5056. doi:10.1038/ncomms6056.

734 [14] Piper C, Pesenacker AM, Bending D, Thirugnanabalan B, Varsani H, Wedderburn LR, et al.
735 Brief Report: T Cell Expression of Granulocyte-Macrophage Colony-Stimulating Factor in
736 Juvenile Arthritis Is Contingent Upon Th17 Plasticity. Arthritis Rheumatol 2014;66:1955–60.
737 doi:10.1002/art.38647.

738 [15] Wicks IP, Roberts AW. Targeting GM-CSF in inflammatory diseases. Nat Rev Rheumatol
739 2015;12:37–48. doi:10.1038/nrrheum.2015.161.

740 [16] Galli E, Hartmann FJ, Schreiner B, Ingelfinger F, Arvaniti E, Diebold M, et al. GM-CSF and
741 CXCR4 define a T helper cell signature in multiple sclerosis. Nat Med 2019;25:1290–300.
742 doi:10.1038/s41591-019-0521-4.

743 [17] Huang H, Wang S, Jiang T, Fan R, Zhang Z, Mu J, et al. High levels of circulating GM-
744 CSF+CD4+ T cells are predictive of poor outcomes in sepsis patients: a prospective cohort
745 study. Cell Mol Immunol 2019;16:602–10. doi:10.1038/s41423-018-0164-2.

746 [18] Zhou Y, Fu B, Zheng X, Wang D, Zhao C, Qi Y, et al. Pathogenic T-cells and inflammatory
747 monocytes incite inflammatory storms in severe COVID-19 patients. Natl Sci Rev 2020;7:998–
748 1002. doi:10.1093/nsr/nwaa041.

749 [19] Chen Z, John Wherry E. T cell responses in patients with COVID-19. Nat Rev Immunol
750 2020:1–8. doi:10.1038/s41577-020-0402-6.

751 [20] Vabret N, Britton GJ, Gruber C, Hegde S, Kim J, Kuksin M, et al. Immunology of COVID-19:
752 Current State of the Science. Immunity 2020;52. doi:10.1016/j.immuni.2020.05.002.

753 [21] Lang FM, Lee KMC, Teijaro JR, Becher B, Hamilton JA. GM-CSF-based treatments in COVID-
754 19: reconciling opposing therapeutic approaches. Nat Rev Immunol 2020;20:507–14.
755 doi:10.1038/s41577-020-0357-7.

756 [22] Lee Y, Awasthi A, Yosef N, Quintana FJ, Xiao S, Peters A, et al. Induction and molecular
757 signature of pathogenic TH17 cells. Nat Immunol 2012;13:991–9. doi:10.1038/ni.2416.

758 [23] Wu C, Yosef N, Thalhamer T, Zhu C, Xiao S, Kishi Y, et al. Induction of pathogenic T17 cells
759 by inducible salt-sensing kinase SGK1. Nature 2013;496:513–7.

760 [24] El-Behi M, Ciric B, Dai H, Yan Y, Cullimore M, Safavi F, et al. The encephalitogenicity of TH17
761 cells is dependent on IL-1- and IL-23-induced production of the cytokine GM-CSF. Nat
762 Immunol 2011;12:568–75. doi:10.1038/ni.2031.

763 [25] Zielinski CE. Autoimmunity beyond Th17: GM-CSF producing T cells. Cell Cycle
764 2014;13:2489–90. doi:10.4161/15384101.2014.946377.

765 [26] Cao Y, Goods BA, Raddassi K, Nepom GT, Kwok WW, Love JC, et al. Functional
766 inflammatory profiles distinguish myelin-reactive T cells from patients with multiple sclerosis.
767 Sci Transl Med 2015;7:287ra74-287ra74. doi:10.1126/scitranslmed.aaa8038.

768 [27] Éliás S, Schmidt A, Kannan V, Andersson J, Tegnér J. TGF-β Affects the Differentiation of
769 Human GM-CSF+ CD4+ T Cells in an Activation- and Sodium-Dependent Manner. Front
770 Immunol 2016;7:603. doi:10.3389/fimmu.2016.00603.

771 [28] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native
772 chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins
773 and nucleosome position. Nat Methods 2013;10:1213–8. doi:10.1038/nmeth.2688.

774 [29] Chang JT, Wherry EJ, Goldrath AW. Molecular regulation of effector and memory T cell
775 differentiation. Nat Immunol 2014;15:1104–15. doi:10.1038/ni.3031.

776 [30] Durek P, Nordström K, Gasparoni G, Salhab A, Kressler C, de Almeida M, et al. Epigenomic
777 Profiling of Human CD4(+) T Cells Supports a Linear Differentiation Model and Highlights
778 Molecular Regulators of Memory Development. Immunity 2016;45:1148–61.
779 doi:10.1016/j.immuni.2016.10.022.

780 [31] Sen DR, Kaminski J, Barnitz RA, Kurachi M, Gerdemann U, Yates KB, et al. The epigenetic
781 landscape of T cell exhaustion. Science (80-) 2016;354:1165–9.
782 doi:10.1126/science.aae0491.

783 [32] Scott-Browne JP, López-Moyado IF, Trifari S, Wong V, Chavez L, Rao A, et al. Dynamic
784 Changes in Chromatin Accessibility Occur in CD8+ T Cells Responding to Viral Infection.
785 Immunity 2016;45:1327–40. doi:10.1016/j.immuni.2016.10.028.

786 [33] Mognol GP, Spreafico R, Wong V, Scott-Browne JP, Togher S, Hoffmann A, et al. Exhaustion-
787 associated regulatory regions in CD8(+) tumor-infiltrating T cells. Proc Natl Acad Sci U S A
788 2017;114:E2776–85. doi:10.1073/pnas.1620498114.

789 [34] Scharer CD, Bally APR, Gandham B, Boss JM. Cutting Edge: Chromatin Accessibility
790 Programs CD8 T Cell Memory. J Immunol 2017;198:2238–43. doi:10.4049/jimmunol.1602086.

791 [35] Yu B, Zhang K, Milner JJ, Toma C, Chen R, Scott-Browne JP, et al. Epigenetic landscapes
792 reveal transcription factors that regulate CD8+ T cell differentiation. Nat Immunol
793 2017;18:573–82. doi:10.1038/ni.3706.

794 [36] Qu K, Zaba LC, Giresi PG, Li R, Longmire M, Kim YH, et al. Individuality and variation of
795 personal regulomes in primary human T cells. Cell Syst 2015;1:51–61.
796 doi:10.1016/j.cels.2015.06.003.

797 [37] Gensterblum E, Renauer P, Coit P, Strickland FM, Kilian NC, Miller S, et al.
798 CD4+CD28+KIR+CD11a(hi) T cells correlate with disease activity and are characterized by a

799 pro-inflammatory epigenetic and transcriptional profile in lupus patients. J Autoimmun
800 2018;86:19–28. doi:10.1016/j.jaut.2017.09.011.

801 [38] Kitagawa Y, Ohkura N, Kidani Y, Vandenbon A, Hirota K, Kawakami R, et al. Guidance of
802 regulatory T cell development by Satb1-dependent super-enhancer establishment. Nat
803 Immunol 2016;18:173–83. doi:10.1038/ni.3646.

804 [39] Satpathy AT, Saligrama N, Buenrostro JD, Wei Y, Wu B, Rubin AJ, et al. Transcript-indexed
805 ATAC-seq for precision immune profiling. Nat Med 2018;24:580–90. doi:10.1038/s41591-018-
806 0008-8.

807 [40] Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, et al. Genetic determinants of
808 co-accessible chromatin regions in activated T cells across humans. Nat Genet 2018;50:1140–
809 50. doi:10.1038/s41588-018-0156-2.

810 [41] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of
811 lineage-determining transcription factors prime cis-regulatory elements required for
812 macrophage and B cell identities. Mol Cell 2010;38:576–89. doi:10.1016/j.molcel.2010.05.004.

813 [42] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning
814 sequence reads to genomic features. Bioinformatics 2014;30:923–30.
815 doi:10.1093/bioinformatics/btt656.

816 [43] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential
817 expression analysis of digital gene expression data. Bioinformatics 2010;26:139–40.
818 doi:10.1093/bioinformatics/btp616.

819 [44] Hansen KD, Irizarry RA, WU Z. Removing technical variability in RNA-seq data using
820 conditional quantile normalization. Biostatistics 2012;13:204–16.
821 doi:10.1093/biostatistics/kxr054.

822 [45] Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S. Wellington: a novel method for the
823 accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res
824 2013;41:e201–e201. doi:10.1093/nar/gkt850.

825 [46] Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive
826 human regulatory lexicon encoded in transcription factor footprints. Nature 2012;489:83–90.
827 doi:10.1038/nature11212.

828 [47] Matys V, Kel-Margoulis O V, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and
829 its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res
830 2006;34:D108-10. doi:10.1093/nar/gkj143.

831 [48] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set
832 enrichment analysis: a knowledge-based approach for interpreting genome-wide expression

833 profiles. Proc Natl Acad Sci U S A 2005;102:15545–50. doi:10.1073/pnas.0506580102.

834 [49] Sergushichev A. An algorithm for fast preranked gene set enrichment analysis using
835 cumulative statistic calculation. BioRxiv 2016.

836 [50] Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, Teichmann SA. RNA
837 sequencing reveals two major classes of gene expression levels in metazoan cells. Mol Syst
838 Biol 2011;7:497. doi:10.1038/msb.2011.28.

839 [51] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking:bringing order to the
840 web. Tech Report, Stanford Digit Libr Technol Proj 1998:1–17. doi:SIDL-WP-1999-0120.

841 [52] Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9:2579–
842 605.

843 [53] Rasouli J, Ciric B, Imitola J, Gonnella P, Hwang D, Mahajan K, et al. Expression of GM-CSF in
844 T Cells Is Increased in Multiple Sclerosis and Suppressed by IFN-β Therapy. J Immunol
845 2015;194:5085–93. doi:10.4049/jimmunol.1403243.

846 [54] Anderson P. Post-transcriptional control of cytokine production. Nat Immunol 2008;9:353–9.
847 doi:10.1038/ni1584.

848 [55] Chen CY, Gherzi R, Ong SE, Chan EL, Raijmakers R, Pruijn GJ, et al. AU binding proteins
849 recruit the exosome to degrade ARE-containing mRNAs. Cell 2001;107:451–64.

850 [56] Shaw G, Kamen R. A conserved AU sequence from the 3' untranslated region of GM-CSF
851 mRNA mediates selective mRNA degradation. Cell 1986;46:659–67.

852 [57] Wong MT, Ong DEH, Lim FSH, Teng KWW, McGovern N, Narayanan S, et al. A High-
853 Dimensional Atlas of Human T Cell Diversity Reveals Tissue-Specific Trafficking and Cytokine
854 Signatures. Immunity 2016;45:442–56. doi:10.1016/j.immuni.2016.07.007.

855 [58] Liu H, Rhodes M, Wiest DL, Vignali DA. On the dynamics of TCR:CD3 complex cell surface
856 expression and downmodulation. Immunity 2000;13:665–75.

857 [59] Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and
858 transcriptomic analyses. Nat Rev Genet 2012;13:227–32. doi:10.1038/nrg3185.

859 [60] Edfors F, Danielsson F, Hallström BM, Käll L, Lundberg E, Pontén F, et al. Gene-specific
860 correlation of RNA and protein levels in human cells and tissues. Mol Syst Biol 2016;12:883.
861 doi:10.15252/msb.20167144.

862 [61] Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong CK, Fumis L, Karamanis N, et al. Open
863 Targets Platform: New developments and updates two years on. Nucleic Acids Res
864 2019;47:D1056–65. doi:10.1093/nar/gky1133.

865 [62] Nylander A, Hafler DA. Multiple sclerosis. J Clin Invest 2012;122:1180–8.
866 doi:10.1172/JCI58649.

867 [63] Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, Cotsapas C, et al.
868 Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis.
869 Nat Genet 2013;45:1353–60. doi:10.1038/ng.2770.

870 [64] International Multiple Sclerosis Genetics Consortium S, Wellcome Trust Case Control
871 Consortium 2 G, Sawcer S, Hellenthal G, Pirinen M, Spencer CCA, et al. Genetic risk and a
872 primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature
873 2011;476:214–9. doi:10.1038/nature10251.

874 [65] Hrckulak D, Kolar M, Strnad H, Korinek V. TCF/LEF Transcription Factors: An Update from the
875 Internet Resources. Cancers (Basel) 2016;8:E70. doi:10.3390/cancers8070070.

876 [66] Hu D, Notarbartolo S, Croonenborghs T, Patel B, Cialic R, Yang T-H, et al. Transcriptional
877 signature of human pro- inflammatory T H 17 cells identifies reduced IL10 gene expression in
878 multiple sclerosis. Nat Commun 2017;8:1600. doi:10.1038/s41467-017-01571-8.

879 [67] Al-Mossawi MH, Chen L, Fang H, Ridley A, de Wit J, Yager N, et al. Unique transcriptome
880 signatures and GM-CSF expression in lymphocytes from patients with spondyloarthritis. Nat
881 Commun 2017;8:1510. doi:10.1038/s41467-017-01771-2.

882

883

884

885

886 **FIGURE LEGENDS**

887

888 **Figure 1. Experimental setup and quality control for human CD4 T cell transcriptome and**
889 **chromatin accessibility analysis.** (**A**) Human PBMCs for each donor were split in three fractions,
890 and memory CD4 T cells, naïve CD4 T cells and bulk (total) CD4 T cells were isolated in parallel by
891 negative ("untouched") magnetic isolation (cell purities, see Supplementary Figure S1). From bulk
892 CD4 T cells, GM-CSF-secreting cells (GM-CSF positive) were captured, and the negative fraction was
893 used as corresponding GM-CSF negative population. The five indicated cell populations were used
894 for molecular profiling by ATAC-seq and RNA-seq. Differentially accessible DNA regions (DAR) and
895 differentially expressed genes (DEG) were determined for the comparison of memory *versus* naïve
896 CD4 cells and for GM-CSF positive *versus* GM-CSF negative CD4 cells respectively. (**B**) High purity
897 of isolated (captured) GM-CSF+ CD4 T cells was confirmed by flow cytometry. The histograms show
898 the signal of the GM-CSF capture construct for the indicated cell populations, pre-gated on live singlet
899 lymphocytes based on forward scatter, side scatter and pulse width. Here, bulk CD4 cells represent
900 an aliquot taken after labelling with the GM-CSF-PE capture construct, but without further capture
901 assay procedure. The left panel shows a representative donor and the right panel shows summarized
902 data for n=6 donors from 4 independent experiments (mean and SEM are indicated in grey). Each
903 donor is represented by a symbol; same symbol shape (but filled or unfilled) indicate donors
904 processed within the same experiment.

905

906 **Figure 2. Explorative analysis of ATAC-seq and RNA-seq samples.** (**A**) Principal Component
907 Analysis (PCA) was carried out on CQN-normalized ATAC-seq data given as $\log_2$(RPKM+1) centered
908 by mean subtraction for each feature (genomic region) across samples. PC1 and PC2 are shown,
909 along with the % variation explained. Symbol colors indicate the given cell populations, and symbol
910 shapes and fillings represent individual donors (n=5-6 donors) as in Figure 1B. Data from technical
911 replicates were pooled before analysis. (**B**) PCA for RNA-seq data given as $\log_2$(FPKM+1) centered
912 by mean subtraction for each feature (gene) across samples. Labels as in (A). (**C, D**) t-SNE
913 dimensionality reduction visualization of ATAC-seq (C) and RNA-seq (D) data, processed and labeled
914 as in (A) and (B), respectively.

915

916 **Figure 3. Differential chromatin accessibility and gene expression in memory *versus* naïve**
917 **CD4 T cells as well as in GM-CSF positive *versus* negative cells.** (**A**) ATAC-seq data were CQN-
918 normalized and differential accessibility between the indicated cell population comparison was
919 calculated (left: memory *vs.* naïve CD4 T cells; right: GM-CSF positive *vs.* negative CD4 T cells).
920 Volcano plots show each consensus peak as a single dot, and lines indicate the threshold for calling a
921 DAR (FDR<0.05 and >25% fold change). Not significantly differentially accessible regions are
922 depicted in grey. (**B**) Differential gene expression was calculated from RNA-seq data for cell

923 populations as in (A). Lines indicate the threshold for calling a DEG (FDR<0.05 and > 25% fold
924 change), not significantly differentially expressed genes are depicted in grey. (**C**) A selection of DARs
925 in Memory *vs.* Naïve (not DAR in GM-CSF positive *vs.* GM-CSF negative) was studied for being
926 assigned to genes known to play a role in T cell memory. Log_2(fold change(Memory/Naïve)) for
927 selected DARs (FDR<0.05) are plotted, and colors indicate the category the respective region is
928 assigned to (TSS: transcription start site; TTS: transcription termination site). (**D**) Known T cell
929 memory "up" (black) or T cell memory "down" (grey) genes based on previous literature were selected
930 if differentially expressed (FDR<0.05 in Memory *vs.* Naïve) in the present data. log_2(fold
931 change(Memory/Naïve)) for these selected DEGs is plotted; values >0 represent up-, values <0
932 represent down-regulation in Memory T cells. Colors indicate whether the gene was previously
933 described to be "up" or "down" in memory T cells.

934

935 **Figure 4. Expression profile and DNA accessibility signatures of human GM-CSF positive CD4**
936 **T cells** (**A**) Differentially accessible DNA regions in GM-CSF positive *versus* GM-CSF negative CD4
937 T cells (FDR<0.05 in this comparison) are plotted as a heatmap for their accessibility in all the five cell
938 populations studied. CQN normalized log_2(RPKM+1) is represented by the color scale, indicating
939 accessibility (blue: low, red: high). Data were row-scaled and -clustered (Euclidean distance,
940 complete linkage clustering). (**B**) Differentially expressed genes (RNA) in GM-CSF positive *versus*
941 GM-CSF negative CD4 T cells (FDR<0.05 in this comparison) are plotted as a heatmap for their
942 expression in all the five cell populations studied. Gene expression is displayed as log_2(FPKM+1) with
943 blue indicating low and red indicating high expression according to the color scale. Data were row-
944 scaled and -clustered (Euclidean distance, complete linkage clustering). (**C**) The mean of log_2(fold
945 change) of GM-CSF+/GM-CSF- cell populations using the median intensity values from CyTOF
946 measurements are shown, using CD4 T cell-gated PBMC data from Wong *et al.* [57] and designated
947 as "CyTOF" (grey bars). For the genes corresponding to the proteins measured in CyTOF, the
948 log_2(fold change) of GM-CSF+/GM-CSF- cell populations from the RNA-seq data of this study is
949 plotted, calculated with the generalized linear model (edgeR) and labelled as "RNA" (red).

950

951 **Figure 5. GM-CSF positive CD4 T cell signatures are associated to autoimmune diseases,**
952 **especially MS.** (**A**) Rank based gene set enrichment analysis is shown using the $-log_{10}$(FDR) $\times$
953 sign(log_2(GM-CSF+/GM-CSF-)) function to rank genes based on differential expression according to
954 the GM-CSF+/- cell populations, and the Open Targets database [61] was used to provide lists of
955 genes associated with diseases. NES: normalized enrichment score; pval: P value; padj: FDR. (**B**)
956 The Heatmap represents the row-scaled log_2(RPKM+1) expression values from RNA-seq data of CD4
957 T cells from MS patients or healthy controls (data from [26]). Groups are separated based on disease
958 status (MS or healthy) and myelin antigen reactivity (reactive: Tetramer+). Genes are selected as
959 those identified in the current RNA-seq study as differentially expressed between GM-CSF positive
960 *versus* GM-CSF negative cells and having detectable expression (log_2RPKM>0) in at least 4 samples

961 in the data from [26]. (**C**) Of the 110 established non-MHC MS susceptibility variants [63], two SNPs
962 mapped to a consensus peak from the current ATAC-seq study. Information about these SNPs is
963 shown in the table, along with the differential accessibility analysis in GM-CSF positive *versus* GM-
964 CSF negative CD4 T cells.

965

966 **Figure 6. Enrichment of ATAC-seq peaks in regions representing certain genomic features,**
967 **and relationship of ATAC-seq and RNA-seq data.** (**A**) The enrichment of ATAC-seq consensus
968 peaks in different types of genomic regions is shown with bars representing the enrichment
969 significance × direction of enrichment (-ln(P value) × sign(enrichment)). (**B**) Genes that were detected
970 on both RNA-seq level and were also assigned to ATAC-seq peaks, were ranked using the –
971 $\log_{10}$(FDR) × sign($\log_2$(fold change)) function based on both RNA-seq and ATAC-seq (using values of
972 the assigned peaks) within a given cell type comparison (left: GM-CSF+ *vs.* GM-CSF-; right: memory
973 *vs.* naïve). The ranks of genes using ATAC-seq *vs.* RNA-seq for ranking are visualized for both cell
974 type comparisons; lines indicate FDR cutoff 0.05 in the respective comparison. P value represents the
975 probability of observing this or more directional agreement between the two datatypes by chance,
976 using only genes significantly differential (FDR<0.05) in both data types. P values were calculated by
977 Fisher's exact test with Monte Carlo simulation. (**C**) Dot plot representation as in (B), but using
978 $\log_2$(fold change) for ranking, and visualizing only (assigned) genes that are significantly differential in
979 both data types.

980

981 **Figure 7. Identification of key TFs with global regulatory effect.** (**A**) Rank based enrichment
982 analysis of TF binding motifs in footprints within peaks that were ranked based on differential
983 accessibility using the –$\log_{10}$(FDR) × sign($\log_2$(fold change)) function. The analysis was performed
984 separately for the GM-CSF+/GM-CSF- and Memory/Naïve cell comparison (left and right
985 respectively). NES: normalized enrichment score. (**B**) Average expression levels of identified key TFs
986 for both cell comparisons are shown as log(Count Per Million reads). To indicate the relative
987 expression level of these TFs (bars) in view of the spectrum of lowly and highly expressed genes, the
988 scaled probability density function is shown based on the distribution of the average expression of all
989 genes. (**C**) Volcano plot shows the differential expression (effect size *vs.* significance) of the TFs
990 identified as key regulators in each of the cell type comparisons (GM-CSF+/GM-CSF- and
991 Memory/Naïve cell comparison; left and right). Key TFs are indicated as black dots, and those which
992 are significantly differentially expressed (FDR<0.05) are labelled with their gene symbol.

993

994 **Figure 8. Network structure identifying shared TF binders across peaks and shared targets of**
995 **key TFs.** (**A, B**) Key TFs were identified as in Figure 7. The binary heatmap shows the binding of key
996 TFs (columns) in footprints with ATAC-seq peaks (rows) that are either differentially accessible and/or
997 that are assigned to a differentially expressed gene in the GM-CSF+/GM-CSF- cell comparison (A) or

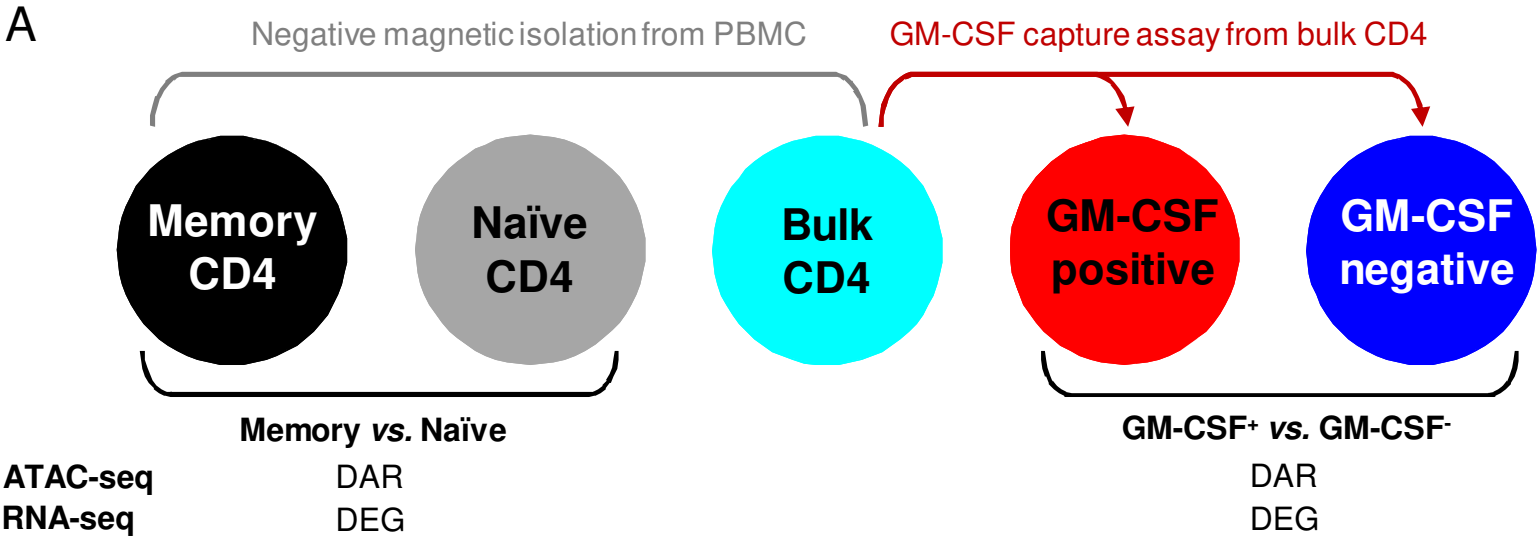
998 in the memory/naïve cell comparison (B). Black indicates that the key TF binds to the respective
999 region. Rows and columns are clustered using binary distance and complete linkage. (**C, D**) Heatmap
1000 representation of the binary distance matrix of the key TFs using the binary vector of binding ("bound"
1001 or "not bound") for distance calculation. The red color scale represents the distance; smaller distance
1002 represents more co-binding. Rows and columns are clustered using Euclidean distance and complete
1003 linkage. The analysis was performed for the GM-CSF+/GM-CSF- cell comparison (C) or the
1004 memory/naïve cell comparison (D).

1005

1006 **Figure 9. Gene regulatory network of GM-CSF+/GM-CSF- or Memory/Naïve CD4 T cells.** (**A**)
1007 Gene regulatory network reflecting the signatures of GM-CSF positive *versus* GM-CSF negative CD4
1008 T cells. Left panel: A directed network representing binding from source TFs to target peaks assigned
1009 to the indicated gene is shown. Source nodes were selected as described (Figure 7, key TFs) and
1010 target nodes were selected to be either differentially accessible (as a peak assigned to the respective
1011 node gene) and/or to be differentially expressed on mRNA level in the GM-CSF+/GM-CSF- contrast.
1012 To calculate the PageRank [51] as a measure of importance of the TFs (with TFs influencing more
1013 genes being more important), the edges were first inverted (not displayed), and the resulting
1014 PageRank value is represented by the color scale (light to dark green for lower to higher PageRank).
1015 Right panel: Those TFs having a PageRank value higher than the 99-percentile in either the GM-CSF
1016 and/or the Memory cell network (B) and their corresponding PageRank value in the GM-CSF network
1017 are displayed. (**B**): Same as (A), but for the Memory/Naïve CD4 T cell contrast.

Figure 1

A


Negative magnetic isolation from PBMC
GM-CSF capture assay from bulk CD4
Memory CD4
Naïve CD4
Bulk CD4
GM-CSF positive
GM-CSF negative
Memory vs. Naïve
GM-CSF+ vs. GM-CSF-
ATAC-seq
DAR
DAR
RNA-seq
DEG
DEG

B

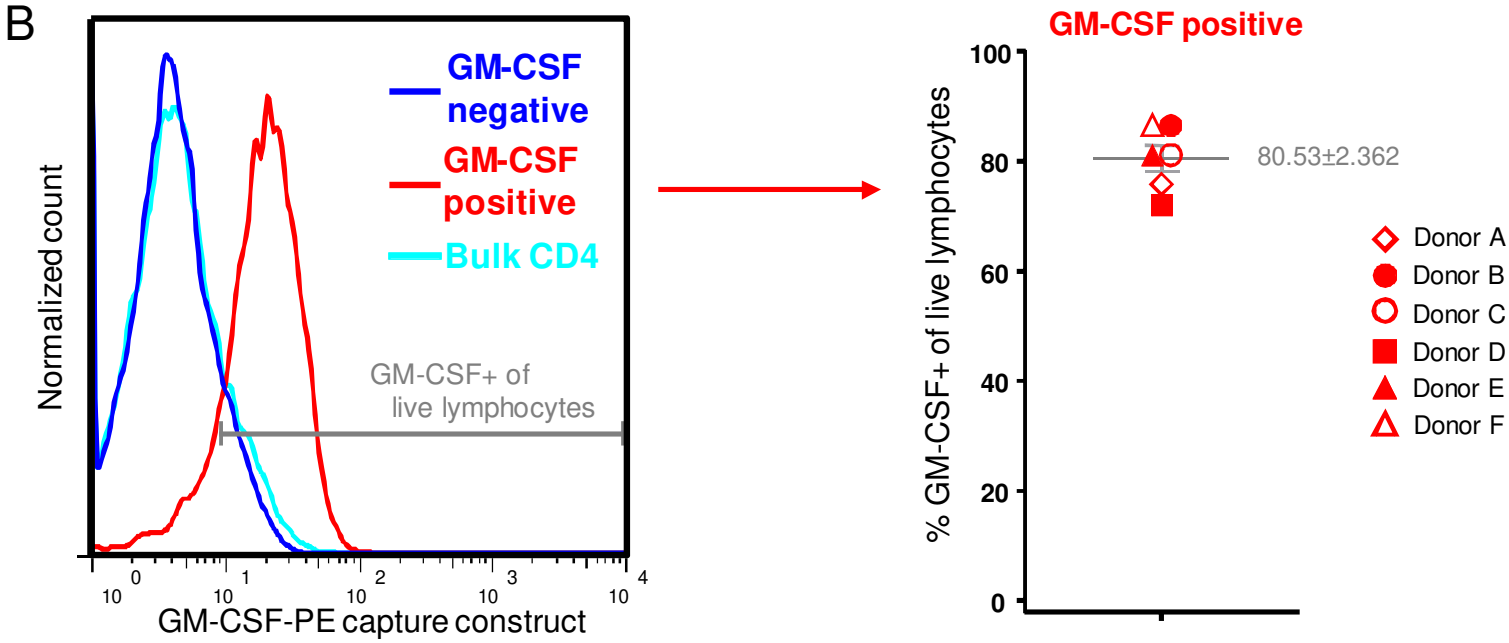

GM-CSF negative
GM-CSF positive
Bulk CD4
GM-CSF+ of live lymphocytes
Normalized count
GM-CSF-PE capture construct
GM-CSF positive
% GM-CSF+ of live lymphocytes
80.53±2.362
Donor A
Donor B
Donor C
Donor D
Donor E
Donor F

Figure 2

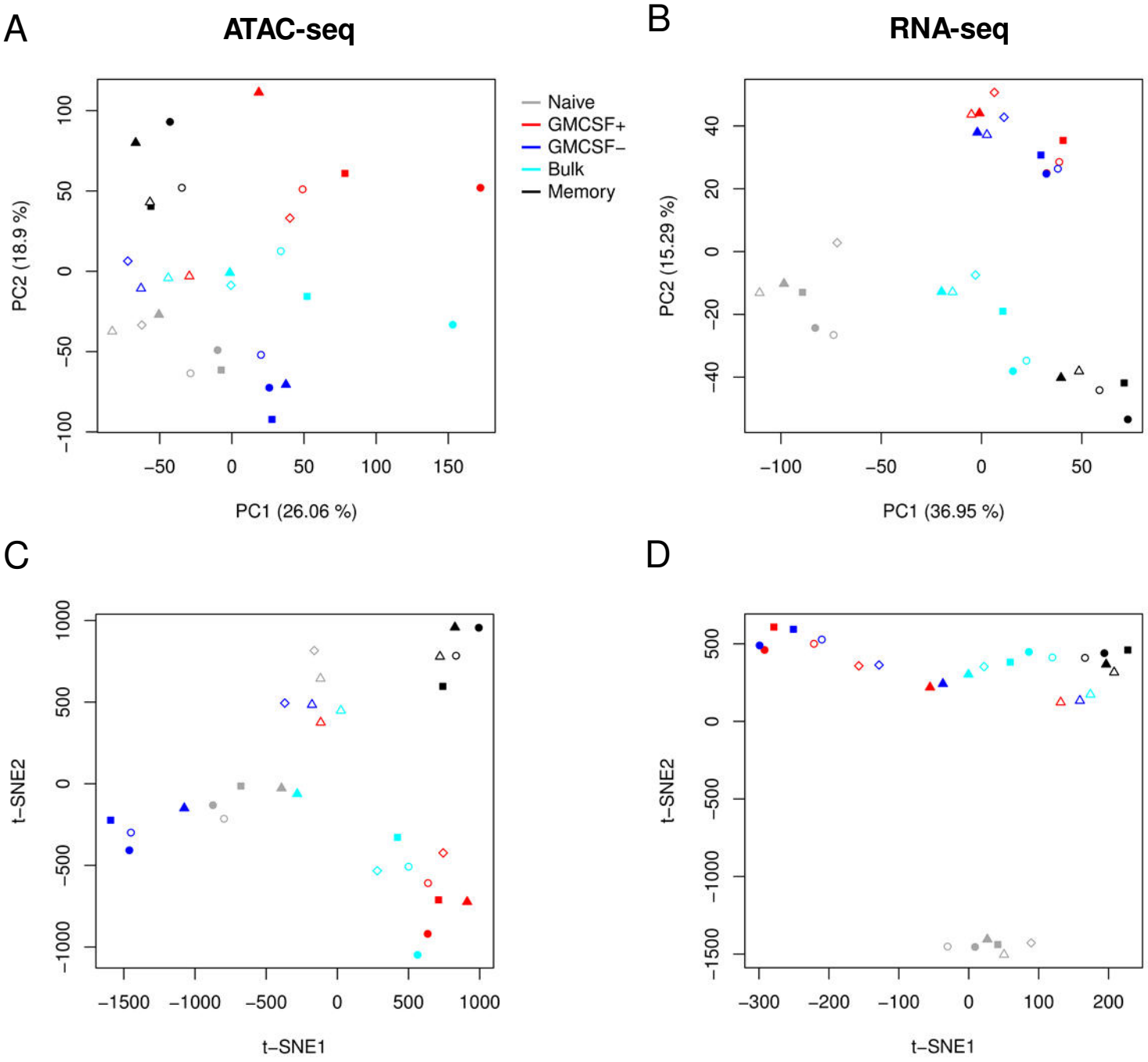
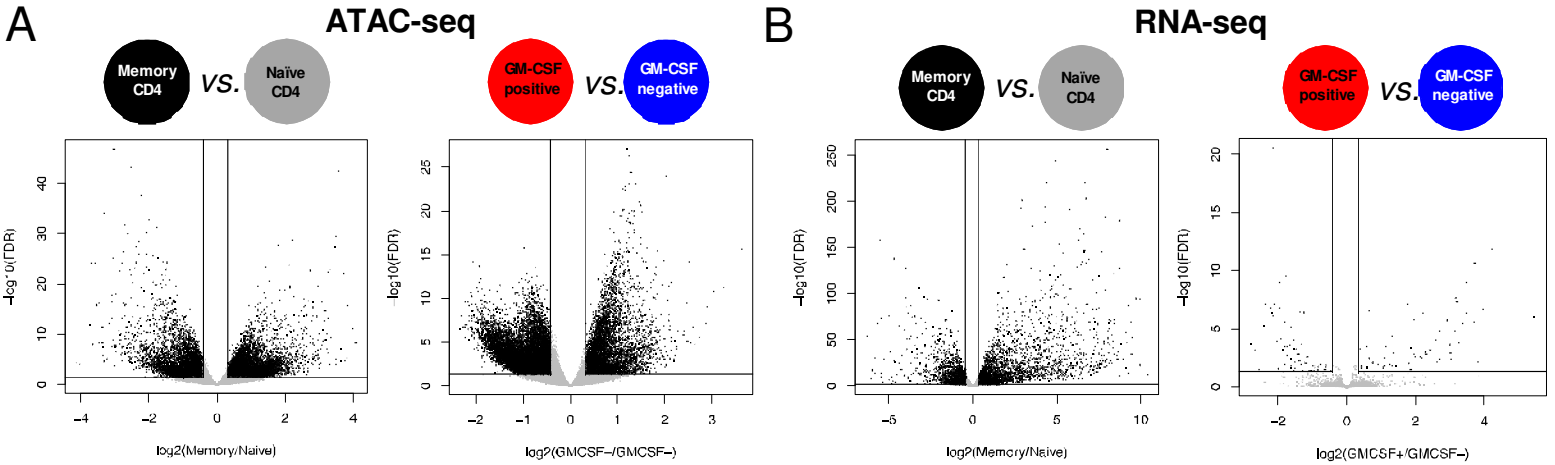

A
ATAC-seq
PC2 (18.9 %)
PC1 (26.06 %)
Naive
GMCSF+
GMCSF−
Bulk
Memory
B
RNA-seq
PC2 (15.29 %)
PC1 (36.95 %)
C
t−SNE2
t−SNE1
D
t−SNE2
t−SNE1

Figure 3


A
ATAC-seq
Memory CD4
vs.
Naïve CD4
GM-CSF positive
vs.
GM-CSF negative
−log10(FDR)
log2(Memory/Naive)
log2(GMCSF−/GMCSF−)
B
RNA-seq
Memory CD4
vs.
Naïve CD4
GM-CSF positive
vs.
GM-CSF negative
−log10(FDR)
log2(Memory/Naive)
log2(GMCSF+/GMCSF−)

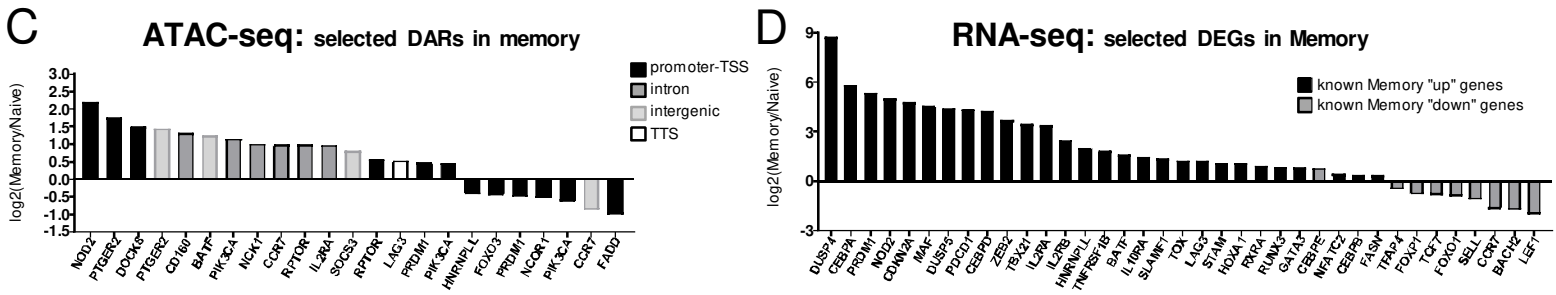
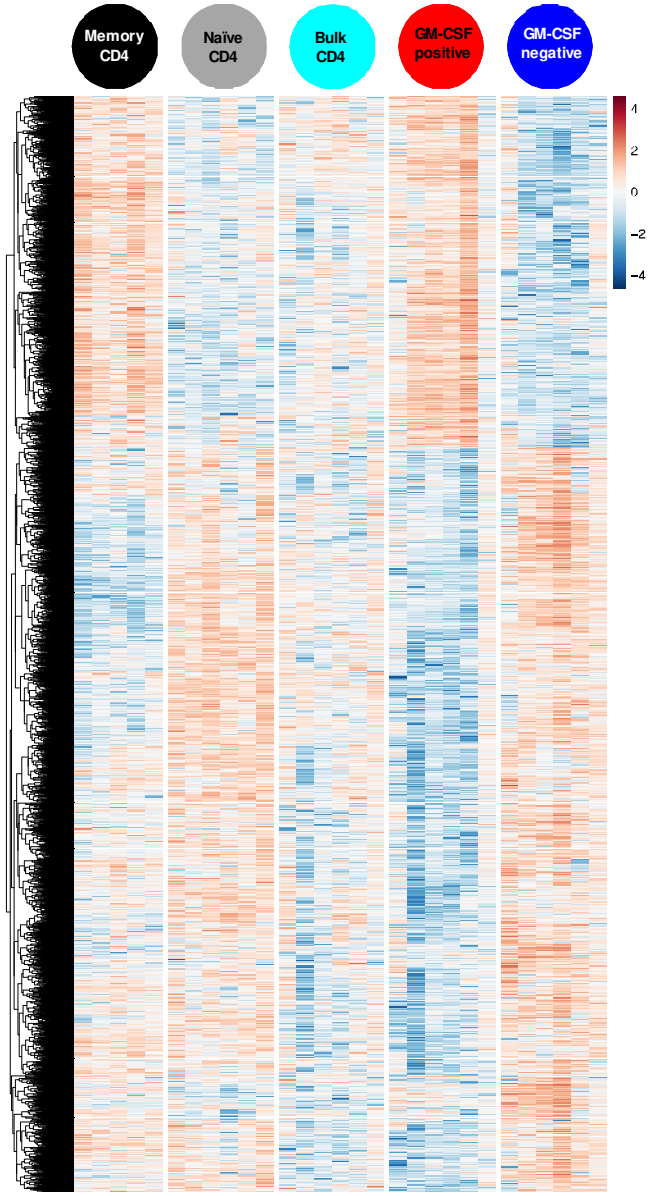
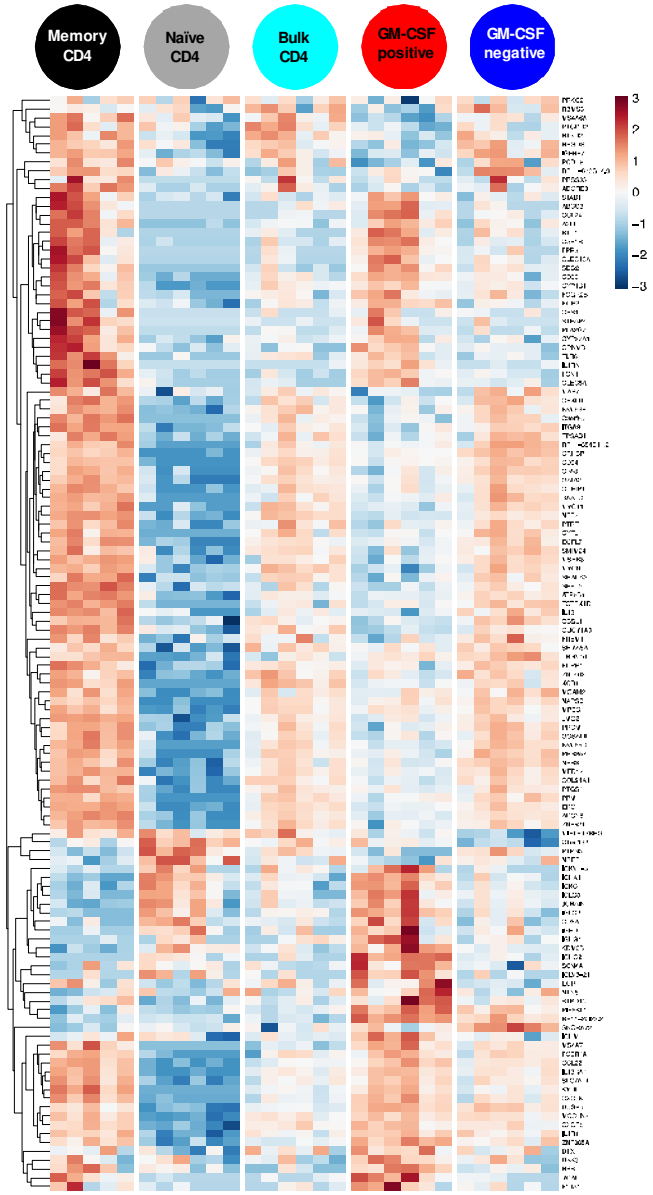

C
ATAC-seq: selected DARs in memory
log2(Memory/Naive)
promoter-TSS
intron
intergenic
TTS
NOD2
PTGER2
DOCK8
PTGER2
CD16p
BATF
PIK3CA
NCK1
CCR7
RPTOR
IL2RA
SOCS3
RPTOR
LAG3
PRDM1
PIK3CA
HNRNPLL
FOXO3
PRDM1
NCOR1
PIK3CA
CCR7
FADD
D
RNA-seq: selected DEGs in Memory
log2(Memory/Naive)
known Memory "up" genes
known Memory "down" genes
DUSP4
CEBPA
PRDM1
NOD2
CDKN2A
MAF
DUSP5
PDCD1
CEBPD
ZEB2
TBX21
IL2RA
IL2RB
HNRNPLL
TNFRSF1B
BATF
IL10RA
SLAMF7
TOX
LAG3
STAM
HOXA1
RXRA
RUNX3
GATA3
CEBPE
NFATC2
CEBPB
FASN
TFAP4
FOXP1
TCF7
FOXO1
SELL
CCR7
BACH2
LEF1

Figure 4

A


ATAC-seq: DARs in GM-CSF positive
Memory CD4
Naïve CD4
Bulk CD4
GM-CSF positive
GM-CSF negative
4
2
0
−2
−4

B


RNA-seq: DEGs in GM-CSF positive
Memory CD4
Naïve CD4
Bulk CD4
GM-CSF positive
GM-CSF negative
3
2
1
0
−1
−2
−3

C

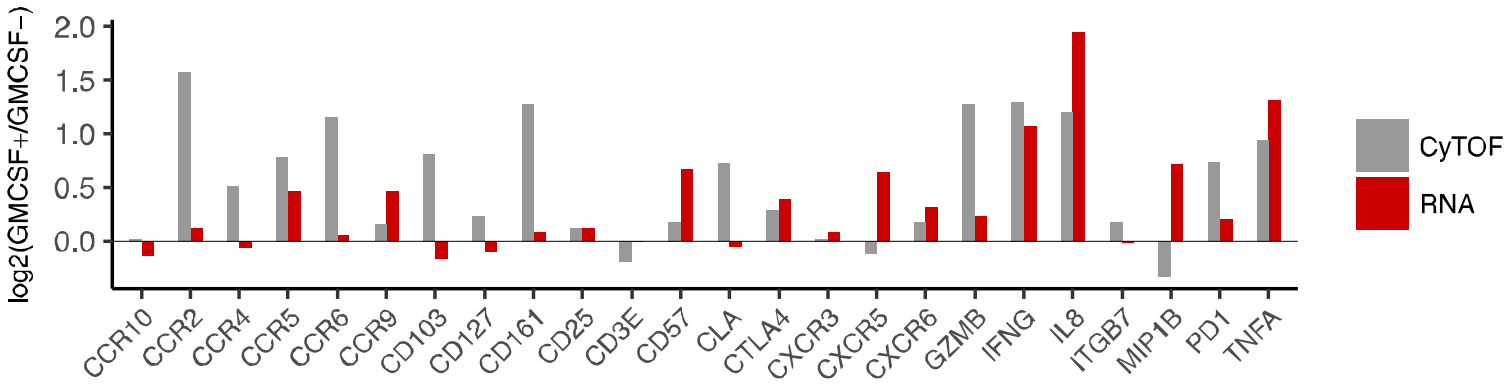
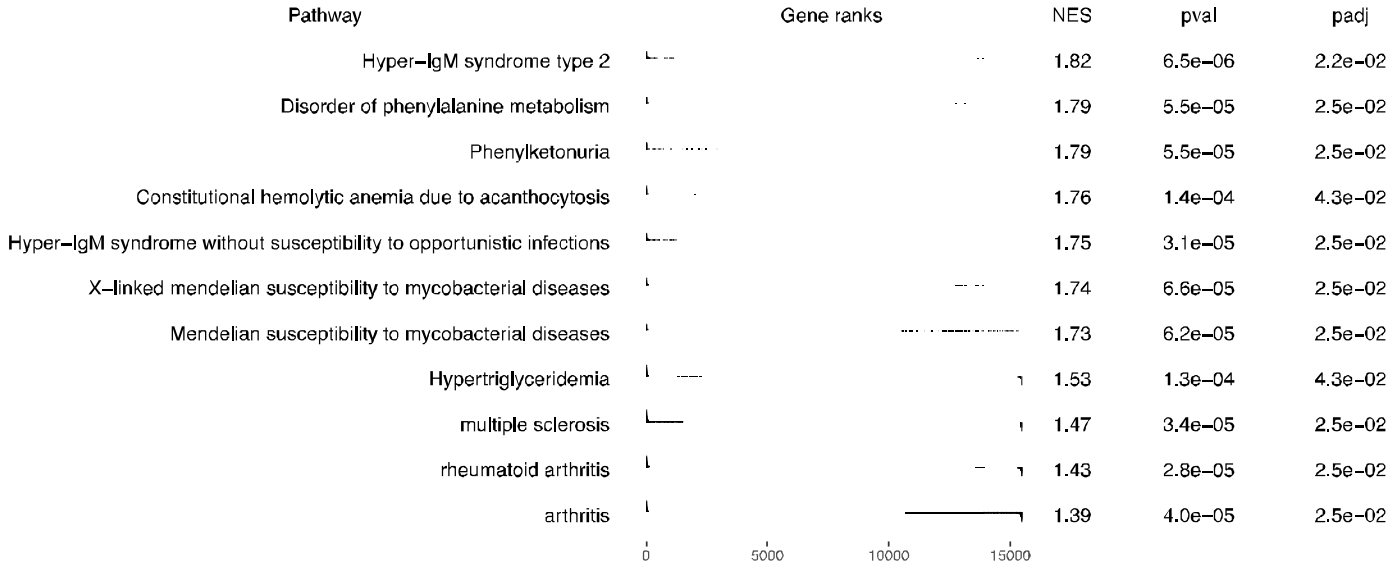
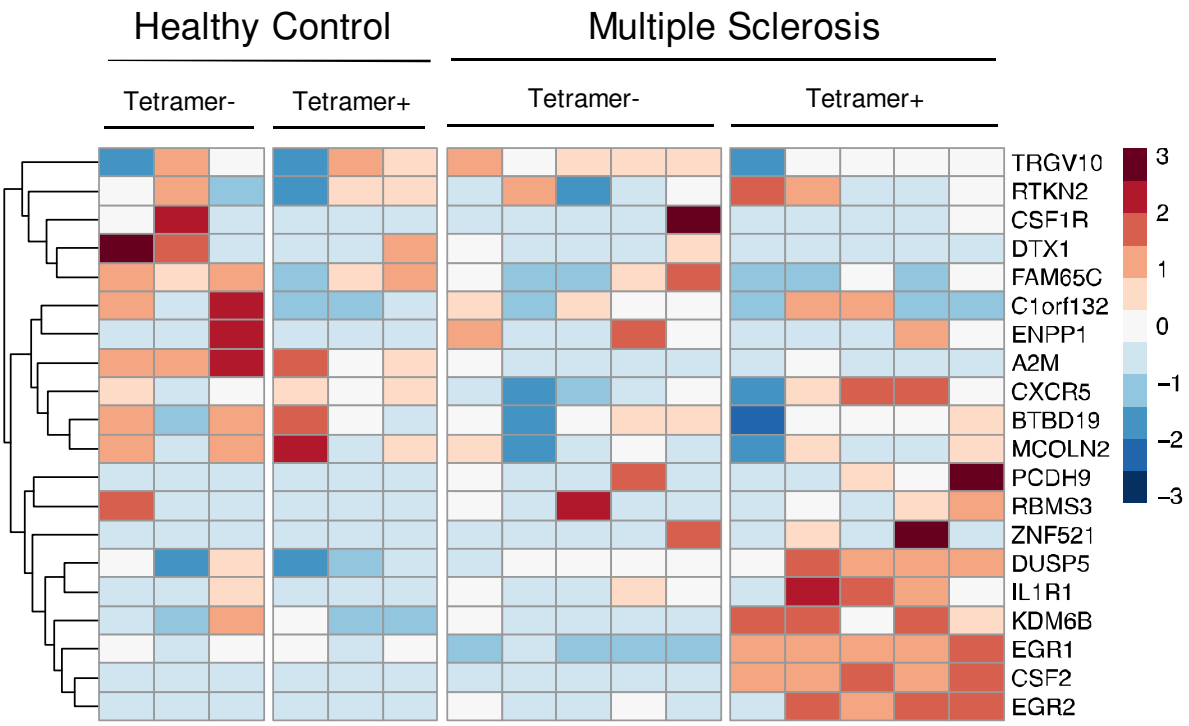

log2(GMCSF+/GMCSF−)
2.0
1.5
1.0
0.5
0.0
CCR10
CCR2
CCR4
CCR5
CCR6
CCR9
CD103
CD127
CD161
CD25
CD3E
CD57
CLA
CTLA4
CXCR3
CXCR5
CXCR6
GZMB
IFNG
IL8
ITGB7
MIP1B
PD1
TNFA
CyTOF
RNA

Figure 5

A



B

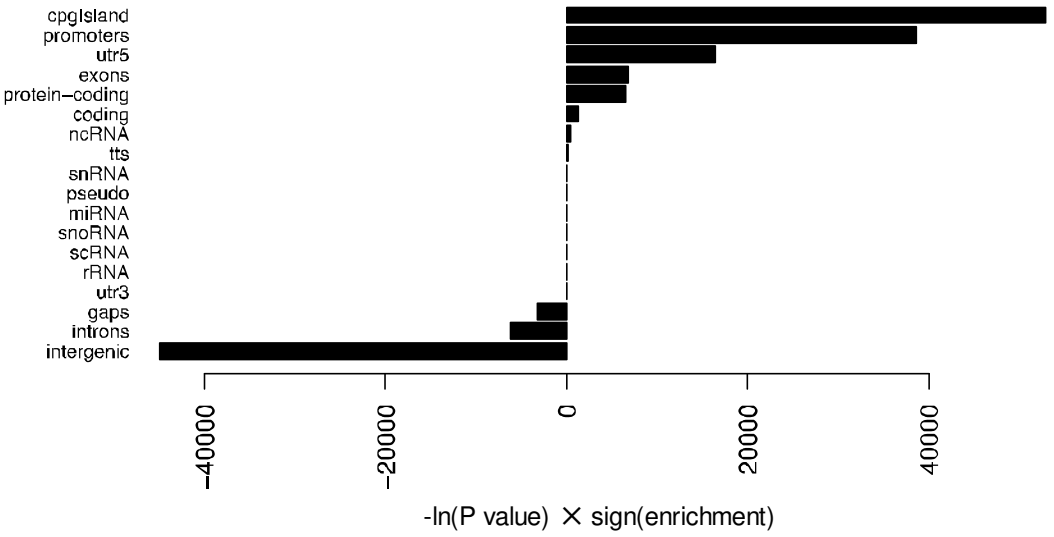


C

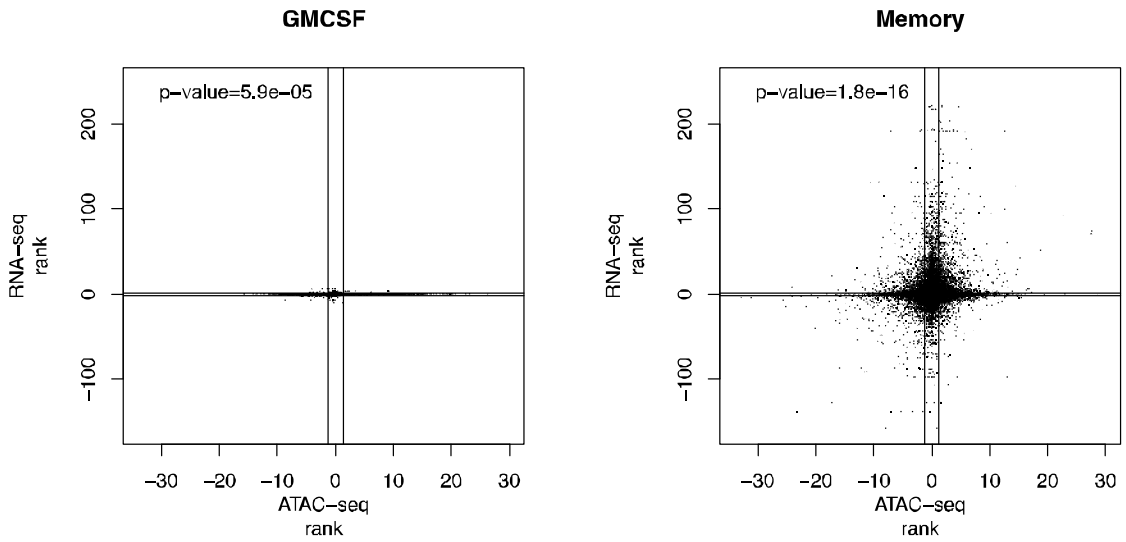
SNP	Gene Name	Annotation	Distance to TSS	log2(GMCSF+/GMCSF-)	FDR
rs1323292	*RGS1*	Intergenic	-3818	-0.972	0.005
rs60600003	*ELMO1*	promoter-TSS	-129	-0.258	0.047

Figure 6

A


cpgIsland
promoters
utr5
exons
protein−coding
coding
ncRNA
tts
snRNA
pseudo
miRNA
snoRNA
scRNA
rRNA
utr3
gaps
introns
intergenic
−40000
−20000
0
20000
40000
-ln(P value) × sign(enrichment)

B


GMCSF
p−value=5.9e−05
RNA−seq
rank
ATAC−seq
rank
Memory
p−value=1.8e−16
RNA−seq
rank
ATAC−seq
rank

C

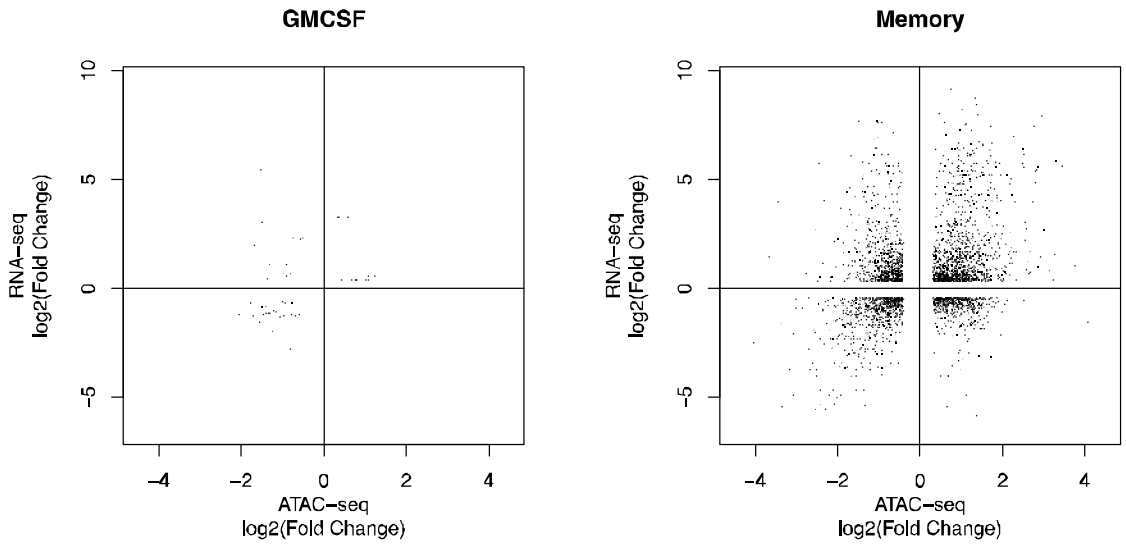

GMCSF
RNA−seq
log2(Fold Change)
ATAC−seq
log2(Fold Change)
Memory
RNA−seq
log2(Fold Change)
ATAC−seq
log2(Fold Change)

Figure 7

A

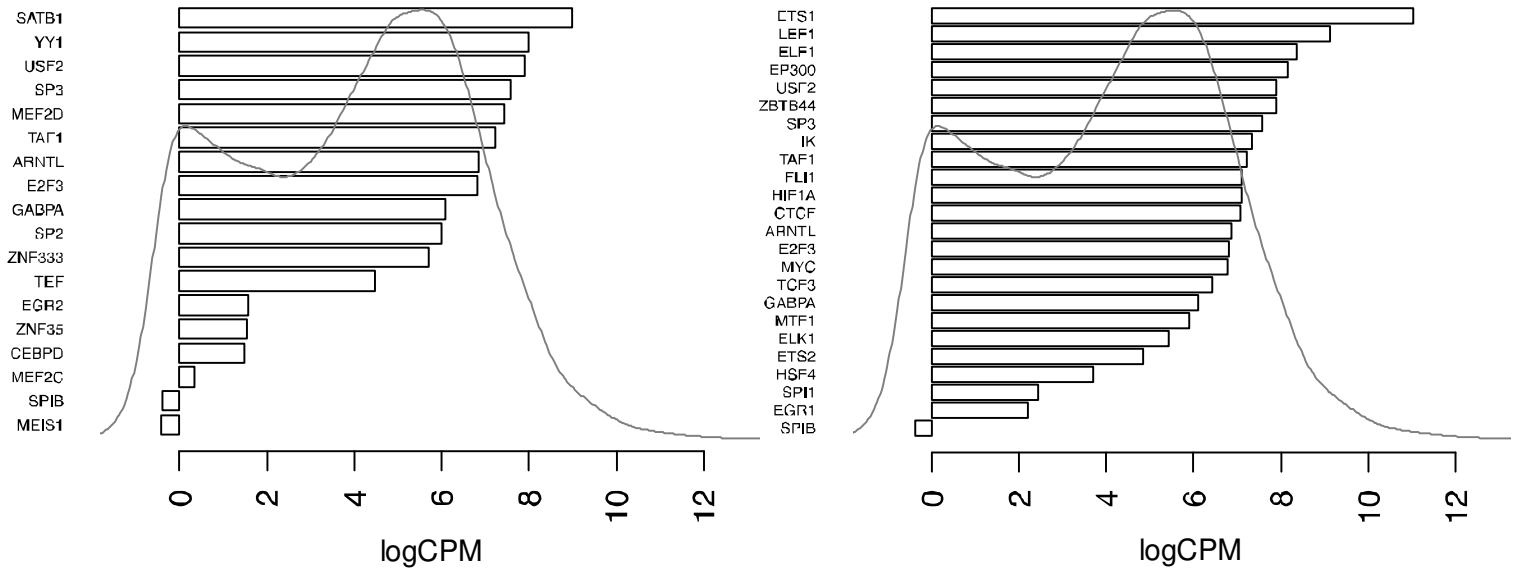
GM-CSF+/GM-CSF-

TF	NES	p value	FDR
YY1	2.25	2.9e−07	4.4e−05
SATB1	2.24	6.1e−07	4.4e−05
ZNF333	2.16	1.1e−05	4.6e−04
MEIS1	1.97	2.3e−04	5.5e−03
MEF2D	1.95	5.1e−04	9.2e−03
EGR2	1.86	1.9e−03	2.7e−02
TEF	1.84	3.6e−04	7.5e−03
TAF1	1.80	2.4e−04	5.5e−03
MEF2C	1.80	3.5e−03	4.0e−02
SP2	1.67	2.9e−05	8.6e−04
ARNTL	1.65	2.8e−03	3.5e−02
USF2	1.62	5.3e−04	9.2e−03
ZNF35	1.61	5.8e−04	9.3e−03
SPIB	1.52	8.4e−07	4.4e−05
GABPA	1.51	1.5e−05	5.2e−04
E2F3	1.49	8.4e−07	4.4e−05
SP3	1.46	2.2e−03	2.9e−02
CEBPD	1.33	8.3e−04	1.2e−02

Memory/Naive

TF	NES	p value	FDR
CTCF	1.82	1.0e−03	1.2e−02
E2F3	−1.51	1.1e−04	2.0e−03
SP3	−1.52	4.9e−03	4.2e−02
HSF4	−1.55	4.4e−03	4.0e−02
ETS2	−1.61	5.3e−03	4.3e−02
ZBTB44	−1.67	5.5e−03	4.3e−02
SPIB	−1.73	1.8e−07	1.2e−05
HIF1A	−1.73	3.9e−03	3.9e−02
ETS1	−1.75	4.5e−03	4.0e−02
EP300	−1.76	3.2e−04	5.0e−03
TAF1	−1.78	1.8e−03	1.9e−02
FLI1	−1.78	7.8e−04	1.0e−02
MTF1	−1.80	7.0e−04	1.0e−02
ARNTL	−1.82	1.1e−03	1.2e−02
IK	−1.86	1.0e−03	1.2e−02
EGR1	−1.88	8.9e−06	2.4e−04
SPI1	−1.90	6.7e−06	2.4e−04
GABPA	−1.90	1.8e−07	1.2e−05
ELF1	−1.94	1.3e−06	6.2e−05
TCF3	−1.94	7.7e−05	1.4e−03
LEF1	−1.94	7.7e−05	1.4e−03
USF2	−1.99	7.6e−06	2.4e−04
ELK1	−1.99	1.9e−07	1.2e−05
MYC	−2.05	1.5e−05	3.5e−04

B



C

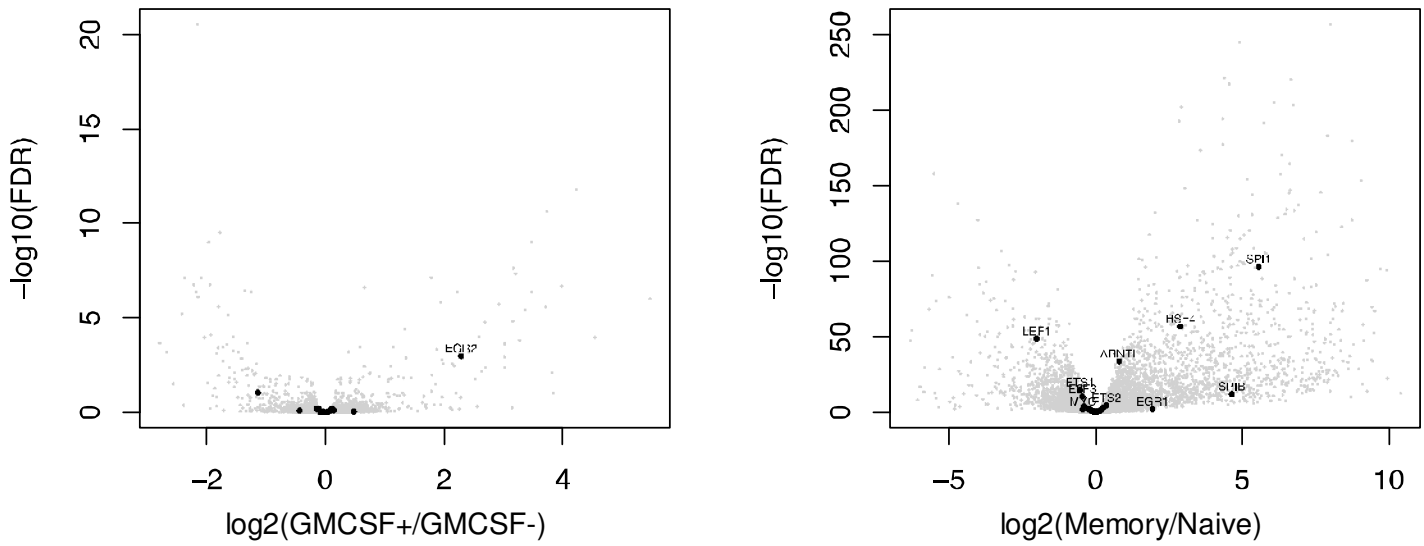


Figure 8

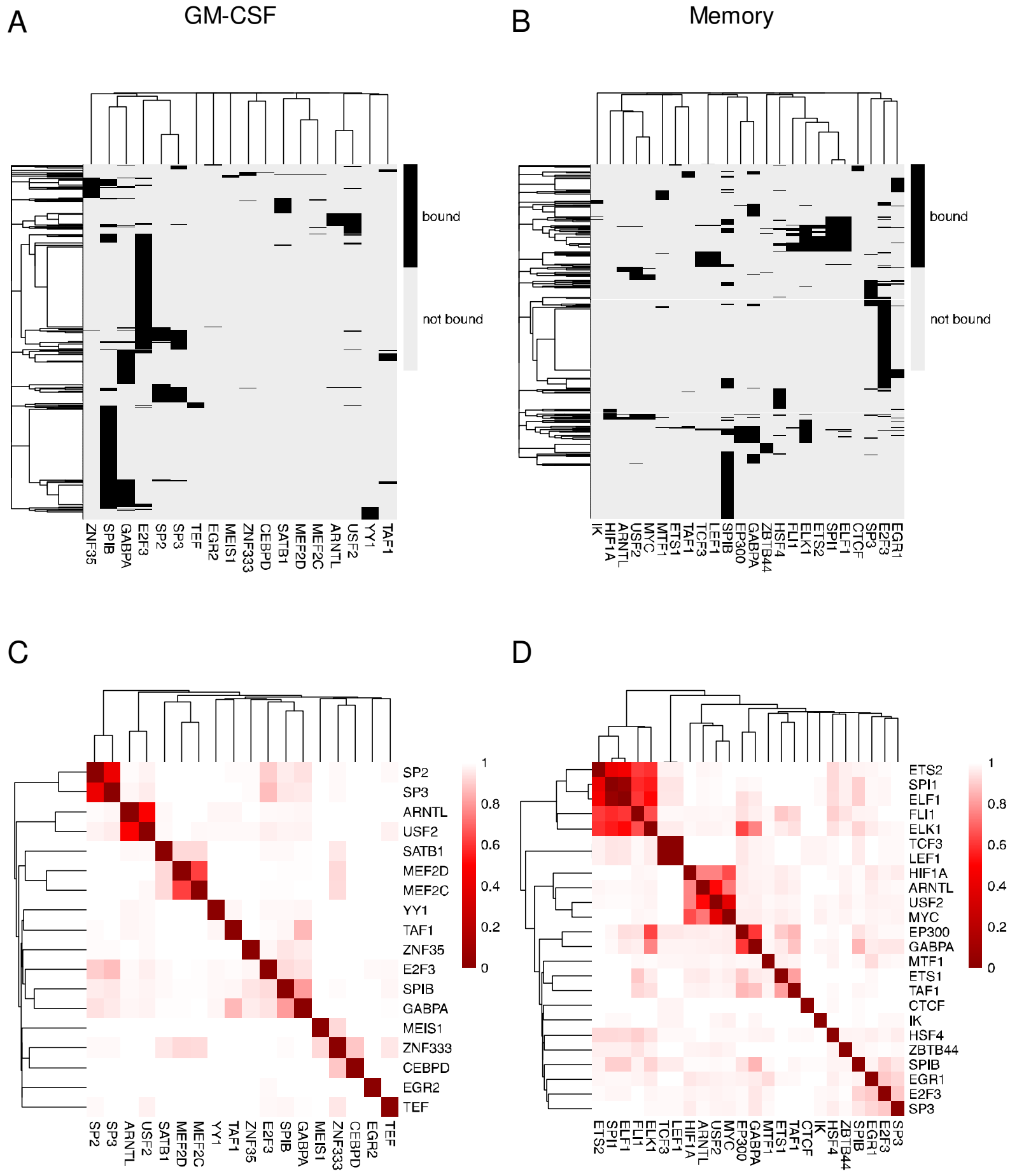
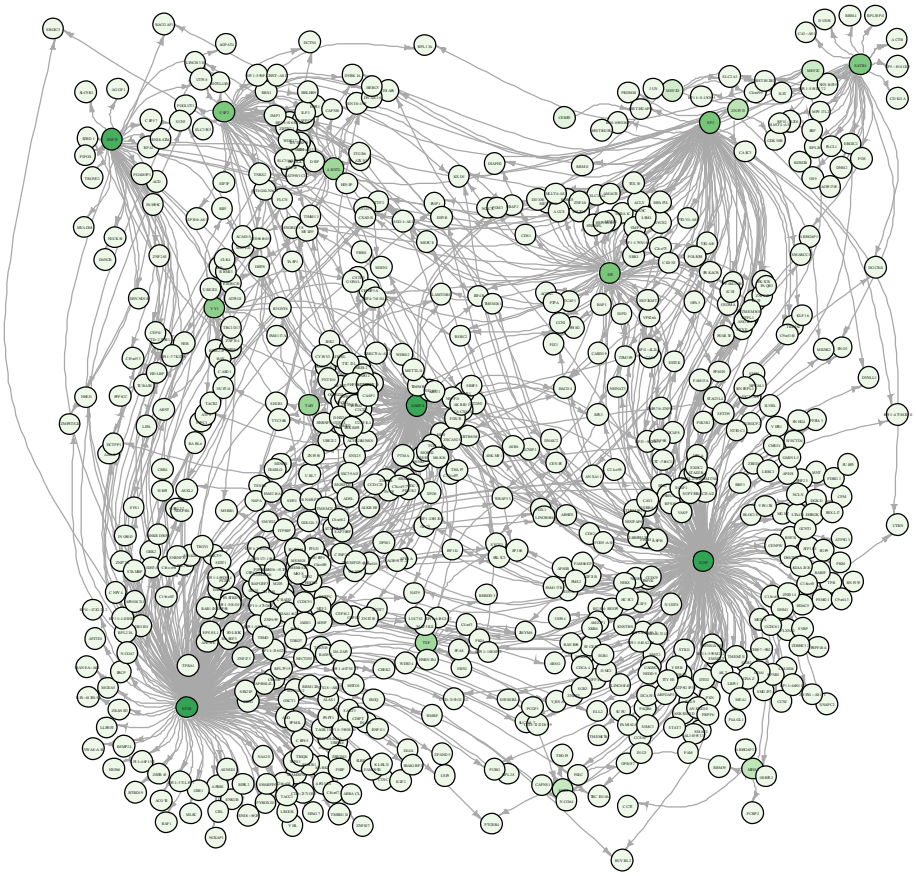
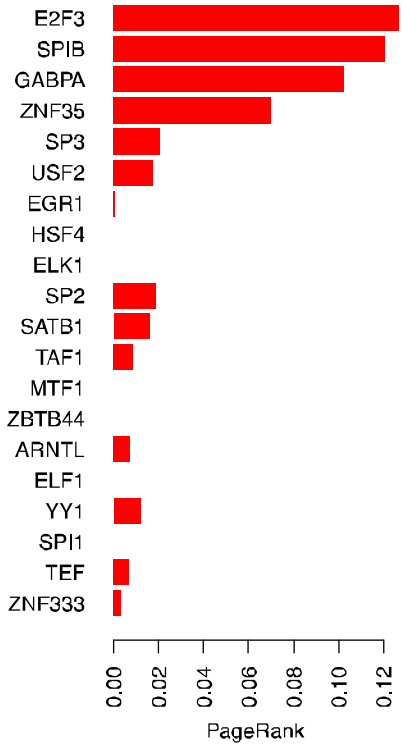

A
GM-CSF
bound
not bound
ZNF35
SPIB
GABPA
E2F3
SP2
SP3
TEF
EGR2
MEIS1
ZNF333
CEBPD
SATB1
MEF2D
MEF2C
ARNTL
USF2
YY1
TAF1
B
Memory
bound
not bound
IK
HIF1A
ARNTL
USF2
MYC
MTF1
ETS1
TAF1
TCF3
LEF1
SPIB
EP300
GABPA
ZBTB44
HSF4
FLI1
ELK1
ETS2
SPI1
ELF1
CTCF
SP3
E2F3
EGR1
C
SP2
SP3
ARNTL
USF2
SATB1
MEF2D
MEF2C
YY1
TAF1
ZNF35
E2F3
SPIB
GABPA
MEIS1
ZNF333
CEBPD
EGR2
TEF
1
0.8
0.6
0.4
0.2
0
D
ETS2
SPI1
ELF1
FLI1
ELK1
TCF3
LEF1
HIF1A
ARNTL
USF2
MYC
EP300
GABPA
MTF1
ETS1
TAF1
CTCF
IK
HSF4
ZBTB44
SPIB
EGR1
E2F3
SP3
1
0.8
0.6
0.4
0.2
0

Figure 9

A

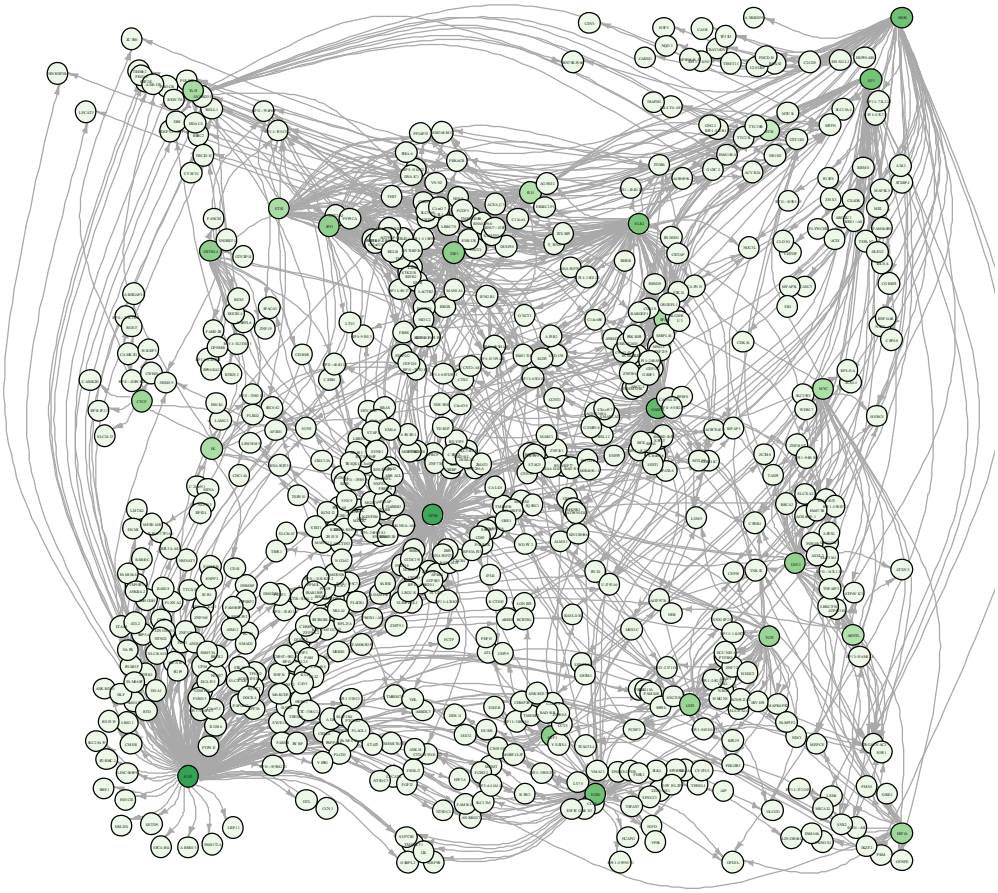
GM-CSF

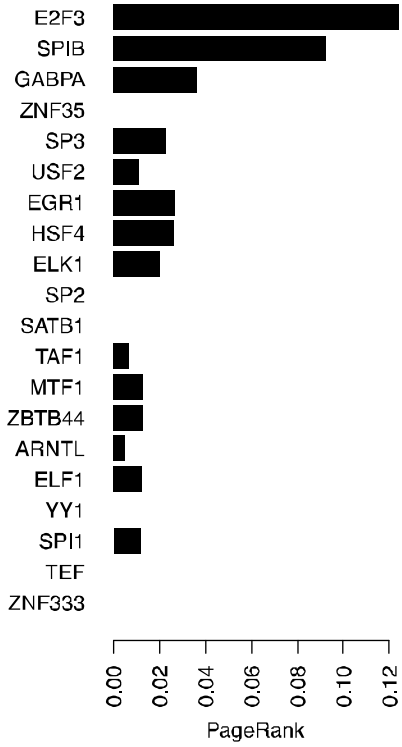




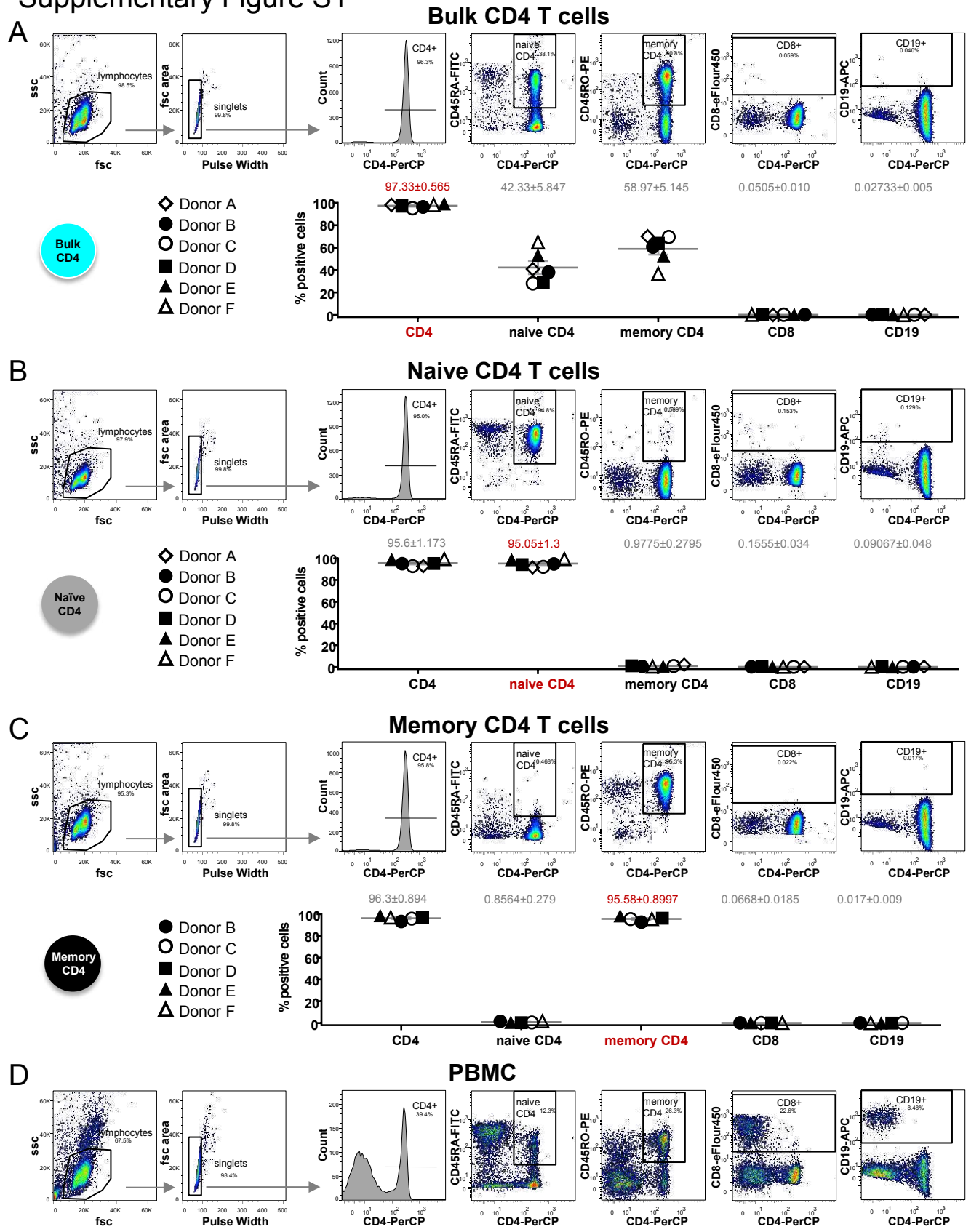
B

Memory



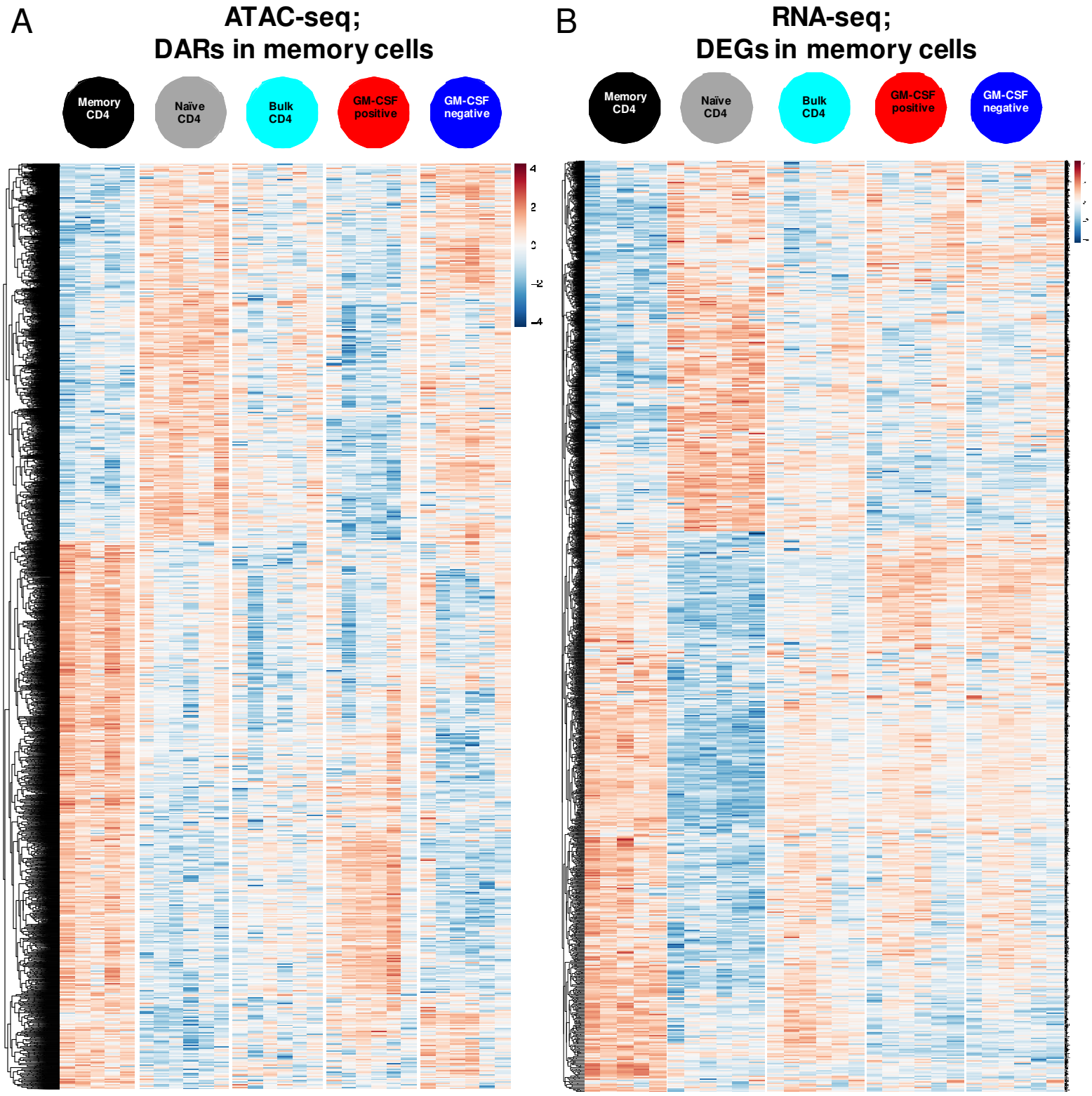


Supplementary Figure S1



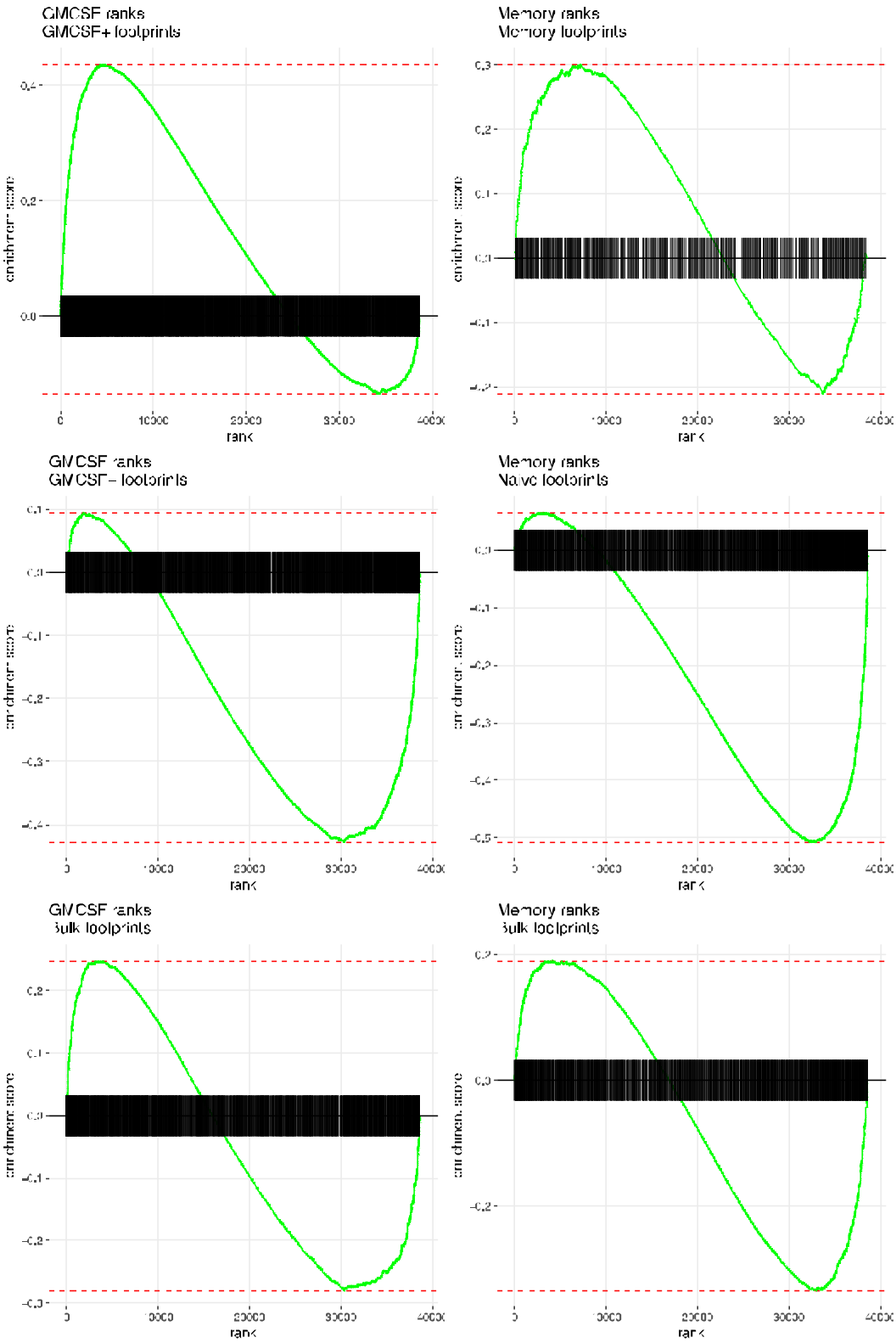
Supplementary Figure S1: Purity of cell populations. Total CD4 T cells (**A**), naive CD4 T cells (**B**) and memory CD4 T cells (**C**) were isolated in parallel by negative magnetic isolation from human PBMCs. Cells were stained for the indicated surface markers and analyzed by flow cytometry. Upper panels show flow cytometry raw data for a representative donor each. The two scatterplots to the left show the pre-gating strategy; the 5 flow cytometry plots to the right are pre-gated on doublet-excluded lymphocytes. Gates to determine percentages of CD4+ cells, naive CD4+ cells, memory CD4+ cells, CD8+ cells and CD19+ cells are indicated. Lower panels show the summary data for the gates defined as above, for n=5 to 6 donors from 4 independent experiments. Same symbols (filled or unfilled) indicate donors processed within the same experiment. Horizontal grey lines represent mean values, and error bars in grey represent SEM. Mean±SEM values are indicated in grey, or highlighted in red font for the most relevant gate for the given isolated cell population. (**D**) shows PBMCs from a representative donor, which were always included in the staining as positive control. Mean±SEM values for PBMCs (n=6 donors) were as follows. CD4+ cell gate: 45.60±3.4%; naive CD4+ cell gate: 19.75±3.6%; memory CD4+ cell gate: 25.53±2.98%; CD8+ cell gate: 22.15±2.44 %; CD19+ cell gate: 7.91±0.64%.

Supplementary Figure S2



Supplementary Figure S2: (**A**) Differentially accessible DNA regions in memory *versus* naive CD4 T cells (FDR<0.05 in this comparison) are plotted as a heatmap for their accessibility in all the five cell populations studied. log2(RPKM+1) is represented by the color scale, indicating accessibility (blue: low, red: high). Data were row-scaled and -clustered (Euclidean distance complete linkage). (**B**) Differentially expressed genes (RNA) in memory *versus* naive CD4 T cells (FDR<0.05 in this comparison) are plotted as a heatmap for their expression in all the five cell populations studied. Gene expression is displayed as log2(FPKM+1) with blue indicating low and red indicating high expression according to the color scale. Data were row-scaled and -clustered (Euclidean distance complete linkage).

Supplementary Figure S3



Supplementary Figure S3. Enrichment of footprints across peaks ranked based on differential accessibility. Peaks were ranked based on their differential accessibility in the indicated contrast using the –log10(FDR) × sign(log2(Fold Change)). Peaks are represented on the horizontal axis and the enrichment score of the set of footprints identified in the indicated cell population is represented on the vertical axis.