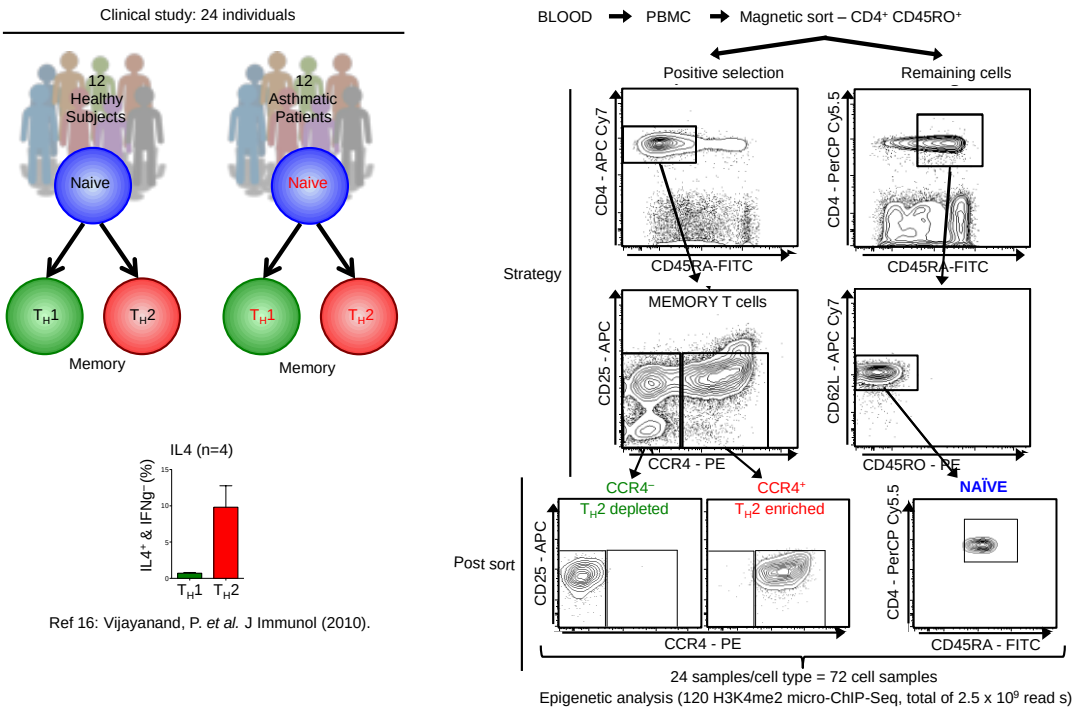
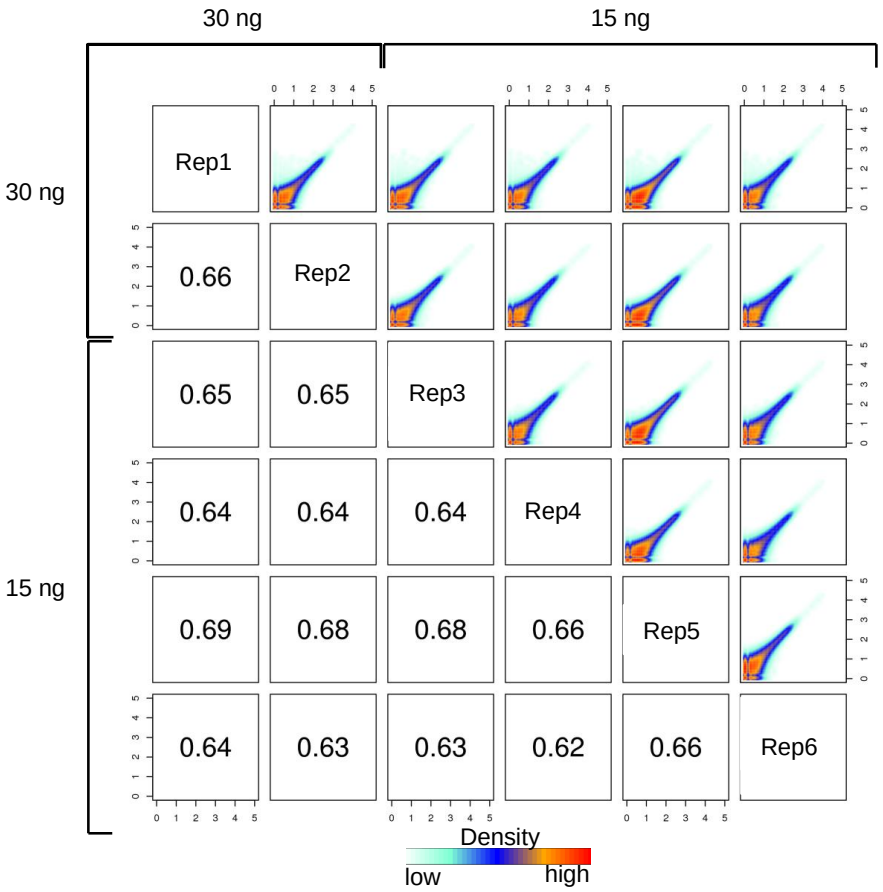


Ref 16: Vijayanand, P. *et al.* J Immunol (2010).

Supplementary Figure 1

Schematic model summarizing the study design and cell sorting strategy.

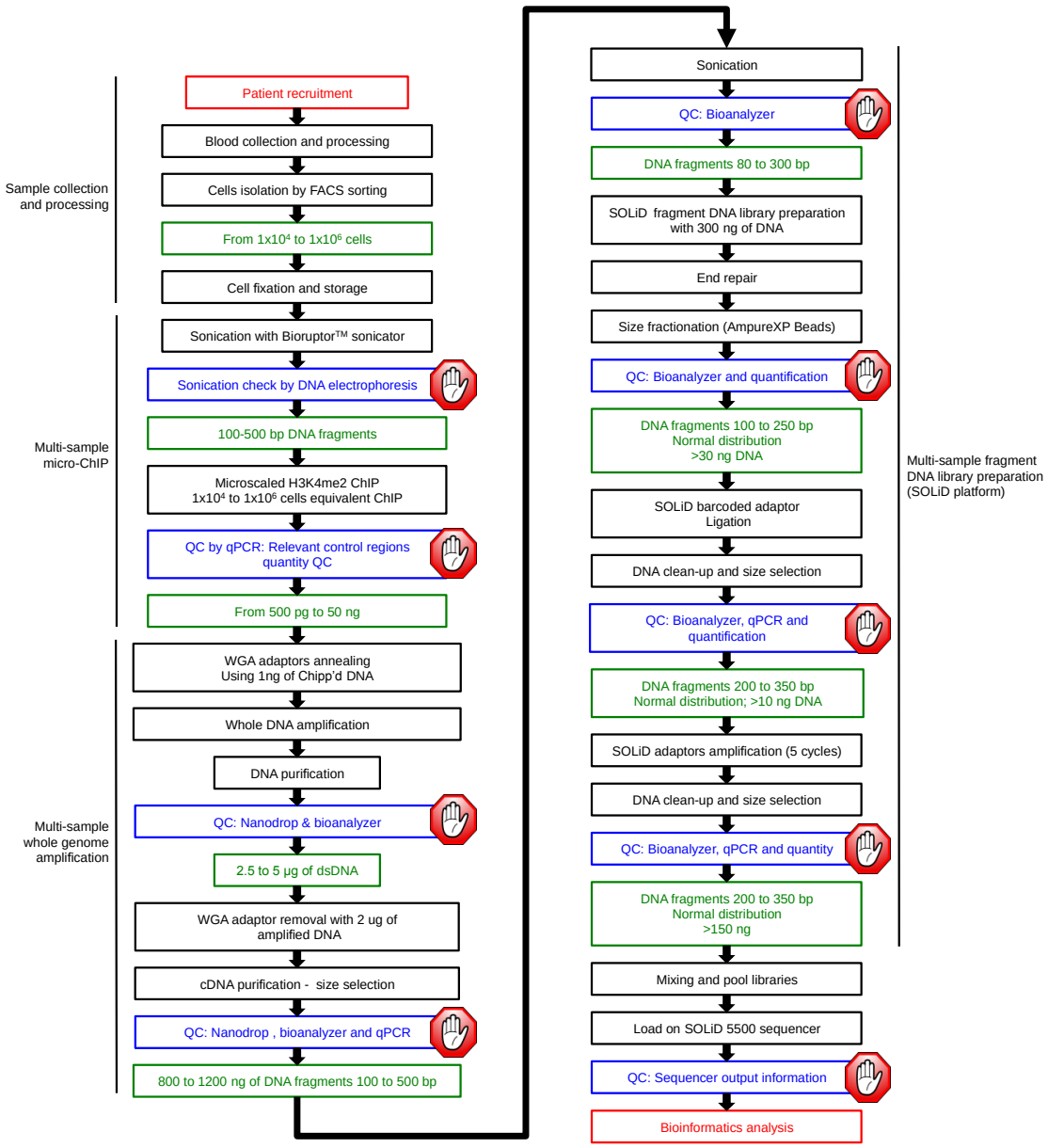
Schematic diagram depicts the cell types isolated from peripheral blood of healthy and asthmatic subjects. Sorting strategy for isolating naïve and CCR4$^-$ (T_H1), CCR4$^+$ (T_H2) memory T cells from peripheral blood mononuclear cells (PBMC), and FACS plots pre and post-sorting are shown. The number of samples processed (passing quality control checks) for H3K4me2 ChIP-Seq assay is shown below.



Supplementary Figure 2

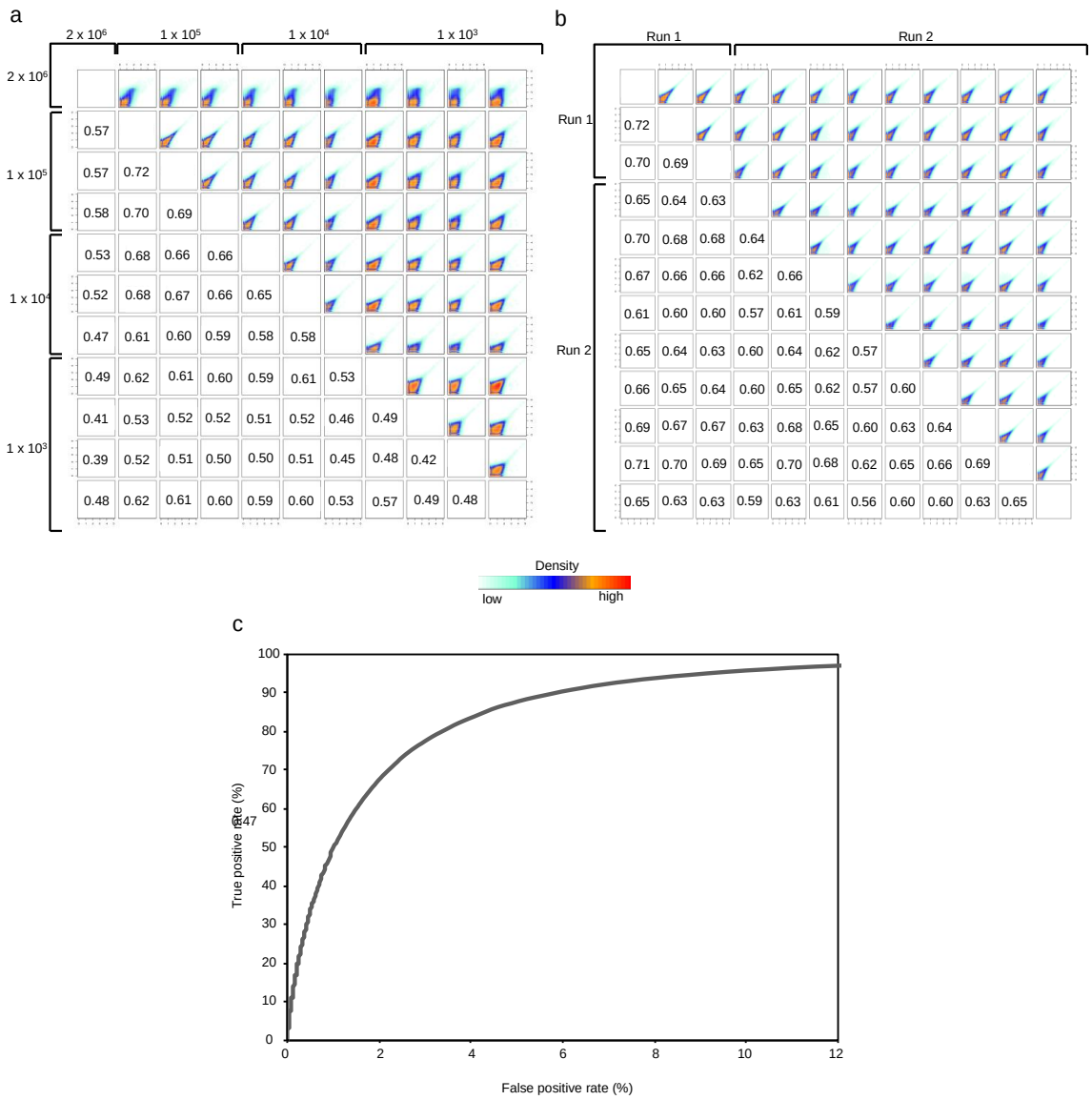
Reproducibility of standard H3K4me2 ChIP-seq assay.

Density plots show pair-wise comparison of sequencing coverage (number of reads) at genome-wide 500 bp windows (*MEDIPS* v.1.10.0 software, **Methods** and **Supplementary Notes**) obtained from 6 independent standard H3K4me2 ChIP-Seq assays performed with 2 x 10^6 D10 cells. The amount of DNA (post H3K4me2 ChIP; 15ng or 30ng) used for whole genome amplification (see **Methods** and **Supplementary Notes**) is shown. Pairwise Spearman correlation values (numbers inside boxes) for genome-wide comparison of H3K4me2 enrichment patterns between replicate samples (labeled as Rep 1-6) are illustrated.



Supplementary Figure 3

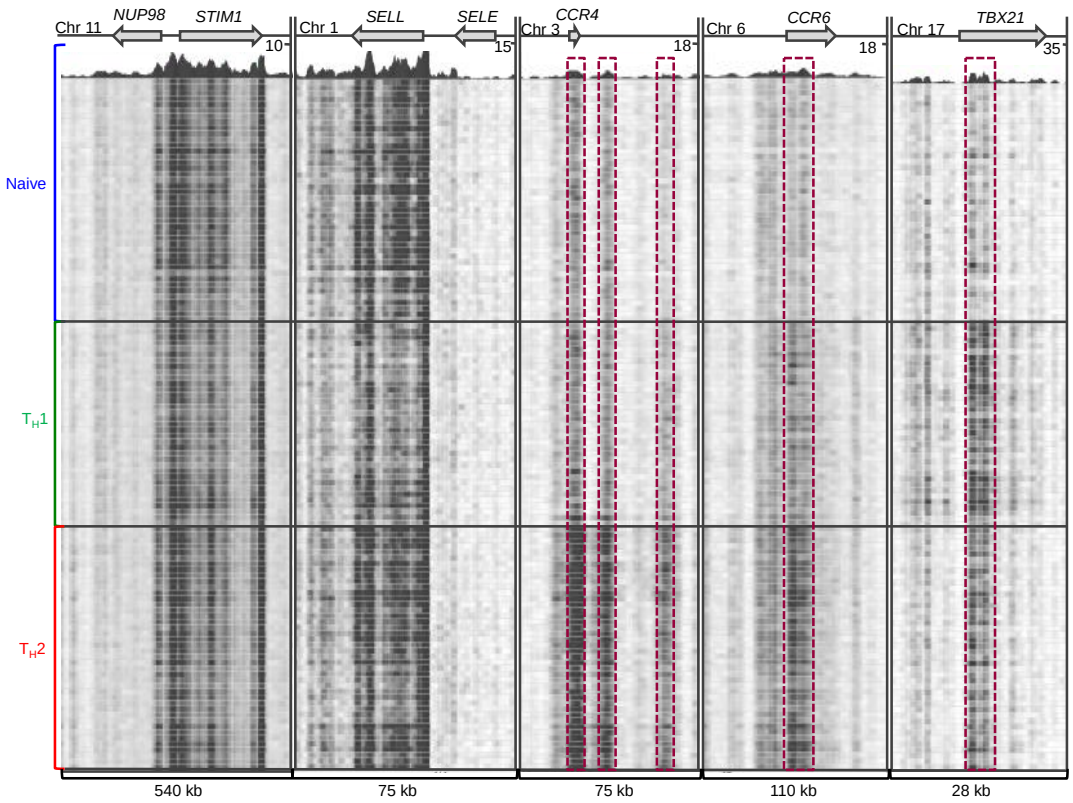
Workflow of microscaled ChIP-seq assay from sample isolation to sequencing.



Supplementary Figure 4

Reproducibility of the micro-scaled H3K4me2 ChIP-seq assay.

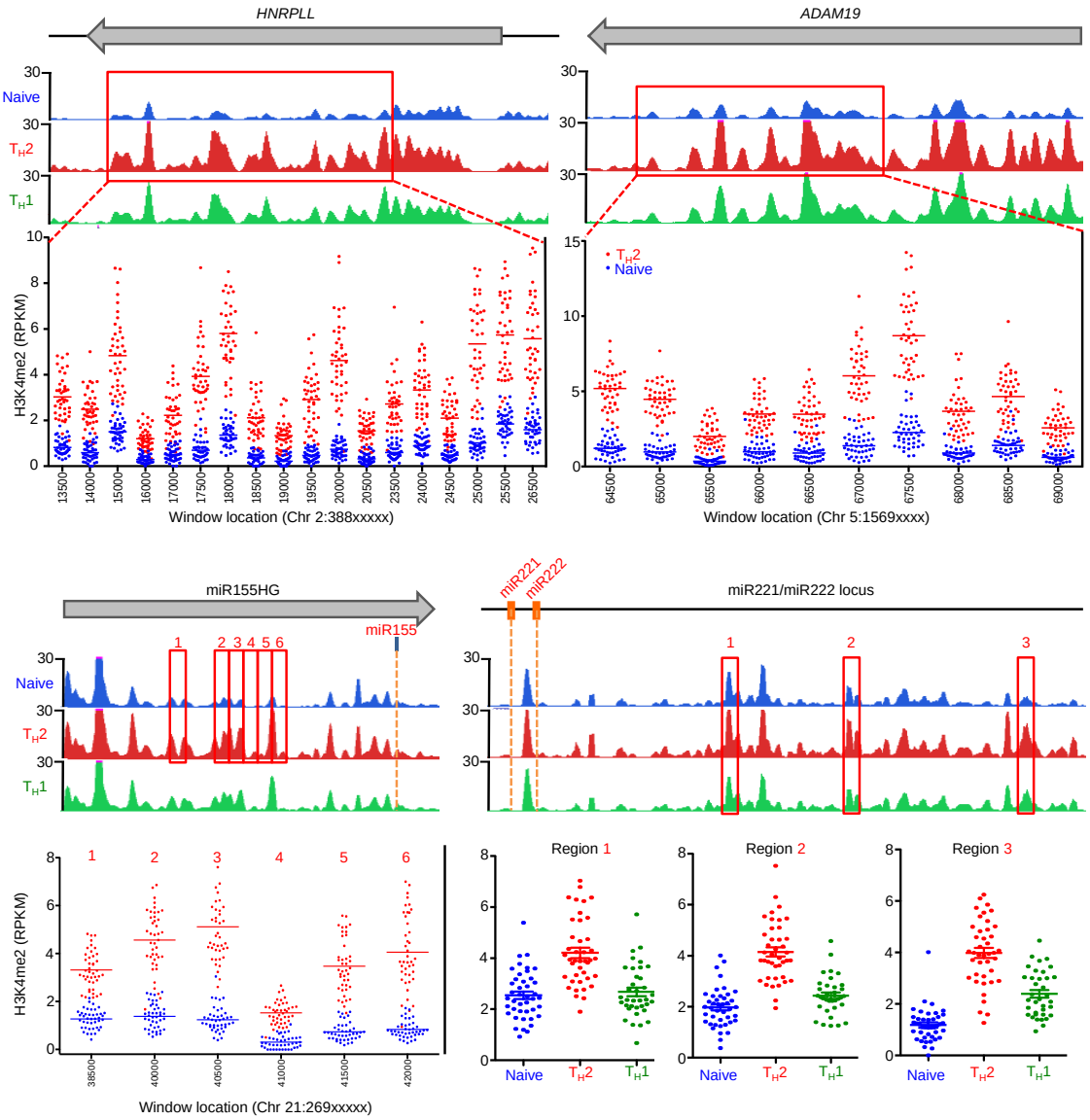
(**a**) Density plots show pair-wise comparison of sequencing coverage at genome-wide 500bp windows (*MEDIPS* v.1.10.0 software, **Methods** and **Supplementary Notes**) obtained from standard ChIP-Seq (2 x 10^6 cells) and micro-scaled ChIP-Seq (10^5, 10^4 and 10^3 cell samples) performed with D10 cells. Pairwise Spearman correlation values (numbers inside boxes) for genome-wide comparison of H3K4me2 enrichment patterns between assays performed with different cell numbers are illustrated. (**b**) Shows density plots and pairwise Spearman correlation between multiple (n = 11) replicate micro-scaled ChIP-Seq assays (10^5 cell samples) performed in two separate sequencing runs (Run1 and Run2). (**c**) ROC analysis (detailed methodology is described in **Methods** and **Supplementary Notes**) shows the percentage of true and false positives identified by micro-scaled ChIP-Seq assay (10^5 cell sample) when tested for the top 1% of enriched windows (true positives) identified in the standard ChIP-Seq assay.



Supplementary Figure 5

Reproducibility of the microscaled H3K4me2 ChIP-seq assay.

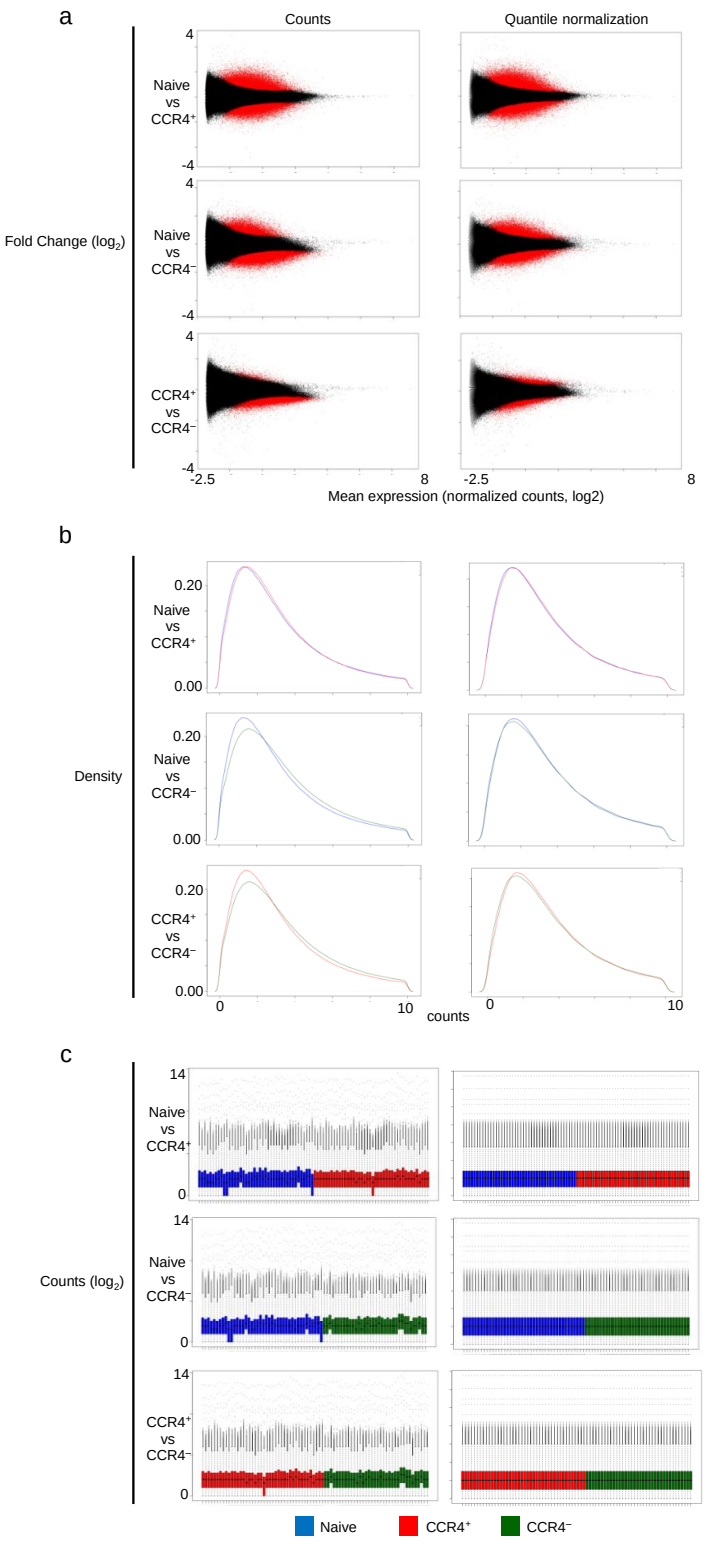
ChIP-Seq analysis showing H3K4me2 enrichment patterns from each assay (rows), for the following gene loci: control regions: *STIM 1, NUP98, SELP and SELL* gene loci*; n*on-expressed gene: *SELE locus*; T_H2 cell-type specific regions: *CCR4* and *CCR6* locus; T_H1 cell-type specific regions: *TBX21* (encoding T-BET), performed in peripheral blood naïve, T_H1 and T_H2 memory T cells from all study subjects. The significant cell-specific H3K4me2 enrichment across (enhancers and promoters) in these loci is highlighted in the red dashed line boxes.



Supplementary Figure 6

Sensitivity of the H3K4me2 ChIP-seq assay.

ChIP-Seq analysis showing cell-specific H3K4me2 enrichment patterns, for the following gene loci: *HNRPLL*, *ADAM19*, miR155 and miR221-222, in naïve, T_H1 and T_H2 cells. For each specific cell-type, data was merged from all donors including assay duplicates. H3K4me2 enrichment values for specific 500 bp windows (highlighted in red boxes) are shown in the graphs below. Each dot represents data from a single assay-donor; error bars indicate mean ± (s.e.m.).

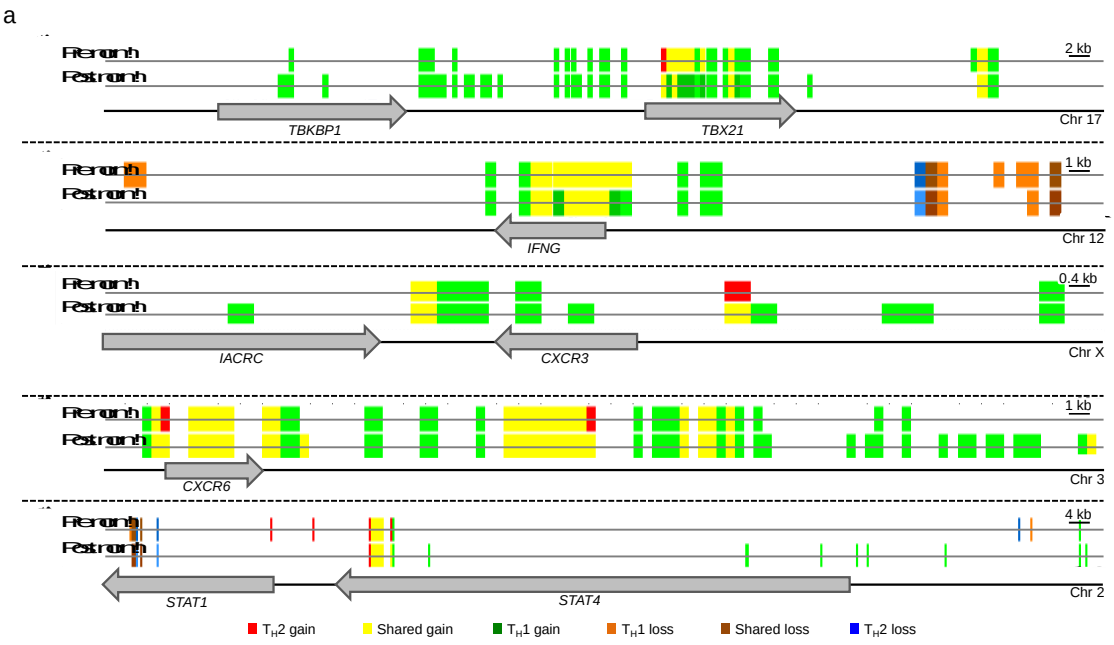


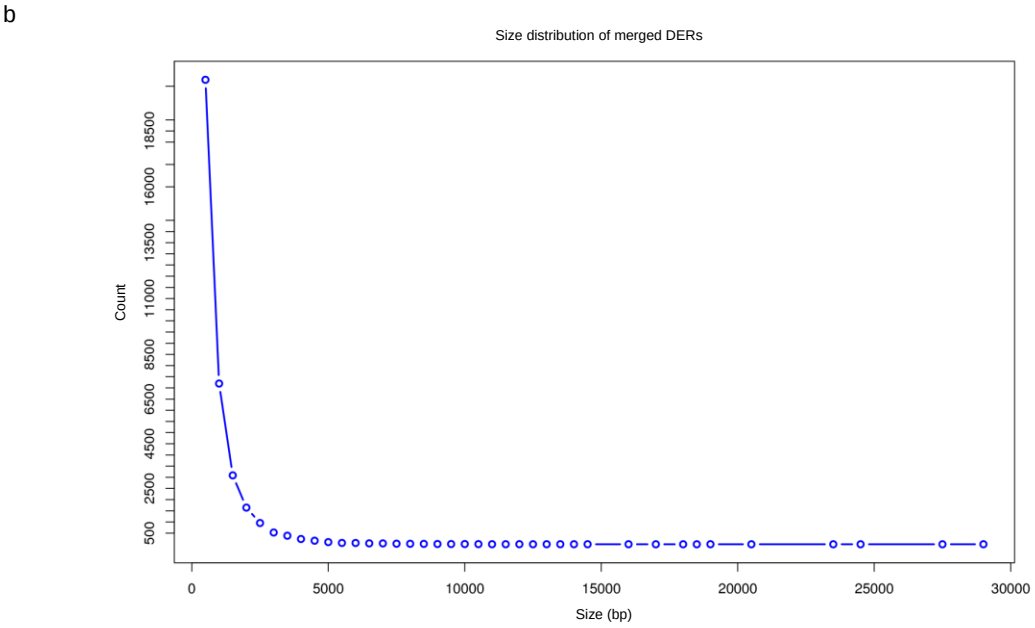
Supplementary Figure 7

Bioinformatic analysis of H3K4me2 ChIP-seq data set.

Diagnostic plots examining different characteristics of the ChIP-Seq data based on raw counts (*left*) and

quantile normalized counts (*right*) resulting from the three pairwise cell type comparisons. The MA plots (*top*) contrast $\log_2$ fold changes (*y*-axis) against mean sequencing coverage (*x*-axis) for all genome wide 500 bp windows. Genomic windows with a Bonferroni adjusted *P*-value <0.05 are indicated in red. The density plots (*middle*) show the relative distribution of read counts (*x*-axis) at genome wide 500 bp windows for windows with read counts ≤10. The boxplots (*bottom*) show read counts at genome wide 500 bp windows for individual assays.

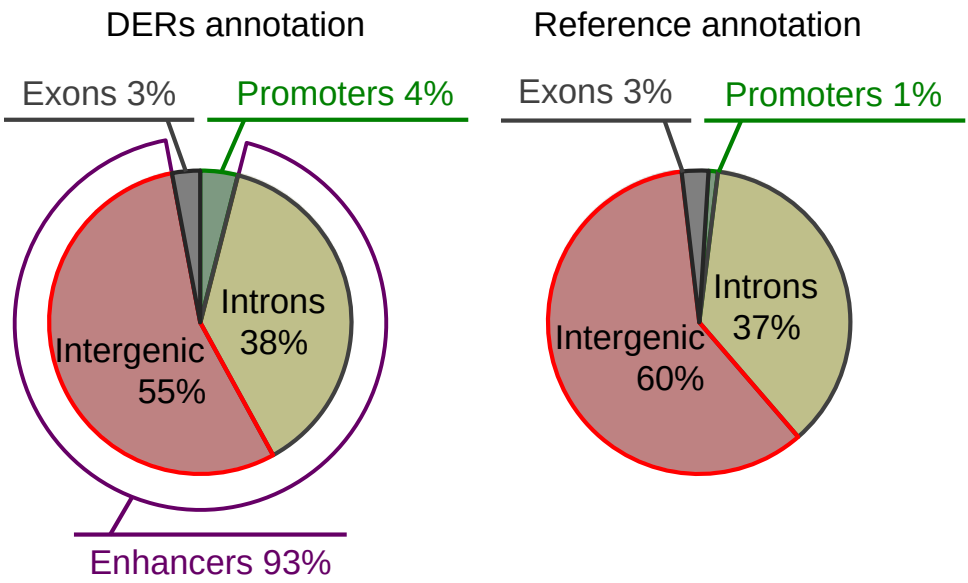




Supplementary Figure 8

Bioinformatic analysis of H3K4me2 ChIP-seq data set.

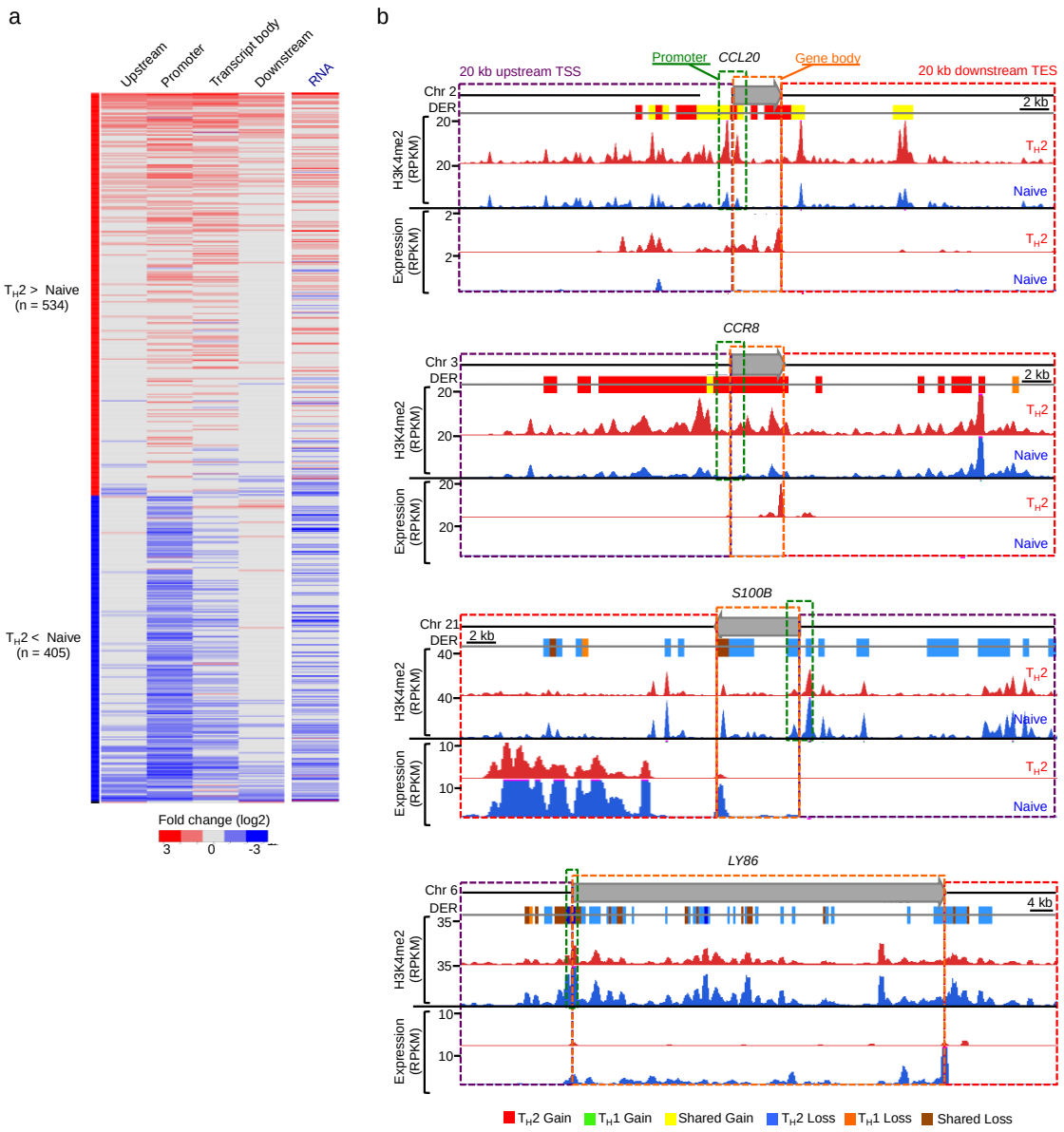
(**a**) As examples, cell-type specific enhancer DER tracks (pre- and post-normalization (norm)) along with UCSC tracks are shown for the following gene loci: *TBX21*, *IFNG*, *CXCR3*, *CXCR6*, *STAT1* and *STAT4*, where additional DERs were detected following quantile normalization of the ChIP-Seq data (**Supplementary Fig. 7**). (**b**) The number and size distribution of DERs after merging consecutive DERs (see **Methods** and **Supplementary Notes**).



Supplementary Figure 9

Genomic distribution of differentially enriched *cis*-regulatory regions (DERs).

Pie chart (*left*) shows the distribution of DERs in different genomic regions, and compared to the reference annotation (*right*)

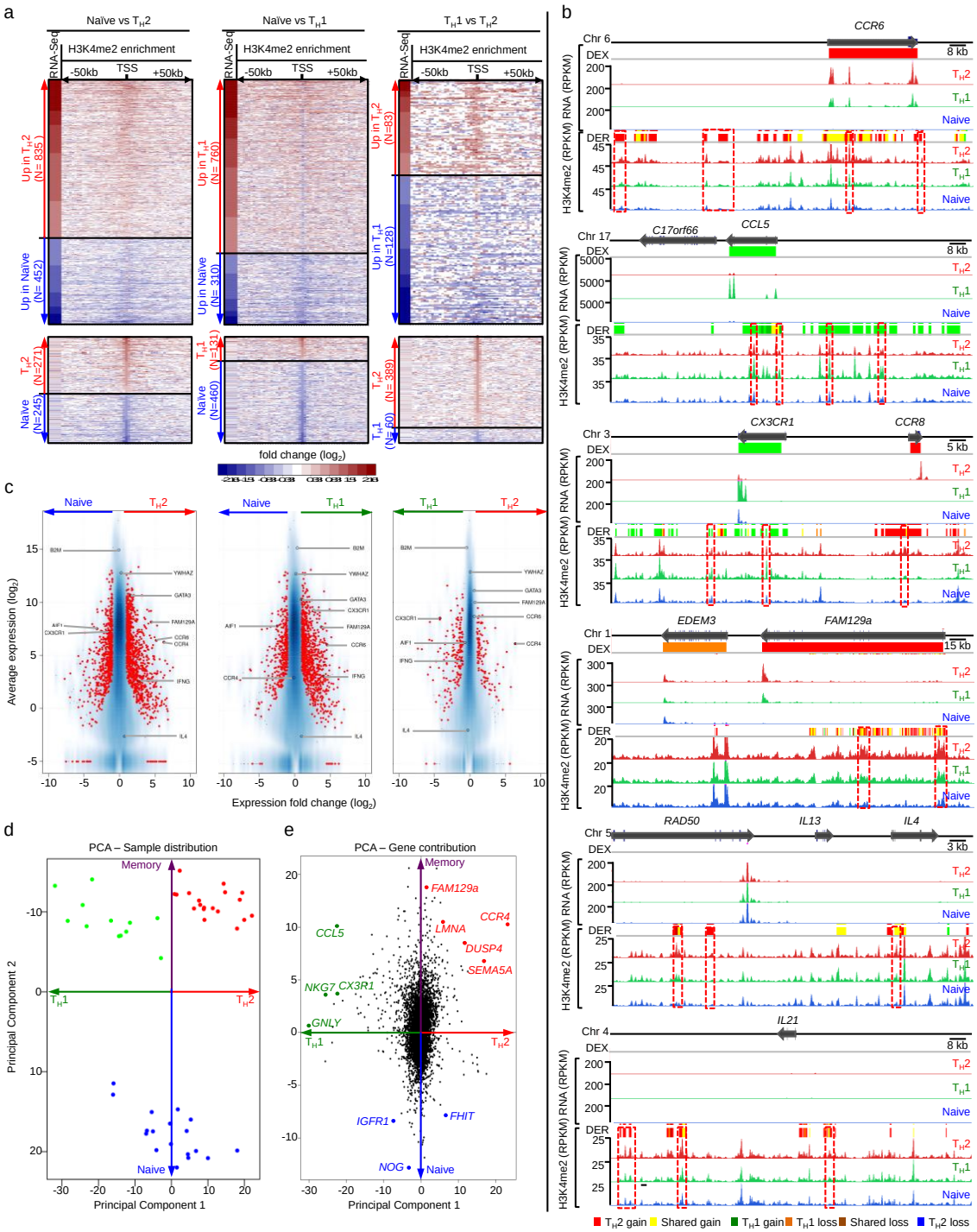


Supplementary Figure 10

Concordant changes in gene expression and H3K4me2 enrichment patterns at promoter and distal *cis*-regulatory elements (enhancers).

(**a**) Heat map shows the comparison of gene expression (RNA-Seq) and H3K4me2 enrichment patterns (across an extended genomic region) for transcripts having a DER in their promoter region when comparing naïve to T_H2 memory $CD4^+$ T cells (each row represents one transcript). Upstream = -20 kb from transcription start site (TSS); Promoter = +/-1 kb around TSS; Transcript body = region between TSS and transcript end site (TES); Downstream = +20 kb from TES; RNA = RNA-Seq data. The heat map indicates concordant gain (*red*) or loss (*blue*) of H3K4me2 enrichment in T_H2 memory compared to naïve $CD4^+$ T cells, at the promoter and at enhancers located in or close to that transcript (see **Methods** and **Supplementary Notes**). Similarly, the last column indicates concordant up- (*red*) or down- (*blue*) regulation of gene expression in T_H2 memory cells compared to naïve $CD4^+$ T cells for the corresponding transcripts. (**b**) As examples, H3K4me2 enrichment and RNA-Seq tracks for naïve and T_H2 memory T cells (data merged from all donors including duplicate assays) are shown for the following gene loci: *CCL20* and *CCR8* (T_H2 gain); *S100B* and *LY86* (T_H1 gain), along with

UCSC gene tracks (*top row*) and cell-type specific enhancer DERs.



Supplementary Figure 11

Concordant changes in gene expression and H3K4me2 enrichment patterns at promoter and distal *cis*-regulatory elements.

(**a**) Heat map shows the comparison of gene expression (RNA-Seq) and H3K4me2 enrichment patterns (across an extended genomic region) for genes that are differentially expressed in any of the three pairwise comparisons of naïve, T_H2 and T_H1 cells (each row represents one gene). First column of the heat map

(labeled RNA-Seq) displays log_2 fold change in gene expression values. Next to the RNA-Seq data, fold change in H3K4me2 enrichment pattern across each extended gene locus (transcription start site (TSS) +/-50 kb) is displayed (see **Methods** and **Supplementary Notes**). The heat map indicates concordant changes in (up-*red* or down-*blue*) gene expression and H3K4me2 enrichment at the promoter and distal enhancer regions. *Bottom panel* shows genes with no significant changes in expression levels, but display a promoter-localized differentially enriched H3K4me2 region (DER) in any of the three pairwise comparisons of naïve, T_H2 and T_H1 cells. (**b**) As examples, RNA-Seq and H3K4me2 enrichment tracks for naïve and memory T cells (data merged from all donors including duplicate assays) are shown for the following gene loci: *CCR6, CCR8* and *FAM129a* (T_H2 active genes); *CCL5* and *CX3CR1* (T_H1 active genes); *IL4 and IL21* (T_H2 poised genes), along with UCSC gene tracks (*top row*), cell-type specific DERs and differentially expressed genes (DEX). Red dashed line boxes highlight cell-specific DERs. (**c**) Shows MA plots (vertically displayed) for the pairwise comparisons of naïve, T_H2 and T_H1 cells, and red dots indicate differentially expressed genes (false discovery rate of 1%). (**d-e**) Shows principal component analysis (PCA) for RNA-Seq data from each sample, and the genes contributing to the PCA.

Supplementary Notes

Additional Methods

Micro-scaled multi-sample ChIP-Sequencing. Additional considerations that improved the reproducibility and consistency of our micro-scaled ChIP-Seq assay are: *(i)* Handling all tissue and cell isolation procedures at 4 ^{o}C to minimize changes from epigenetic patterns that are present *in vivo*. *(ii)* Formaldehyde-fixed cell pellets were snap frozen in liquid nitrogen and stored at –80 ^{o}C prior to the ChIP assay. This storage step gave us the flexibility to perform ChIP assays in large batches of samples (n ~ 24) from subjects in disease and control groups, thus minimizing assay-to-assay variability. *(iii)* Validating sonication conditions for reproducible shearing of chromatin from small numbers of cells. We found that using the BioRuptor^{TM} sonicator, which has a 12-tube holder, aided equivalent shearing of chromatin among the multiple samples that were to be compared. The extent of chromatin shearing was verified for all samples by using highly sensitive SYBR gold^{TM} dye to stain DNA run on agarose gels. *(iv)* We validated a picogreen^{TM}-based assay (Life Technologies) for measuring picogram quantities of DNA, to ensure the use of equal starting amounts of template chromatin from all the samples when undertaking multi-sample ChIP assays or whole genome amplification steps post-ChIP. *(v)* Switching to tubes with low DNA and protein binding, and changing tubes during post IP washes significantly reduced assay background while preserving specific enrichment in ChIP assays. *(vi)* qPCR-based measurement of enrichment at control genes was used to guarantee comparable ChIP performance across multiple samples. *(vii)* Using equivalent starting amounts of post-ChIP DNA (1 ng) for whole genome amplification (WGA, Sigma), ensured that the number of PCR cycles for amplification remained similar across the samples to be compared.

Additional considerations that improve the validity of our micro-scaled ChIP-Seq approach *(i)* The whole-genome amplification step that is essential for preparing libraries from micro-scaled ChIP assays does not introduce random noise. Instead, the observed differences in amplification yield for different regions of the genome are highly reproducible as can be seen in the high correlation between different amplification runs (**Supplementary Fig. 4b**). Therefore, differences between samples (using 10^5 primary

cells) that have undergone the same amplification procedure, which is what we are looking at with our DER analysis, will reflect differences in the un-amplified samples.

(ii) The majority of regions with a high number of reads in the unamplified sample also have a high number of reads in the amplified sample. To quantify this, we have performed an receiver operating characteristic (ROC) analysis in which the top 1% of windows in the standard sample was identified using the read counts in windows from the amplified sample (excluding windows with no reads in either sample to avoid overestimation of performance due to genomic regions in which no reads can be mapped). This gave a very high area under the curve (AUC) of 0.976, with 70% of the high scoring windows in the standard sample recovered in the top 2% of windows in the amplified sample (**Supplementary Fig. 4c**). At the same time, only 5% of the high-scoring windows of the standard sample were not included in the top 10% of the amplified sample, indicating that only a small fraction of genomic regions were amplified poorly.

Identification of differentially enriched *cis*-regulatory regions (DERs) comparing cell types. We have calculated coefficients of variability (CV; standard deviation and average) between replicates from the same donor and between averages between different donors. For all cell types, the CV was higher in the inter-donor comparison but not greatly so, e.g. for naïve T cells, CV = 32.9% in the intra-donor comparison vs. 35.3% in the inter-donor comparison. So there is more variability between donors, yet there is also considerable variability between replicates. To account for both, the biological and ChIP assay-to-assay variability in determining statistically significant DERs, replicate ChIP assays from the same individuals were not pooled into an aggregate profile. For each sample, sequencing coverage was calculated at genome wide 500 bp windows, and pairwise differential coverage between groups of cell-types was calculated using *MEDIPS* v.1.12.0 (*extend = 120, uniq = F, window_size=500, BSgenome = "BSgenome.Hsapiens.UCSC.hg19", minRowSum=1, p.adj = "bonferroni", diff.method="edgeR"*)[1,2].

Inspection of the resulting MA plots indicated a varying efficiency of immunoprecipitation in the group of T_H1 cell samples compared to the groups of naïve $CD4^+$ T cells and T_H2 cell samples (**Supplementary Fig. 7**). To correct for this effect, we have performed quantile normalization on the genome-wide count tables across the samples prior to calculation of pairwise differential coverage between groups. MA plots (**Fig. 2a,** *left*) as

well as density plots of group-wise mean raw and quantile normalized sequencing coverage (**Supplementary Fig. 7**), indicate improved comparability of genome-wide enrichment distributions after quantile normalization. In addition, several well-known regulatory elements that contrast T_H1 cells from T_H2 cells became significant only after this normalization procedure (**Supplementary Fig. 8a**). Consequently, we performed quantile normalization as a standard prior to identification of DERs, resulting in 51,261 DERs for naïve $CD4^+$ T cells vs. T_H2 cells (23,583 naïve > T_H2 and 27,678 naïve < T_H2), 38,665 DERs for naïve $CD4^+$ T cells vs. T_H1 cells (22,494 naïve > T_H1, and 16,171 naïve < T_H1), and 9,737 DERs for T_H2 cells vs. T_H1 cells (5,847 $T_H2 > T_H1$, and 3,890 $T_H2 < T_H1$) (Bonferroni adjusted *P*-value < 0.05).

To test the influence of potential interdependence between replicate ChIP assays from the same individual, we repeated the comparison between naïve and T_H2 cells by pooling ChIP replicates from the same donor into an aggregate profile. As expected, due to the reduced statistical power caused by a smaller sample size (due to pooling of replicate assays), we detected a slightly lower number of DERs (*n* = 46,912) compared to our original approach (*n* = 51,261). Most importantly, at the same significance level (adjusted *P*-value < 0.05), 44,425 (87%) of DERs identified by our original approach were also detected when pooling ChIP replicates. We have further characterized the location of DERs that were not re-discovered by pooling replicates (*n* = 6,836) and found that 3,922 (57%) are individual 500 bp windows that are part of a group of adjacent significant DERs (*i.e.* consecutive DERS) identified by our approach (data not shown). Consequently, we observed an overall agreement of 94% between the distinct genomic loci identified by our approach compared to the pooled approach. Thus, we conclude that the presented results obtained by separate processing of ChIP replicates are not confounded by a potential interdependence between replicate ChIP assays from the same individuals.

A recently published method ("DiffReps") for analyzing quantitative differences of broad histone marks across groups of ChIP-Seq samples was compared to a number of other available tools and shown to perform comparably on a small set of ChIP-Seq samples[3]. One method included in this comparison is very similar to the *MEDIPS* methodology employed in our study and the authors reported high agreement between their "DiffReps" method and this *MEDIPS*-like approach. As a further comparison, we applied the "DiffReps" tool to analyze our larger data set and then compared the results with those obtained by our

method (*MEDIPS*) using the same statistical cut off (Bonferroni adjusted *P*-value < 0.05). We observed that around 56% of the DERs detected by our method (*MEDIPS*) were also detected by the "DiffReps" method, suggesting a significant overlap between the two methods (*P*-value < 1.67e-179, *hypergeometric* test). Further, a number of likely false positives and false negatives were also observed with the "DiffRep" method compared to the *MEDIPS* method (data not shown).

Classification of DERs into sub-groups. As described above there was a total of 99,663 DERs identified in the 3 cell type comparisons. However many DERs that were identified in different cell-type comparisons had the same genomic location. The total number of DERs at distinct locations was 71,640. Merging consecutive windows (DERs), we identified a total of 35,428 distinct genome regions ranging in size from 500 bp to 29 Kb (**Supplementary Fig. 8b**). Next, we wanted to assign an overall state to each distinct DER based on the combined results in the 3 cell type comparisons. In each comparison, there are 3 possible outcomes (up, down, or non-significant). Our 71,640 DERs showed sixteen different patterns of outcomes, which are shown in **Supplementary Table 3**. We grouped DERs with patterns of outcomes that have similar biological meaning into six merged groups. Specifically, we defined a group of DERs with T_H2 and T_H1 cells gain comparing to naïve cells but no significant difference between T_H2 and T_H1 cells as *shared gain* (9,317 DERs); a group of DERs with T_H2 and T_H1 cells loss comparing to naïve T cells but no significant difference between T_H2 and T_H1 cells as *shared loss* (12,870 DERs); groups with T_H2 cells gain (20,679) as *T_H2 gain*; groups with T_H2 cells loss (10670) as *T_H2 loss*; groups with T_H1 cells gain (8534) as *T_H1 gain*; and groups with T_H1 cells loss (9570) as *T_H1 loss*.

Assignment of target genes to DERs. There is a considerable overlap between this list of genes and those previously identified by Hawkins *et al.*[4] which describes H3K4me1-enriched enhancers in T_H1 and T_H2 cells generated *in vitro*: if we use the Hawkins *et al*. criteria to assign genes to enhancers, our study identifies 80% of the genes also identified by Hawkins *et al*. as linked to T_H1 and T_H2-specific enhancers (significance of the overlap: *P*-value < 2.2e-16, Fisher's exact test). However, our large sample size (>100) and our use of *in vivo*-generated T_H1 and T_H2 cells, we have been able to identify a number of additional

novel candidate genes, enhancers and miRNAs that were not described by Hawkins *et al*. (see **Supplementary Table 6**, and data not shown). We attribute these differences to several factors: Hawkins *et al.* used naïve T cells polarized with T_H1 or T_H2 cytokines *in vitro* for a relatively short time (72 h), whereas T_H2 memory development *in vivo* is likely to be a more sustained and lengthy process. Moreover, the Hawkins *et al.* study used only a single sample of T_H1 and T_H2 cells for their analyses, potentially reducing the statistical power to detect differences in enhancer profiles at the genome-wide level.

Correlation of gene expression with H3K4me2 enrichment patterns at promoters and enhancers.

In **Supplementary Figure 10** (*right*)**,** we show gene expression changes between naïve and T_H2 cells for the transcripts included in the heat map of H3K4me2 profiles (*left*). For all genes that are significantly differentially expressed (adjusted *P*-value ≤ 0.01), we have calculated the $\log_2$ fold change of T_H2 / naïve $CD4^+$ T cells based on the library size normalized count values per group as reported by *DESeq*. In order to avoid division by zero, we have added a pseudo count of five to each group. Transcripts are given in the same order as in the heat map of H3K4me2 profiles. In cases where a transcript (linked to promoter-localized DERs) is associated with a gene that is not differentially expressed, the expression change for that gene is represented as grey.

To further examine the dependency of H3K4me2 enrichment profiles at potential distal *cis*-regulatory regions (beyond promoter-localized DERs) on gene expression, we have examined H3K4me2 profiles in the extended gene loci (TSS +/- 50 kb) for all differentially expressed genes (**Supplementary Fig. 11**). Here, we have divided the +/- 50 kb region around the transcription start sites of differentially expressed genes into consecutive 2 kb windows. For each 2 kb region, the previously calculated H3K4me2 $\log_2$ fold change of the 500 bp window with the most significant *P*-value was selected, where only windows with a non-adjusted *P*-value ≤ 0.05 were considered. In case a 2 kb window does not contain a significant 500 bp window (non-adjusted *P*-value ≤ 0.05), the window is represented as grey.

Detecting enrichment of transcription factor binding motifs at cell-specific enhancers. We examined if there was a significant enrichment in known TF binding motifs in the subsets of DERs defined above. In each DER category, we retrieved the genomic sequence corresponding to the DERs, and used the findMotifsGenome.pl of the HOMER v 4.3 software package[5] with default parameter settings (**Fig. 5a** and **Supplementary Table 9).** The motifs with *P*-values less or equal to 1.00E-0.3 and ratio of target sequences with motif *vs* background sequences with motif >1.1 were defined as significantly enriched (**Supplementary Table 9**).

Detecting enrichment of transcription factor binding sites at cell-specific enhancers. To investigate, if DERs occur at sites that are known to be bound by transcription factors, we have downloaded ChIP-Seq derived binding sites of 161 transcription factors from 91 different human cell types generated by the ENCODE project and provided via the Table Browser at the UCSC Genome Bioinformatics Site (track *Txn Fac ChIP V3,* table *wgEncodeRegTfbsClusteredV3*). Moreover, we downloaded ChIP-Seq derived binding sites of 18 additional transcription factors provided via GEO-NCBI [GSM776559, GSM776558, GSM776557, GSE37589, GSM1208648, GSM1208734, GSM1208702, GSM1208676, GSM1208658, GSM1208799, GSM1208792, GSM1208791, GSM1208776, GSM1208775, GSM1208774, GSM1031240, GSM831021, GSE38567]. Binding sites provided in *hg18* were converted to *hg19* using the *liftOver* tool at the UCSC Genome Bioinformatics Site[6,7].

For each of the 179 transcription factors, we tested if there is an enrichment of their binding sites within any of the predefined groups of DERs compared to random background. To assemble background datasets for each DER group, we sampled 100 sets of random genome wide 500 bp windows, each with the same amount of regions as the number of DERs. To avoid artificially enhanced enrichments due to inclusion of genomic regions in the background that are inaccessible or for which reads cannot be mapped, we restrict the genomic space for background sampling to windows that show sequencing coverage in any of our H3K4me2 samples, and further exclude genomic regions known to harbor signal artifacts in sequencing experiments as provided by the Table Browser of the UCSC Genome Bioinformatics Site (*hg19*, track *Mapability,* table *DAC Blacklist wgEncodeDacMapabilityConsensusExcludable*). Random regions were

generated using the function *shuffleBed* of the *bedtool* suite v.2.16.2 providing the genomic regions to be excluded via the parameter *excl*. Overlaps between the midpoint of DERs or random regions, respectively, and transcription factor binding sites were calculated using the function *IntersectBed* of the *bedtools* suite v.2.16.2 (ref. 8) For each transcription factor, and for each group of DERs, we calculate the mean and standard deviation of overlaps observed for the 100 background sets of genomic regions. We use the mean and standard deviation obtained from random sampling to calculate an odds ratio and to model a normal distribution and calculate *P*-values for the observed overlap between DERs and binding sites compared to random sampling (**Supplementary Table 10**). We report all significance values with an odds ratio >1 and with a *P*-value ≤ 0.01.

SNPs datasets and enrichment calculation. For a given set of disease associated SNPs, we generated a specific control set of SNPs with similar properties. For each SNP in the disease set, we assigned a 'functional code' similar to the approach in Yu *et al.*[9] The code was composed of the functional assessment in the snp138Common table (set to one of the following: 0='nonsense', 100='missense', 200='stop-loss', 300='frameshift', 400='cds-indel', 500='coding-synon', 600='intron', 700='ncRNA', 800='splice-3', 900='splice-5', 1000='untranslated-3', 1100='untranslated-5', 1200='near-gene-3', 1300='near-gene-5', 1400='unknown') plus the minor allele frequency divided into intervals of 5 percentage points (e.g. '0' = 0-5%, '5' = 5-10%, '10' = 10-15% etc.). We then randomly chose 40 control SNPs with the same functional code from the SNPs tested for by the Illumina 1 MDuo genotyping array as specified from the SNP Genotyping Array track for *hg18* of the UCSC genome browser[6,7]. The Illumina platform was chosen as it was most frequently used in the asthma GWAS studies. If less than 40 exactly matching control SNPs were available, we also included control SNPs with up to 10% difference in their allele frequency. For some of the disease datasets other than asthma, there were individual disease-associated SNPs in a functional code category for which there still was not a sufficient number of control SNPs available, and such singletons were excluded from the control set generation.

To calculate SNP enrichment in a given DER category, we determined the fraction of disease-associated SNPs located in the corresponding genomic region, and compared it to the fraction of non-disease

associated SNPs from a background set to calculate enrichments (**Fig. 6a**). To calculate overlapping and intersecting genome coordinates, the BEDTools utilities package[8] was utilized. Significance was assessed using the chi-square test. In addition to the DER analysis in T cell subsets, we also calculated enrichment of disease-associated SNP sets in H3K4me1 peaks of eight different tissues: adipose nuclei, liver, adipose-derived mesenchymal stem cells (ADMSC), kidney, breast myo-epithelial cells, brain, $CD4^+$ T cells and skeletal muscle cells (**Fig. 6a** and **Supplementary Table 11c**). These H3K4me1-peaks were calculated in our previous analysis[10] based on ChIP-Seq data together with input samples obtained from the Human Epigenome Atlas[11,12].

Haploblock-based SNP enrichment analysis. We performed a second analysis to examine the overlap of asthma associated SNPs and enhancers of different cell types based on previously published methods[13] with a few relevant modifications. As a source of SNPs, we used a recent analysis of asthma-associated SNPs across multiple study populations in the European population[14]. We collected SNPs that were shown in **Supplementary Table 1** of Moffatt study[14] as showing genome wide significant association with either childhood onset asthma or later onset asthma (66 SNPs in total). We used HapMap release 27 data (compiled from merged genotype data from phases I+II+III (HapMap rel #27, NCBI B36) downloaded from (ftp://ftp.ncbi.nlm.nih.gov/hapmap/ld_data/2009-04_rel27/) as a source for Linkage Disequilibrium (LD) information between SNPs, as this was the only source for which genome wide linkage information was available in a pre-calculated form (which was needed for the automated calculation of a large set of randomized controls as described below). CEU (Utah residents with ancestry from northern and western Europe) was chosen as the reference population. For 3 lead SNPs, no linkage information was included in Hapmap, and they were excluded from the dataset.

To construct haploblocks, we grouped SNPs together that were in linkage with $r^2 > 0.1$, resulting in 11 haploblocks. For each haploblock, we added all SNPs that were in linkage with $r^2 > 0.8$ for any of the lead SNPs For each asthma-associated haploblock, we calculated a set of 20 randomized controls choosing a set of control-lead SNPs that matched the frequency and properties of the lead-SNPs from meta-analysis. We chose control lead SNPs from the Illumina 1 MDuo genotyping array as specified from the SNP Genotyping Array track for *hg18* of the UCSC genome browser[6,7]. The Illumina platform was chosen as it

was most frequently used in the asthma GWAS studies. As described in the SNP enrichment above, we assigned a 'functional code, composed of the functional assessment in the snp138Common table and the allele frequency. Control haploblocks were randomly drawn by selecting first a matching control SNP for the asthma-SNP lead-SNP with the highest *P*-value, and then examining if the control had matching SNPs in linkage with $r^2 > 0.1$. For asthma-haploblocks with up to 4 SNPS, controls matching all these criteria for all SNPs were found. For larger haploblocks, at least 4 SNPs and more than half of the contained lead-SNPs had to perfectly match.

For a given set of haploblocks and a set of enhancers, we determined the number of haploblocks that overlapped with at least one enhancer. We determined the average number of overlapping haploblocks in the 20 control datasets, and used the binomial distribution to calculate *P*-values for the number of overlapping haploblocks in the meta-analysis to have occurred by chance (**Supplementary Table 11d**).

REFERENCES

1. Chavez, L. et al. Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res* **20**, 1441-1450 (2010).
2. Lienhard, M., Grimm, C., Morkel, M., Herwig, R. & Chavez, L. *MEDIPS*: genome wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* **30**, 284-286 (2014).
3. Shen L, Shao NY, Liu X, Maze I, Feng J, Nestler EJ. diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS One*, **8** e65598 (2013).
4. Hawkins, R.D. et al. Global chromatin state analysis reveals lineage-specific enhancers during the initiation of human T helper 1 and T helper 2 cell polarization. *Immunity* **38**, 1271-1284 (2013).
5. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).
6. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-496 (2004).

7. Karolchik, D., Hinrichs, A.S. & Kent, W.J. The UCSC Genome Browser. *Curr Protoc Bioinformatics* **Chapter 1**, Unit1 4 (2009).

8. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

9. Yu, W. et al. GWAS Integrator: a bioinformatics tool to explore human genetic associations reported in published genome-wide association studies. *Eur J Hum Genet* **19**, 1095-1099 (2011).

10. Gerasimova, A. et al. Predicting cell types and genetic variations contributing to disease by combining GWAS and epigenetic data. *PLoS One* **8**, e54359 (2013).

11. Bernstein, B.E. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-1048 (2010).

12. Milosavljevic, A. Emerging patterns of epigenomic variation. *Trends Genet.* **27**, 242-250 (2011).

13. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res* **22**, 1748-1759 (2012).

14. Moffatt, M.F. et al. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med* **363**, 1211-1221 (2010).