Online Supplemental Material

Differential gene expression analysis

Each cell type was analysed as described below.

B cells, NK cells and CD8$^+$ T cells Counts were first transformed to log2-counts per million (CPM) and the correlation between samples from the same subject was then estimated using limma's duplicateCorrelation function blocking on subject. Samples weights were calculated using limma's arrayWeights function. Differential expression between groups was then assessed using linear models and robust empirical Bayes moderated t-statistics with a trended prior variance (limma-trend pipeline).

CD4$^+$ T cells To determine the correlation between samples from the same subject, an iterative approach was used. Limma's voomWithQualityWeights function was first applied to transform the data to log2-CPM and calculate sample and observation level weights. This was used to estimate the correlation between samples from the same subject using the duplicateCorrelation function. voomWithQualityWeights was then applied again to the data, incorporating the correlation estimate. Differential expression between groups was then assessed using linear models and robust empirical Bayes moderated t-statistics (limma-voom pipeline).

In each analysis, the linear models incorporated an adjustment for sample sequencing batch and subject age to increase precision, as well as the correlation estimate from samples from the same subject. The Benjamini and Hochberg method was used to control the false discovery rate (FDR) below 5%.

Pathway analyses were conducted on the Gene Ontology (GO) and KEGG databases using limma's goana and kegga functions respectively. Analyses on the Molecular Signatures Database were carried out using limma's fry gene set test.

Differential gene expression analysis of publicly available data

Longitudinal RNA-seq data of CD4$^+$ T cells from seven children who developed beta-cell autoimmunity at a young age and matched control subjects was downloaded from the European Genome-Phenome Archive (accession number EGAS00001004071), and differential expression analysis was conducted using the limma v3.44.3 and edgeR v3.30.3. Expression-based gene filtering was first performed such that genes needed to achieve a CPM greater than 2 in at least one sample. The TMM normalization method was then applied to normalize sample composition differences. For each cell type, the data were transformed to log2-CPM and the correlation between samples from the same subject was estimated using limma's duplicateCorrelation function. Sample weights were also calculated using the arrayWeights function. The limma-trend pipeline, as described previously, was then applied to identify differentially expressed genes between the groups. The FDR was controlled below 5% using the Benjamini and Hochberg method.

ATAC-seq data pre-processing

Libraries from technical replicates were first concatenated. Reads were trimmed with trim_galore (v0.4.5) and analysed for quality with fastqc (v0.11.8). ATAC-seq reads were then aligned to the human genome assembly (hg38) using Bowtie2 v2.2.5 bowtie2 --very-sensitive -X 1000). For each sample, mitochondrial reads were removed from the aligned BAM file using the removeChrom python script developed by Harvard Informatics (https://github.com/harvardinformatics/ATAC-seq), and only properly paired reads were used for downstream analysis. Library complexities were then calculated using the estimateLibComplexity function in the ATACseqQC package, and a stochastic subsampling process then performed in order to standardize all samples to equivalent molecular complexity using samtools view. PCR duplicates were then removed using Picard MarkDuplicates (http://broadinstitute.github.io/picard/). Read mates were fixed using samtools fixmate and fixed reads were then shifted 9 bp to compensate for Tn5 transposase adapter insertion. Peak calling was then performed with MCAS2 (v 2.1.0). Peaks in blacklisted genomic regions as defined by ENCODE for hg38 as well as those at unplaced chromosome contigs were removed.

Annotation of ATAC peaks Peaks were annotated as 5’ UTR, 3’ UTR, promoter-transcription start site (TSS), exonic, intronic, TTS, non-coding or intergenic using the Homer suite annotatePeaks.pl function and the default setting. Chromatin state(s) of the differentially accessible (DA) peaks were annotated using the ChIP-seq-defined ChromHMM states from the Roadmap Epigenomics Project, following the method in Corces et al., 2018 (Corces et al.The chromatin accessibility landscape of primary human cancers. Science 2018;362). In brief, 15 state models were downloaded from the chromatin state learning site for ‘Primary T helper naïve cells from peripheral blood’ (E038) (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html). We then identified the regions of each ChromHMM state that were overlapped by a peak. To determine the significance of these overlaps for each ChromHMM state, we compared the total length of the peaks covered by the given ChromHMM state to the expected background determined by the total length of the universe of ATAC-seq peaks covered by the ChromHMM state, using a binomial test in R. H3K27ac ChIP-seq data from naïve and activated (5 and 24 hour of activation with anti-CD3/28 beads) human CD4$^+$ T cells was downloaded from Gene Expression Omnibus (GEO) (Accession number GSE116698). The GWAS SNP set used for analysis was derived from the NHGRI GWAS Catalog, downloaded September 2019. The promoter-capture Hi-C (pHiC) data used to define promoter-enhancer interactions in primary CD4$^+$ T cells was derived from Javierre et al., 2016 (Javierre et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* 2016;167:1369-1384).