Title: Sequence-Dependent DNA Shape Features Regulate IL-4 Induced Chromatin Accessibility in Alternatively Activated Macrophages

Authors: Mei San Tang[1§], Emily R. Miraldi[2,3], Natasha M. Girgis[1], Richard A. Bonneau[4,5], P'ng Loke[1§]

Affiliations:

[1]Department of Microbiology, New York University School of Medicine, New York, NY 10016, USA.

[2] Divisions of Immunobiology and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA

[3] Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, 45267, USA

[4] Department of Biology, Center for Genomics and Systems Biology, New York University, New York, NY 10003, USA.

[5] Simons Center for Data Analysis, Simons Foundation, New York, NY 10011, USA.

[§]Correspondence to:
MeiSan.Tang@nyumc.org
Png.Loke@nyumc.org

Abstract

Interleukin-4 (IL-4) activates macrophages to adopt a distinct phenotype important for tissue repair and helminth infection, but the molecular basis of chromatin remodeling in response to IL-4 stimulation is not understood. We find that IL-4 activation of terminally differentiated macrophages in mice is accompanied by cell-type-specific chromatin remodeling in regions enriched with binding motifs of the pioneer transcription factor PU.1. Mutation studies based on natural genetic variation between C57BL/6 and BALB/c mouse strains demonstrate that accessibility of these IL-4 induced regions can be regulated through differences in DNA shape, without disruption of pioneer factor motifs. We propose a model whereby DNA shape features of stimulation-dependent genomic elements contribute to differences in the accessible chromatin landscape of alternatively activated macrophages on different genetic backgrounds.

Main Text

Macrophage activation is a process by which macrophages transition from a resting state to adopt different phenotypes, in response to specific external stimuli that can either be danger signals or homeostatic and metabolic signals (*1*). Macrophages alternatively activated (AAM) by the type 2 cytokines interleukin-4 (IL-4) and IL-13 adopt a distinct phenotype that can promote resolution of helminth infections, as well as tissue repair during chronic inflammation (*2*). Notably, different lineages and phenotypes of tissue macrophages will respond differently to IL-4 stimulation *in vivo* (*3, 4*). The accessible chromatin landscape of terminally differentiated cells, shaped during differentiation, may determine cellular functions and phenotype (*5, 6*). It has been shown that different types of tissue macrophages with distinct functions have distinct accessible chromatin profiles (*7*). The macrophage activation process is accompanied by changes in transcriptional activities and histone modifications genome-wide, orchestrated by combinatorial actions of different transcription factors (TFs) and have been best demonstrated in the context of macrophage response to toll-like receptor (TLR) signaling (*1, 8-11*). However, the molecular basis of stimulation-dependent chromatin remodeling is incompletely understood, especially in the context of alternative activation in response to type 2 cytokines.

To examine chromatin remodeling on different types of tissue macrophages, we injected recombinant IL-4–antibody complex (IL-4c) into the peritoneal cavity of C57BL/6 mice to induce accumulation of alternatively activated F4/80^{hi}CD206^{-} macrophages of embryonic origin (AAM^{res}) and compared these with F4/80^{int}CD206^{+} macrophages derived from Ly6C^{hi} inflammatory blood monocytes (AAM^{mono}) from mice injected with IL-4c and thioglycollate (*4*). We then used ATAC-seq (*12*) to profile the open chromatin landscape of these macrophages, in comparison to unstimulated F4/80^{hi}CD206^{-} macrophages of naïve mice and F4/80^{int}CD206^{+} macrophages from thioglycollate-treated mice (*4*). The overall differences in accessible chromatin landscape (a total of 61,713 open chromatin regions) could be attributed mainly to the type of macrophage (27% of total variance), but alternative activation by IL-4 also altered the accessible chromatin profiles (Figures 1A, B). *Arg1* and

Ucp1, which are known to be IL-4 inducible (*4*), had constitutively accessible chromatin regions, whereas *Retnla*, another IL-4 inducible gene, had chromatin regions that gained accessibility in response to IL-4 (Figure 1B). This IL-4 induced chromatin remodeling process can be cell-type-specific at certain regions (e.g. *Tgfb2*, *Ccl2*) (Figure 1B). Of the 61,713 total accessible regions, we identified 1572 regions induced by IL-4 for AAM^{res} and 1462 regions for AAM^{mono} (Figure S1A). IL-4-dependent regions also had the largest contribution to the differences in open chromatin profiles between unstimulated and IL-4 stimulated macrophages (Figure 1C). The IL-4 induced regions almost all (99% in AAM^{res} and 97% in AAM^{mono}) gained accessibility from undetectable levels at baseline (Figure 1D). When compared to constitutively accessible regions (Figure S1B), IL-4 induced regions have distinct sequence properties – they were significantly over-represented with non-coding intronic elements (Figure 1E), were closer to IL-4 induced genes (Figure 1F), were lower in GC content (Figure 1G) and less likely to overlap with a CpG island (Figure 1G). These differences between constitutively accessible and IL-4 induced regions were observed in both monocyte-derived and tissue resident macrophages.

Even though both AAM^{mono} and AAM^{res} received the same stimulation, the regions that were remodeled by IL-4 were largely cell-type-specific (Figure 2A). Of all the 2855 IL-4 induced regions, only 179 regions (6% of total IL-4 induced regions) were shared between both AAM^{mono} and AAM^{res} (Figure S1C). While the IL-4 induced regions from AAM^{mono} and AAM^{res} were largely distinct, the TF motifs discovered in these regions were largely similar, and these included PU.1, KLF and the AP-1 family of motifs (Figure 2B). To quantify the accuracy of motif prediction with actual TF binding, we compared our predicted PU.1 motifs with ChIP-seq-defined PU.1 binding sites of thioglycollate-induced macrophages (*11, 13*) and found that 78% of the PU.1 motif sites predicted from thioglycollate-induced macrophages in our study (4,282 of total 5,492 predicted PU.1 motif sites) overlapped with a PU.1 binding site defined by ChIP-seq (Figure S2A). While PU.1 is better characterized as a pioneer factor that shapes the chromatin landscape during macrophage development (*14, 15*), it can also unveil latent enhancers during macrophage activation, through cooperative binding with

stimulus-dependent TFs (such as STAT1, STAT6) (*10*). To determine if similar molecular events could be associated with IL-4 induced chromatin regions, we examined regions of the IL-4-induced PU.1 motif along with ±25bp flanking sequences (*16*) from both AAM^{mono} and AAM^{res} (Figure S2B). STAT TF family motifs were highly enriched for both types of macrophages (Figure 2C). Motifs from other stimulus-dependent TF families (*17*) important in macrophage activation (such as IRF, RXR) were also highly enriched (Figure 2C) (*18-20*). Notably, there was enrichment of a motif from a TF family of "More than 3 adjacent zinc finger factors" that was AAM^{res}-specific and within this TF family, MZF1, which is a TF with growth suppressing function on hematopoietic cells (*21*), had the highest enrichment level. However, the majority of TF families were enriched at similar levels in both AAM^{res} and AAM^{mono}, and thus differential accessibility of putative TF binding sites alone could not explain the macrophage-specific chromatin remodeling patterns. We hypothesized that other sequence features around PU.1 motifs could be important.

Since functional PU.1 binding can be most accurately determined by TF motifs neighboring the PU.1 motif and by the local three-dimensional shape of the DNA at PU.1 binding sites (*22*), we next determined if such sequence properties could be associated with differences in IL-4 induced PU.1 regions between AAM^{mono} and AAM^{res}. We computationally predicted four DNA shape configurations (minor groove width - MGW, propeller twist – ProT, helical twist – HelT and roll) at the IL-4 induced PU.1 regions (*23*) and found that IL-4 induced PU.1 regions of AAM^{mono} and AAM^{res} differed significantly in shape profiles (Figure 2D). These macrophage-specific DNA shape profiles were specific to the PU.1 motif regions, as IL-4 induced KLF regions did not display AAM^{mono} and AAM^{res} specific DNA shape profiles (Figure S2C). To determine if cell-type-specific local DNA shape of PU.1 binding sites can be observed in an independent setting, we predicted DNA shape profiles of PU.1 binding sites defined by ChIP-seq from different cell types (splenic B cells, thioglycollate-elicited-peritoneal macrophages and a microglial cell line) (*24, 25*) (Figure S3A). The shape of PU.1 binding sites specific to each of these cell types were significantly distinct (Figure S3B). Together, these results point towards the contribution of local DNA shape on the cell-

type-specific pattern of PU.1 motif accessibility and binding. Thus, local DNA shape features of PU.1 binding sites might contribute towards cell-type-specific chromatin remodeling, not only for AAMs in response to IL-4 stimulation, but also for other PU.1-expressing cell types.

To determine whether altering the local DNA shape would change accessibility of the PU.1 motifs, we designed a mutation study using natural variants between the C57BL/6 and BALB/c mouse strains, by performing ATAC-seq on AAM^{mono} and AAM^{res} generated from mice of these two different genetics backgrounds. We first defined "strain-specific" (significant differences in accessibility counts between C57BL/6 and BALB/c at FDR 10%) or "strain-common" (no significant differences in accessibility count between C57BL/6 and BALB/c) IL-4 induced PU.1 regions from both AAM^{mono} and AAM^{res}. We next focused only on strain-specific IL-4 induced PU.1 regions with SNP(s) flanking the PU.1 motif (Figure S4A). This allowed us to ask if the differences in chromatin accessibility between strains were only due to the flanking sequence mutation(s) and not confounded by disruption of a potential PU.1 binding site. We then correlated these strain-specific differences in accessibility with phenotypic differences in gene expression. We highlight this DNA shape dependent regulatory mechanism at two independent genomic regions that we infer to be regulating biologically relevant functions. Our first example is an intergenic accessible region in AAM^{res}, where two different SNPs occurring at the 19th and 24th base pairs upstream of a PU.1 motif led to changes in DNA shape, most drastically with the ProT and HelT configurations (Figure 3A, Table 1). This was accompanied by decrease in chromatin accessibility counts ($\log_2$ fold change BALB/c vs. C57BL/6 = -2.57, $p = 0.01$) and reduced expression of the nearby *Slc30a4* gene in the BALB/c mice ($\log_2$ fold change BALB/c vs. C57BL/6 = -3.43, $p = 0.006$). *Slc30a4* is a zinc exporter that affects the survival of *Histoplasma capsulatum* in IL-4 stimulated macrophages (*26*). Our second example is an intronic region on the *P4hb* gene in AAM^{mono}, where SNPs occurring at the 8th and 24th base pairs upstream of a PU.1 motif led to changes in DNA shape, with the former causing larger shifts in the MGW and roll configurations (Figure 3B, Table 1). This was, however, accompanied by increased accessibility instead ($\log_2$ fold change BALB/c vs. C57B6L/6 =

1.87, p = 0.0085) and increased expression of the *P4hb* gene ($\log_2$ fold change BALB/c vs. C57BL/6 = 0.34, p = 0.048). Overall, PU.1 motif-flanking variants occurred at higher frequency then variants located directly within the motif (24.3% vs. 5.4% in AAM^{res} and 20.7% vs. 6.9% in AAM^{mono}) (Figure 3C, top panel). This suggests that DNA shape readout (a proxy for local environment), instead of direct base readout during PU.1 binding, is likely to be a more important sequence property that could influence PU.1 binding. The frequency of local variants was also greater in strain-specific genomic elements when compared to strain-common regions (Fisher's exact test, $p = 4.59\times10^{-4}$ in AAM^{res}, $p = 8.12\times10^{-6}$ in AAM^{mono}), even though strain-common PU.1 motifs could also have sequence variants present within ±25bp of the PU.1 motif (Figures 3C, bottom panel; Table 2). When we extended this analysis to all accessible regions in IL-4 stimulated macrophages, regardless of the presence of PU.1 motif or IL-4 inducibility, we see that 17.5% of all strain-common regions carry local variants. This suggests that the presence of sequence variants do not always have an impact on chromatin accessibility and additional factors must influence whether accessibility at specific chromatin regions are more likely to be altered by sequence variants.

Given the potential role of DNA shape in regulating chromatin accessibility, we first asked if the amount of DNA shape change caused by a sequence variant was associated with its impact on chromatin accessibility. We identified all accessible regions carrying SNP(s), regardless of the presence of a PU.1 motif or IL-4 inducibility, and categorized them into either "strain-specific" or "strain-common", as defined above (Figure S4B). The strain-specific regions had a significantly greater change in DNA shape secondary to sequence variants and this was most significant in the MGW, ProT and roll DNA shape configurations (Figure 3D). We next used a logistic regression model to systematically examine the contribution of various DNA sequence features in altering chromatin accessibility (detailed in **Methods**). We found IL-4 inducibility, SNP frequency, as well as presence of a nearby CpG island to be significant predictors of strain-specific accessible regions (Figure 3E). Accessible regions that were IL-4-inducible and regions that contained higher frequency of SNPs were more likely to be strain-specific, while accessible regions overlapping a CpG island were

more likely to be strain-common. Consistent with this finding, strain-specific regions (regardless of the presence of variants) were significantly more enriched with IL-4-induced peaks overall, and this was observed in both AAM^{mono} and AAM^{res} (Figure 3F). This suggests that differences in the accessible chromatin landscape of activated macrophages on different genetics background were more likely to be driven by stimulation-dependent genomic elements, further supporting the functionality of IL-4-induced genomic elements during alternative activation. We also asked if the higher frequency of SNPs in a local chromatin region could have contributed to alteration of chromatin accessibility by inducing a greater change in DNA shape. Indeed, SNP frequency positively correlated with the amount of SNP-induced DNA shape change, especially for the shape features MGW and ProT (Figure 3G).

We next compared the global transcriptional profiles of AAM^{res} and AAM^{mono} from C57BL/6 and BALB/c mice to identify functional differences in these AAMs of different genetic backgrounds. Even though most of the differences in transcriptional profiles were predominantly driven by the types of macrophages (45% of total variance), strain differences also contributed to the variation in transcriptional profiles (18% of total variance) (Figure 4A). Consistent with this finding, most of the macrophage-specific functions were conserved across mouse strains and not affected by differences in genetics backgrounds (Figure 4B, top panel). In contrast, functional differences secondary to genetics were largely specific to the different types of macrophages (Figure 4B, bottom panel). We next examined the strain-specific functional differences in AAM^{res} and AAM^{mono}, respectively (Figure 4C). Notably, BALB/c AAM^{res} expressed lower levels of cell-cycle-related genes, in line with the previously reported observation that peritoneal AAM^{res} have lower proliferation capacity during *Litomosoides sigmodontis* infection (*27*). In addition, our paired ATAC-seq and RNA-seq data from AAMs of different genetic backgrounds allowed us to examine the functionality of specific accessible regulatory elements, as the BALB/c sequence variants could act as perturbations to induce changes in specific accessible regions (analogous to a "knock-out" or "knock-in" experiment of specific regulatory elements). The impact of deleting or inducing a specific regulatory element can then be examined by changes in

proximal gene expression levels. We therefore endeavored to elucidate the functions of specific IL-4 induced regions. Specifically, we identified regions with strain-specific accessibility pattern, located on the gene bodies of strain-specific genes. We found 8 and 20 of such region-gene pairs in AAM^{res} and AAM^{mono}, respectively (Figure 4D). Notably, the regions associated with *Rab31*, *P4hb* and *Arhgef3* were also IL-4 induced regions where DNA shape changes flanking a PU.1 motif led to alteration in chromatin accessibility. Nonetheless, the directions of changes in the expression levels of these genes were not always in the same direction as the changes in the chromatin accessibility of the matching regions, as a reduction in chromatin accessibility sometimes correlated with increased gene expression (for example, *Arhgef3* in AAM^{res} and *Sptbn1* in AAM^{mono}). Many of these genes directly associated with IL-4 induced regions were cytoskeleton genes (*Sfi1*, *Ank2*, *Plekhg6*, *Arhgap39*, *Sptbn1*, *Mapt*) and components of the Rho GTPase signal transduction pathway (*Plekhg6*, *Arhgef3*, *Arhgef10l*, *Vav3*), which are cellular functions important to specific macrophage functions such as phagocytosis and efferocytosis (*28*), suggesting the possibility that AAMs from C57BL/6 and BALB/c could have differences in their phagocytic capability (*29, 30*).

In this study, we define and characterize IL-4 induced chromatin accessibility in both tissue-resident and monocyte-derived macrophages *in vivo*. Although the pioneer factor PU.1 motif is similarly associated with IL-4 induced regions in both macrophages lineages, the local DNA shape features at these PU.1 motifs are macrophage-specific. By leveraging natural genetic variation between mouse strains, we demonstrate that the accessibility of putative PU.1 binding sites correlates with the presence of SNPs flanking the PU.1 binding site, local DNA shape predictions, and altered expression of proximal genes. However, many critical questions remain unanswered. Do these *de novo* accessible regions persist after resolution of the external stimulation to cause differences in activation kinetics with a repeated stimulation (*10, 31*), or would they lose accessibility to allow for plasticity in macrophage activation? In addition, what are the intermediate molecular events that could explain cell-type-specific local DNA shape at PU.1 regions? Possible mechanisms include

differences in specific TFs cooperatively binding with PU.1 (*25*), different protein-protein interactions occurring at different functional domains of the PU.1 protein in a macrophage-specific manner (*32*), or variants of the PU.1 protein secondary to different post-translational modifications of PU.1 (*33*). Finally, since traits-associated variants in the human population are concentrated in non-coding regulatory DNA and disruption of these regulatory elements could be the underlying mechanism of how variants lead to specific phenotypes (*34*), we also propose that DNA shape change could be a generalizable means to predict which regulatory element are more likely to be disrupted by sequence variants, leading to phenotypic predictions in human health and disease (*35*).

Acknowledgements

We thank the NYU School of Medicine Genome Technology Center and Cytometry and Cell Sorting Laboratory core facilities. These shared resources are partially supported by the Laura and Isaac Perlmutter Cancer Center support grant P30CA016087. We also thank Shenglong Wang and other staff expertise from the NYU IT High Performance Computing for their excellent technical support with software and other computing resources. This work was supported through the NIH, NIAID grants AI093811 and AI094166 (P.L.) and NIDDK grant DK103788 (P.L.), American Association of Immunologists (M.S.T.), Vilcek Foundation (M.S.T.).

Author contributions

Conceptualization, M.S.T. and P.L.; Methodology, M.S.T, E.R.M., N.M.G., R.A.B. and P.L.; Formal Analysis, M.S.T.; Investigation, M.S.T. and N.M.G.; Writing – Original Draft, M.S.T. and P.L.; Writing – Review and Editing, E.R.M. and P.L.; Visualization, M.S.T.; Data Curation, M.S.T.; Supervision, R.A.B. and P.L.; Funding acquisition, P.L. All authors read and approved of the final draft.

Data availability

ATAC-Seq and RNA-Seq data are being prepared for deposition on the NCBI Gene Expression Omnibus (GEO) database. All analysis codes will be made publicly available for reproducibility purpose.

Materials and Methods

Experimental methods

Mice Wild type (WT) C57BL/6 mice were purchased from Jackson laboratory and bred onsite for the first set of experiments that compared the effect of IL-4 stimulation on the accessible chromatin profiles from M^{res}, AAM^{res}, M^{mono} and AAM^{mono}. For experiments directly comparing AAMs of C57BL/6 and BALB/c backgrounds, mice of both strains were purchased from Jackson laboratory and immediately used for experiments. Mice were age (7-8 weeks of age) and gender matched. IL-4/anti-IL-4 monoclonal antibody (mAb) complexes (IL-4c) were prepared as described previously (*36*). To induce AAM^{res}, mice were injected intraperitoneally (i.p.) with IL-4c on days 0 and 2. Mice were also treated with 4% thioglycollate alone (to induce M^{mono}) or in combination with IL-4c to induce AAM^{mono} (*4*). All animal procedures were approved by the New York University Institutional Animal Care and Use Committee (IACUC) under protocol numbers 131004 and 130504.

Peritoneal cell isolation and cell sorting Peritoneal cells were isolated by washing the peritoneal cavity twice with cold PBS 1x. Peritoneal exudate were then treated with ACK Lysis buffer to lyse red blood cells and washed once with PBS. Cells were then re-suspended to single-cell suspensions for staining with fluorescently conjugated antibodies at 1:100 dilutions, unless otherwise noted. Antibodies were diluted using 2% fetal bovine serum (FBS). Cells were stained with LIVE/DEAD™ Blue (Invitrogen), blocked with 4µg/ml anti-CD16/32 (2.4G2; Bioxcell) and stained with anti-CD11b Pacific Blue (M1/70; Biolegend),

F4/80 PECy7 (BM8; Biolegend), CD206 APC (C068C2; Biolegend), Siglec-F PE (E50-2440; BD Biosciences), CD3 PE (145-2C11; Biolegend), CD19 PE (6D5; Biolegend), CD49b PE (DX5; Biolegend), Ly6G (1A8; Biolegend). Cells were gated on singlet, live, Dump-negative (CD3−, CD19−, DX5−, Siglec-F−, Ly6G−), CD11b+, then subsequently gated on their M^{res} and AAM^{res} (F4/80hi, CD206−) or M^{mono} and AAM^{mono} (F4/80int, CD206+) phenotype. Cells were sorted using 100µm nozzle into FBS, on either BD FACSAriaII or SONY HAPS1, depending on instrument availability.

Assay for Transposase-Accessible Chromatin with Sequencing (ATAC-seq) ATAC-seq was performed as described by Buenrostro et al (*12*). 50,000 FACS-purified cells per sample were spun down at 400g for 5 min at 4°C and washed once with 50µl cold PBS. Cells were lysed with 50µl lysis buffer (10 mM Tris-HCl, pH7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630) and immediately spun down at 1500rpm for 10 min at 4°C. The isolated cell nuclei were then incubated for 30 min at 37°C with 50µl of transposase reaction, which contained 25µl Tagment DNA buffer (Illumina), 2.5µl Tagment DNA enzyme (Illumina) and 22.5µl nuclease-free water. The transposed DNA was immediately purified using the Qiagen MinElute PCR Purification Kit (Qiagen) following manufacturer's guide and eluted at 10ul volume. PCR amplification of the transposed DNA was done using a low-cycle number protocol and with primers published by Buenrostro et al (*12*). Each PCR mix contained of 25µl of NEB 2x PCR Mix (New England Biolabs), 2.5µl of 25µM forward primer (Primer Ad1_noMX), 2.5µl of 25µM reverse barcoded primer, 0.3µl of 100x SYBR Green (Invitrogen) and 10µl of transposed DNA. PCR was carried out with the cycling protocol: 72°C for 5 min, 98°C for 30s, followed by 5 cycles of 98°C for 10s, 63°C for 30s, 72°C for 1 min. The reaction was held at 4°C after the 5th cycle. A side qPCR was set up using the PCR product from these 5 cycles of amplification. Each qPCR mix contained 5µl NEB 2x PCR Mix, 0.25µl 25µM forward primer, 0.25µl 25µM reverse barcoded primer, 0.06µl 100x SYBR Green, 4.44µl nuclease-free water and 5µl of the PCR-amplified product. qPCR was carried out using the cycling protocol: 98°C for 30s, followed by 25 cycles of 98°C for 10s, 63°C for

30s, 72°C for 1 min and plate read. The qPCR amplification plot was then used to calculate the additional number of cycles needed for the PCR to achieve maximum amount of product without going into saturation. Each sample was amplified for a total of 14-16 cycles. The amplified libraries were then purified using Qiagen MinElute PCR Purification kit following manufacturer's guide and eluted at 20µl volume. Libraries were sequenced on the HiSeq 2000 with 2 x 50 cycles and for an average of 50 million paired-end reads per sample. We performed the IL-4c stimulation experiment twice and generated two independent sets of libraries to obtain an optimal number of biological replicates for each macrophage population. The two independent sets of libraries are referred here after as "Run 1" and "Run 2", respectively. ATAC-seq libraries for C57BL/6 and BALB/c AAMs were generated using the same protocol and sequenced in a single run.

Transcriptional profiling of BALB/c and C57BL/6 AAMs 100,000 cells were sorted per sample as described above. FACS-purified cells were spun down and washed once with PBS before lysis with 350µl of Buffer RLT from the RNeasy Mini Kit (QIAgen). RLT lysates were homogenized by 1 minute of vortexing and were immediately stored at -80°C until RNA isolation. RNA was isolated using the RNeasy Mini Kit (QIAgen) based on manufacturer's protocol, with an additional DNase digestion step using the RNase-free DNase set (QIAgen). Transcriptional profiling was done using the CEL-seq2 protocol (*37*) and library preparation was performed at the NYU School of Medicine Genome Technology Center core facility. CEL-seq2 libraries were sequenced on the HiSeq 4000 with 2 x 50 cycles. While CEL-seq2 was originally developed as a single-cell assay, we used this protocol in this study as a bulk transcriptional profiling assay and use the more commonly-used terminology "RNA-seq" to describe data generated from this assay.

Bioinformatics and computational methods

ATAC-seq sequence processing Raw ATAC-Seq reads were aligned to the reference mouse genome mm9 using bowtie2 (v2.2.9) (*38*), with the parameters --maxin 2000 and --local,

while keeping all other parameters at default settings. To keep only highly unique alignments, reads with MAPQ score less than 30 were removed. We further removed all duplicate reads, as well as reads mapping to mitochondrial DNA and chromosome Y. Read filtering steps were done using the suite of tools from samtools (v1.2 and v1.3.1) (*39*), ngsutils (v0.5.9) (*40*) and picard-tools (http://broadinstitute.github.io/picard/, v1.1.1 and v2.8.2). After all filtering steps, reads were merged across all replicates from the same macrophage population. This resulted in a median depth of 15,235,324 reads per macrophage population in Run 1 and 9,865,310 reads per macrophage population in Run 2. For visualization of accessibility reads on the Integrative Genomics Viewer (IGV), we merged reads from the same macrophage population across samples from both runs, generated tiled data format (TDF) files using IGVtools and finally normalized the merged reads to reads per million (RPM) (*41*).

Identification of accessible chromatin regions We used the merged reads for each macrophage population to identify accessible chromatin regions, using the PeaKDEck (v1.1) peak calling algorithm, which measures signal density from randomly sampled bins genome-wide before generating a data set-specific probability distribution to identify regions with significant signal enrichment (*42*). We ran PeaKDEck using sampling bins that consist of a 75bp central bin (-bin) and a 10000bp background bin (-back). Sampling along the genome was done in steps (-STEP) of 25bp and the background probability distribution was generated using 100000 randomly selected sites (-npBack). Significance was defined using a p-value of less than 0.0001 and regions with significant p-values were defined as a "peak" (i.e. an accessible chromatin region). Peak calling was done independently on libraries generated from Run 1 and Run 2.

Generation of a union set of accessible chromatin regions We next counted the number of reads present at each accessible region in order to analyze the ATAC-Seq data using quantitative approaches downstream. To do this, we first generated a set of consensus peaks across the data set by taking the union of peaks called from each macrophage population.

Peaks were merged if overlapping by 1bp or more. The number of reads at each peak within the union peak sets were then counted for each sample. Finally, each peak was re-centered ±100bp on its summit, defined as the position with maximum pile up of reads. Re-centering on peak summits was performed as this should coincide with the binding event of a transcription factor within an accessible chromatin region. We implemented the read counting and peak summit re-centering steps directly using the dba.count function from the Bioconductor package DiffBind (version 1.14.2) (*43*). The final count matrix, which consisted of 61,713 peaks, was used for downstream analyses.

Quantitative ATAC-seq analysis ATAC-seq read counts were transformed using the regularized logarithmic (rlog) transformation as implemented in the Bioconductor package DESeq2 (*44*). To manage batch effect from the two separate libraries, we first modeled the rlog accessibility read counts to the batch variable using a linear model and subtracted out the coefficient contributed by the batch variable – this was implemented directly using the removeBatchEffect function in limma (*45*). We next chose a set of 30,856 regions with high variance, using the varFilter function in the genefilter package with default parameters, which keeps only features with variance inter-quartile range > 0.5 (*46*). We performed principal component analysis (PCA) using the batch-subtracted rlog read counts of these regions with high variance using the prcomp function in R.

To identify IL-4 dependent accessible regions, we directly compared the ATAC-seq profiles of IL-4 stimulated macrophages to their reference unstimulated macrophages, using a differential analyses workflow directly implemented through DESeq2. We fit the negative binomial model in DESeq2 using the raw accessibility reads from all 61,713 regions, with the model ~ Batch + Population**,** where Batch is a variable describing if a sample belonged in Run 1 or Run 2, while Population is a variable describing if the sample is M^{res}, AAM^{res}, M^{mono} or AAM^{mono}. IL-4 dependency was defined using a significance threshold of False Discovery Rate (FDR) of 10%. To visualize IL-4 dependent regions, we scaled the batch-subtracted rlog read counts of these IL-4 dependent regions by z-score transformation and next performed k-

means clustering on these scaled, rlog-transformed reads (K = 4). The clustered matrix was visualized as a heatmap.

Comparison of sequence properties between constitutively accessible and IL-4 induced regions *Identification of constitutively accessible and IL-4 induced regions*: To define a set of constitutively accessible regions, we used only peaks from M^{res} and M^{mono}, respectively, that were identified in both Run 1 and Run 2. This resulted in 8061 constitutively accessible regions in M^{res} and 14,045 constitutively accessible regions in M^{mono}. IL-4 induced peaks were defined using the differential analysis outlined above. All region overlap analyses throughout this study were performed using the intersect function from the BEDTools suite (*47*) and overlaps were defined as any regions overlapping by at least 1bp, unless otherwise noted.

Genomic elements enrichment analysis: We downloaded genome-wide annotations of five different genomic elements (promoter, start exon, coding exon, end exon, intron) from the UCSC Known Gene database for mm9 (*48*). We defined promoter elements as the 200bp-region upstream of a transcriptional start site (TSS). We next assigned each of the 61,713 accessible regions in our data set to a unique genomic element label. Where an accessible region overlapped two different types of genomic elements, we assigned it to the element with higher number of overlapping base pairs. Finally, any chromatin regions not assigned to one of these five genomic elements were labeled as intergenic. To determine the enrichment levels of a particular type of genomic element G within a given set of accessible regions A (either constitutively accessible or IL-4 induced regions in AAM^{res} or AAM^{mono}), we used the binomial cumulative probability distribution, $b(x; n,p)$, where x = number of success, n = number of trials and p = background probability of success. We used the pbinom function in R. We defined x to be the number of accessible regions in A that were labeled as the genomic element G that was being tested, n to be the total number of genomic elements G detected in our combined data set and p to be the proportion of the accessible region A to the total 61,713

accessible regions. This then gave the enrichment levels of *G* in *A*, relative to all the accessible regions identified across the different macrophage populations.

G/C content analysis: To calculate percentage GC, we first used the hgGcPercent function from the kentTools suite (v20170111, UCSC Genome Bioinformatics Group, https://github.com/ucscGenomeBrowser/kent) to quantitate the number of G or C bases in each accessible region. This value was next normalized using the length of the accessible region. CpG island track was downloaded from the UCSC Genome Annotation Database for mm9 (http://hgdownload.soe.ucsc.edu/goldenPath/mm9/database/). Enrichment levels of CpG island in a given set of accessible regions *A* was based on the binomial cumulative probability as described above, where x = number of accessible regions in *A* that overlapped a CpG island, n = number of CpG island in the total data set of 61,713 regions and p = proportion of *A* to the total 61,713 regions.

Calculation of distance to IL-4 induced genes: IL-4 induced genes were first identified for AAM^{res} and AAM^{mono} from the microarrays generated by Gundra and Girgis *et al* (*4*), using the linear model and empirical Bayes statistics as implemented in limma, with genes significantly induced by IL-4 defined using the thresholds FDR 10% and $\log_2$ fold change > 1. The distance between each accessible region and its closest IL-4 induced gene body was calculated using the closest function in BEDTools.

Transcription factor (TF) motif analysis Sequences were fetched using the BEDTools getfasta function for TF motif analysis with the MEME Suite tools (*49*). Whole genome fasta file for mm9 was downloaded from the Illumina igenome database (https://support.illumina.com/sequencing/sequencing_software/igenome.html). Background file was generated using the function fasta-get-markov in MEME, based on the total 61,713 accessible regions at a Markov model order of 3. TF motif databases were curated as described in (*50*). We first performed *de novo* motif discovery by running MEME (as part of

MEMEChIP, which randomly sampled 600 sequences) with the parameters: -mod zoops -nmotifs 3 -minw 6 -maxw 30. We next identified IL-4 induced PU.1 motif sites by performing motif scanning with FIMO (*51*), using the PU.1 motif discovered *de novo* from the IL-4 induced peaks of AAM^{res} and AAM^{mono}, respectively. FIMO was run with a p-value threshold of 0.0001 (as part of MEMEChIP). These IL-4 induced PU.1 motif sites were then extended ±25bp using the BEDTools slop function. We used the Centrality of Motifs (CentriMo) algorithm (*16*) to identify motifs centrally enriched around these PU.1 motif ± 25bp regions. To visualize the CentriMo results, we summarized TF motifs at TF family level as defined in (*17*). Only non-ETS family of TFs were included in the CentriMo visualization, since PU.1 belongs to the ETS-family of TFs. DNA shape features of these PU.1 motif sites ± 25bp regions were analyzed using the DNAshape algorithm (*23, 52*) for 4 different DNA shape configurations at single nucleotide resolution. The same analysis steps were repeated for the KLF motif ± 25bp regions as presented in Supplemental Figure 2B.

Comparisons between predicted PU.1 motif and ChIP-seq defined PU.1 binding sites

PU.1 ChIP-seq regions identified in thioglycollate-induced macrophages, generated from two different experiments, were downloaded as BED files that had been directly deposited on Gene Expression Omnibus (GEO) (GSM1131238 and GSM1183968) (*11, 13*). A set of 55,386 reproducible PU.1 binding sites were defined by intersecting these two sets of PU.1 ChIP-seq regions. We used the PU.1 motif discovered *de novo* from all the constitutively accessible regions in M^{mono} and ran FIMO to identify all PU.1 motif sites from M^{mono}, using a p-value threshold of 0.0001 and the background file generated as described above. Since the published PU.1 ChIP-seq regions were of 200bp length, we extended the predicted PU.1 motifs from M^{mono} by ±100bp to match the comparison. Overlapping rate was calculated as total number of predicted PU.1 motifs from M^{mono} overlapping a reproducible PU.1 ChIP-seq region / total number of predicted PU.1 motifs from M^{mono} × 100%.

Identification of cell-type-specific PU.1 binding sites We first downloaded PU.1 ChIP-seq regions as BED files directly deposited on GEO (GSM537983, GSM537989, GSM1315477) (*24, 25*) and lifted over regions annotated on the mm8 mouse reference genome to the mm9 mouse reference genome using the UCSC liftover tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). We next defined cell type specific PU.1 binding sites as PU.1 ChIP-seq regions identified in only one cell type and not in the other two. Finally, we fetched the sequences for each set of cell type specific PU.1 binding sites using the getfasta function in BEDtools. We performed *de novo* motif discovery using MEME (as part of MEMEChIP) with the parameters -mod zoops -nmotifs 3 -minw 6 -maxw 30 and a background Markov model of order 3. We used the PU.1 motif discovered *de novo* and ran FIMO to identify cell type specific PU.1 motif sites that were ChIP-seq defined. FIMO was run using a raw p-value threshold of 0.0001 and only the top 1000 motif sites (based on motif score) were included. Finally, these cell type specific PU.1 motif sites were extended ±25bp for DNA shape analysis. We used only the top 1000 motif sites to keep the number of motif sites comparable to the number of IL-4 induced PU.1 motif sites from our own study.

ATAC-seq sequence processing from C57BL/6 vs. BALB/c experiment Raw ATAC-seq reads were aligned to the mm10 mouse reference genome (for compatibility with available BALB/c sequence variant data, details below). Reads with MAPQ < 30, as well as duplicate and mitochondrial reads were all removed before peak calling. Peak calling in PeaKDEck was done using the same parameters in the IL-4c stimulation experiment, except for a p-value threshold of 0.00001. A consensus count matrix consisting of 40,981 accessible regions from all samples was generated as described above using DiffBind and used to identify strain-specific regions using the differential analysis workflow implemented through DESeq2. We fit the negative binomial model using ~ Strain + CellType + Strain:CellType, where Strain is one of either C57BL/6 or BALB/c and CellType is one of either AAM^{res} or AAM^{mono}, and extracted strain-specific regions from the comparisons: (1) BALB/c vs. C57BL/6 in AAM^{res}

and (2) BALB/c vs. C57BL/6 in AAM^{mono}. Significance was defined at a threshold of FDR 10%.

Characterizing different types of BALB/c variants at PU.1 motif regions BALB/c-specific sequence variants, including single nucleotide polymorphisms (SNPs) and insertion/deletion (indels), were downloaded as variant call format (VCF) files from the Sanger Mouse Project (ftp://ftp-mouse.sanger.ac.uk/). Variants that were not labeled as "PASS" under the Filter flag were removed using vcftools (*53*). These VCF files were then converted to the BED file format using the vcf2bed function from bedops (*54*) for overlap analysis with the IL-4 induced PU.1 regions. Since these BALB/c variants were identified on the mm10 genome, we lifted over IL-4 induced PU.1 regions defined initially on the mm9 coordinates to the mm10 coordinates. PU.1 motif sites that directly overlapped sequence variants were identified using the BEDTools intersect function. To define PU.1 motif sites with flanking variants, we use the BEDTools closest function to calculate distances between a PU.1 motif and its closest BALB/c variant and identified motifs with a closest variant at an absolute distance of <25bp, but not overlapping a variant directly. We also limited the DNA shape prediction to SNPs only, to avoid the shift in position due to the insertions/deletions.

Generation of BALB/c whole genome fasta To construct a BALB/c whole genome fasta for DNA shape prediction, we took the list of BALB/c variants filtered from above and used the vcf2diploid tool (*55*) to integrate these BALB/c variants into the mm10 reference genome. To annotate PU.1 regions on the BALB/c coordinates, we used the accompanying chain file generated from the vcf2diploid tool in the above step to liftover the PU.1 regions from mm10 coordinates to BALB/c coordinates. The mm10 and BALB/c whole genome fasta, as well as PU.1 regions on corresponding C57BL/6 and BALB/c coordinates, were then used with the getfasta function in BEDTools to fetch the sequences of specific IL-4 induced PU.1 regions with flanking SNPs for shape prediction with DNAShape algorithm.

Quantitating the amount of shape change induced by SNP To quantitate the amount of shape change induced by the presence of BALB/c variant(s) at an accessible region, we calculated the Euclidean distance between the two vectors of shape values that corresponded to shape values predicted from a C57BL/6 region and a BALB/c region, respectively (*35*). A larger Euclidean distance would indicate a greater shift in shape. We first identified all accessible regions that carried SNP(s) only from across the data set (i.e. regions that did not contain any sequence variants, or regions that contained both SNPs and insertion/deletions were discarded), as including insertion/deletions in the analysis would result in comparisons between sequences of unequal length and it would not be feasible to calculate the Euclidean distance at these regions. We then grouped regions as "strain-specific" and "strain-common" based on results from the differential analysis (922 strain-specific regions and 4776 strain-common regions). Shape prediction was done using the DNAShape algorithm and Euclidean distance was calculated for each pair of shape values using the dist function in R.

Identification of predictors for strain-specific and strain-common regions To identify features associated with the 922 strain-specific and 4776 strain-common SNP-carrying accessible regions, we first assessed each of these accessible regions for the following characteristics:

- $\text{EuclideanDistance}_{\text{ProT}}$ = amount of shape change induced by the BALB/c SNP(s) in the ProT shape configuration
- $\text{EuclideanDistance}_{\text{Roll}}$ = amount of shape change induced by the BALB/c SNP(s) in the Roll shape configuration
- $\text{EuclideanDistance}_{\text{MGW}}$ = amount of shape change induced by the BALB/c SNP(s) in the MGW shape configuration
- $\text{GCPercent}_{\text{B6}}$ = % GC calculated using C57BL/6 sequences
- CpG_{B6} = A binary variable (Yes / No) indicating if the region overlapped a CpG island

- GenomicElements = The type of genomic element that the region is labeled as (one of promoter, start exon, end exon, coding exon, intron or intergenic)
- SnpFrequency = Number of SNPs carried by the region
- IL4Inducibility = A binary variable (Yes / No) indicating if the region was IL-4 inducible

We then fit a logistic regression model as follows using the glm function in R, with the parameter family = "binomial":

Class ~ $EuclideanDistance_{ProT}$ + $EuclideanDistance_{Roll}$ + $EuclideanDistance_{MGW}$ + $GCPercent_{B6}$ + CpG_{B6} + GenomicElements + SnpFrequency + IL4Inducibility

The dependent variable "Class" is a label describing if the region is "strain-specific" or "strain-common". Significance of predictor was determined by analysis of variance for the coefficients of the predictors, using a p-value cutoff of < 0.05.

Processing of CEL-seq reads CEL-seq reads were first demultiplexed using the bc_demultiplex script from https://github.com/yanailab/CEL-Seq-pipeline (*37*). Demultiplexed reads were aligned to the mm10 mouse reference genome using bowtie2 (version 2.2.9). Aligned reads were counted for each gene using a modified htseq-count script (from https://github.com/yanailab/CEL-Seq-pipeline) adapted for CEL-seq reads with unique molecular identified (UMI). We included only reads with MAPQ score > 30 and removed singleton genes. This resulted in a final median read depth of 737,848 reads per sample, covering a median of 11,096 genes per sample. PCA was performed using 7431 genes with high variance, defined using the varFilter function in the genefilter package with default parameters, which keeps only features with variance inter-quartile range > 0.5. Differential analysis was done using DESeq2 by fitting the negative binomial model using ~ Strain + CellType + Strain:CellType. Significantly differential genes were extracted using a threshold of FDR 10% for the four different comparisons of: (1) AAM^{res} vs. AAM^{mono} in C57BL/6, (2) AAM^{res} vs. AAM^{mono} in BALB/c, (3) BALB/c vs. C57BL/6 in AAM^{res} and (4) BALB/c vs.

C57BL/6 in AAM^{mono}. Overlapping genes were defined as genes identified as differential in two different comparisons. Pathway enrichment analysis was done using strain-specific genes in AAM^{res} and AAM^{mono}, respectively, through Ingenuity Pathway Analysis (IPA) with the parameter Organism = Mouse and keeping all other parameters at default settings. The Gene Ontology annotation terms associated with each of the genes listed in Figure 4D were fetched through the BioMart project as implemented through the Bioconductor biomaRt package (*56*).

References

1. C. K. Glass, G. Natoli, Molecular control of activation and priming in macrophages. *Nature immunology* **17**, 26 (2016).
2. J. E. Allen, R. M. Maizels, Diversity and dialogue in immunity to helminths. *Nature Reviews Immunology* **11**, 375 (2011).
3. U. M. Gundra *et al.*, Vitamin A mediates conversion of monocyte-derived macrophages into tissue-resident macrophages during alternative activation. *Nature immunology* **18**, 642-653 (2017).
4. U. M. Gundra *et al.*, Alternatively activated macrophages derived from monocytes and tissue macrophages are phenotypically and functionally distinct. *Blood* **123**, 22 (2014).
5. R. E. Thurman *et al.*, The accessible chromatin landscape of the human genome. *Nature* **489**, 75 (2012).
6. A. B. Stergachis *et al.*, Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**, 888-903 (2013).
7. Y. Lavin *et al.*, Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. *Cell* **159**, 1312-1326 (2014).
8. S. Ghisletti *et al.*, Identification and Characterization of Enhancers Controlling the Inflammatory Gene Expression Program in Macrophages. *Immunity* **32**, 317-328 (2010).
9. G. Natoli, S. Ghisletti, I. Barozzi, The genomic landscapes of inflammation. *Genes & development* **25**, 101-106 (2011).
10. R. Ostuni *et al.*, Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157-171 (2013).
11. M. U. Kaikkonen *et al.*, Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Molecular cell* **51**, 310-325 (2013).
12. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218 (2013).
13. S. Heinz *et al.*, Effect of natural genetic variation on enhancer selection and function. *Nature* **503**, 487-492 (2013).
14. J. C. Walsh *et al.*, Cooperative and antagonistic interplay between PU.1 and GATA-2 in the specification of myeloid cell fates. *Immunity* **17**, 665-676 (2002).
15. S. R. McKercher *et al.*, Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *The EMBO Journal* **15**, 5647-5658 (1996).
16. T. L. Bailey, P. Machanick, Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research* **40**, e128-e128 (2012).
17. E. Wingender, T. Schoeps, M. Haubrock, M. Krull, J. Dönitz, TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Research* **46**, D343-D347 (2018).

18. T. Satoh *et al.*, The Jmjd3-Irf4 axis regulates M2 macrophage polarization and host responses against helminth infection. *Nature immunology* **11**, 936-944 (2010).
19. A. Chawla, Control of macrophage activation and function by PPARs. *Circulation research* **106**, 1559-1569 (2010).
20. B. Daniel *et al.*, The IL-4/STAT6/PPARγ signaling axis is driving the expansion of the RXR heterodimer cistrome, providing complex ligand responsiveness in macrophages. *Nucleic Acids Research*, gky157-gky157 (2018).
21. M. Gaboli *et al.*, Mzf1 controls cell proliferation and tumorigenesis. *Genes & Development* **15**, 1625-1630 (2001).
22. I. Barozzi *et al.*, Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Molecular cell* **54**, 844-857 (2014).
23. T. Zhou *et al.*, DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic acids research* **41**, W56-W62 (2013).
24. A. Crotti *et al.*, Mutant Huntingtin promotes autonomous microglia activation via myeloid lineage-determining factors. *Nature neuroscience* **17**, 513 (2014).
25. S. Heinz *et al.*, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589 (2010).
26. K. S. Vignesh *et al.*, IL-4 induces metallothionein 3-and SLC30A4-dependent increase in intracellular Zn2+ that promotes pathogen persistence in macrophages. *Cell reports* **16**, 3232-3246 (2016).
27. S. M. Campbell *et al.*, Myeloid cell recruitment versus local proliferation differentiates susceptibility from resistance to filarial infection. *eLife* **7**, (2018).
28. C. J. Martin, K. N. Peters, S. M. Behar, Macrophages Clean Up: Efferocytosis and Microbial Control. *Current opinion in microbiology* **0**, 17-23 (2014).
29. R. Khare, M. L. Hillestad, Z. Xu, A. P. Byrnes, M. A. Barry, Circulating antibodies and macrophages as modulators of adenovirus pharmacology. *Journal of virology* **87**, 3678-3686 (2013).
30. S. Soudi *et al.*, Comparative Study of The Effect of LPS on The Function of. **15**, 45-54 (2013).
31. Y. Qiao *et al.*, Synergistic Activation of Inflammatory Cytokine Genes by Interferon-γ-induced Chromatin Remodeling and Toll-like Receptor Signaling. *Immunity* **39**, 10.1016/j.immuni.2013.1008.1009 (2013).
32. C. Nishiyama *et al.*, Functional analysis of PU.1 domains in monocyte-specific gene regulation. *FEBS Letters* **561**, 63-68 (2004).
33. P. Mazzi, M. Donini, D. Margotto, F. Wientjes, S. Dusi, IFN-gamma induces gp91phox expression in human monocytes via protein kinase C-dependent phosphorylation of PU.1. *J Immunol* **172**, 4941-4947 (2004).
34. M. T. Maurano *et al.*, Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190-1195 (2012).
35. S. C. J. Parker, L. Hansen, H. O. Abaan, T. D. Tullius, E. H. Margulies, Local DNA Topography Correlates with Functional Noncoding Regions of the Human Genome. *Science* **324**, 389 (2009).
36. S. J. Jenkins *et al.*, Local macrophage proliferation, rather than recruitment from the blood, is a signature of TH2 inflammation. *Science* **332**, 1284-1288 (2011).
37. T. Hashimshony *et al.*, CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome biology* **17**, 77 (2016).
38. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**, 357-359 (2012).
39. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078-2079 (2009).
40. M. R. Breese, Y. Liu, NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics (Oxford, England)* **29**, 494-496 (2013).

41. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**, 178-192 (2013).
42. M. T. McCarthy, C. A. O'Callaghan, PeaKDEck: a kernel density estimator-based peak calling program for DNaseI-seq data. *Bioinformatics (Oxford, England)* **30**, 1302-1304 (2014).
43. R. Stark, G. Brown. (2011).
44. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).
45. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
46. R. Gentleman, V. Carey, W. Huber, F. Hahne. (2016).
47. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841-842 (2010).
48. F. Hsu *et al.*, The UCSC Known Genes. *Bioinformatics (Oxford, England)* **22**, 1036-1046 (2006).
49. T. L. Bailey *et al.*, MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208 (2009).
50. K. Karwacz *et al.*, Critical role of IRF1 and BATF in forming chromatin landscape during type 1 regulatory cell differentiation. *Nature immunology* **18**, 412 (2017).
51. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)* **27**, 1017-1018 (2011).
52. T.-P. Chiu *et al.*, DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics (Oxford, England)* **32**, 1211-1213 (2016).
53. P. Danecek *et al.*, The variant call format and VCFtools. *Bioinformatics (Oxford, England)* **27**, 2156-2158 (2011).
54. S. Neph *et al.*, BEDOPS: high-performance genomic feature operations. *Bioinformatics (Oxford, England)* **28**, 1919-1920 (2012).
55. J. Rozowsky *et al.*, AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522 (2011).
56. D. Smedley *et al.*, The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research* **43**, W589-W598 (2015).

Figures

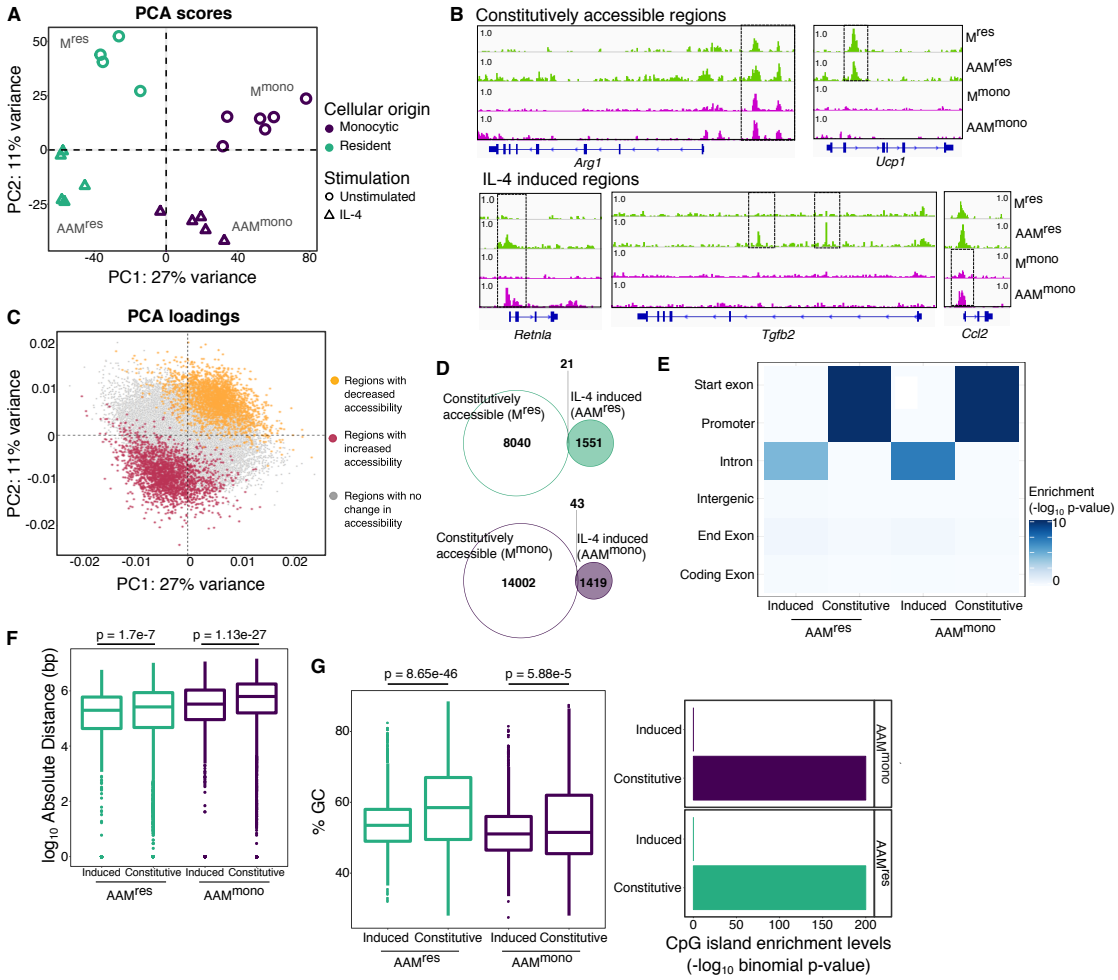


Figure 1: IL-4 stimulation leads to remodeling of open chromatin landscape in peritoneal macrophages.

(A) PCA scores of individual ATAC-seq samples. PCA was performed using rlog-transformed ATAC-seq read counts of 30,856 regions with high variance (only regions with variance inter-quartile range > 0.5 were retained). (B) Genome browser views of representative (boxed) constitutively accessible and IL-4 induced regions. (C) The contributions of individual accessible regions to PCs 1 and 2 are represented in the PCA loadings plot. Each data point is color-coded based on the direction of its IL-4 dependency. Hence, IL-4 induced regions (red) are highly associated with IL-4 stimulated macrophages, while IL-4 repressed regions (yellow) are highly associated with unstimulated macrophages. We compared (D) enrichment levels for different types of genomic elements, (E) distance from a closest IL-4-induced gene and (F) G/C content between constitutively accessible and

IL-4-induced regions in AAM^{mono} and AAM^{res}, respectively. G/C content information is represented in two different ways – percentage of G/C bases in an accessible region (F, left panel) and CpG island enrichment for a given group of accessible regions (F, right panel). Enrichment p-values are from binomial test while two-class comparison p-values are from Mann-Whitney test.

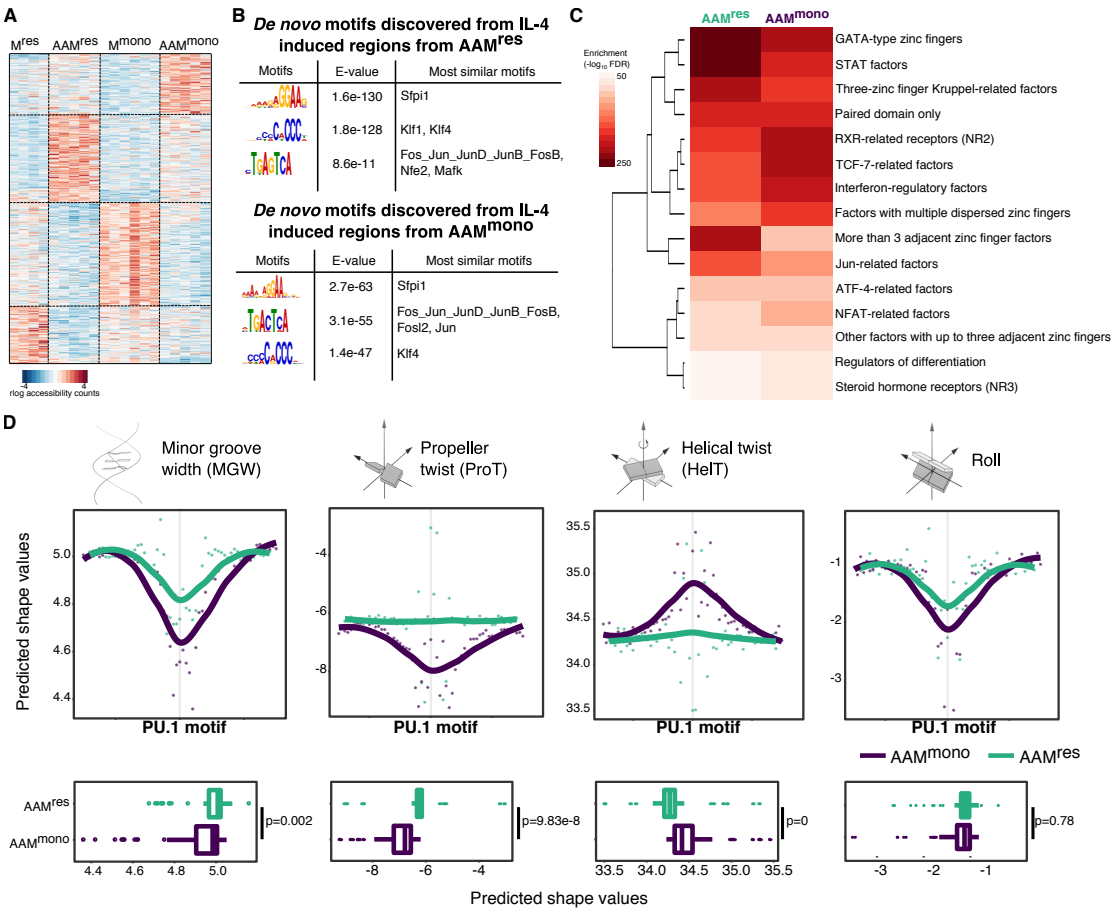


Figure 2: IL-4 induced regions are associated with PU.1 motif that has macrophage-specific DNA shape features.

(A) Heatmap visualizing the macrophage-specific IL-4 dependent regions. Each row represents one of the 2855 IL-4 dependent regions and each column a unique sample. Values are rlog-transformed, batch-subtracted read counts, scaled using a z-score transformation for each region. (B) Motifs discovered *de novo* from 1572 IL-4 induced regions in AAM^{res} and 1462 IL-4 induced regions in AAM^{mono}. (C) Heatmap visualizing enrichment levels of different TF families that were centrally enriched around IL-4 induced PU.1 regions in AAM^{mono} and AAM^{res}. Only TF families with $-\log_{10}$ FDR > 20 are visualized. FDR values are based on the Centrality of Motifs (CentriMo) algorithm. (D) Average of DNA shape features at IL-4 induced PU.1 regions of AAM^{mono} and AAM^{res}. Scatter plots are centered on IL-4 induced PU.1 motifs (x-axes), with the solid lines representing Loess fit of predicted shape values at single-nucleotide resolution (dots). Boxplots are average predicted shape values over the PU.1 motif +/- 25bp windows and p-values are from Kolmogorov-Smirnov test.

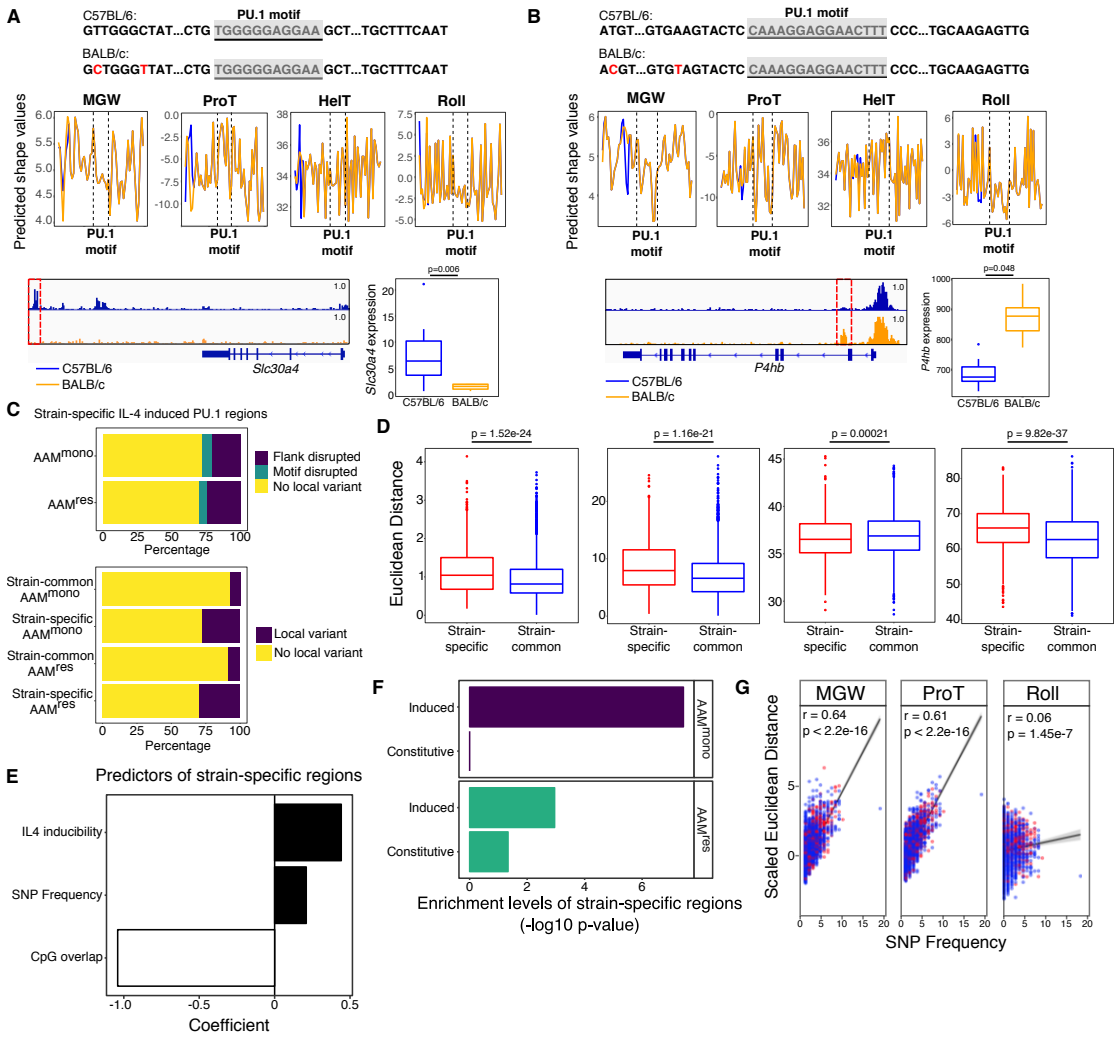
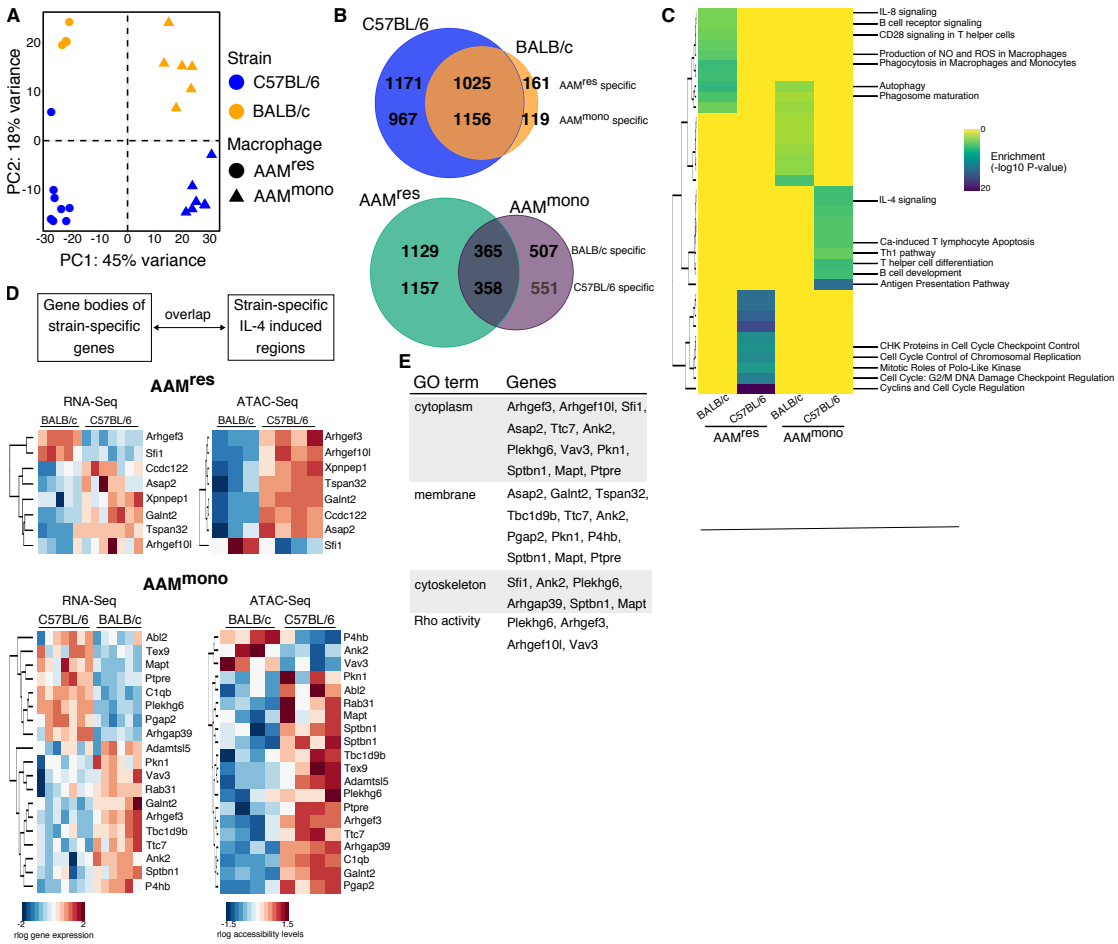


Figure 3: Accessibility of IL-4 induced PU.1 regions can be altered through DNA shape change induced by sequence mutation, without disruption of the PU.1 motif.

(A-B) Selected examples from (A) AAM^{res} and (B) AAM^{mono} of an IL-4 induced PU.1 motif with flanking BALB/c SNPs. In both examples, sequences represent a unique IL-4 induced PU.1 motif region, where the PU.1 motif is underlined and shaded in grey, while BALB/c SNPs are highlighted with red font. Line graphs of predicted shape values are centered on the PU.1 motif (sequences between the vertical lines). Genome browser tracks illustrate the strain-specific accessibility of the region containing the matching sequence (boxed in red). Boxplots represent size-factor normalized read counts in C57BL/6 and BALB/c macrophage. P-values are from DESeq2 (N = 4-8 mice per group). (C) Frequency of different types of sequence variants in strain-specific IL-4 induced PU.1 regions (top) and all IL-4 induced PU.1 regions (bottom). P = 58 strain-specific IL-4 induced PU.1 regions (AAM^{mono}), 37

strain-specific IL-4 induced PU.1 regions (AAM^{res}), 754 strain-common IL-4 induced PU.1 regions (AAM^{mono}) and 828 strain-common IL-4 induced PU.1 regions (AAM^{res}). (D) Comparison of the amount of SNP-induced DNA shape change in 922 SNP(s)-containing strain-specific and 4776 SNP(s)-containing strain-common accessible regions, respectively. The shift in DNA shape for each sequence is quantitated using Euclidean distance, where a larger Euclidean distance value indicates a larger shift in shape. P-values are from Mann-Whitney tests. (E) Coefficient values representing the relative contribution of each significant predictor in distinguishing between strain-specific and strain-common accessible regions. (F) Enrichment values of IL-4 induced and constitutively accessible regions in strain-specific regions for AAM^{mono} and AAM^{res}. Enrichment p-value is based on binomial test. (G) Scatter plots illustrating the positive correlations between SNP frequency and amount of DNA shape change undergo by a specific region. Red = strain-specific regions, blue = strain-common regions. Euclidean distances here are scaled across each DNA shape feature for visualization purpose. Correlation coefficient and p-values are based on the Spearman's test using un-scaled Euclidean distances.

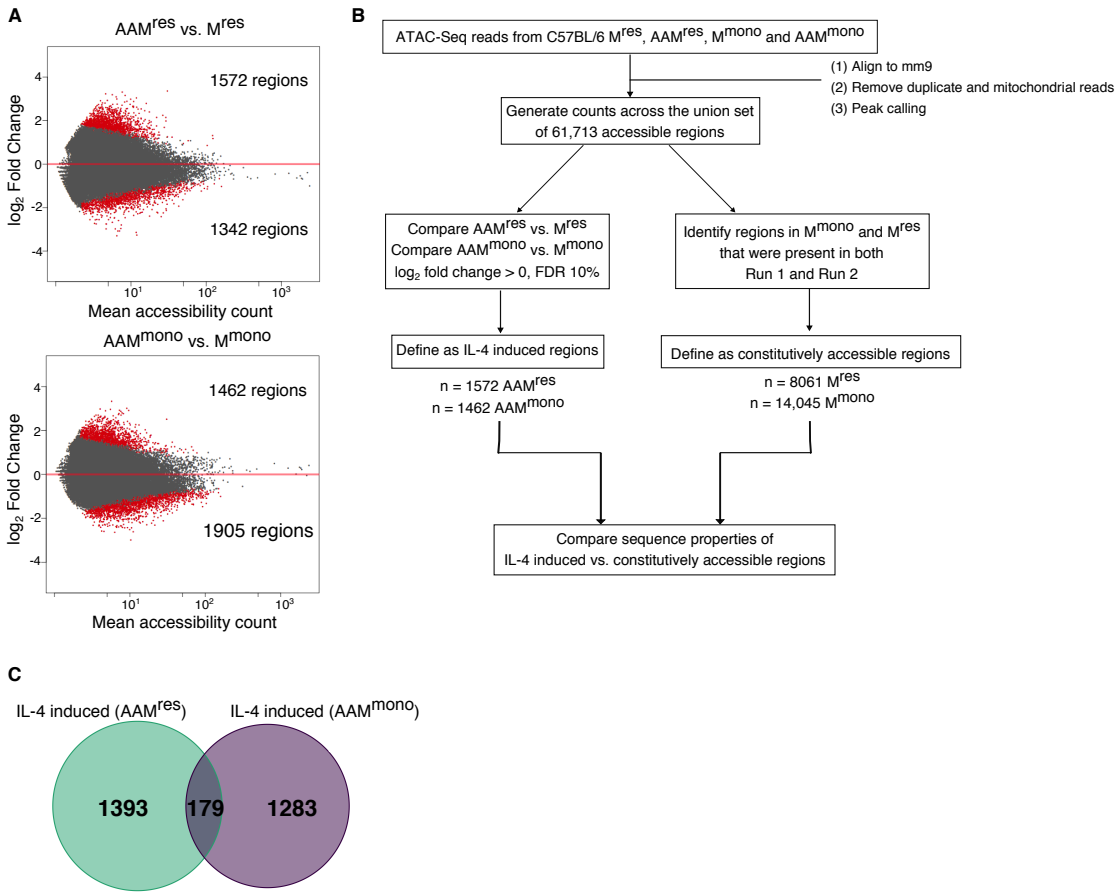


GO term	Genes
cytoplasm	Arhgef3, Arhgef10l, Sfi1, Asap2, Ttc7, Ank2, Plekhg6, Vav3, Pkn1, Sptbn1, Mapt, Ptpre
membrane	Asap2, Galnt2, Tspan32, Tbc1d9b, Ttc7, Ank2, Pgap2, Pkn1, P4hb, Sptbn1, Mapt, Ptpre
cytoskeleton	Sfi1, Ank2, Plekhg6, Arhgap39, Sptbn1, Mapt
Rho activity	Plekhg6, Arhgef3, Arhgef10l, Vav3

Figure 4: AAMs from C57BL/6 and BALB/c are functionally distinct.

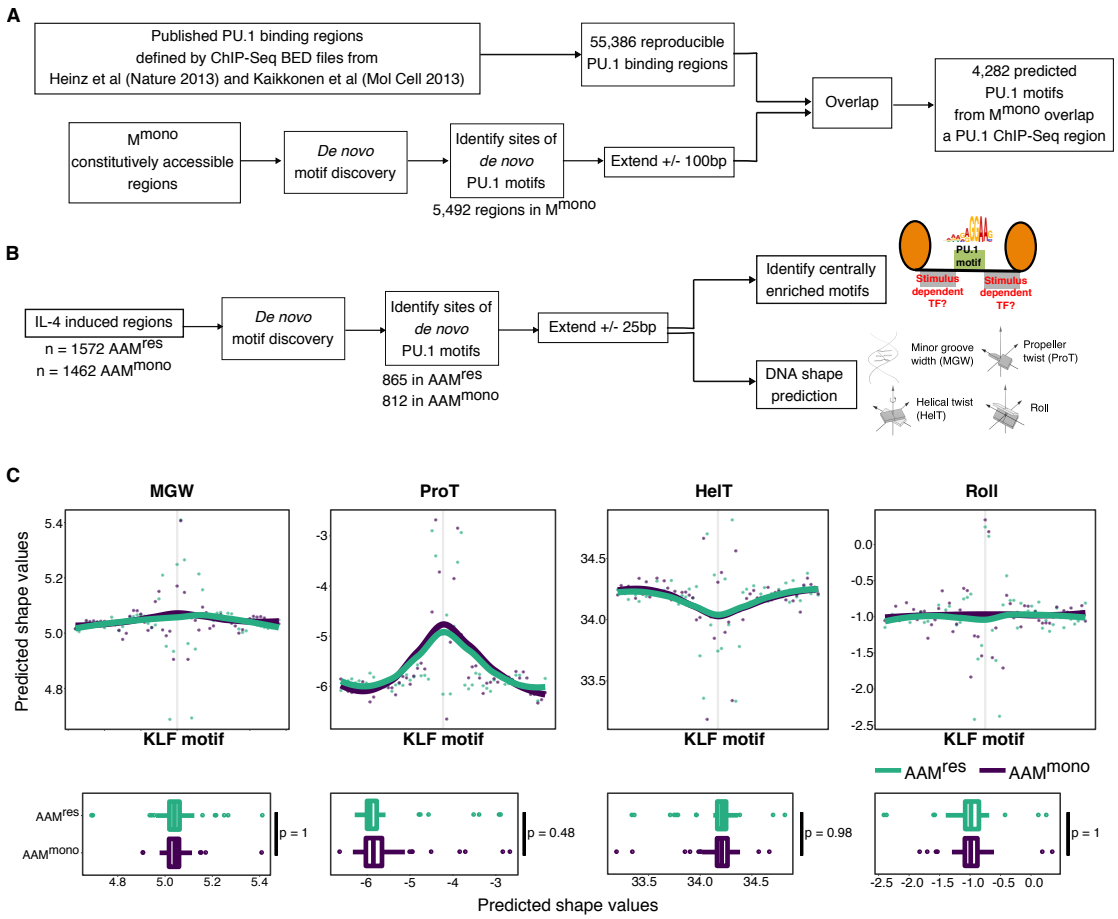
(A) PCA of 7431 genes with high variance (only genes with variance inter-quartile range > 0.5 were retained). (B) Venn diagrams indicating the number of genes that were commonly and uniquely identified as significantly differential in different comparisons. Top panel: Identification of macrophage-specific genes in C57BL/6 and BALB/c AAMs, respectively. Bottom panel: Identification of strain-specific genes in AAM^{res} and AAM^{mono}, respectively. (C) Enrichment values from Ingenuity Pathway Analysis visualized as $-\log_{10}$ P-value for the four different groups of genes – (1) BALB/c specific in AAM^{res}, (2) C57BL/6 specific in AAM^{res}, (3) BALB/c specific in AAM^{mono} and (4) C57BL/6 in AAM^{mono}. Only the top 10 pathways (as defined by enrichment p-values) are included in this visualization. Specific pathways are highlighted for clarity purpose. The complete list of enriched pathways is provided as Supplemental Materials Data S2. (D) List of strain-specific genes that also contained one or more strain-specific IL-4 induced regions on their gene bodies. Heatmaps

correspond to scaled, rlog-transformed RNA-seq and ATAC-seq reads from the matching genes and IL-4 induced regions. (E) Gene Ontology (GO) terms associated with specific genes listed in (D).



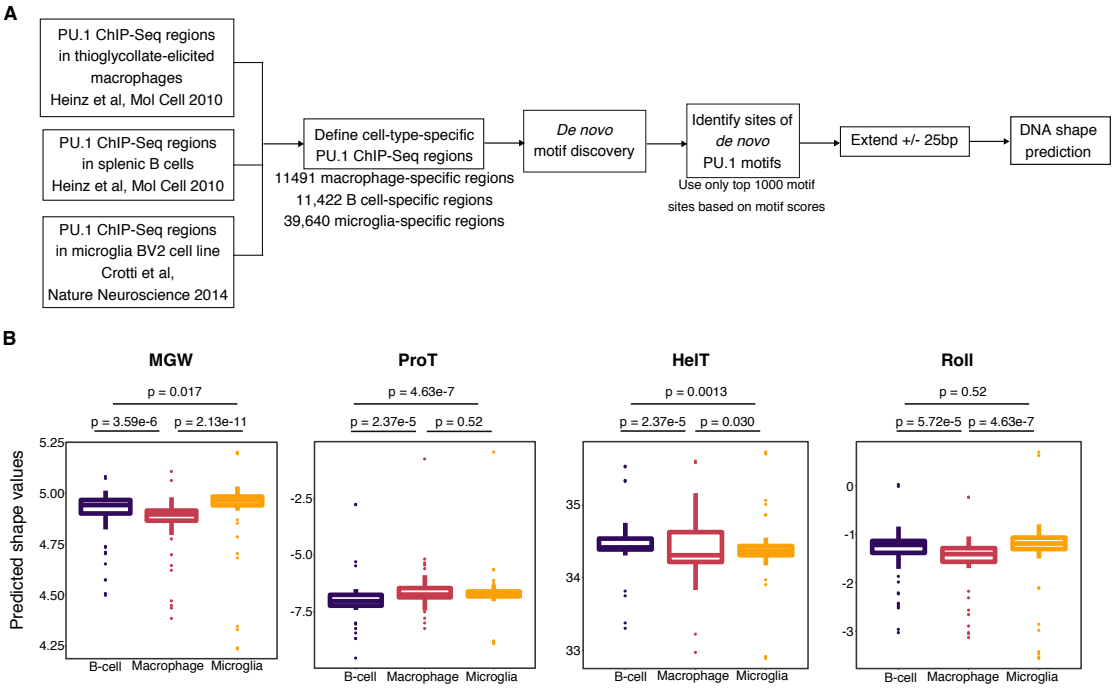
Supplemental Figure 1: IL-4 stimulation leads to remodeling of open chromatin landscape in peritoneal macrophages.

(A) Comparisons between the accessible chromatin profiles of IL-4 stimulated macrophages and unstimulated macrophages, presented as MA plots (left – AAM^{res}; right – AAM^{mono}). Differential chromatin regions (FDR 10%, |LFC| > 0) are highlighted in red. (B) Venn diagram illustrating the minimal overlap between IL-4 induced regions in AAM^{res} and AAM^{mono}. Values in Venn diagrams represent the number of unique accessible regions in each corresponding set. (C) Schematic outlining the workflow to identify IL-4 induced regions and constitutively regions for comparisons of the sequence properties associated with these two classes of genomic elements.



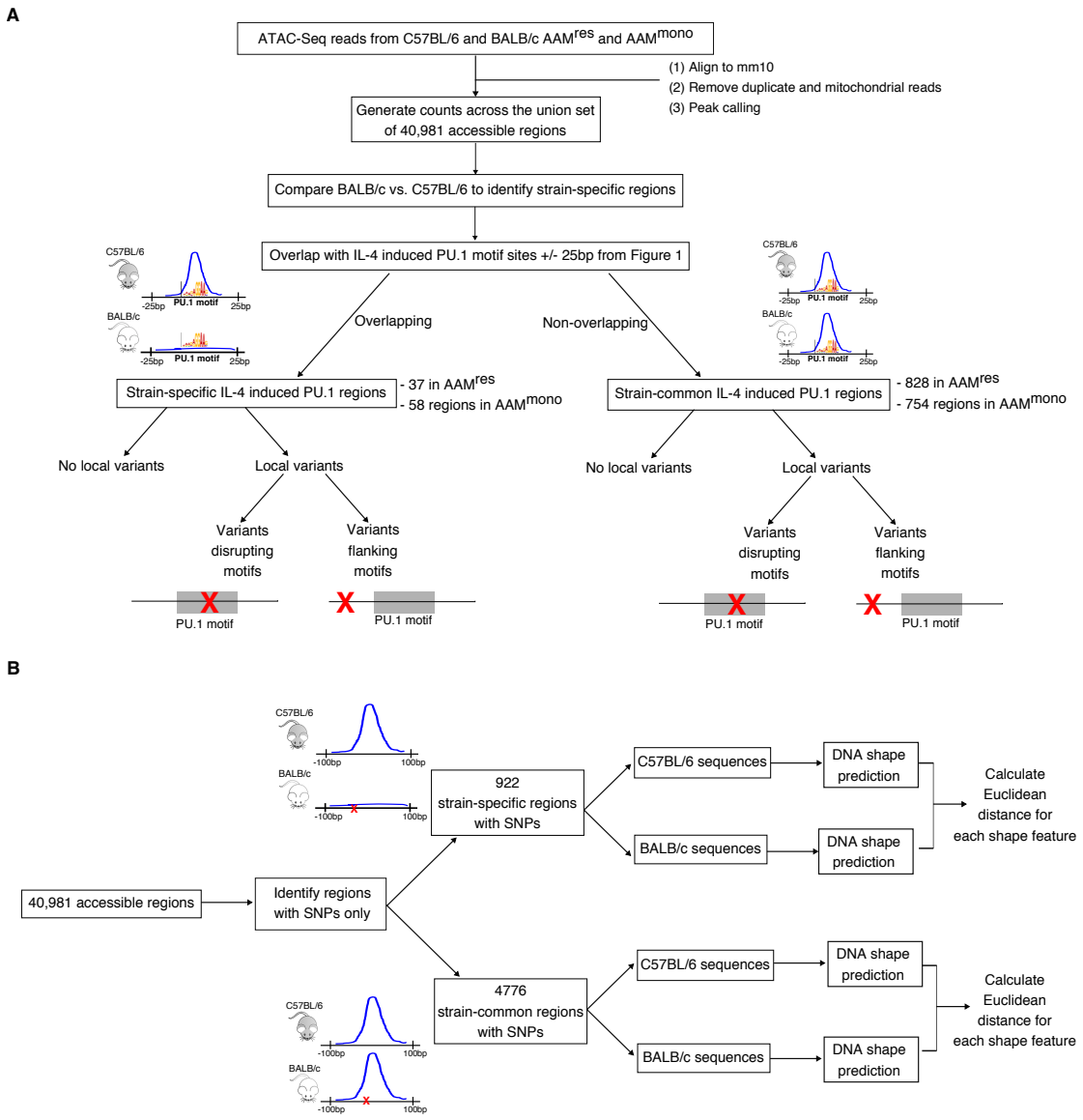
Supplemental Figure 2: Cell-type-specific local DNA shape features are specific to PU.1 motif regions.

(A) Schematic outlining the workflow used to verify the accuracy of predicted PU.1 motif sites by comparing to published ChIP-Seq data sets. (B) Schematic outlining the workflow used to analyze features of IL-4 induced PU.1 motif sites. (C) Average of DNA shape features at IL-4 induced KLF regions of AAM^{mono} and AAM^{res}. P = 902 regions in AAM^{res}, 627 regions in AAM^{mono}. Scatter plots are centered on IL-4 induced KLF motifs (x-axes), with the solid lines representing Loess fit of predicted shape values at single-nucleotide resolution (dots). Boxplots are average predicted shape values over the KLF motif ± 25bp windows and p-values are from Kolmogorov-Smirnov test.



Supplemental Figure 3: Distinct local DNA shape at cell-type-specific PU.1 binding regions can be reproduced in an independent setting.

(A) Schematic outlining the workflow used to identify cell-type-specific PU.1 binding regions from published ChIP-seq data sets for DNA shape prediction. (B) Average predicted DNA shape values over PU.1 motif ± 25bp derived from published PU.1 ChIP-seq data. P-values are from Kolmogorov-Smirnov test.



Supplemental Figure 4: Leveraging on natural genetic variation in C57BL/6 and BALB/c mouse strains to perform a mutation study.

(A) Schematic with graphical illustrations outlining the strategy used to identify IL-4 induced PU.1 motif sites with altered chromatin accessibility due to flanking SNP(s). Blue line indicates read pile-ups from an accessible region, while the "X" in red represent a motif-flanking SNP. (B) Schematic with graphical illustrations outlining the strategy used to classify strain-specific and strain-common accessible regions carrying SNP(s), as well as to quantitate the amount of DNA shape change induced by SNP(s) in these regions.

Table S1.

AAM^{res} (Figure 3A) **C57BL/6 (chr2:122656227-122656288)** GTTGGGCTATGGTCACACTGTCCTG TGGGGGAGGAA GCTGAGATCATTGTCTGCTTTCAAT **BALB/c (chr2:122639812-122639873)** GCTGGGTTATGGTCACACTGTCCTG TGGGGGAGGAA GCTGAGATCATTGTCTGCTTTCAAT
AAM^{mono} (Figure 3B) **C57BL/6 (chr11:120571348-120571413)** ATGTGACACTTCCTGTGAAGTACTC CAAAGGAGGAACTTT CCCCAGCTCCCACATGCAAGAGTTG **BALB/c (chr11:120542029-120542094)** ACGTGACACTTCCTGTGTAGTACTC CAAAGGAGGAACTTT CCCCAGCTCCCACATGCAAGAGTTG

Full length sequences of PU.1 motif with flanking SNPs used as example in Figures 3A-B.

SNPs are highlighted in red font, while PU.1 motifs are underlined.

Table S2.

Type of variants	Strain-specific IL-4 induced PU.1 regions in AAM^{res}	Strain-common IL-4 induced PU.1 regions in AAM^{res}	Strain-specific IL-4 induced PU.1 regions in AAM^{mono}	Strain-common IL-4 induced PU.1 regions in AAM^{mono}
Motif disrupting	2	16	4	9
Motif flanking	9	53	12	45
No local variants	26	702	42	699

Frequency count of different types of sequence variants in all IL-4 induced PU.1 regions.

Data S1.

Metadata, coordinates and sequences of all strain-specific IL-4 induced PU.1 regions identified in AAM^{res} and AAM^{mono} (related to Figure 3A-B).

Data S2.

List of pathways enriched among strain-specific genes from AAM^{res} and AAM^{mono}.