Methods

Transcription factor binding site (TFBS) prediction.

The package CREAD was employed to identify TFBS in proximal promoters of all known human genes. Known genes, their genomic location and 28 species conservation were obtained from UCSC Genome Browser. For each protein-coding gene, the proximal promoter was defined as 1000 bases upstream and 300 bases downstream of the transcription start site. Position weight matrices (PWMs) were obtained from TRANSFAC (v8.3) and JASPAR (v3.0) and each matrix was mapped to a transcription factor. Only PWMs for which we were able to identify a transcription factor with a NCBI Gene ID were accepted. We employed the program STORM in CREAD for matching PWMs to the promoter sequences with p < 10^{-4}. Next, multiSTORM was employed to all other species aligned to the human genome with the same criteria used in STORM. Only TFBS identified in at least 4 species were accepted. Finally, the program site-cons was employed to select for sites significantly (p<0.05) more conserved compared to the 100 flanking bases. This strategy was applied to all transcripts in the human genome. TFBS enrichment for a set of promoters was calculated by generating 1000 random backgrounds with the same number of promoters and base composition. The background sets were used to generate a distribution for each PWM and the score for each PWM in the foreground promoters was calculated as a z-score (number of standard deviations from the mean).

Compilation of gene lists

A list of all human TFs was compiled by curating the collection of TFs previously compiled by Roach et al. and by adding other known TFs based on literature, annotations and domains (Supplementary Table S4A). All human enzymes were obtained from Gene Ontology by accepting all genes assigned terms from the enzyme nomenclature database (EC) that are children of the term "enzyme activity" (GO:0003824) in the GO tree structure (Supplementary Table S4B). All human kinases were obtained from the human Kinome collection (Supplementary Table S4C). Plasma membrane proteins were compiled by accepting all genes annotated with Gene Ontology term "integral to plasma membrane" (GO:0005887) or any of its child terms in the tree structure (Supplementary Table S4D). Gene Ontology version 5778 was used for all lists.

Plasmids and antibodies. Human ZBTB25 gene (GenBank accession number: BC035804) was obtained from Open Biosystems and was subcloned into pCMV-GFP-N (ZBTB25-GFP; pCMV-GFP-N vector was generated by modifying ClonTech pCMV-Myc vector with EGFP epitope and MCS region modifications).
NF-ATc2 luciferase reporter contained 3 copies of the NF-AT binding site in IL-2 and was obtained from Emmel et al. 7. The AP-1 –luciferase reporter contained 4 copies of the AP-1 binding site and was obtained from R. N. Niles 8. The RNAi Consortium (TRC) lentiviral shRNA plasmids against human ZBTB25 were from Broad Institute of MIT and Harvard (Cambridge, MA). The antibodies used were: anti human CD3 antibody (Mass. General Hospital, MA); anti CD28 antibody (Caltag Laboratories, Carlsbad, CA)

**Electroporation and Reporter Assays**

Jurkat cells were maintained at 37°C and 5 % CO₂ in Iscove’s Modified Dulbecco’s Medium (IMDM) supplemented with 10 % fetal calf serum, 5×10⁻⁴M beta-mercaptoethanol, and 15µg/ml gentamycin sulfate. Electroporation (250V, 950 uF, Gene Pulser XcellTM electroporation system, Bio-Rad, Hercules, CA) was employed for transfection of Jurkat E-6 cells. One day prior to electroporation, cells were diluted into fresh growth medium (generally about 5×10⁵ cells/ml). 5×10⁶ Jurkat cells were used per electroporation. Reporter luciferase activity was examined by a standard protocol as supplied by the manufacturer (Promega, Madison, WI).

**Reverse transcription and real–time quantitative PCR**

Total RNA from primary cells was isolated from human peripheral blood (CD4⁺, CD8⁺, CD19⁺, CD14⁺, CD4⁺CD25⁺ cells) purchased from Miltenyi Biotech. These cells were pooled from several individuals. In addition, we obtained CD4⁺ T cells from 2 individual donors. RT–PCR was used for validation and genes identified here as highly enriched in specific cell types were expected to have minimal variation between individuals. The RNA extraction from cells was performed using RNeasy kit (Qiagen) according to the manufacturer’s instructions. One µg of total RNA was subjected to reverse–transcription using an iScript™ cDNA synthesis kit (Bio–Rad). Real time quantitative PCR was performed in a Bio–Rad iCycler thermal cycler equipped of an iQ5 optical module using the iQTMSYBR®Green super mix (Bio–Rad). Briefly, 100 ng of reverse transcribed cDNA were used for each PCR with 250 nM forward and reverse primers. The thermal cycling conditions were 4 min at 95°C, followed by 40 cycles at 94°C for 15 s, 59°C for 1 minute. Amplicons were visualized on a 2% agarose gel to confirm correct band sizes. All data was normalized to GAPDH expression. Primers:

GATGAATGTGGCATTAGTGGACT-3', GPR18-r1: 5'–GCTTGGGTAAAACACTGTGAGAG-3'; EGR1-f3: 5'–CACGCCGAACACTGACATTTT-3', EGR1-r3: 5'–GAGGGGTTAGCGAAGGCTG-3'; beta-actin-f: 5'–GCGGGAAATCGTGCGTGACATT-3', beta-actin-r: 5'–GATGGAGTTGAAGGTAGTTTCGTG-3'; GAPDH (in ZBTB25 quantification)-f: 5'–GAAGGTGAAGGTCGGAGTC-3', and GAPDH -r: 5'–GAAGATGGTGATGGGATTTC-3'. GAPDH (in BCL11B/TSPAN7 quantification), forward primer 5'–TCATCTCTGCCCCCTCTGCT-3', reverse primer 5'–CGACGCCTGCTTCACCACCT-3'. BCL11B, forward primer 5'–TGGGTGCCTGCTATGACAAG-3', reverse primer 5'–GCACGCAGAGGTGAAGTGAT–3'. TSPAN7, forward primer 5'–ACCAAAACCTGTGATAACCTGTCT-3', reverse primer 5'–AGGGAGATATAGGTGCCCAGA-3'.

Knockdown of ZBTB25

The production and infection of lentivirus were performed according to the standard protocols from Broad Institute (the protocols are available from http://www.broadinstitute.org/rnai/trc/lib). Selection with puromycin (2 µg/ml) began one day after lentivirus infection of Jurkat E-6 cells. The knockdown efficiency of endogenous ZBTB25 of each stable cells was examined by real–time RT–PCR.

References
Figure S1. Schematic illustration of the enrichment score calculation. Raw microarray data was normalized using GCRMA algorithm within R & Bioconductor. To compute the enrichment score the module Limma was employed. Limma uses linear models and Bayes methods to assess differential expression. In this example, the enrichment score is computed for a group of replicates labeled G1. Limma is employed pairwise to compare expression levels in G1 to expression levels in each of the other groups in the panel (in this example 10 groups), generating 10 linear model coefficients for each probe (orange) and 10 associated p-values (green). P-values are adjusted using the Bonferroni correction. The enrichment score is calculated for each probe (row) and is defined as the sum of all linear model coefficients with p<0.05. Thus, probes highly expressed in G1 only will result in a very high enrichment scores, due to the sum of 10 large linear model coefficients.
Figure S2. Transcription factors highly enriched (>97.5 percentile) in T cell developments for which the range (highest - lowest enrichment score) is above the 85th percentile.
Figure S3. Gene enrichment profiles for all genes enriched above the 98th percentile in at least one developmental stage. The range (highest score – lowest score) of all genes was calculated and those with a range larger than the median range are shown here clustered using k-means to 40 groups. The first column shows individual gene patterns and the second column shows the mean enrichment score in each development stage. Some patterns of interests with at least 10 probes were further tested for enrichment of transcription factor binding sites shown in a table in the 3rd column. The top 20 scoring position weight matrices are shown in each table.

Figure S4. Transcription factors scoring above the 98th percentile in at least one of the differentiated cell types. Transcription factors were clustered using kmeans to 20 groups. The left columns shows individual gene patterns and the right column the mean score in each cell type.
Figure S3: Gene expression profiles of individual genes in each cluster.

- Cluster 37:
  - CD34+ CD38+ CD1A-
  - CD4+ CD8+ CD3+
  - CD4+ SPCD8+ SP

- Cluster 38:
  - CD34+ CD38+ CD1A+
  - CD4+ CD8+ CD3+
  - CD4+ SPCD8+ SP

- Cluster 39:
  - CD34+ CD38+ CD1A-
  - CD4+ CD8+ CD3+
  - CD4+ SPCD8+ SP

- Cluster 40:
  - CD34+ CD38+ CD1A+
  - CD4+ CD8+ CD3+

**Matrix Genes Z-score**:

- FOXI1 1.891
- ETV4 1.891
- PAX4 2.938
- ELF2 2.858
- PPARG 2.562

**Binding site enrichment**:

- ETV4; ETV7; ETV6; IRF3; IRF2; IRF1; ELF2; ELK1; FLI1; SPIB
Figure S4
Figure S5. (A) Enrichment profiles of transcription factors across T cell differentiation into Th1, Th2 and Treg. Genes with similar profiles were clustered together. (B) Principal component analysis for T cells using transcription factors. Components 1 and 3 are shown and cells are color coded based on their location in the body.

Figure S6. (A) Cell/tissue enrichment profiles for BCL11B. (B) Quantitative PCR mRNA expression levels for BCL11B in primary T cells, primary B cells and primary monocytes. Expression levels were normalized to monocytes, the lowest expressing cells in the panel. (C) Visualization of PCR products on 2% agarose gel.
Figure S7. A heatmap representation all POZ-BTB domain containing genes present in the microarray.
Figure S8. (A) Knockdown efficiency of ZBTB25 in ZBTB25 knockdown Jurkat E-6 cells as determined by real-time RT-PCR. Error bars represent ± Standard Deviation. (B) Effect of ZBTB25 depletion on TCR stimulated AP-1 signaling. 5x10^6 indicated ZBTB25 knockdown Jurkat E-6 stable cells were electroporated with 4 µg of AP-1-luc reporter and 1 ng of renilla-luc reporter. After 18 hours of electroporation, cells from each sample were dispensed into 2 equal aliquots with 1 ml complete IMDM media with or without PMA (50 ng/ml) plus anti-CD28 antibody (1 µg/ml). After another 7 hours of incubation, cells were harvested and examined for luciferase activity. The experiment was done in triplicates.
Figure S9. (A) Quantitative PCR results of TSPAN7 RNA levels in various cell lines and primary cells. Expression levels were normalized to Raji cells, the lowest expressing cells in the panel. (B) Visualization of PCR products on 2% agarose gel.