Scalable ultra-high-throughput single-cell chromatin and RNA sequencing reveals gene regulatory dynamics linking macrophage polarization to autoimmune disease

Sara Lobato-Moreno[*,1,2,5], Umut Yildiz[*,2,5], Annique Claringbould[*1,2], Nila H. Servaas[1], Evi P. Vlachou[1], Christian Arnold[1], Hanke Gwendolyn Bauersachs[1], Víctor Campos-Fornés[2,5], Karin D. Prummel[1,2], Kyung Min Noh[+,2], Mikael Marttinen[+,1,2,3], Judith B. Zaugg[+,1,2,4]

1) European Molecular Biology Laboratory, Structural Computational Biology Unit
2) European Molecular Biology Laboratory, Genome Biology Unit
3) current address: Faculty of Medicine and Health Technology, Tampere University
4) Molecular Medicine Partnership Unit (MMPU), Heidelberg, Germany
5) Faculty of Biosciences, Collaboration for Joint PhD Degree between EMBL and Heidelberg University, Heidelberg, Germany
*) these authors contributed equally; position of S.L. and U.Y. was determined by coin flipping
+) Co-correspondence: mikael.marttinen@embl.de, kyung.min.noh@embl.de, judith.zaugg@embl.de

Abstract

Enhancers and transcription factors (TFs) are crucial in regulating cellular processes, including disease-associated cell states. Current multiomic technologies to study these elements in gene regulatory mechanisms lack multiplexing capability and scalability. Here, we present SUM-seq, a cost-effective, scalable **S**ingle-cell **U**ltra-high-throughput **M**ultiomic sequencing method for co-assaying chromatin accessibility and gene expression in single nuclei. SUM-seq enables profiling hundreds of samples at the million cell scale and outperforms current high-throughput single-cell methods. We applied SUM-seq to dissect the gene regulatory mechanisms governing macrophage polarization and explored their link to traits from genome-wide association studies (GWAS). Our analyses confirmed known TFs orchestrating M1 and M2 macrophage programs, unveiled key regulators, and demonstrated extensive enhancer rewiring. Integration with GWAS data further pinpointed the impact of specific TFs on a set of immune traits. Notably, inferred enhancers regulated by the STAT1/STAT2/IRF9 (ISGF3) complex were enriched for rheumatoid arthritis-associated genetic variants, and their target genes included known drug targets. This highlights the potential of SUM-seq for dissecting molecular disease mechanisms. SUM-seq offers a cost-effective, scalable solution for ultra-high-throughput single-cell multiomic sequencing, excelling in unraveling complex gene regulatory networks in cell differentiation, responses to perturbations, and disease studies.

Introduction

Gene regulatory elements in enhancer regions play a major role in disease development, as evidenced by the fact that most genome-wide association studies (GWAS) single nucleotide polymorphisms (SNPs) fall in non-coding gene regulatory regions (reviewed in[1]). Moreover, many diseases manifest through an imbalance of cellular differentiation, highlighting the importance of studying enhancer dynamics in the context of cellular differentiation. For example, many autoimmune diseases are characterized by a disbalance of pro- and anti-inflammatory immune cells. Thus, studying such disease mechanisms requires a technology that allows the joint analysis of enhancer and transcription factor (TF) dynamics along a differentiation time course at single-cell resolution.

Recent developments in scalability of single-cell omics technologies, specifically single-cell RNA-seq (scRNA-seq) and single-nucleus (sn)ATAC-seq have revolutionized our understanding on the diversity of cell states and cellular responses to perturbations (reviewed in[2]). Particularly multimodal profiling has enhanced our ability to unravel gene regulatory dynamics governing fundamental biological processes[3–5]. However, while scalability for individual modalities has become available (e.g. scifi-RNA-seq[6], sci-RNA-seq[7], dsci-ATAC-seq[8], sci-ATAC-seq[9]), the majority of current multiomic methods are limited in scalability, multiplexing capability, or cost effectiveness (e.g., 10x multiome, ISSAAC-seq[10]), with a handful providing scalability in terms of number of cells assayed (e.g., SHARE-seq[11], Paired-seq[12]), but at the expense of complexity (**Table 1**).

Here, we present SUM-seq (**S**ingle-cell **U**ltra-high-throughput **M**ultiomic sequencing): a cost effective and scalable sequencing technique for multiplexed multiomics profiling. SUM-seq enables simultaneous profiling of chromatin accessibility and gene expression in single nuclei at ultra-high-throughput scale (up to millions of cells and hundreds of samples[6]). To this end we refined the two-step combinatorial indexing approach, originally introduced by Datlinger et al.[6] for snRNA-seq, tailoring it to the multiomic setup. In brief, accessible chromatin and nuclear RNA first receive a sample-specific index by transposition and reverse transcription, respectively. Subsequently, a second index is introduced using a droplet-based microfluidic platform (e.g. 10X Chromium). The dual indexing allows overloading of microfluidic droplets while retaining the capacity to assign matched RNA- and ATAC-seq reads to the same individual cell.

We demonstrate the application of SUM-seq in two independent experimental setups, a species-mixing benchmarking experiment and a macrophage M1 and M2 polarization time course experiment, exemplifying its flexibility to accommodate complex experimental setups. We use both modalities to resolve temporal patterns of gene regulation, and integrate with genetic evidence to link regulatory networks to disease. Additionally, we validated the feasibility of decoupling sample collection from library generation, rendering it well-suited for extensive atlas projects, time course experiments, and other experimental designs that involve prolonged sample collection periods.

Results

Overview of the SUM-seq method

SUM-seq is built upon the concept of a recently developed combinatorial fluidic indexing scRNA-seq approach, scifi-RNA-seq[6], and a low-throughput multiomic assay, ISSAAC-seq[10]. First, nuclei are isolated and fixed with glyoxal, and split into equal bulk aliquots in a multi-well format. In the second and third steps, unique sample indices (one per well) are introduced for the ATAC and RNA modalities. For ATAC, accessible genomic regions are indexed by Tn5 molecules loaded with barcoded oligos. For RNA, the mRNA molecules are indexed with barcoded oligo-dT primers via reverse transcription (**Figure 1a, detailed library structures in Extended Data Figure 1a**,**b, Supplementary table 1**). In step four, samples are pooled for Tn5 tagmentation of the cDNA-mRNA hybrids to introduce a primer binding site necessary for binding of the barcoded oligonucleotides in the microfluidic system. In step five, nuclei are overloaded onto the microfluidic system (e.g., 10X Chromium), facilitating the encapsulation of multiple nuclei in a single droplet. In the droplets, fragments are barcoded with the microfluidic barcode, resulting in dual barcoding of fragments with both a sample and a droplet barcode enabling downstream demultiplexing and assignment of sequencing reads to individual nuclei. This is followed by a PCR reaction using a combination of three primers to

amplify fragments of both the ATAC and RNA modalities, after which the library is split into two equal proportions and subjected to modality-specific amplification. At this stage, a library index is introduced to enable multiplexing libraries on the sequencer. snATAC and snRNA libraries are sequenced using standard Illumina sequencing primers. A detailed protocol for SUM-seq is available as a supplementary document. Additionally, we have developed a scalable and reproducible Snakemake pipeline for processing SUM-seq data (https://git.embl.de/grp-zaugg/SUMseq - accessible upon request/publication), in which reads are assigned to sample indices and further demultiplexed to single-cell resolution by the droplet barcode (**Figure 1b**). From here, reads are mapped and a gene expression matrix and tile matrix are generated for RNA and ATAC, respectively. Modalities can be matched based on the assigned sample index-cell barcode combinations. The pipeline provides reproducibility, high scalability and flexibility for execution on any system (local, cluster, cloud).

SUM-seq enables efficient and scalable single-cell multiomics analysis

To evaluate performance of our method and quality of the data, we performed SUM-seq on an equal mixture of human leukemia (K562) and mouse fibroblast (NIH-3T3) cell lines. For this species-mixing benchmarking experiment we loaded 100,000 nuclei into a single channel of the Chromium system which equates to ~7-fold overloading compared to the standard 10X workflow. In order to limit the sequencing requirements, we processed only 20% (20,000 nuclei) of the generated microfluidic droplets for final library preparation. After combinatorial index demultiplexing, human and mouse reads were well separated with a collision rate of 0.1 % (UMIs) and 3.8 % (ATAC fragments), resulting in 6,215 human cells and 7,607 mouse cells with data from both modalities (**Figure 1c**,**d**). This translates to a ~70 % recovery rate of input and represents a ~7-fold increase in throughput at low collision rates compared to the standard 10X workflow (**Extended Data Figure 1d**). Key performance metrics of snRNA (UMIs and genes per cell) and snATAC (fragments in peaks per cell, TSS enrichment score, fragment size distribution) were consistently of very high quality for both modalities in SUM-seq and outperformed other high-throughput assays for scRNA, snATAC, and multiomic approaches (**Figure 1e, Extended Data Figure 2c-g**). Notably, there was no drop in data quality for nuclei in overloaded droplets compared to droplets containing a single nucleus (**Figure 1f**). Lastly, the aggregate of snRNA and snATAC data resembled the published bulk RNA-seq and ATAC-seq in the K562 benchmarks from ENCODE data (**Figure 1g**).

During protocol optimization, we evaluated the effect of: I) adding polyethylene glycol (PEG), II) strategies to mitigate barcode hopping, and III) freezing of samples on the data quality. (I) Consistent with previous studies[14,15], the addition of 12% PEG to the reverse transcription reaction increased the number of UMIs and genes detected per cell (~2.5- and ~2-fold, respectively) with minor impact on the quality metrics of the ATAC modality (**Extended Data Figure 2h**,**i**). (II) Recently, it has been reported that barcode hopping can occur within multinucleated droplets[16,17]. In SUM-seq, barcode hopping primarily affects the ATAC modality and is mitigated by two complementary strategies: adding a blocking oligonucleotide[17] in excess to the droplet barcoding step and reducing the number of linear amplification cycles from twelve to four (**Methods**) (**Figure 1b,c** and **Extended Data Figure 2a**). (III) To assess the suitability of SUM-seq for asynchronous sample collection (e.g. time course data, clinical samples, etc.), we tested glycerol-based cryopreservation following glyoxal fixation and found that it has minimal impact on the performance metrics of the assay (**Extended Data Figure 2j**). Thus, SUM-seq can be employed for gathering, fixing and cryopreserving samples prior

to library construction, which enables multiplexing samples in one experiment thereby reducing costs and batch effects.

Figure 1. SUM-seq allows simultaneous profiling of chromatin accessibility and gene expression in single cells at ultra-high-throughput scale. a, Schematic depiction of the SUM-seq workflow. Key steps and detailed structures are described in the main text and Extended Data Figure 1. **b,** Schematic overview of the computational analysis pipeline. **c, d,** Species-mixing plots for the ATAC (**c**) and the RNA (**d**) modality, indicating the fraction of reads from singlets and multiplets assigned to the mouse genome (mm10, y-axes) and the human genome (hg38, x-axes). Genomic DNA fragments (**c**) and transcripts (**d**) are demultiplexed based on the combination of the sample index and the microfluidic index. The pie chart shows the fraction of human and mouse cells as well as the frequency of multiplets and singlets. Collision rates are highlighted in red. **e,** Number of accessible DNA fragments in peaks

(FRiP = fragments in peaks; upper panel) and genes detected (lower panel) per cell are shown as violin plots for SUM-seq and other methods (dsci-ATAC-seq[8]: GSM3507387; SHARE-seq[11]: GSM4156590 (ATAC), GSM4156602 and GSM4156603 (RNA); Paired-seq[12]: GSM3737488 (ATAC), GSM3737489 (RNA); scifi-RNA-seq[6]: GSM5151362). Median values are shown on top. **f,** Distribution of accessible DNA fragments in peaks per cell and the TSS score per cell (upper panels) as well as UMIs and genes detected per cell (lower panels) are shown as violin plots, split by the number of nuclei in a droplet. The number of droplets with N nuclei encapsulated is shown on the top. **g,** A representative genome browser view (*chr12*: *52,334,173 - 55,349,410*) of SUM-seq data is shown for K562 as aggregated transcript and ATAC fragment mappings. For comparison, K562 ATAC-seq and RNA-seq data tracks were retrieved from the ENCODE database (tracks labeled in green; GSE86660[13] (RNA), GSE170214[13] (ATAC)). The bottom panel shows binarised accessibility for 200 randomly selected K562 nuclei at single-cell resolution. Boxplots in **e**,**f**: center line represents median; lower and upper hinges represent the 25th and 75th quartiles respectively.

SUM-seq recapitulates regulatory dynamics during M1/M2 macrophage polarization

Macrophages are innate immune cells that can polarize towards a pro-inflammatory M1 or an anti-inflammatory M2 state depending on microenvironmental signaling. Despite the inherent complexity and heterogeneity observed *in vivo*, the M1/2 dichotomy serves as a useful framework for elucidating major transcriptional events and regulatory elements governing macrophage polarization. Several key TFs, such as STAT1, IRF5 and NF-κB for M1, and STAT6, PPARγ, and IRF4 for M2 have been described[18–20]. However, the complex regulatory networks orchestrating macrophage polarization over time remain elusive.

To unravel these intricate regulatory networks, we applied SUM-seq to profile hiPSC-derived macrophages across a polarization time course from the naive M0-state to the M1- and M2-states. We stimulated the M0 macrophages with LPS and IFN-γ to induce M1 polarization, or IL4 to induce M2 polarization. To discern early and sustained responses at chromatin accessibility and gene expression levels, we collected samples at five time points along the two polarization trajectories; prior to stimulation (M0) and at 1-hour, 6-hour, 10-hour, and 24-hour intervals, each sampled in duplicates totaling 18 samples **(Figure 2a)**.

Leveraging the overloading capacity of SUM-seq, we loaded 150,000 nuclei into a single microfluidic channel of the 10X Chromium system, resulting in 51,750 high-quality nuclei evenly distributed across samples (**Figure 2b**). Cells passing the QC filters (**Methods**) exhibited an average of 11,900 unique fragments, a TSS score of 8, and 40% reads in peaks. For the snRNA-seq readout, cells displayed on average 407 UMIs and 342 genes (**Extended Data Figure 3a-c**). The lower average UMI and gene counts per cell in the macrophage experiment, in comparison to the mixed-species experiment, are a consequence of the absence of PEG during reverse transcription due to a technical issue (**Methods**). However, the high number of nuclei that passed the quality filter enabled effective downstream analyses.

We obtained a joint visualization of the two modalities by constructing a weighted nearest-neighbor (WNN) graph, which was further used to determine a uniform manifold approximation and projection (UMAP) (**Methods**). The resulting UMAP separated cells along the M1 and M2 polarization trajectories (**Figure 2b**). By calculating M1 and M2 scores from the expression of the literature-based marker genes (**Supplementary Table 2; Methods**), and mapping them onto the UMAP, we observed an increased M1 and M2 score during the respective polarization trajectories (**Figure 2c**), which confirms the validity of our experimental setup. Notably, M2 cells were less distinguished from the M0 state, and underwent only gradual changes throughout the time course. In contrast, during M1 polarization, we observed a clear separation of cells from the M0 state, with a substantial shift between the 1-hour and

6-hour time points. This underscores the capacity of our system to effectively capture distinct early and late polarization stages.

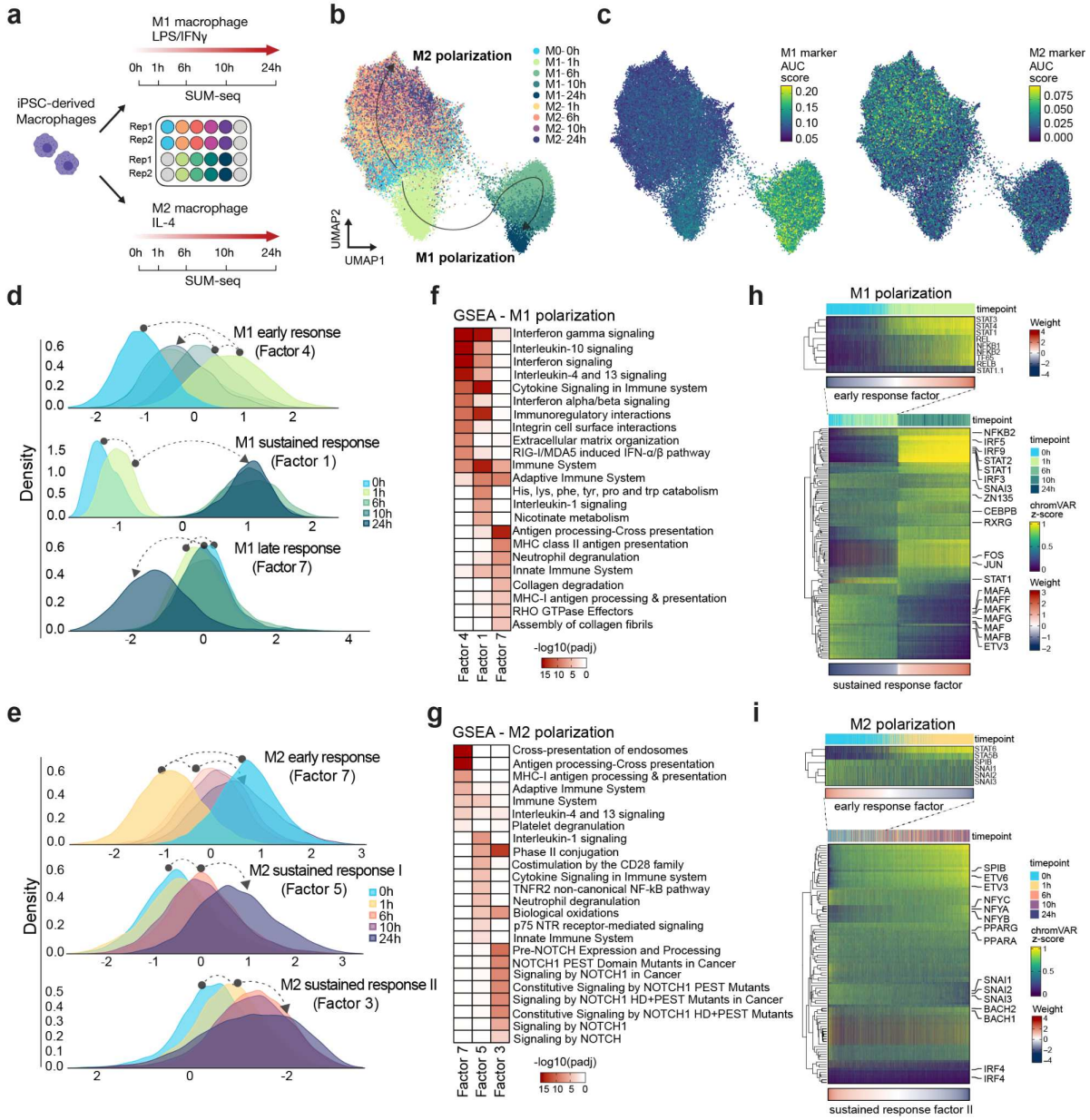

Figure 2. Integrated chromatin accessibility and gene expression data at single-nucleus resolution via SUM-seq characterizes differentiation trajectories along M1/M2 macrophage polarizations. a, Schematic overview of the macrophage polarization experiment. hiPSC-derived macrophages were stimulated with LPS and IFN-γ (M1) or IL4 (M2). Nuclei were fixed with glyoxal and collected for SUM-seq at 0h, 1h, 6h, 10h, and 24h (*n*=2 per time point). **b-c,** Weighted nearest neighbor (WNN) UMAP projection of integrated SUM-seq data of macrophage polarization. Cells are annotated and labeled according to their sample index (**b**), AUC score of M1 signature genes (**c** left panel) and AUC score of M2 signature genes (**c** right panel; Supplementary table 2). **d**,**e,** Distributions of cells from each time point across three MOFA factor weights associated with M1 polarization (**d;** early response, sustained response, late response) and M2 polarization (**e**; early, sustained response I and II). Dotted arrows indicate the direction of the time-resolved response. **f**,**g,** Gene set enrichment analysis (Methods) for M1 (**f**) and M2 (**g**) polarization factors are shown as heatmaps. **h,** Motif activity for TFs associated with M1 early response (Methods) across M0 and M1-1h cells sorted by M1 early response factor weights (top). Motif activity for TFs associated with M1 sustained response across all M0 and M1 cells sorted by sustained factor weights (bottom). **i** Motif activity for TFs associated with M2 early

response across M0 and M2-1h cells sorted by M2 early response factor weights (top). TF motif activity for TFs associated with M2 sustained response II across all M0 and M2 cells sorted by sustained factor II weights (bottom).

To delineate the key features in terms of gene expression and chromatin accessibility that are influencing macrophage M1/M2 polarization, we performed Multi-Omics Factor Analysis (MOFA)[21] to establish a shared low-dimensional representation for the snATAC- and snRNA-seq data (**Extended Data Figure 3d,e**). Investigating the association of the MOFA factors with metadata, we pinpointed three factors associated with M1 polarization, explaining the early response (Factor 4), sustained response (Factor 1) and late response (Factor 7) (**Figure 2d, Extended Data Figure 3d**). Notably, the early response was predominantly driven by chromatin accessibility, while the sustained and late responses were influenced by both gene expression and chromatin accessibility (**Extended Data Figure 3d**). For M2 we identified an early response (Factor 7) and two distinct sustained response factors (Factors 3 and 5) (**Figure 2e**). For functional interpretation of the MOFA factor associated genes, we performed gene set enrichment analysis (**Supplementary Table 3, Methods**). Genes associated with the early M1 response were enriched in IFN-γ and cytokine signaling (**Figure 2f**). IFN signaling was also enriched in genes associated with the sustained response (Factor 1), along with general immune system processes, IL1 signaling, and metabolic terms in line with metabolic reprogramming previously observed in M1 macrophages[22]. Genes associated with the M1 late response were predominantly enriched for terms related to antigen (cross-) presentation and RHO signaling (**Figure 2f**). The early M2 polarization factor was enriched for IL4 signaling, while the two M2 sustained factors were enriched in biological oxidation - likely signifying the switch in metabolism of M2 macrophages[23], and non-canonical NF-κB-signaling and Notch signaling, respectively (**Figure 2g**).

SUM-seq uncovers key TF motif activity dynamics along the M1/ M2 polarization

We investigated what TFs underlie the chromatin remodeling during M1 polarization. For this, we quantified TF motif accessibility using chromVAR[24], referred to as TF motif activity hereafter, and selected the top 10% most variable TFs. We further filtered for TF motifs that are enriched in peaksets associated with either the early, the sustained, or the late response of M1 (**Extended Data Figure 4a**), resulting in 103 unique TFs (139 motifs; **Supplementary table 4**).

When plotting the chromatin-based TF motif activity in cells from M0 and M1-1h along the early response trajectory, we observed increasing motif activity for prototypical M1 TFs, such as STAT1 and IRF5[20], followed by NF-κB (NFKB1, NFKB2, RELA, REL, TF65; **Figure 2h**). Notably, activity of STAT1, as well as other IFN-II response TFs such as IRF7, preceded IRF5 and NF-κB activity, highlighting their role as immediate response factors (**Figure 2h**). Furthermore, we observed a transient upregulation of the AP-1 complex motifs (JUN, FOS, etc), indicative of a cellular activation state, even in the unpolarized M0 state. Meanwhile, the myeloid differentiation factors of the ETS family, including SPI1 (PU.1)[25], were already active in M0 macrophages and slightly decreased their activity upon polarization. To investigate the M1 response at later time points, we ordered all M1 cells along the sustained response factor. This revealed a switch-like increase in motif activity for many prototypical M1 TFs, including STATs and IRFs, such as IRF5[19]. STAT1 can function as a homodimer or together with STAT2 and IRF9 to form the ISGF3 complex[26]. These two STAT1 states bind to different regions in the DNA and are consequently characterized by different binding motifs. Notably, we observed a steady drop in STAT1 homodimer motif activity over time, which coincided with a marked

increase of its heterodimer motif as well as those of STAT2 and IRF9. These observations highlight the role of STAT1 homodimer in initiating chromatin remodeling in response to IFN-γ. Meanwhile the sustained response is maintained by partnering with STAT2 and IRF9 in the ISGF3 complex, the driver of type I IFN response[26]. In contrast to the switch-like dynamic of the IRFs and STATs, the NFKB2 motif activity presented a more gradual increase along the M1 polarization. Additional factors like AP-1 and ETS family members also exhibited switch-like activity patterns, increasing and decreasing with time, respectively. In the late response phase, we observed only subtle changes, including a slight increase in NFY motifs and a slight decrease in CEBP motifs (**Extended Data Figure 4c**). This pattern suggests that most TF activity stabilizes between 6 and 24 hours of M1 polarization.

A similar analysis of TFs along the M2 polarization focused on 93 TFs (121 motifs) (**Extended Data Figure 4b, Supplementary table 4**). Overall, we observed less striking dynamics along the M2 polarization trajectories compared to the M1 response. The most notable pattern was observed for the IL4-responsive TF STAT6[27], which showed increasing motif activity along the early M2 response trajectory (**Figure 2i**). The sustained response patterns were mostly characterized by a steady increase in motif activity for ETS family members, most notably SPI1[28], ETV transcription factors, such as ETV3 and ETV6[29], and increased activity of NFY factors (**Figure 2i, Extended Data Figure 4d**). Conversely, SNAIL family member motif activity decreased along the sustained response. Likewise, CTCF accessibility went down upon M2 polarization, suggesting a global rewiring of chromatin architecture upon macrophage polarization[30].

Gene regulatory network analysis reveals sustained response by STAT1/STAT2/IRF9 in M1 macrophages

To understand the regulatory network dynamics, including the genes that are regulated by the TFs identified above, we constructed an enhancer-mediated gene regulatory network (eGRN) using GRaNIE[31] (**Supplementary Table 5; Methods**). Of the 103 TFs we found associated with M1 polarization, we recovered 40 in the eGRN either as TF itself or as target gene of the M1-TFs (**Figure 3a; Extended Data Figure 5b**). The eGRN shows that the most connected TFs share many of their target genes. For example, IRF1 and IRF8, connected to 1385 and 776 genes respectively, share 679 of their regulon genes.

We further resolved the hierarchy of a TF's activity on chromatin accessibility and their effect on inducing transcription by comparing motif and regulon activity similarity along the M1 immediate response factor (**Methods**). For the majority of TFs, we found a positive relation between their motif activity and expression of their target genes derived from the eGRN (**Extended Data Figure 6a**). The two evident exceptions to this were STAT1 and NFKB1. For NFKB1, we observed a delayed expression of the regulon despite the gradual increase in NFKB motif activity from an early time point (**Figure 3b**). STAT1 has two motifs in the HOCOMOCO database: one binds the DNA as a STAT1 homodimer (STAT1.H12INVIVO.0.P.B), and the other (STAT1.H12INVIVO.1.P.B) is highly similar to the motif of STAT2. In the JASPAR database[32], this second motif is classified as the interaction of STAT1 and STAT2. As such, herein, we refer to the second motif as the STAT1 heterodimer motif. We found that STAT1 regulon activity was discordant with the chromatin activity for the STAT1 homodimer motif, but it closely tracked the STAT1 heterodimer motif activity (**Figure 3b,c**).

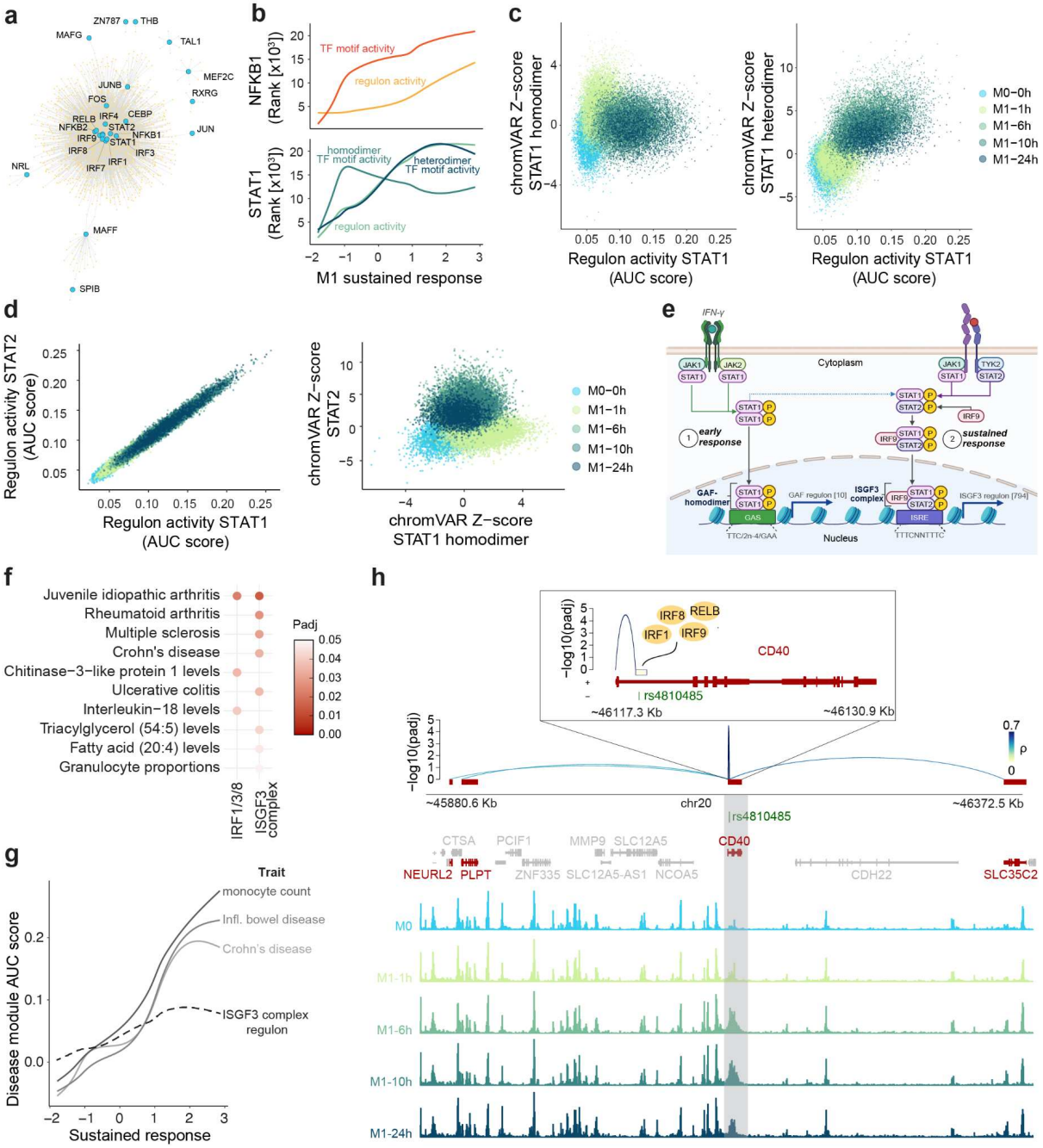

Figure 3. Gene regulatory network analysis of SUM-seq data coupled with genetic evidence reveals TF and regulon hierarchy, linking TFs to immune traits. a, Gene regulatory network visualization of the 24 M1 polarization-associated TFs and their assigned target genes. **b,** TF motif activity (chromVAR Z-score) and regulon activity (AUC cell score) for NFKB1 (top) and STAT1 (bottom) motifs along the M1 sustained response factor. Cells are ranked by their respective value for activity measures, lines show best fit (generalized additive model). **c,** Scatter plots of the correlation between the STAT1 regulon activity and STAT1 homodimer (left) and STAT1 heterodimer (right) TF motif activity (chromVAR Z-score). Points represent cells and are coloured by their experimental time point. **d,** Scatter plots of each cell across the M1 macrophage polarization showing the correlation between STAT1 and STAT2 regulons (left) and STAT1 homodimer and STAT2 TF motif activity (right). Points represent cells and are coloured by their experimental time point. **e,** Schematic type II IFN-associated STAT1 homodimer (known as GAF) response (early response) and type I IFN-associated STAT1/STAT2/IRF9 (known as ISGF3 complex) response (sustained response). IFNγ binding to its receptor triggers an early response driven by the phosphorylation and subsequent dimerization of STAT1, forming GAF. The GAF-homodimer enters the nucleus and binds the GAS motif, activating the

expression of 10 downstream targets identified in our eGRN. As a secondary response, STAT1, STAT2 and IRF9 get phosphorylated and activated forming a complex named ISGF3 complex. The ISGF3 complex binds the ISRE motif, potentially activating the expression of 794 downstream genes identified in our eGRN. **f,** Identification of diseases and traits associated with genetic variants enriched in open regions of two highly connected regulons in the eGRN (IRF1/3/8 and the STAT1/STAT2/IRF9 - putative ISGF3 complex) using linkage disequilibrium score regression (LDSC). All open regions in macrophages are used as background. **g,** AUC cell scores for disease modules derived by intersection of putative genome-wide SNPs for the top enriched diseases with the putative ISGF3 complex regulon (union of STAT1, STAT2, IRF9 gene targets) (**Methods**), compared to AUC cell scores for all ISGF3 regulon genes along the M1 sustained response factor. **h,** eGRN peak-gene interactions for the CD40 intronic peak intersecting RA-associated SNP rs4810485 (top), zooming in to the CD40 region with the TFs binding to this peak (insert). M0 and M1 cell aggregate ATAC-seq tracks split by timepoint highlighting the CD40 intronic peak (bottom).

While the STAT1 homodimer (also called IFN-y activated factor (GAF)) motif response on the chromatin level was strongest at the early time point of 1h post stimulation and decreased from 6h onwards, the STAT1 heterodimer motif, along with STAT2 and IRF9 motif activity, increased gradually over time (**Extended Data Figure 6b,c**). This pattern was also reflected by the STAT1 regulon. The shared TF activity patterns for the STAT1 heterodimer, STAT2 and IRF9 motifs can potentially be attributed to the fact that these TFs together form the ISGF3 complex, which binds to the DNA in response to type-I interferon[26] (**Figure 3e**). Moreover, most genes in the STAT1 regulon were connected via the heterodimer motif, with only 10 target genes connected to the homodimer motif (**Supplementary Table 6**). Notably, the homodimer targets included genes that suppress NF-κB signaling, including TNFAIP3[33,34] and ZNF598[33,34].

The early STAT1 homodimer response is recapitulated by its phosphorylation pattern: reanalysing data from He et al[35], we observed that STAT1 becomes phosphorylated at sites T699 and Y701, almost immediately upon IFN-y/LPS stimulation of THP1 derived macrophages, followed by a gradual decrease over time (**Extended Data Figure 6g**). Indeed, phosphorylation of these sites is known to be induced by IFN signaling and trigger accumulation in the nucleus and activation of STAT1 DNA binding activity[36,37]. In contrast to this, the gradual increase in STAT2 and IRF9 activity is reflected by a delayed but sustained increase in the phosphorylation of IRF9 at sites S131 and S253 (**Extended Data Figure 6h**). S253 is known to be induced by IFNβ and may play a role in regulating the expression of interferon stimulated genes (ISGs) and interactions with STAT2[38]. Although the target genes of STAT1, STAT2 and IRF9 did not overlap completely, their regulon activity was highly correlated, providing evidence that the ISGF3 complex was captured by the eGRN (**Figure 3d**; **Extended Data Figure 6d-f**).

In summary, integrating chromatin accessibility and gene expression data at single-cell resolution over a time course, we characterized the M1 and M2 polarization trajectories. This highlights SUM-seq's capability to provide detailed insights into cellular states and TF activities. Using both ATAC and RNA modalities along a time course enabled us to dissect the regulatory layers of M1 polarization, leading to the identification of a putative hierarchy of IFN-driven responses. Specifically, we observed a marked shift in STAT1-mediated regulation, from its homodimer-driven chromatin remodeling during the early M1 polarization, to an ISGF3-driven response reflected in both accessibility and transcription at later time points. This switch is in line with previous reports on the timing of IFN stimulations[39,40] (**Figure 3e**).

GWAS enrichment analyses of the SUM-seq identified eGRN

To further investigate the role of the macrophage network in disease, we tested whether the chromatin accessible regions linked to regulon genes in the network were enriched for heritability of specific diseases using linkage disequilibrium score regression (LDSC)[41]. Overall, we found all open regions in our macrophage data set strongly enriched for white blood cell count traits and autoimmune diseases (**Extended Data Figure 7a**). Using all these open regions as background, the peaks in the putative ISGF3 regulon (union of STAT1, STAT2 and IRF9 gene targets) were enriched for specific autoimmune related traits including juvenile idiopathic arthritis (JIA), ulcerative colitis and Crohn's disease, multiple sclerosis and rheumatoid arthritis (RA; **Figure 3f**). The IRF1/3/8 regulon (another set of highly connected TFs), in turn, was enriched for JIA, chitinase-3-like protein 1 levels and IL18 levels. The enrichment for chitinase-3-like protein 1 levels showcases an example where the eGRN picked up a direct molecular association since the SNP (rs872129) is located upstream of the CHI3L1 gene encoding for chitinase-3-like protein 1 (**Extended Data Figure 7b**). IL18 is a pro-inflammatory cytokine that is transcribed by a complex including IRF1[42] and was identified as a direct IRF1 target in our network. IL18 mediates the autoinflammatory macrophage activation syndrome involved in JIA pathogenesis, and it is currently investigated as a therapeutic target of inflammatory diseases[43].

These enrichments suggest that we identify the TFs that play an important role in autoimmune diseases involving macrophages. Indeed, mapping the putative genome-wide SNPs (p < 10^-6) for the top enriched diseases (p < 0.1) to the STAT1-STAT2-IRF9 regulon resulted in prioritizing 15 genes each for Crohn's disease, inflammatory bowel disease and monocyte counts, as well as 4 genes for RA (**Supplementary Table 7; Methods**). We then calculated the expression activity for these disease genes (AUC cell score) and projected them along the M1 sustained response factor. We observed that the disease gene sets show a stronger increase in activity over time than the general STAT1-STAT2-IRF9 regulon (**Figure 3g**). The strongest (most significant) SNP in RA that overlapped with our eGRN, rs4810485, is located in the intron of CD40, and is linked to CD40, PLTP, NEURL2 and SLC35C2 in our eGRN (**Figure 3h**). Previous work has linked this SNP to CD40 as an expression, protein and splice quantitative trait locus (e-, p-, sQTL) in blood and monocyte-specific datasets, as well as to PLTP as a blood eQTL[44] (ref) (**Supplementary Table 8)**. PLTP is overexpressed in synovial tissue from RA patients, and is thought to regulate NF-κB/STAT3 in macrophages[45]. CD40 is a cell surface receptor that interacts with CD40 ligand (CD40-L), involved in the pathogenesis of RA and other autoimmune diseases[46]. CD40 expression can be induced by IFN-γ stimulation in macrophages and is expressed in synovial monocytes in RA[47,48]. Our findings recapitulate the link between the ISGF3 complex, CD40, and RA, and provide further evidence for the importance of PLTP in this disease.

In summary, integration of our high-resolution single-cell multimodal time course data of M1 polarization with genetic disease evidence revealed molecular disease mechanisms.

Discussion

SUM-seq is a single-cell multiomic method that allows highly multiplexed and scalable experimental setups to investigate the molecular underpinnings of gene regulation via profiling of chromatin accessibility and gene expression from the same cell. Here, we applied SUM-seq to study macrophage polarization. We showcase how the multimodal data from a single multiplexed experiment can be used to study the molecular underpinnings of TF-driven disease mechanisms, integrating TF response dynamics with genetic evidence from GWAS.

We benchmarked SUM-seq showing its capability to effectively capture both chromatin accessibility and gene expression of single cells, outperforming published scalable high-throughput methods. We demonstrate its capability to generate high quality data from both fresh and cryopreserved samples, making it adaptable for e.g. atlas studies, where sample collection can span a long period of time and requires multi-center efforts. While SUM-seq doesn't reach the gene expression complexity of lower throughput droplet-based single-cell multiomic methods, it provides matching chromatin accessibility complexity paired with a significant improvement in cost-efficient scalability. Furthermore, we anticipate that through minor optimisations 1) gene expression complexity will further improve and 2) SUM-seq can expand to incorporate surface proteomic readouts by use of single cells instead of single nuclei[49]. Furthermore, it should be noted that although the current format measures chromatin accessibility and gene expression, the strategy can be adopted for the measurement of omic layers such as TF binding and histone modifications (Cut&Tag[50]), DNA methylation (sci-MET[51]), and whole genome sequencing (s3-WGS[52]).

By analyzing the activity of TF motifs on the chromatin versus the expression of associated TF target genes, we observed a hierarchical activation of a TF cascade upon IFN-γ/LPS stimulation of M0 macrophages. Specifically, we identified TFs that clearly act in a stepwise manner (i.e. STAT1), whereas others have a continuous activation profile (i.e. NFKB1). We could recapitulate the hierarchy of STAT1 homodimer activity preceding STAT1/STAT2/IRF9 activity. Moreover, some of the STAT1 homodimer target genes (including TNFAIP3[33], and ZNF598[34]) are involved in modulating NF-κB signaling through the ubiquitin-mediated degradation pathway, potentially explaining the long time gap in NF-κB chromatin activity and the expression of its target genes. An alternative explanation for this delay is the previously reported observation that enhancers of NF-κB response genes are primed by NF-κB but their expression additionally requires activity of the IFN-I-induced ISGF3 complex[53].

Genes with genetic links to disease have been shown to be better drug targets [54]. Yet for the majority of disease-associated genetic variants, we lack molecular mechanisms, and it is often unclear in which cell type or disease state a variant exerts its effect. Most disease-associated common genetic variants lie in non-coding regulatory elements that are presumably regulated by TFs. Indeed, TFs are often identified as main drivers of disease-relevant phenotypes in genetic screens[55]. Thus, to understand disease mechanisms it is crucial to integrate genetic evidence with disease-specific gene-regulatory dynamics. We showcase how SUM-seq can be used in a framework that pinpoints gene-regulatory dynamics and links them to genetic disease evidence in one single experiment. Specifically, profiling the macrophage polarization response in a single SUM-seq experiment, followed by integrative analysis of gene regulatory dynamics and publicly available summary statistics from GWAS, identified several known disease mechanisms, as well as novel associations. For example, we find an association between the GWAS of IL18 levels in blood and IRF1, a known regulator of IL18. In addition, our findings illustrate a genetic correlation between JAK/STAT signaling, specifically through the d complex, and the manifestation of autoimmune diseases. JAK inhibitors (JAKi, which work by modulating interferon-stimulated JAK-STAT signaling[56]) have already been approved for the treatment of RA and other immune-mediated diseases[57], highlighting the power of our framework to uncover new drug targets.

In conclusion, our study showcases the effective implementation of SUM-seq in the context of a time course experiment. Importantly, our findings suggest broad applicability of this method to diverse experimental settings that demand gene regulatory analyses at single-cell resolution. Specifically, SUM-seq is a promising tool for arrayed CRISPR, drug or

perturbation screens, and large-scale atlas projects, underscoring its potential as a versatile and powerful technique across a spectrum of biological studies. To facilitate wide-spread adoption of SUM-seq we provide an extensive experimental protocol (**Supplementary Material** - available upon request/publication), as well as a scalable and reproducible Snakemake data pre-processing pipeline (https://git.embl.de/grp-zaugg/SUMseq - available upon request/publication) to make the method user-friendly.

Overall, we envision SUM-seq to provide an easy-to-adopt scalable method for projects requiring multiomic single-cell profiling up to millions of cells from hundreds of samples.

Acknowledgements

We thank the EMBL Genomic Core facility for help with sequencing, Protein Purification Core facility for Tn5 and RNasin production, Michael Snyder for providing hiPSCs, and Yakov Tsepilov and Xiangyu (Jack) Ge for providing access to the cleaned and harmonized database of GWAS summary statistics. We further thank members of the Zaugg and the Noh group for extensive discussions, Nadine Fernandez-Novel Marx for experimental support, and Josephine Brysting for providing cell lines.

A.C. was supported by an EIPOD4 Fellowship from the EC Horizon2020 MSCA (grant agreement number 847543). N.S. and J.B.Z. acknowledge funding from GSK through the EMBL-GSK collaboration framework (3000032294). N.H.S. and J.B.Z. acknowledge funding from the EMBL Infection Biology Transversal theme. K.D.P. was supported by SNSF (P2ZHP3_199669) and EMBO (ALTF538) Postdoctoral Fellowships. This work was supported by the Cariplo foundation grant, the GSK basic research fund, and the EMBL research fund (to K.M.N.). M.M. was supported by the Research council of Finland (grant number 347543), Sigrid Jusélius foundation, and Instrumentarium Science foundation. This project is co-funded by the European Union (ERC, EpiNicheAML,101044873) to J.B.Z.. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Author Information

These authors contributed equally: Sara Lobato-Moreno, Umut Yildiz, and Annique Claringbould
Correspondence to Kyung-Min Noh, Mikael Marttinen or Judith B. Zaugg

Author Contributions

M.M. conceived the project with input from U.Y., K.M.N. and J.B.Z. M.M., S.L. and U.Y. established and implemented the protocol and conducted experiments with the assistance of H.G.B. and K.D.P. V.C. assisted in experiments. M.M. and A.C. performed computational analyses. C.A. and E.P.V. constructed the pre-processing pipeline. N.H.S. assisted in data interpretation and manuscript writing. S.L., U.Y., A.C., M.M., J.B.Z. and K.M.N wrote the manuscript. M.M., J.B.Z. and K.M.N. supervised and directed the project. J.B.Z., K.M.N and M.M. secured financial support.

Competing interests

The authors declare no competing interest.

Ethics declarations

hiPSCs used for macrophage generation were derived from peripheral blood mononuclear cells with institutional review board approval (Stanford University, reference numbers 29904, 30064). The use of hiPSCs was approved by the EMBL Research Ethics Committee.

Table 1: Comparison between SUMseq and other ATAC+RNA single cell methods

Method	Microfluidics/ plate-based	Modalities	Multiplexing	Costs	No° cells	Sensitivity
SUM-seq	mf	ATAC/RNA	unlimited	$	ultra-high	high
10X multiome	mf	ATAC/RNA	no	$$$$	high	high
SNARE-seq2	mf	ATAC/RNA	no	$$	low	high
Paired-seq	mf	ATAC/RNA	no	$$	ultra-high	low
ASTAR-seq	mf	ATAC/RNA	no	$$	low	high
ISSAAC-seq	mf	ATAC/RNA	no	$$$	high	high
Smart3-ATAC	plate	RNA	yes-plate	$$	low	high
SHARE-seq	plate	ATAC/RNA	yes, but limited to first index	$$$	ultra-high	low
scCAT-seq	plate	ATAC/RNA	no	$$$	low	high

mf = microfluidics

Methods

Cell culture

Human myelogenous leukemia cells (K562) were cultured in RPMI 1640 medium (Gibco, catalog no. 11875093) supplemented with 10% FBS (Gibco, catalog no. 10270-106), 100 U/ml penicillin/streptomycin (PenStrep, Gibco, catalog no. 15140122) and 1x non-essential amino acids (Gibco, catalog no. 11140050) at 37°C with 5% CO2. Mouse fibroblast cells (NIH-3T3) were cultured in DMEM medium (high glucose, Gibco, catalog no. 11965092) supplemented with 10% FBS, 100 U/ml penicillin/streptomycin and 1x non-essential amino acids at 37 °C with 5% CO2.

Macrophage differentiation

hiPSCs were cultured in Essential 8 medium with supplement (Gibco, catalog no. A1517001) in a vitronectin (ThermoFisher, catalog no. A14700) coated dish. hiPSCs were differentiated to macrophages as previously described by van Wilgenburg et al[58]. In brief, 4 million hiPSCs were resuspended in EB medium (Essential 8 medium (Gibco, catalog no. A1517001), 50 ng/ml Recombinant Human BMP4 (Peprotech, catalog no. 120-05ET), 50 ng/ml Recombinant Human VEGF (Peprotech, catalog no. 100-20), 50 ng/ml Recombinant Human SCF (Peprotech, catalog no. 300-07)) with 10 µM Y-27632 (AbCam Biochemicals, catalog no. ab120129), seeded in 400-microwell Aggrewells (Stemcell Technologies, catalog no. 34450) and centrifuged at 100 x g for 3 min to evenly distribute the cells in microwells. 75% of the medium was replaced by fresh EB medium on the next 2 consecutive days. On the third day, EBs were transferred to a low-attachment plate (Sigma-Aldrich, catalog no. CLS3471-24EA) and cultured for two additional days with minimal disruption. On the fifth day, EBs were transferred to the final format of choice and left undisturbed for 1 week in factory media (*X-VIVO*™-15 (Lonza, catalog no. BE02-060F) 1% GlutaMax (Thermo Scientific, catalog no. 35050061), 1% PenStrep (Gibco, catalog no. 15140122), 50 µM ß-mercaptoethanol, 50 µg/ml Normocin (Invivogen, catalog no. ant-nr-05) and 100 ng/ml Recombinant Human M-CSF (Peprotech, catalog no. 300-25), 25 ng/ml Recombinant Human IL3 (Peprotech, catalog no. 200-03)) to allow them to attach. Fresh media was added weekly until macrophage precursor production started around week 4, as detected by the presence of large suspension cells with spherical morphology. From that point on, factories were maintained for a total of 8-10 weeks. Precursors were harvested weekly, taking approximately 1/4th of the total volume containing precursors and replacing it with fresh media. For the terminal differentiation of precursors to macrophages, harvested precursors were resuspended in macrophage media (*X-VIVO*™-15, 1% GlutaMax, 1% PenStrep, Recombinant Human M-CSF) and plated in the desired format at a density of 160.000 cells/cm2 for 7 days.

Polarization to M1 or M2 macrophages

At day 7, mature macrophages were polarized to M1 or M2 by adding specific polarizing molecules. For M1 polarization, macrophages were cultured in macrophage media supplemented with 20 ng/ml human recombinant IFN-γ (PeproTech, catalog no. 300-02) and 25 ng/ml LPS (Invivogen, catalog no. tlrl-3pelps). For M2 polarization macrophages were cultured in macrophage media supplemented with 20 ng/ml recombinant human IL-4 (PeproTech, catalog no. 200-04). Cells were harvested after 24, 10, 6 and 1h, including an undifferentiated control (M0) in which medium was replaced by fresh macrophage medium.

Nuclei preparation

For the species mixing experiments, mouse NIH-3T3 cells were washed once with PBS and dissociated with Accutase (Stem Cell Technologies, catalog no. 07920). Two million NIH-3T3 and K562 cells were collected and fixed in a 3% glyoxal solution (40% glyoxal (Merck, catalog no. 128465) 0.75% acetic acid; adjusted to pH 5 by addition of 1 M NaOH) for 7 min at room temperature. For cryopreservation, cells were washed after fixation with RSB-1%BSA-RI (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 1 mM DTT, 1% BSA, 20 µg/ml in-house produced RNasin (referred to RNasin hereafter)) and slowly frozen (Freezing buffer: 50 mM Tris-HCl pH 7.5, 5 mM MgAc, 0.1 mM EDTA and 25% glycerol). On the day of pre-indexing, fresh cells were processed as described above, while cryopreserved cells were thawed slowly on ice and washed with RSB-1%BSA-RI. Nuclei were extracted with 1x lysis buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM MgCl2, 0.1% Tween20, 20 µg/ml RNasin, 1 mM DTT, 1% BSA, 0.025% IGEPAL CA-630 and 0.01% Digitonin) incubating samples for 5 min on ice. Next, nuclei were washed with wash buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween20, 20 µg/ml RNasin, 1 mM DTT, 1% BSA) and filtered through a 40 µm cell strainer (Falcon, catalog no. 352340).

After the desired polarization time, macrophages were harvested by replacing the medium with ice-cold PBS-1%BSA and 2 mM EDTA and pipetting. Nuclei were extracted by resuspending the cell pellet in the 1x lysis buffer for 4 min and washed with the wash buffer. To maintain nuclei integrity, nuclei were fixed with a 3% glyoxal solution for 7 min at room temperature and washed with RSB-1%BSA-RI.

Transposome generation

Tn5 was produced in-house according to a previously described protocol with minor modifications. In short, the pETM11-Sumo3-Tn5(E54K,L372P) plasmid was transformed into *Escherichia coli* BL21(DE3) codon + RIL cells (Stratagene). Cells were grown in TB-FB media (1.2% Bacto-Tryptone, 2.4% yeast-extract, 0.4% glycerol, 17 mM KH_2PO_4, 72 mM K_2HPO_4, 1.5% Lactose, 0.05% Glucose, 2 mM $MgSO_4$) supplemented with kanamycin and chloramphenicol at 37 °C until OD600 ~0.5. The temperature was then lowered to 18 °C and cells were grown overnight, wherafter cells were harvested by centrifugation. The cell pellet was resuspended in running buffer (20 mM HEPES-NaOH pH 7.2, 800 mM NaCl, 5 mM imidazole, 1 mM EDTA, 2 mM DTT, and 10% glycerol) supplemented with cOmplete protease inhibitors (Roche) and lysed using a Microfluidizer. Polyethyleneimine (PEI) pH 7.2 was added dropwise to a final concentration of 0.5% to remove nucleic acids. The cleared lysate was loaded onto a cOmplete His-tag purification column (Roche) and the His6-Sumo3-Tn5 was eluted with a running buffer containing 300 mM imidazole. To remove the fusion tag, His6-tagged SenP2 protease was added to the elution fractions. The sample was digested overnight at 4 °C while being dialyzed back to the running buffer. The next morning, the dialyzed sample was loaded again onto a cOmplete His-tag purification column (Roche), and the untagged Tn5 was collected in the flow-through. The concentrated flow through was then loaded onto a HiLoad Superdex200 16/600 pg column (GE Healthcare) equilibrated with 50 mM Tris pH 7.5, 800 mM NaCl, 0.2 mM EDTA, 2 mM DTT, and 10% glycerol. Elution fractions corresponding to the Tn5 dimer peak were pooled, aliquoted, snap frozen in liquid nitrogen and stored at -80 °C. For transposome assembly, 100 µM of oligonucleotides with the Nextera Read1 sequence and 100 µM of oligonucleotides with a Read2-sample_index-spacer structure was annealed with 100 µM mosaic end-complement oligonucleotides with a 3' dideoxynucleotide end (ddc) at a 1:1:2 ratio by heating for 95 °C for 3 minutes followed by

cooling down to 25 °C at a ramp rate of -1 °C/min. Annealed oligos were mixed with an equal volume of 100% glycerol and stored at -20 °C until use. Finally, the annealed oligos were mixed with the in-house produced Tn5 (at 1 mg/ml in 50 mM Tris, 100 mM NaCl, 0.1 mM EDTA, 1 mM DTT, 0.1% NP-40, and 50% glycerol), at a 1:1 ratio and incubated for 30-60 min at room temperature. For the species-mixing experiment, annealed oligos were first diluted with H_2O at a 1:1 ratio before generating the transposomes. Assembled Tn5 was stored at -20 °C until use.

Data analysis procedures

ATAC sequencing data preprocessing

Base calls were converted to fastq format and demultiplexed by i7, allowing for one mismatch, using bcl-convert (v4.0.3). Demultiplexed reads were aligned to the hg38 genome and fragment file generation was performed with chromap (v0.2.3)[59]. For the species mixture experiment to evaluate collision rates, a combined hg38/mm10 reference was used and fragments aligning to either genome were counted. Using the R package ArchR (v1.02)[60], generated fragment counts for each cell were computed in 500 base genome bins. The basic preprocessing pipeline for data processing until the ArchR object generation is available at: https://git.embl.de/grp-zaugg/SUMseq.

Cell barcode filtering was performed based on number of unique fragments, TSS enrichment score, and fraction of reads in promoter regions for each sample. Peak calling was performed for each sample separately. In brief, MACS2 peak caller is used to identify peaks for each group of cells, whereafter an iterative overlap peak merging procedure is done to derive a consensus peakset between all samples. BigWig files for trace plots were generated with ArchR::plotBrowserTrack normalizing for reads in TSS.

M1 and M2 signature score generation

To determine M1 and M2 signature genes, we obtained publicly available datasets and used DESeq2[61] to identify genes that were differentially expressed in M1 and M2 macrophages compared to M0 (adjusted p-value < 0.05 and fold change > 0). The following datasets were utilized: GSE159112[62], where THP1-derived macrophages were stimulated with LPS and IFN-γ or IL-4 and IL-13 for 24 hours to obtain M1 and M2 macrophages, respectively; and GSE55536[63], involving iPSC-derived macrophages and human monocyte-derived macrophages stimulated with IFN-γ and LPS for M1 polarization, or IL-4 for M2 polarization. Genes identified as differentially expressed in both datasets were used as genesets for input to the ranking based scoring-approach AUCell[64].

Transcription factor motif accessibility

Position weight matrices (PWMs) of TF-binding motifs were obtained from HOCOMOCO v12 (*in vivo* subcollection)[65]. Motif positions in the accessible chromatin regions were determined using the function ArchR::addMotifAnnotations. Z-score of bias-corrected (GC content and mean peak accessibility) per-cell motif accessibility was calculated with the function ArchR::addDeviationsMatrix, which leverages the chromVAR framework[24].

RNA sequencing data preprocessing

Base calls were converted to fastq format using bcl2fastq (v2.20.0) or in the case of multiple pooled libraries base calls were converted to fastq format and demultiplexed by library index (i7) using bcl-convert (v4.0.3). The cell barcode (i5) was concatenated to the sample index and UMI (Read2), whereafter reads were demultiplexed by sample index using Je (v2.0.RC), allowing for two mismatches. Read alignment to Hg38 and gene expression matrix generation for each demultiplexed sample was conducted with STARsolo. For the species mixing experiment, a combined Hg38/mm10 reference was used and number of UMIs aligning to each genome was determined for each cell to estimate collision rates. For each sample, cell calling was either performed with EmptyDrops (v.1.16) or by determining inflection points on a rank vs UMI plot. Additionally, cells were filtered by mitochondrial and ribosomal read percentage. Feature by cell matrices were merged between samples, and finally features found in a minimum of 10-25 cells were retained.

Comparison to other technologies

We compare the performance of SUM-seq to dsci-ATAC-seq[8]GSM3507387), SHARE-seq[11]GSM4156590, RNA: GSM4156602 and GSM4156603), Paired-seq[12] (ATAC: GSM3737488, RNA: GSM3737489), and scifi-RNA-seq[6] (GSM5151362) using cell line data. We used 3T3 data for SUM-seq, sequenced to approximately 30.000 reads/cell for each modality. For other methods, we use authors' count matrices provided in the indicated repositories.

Dimensionality reduction and modality co-embedding

The peak matrix was normalized using TF-IDF, whereafter singular value decomposition was applied to derive a low-dimensional representation. The number of components considered was based on the proportion of variance explained (range 30-50), discarding the first component from downstream analysis due to high correlation with number of fragments per cell.

The gene expression matrix was normalized by proportional fitting (cell depth normalization to the mean cell depth), followed by logarithmic transformation (log(x + 0.5)), and another round of proportional fitting[66]. Next, a low dimensional representation of the normalized gene expression matrix was determined with PCA. The number of components considered was based on the proportion of variance explained (range 30-50).

Derived low-dimensional representations for each modality was used as input to learn cell modality weights, from which a WNN graph was constructed [67]. The WNN graph was further used to define a common UMAP visualization of the data modalities.

Multiomic factor analysis

To generate a common low-dimensional latent space for ATAC and RNA in the M1 and M2 polarizations, we utilized the multiomics integration platform MOFA[21]. First, ATAC data was collapsed from peaks to cis-regulatory topics with cisTopic (v0.3)[68], determining a suitable number of topics based on the maximum on the second-derivative of likelihood curve and minimum on the model perplexity curve. All topics were used as input for MOFA. For RNA, top 4000 (default) most variable genes were used as input.

To determine polarization-associated latent factors, latent factors were correlated with biological metadata (state (M0/M1/M2), timepoint) as well as technical metadata (UMI/cell,

fragments/cell). Factors with a high correlation against biological metadata, but not technical metadata were chosen as factors for downstream analysis.

Gene set enrichment analysis

Gene set enrichment analysis was performed for Reactome database genesets (v59). For every gene set G, significance is evaluated via a parametric *t*-test, where weights of the foreground set (features in set G) are contrasted against a background set (weights of features not in set G). *P* values were adjusted for multiple testing using the Benjamini–Hochberg procedure. Enrichments with false discovery rate 5% were considered significant.

Motif enrichment analysis

Motif enrichment analysis was performed with R package monaLisa (v1.8) against HOCOMOCO v12 (*in vivo* subcollection) PWMs. Peaks within topics in the top and bottom 3% quantile of feature weights for MOFA factors of interest were considered for analysis. TFs with FDR 0.1% and log2 enrichment in the top or bottom 10% quantiles were considered significant.

Gene regulatory network inference and transcription factor prioritization

We inferred an enhancer-based gene regulatory network (eGRN) using GRaNIE[31] v1.5.3 with TFBS predictions based on HOCOMOCO v12[65] database (*in vivo* subset) that were generated using PWMScan (see[69] for methodological details how these were produced). The single-cell data from all time points and stimulations was first clustered using the smart local moving (SLM) algorithm with a resolution of 1, giving rise to 31 clusters that have at least 25 cells. These clusters largely align with the stimulations and time points (**Extended Data Figure 5a**). We then calculated the mean RNA and ATAC values for each of these clusters as a pseudobulk, and input these clusters as 'samples' to create a eGRN following the GRaNIE single cell vignette (https://grp-zaugg.embl-community.io/GRaNIE/articles/GRaNIE_singleCell_eGRNs.html). We kept links if the TF-peak connection had an FDR < 0.2 and the peak-gene connection had an FDR < 0.1, the defaults in the GRaNIE package. We collapsed TF motif connections to the level of TFs to construct TF regulons.

Residual analysis

To identify discrepancy and concordance between TF motif accessibility and regulon activity scores, values were rank min-max normalized and the difference between the two was computed, and is referred to as the residual value. AUCell was used to infer single-cell activity scores for regulon geneset

GWAS integration

LDSC

To integrate GWAS with our eGRN, we performed stratified linkage disequilibrium score regression (S-LDSC; [41]). We added all open regions in all macrophage cells as background, on top of the default baseline model that includes genic regions, enhancer regions, and conserved regions. We then tested enriched heritability for all peaks that were part of the STAT1-STAT2-IRF9 eGRN, the IRF1-IRF3-IRF8 eGRN, all peaks in the eGRN, and all peaks

in M1 and M2 cells only. We included 8890 traits from UK biobank, FinnGen and several GWAS repositories, and tested S-LDSC only if the overall SNP heritability (h2) was > 0.05, as calculated by LDSC, resulting in a list of 1230 traits. We corrected the enrichment p-values for the number of peaksets tested in each trait.

Gene modules and SNP overlap

For those traits that were enriched for heritability (adjusted p < 0.05), the SNPs associated to the particular trait (suggestive GWAS p-value < 5*10^-6) were intersected with the peaks in the STAT1-STAT2-IRF9, IRF1-IRF3-IRF8, and full eGRNs (Supplementary Tables 7, 9, 10). We selected the genes connected (in the STAT1-STAT2-IRF9 eGRN) to the peaks that overlapped a SNP to obtain disease gene modules for CD, IBD and monocyte counts.

Code availability

A Snakemake preprocessing pipeline for SUM-seq data is available in a public Git repository (https://git.embl.de/grp-zaugg/SUMseq, accessible upon request/publication). All analysis source code will be available on a separate Git repository.

Data availability

All data will be deposited to the NCBI GEO or EGA databases.

References

1. Claringbould, A. & Zaugg, J. B. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol. Med.* **27**, 1060–1073 (2021).
2. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.* **24**, 494–515 (2023).
3. Sun, Z. *et al.* Joint single-cell multiomic analysis in Wnt3a induced asymmetric stem cell division. *Nat. Commun.* **12**, 5941 (2021).
4. Wang, Y. *et al.* Single-cell multiomics sequencing reveals the functional regulatory landscape of early embryos. *Nat. Commun.* **12**, 1247 (2021).
5. Bian, S. *et al.* Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science* **362**, 1060–1063 (2018).
6. Datlinger, P. *et al.* Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat. Methods* **18**, 635–642 (2021).
7. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
8. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).
9. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).
10. Xu, W. *et al.* ISSAAC-seq enables sensitive and flexible multimodal profiling of chromatin accessibility and gene expression in single cells. *Nat. Methods* **19**, 1243–1249 (2022).
11. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).
12. Zhu, C. *et al.* An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019).
13. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
14. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 (2020).
15. Bagnoli, J. W. *et al.* Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 2937 (2018).
16. Zhang, X., Marand, A. P., Yan, H. & Schmitz, R. J. Massive-scale single-cell chromatin accessibility sequencing using combinatorial fluidic indexing. *BioRxiv* (2023) doi:10.1101/2023.09.17.558155.
17. Zhang, H. *et al.* txci-ATAC-seq, a massive-scale single-cell technique to profile chromatin accessibility. *BioRxiv* (2023) doi:10.1101/2023.05.11.540245.
18. Satoh, T. *et al.* The Jmjd3-Irf4 axis regulates M2 macrophage polarization and host responses against helminth infection. *Nat. Immunol.* **11**, 936–944 (2010).
19. Krausgruber, T. *et al.* IRF5 promotes inflammatory macrophage polarization and TH1-TH17 responses. *Nat. Immunol.* **12**, 231–238 (2011).
20. Kawashima, T. *et al.* STAT5 induces macrophage differentiation of M1 leukemia cells through activation of IL-6 production mediated by NF-kappaB p65. *J. Immunol.* **167**, 3652–3660 (2001).
21. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
22. Kelly, B. & O'Neill, L. A. J. Metabolic reprogramming in macrophages and dendritic cells in innate immunity. *Cell Res.* **25**, 771–784 (2015).

23. Huang, S. C.-C. *et al.* Metabolic Reprogramming Mediated by the mTORC2-IRF4 Signaling Axis Is Essential for Macrophage Alternative Activation. *Immunity* **45**, 817–830 (2016).
24. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
25. Karpurapu, M. *et al.* Functional PU.1 in macrophages has a pivotal role in NF-κB activation and neutrophilic lung inflammation during endotoxemia. *Blood* **118**, 5255–5266 (2011).
26. Au-Yeung, N., Mandhana, R. & Horvath, C. M. Transcriptional regulation by STAT1 and STAT2 in the interferon JAK-STAT pathway. *JAKSTAT* **2**, e23931 (2013).
27. Czimmerer, Z. *et al.* The transcription factor STAT6 mediates direct repression of inflammatory enhancers and limits activation of alternatively polarized macrophages. *Immunity* **48**, 75-90.e6 (2018).
28. Huang, Q. *et al.* Spi-B Promotes the Recruitment of Tumor-Associated Macrophages via Enhancing CCL4 Expression in Lung Cancer. *Front. Oncol.* **11**, 659131 (2021).
29. Villar, J. *et al.* ETV3 and ETV6 enable monocyte differentiation into dendritic cells by repressing macrophage fate commitment. *Nat. Immunol.* **24**, 84–95 (2023).
30. Ong, C.-T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* **15**, 234–246 (2014).
31. Kamal, A. *et al.* GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks. *Mol. Syst. Biol.* **19**, e11627 (2023).
32. Rauluseviciute, I. *et al.* JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* (2023) doi:10.1093/nar/gkad1059.
33. Wertz, I. E. *et al.* De-ubiquitination and ubiquitin ligase domains of A20 downregulate NF-kappaB signalling. *Nature* **430**, 694–699 (2004).
34. Wang, G., Kouwaki, T., Okamoto, M. & Oshiumi, H. Attenuation of the Innate Immune Response against Viral Infection Due to ZNF598-Promoted Binding of FAT10 to RIG-I. *Cell Rep.* **28**, 1961-1970.e4 (2019).
35. He, L. *et al.* Global characterization of macrophage polarization mechanisms and identification of M2-type polarization inhibitors. *Cell Rep.* **37**, 109955 (2021).
36. Viengkhou, B., White, M. Y., Cordwell, S. J., Campbell, I. L. & Hofer, M. J. A novel phosphoproteomic landscape evoked in response to type I interferon in the brain and in glial cells. *J. Neuroinflammation* **18**, 237 (2021).
37. Sadzak, I. *et al.* Recruitment of Stat1 to chromatin is required for interferon-induced serine phosphorylation of Stat1 transactivation domain. *Proc Natl Acad Sci USA* **105**, 8944–8949 (2008).
38. Paul, A., Ismail, M. N., Tang, T. H. & Ng, S. K. Phosphorylation of interferon regulatory factor 9 (IRF9). *Mol. Biol. Rep.* **50**, 3909–3917 (2023).
39. Platanias, L. C. Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nat. Rev. Immunol.* **5**, 375–386 (2005).
40. Sekrecka, A. *et al.* Time-dependent recruitment of GAF, ISGF3 and IRF1 complexes shapes IFNα and IFNγ-activated transcriptional responses and explains mechanistic and functional overlap. *Cell. Mol. Life Sci.* **80**, 187 (2023).
41. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
42. Hurgin, V., Novick, D. & Rubinstein, M. The promoter of IL-18 binding protein: activation

by an IFN-gamma -induced complex of IFN regulatory factor 1 and CCAAT/enhancer binding protein beta. *Proc Natl Acad Sci USA* **99**, 16957–16962 (2002).
43. Baggio, C. *et al.* IL-18 in Autoinflammatory Diseases: Focus on Adult Onset Still Disease and Macrophages Activation Syndrome. *Int. J. Mol. Sci.* **24**, (2023).
44. Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, D1311–D1320 (2021).
45. Vuletic, S., Dong, W., Wolfbauer, G., Tang, C. & Albers, J. J. PLTP regulates STAT3 and NFκB in differentiated THP1 cells and human monocyte-derived macrophages. *Biochim. Biophys. Acta* **1813**, 1917–1924 (2011).
46. Peters, A. L., Stunz, L. L. & Bishop, G. A. CD40 and autoimmunity: the dark side of a great activator. *Semin. Immunol.* **21**, 293–300 (2009).
47. Nguyen, V. T. & Benveniste, E. N. Involvement of STAT-1 and ets family members in interferon-gamma induction of CD40 transcription in microglia/macrophages. *J. Biol. Chem.* **275**, 23674–23684 (2000).
48. Sekine, C., Yagita, H., Miyasaka, N. & Okumura, K. Expression and function of CD40 in rheumatoid arthritis synovium. *J. Rheumatol.* **25**, 1048–1053 (1998).
49. Swanson, E. *et al.* Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* **10**, (2021).
50. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.* **39**, 825–835 (2021).
51. Nichols, R. V. *et al.* High-throughput robust single-cell DNA methylation profiling with sciMETv2. *Nat. Commun.* **13**, 7627 (2022).
52. Mulqueen, R. M. *et al.* High-content single-cell combinatorial indexing. *Nat. Biotechnol.* **39**, 1574–1580 (2021).
53. Wienerroither, S. *et al.* Cooperative Transcriptional Activation of Antimicrobial Genes by STAT and NF-κB Pathways by Concerted Recruitment of the Mediator Complex. *Cell Rep.* **12**, 300–312 (2015).
54. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
55. Freimer, J. W. *et al.* Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of immune networks. *Nat. Genet.* **54**, 1133–1144 (2022).
56. Tanaka, Y., Luo, Y., O'Shea, J. J. & Nakayamada, S. Janus kinase-targeting therapies in rheumatology: a mechanisms-based approach. *Nat. Rev. Rheumatol.* **18**, 133–145 (2022).
57. Serra López-Matencio, J. M., Morell Baladrón, A. & Castañeda, S. JAK-STAT inhibitors for the treatment of immunomediated diseases. *Medicina Clínica (English Edition)* **152**, 353–360 (2019).
58. van Wilgenburg, B., Browne, C., Vowles, J. & Cowley, S. A. Efficient, long term production of monocyte-derived macrophages from human pluripotent stem cells under partly-defined and fully-defined conditions. *PLoS ONE* **8**, e71098 (2013).
59. Zhang, H. *et al.* Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat. Commun.* **12**, 6566 (2021).
60. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
61. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion

for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

62. Xu, M. *et al.* Arachidonic acid metabolism controls macrophage alternative activation through regulating oxidative phosphorylation in pparγ dependent manner. *Front. Immunol.* **12**, 618501 (2021).
63. Zhang, H. *et al.* Functional analysis and transcriptomic profiling of iPSC-derived macrophages and their application in modeling Mendelian disease. *Circ. Res.* **117**, 17–28 (2015).
64. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
65. Vorontsov, I. E. *et al.* HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res.* (2023) doi:10.1093/nar/gkad1077.
66. Booeshaghi, A. S., Hallgrímsdóttir, I. B., Gálvez-Merchán, Á. & Pachter, L. Depth normalization for single-cell genomics count data. *BioRxiv* (2022) doi:10.1101/2022.05.06.490859.
67. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587. (2021).
68. Bravo González-Blas, C. *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
69. Berest, I. *et al.* Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF. *Cell Rep.* **29**, 3147-3159.e12 (2019).