Genome-Wide CRISPR Screens Illuminate Cross-Talk between T Helper Cell Activation and Differentiation

Amelia Robinson[1], William Harris[1], Ryan Parker[2], Jack Jenkins[3], and Samuel Adams*

[1] Department of Computer Science, University of Oxford, UK

[2] Department of Civil Engineering, University of Manchester, UK

[3] Department of Mechanical Engineering, University of Sydney, Australia

keywords: CD4 T helper cell, mouse, Cas9, CRISPR, pooled screen, library, retrovirus, knock-out, overexpression, ATAC-seq, ChIP-seq, time-course, overexpression

Abstract: T helper type 2 (Th2) cells are important regulators of mammalian adaptive immunity and have relevance for infection, auto-immunity and tumour immunology. Using a newly developed, genome-wide retroviral CRISPR knock-out (KO) library, combined with RNA-seq, ATAC-seq and ChIP-seq, we have dissected the regulatory circuitry governing activation/proliferation and differentiation of these cells. Our experiments distinguish cell activation *versus* differentiation in a quantitative framework. We demonstrate that these two processes are tightly coupled and are jointly controlled by many transcription factors, metabolic genes and cytokine/receptor pairs. There are only a small number of genes regulating differentiation without any role in activation. By combining biochemical and genetic data, we provide an atlas for Th2 differentiation, validating known regulators and identifying novel players, such as *Pparg* and *Bhlhe40*, that are part of the core regulatory network governing Th2 helper cell fates.

Introduction

CD4+ T helper (Th) cells are a central part of the adaptive immune system and play a key role in infections, autoimmunity and tumour repression. During the immune response, Th cells become activated and differentiate from a naive state into different effector subtypes, including T helper type 1 cells (Th1), Th2, Th17 and regulatory T cells (Treg). Different subtypes have distinct functions and molecular characteristics[1]. Th2 cells are primarily responsible for eliminating helminths and other parasites and are strongly associated with allergies. Thus, a better understanding of Th2 cell development is key to combating a range of clinical conditions[2] .

Th2 differentiation is characterized by the production of the cytokines *Il4*, *Il5* and *Il13*. *In vitro*, *Il4* is crucial for the activation of the signalling transducer *Stat6*[3–5], which in turn induces the Th2 master regulator *Gata3*[6–9]. *Gata3* activates *Il4*, forming a positive feedback loop. Th1 cells possess an equivalent mechanism for their defining transcription factor (TF), *Tbx21*, which represses *Gata3*. *Gata3* is able to inhibit *Ifng*, the main cytokine driving Th1 differentiation. Thus, the balance of the two TFs *Tbx21* and *Gata3* defines the Th1-Th2 axis[10]. There are however many genes affecting this balance, and alternative Th fates are frequently affected by overlapping sets of regulatory genes. All T cell fates have in common the requirement of activation via the T cell receptor and a co-stimulatory molecule, for example CD28. Additional signalling via cytokines then determines which T cell fate is adopted. Therefore, a delineation of activation versus differentiation is critical for our understanding of Th subtype development. Despite the importance of different T helper subtypes, so far only the Th17 subtype has been examined systematically [11,12]. Here we dissect Th2 differentiation with a special emphasis on differentiation versus activation signals.

A major challenge in performing genetic studies in primary mouse T cells is the lack of efficient genetic perturbation tools. To date, only a small-scale RNA interference screen has been performed *in vivo* on mouse T cells[13] . However, recently-developed CRISPR technology has the advantages of higher specificity and greater flexibility, allowing knock-out[14,15], repression[16–18] and activation[19–21]. Currently all existing CRISPR libraries are lentiviral-based[22,23] and therefore unable to infect murine Th cells[24]. To overcome this limitation, we created a genome-wide retroviral CRISPR sgRNA library. By using this library on T cells from mice constitutively expressing *Cas9,* we obtained high knock-out efficiency. In addition, we established an arrayed CRISPR screening protocol that is scalable and cost-efficient.

After library transduction, we screened for and characterized genes strongly affecting Th2 differentiation and activation, with *Il4*, *Il13*, *Gata3*, *Irf4* and *Xbp1* as our primary screen read-outs. *Il4*, *Il13*, *Gata3* are at the core of Th2 differentiation[10] while *Irf4* and *Xbp1* have been suggested to have supporting roles in keeping the chromatin accessible and in overcoming the stress response associated with rapid protein synthesis during T cell activation[25–27]. *Gata3* is involved in both activation and differentiation, as mice deficient in *Gata3* are unable to generate single-positive CD4 T cells[28], showing a role beyond regulating the Th1/Th2 differentiation axis. Selected genes discovered by the screen were validated in individual knock-outs and assayed by RNA-seq. To place the discovered genes into the context of Th2 differentiation, we profiled

developing Th2 cells using RNA-seq for gene expression, ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) for chromatin accessibility and ChIP-seq of three key TFs: GATA3, IRF4 and BATF. We further acquired corresponding data from human donors to study the conservation of the regulatory pathway.

The regulatory function of all genes has been assessed by combining state of the art gene regulatory network analysis, comparison of Th2 *versus* Th0, early *versus* late, literature curation and genome-wide screen enrichment. Selected hits were validated in individual KO and overexpression experiments. The function of novel key regulators of Th2 differentiation was further explored by performing additional ChIP-seq experiments. We characterize genes in terms of their impact on activation and differentiation, and provide a more comprehensive, multi-factor model for Th2 cell fate determination For ease of visualization, the integrated dataset is provided online at http://data.teichlab.org.

Results and Discussion

Genome-wide CRISPR/Cas9 screens recapitulate known genes and reveal novel genes in primary mouse T cell differentiation

Figure 1 depicts an overview of our experimental approach. First, a high complexity retroviral sgRNA library was generated (Figure 1b). We activated naive CD4+ T cells purified T cells from mouse spleens with anti-CD3 and anti-CD28 together with IL4 at day 0. At day 1, T cells were transduced with the retroviral libraries and selected with puromycin from day 3. After dead cell removal, the screens were carried out on day 4.

Our screening strategy used two different approaches. For *Il4*, *Il13*, *Xbp1* and *Gata3*, we used T cells from transgenic mice carrying a fluorescent reporter driven by the promoter of the respective genes. In this protocol, cell populations with high or low fluorescence of the gene of interest were enriched with sgRNAs for upstream genes inhibiting or promoting Th2 cell differentiation, respectively. In addition, we carried out screens in which T cells were stained with antibodies for IRF4, XBP1 or GATA3. Most CRISPR screens to date are “drop-out” screens where the sgRNAs from an early time point are compared to those in the final surviving cell population. In contrast, here we identify differentiation-related genes by comparing the sgRNAs in the selected target-high *versus* -low fractions. We will refer to the most highly enriched or depleted genes (defined in more detail below) as “hits”.

In total, we carried out 11 genetic screens and analyzed them using the CRISPR screen hit calling software MAGeCK[29]. As an illustration of the results obtained, Figure 2a shows the hits in a screen using anti-Gata3 antibody staining (*i.e.* sgRNA for specific target genes), ranked by MAGeCK *p*-value, against the fold change (Th2, 0h *versus* 72h, described later) of those sgRNA targeted genes. Reassuringly, *Gata3* is recovered as a top hit in its own screen, as expected. Another top hit is a known signal transducer from the IL4-receptor to *Gata3*, the TF *Stat6*. Previous work has shown *Stat6* to be required for the majority of Th2 response genes in mouse and human[4,5]. This gives us confidence that relevant genes are recovered.

As a further quality control we also compared the screens with an orthogonal hit calling model, BaIOPSE (Bayesian Inference Of Pooled Screen Enrichment, Figure 2b, further described in Supplemental methods). In short, size factors, screen efficiencies, probe efficiencies and gene KO effects are fitted simultaneously. Qualitatively, we find that there is reasonable overlap with MAGeCK[29] and BaIOPSE (BaIOPSE scores in Supplemental Files). In a GO analysis of top hits from all screens (Figure 2c), the categories for calcium and MAPK signaling have the lowest *p*-values. While BaIOPSE allows a more consistent integration of multiple screen replicates than MAGeCK, we use MAGeCK for the remainder of this paper because of its pre-existing community acceptance and because BaIOPSE relies on informative priors.

In all subsequent descriptions of hits, we will refer to the expression of the targeted gene, rather than the level of sgRNA enrichment or depletion. For the sake of brevity, in this paper we will use the nomenclature $X^{\rightarrow y}$, when gene X is in the top 5% of hits in the screen Y, either positively or negatively enriched. If gene X falls within the top 1% of ranked hits, we denote this as $X^{\rightarrow y!}$. A comprehensive list of all genes is included in Supplemental Files and results are summarised in Figure 2d.

Next we identified hits that were consistent between screens (see Methods for details). Some genes appear to have a particularly strong impact on Th2 development as they are hits in multiple screens. Some affect both activation regulators (*Irf4, Xbp1, Gata3*) and differentiation regulators (*Gata3, Il4, Il13*). This includes the known genes *Il27ra*$^{\rightarrow Il4,Il13!}$ and *Lag3*$^{\rightarrow Il4,Il13,Xbp1!}$ but also genes not previously connected to T cells *e.g.* *Trappc12*$^{\rightarrow Il4!,Irf4,Gata3!}$, *Mpv17l2*$^{\rightarrow Il4!,Il13!,Xbp1}$ and the TF *Pou6f1*$^{\rightarrow Il4!,Gata3}$. The cytokine-like gene *Ccdc134*$^{\rightarrow Il4!,Irf4!,Gata3}$ is also a major hit. It has so far received little attention in the literature, but has been linked to arthritis[30] and shown to promote CD8+ T cell effector functions[31]. In short, we have discovered many new genes with a broad effect on Th2 differentiation and activation that deserve further investigation.

Time-course analysis of gene expression and human-mouse comparison highlight metabolic genes

To place our hits into the context of Th2 development, we generated time-course data on mouse and human Th cells during both activation and differentiation (Figure 3a). Mouse and human primary Th cells were isolated from spleen and cord blood respectively and activated with anti-CD3 and CD28. The addition of IL4 to the medium resulted in the maturation into Th2 cells, while absence of IL4 resulted in activated “Th0” cells which proliferate but do not differentiate into a Th subtype. We performed time-course bulk RNA-seq profiling on Th2 and Th0, and ATAC-seq at several time points during Th2 differentiation. The large number of data points allowed us to reconstruct the time-course trajectory of Th2 differentiation by principal component analysis (PCA), using RNA-seq data or ATAC-seq data alone (Figure 3b, c).

When carrying out differential gene expression (DE) analysis between the Th0 and Th2 populations, we split the time-course into an the early/fine-grained (0h-6h) period, and a late/coarse-grained period (0h + 6h-72h), as shown in Figure 3a. The number of DE genes is shown in Figure 3d. Importantly, a sizeable fraction of these (21%) were also identified in at least one of our genetic screens, providing orthogonal evidence for their importance (DE scores are in Supplemental Files).

Evolutionary conservation supports functional relevance, so we carried out an equivalent RNA-seq analysis across ten time points in cultured human primary T cells. Fewer DE genes were identified, possibly because genetic diversity between individuals may obscure some gene expression changes, but more than 1/5th of the human DE genes had direct orthologues in the mouse response (Figure 3d). For the remainder of this paper we will refer to any gene being DE in either human or mouse, at any time, as simply DE.

A total of 216 genes were DE in both mouse and human, either early or late (p=10^{-4}). DE genes that also are top hits in our CRISPR screens are shown in Figure 3d. We note the presence of the well-known cytokines *Ccl17*$^{\rightarrow Il4,Il13,Xbp1}$, *Il13*$^{\rightarrow Il4,Xbp1}$, *Il2*$^{\rightarrow Irf4,Gata3}$ and its receptor *Il2rb*$^{\rightarrow Irf4}$, and the TFs *Gata3*$^{\rightarrow Xbp1!,Gata3!}$, *Tbx21*$^{\rightarrow Il13,Xbp1}$ and *Pparg*$^{\rightarrow Il13,Gata3}$. Several of these are canonical Th2 genes, but several novel hits were also found. Notably, several of these are related to metabolism, such as Pparg, which is thought to signal through mTOR and control fatty acid uptake[32]. Another metabolic gene, related to fatty acid transport[33], with a strong phenotype in our screen, is *Abcd3*$^{\rightarrow Il13,Ir4,Gata3!}$ which has not yet been studied in T cells. The Th1-repressor *Mapkapk3*$^{\rightarrow Il4,Gata3!}$ is also a metabolic gene[34].

Other hits have more diverse functions in T cell development. Hits include the known T cell regulator Stat-inhibitor *Socs1*$^{\rightarrow Irf4,Xbp1}$. The *Il13* hit *Rasgrp1*$^{\rightarrow Il13,Irf4}$ is known to be involved in T cell maturation[35] and links Guanyl to the RAS pathway. Interestingly another Guanylate protein, *Gbp4*$^{\rightarrow Il13}$, is also a novel Il13 hit (but with higher DE p-value). The novel *Il4* candidate regulator *Uhrf1bp1l*$^{\rightarrow Il4}$ has been connected to hypomyelination but could act through the chromatin regulator *Uhrf1* which is required for Treg maturation[36].

In conclusion, a human-mouse comparison of DE genes highlights cytokines and TFs known to be important for both Th2 activation and differentiation, and suggests additional hits in our screens that are likely to be of functional importance, in particular novel genes that act as metabolic regulators (*e.g. Abcd3*).

Analysis of chromatin dynamics reveals different TF binding patterns during activation and differentiation

To gain further insight into the regulation of gene expression we examined chromatin accessibility using ATAC-seq. We performed ATAC-seq of developing Th2 cells across 0, 2, 4

24, 48 and 72 hours time points in both human and mouse (Figure 3a). The chromatin of naive T cells is condensed until activation. It has previously been shown that some TFs, for example *Stat5,* can only access the promoters of its target genes after T cell activation[37]. Th2 differentiation is classically thought to be driven by *Stat6* which in turn upregulates *Gata3*. We examined these dynamics over the time-course of the Th2 response.

The ATAC peaks were first called using MACS2[38]. Overall there is a massive gain of chromatin accessibility from 0 to 2h (Supplemental Figure 1). After this initial opening, the chromatin appears to recondense continuously and this process progresses past 72 hours, as indicated by the reduced total number of ATAC-seq peaks in each time point. We speculate that the regulatory network shifts from a general T cell network to subtype-specific network, and that cell identity becomes less plastic and less responsive to external perturbation over time.

We next compared TF binding predictions between human and mouse. Using FIMO[39] we predicted TF binding sites within ATAC-seq peaks. To reduce the number of potential false positive peaks we concentrated on ATAC peaks that are conserved between mouse and human by calculating the percentage of overlapping peaks between species (10-15%) (Figure 3e) and used these conserved binding sites for the rest of the analysis.

For different TFs, we examined how ATAC peaks, in which the relevant TF motif is found, are changing over time (Figure 3f). As expected for Th2, chromatin accessibility over GATA3 motifs increases strongly with time, correlating with the increase in GATA3 (confirmed by western blot, Supplemental Figure 5, and RNA-seq, Supplemental File). However, the (composite) motif that is most associated with relative peak size increase is BATF::JUN. This is consistent with the suggestion that BATF can act as a pioneer factor to open chromatin[11]. The functional importance of *Baft*/*Jun* is supported by our genetic screens: *Jun*$^{\rightarrow Il13}$, *Fos*$^{\rightarrow Irf4,Xbp1}$ and *Fosl2*$^{\rightarrow Gata3!}$ are all associated with increasing peak height. Since Jun/Fos and Fosl2 all recognize the same AP-1 motif, the exact TF composition at these peaks is likely to depend on their expression level. Notably, *Fosl2* expression is highest at the time points of 1 and 2h in Th0/2 with largely unchanging levels in Th1/2/17/Treg[40]. Overexpression of *Fosl2* has been shown to block IL17A production in Th17 by competing for AP1-sites[11], but overall *Fosl2* expression is low in lymphoid cells[41]. *Fos* and *Jun* are transiently expressed during the first 6 hours. *Jund,* another classical AP-1 factor, displays slowly increasing expression over time. As most AP-1 factors are expressed at low levels, *Batf*, whose expression increases continuously, is the most likely driver behind these peaks.

At the other extreme, some TF motifs are overrepresented in peaks that decrease over time, such as *Hoxd9*$^{\rightarrow Il4}$, *Atf3*$^{\rightarrow Gata3}$, *Atf4*$^{\rightarrow Il4!}$, *Foxj2*$^{\rightarrow Gata3}$, *Dmbx1*$^{\rightarrow Irf4}$, *Foxa2*$^{\rightarrow Il4!}$, *Foxo3*$^{\rightarrow Il4}$ and *Foxc2*$^{\rightarrow Il13!}$. Several of these TFs also have low or decreasing expression levels. We have previously shown that *Atf3*$^{\rightarrow Gata3}$ positively regulates *Ifng*[42] and promotes Th1 differentiation in humans. *Atf4*$^{\rightarrow Il4!}$ has been shown to be important for Th1 function as stress regulator[43] but the impact on *Il4* extends this claim to Th2. *Foxo1*$^{\rightarrow Il13!,Xbp1!}$ is a highly expressed TF but peaks

containing this motif are also decaying. *Foxo1* has recently been shown to inhibit H3K27me3 deposition at pro-memory T-cell genes[44]. *Foxj2* has similar behaviour to *Foxo1* but has not been studied in T cells. However, a link has been made between *Stat6*, *Foxj2* and cholesterol in lung cancer cell lines[45].

Inferred STAT6 binding sites were also compared with previous mouse and human data[4,5], and we found that the vast majority of the previous target genes are also DE in our time course analysis (Supplemental Files). A list of all TFs and the average height of peaks containing their cognate motif is provided in Supplemental Files.

To further characterise the dynamics of the Th2 response, we generated ChIP-seq data at several time points (Figure 3a) for the Th2 master regulator GATA3, as well as BATF and IRF4 which we found to be involved in increasing ATAC peaks. We created a mouse strain with a 3xFLAG-mCherry GATA3 construct (T2A fusion) for this purpose (see Methods for details). The ChIP-seq peaks for *Batf* and *Irf4* have a large overlap as previously reported[11] (Figure 3g) (Jaccard index=0.35). However, we saw no significant overlap of these two factors with GATA3 (Jaccard index=0.028 and 0.032), suggesting that any collaboration between GATA3 and either of BATF/IRF4 is not due to direct protein-protein contact. MEME[46] was applied to the sequences in the GATA3 peaks to find other potential binding partners, and we found enrichment of $YY1^{\rightarrow Il4,Il13,Xbp1,Gata3}$ (p=2.5*10^{-58}) binding motifs. This is consistent with previous reports that *Yy1* is required but not sufficient for Th2 cytokine expression[47]. Indeed, ATAC-seq peaks containing the YY1 motif are stable or decrease slightly (Fig 3d). This finding, together with Yy1 being a strong hit in our screen, reiterates *Yy1* as a key supporter, but not driver, of Th2 differentiation.

Focusing on GATA3 with its 10,203 peaks, a GO-term analysis of its nearby genes yielded “natural killer cell activation” (p=6*10^{-3}), but included few other immune-related terms. This is likely due to the fact that *Gata3* has distinct roles in other cell types[48,49] (a survey has shown that its expression is highest in breast cancer cell lines[41]). Since we performed time-course ChIP-seq we were able to selectively investigate peaks based on their dynamics. We calculated the ChIP peak height ratio at 72h *vs* 24h, and defined the most increasing/decreasing GATA3 peaks as the top/bottom 1000 peaks ordered by ratio. Genes near peaks decreasing over time were not linked to any particular immune-related GO-terms, but a GO term enrichment for genes near increasing peaks revealed “defense to bacterium”/”viral life cycle” (p=5*10^{-3}) as the top term, and included other terms such as “myeloid leukocyte activation” (p=2*10^{-2}). A ranking of peaks and nearby genes, as well as GO terms, are provided in Supplemental Files.

Taken together, the early change in ATAC-seq peak size reflects a rapid increase in accessibility for all TFs, that is further increased for specific Th2 related TFs (e.g. *Batf*/*Jun*, *Gata3*), followed by a progressive loss of peaks as the cells differentiate. Interestingly, our screen hits were found in both categories of TFs. These may be functionally important as either activators or repressors for the specific T helper type.

Motif activity analysis quantifies transcription factors controlling activation *versus* differentiation

To get a broader perspective on how genes affect differentiation and activation we chose to perform a network analysis to compare their downstream effect. For this purpose, we used the ISMARA[50] algorithm, which builds a network by linking TFs to potential target genes based on the presence of the relevant motif in an ATAC-seq peak within the vicinity of the transcription start site (TSS) of that target gene (Figure 4a). In short, a TF has a high MARA activity score if the TF consistently can explain the upregulation of all it's putative downstream genes (or negative score, if it is a suppressor). Interestingly, we found a very high correlation in the networks predicted using ATAC-seq and ChIP-seq data (Figure 4b), suggesting that the algorithm performs well on ATAC-seq input data, allowing us to analyse many TFs besides those with ChIP-seq data. It should be noted that this method struggles to separate TFs with highly similar binding motifs (such as most STAT proteins), and may underestimate the activity of TFs with degenerate motifs. In our interpretation, we associate motifs with the most likely target gene based on known literature, hit score, and expression level in our RNA-seq data.

To obtain an overview of the role of all TFs, we have categorized TFs according to their activity over time within the Th2 differentiation pathway, and whether their activity differs between Th2 and Th0 cells. In other words, two distinct comparisons are made: Firstly t=0h *versus* t=72h within the Th0 compartment, which we term "activation", and secondly Th0 *versus* Th2 cells at t=72h, which we term "differentiation". Figure 5c illustrates this analysis by showing MARA activity scores independently calculated for Th2 and Th0 cells for a number of selected TFs. An example of a TF strongly associated with differentiation (*i.e.* large difference between black and green lines) is *Fos*$^{\rightarrow Irf4,Xbp1}$, while an activation phenotype, reflected in a large difference between t=0h and t=72h, is observed for *E2f1*$^{\rightarrow Irf4}$. The majority of TFs display a behaviour reflecting both activation and differentiation (Figure 4d). We note that for many TFs the activity score does not reflect an increase in expression. Indeed, this is a key strength of the MARA analysis, which calculates a score based on the activity of downstream target genes and can therefore include post-transcriptional regulation or protein-protein interactions affecting TF activity (Figure 4c, e). An example of this is the Th2-defining TF *Gata3*$^{\rightarrow Gata3!,Xbp1!}$, which shows a transient increase in activity, yet its expression levels continually increase with time. *Gata3* is one of the strongest mediators of both activation and differentiation, although its differentiation activity appears to be exerted early. *Stat6*$^{\rightarrow Gata3!}$ is also thought to act early in differentiation, after its activation via the Il4 receptor *Il4ra*$^{\rightarrow Gata3}$. We previously showed that during Th2 differentiation, signals from IL4R are predominantly transduced through STAT6[5]. Consistent with those findings, our data suggests that *Stat6* activity continues to increase throughout differentiation. Interestingly all the STAT proteins map closely together in Figure 4d, affecting primarily differentiation but also activation, possibly all contributing to different extents depending on their expression, phosphorylation status and interactions with other proteins and regulatory elements. *Irf4* is also in this cluster (Figure 4e). *Foxo1*$^{\rightarrow Il13,Xbp1}$ and *Xbp1*$^{\rightarrow Il4}$ are also strongly connected to activation and differentiation, but with *Foxo1* in the opposite direction. Previous work suggested that the

primary role of *Batf* is to open the chromatin together with *Irf4*[11], and this is consistent with our analysis in Figure 3g. Here, it is one of the strongest differentiators, suggesting that chromatin opening is restricted to sites required for differentiation. The roles of other genes are more diffuse. TFs that were identified as hits include *Atf4*$^{\rightarrow Il4!,43}$ and *Yy2*$^{\rightarrow Gata3}$, *Id4*$^{\rightarrow Il13,(Il4),Xbp1}$, *Ebf1*$^{\rightarrow Irf4}$, *Foxp2*$^{\rightarrow Gata3}$, *Yy1*$^{\rightarrow Il4,Il13,Xbp1,Gata3}$ and *Fli1*$^{\rightarrow Il4!}$ affecting both activation and differentiation, but with weaker effects. The identification of a cluster of E2F-proteins as strongly and purely activation-related is consistent with their role in cell cycle control.

The MARA approach allowed us to extract canonical Th2 TFs, such as *Stat6*, *Gata3* and *Batf*, and in addition highlighted newly identified TF-hits (*E2f1*, *Foxo1*) likely to be relevant for Th2 development. Since MARA is not directly dependent on TF-target gene co-variation, the output is complementary to the previous DE approach. It shows again that several TFs are involved in both activation and differentiation, with *Gata3* being a notable example consistent with literature.

Validation of hits by individual CRISPR KO assess gene influence on activation *vs* differentiation

Next we used the results described so far, related this to the existing literature, and chose a panel of 45 genes (40 by screen scores), which were then validated by individual CRISPR KO. Several of the chosen genes have been studied before though not specifically in T cells. Our selection of interesting genes for further characterization is by no means comprehensive, and additional genes can be found by browsing our online resource.

For each KO, cells were grown under Th2 differentiation conditions and RNA-seq carried out on day 4. For each gene a DE list of KO vs non-targeting control was derived and compared to the activation and differentiation axes (Figure 5a). As before, we defined the activation axis as the DE genes from Th0(0h) *vs* Th0(72h) and the differentiation axis as the DE genes from Th0(72h) *vs* Th2(72h) (Figure 3a,b). It should be noted that some genes might not be consistently higher or lower in Th2 *vs* Th0 cells over time. To find if a KO aligns with one of these axes, we determine the projection of the DE genes of the particular KO to the aforementioned axes (see Supplemental methods). Figure 5a shows that all genes tested map away from the neutral centre of the plot (shaded in grey), indicating that the hits contributed to either differentiation or activation, but mostly both, thereby validating the relevance of these genes in Th2 maturation. In this analys is, *Il4* however shows little effect but we believe this is due to IL4 being supplemented in the media. Consistent with the MARA analysis, *Stat6* is primarily driving differentiation. By basing this analysis just on expression, *Gata3* now appears to be primarily driving activation. However, for TFs the MARA analysis accounts for activity and is therefore likely to be more accurate whenever the two analyses diverge. The majority of KO genes affect both differentiation and activation. Examples of genes which have not been studied extensively before in T cells are *Pgk1*$^{\rightarrow Il4}$, *Lrrc40*$^{\rightarrow Gata3}$, *Slc25a3*$^{\rightarrow Irf4}$ and *Ccdc134*$^{\rightarrow Il4!,Irf4!}$.

Validation of key genes by overexpression and ChIP-seq shows the importance of additional transcription factors

To further validate the function and gain mechanistic insights for some of the more novel upstream genes identified in the screen, we performed overexpression by cloning the coding sequence of *Bhlhe40, Pparg, Ccdc134, Gata3, Lrrc40 and Scara3* into the MSCV-gene-IRES-BFP vector. We performed individual transduction and RNA-seq; individual DE genes are listed in the Supplementary File. To summarise the data and allow comparison to the knock-out experiments, we repeated the activation-differentiation analysis (Fig 5b). Since overexpression is approximately the opposite of KO, the sign of the axis in this panel has been reversed for easy comparability. Qualitatively we find agreement between KO and overexpression. The previously unpublished *Lrrc40*$^{\rightarrow Gata3}$ is in particularly good agreement with the KO analysis. We found that it upregulates *Il4* ($p=2*10^{-17}$) and *Il5* ($p=2*10^{-11}$), supporting its role in differentiation. It also regulates *Igfbp4*$^{\rightarrow Gata3}$ ($p=7*10^{-17}$). Overexpression of *Igfbp4* has been shown to inhibit the growth of the thymus(Zhou et al. 2004), which is the same phenotype as the *Gata3* KO mice, suggesting a link *Lrrc40→Igfbp4→Gata3*. The direct function of *Lrrc40* is however unclear. It is present in all cell types and is expressed at the same level across CD4 T cell types. Orthology analysis does not strongly suggest a function but the presence of leucine rich repeats (LRR), shared with the Toll-like receptor, suggest a function in the innate immune system$^{\rightarrow Irf4,51}$. Alternatively it may regulate cell volume[52] or participate in Ca^{2+} regulated channels[53].

Whilst overexpression validated our hits, we wished to gain further insight into the mechanisms by which some of the validated genes function. To this end we added 3XFLAG tag at the 5' end of the two TFs, *Bhlhe40* and *Pparg,* to allow us to find direct targets by ChIP-seq using a FLAG antibody. Using this method we analyzed the genome-wide binding events of *Bhlhe40*, as well as *Pparg*$^{\rightarrow Il13,Gata3}$*.* For both *Bhlhe40* and *Pparg*, we find that the expected motifs are highly enriched (Fig 5c). Under the PPARG peaks we find strong enrichment of several other motifs (Fig 5c, Supplemental Fig 7), including AP1, ETS1, RUNX1, IRF:BATF, GATA3 and STAT5. The identification of motifs for known T-cell program-related genes prompted us to extend our analysis. We compared the 3xFLAG ChIP-seq to our endogenous GATA3/IRF4/BATF ChIP-seq, and all the previously published relevant T cell ChIP-seq datasets (TFs and other DNA binding proteins, see methods). A clustering based on their target genes is shown in Fig 5d. We see that GATA3, PPARG, BARF and IRF4 appear in one cluster, and this holds true for a range of comparison methods (*e.g.,* Pearson correlation, Jaccard index, among other, with and without normalizing for the number of peaks, data not shown). This may explain why *Pparg* is one of the most significant hits in our CRISPR screen: After *Stat6* and *Gata3*, it has the highest *Gata3* upstream screen score of all TFs. *Xbp1* and *Bhlhe40* cluster near *Gata3*. The close relationship of *Bhlhe40* and *Pparg* with canonical Th2 regulators identified these two TFs as new members of the core Th2 regulatory network.

Interestingly STAT6 is separate from the GATA3-cluster. To understand why, we looked closer at how the main Th2 TFs connect together. A graph of TF connectivity is shown in Fig. 6a, where ChIP peaks where associated to their closest genes whenever their distance to the TSS is less than 20kb. Other metrics gave a similar result (data not shown). As in Fig 5d, PPARG-IRF4-BATF-GATA3 form a very tight cluster of TFs that regulate each other and shared target genes. PPARG is thus part of the of the Th2 network. STAT6, which controls the Th2 program after input from IL4, feeds into this program, but also directly regulates the downstream cytokines. BHLHE40 and XBP1 are connected but mainly reside at the downstream. The same overall picture is obtained for a network focusing on some of the most DE activation marker genes (Supplemental Fig. 6).

We next looked closer at the overexpression data, focusing on genes agreeing with the KO data in fold change direction, and having ChIP-seq direct link (DE genes having ChIP-seq peaks, Supplemental Files). *Bhlhe40*$^{\rightarrow Irf4}$ has been shown to suppress inflammation in Th1 through *Il10* which our data weakly supports (direct target, Fig. 6b, DE p=0.015). In another model, ChIP-PCR in iNKT of *Bhlhe40* has shown that it binds near the *Ifng* locus and binding is facilitated by *Tbx21*[54]. We however do not see any peak near *Ifng* in Th2, despite overexpression, nor does DE analysis suggest any effect. Our data suggests alternative mechanisms. For example, *Tnfrsf13b*$^{\rightarrow Il13,Irf4}$ and *Tnfsf13b* are both DE, in opposite directions, with *Tnfrsf13b* a direct target. A peak and weak downregulation (p=0.015) however supports *Bhlhe40* as a negative regulator of inflammation through *Il10*[55].

Pparg$^{\rightarrow Il13,Gata3}$ has recently been shown to be essential for Th2 development[56], which our screen confirms. *Il5* ($p=10^{-14}$), *Il4* ($p=2\ 10^{-6}$) are highly DE direct targets (Fig. 6b). This fits with previous experiments as PPARs has been noted to influence T cell activation and differentiation[57].

To conclude, we have investigated several genes individually by overexpression and mapped their impact on activation and differentiation. Their location agrees qualitatively with the CRISPR KO results. We have studied two upstream TFs by ChIP-seq and found *Pparg* to be particularly overlapping with the central Th2 genes GATA3/BATF/IRF4.

Conclusions

In this study, we demonstrated, for the first time, the applicability of CRISPR to primary murine T cells. By carrying out *in vitro* genome-wide screens we have created a resource of genes important for Th2 helper cell differentiation. We provide optimized protocols for performing additional screens as well as individual KOs. In our hands, these methods have not only worked better than RNA interference, but CRISPR also has advantages in terms of improved targeting and gene disruption instead of down-regulation. In our analysis, we have chosen 5 different read-outs (*Gata3, Il4, Il13, Xbp1* and *Irf4*), which we thought would represent Th2 differentiation and/or activation.

Our unbiased approach of discovering Th2 regulators show that the identified hits belong to many different classes of proteins, including cytokines, TFs, proteins involved in calcium

signalling and metabolic genes. We have performed regulatory network analysis (MARA) to get deeper insight into the upstream genes. We see that many regulatory genes appear to be involved in both differentiation and activation. A summary of this analysis and the later validation is shown in Figure 7. Our analysis of Th2 differentiation has allowed us to re-discover known regulators, such as *Gata3*, *Stat6* and many others. In addition, we have highlighted a number of novel or only poorly studied genes that impact Th2 cell formation. Amongst TFs, examples include *Foxo1*$^{\rightarrow Il13,Xbp1}$, *Bcl11b*$^{\rightarrow Il13!}$, and *Bhlhe40*$^{\rightarrow Irf4}$ (Figure 6b). Non-TFs have also been highlighted, including the cytokine *Ccdc134*$^{\rightarrow Il4!,Irf4!,Gata3}$.

For 46 genes, we also generated specific knock-outs and overexpresson, and validated their impact on differentiation and activation through RNA-seq. *Bhlhe40*$^{\rightarrow Irf4}$ and *Pparg*$^{\rightarrow Il13,Gata3}$ were studied further through ChIP-seq. *Pparg* appear to be particularly central to the Th2 program.

Our results yield a list of genes involved in T helper cell differentiation that deserve further analysis, and an efficient protocol for CRISPR-mediated KO. Both of these are key tools that will enable a more complete understanding of T helper cell biology. By combining our CRISPR KO screen with time-course data, ChIP-seq and overexpression, we have been able to provide a comprehensive map of the most important genes for Th2 differentiation and activation. These genes, along with their expression dynamics and chromatin accessibility, can be browsed on our Supplemental website, http://data.teichlab.org.

Methods

See on-line methods for further information

Author contributions

R.M. helped with cell culturing and early trials of virus production and infections. X.C. and U.U. designed and performed the time-course RNA-seq, ATAC-seq and ChIP-seq. X.C. further helped with western blots, antibody staining and DNA isolation from fixed cells. T.G. performed the time-course data spline DE and PCA. K.M. critically discussed results and helped to write the manuscript. G.D. derived the ES cells from the GATA3-FLAG mouse. K.Y. provided the *Cas9* mice, the sequencing protocol and advised on cloning. R.L. and S.T. developed the experimental design for the RNA-seq, ChIP-seq and ATAC-seq time course for human and mouse cells. R.L. and S.T. initiated and supervised the study. S.T. contributed to the experimental design, data interpretation and writing of the manuscript. J.P. helping with T cell culturing, antibody staining and mouse maintenance. J.H. did the remaining work, in particular, the cloning, the CRISPR screening, the individual CRISPR KO, the RNA-seq, the analysis, and wrote most of the paper.

Acknowledgements

We would like to thank Natalia Kunowska, Xin Xie, Andrew Knights and Sebastian Łukasiak for discussions about 293T culturing, CRISPR and virus production. Bee Ling Ng, Chris Hall and Jennie Graham helped with cell sorting. We thank James Watts, Sean Wright and Amanda Logan for help with the mice. We thank all voluntary blood donors and personnel of Turku University Hospital, Department of Obstetrics and Gynaecology, Maternity Ward (Hospital district of Southwest Finland) for the cord blood collection. We thank Marjo Hakkarainen and Sarita Heinonen for their technical assistance.

Funding

J.H. is funded by the Swedish Research Council, T.G. by the European Union's H2020 research and innovation programme "ENLIGHT-TEN" under the Marie Sklodowska-Curie grant agreement 675395, and S.A.T. by the European Research Council grant ThDEFINE, and X.C. by the FET-OPEN grant MRG-GRAMMAR. R.L. is funded by the Academy of Finland Centre of Excellence in Molecular Systems Immunology and Physiology Research 2012-2017 (AoF grant 250114); the Academy of Finland grants 294337, 292335, and the Sigrid Jusélius Foundation. Wellcome trust core facilities are supported by grant WT206194.

Competing Interests

None declared

Data and materials availability

All the plasmids including the plasmid library are available from Addgene (see Online Methods for accession numbers). Selected parts of the data are also available for online visualization at http://data.teichlab.org. The sequencing data has been deposited at ArrayExpress (E-MTAB-6276, E-MTAB-6285, E-MTAB-6292 and E-MTAB-6300) and the R code used for the analysis is available on Github (https://github.com/mahogny/th2crispr).

Supplemental files

S15. Supplemental Files

SYY. Cell count statistics for each screen and kill rate statistics

SXX: TODO FC and p-value for the OE genes, with ChIP overlap

References

1. Zhu, J., Yamane, H. & Paul, W. E. Differentiation of effector CD4 T cell populations (*). *Annu. Rev. Immunol.* **28,** 445–489 (2010).
2. DuPage, M. & Bluestone, J. A. Harnessing the plasticity of CD4+ T cells to treat immune-mediated disease. *Nat. Rev. Immunol.* **16,** 149–163 (2016).
3. Kaplan, M. H., Schindler, U., Smiley, S. T. & Grusby, M. J. Stat6 is required for mediating responses to IL-4 and for development of Th2 cells. *Immunity* **4,** 313–319 (1996).
4. Chen, Z. *et al.* Identification of novel IL-4/Stat6-regulated genes in T lymphocytes. *J. Immunol.* **171,** 3627–3635 (2003).
5. Elo, L. L. *et al.* Genome-wide profiling of interleukin-4 and STAT6 transcription factor regulation of human Th2 cell programming. *Immunity* **32,** 852–862 (2010).
6. Swain, S. L., Weinberg, A. D., English, M. & Huston, G. IL-4 directs the development of Th2-like helper effectors. *J. Immunol.* **145,** 3796–3806 (1990).
7. Zhang, D. H., Cohn, L., Ray, P., Bottomly, K. & Ray, A. Transcription factor GATA-3 is differentially expressed in murine Th1 and Th2 cells and controls Th2-specific expression of the interleukin-5 gene. *J. Biol. Chem.* **272,** 21597–21603 (1997).
8. Zheng, W. & Flavell, R. A. The transcription factor GATA-3 is necessary and sufficient for Th2 cytokine gene expression in CD4 T cells. *Cell* **89,** 587–596 (1997).
9. Gros, G. L., Ben-Sasson, S. Z., Seder, R., Finkelman, F. D. & Paul, W. E. Generation of interleukin 4 (IL-4)-producing cells in vivo and in vitro: IL-2 and IL-4 are required for in vitro generation of IL-4-producing cells. *The Journal of immunology* **181,** 2943–2951 (2008).
10. Kanhere, A. *et al.* T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements. *Nat. Commun.* **3,** 1268 (2012).
11. Ciofani, M. *et al.* A validated regulatory network for Th17 cell specification. *Cell* **151,** 289–303 (2012).
12. Yosef, N. *et al.* Dynamic regulatory network controlling TH17 cell differentiation. *Nature* **496,** 461–468 (2013).

13. Chen, R. *et al.* In Vivo RNA Interference Screens Identify Regulators of Antiviral CD4+ and CD8+ T Cell Differentiation. *Immunity* **41,** 325–338 (2014).
14. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–826 (2013).
15. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823 (2013).
16. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154,** 442–451 (2013).
17. Kearns, N. A. *et al.* Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat. Methods* **12,** 401–403 (2015).
18. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159,** 647–661 (2014).
19. Maeder, M. L. *et al.* CRISPR RNA-guided activation of endogenous human genes. *Nat. Methods* **10,** 977–979 (2013).
20. Perez-Pinera, P. *et al.* RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat. Methods* **10,** 973–976 (2013).
21. Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33,** 510–517 (2015).
22. Tzelepis, K. *et al.* A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep.* **17,** 1193–1205 (2016).
23. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34,** 184–191 (2016).
24. Baumann, J. G. *et al.* Murine T cells potently restrict human immunodeficiency virus infection. *J. Virol.* **78,** 12537–12547 (2004).
25. Glasmacher, E. *et al.* A genomic regulatory element that directs assembly and function of immune-specific AP-1-IRF complexes. *Science* **338,** 975–980 (2012).

26. Li, P. *et al.* BATF-JUN is critical for IRF4-mediated transcription in T cells. *Nature* **490,** 543–546 (2012).
27. Kemp, K. L. *et al.* The serine-threonine kinase inositol-requiring enzyme 1α (IRE1α) promotes IL-4 production in T helper cells. *J. Biol. Chem.* **288,** 33272–33282 (2013).
28. Pai, S.-Y. *et al.* Critical roles for transcription factor GATA-3 in thymocyte development. *Immunity* **19,** 863–875 (2003).
29. Li, W. *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15,** 554 (2014).
30. Xia, P. *et al.* Amelioration of adjuvant-induced arthritis in CCDC134-overexpressing transgenic mice. *Biochem. Biophys. Res. Commun.* **490,** 111–116 (2017).
31. Huang, J. *et al.* Cytokine-like molecule CCDC134 contributes to $CD8^+$ T-cell effector functions in cancer immunotherapy. *Cancer Res.* **74,** 5734–5745 (2014).
32. Angela, M. *et al.* Fatty acid metabolic reprogramming via mTOR-mediated inductions of PPARγ directs early activation of T cells. *Nat. Commun.* **7,** 13683 (2016).
33. Dean, M., Rzhetsky, A. & Allikmets, R. The human ATP-binding cassette (ABC) transporter superfamily. *Genome Res.* **11,** 1156–1166 (2001).
34. Köther, K. *et al.* MAPKAP kinase 3 suppresses Ifng gene expression and attenuates NK cell cytotoxicity and Th1 CD4 T-cell development upon influenza A virus infection. *FASEB J.* **28,** 4235–4246 (2014).
35. Priatel, J. J., Teh, S.-J., Dower, N. A., Stone, J. C. & Teh, H.-S. RasGRP1 transduces low-grade TCR signals which are critical for T cell development, homeostasis, and differentiation. *Immunity* **17,** 617–627 (2002).
36. Obata, Y. *et al.* The epigenetic regulator Uhrf1 facilitates the proliferation and maturation of colonic regulatory T cells. *Nat. Immunol.* **15,** 571–579 (2014).
37. Rawlings, J. S., Gatzka, M., Thomas, P. G. & Ihle, J. N. Chromatin condensation via the condensin II complex is required for peripheral T-cell quiescence. *EMBO J.* **30,** 263–276

(2011).

38. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
39. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–1018 (2011).
40. Stubbington, M. J. T. *et al.* An atlas of mouse CD4+ T cell transcriptomes. *Biol. Direct* **10,** 14 (2015).
41. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science* **347,** 1260419–1260419 (2015).
42. Filén, S. *et al.* Activating transcription factor 3 is a positive regulator of human IFNG gene expression. *J. Immunol.* **184,** 4990–4999 (2010).
43. Xia, R., Lu, B. & Yang, X. ATF4 reprograms T cell metabolism in response to the environmental stress and is required for Th1 immune responses (IRM9P.459). *The Journal of Immunology* **194,** 130.4–130.4 (2015).
44. Gray, S. M., Amezquita, R. A., Guan, T., Kleinstein, S. H. & Kaech, S. M. Polycomb Repressive Complex 2-Mediated Chromatin Repression Guides Effector CD8(+) T Cell Terminal Differentiation and Loss of Multipotency. *Immunity* **46,** 596–608 (2017).
45. Dubey, R., Chhabra, R. & Saini, N. Small interfering RNA against transcription factor STAT6 leads to increased cholesterol synthesis in lung cancer cell lines. *PLoS One* **6,** e28509 (2011).
46. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** 202–208 (2009).
47. Hwang, S. S. *et al.* Transcription factor YY1 is essential for regulation of the Th2 cytokine locus and for Th2 cell differentiation. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 276–281 (2013).
48. Wei, G. *et al.* Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. *Immunity* **35,** 299–311 (2011).
49. Van de Walle, I. *et al.* GATA3 induces human T-cell commitment by restraining Notch

activity and repressing NK-cell fate. *Nat. Commun.* **7,** 11171 (2016).

50. Balwierz, P. J. *et al.* ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.* **24,** 869–884 (2014).
51. Sun, J., Wang, Z. & Wang, X. Suppression of LRRC19 promotes cutaneous wound healing in pressure ulcers in mice. *Organogenesis* **14,** 13–24 (2018).
52. Kasuya, G. *et al.* Cryo-EM structures of the human volume-regulated anion channel LRRC8. *Nat. Struct. Mol. Biol.* (2018). doi:10.1038/s41594-018-0109-6
53. Yang, C. *et al.* Knockout of the LRRC26 subunit reveals a primary role of LRRC26-containing BK channels in secretory epithelial cells. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E3739–E3747 (2017).
54. Kanda, M. *et al.* Transcriptional regulator Bhlhe40 works as a cofactor of T-bet in the regulation of IFN-γ production in iNKT cells. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E3394–402 (2016).
55. Yu, F. *et al.* The transcription factor Bhlhe40 is a switch of inflammatory versus antiinflammatory Th1 cell fate determination. *J. Exp. Med.* **215,** 1813–1821 (2018).
56. Chen, T. *et al.* PPAR-γ promotes type 2 immune responses in allergy and nematode infection. *Sci Immunol* **2,** (2017).
57. Choi, J.-M. & Bothwell, A. L. M. The nuclear receptor PPARs as important regulators of T-cell functions and autoimmune diseases. *Mol. Cells* **33,** 217–222 (2012).
58. Naviaux, R. K., Costanzi, E., Haas, M. & Verma, I. M. The pCL vector system: rapid production of helper-free, high-titer, recombinant retroviruses. *J. Virol.* **70,** 5701–5705 (1996).
59. Grote, D., Souabni, A., Busslinger, M. & Bouchard, M. Pax 2/8-regulated Gata 3 expression is necessary for morphogenesis and guidance of the nephric duct in the developing kidney. *Development* **133,** 53–61 (2006).
60. Barlow, J. L. *et al.* Innate IL-13-producing nuocytes arise during allergic lung inflammation

and contribute to airways hyperreactivity. *J. Allergy Clin. Immunol.* **129,** 191–8.e1–4 (2012).

61. Hu-Li, J. *et al.* Regulation of expression of IL-4 alleles: analysis using a chimeric GFP/IL-4 gene. *Immunity* **14,** 1–11 (2001).
62. Skarnes, W. C. *et al.* A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474,** 337–342 (2011).
63. Parks, G. D., Duke, G. M. & Palmenberg, A. C. Encephalomyocarditis virus 3C protease: efficient cell-free expression from clones which link viral 5' noncoding sequences to the P3 region. *J. Virol.* **60,** 376–384 (1986).
64. Bürglin, T. R. & Henriksson, J. FACSanadu: Graphical user interface for rapid visualization and quantification of flow cytometry data. *bioRxiv* 201897 (2017). doi:10.1101/201897
65. Quail, M. A. *et al.* SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. *BMC Genomics* **15,** 110 (2014).
66. Heng, T. S. P. *et al.* The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9,** 1091–1094 (2008).
67. Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles* **76,** 1–32 (2017).
68. Carlson, M. GO.db: A set of annotation maps describing the entire Gene Ontology. (2016).
69. Pramanik, J. *et al.* The IRE1a-XBP1 pathway promotes T helper cell differentiation by resolving secretory stress and accelerating proliferation. *bioRxiv* 235010 (2017). doi:10.1101/235010
70. Yssel, H., De Vries, J. E., Koken, M., Van Blitterswijk, W. & Spits, H. Serum-free medium for generation and propagation of functional human cytotoxic and helper T cell clones. *J. Immunol. Methods* **72,** 219–227 (1984).
71. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–1218 (2013).

72. Schmidl, C., Rendeiro, A. F., Sheffield, N. C. & Bock, C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat. Methods* **12,** 963–965 (2015).
73. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
74. 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).
75. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
76. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* (2017). doi:10.1038/nmeth.4197
77. Pimentel, H. J., Bray, N., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-Seq incorporating quantification uncertainty. *bioRxiv* 058164 (2016).
78. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
79. Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* **36,** D102–6 (2008).
80. Schmitges, F. W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.* **26,** 1742–1752 (2016).
81. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9,** e1003118 (2013).
82. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9,** 171–181 (2014).
83. Wu, T. D., Reeder, J., Lawrence, M., Becker, G. & Brauer, M. J. GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. in

Methods in molecular biology (Clifton, N.J.) **1418,** 283–334 (2016).

84. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).
85. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
86. Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31,** 3497–3500 (2003).
87. Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12,** R83 (2011).
88. Liu, X. *et al.* Transcription factor achaete-scute homologue 2 initiates follicular T-helper-cell development. *Nature* **507,** 513–518 (2014).
89. Kuwahara, M. *et al.* Bach2–Batf interactions control Th2-type immune response by regulating the IL-4 amplification loop. *Nat. Commun.* **7,** 12596 (2016).
90. Hertweck, A. *et al.* T-bet Activates Th1 Genes through Mediator and the Super Elongation Complex. *Cell Rep.* **15,** 2756–2770 (2016).
91. Garber, M. *et al.* A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol. Cell* **47,** 810–822 (2012).
92. Samstein, R. M. *et al.* Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* **151,** 153–166 (2012).
93. Humblin, E. *et al.* IRF8-dependent molecular complexes control the Th9 transcriptional program. *Nat. Commun.* **8,** 2085 (2017).
94. Liao, W., Ouyang, W., Zhang, M. Q. & Li, M. O. Genome Wide Mapping of Foxo1 Binding-sites in Murine T Lymphocytes. *Genom Data* **2,** 280–281 (2014).
95. Hayatsu, N. *et al.* Analyses of a Mutant Foxp3 Allele Reveal BATF as a Critical Transcription Factor in the Differentiation and Accumulation of Tissue Regulatory T Cells.

Immunity **47,** 268–283.e9 (2017).

96. Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.* **18,** 956–963 (2011).
97. Kim, H.-J. *et al.* Stable inhibitory activity of regulatory T cells requires the transcription factor Helios. *Science* **350,** 334–339 (2015).
98. Escobar, T. M. *et al.* miR-155 activates cytokine gene expression in Th17 cells by regulating the DNA-binding protein Jarid2 to relieve polycomb-mediated repression. *Immunity* **40,** 865–879 (2014).
99. Bevington, S. L. *et al.* Inducible chromatin priming is associated with the establishment of immunological memory in T cells. *EMBO J.* **35,** 515–535 (2016).
100. Seitan, V. C. *et al.* Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Res.* **23,** 2066–2077 (2013).
101. Riley, J. P. *et al.* PARP-14 binds specific DNA sequences to promote Th2 cell gene expression. *PLoS One* **8,** e83127 (2013).
102. Jain, R. *et al.* Interleukin-23-Induced Transcription Factor Blimp-1 Promotes Pathogenicity of T Helper 17 Cells. *Immunity* **44,** 131–142 (2016).
103. Brown, C. C. *et al.* Retinoic acid is essential for Th1 cell lineage stability and prevents transition to a Th17 cell program. *Immunity* **42,** 499–511 (2015).
104. Kakugawa, K. *et al.* Essential Roles of SATB1 in Specifying T Lymphocyte Subsets. *Cell Rep.* **19,** 1176–1188 (2017).
105. Zhang, S. *et al.* Reversing SKI-SMAD4-mediated suppression is essential for TH17 cell differentiation. *Nature* **551,** 105–109 (2017).
106. De, S. *et al.* Dynamic BRG1 recruitment during T helper differentiation and activation reveals distal regulatory elements. *Mol. Cell. Biol.* **31,** 1512–1527 (2011).
107. Ing-Simmons, E. *et al.* Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. *Genome Res.* **25,** 504–513 (2015).

108. Wei, L. *et al.* Discrete roles of STAT4 and STAT6 transcription factors in tuning epigenetic modifications and transcription during T helper cell differentiation. *Immunity* **32,** 840–851 (2010).

109. Vahedi, G. *et al.* STATs shape the active enhancer landscape of T cell populations. *Cell* **151,** 981–993 (2012).

110. Hirahara, K. *et al.* Asymmetric Action of STAT Transcription Factors Drives Transcriptional Outputs and Cytokine Specificity. *Immunity* **42,** 877–889 (2015).

111. Iwata, S. *et al.* The Transcription Factor T-bet Limits Amplification of Type I IFN Transcriptome and Circuitry in T Helper 1 Cells. *Immunity* **46,** 983–991.e4 (2017).

112. Villarino, A. *et al.* Signal transducer and activator of transcription 5 (STAT5) paralog dose governs T cell effector and regulatory functions. *Elife* **5,** (2016).

113. Nakayamada, S. *et al.* Early Th1 cell differentiation is marked by a Tfh cell-like transition. *Immunity* **35,** 919–931 (2011).

114. Krijthe, J. H. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation. (2015).

115. Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171,** 1573–1588.e28 (2017).

Figures

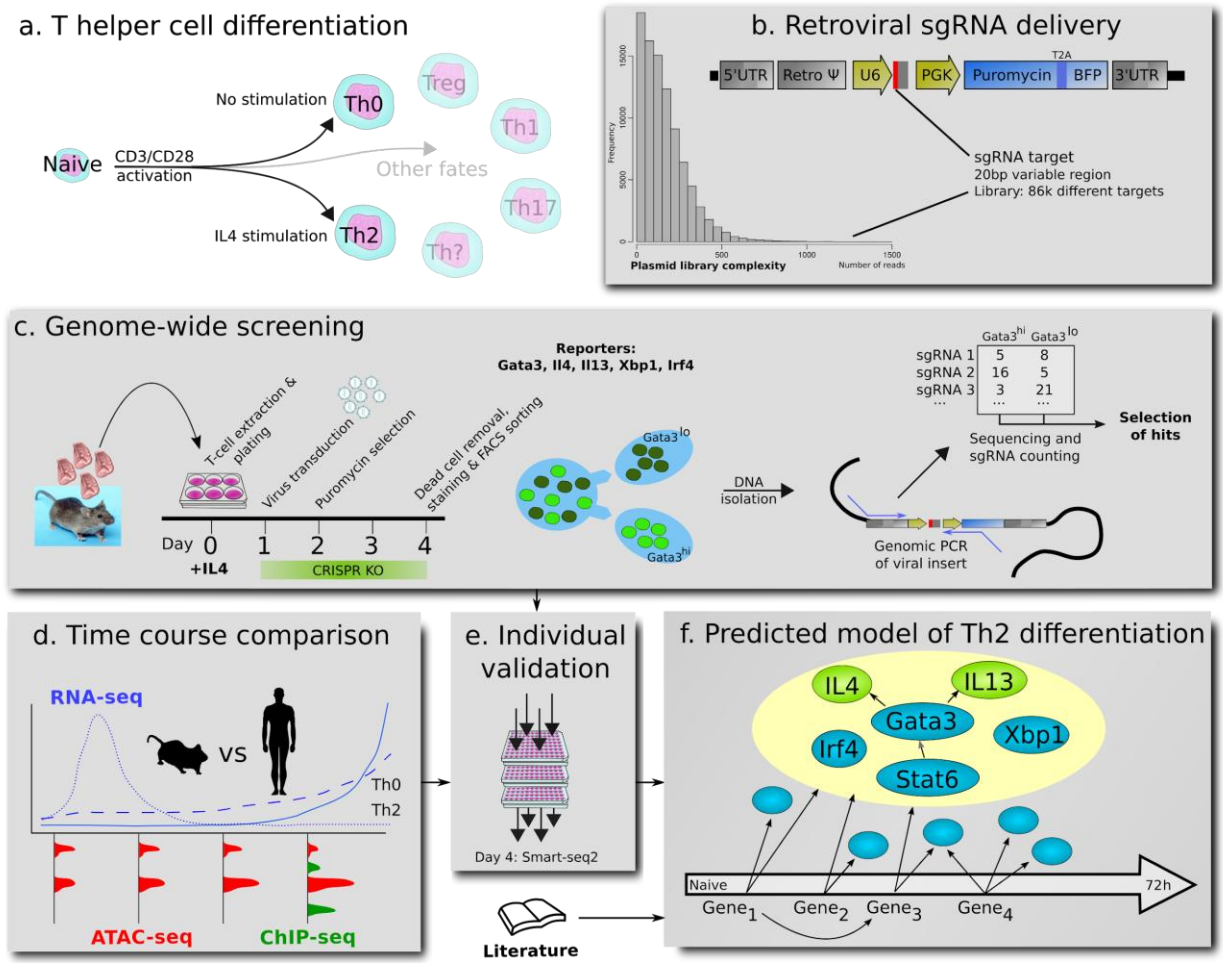

Figure 1: Overview of the experimental KO screening strategy. (a) In our culture system, naive, *ex vivo* T cells are differentiated into Th2 cells by IL4. Potential alternative T cell fates that may be open to genetically perturbed cells are indicated. *In vivo*, T cells develop into different subtypes dependent on stimuli. **(b)** The retrovirus is based on murine stem cell virus (MSCV), encoding one sgRNA per virus, and allows for BFP and puro selection. For the screening we have used a pool of plasmids, encoding over 86 000 sgRNAs, from all of which we produced viruses. **(c)** For genome-wide screens, we pool cells from up to 30 mice. After infection and puromycin selection, the cells are sorted based on fluorescence for the investigated gene. sgRNAs affecting gene expression are identified by genomic PCR. Differential sgRNA expression analysis then allows us to find genes affecting either viability (drop-out screen) or differentiation. **(d)** The top enriched and depleted genes ("hits") were analyzed based on their dynamics measured by RNA-seq, ATAC-seq and ChIP-seq. **(e)** Particularly interesting genes were then further validated by individual KO and RNA-seq. **(f)** By using all this data and a curating the literature we provide a Th2 gene regulatory network.

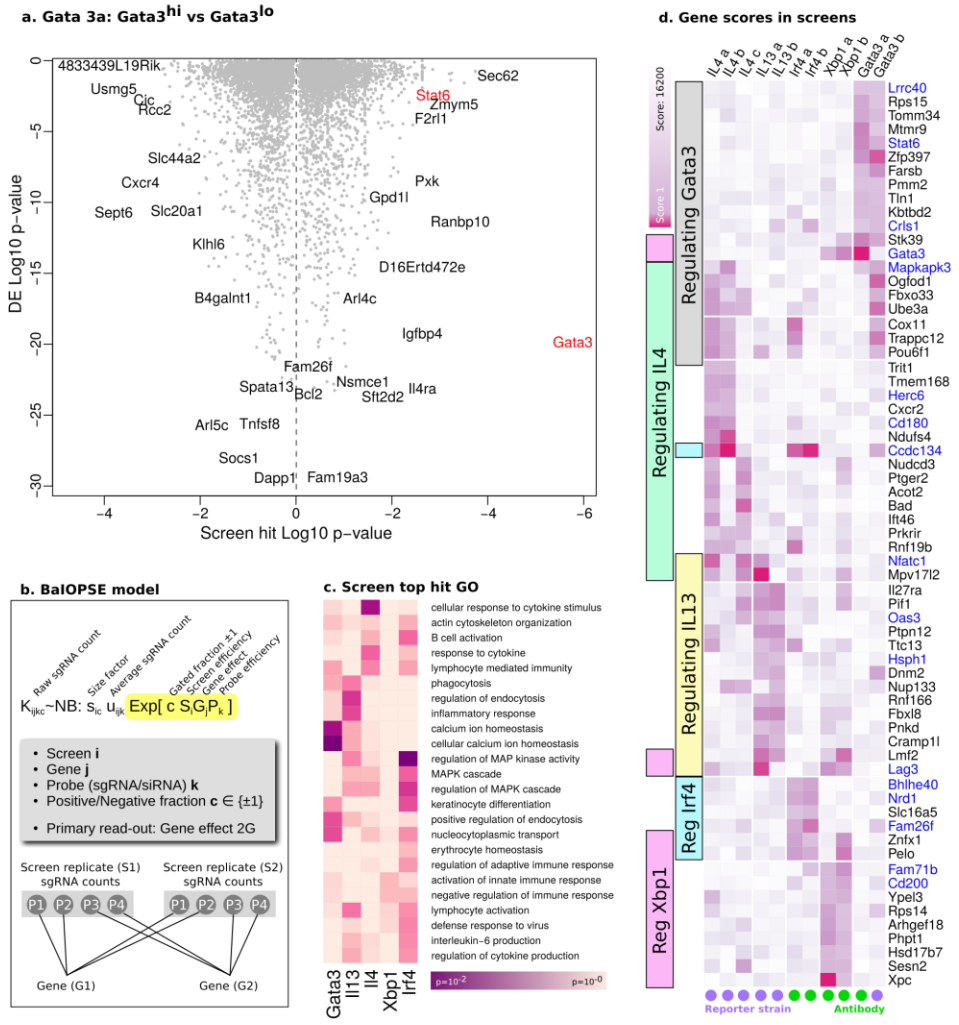

Figure 2: Results from genome-wide Th2 differentiation screen. (a) Hits from screen for *Gata3* expression measured by antibody staining. The x-axis denotes the *p*-value for differential expression obtained by MAGeCK (hits of high relevance toward both sides). The y-axis shows the *p*-value comparing Th2 and Th0 gene expression level (explained later). Highlighted in red are *Gata3* and *Stat6*, since these are known to control *Gata3* expression. **(b)** The alternative BaIOPSE (Bayesian Inference Of Pooled Screen Enrichment) hit calling model. This model is in essence an extended negative binomial differential expression model over sgRNA counts K. Each sgRNA has an efficiency term P, and each screen has an efficiency term S. The interesting read-out is the gene effect 2G. **(c)** GO annotation of top hits for each screen as defined by BaIOPSE. The color represents Log_{10} p-value. **(d)** Summary results of all 11 screens carried out. Genes that were consistent hits in multiple screens are shown (see methods for gene selection). The purple colour shows the Log_{10} combined MAGeCK rank (positive and negative enrichment combined). Screens that relied on antibody staining are marked by a green circle, and those based on fluorescent gene reporters are marked by a purple circle. Genes in blue have been KO:ed individually (see Figure 6).

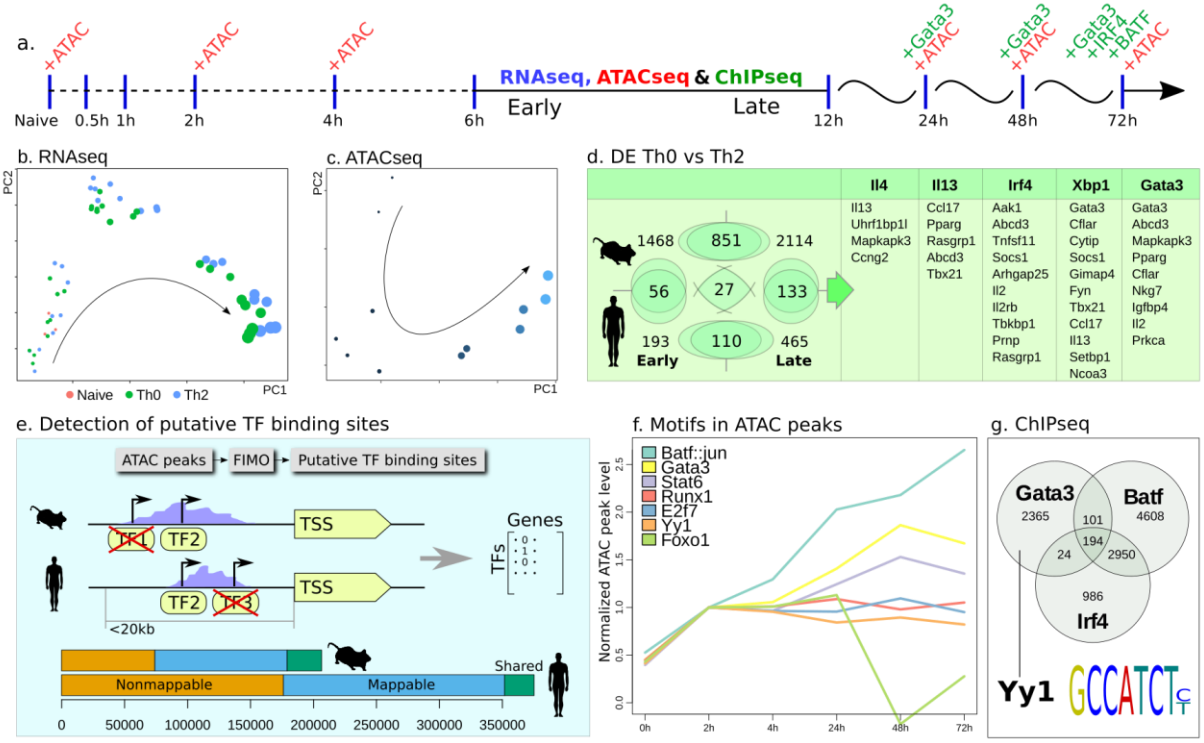

Figure 3. Molecular characterisation and assessment of hits over the time-course of Th2 differentiation (a) The chosen time-points for RNA-seq, ATAC-seq and ChIP-seq. **(b)** PCA projection of bulk RNA-seq and **(c)** ATAC-seq samples. The size of the circle represents time. The naive samples separate in the third principal component not shown. **(d)** Number of differentially expressed genes in the early and late response, in human and mouse ($p=10^{-4}$). DE genes in both human and mouse that are also hits in the genetic screens sorted by rank in their respective screen. **(e)** Workflow for finding conserved putative TF binding sites in human and mouse. The green region represents conserved (overlapping) peaks. The blue region represents peaks in regions with a corresponding sequence in the other species, but without peak conservation. The orange region depicts peaks lying in non-syntenic (unmappable) regions. **(f)** Examples of ATAC-seq peak dynamics associated with different TFs. **(g)** Overlap of peaks in different ChIP-seq experiments at 72 hours. We note the presence of the YY1 motif within the GATA3 peaks.

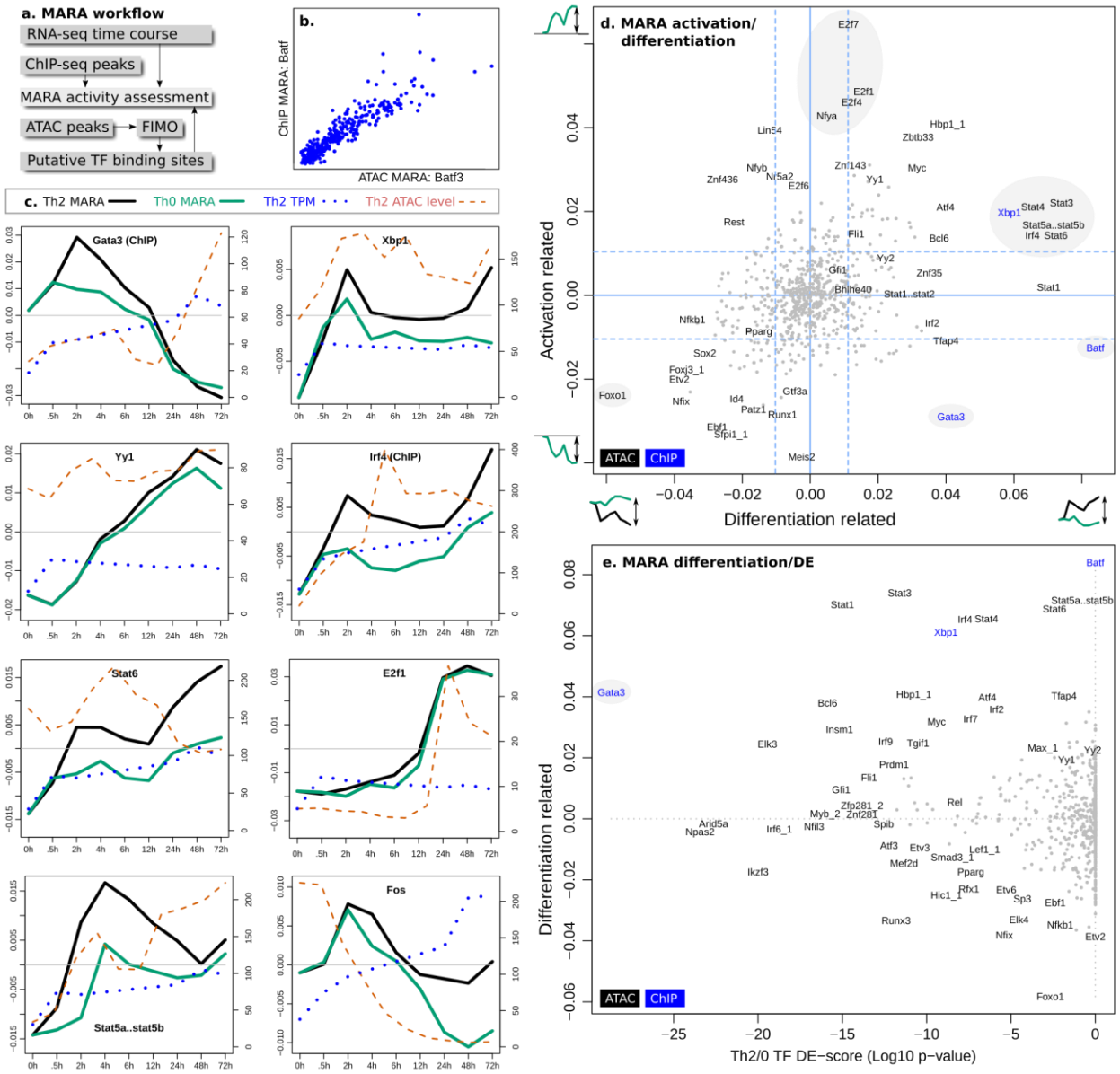

Figure 4: Analysis of transcription factor activity using "Motif Activity Response Analysis" (MARA) (a) Workflow for combining putative binding sites with time-course RNA-seq. **(b)** Comparison of BATF activity predictions for individual genes, by ATAC-seq predicted binding sites and ChIP-seq peaks. **(c)** Dynamics of selected TFs, comparing their expression level, activity in Th2 (black line) and Th0 (green line) and chromatin accessibility. **(d)** MARA activation vs differentiation scores (as defined in text) of all TFs. **(e)** Comparison of differentiation score and DE *p*-value Th2 vs Th0.

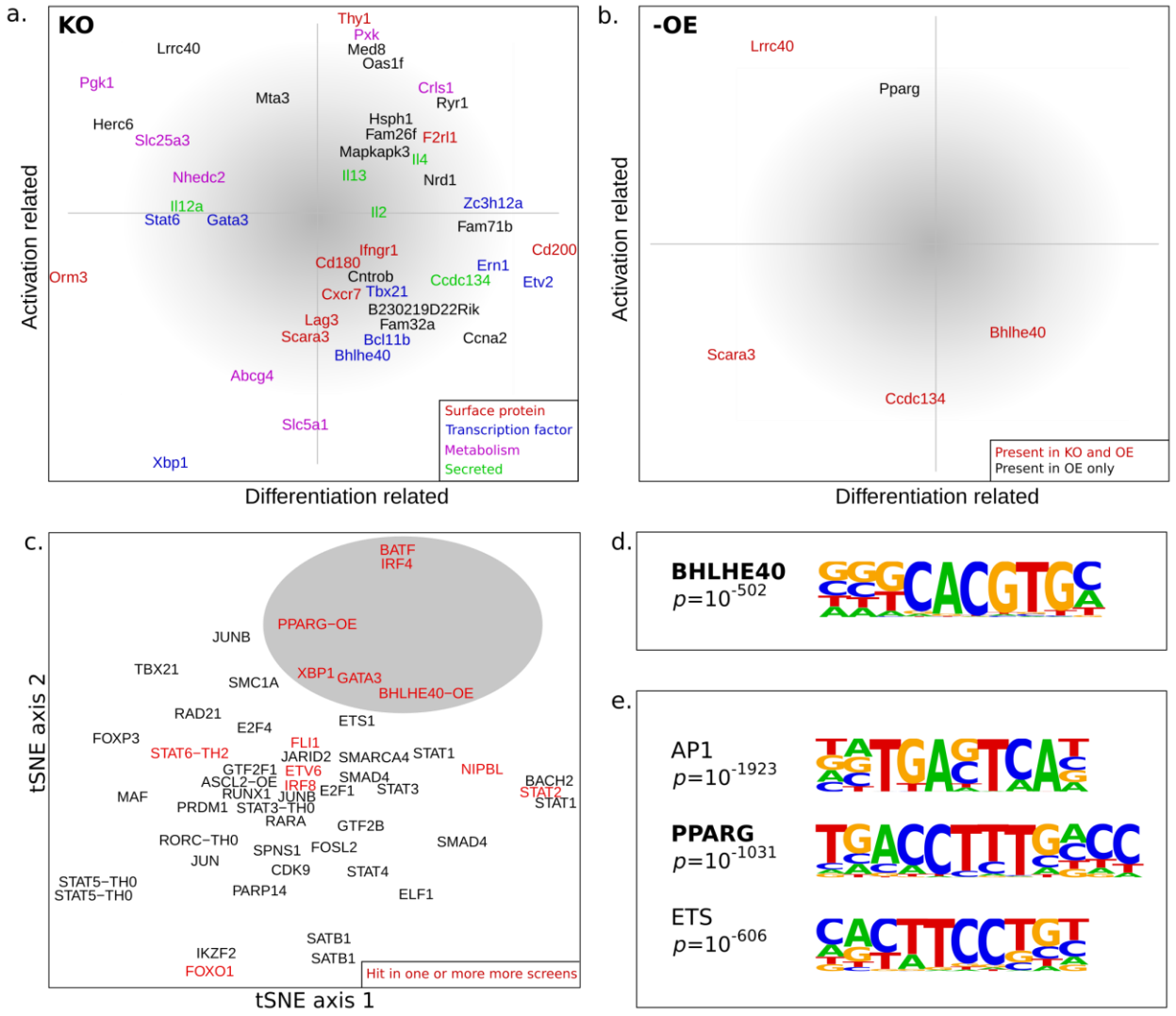

Figure 5: KO and overexpression effect on activation and differentiation (a) The effect on differentiation and activation of gene KOs. DE genes between KO and WT were quantified and projected onto axises for activation and differentiation. These axes are defined as the fold change of DE genes from the RNA-seq time course, with activation as Th0 (t=0) vs (t=72h), and differentiation as Th2 vs Th0 (t=72h). Thus genes/KOs toward the middle of the plot have the least effect. **(b)** Verification of the KO effect by overexpression. The same projection was made onto the activation and differentiation axes. To facilitate comparison the axes have been flipped, thus genes should appear in the same position as in the KO analysis. We find a qualitative agreement. Some genes (black) were not tested by KO. **(c)** t-SNE clustering of transcription factors based on their nearest genes from the ChIP-seq peaks. A particular cluster of Th2 genes is highlighted. **(d)** The motif found under peaks after overexpression and ChIP of BHLHE40. **(e)** Motifs found under peaks after overexpression and ChIP of PPARG. The most significant motifs are listed here. Further motifs are shown in Supplemental Figure S7.

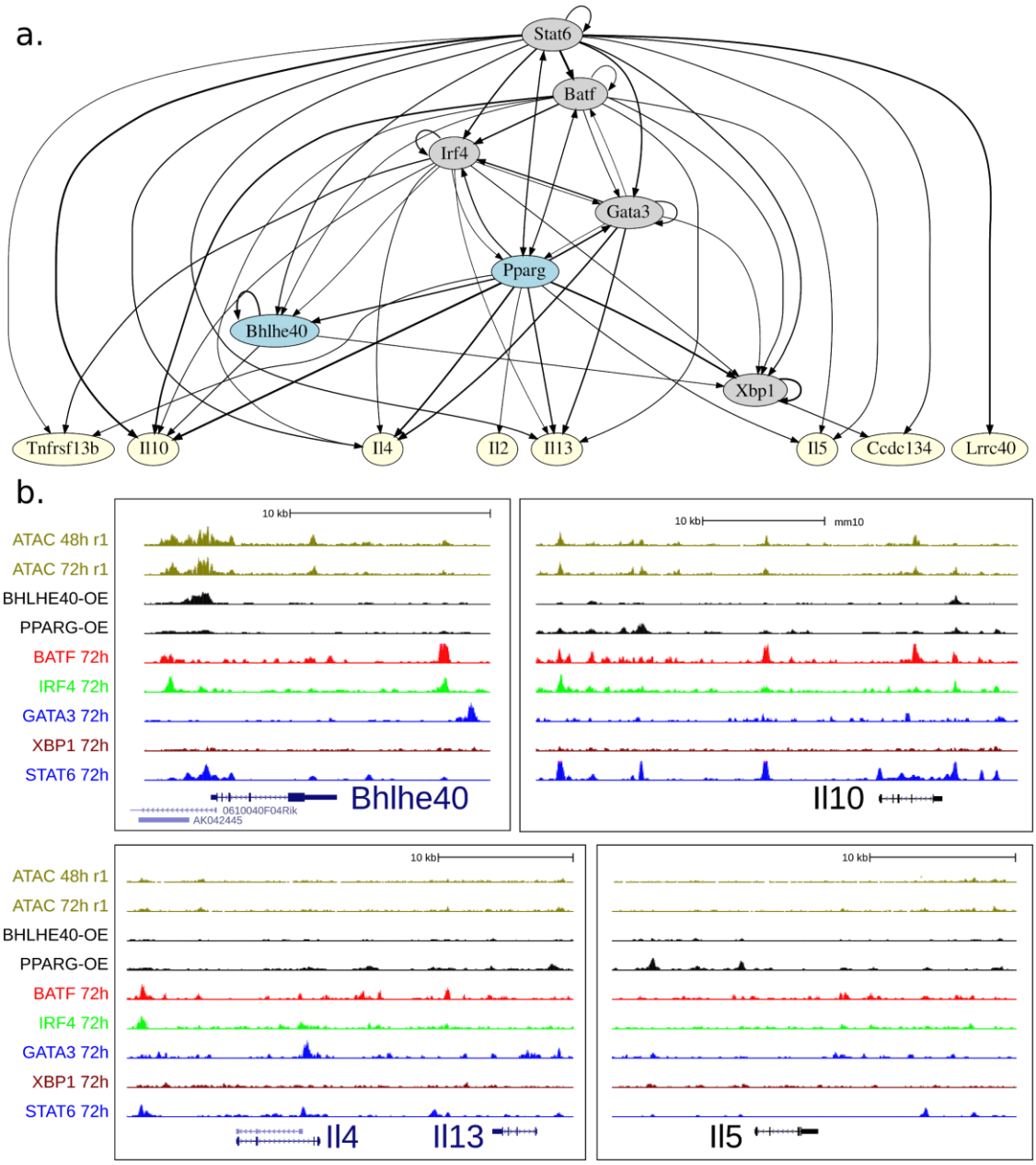

Figure 6: Validation of the Th2 TF network by ChIP-seq (a) Network of Th2 transcription factors based on ChIP-seq peaks. From our validation data, the characterized TFs *Bhlhe40* and *Pparg* are highlighted. *Batf*, *Irf4*, *Pparg* and *Gata3* cluster together as in the Fig. 5 tSNE. **(b)** Binding of these transcription factors to key Th2 genes.

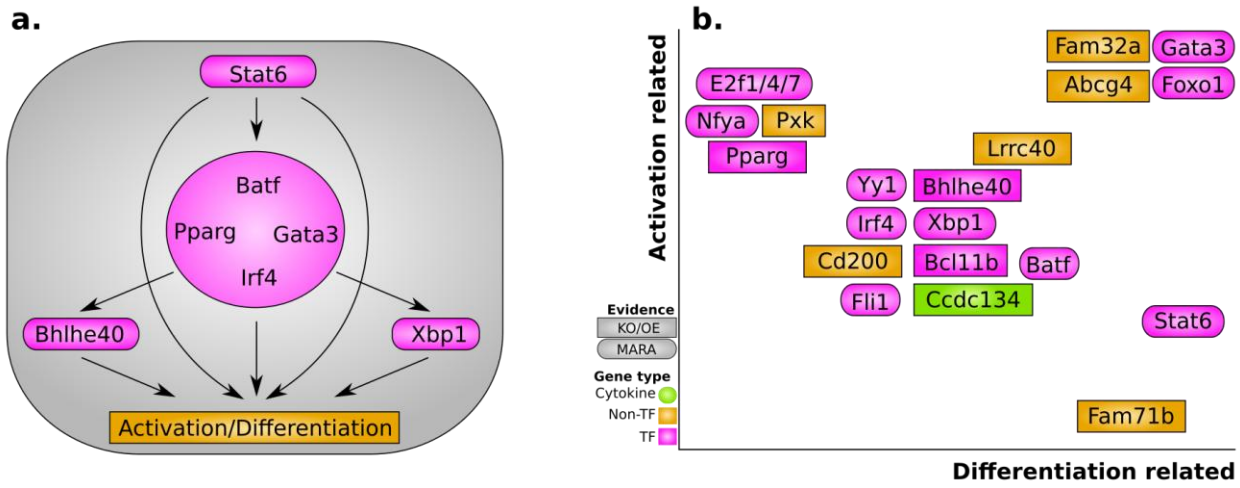

Figure 7: A conceptual view of Th2 differentiation (a) A broad overview of the core transcription factors. We have noted that *Pparg* is particularly integrated in the core program, while *Stat6* is the Th2 entry point that both go through the core program but also connect directly to the cytokines. Several transcription factors work downstream but are less integrated. **(b)** While the genes controlling Th2 fate are from a wide range of programs, their behaviour can be categorized into the modes of activation and differentiation. Here we display some of the highlighted genes according to these categories based on our MARA, CRISPR knock-out and overexpression. The scale is qualitative.

Online methods

Ethics statement

The mice were maintained under specific pathogen-free conditions at the Wellcome Trust Genome Campus Research Support Facility (Cambridge, UK). These animal facilities are approved by and registered with the UK Home Office. All procedures were in accordance with the Animals (Scientific Procedures) Act 1986. The protocols were approved by the Animal Welfare and Ethical Review Body of the Wellcome Trust Genome Campus. The usage of the cord blood of unknown donors was approved by the Ethics Committee of Hospital District of Southwest Finland.

Cloning

The software Collagene (http://www.collagene.org/) was used to design and support the cloning. Phusion polymerase (NEB #M0531L) was used for all cloning PCR reactions.

The entire BFP/puromycin and sgRNA system was PCR-amplified from pKLV-U6sgRNA(BbsI)-PGKpuro2ABFP (primers: kosuke_mfei_fwd/kosuke_clai_rev). The plasmid pMSCV-IRES-mCherry FP (Addgene #52114) grown in dam-/dcm- competent *E. coli* (NEB #C2925I), was digested with NEB ClaI/MfeI and the backbone was gel purified using the QIAquick Gel Extraction Kit (Qiagen #28704). Ligation was done with T4 ligase (NEB #M0202T). The resulting plasmid that can be used to target individual genes was named pMSCV-U6sgRNA(BbsI)-PGKpuro2ABFP (Addgene #102796).

To produce the pooled library pMSCV-U6gRNA(lib)-PGKpuroT2ABFP (Addgene: #104861) the sgRNA part of a previous mouse KO sgRNA pooled library[22] (Addgene: #67988) was PCR-amplified using the primers gib_sgRNAlib_fwd/rev. Up to 1ug was loaded in a reaction and run for 10 cycles. The insert was gel purified, and then repurified/concentrated using the MinElute PCR Purification Kit (Qiagen #28006). The backbone from pMSCV-U6sgRNA(BbsI)-PGKpuro2ABFP was obtained by BamHI-HF (NEB) digestion. The final product was produced by Gibson assembly (NEB #E2611S) and combining the output of 10 reactions. 6 tubes of 5-alpha Electrocompetent *E. coli* (NEB #C2989K) were transformed using electroporation and the final library obtained by combining 4 maxipreps. The library complexity was confirmed by streaking diluted bacteria onto plates and counting colonies. The total number of colonies was >100x the size of the library which according to simulations in R is far beyond the requirement for faithful replication of a library (data not shown).

Two *Cas9* control viruses were also derived from pKLV2(W-)U6sgRNA5(BbsI)-PGKpuroBFP and pKLV2(gfp)U6sgRNA5(BbsI)-PGKpuroBFP. The new plasmids are correspondingly named pMSCV(W-)U6sgRNA5(BbsI)-PGKpuroBFP and pMSCV(gfp)U6sgRNA5(BbsI)-PGKpuroBFP (Addgene #102797, #102798). The cloning was performed in the same manner as for pMSCV-U6sgRNA(BbsI)-PGKpuro2ABFP.

Virus production

293T-cells were maintained in Advanced DMEM/F12 (Gibco #12491015) supplemented with geneticin (500ug/ml, Gibco #10131035). At least one day before transfection, cells were kept in media without geneticin. When at roughly 80% confluency (day 1), the cells were transfected using Lipofectamine LTX. To a 10cm dish with 5ml advanced DMEM, we added 3ml OPTI-MEM (Gibco #31985062) containing 36ul LTX, 15ul PLUS (Thermofisher #15338030), and a total of 7.5ug library plasmid and 7.5ug pcl-Eco plasmid[58] (Addgene #12371). The OPTI-MEM was incubated for 30 minutes prior to addition. The media was replaced with 5ml fresh Advanced DMEM/F12 the day after transfection (day 2), and virus harvested on day 3. Cells were removed by filtering through a 0.45um syringe filter. Virus was either snap frozen or stored in 4°C (never longer than day 5 before being used).

Making of mouse strains

Rosa26$^{Cas9/+}$ mice[22] were crossed with other mice carrying fluorescent reporters. These strains were Gata3^{GFP,59}, Il13$^{+/Tom,60}$ and Il4$^{tm1.1Wep,61}$. For the screens we then pooled mice, both heterozygous and homozygous for *Cas9* expression, male and female, of 8-12 weeks age.

The GATA3-3xFLAG-mCherry mouse strain was produced briefly as follows. The targeting construct was generated by BAC liquid recombineering[62] such that a CTAP TAG element was linked via a Picornavirus “self-cleaving” T2a peptide[63] to mCherry red fluorescent protein and placed upstream of a LoxP/Frt flanked promoter driven Neomycin cassette (CTAP-T2a-mCherry-Neomycin). The cassette was flanked by arms of homology and designed to fuse the tagged fluorescent cassette to the terminal Gata3 coding exon, replacing the stop coding and a portion of the endogenous 3’UTR (Supplemental Figure 3). Two sgRNAs, left 5’CATGCGTGAGGAGTCTCCAA and right 5’CTTCTACTTGCGTTTTTCGC, were designed to generate double-strand breaks 3’ to the terminal stop codon. The respective complementary oligos (Sigma Genosys) were annealed and cloned into a U6 expression vector.
The targeting construct (2ug), along each U6 guide (1.5*2ug) and wild-type Cas9 (3ug, kind gift from George Church) were nucleofected into 3*10^7 JM8 F6 C57Bl/6 ES cells using Amaxa Human Stem Cell Kit 2 (Lonza #VPH-5022) and the Amaxa nucleofector B. Subsequent ES cell injections and animal husbandry were carried out by the Sanger Animal facility.

Validation of *Cas9* mouse

Expression of *Cas9* was confirmed by western blot (anti-Cas9, BioLegend 7A9, #844301) as well as by RT-PCR (primers: cas9_qpcr1/2/r/f). Qualitatively, *Cas9* expression appears to increase during activation of cells (data not included). The function of *Cas9* was also validated using the two control viruses and cytometric analysis. The resulting viruses express both GFP and BFP but only one of them contains a sgRNA targeting its own GFP sequence FACS analysis confirmed a reduction in GFP signal in T-cells infected with the self-targeting virus, as compared to T-cells infected with the control virus (data not included).

T-cell extraction for CRISPR screening

6-well plates were first prepared at least 2 hours before by adding anti-CD3e (1ul/ml, eBioscences #16-0031-81) in PBS, at least 1.2ml/well, and then kept at 37°C.

Cells were extracted from spleens of up to 30 mice by the following procedure: Spleens were massaged through a 70um strainer into cold IMDM media (strainer slanted to avoid crushing the cells). Cells were spun down at 5min/400g and then resuspended in 5ml red blood cell lysis media (3-4 spleens per 50ml falcon tube). After 4 minutes PBS was added up to 50ml and cells spun again. Cells were then resuspended in cold PBS and taken through a 70um strainer. The cells were counted and spun down again. Finally, the cells were negatively selected using EasySep™ Mouse Naïve CD4+ T Cell Isolation Kit (Stem Cell Technologies, #19765) except for the following modifications: The volume and amount of antibodies were scaled down to 20% of that specified by the manufacturer. Up to the equivalent of the cells of 6-7 spleens can be loaded on one "The Big Easy" EasySep Magnet (Stem Cell Technologies, #18001). Overloading it will cause a severe drop in output cells.

On day 0, the cells were then resuspended in warm IMDM supplemented with 2-Mercaptoethanol “BME” (50 uM Gibco #31350010), IL4 (10ng/ml, R&D Systems 404-ML), IL2 (6ng/ml) and anti-CD28 (3ug/ml, eBioscience #16-0281-86) and Pen/Strep, before being seeded onto the 6-well plates (30-40M cells per plate).

T-cell culturing for CRISPR screening

On day 1, the cells were infected by the following procedure. To each well, 1.2ml media was added. This media consisted of 80% virus, 20% IMDM, supplemented with BME/IL2/IL4/anti-CD28 at concentrations as before. In addition, the media contained 8ug/ml polybrene. The plate was put in a zip-lock lag and spun at 1100g for about 2 hours at 32°C. The plate was then put in an incubator overnight (never more than 24h in total). The cells in the media were spun down (the cells attached kept in place) and resuspended with media as after the T cell extraction except with the addition of 2ug/ml puromycin. Each well required 3-4ml media. For the 7 day culturing the media had to be replenished after half the day. We estimate that the MOI was about 0.2. The use of puromycin is essential to keep the FACS time down to reasonable levels (commonly 2ng/ml).

Sorting and genomic DNA extraction

On the day of sorting, cells were extracted and spun down. To eliminate dead cells we performed a “low-g spin”, 5 minutes at 200g. This brought the viability up to roughly 50%. We have in addition tried other methods such as Ficoll (works slightly better but takes 30 minutes and is harder to reproduce) and Miltenyii Dead Cell Removal Kit. In our experience, the Miltenyii kit works great on uninfected cells but effectively removed almost every infected cell when attempted on the real sample. This is most likely because the kit does negative selection against Annexins which might be promoted by the virus or the puromycin.

In the cases when we used antibody reporters, we first fixed and permeabilized using the Foxp3 Transcription Factor Staining Buffer Kit (eBioscience, #00-5523-00). We then used the following

antibodies: PE Mouse anti-XBP-1S (BD Biosciences, #562642), FITC anti-IRF4 (BD Biosciences, #11-9858-80) and Alexa Fluor 488 Mouse anti-GATA3 (BD Biosciences, #560077).

For sorting, cells were resuspended at 40M/ml in IMDM with BME and 3mM EDTA (PBS for the stained cells). The use of EDTA is essential to ensure singlet events at this high cell concentrations. The cells were then sorted into IMDM using either a Beckman MoFlo or MoFlo XDP, or BD Influx. For non-stained screens we could use BFP to ensure that the cells passed were infected. For the stained screens the BFP signal was disrupted by the staining and we performed it blindly. The subsequent steps are not affected by the addition of uninfected cells. During protocol development, the FACS data was analysed using the software FACSanadu[64] (http://www.facsanadu.org).

After sorting the cells, we performed DNA extraction in two different ways. When using fluorescent reporter strains we used the Blood & Cell Culture DNA Midi Kit (Qiagen #13343). For the fixed cells, due to lack of suitable commercial kits (The FFPE kits we have seen are for low amounts of DNA only), we instead performed DNA extraction as follows. Sorted cells were pellet using a table-top centrifuge at 2000g, 5 minutes. Cell pellet was resuspended in 500 ul Lysis Buffer I (50 mM HEPES.KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% Glycerol, 0.5% NP-40, 0.25% Triton X-100) and rotate at 4°C for 10 minutes. Cells were spun down at 2000g, 5 minutes, resuspended in 500 ul Lysis Buffer II (10 mM Tris.Cl, pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA) and rotate at 4°C for 10 minutes. Then the cells were pelleted again at 2000g for 5 minutes, and the cell pellet was resuspended in 25 ul Lysis Buffer III (Tris.Cl, pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-Deoxycholate, 0.5% N-lauroylsarcosine). Then 75 ul TES Buffer (50 mM Tris.Cl pH 8.0, 10 mM EDTA, 1% SDS) was added to the cell suspension. This 100 ul reaction was put on a thermomixer to reverse crosslinking at 65°C, overnight. Then 1 ul proteinase K (20 mg/ml, ThermoFisher #100005393) was added, and protein was digested at 55°C for 1 hour. DNA was purified using MinElute PCR purification kit (Qiagen) according to the manufacturer's instruction. DNA concentration was measured by a Nanodrop.

Sequencing of CRISPR virus insert

The genomic DNA was first PCR-amplified (primers: gLibrary-HiSeq_50bp-SE_u1/l1[22]) in a reaction with Q5 Hot Start High-Fidelity 2× Master Mix (NEB #M0494L). In each 50ul reaction, we loaded up to 3ug DNA. From each reaction we pipetted and pooled 5ul, before purifying it using the QIAquick PCR purification kit (Qiagen #28104). The purified product was then further PCR-amplified using KAPA HiFi HotStart ReadyMix (Kapa #KK2602) and iPCRtag sequencing adapters[65] . After Ampure XP bead purification (beads made up 70% of the solution) and Bioanalyzer QC, the libraries were pooled and sequenced with a HiSeq 2500 (Illumina #SY-401-2501, 19bp SE). The custom primers U6-Illumina-seq2 (R1) and iPCRtagseq (index sequencing) were used for this purpose. The original sgRNA library contained 86,035 distinct sgRNAs. In a representative sequencing run (*Gata3*, using antibody selection) the sgRNAs with fewer than 500 reads encompassing 91% of the total complexity.

Analysis and QC of CRISPR hits

Sequencing BAM-files were transformed into FASTQ using samtools and bamToFastq. A custom Java program was then used extract per-sgRNA read counts. From these, per-gene p-values were calculated using MAGeCK[29] using the positive and negative cell fraction from each screen. The hit rankings were then compared using R. To obtain a total per-gene score, we first calculate the total rank from one screen as $r=\min(r_{pos},r_{neg})$, using the ranks from the positive and negative enrichments respectively. Then, to calculate the composite score of two or more screens, we used the geometric mean $(r_1r_2r_3...r_n)^{1/n}$. Follow-up hits were manually picked as those scoring high between the replicates, with genes of low expression level qualitatively filtered out using ImmGen[66] .

The BaIOPSE model was implemented in STAN[67] using the RStan interface. For the full model implementation and parameters, with variances rather defined by the exponentials over the priors, we refer to the source code. 12 Markov chains were run 800 steps and convergence was checked by the r-value. The top 300 hits were used to calculate GO term *p*-values. GO terms were obtained in R by GO.db[68] and assessed individually using a Fisher exact test.

Mouse time-course RNA-seq

$CD4^+CD62L^+$ naive T cells were purified from spleens of wild-type C57BL/6JAX adult (6 - 8 weeks) mice using the CD4+CD62L+ T Cell Isolation Kit II (Miltenyii #130-093-227). Cell culture plates were coated with anti-CD3e antibody (1 ug/ml, eBioscences #16-0031-81) in 1X DPBS (Gibco) at 4°C overnight. Purified naive T cells were seeded at a concentration of 1 M cells/ml on the coated plates in IMDM (Gibco) with 10% heat-inactivated FBS (Sigma #F9665-500ML), supplied with 5 ug/ml anti-CD28 (eBioscience #16-0281-86) with (Th2) or without (Th0) 10 ng/ml mouse recombinant IL-4 (R&D Systems #404-ML-050). Cells were cultured in plates for up to 72 hours.

Total RNA was purified by Qiagen RNeasy Mini Kit according to manufacturer's instruction, and concentration was determined by a Nanodrop. A total of 500 ng RNA was used to prepare sequencing libraries using KAPA Stranded mRNA-seq Kit (KAPA #07962193001) according to manufacturer's instructions. Sequencing was performed on an Illumina HiSeq 2000 (125bp PE, v4 chemistry).

The efficiency of the Th2 differentiation was confirmed by antibody staining and FACS (Supplemental Figure 5). *In vitro* differentiated Th2 cells were fixed, permeabilized and stained with fluorescent dye conjugated antibodies to detect intracellular cytokine expression following eBioscience intracellular staining protocol as previously described[69] (also described in http://tools.thermofisher.com/content/sfs/manuals/staining-intracellular-antigens-for-flow-cytometry.pdf). Fluorescent dye-conjugated primary antibodies used: IL4 (eBioscience clone #11B11), IL13 (eBioscience clone #eBio13A) and CD4 (eBioscience clone #GK1.5 or #RM4-5). Stained cells were analysed by flow cytometry on a Fortessa (BD Biosciences) using FACSDiva and FlowJo software. CompBeads (BD Biosciences) were used for compensation where distinct positively stained populations were unavailable.

Human time-course RNA-seq

Mononuclear cells were isolated from the cord blood of healthy neonates at Turku University Central Hospital using Ficoll-Paque PLUS (GE Healthcare, #17-1440-02). CD4+ T cells were then isolated using the Dynal bead-based positive isolation kit (Invitrogen). CD4+ cells from three individual donors were activated directly in 24w plates with plate-bound anti-CD3 (500ng/well, Immunotech) and soluble anti-CD28 (500 ng/mL, Immunotech) at a density of 2×10^6 cells/mL of Yssel's medium[70] containing 1% human AB serum (PAA). Th2 cell polarization was initiated with IL-4 (10 ng/ml, R&D Systems). Cells activated without IL-4 were also cultured (Th0). At 48 hr, IL-2 was added to the cultures (17 ng/ml, R&D Systems). Cells were harvested at respective time points and RNA was isolated using RNeasy Mini Kit (Qiagen #74106) for library preparation. The efficiency of the Th2 differentiation was confirmed by measuring GATA3 levels using western blot (WB) and RT-qPCR.

For WB, cells were lysed in Triton-X-100 lysis buffer (TXLB) (50 mM Tris-HCl (pH 7.5), 150 mM NaCl, 0.5% Triton-X-100, 5% Glycerol, 1% SDS) and sonicated for 5 min using a Bioruptor sonicator. Protein concentration was then estimated using DC Protein assay (Biorad #500-0111). Equal protein amounts were loaded onto acrylamide gel (Bio-Rad Mini or Midi PROTEAN TGX precast gels). For protein transfer to PVDF membranes, mini or a midi transfer packs from Bio-Rad were used, depending on the gel size. Primary and secondary antibody incubations were performed in 5% Non-Fat milk or BSA in TBST buffer (0.1%Tween 20 in Tris-buffered saline). The following antibodies were used: GATA3 (BD Pharmingen #558686); GAPDH (Hytest #5G4MAB6C5); Actin (Sigma, #A5441).

For RT-qPCR, RNA was isolated (RNeasy Mini Kit, #74106, Qiagen) and treated in-column with DNase (RNase-Free Dnase Set, #79254, Qiagen) for 15 minutes. The removal of genomic DNA was ascertained by treating the samples with DNase I (Invitrogen, #18068-015) before cDNA synthesis with SuperScript II Reverse Transcriptase (Invitrogen, #18064014). RT-qPCR was performed using the following *Gata3*-specific primers and a double labelled (FAM-reporter, TAMRA-quencher) probe: 5'-GGACGCGGCGCAGTAC-3' (left primer), 5'-TGCCTTGACCGTCGATGTTA-3' (right primer), 5'-TGCCGGAGGAGGTGGATGTGCT-3' (probe). KAPA Probe Fast Rox Low master mix (KAPA Biosystems, #kk4718) was used and amplification was monitored with QuantStudio 12K Flex Real-Time PCR System (ThermoFisher Scientific).The C_t values were normalized against the signal acquired with *EF1α*, using the following primers and probe: 5' -CTGAACCATCCAGGCCAAAT- 3' (left primer), 5' -GCCGTGTGGCAATCCAAT- 3' (right primer), 5'-AGCGCCGGCTATGCCCCTG- 3' (probe).

ATAC-seq was performed from same cultures for better comparability. ATAC-seq libraries were prepared as described below.

Time-course ATAC-seq data generation

Experiments were done according to the published protocol[71] with some modification. Briefly, 50,000 cells were washed with ice cold 1X DPBS twice, and resuspended in a sucrose swelling buffer (0.32 M sucrose, 10 mM Tris.Cl, pH 7.5, 3 mM $CaCl_2$, 2 mM $MgCl_2$, 10% glycerol). The cell suspension was left on ice for 10 minutes. Then, a final concentration of 0.5% NP-40 was

added, and the cells suspension was vortexed for 10 seconds and left on ice for 10 minutes. Nuclei was pelleted at 500 g at 4°C for 10 minutes. Nuclei were washed once with 1X TD buffer (from Nextera DNA Library Preparation Kit, Illumina, #FC-121-1030), and resuspended in 50 ul tagmentation mix containing:

- 25 ul 2X TD buffer (Nextera DNA Library Preparation Kit, Illumina #FC-121-1030)
- 22.5 ul H_2O
- 2.5 ul TDE1 (Nextera DNA Library Preparation Kit, Illumina #FC-121-1030)

The tagmentation reaction was carried out on a thermomixer at 37°C, 800 rpm, for 30 minutes. The reaction was stopped by the addition of 250 ul (5 volumes) Buffer PB (from Qiagen MinElute PCR Purification Kit), The tagmented DNA was purified by Qiagen PCR Purification Kit according to manufacturer's instructions and eluted in 12.5 ul Buffer EB from the kit, which yielded ~10 ul purified DNA.

The library amplification was done in a 25 ul reaction include:

- 10 ul purified DNA (from above)
- 2.5 ul PCR Primer Cocktail (Nextera DNA Library Preparation Kit, Illumina #FC-121-1030)
- 2.5 ul N5xx (Nextera index kit , Illumina #FC-121-1012)
- 2.5 ul N7xx (Nextera index kit , Illumina #FC-121-1012)
- 7.5 ul NPM PCR master mix (Nextera DNA Library Preparation Kit, Illumina #FC-121-1030)

PCR was performed as follows:

- 72°C 5 minutes
- 98°C 2 minutes
- [98°C 10 secs, 63°C 30 secs, 72°C 60 secs] x 12
- 10°C hold

Amplified libraries were purified by double Agencourt AMpureXP beads purifications (Beckman Coulter, #A63882). 0.4X beads:DNA ratio for the first time, flow through was kept (removing large fragments); 1.4X beads:DNA ratio for the second time, beads were kept. Libraries were eluted from the beads by elution in 20 ul Buffer EB (from Qiagen PCR Purification Kit).

1 ul library was run on a Agilent Bioanalyzer to check size distribution and quality of the libraries.

Sequencing was done with an Illumina Hiseq 2500 (75 bp PE).

ChIP-seq data generation

ChIPmentation[72] was used to investigate the TF binding sites. 1 million cells from each sample were crosslinked in 1% HCHO (prepared in 1X DPBS) at room temperature for 10 minutes, and HCHO was quenched by the addition of glycine at a final concentration of 0.125 M. Cells were pelleted at 4°C at 2000 x g, washed with ice-cold 1X DPBS twice, and snapped frozen in liquid nitrogen. The cell pellets were stored in -80°C until the experiments were performed. ChIPmentation was performed according to the version 1.0 of the published protocol (http://www.medical-epigenomics.org/papers/schmidl2015/) with some modifications at the ChIP stage. The antibody used were IRF4 (sc-6059), BATF (sc-100974) and FLAG (Sigma M2, #F3165).

Briefly, cell pellets were thawed on ice, and lysed in 300 ul ChIP Lysis Buffer I (50 mM HEPES.KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, pH 8.0, 10% Glycerol, 0.5% NP-40, 0.25% Triton X-100) on ice for 10 minutes. Then cells were pelleted at 4°C at 2000 x g for 5 minutes, and washed by 300 ul ChIP Lysis Buffer II (10 mM Tris.Cl, pH 8.0, 200 mM NaCl, 1 mM EDTA, pH 8.0, 0.5 mM EGTA, pH 8.0), and pelleted again at 4°C at 2000 x g for 5 minutes. Nuclei were resuspended in 300 ul ChIP Lysis Buffer III (10 mM Tris.Cl, pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Sodium Deoxycholate, 0.5% N-Lauroylsarcosine). Chromatin was sonicated using Bioruptor Pico (Diagenode) with 30 seconds ON/30 seconds OFF for 10 cycles. 30 ul 10% Triton X-100 was added into each sonicated chromatin, and insoluble chromatin was pelleted at 16,100 x g at 4°C for 10 minutes. 1 ul supernatant was taken as input control. The rest of the supernatant was incubated with 10 ul Protein A Dynabeads (Invitrogen) pre-bound with 1 ug anti-FLAG in a rotating platform in a cold room overnight. Each immunoprecipitation (IP) was washed with 500 ul RIPA Buffer (50 mM HEPES.KOH, pH 7.5, 500 mM LiCl, 1 mM EDTA, 1% NP-40, 0.7% Sodium Deoxycholate, check components) for 3 times. Then, each IP was washed with 500 ul 10 mM Tris, pH 8.0 twice, and resuspended in 30 ul tagmentation reaction mix (10 mM Tris.Cl, pH 8.0, 5 mM Mg2Cl, 1 ul TDE1 (Nextera)). Then, the tagmentation reaction was put on a thermomixer at 37°C for 10 minutes at 800 rpm shaking. After the tagmentation reaction, each IP was washed sequentially with 500 ul RIPA Buffer twice, and 1X TE NaCl (10 mM Tris.Cl, pH 8.0, 1 mM EDTA, pH 8.0, 50 mM NaCl) once. Elution and reverse-crosslinking were done by resuspending the beads with 100 ul ChIP Elution Buffer (50 mM Tris.Cl, pH 8.0, 10 mM EDTA, pH 8.0, 1% SDS) on a thermomixer at 65°C overnight, 1,400 rpm. DNA was purified by MinElute PCR Purification Kit (QIAGEN, #28004 and eluted in 12.5 ul Buffer EB (QIAGEN kit, #28004), which yielded ~10 ul ChIPed DNA.
The library preparation reactions contained the following:

- 10 ul purified DNA (from above)
- 2.5 ul PCR Primer Cocktails (Nextera DNA Library Preparation Kit, Illumina #FC-121-1030)
- 2.5 ul N5xx (Nextera Index Kit , Illumina #FC-121-1012)
- 2.5 ul N7xx (Nextera index kit , Illumina #FC-121-1012)

- 7.5 ul NPM PCR Master Mix (Nextera DNA Library Preparation Kit, Illumina #FC-121-1030)

PCR was set up as follows:

- 72°C, 5 mins
- 98°C, 2 mins
- [98°C, 10 secs, 63°C, 30 secs, 72°C, 20 secs] x 12
- 10°C hold

The amplified libraries were purified by double AmpureXP beads purification: first with 0.5X bead ratio, keep supernatant, second with 1.4X bead ratio, keep bound DNA. Elution was done in 20 ul Buffer EB (QIAGEN).

1 ul of library was run on an Agilent Bioanalyzer to see the size distribution. Sequencing was done on an Illumina Hiseq 2000 platform (75 bp PE, v4 chemistry).

ChIP-seq peak analysis

The reads were first trimmed using Trimmomatic 0.36[73] with settings ILLUMINACLIP:NexteraPE-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:30. Peaks were then called using MACS2[38], merged over time, and annotated using HOMER[74].

The quality of the peaks was assessed using the two available replicates for each time point. While the trend over time agreed, the number in each time point did not. For this reason we decided to consider the union of the peaks rather than the common peaks.

The sequences of the detected ChIP-seq peaks were extracted using “bedtools getfasta”[75], for 200, 300, 400, 500bp regions around the peaks. These were fed into MEME[46] for additional motif discovery.

To compare genes associated to increasing and non-increasing GATA3 peaks, we calculated the relative peak height at 72h vs 24h. We define the most/least increasing as the peaks with top/bottom 1000 ratios, and then included the genes the peaks are closest to. GO analysis were performed to compare these groups.

Time-course RNA-seq differential expression

Gene expression from RNA-seq data was quantified in TPM using Salmon v0.6.0[76] , with the parameters --fldMax 150000000 --fldMean 350 --fldSD 250 --numBootstraps 100 --biasCorrect --allowOrphans --useVBOpt. The cDNA sequences supplied contain genes from GRCm38 (mouse), GRCh38 (human) and sequences from RepBase, as well as ERCC sequences and an eGFP sequence.

Differentially expressed (DE) genes were found using the Sleuth R package[77] , using the wasabi R package (https://github.com/COMBINE-lab/wasabi) to allow it to accept Salmon input data. To strengthen the test of differential dynamics between Th2 and Th0 culture conditions, instead of testing each time point individually (with few replicates), we separated time into early (<=6h) and late (>6h). The DE test consisted of a likelihood-ratio test using the sleuth_lrt function, where the full model contained terms accounting for the culture condition, for the temporal effect (modelled as a spline with 5 degrees of freedom) and for an interaction of both terms. To capture the Th0/Th2 difference, the reduced model only contained a term accounting for the time variation, modelled as before. A gene is considered differentially expressed for p-value < 0.01.

Human/Mouse *Stat6* comparison

Targets of Stat6 and Il4 as defined by time-course microarray and ChIP-seq data were downloaded from a previous study[5].

ATAC-seq motif extraction

ATAC-seq reads were aligned using Bowtie 2[78] with the parameter –X 2000 and the mouse genome mm10. This was followed by peak calling on each replicate individually using MACS2[38] with the function “callpeak” and the parameters -B --SPMR --call-summits. The peaks obtained were kept if they overlapped a peak from the other replicate of the same time point by at least 50%. In these cases, the new peak would equal the combined coordinates of all the overlapping peaks considering all replicates and time points.

Peaks were classified (annotatePeaks.pl --annStats) as intronic, exonic, upstream or intergenic, according to the gene feature they intersected. Intersection is scored first considering the number of bases overlapped, and then the closeness in size between the peak and the feature.

Known motif detection was performed on the peaks’ sequences using FIMO[39] , and motifs from the JASPAR 2016 database[79] considering only those starting in MA or PB. In addition, we supplemented with a more recent list of C2H2 motifs[80]. To make the analysis more targeted, only motifs from TFs DE between Th2 and Th0 were considered, and for each of them a single motif was selected, prioritizing the longest ones with the lowest mean entropy.

The overlap between human and mouse was calculated using liftOver -minMatch=0.03 -multiple. Roughly 100 peaks mapped to multiple sites and were thus ignored. LiftOver was also performed on individual TF sites from FIMO. The overlap between organisms was calculated using R GenomicRanges[81]. The overlap procedure was done at the peak and detected motif levels.

We found that the analyses throughout the paper appear to give similar results when using all mouse peaks as opposed to only using the conserved (overlapping) peaks. However, the ChIP-seq peaks of GATA3, IRF4 and BATF appear more comparable to ATAC-seq predicted sites if only the conserved sites are used are used in the MARA.

ATAC-seq chromatin dynamics analysis

The height of the peaks, as well as any reads outside the peaks, were quantified using bedtools[75]. The peak levels were divided by the background signal for normalization. Further, to make the contributions from different peaks comparable, they were normalized to the level of the second time point. The contribution of motifs over time is defined as the average peak signal in which they are present.

UCSC visualization of ChIP-seq and ATAC-seq

The MACS2 BedGraph files were prepared for UCSC visualization using bedSort and bedGraphToBigWig.

MARA analysis

The MARA analysis was performed as follows. Early and late times were analyzed independently. For each of the two durations, the connectivity matrix was constructed based on if a motif peak was present for a gene at any time. The number of such peaks, ignoring time fluctuations, were entered as the connectivity value. The full RNA-seq time-course data for either Th0 or Th2 was used as the signal. These two files were uploaded to ISMARA[50] using expert mode.

In the MARA comparison over time, Th0 and Th2 difference is calculated as the average MARA activity difference over time. The activity increase is taken as the difference in activity at the first and last time points for Th0.

Follow-up knock-out RNA-seq data generation

The backbone pMSCV-U6sgRNA(BbsI)-PGKpuro2ABFP was digested using BbsI and purified on a gel. 96*2 desalted primers for the sgRNA insert were obtained from Sigma in premixed and diluted format. They were diluted to 10uM in T4 ligation buffer (NEB, #M0202T) and annealed (cooling from 98°C to 4°C during 1 hour on a PCR block). Ligations were performed in 10ul volume, in a 96w PCR on ice. Transformed *E. coli* (Stbl3, made competent in lab) were streaked onto 10cm ampicillin agar plates using an 8 channel pipette.

To avoid validating individual colonies, a mixture of at minimum 10+ colonies were picked and mixed for each clone. Digest by BbsI of a few representative shows at the minimum presence of clones without original bbsI spacer. Bacteria were grown overnight in a 96w deep-well plate having an air-permeable seal. Minipreps were made using a homemade gravity manifold holding miniprep tubes (blueprint for laser cutting available on request). The virus was subsequently made in 293T-cells, in 24w format. The virus was then harvested into a 96w deep-well plate and any 293T removed by centrifugation.

Naive T cells were extracted from 3 mice independently, this time with the Naive CD4+ T Cell Isolation Kit (Miltenyii #130-104-453) according to manufacturer's instruction. Cells were seeded at 200k/well density in 96w format. Infection and puromycin selection was then performed as

before. On day 5, cells were washed and dead cells removed by low-G spin. This typically raised the viability from roughly 10-20% to 60% according to Trypan blue. Cells were spun down and as much of the media removed as possible. Up to 100ul of buffer RLT+ was then added to each well and plates frozen. Later, plates were thawed and RNA extracted by adding 100ul of Ampure XP beads. Purification was done by a robot, with 2x200ul EtOH wash and final suspension in RNAse-free water. RNA was then diluted to 500ng/ul and 2ul was taken as input into non-capping DogSeq (manuscript in preparation). This protocol is for this application roughly equivalent to Smartseq-2[82]. Libraries were made using Nextera XT and all 96*3 libraries sequenced with a HiSeq 2500 (150bp PE).

Follow-up knock-out RNA-seq analysis

Reads were filtered using Cutadapt for the Smart-seq2 TSO and mapped using GSNAP[83] . The software featureCounts was then used to produce a final count table[84] . The effects of the KO was studied using an EdgeR[85] linear regression model using the KO with scrambled sgRNA as reference point. We studied the impact of the virus infection level, measured as a function of BFP, and found it to be confounding. To obtain stronger DE effect for future KO experiments we recommend that non-infected cells are removed by FACS sorting rather than puromycin selection. Individual replicates were compared in terms of p-value and correlation of DE genes when one sgRNA was used *versus* when several sgRNAs targeting the same gene were pooled. Libraries with low replicability or low virus infection were manually removed.

We define a differentiation axis as the DE genes (using DEseq2) from the RNA-seq time-course, Th0 vs Th2 at t=72h, with $p<10^{-10}$. The activation axis is similarly defined as the DE genes Th2 at t=0h vs Th2 at t=72h, with $p<10^{-10}$.

Similarly, we define axes for each T helper type using DE data from Th-express[40] with $p<10^{-2}$. For Figure 6ab we then calculate the similarity score based on the fold changes as $s_i=\sum_{i\in g} ref_i\ ko_i/|g|$, where g is the set of genes which in the KO have at least a 2-fold change.

The plasticity dimension reduction Figure 5c is produced as follows. A standard pinwheel type diagram cannot be produced because of a lack of absolute T cell type references (cultured under the same conditions as the KO, with scrambled sgRNA, and same RNA-seq protocol). The diagram was instead based on the previous Th-Express DE vectors. The position of each KO is calculated as $\alpha\sum s_i\ \mathbf{v}_i$ where **v** is the vector from Th2 to Th_i, and α an arbitrary constant. As a result, the labels Th_i are merely directions toward T helper cell types, and not absolute coordinates in the cell type state space.

Follow-up overexpression RNA-seq and ChIP-seq data generation

The plasmid pMSCV-IRES-Blue FP (Addgene #52115, gift from Dario Vignali) was digested with NEB MfeI-HF and BamHI-HF and purified with the QIAquick PCR Purification Kit.

The cDNA for cloning was generated by RLT lysis of *in vitro* Th2 cells (naive and day 5), Ampure XP bead purification, and SmartSeq2 first strand synthesis without the addition of the TSO.

For each gene, PCR primers were generated using an R script. The CDS of each gene was downloaded from Ensembl, and the most DE transcript as given by our RNA-seq time course was selected. Genes that later proved hard to clone from cDNA were ordered as gBlocks, with optional codon optimization using the IDTDNA web interface.

The first (genefwd) and last 30bp (generev) were used as gene-specific PCR primer part. We created the primers by concatenating sequences as follows, where RC denotes reverse complement: primer_fwd=(overlapFWD, seqkozak, genefwd), primer_3xflag_fwd=(seq3xflag, flagspacer, genefwd) and primer_rev=(overlapREV, rc(seqStop), generev). Further, to genes not starting with the codon G after ATG, the sequence GAG was added. All primers mentioned here were ordered PAGE purified from IDTDNA. The sequences were fwd_3x=TCTTACGTAGCTAGCGGATCttaaccatggactacaaagaccatgacggtgattataaagatcatgacatc gattacaaggatg, seq3xflag=cggtgattataaagatcatgacatcgattacaaggatgacgatgacaag, rev_3x=AATTGATCCCGCTCGAGCCTACTTGTCATCGTCATCCTTGTAATCGATGTCATGATC TTTATAATCACCG, seqkozak=ttaaccatg, seqStop=tag. The gene specific forward sequences are in the Supplemental Files. The specific gene product was first obtained by PCR of the cDNA with Phusion master mix, gene_fwd and gene_rev primers (25 cycles, 2 minute extension, 72C annealing). The result was run on an 1% agarose gel, the band cut, and purified with Qiaquick gel purification kit and Qiaquick PCR purification kit.

To obtain 3xFLAG versions of the insert, a phusion PCR reactions were set up (8 cycles, 65C annealing, 2min extension) over the previously amplified non-3xFLAG fragments, primers fwd_3x, primer_3xflag_fwd and primer_rev. The gene specific inner primer was used at 0.5uM concentration as opposed to 10uM for the outer primers. The products were purified by Ampure XP and eluted into 10ul NFW.

The inserts and the MSCV backbone were joined with NEB Gibson assembly master mix in 10ul reactions. 1ul of the ligated product was transformed in 25ul NEB DH5a competent cells, according to manufacturer's specification.

Colonies were picked, amplified in 5ml LB (50ml Falcon tube) and plasmid purified by miniprep. Validation was done by two rounds of Sanger sequencing, forward (mscv_seq2,CTTGAACCTCCTCGTTCGAC) and reverse (mscv_seq3,TAACATATAGACAAACGCACACCG). A custom Java program was written to find the best matching expected sequence (generated by previous R script) and output a FASTA file with the reference and reads (reverse read reverse-complemented). The sequences for each clone were aligned by CLUSTALW[86]. The result was visualized with CLUSTALX and/or a plain text editor. The sanger sequencing reads of all clones are available as-is on GitHub. The finally selected clones are available on AddGene with IDs #117263-117267.

Naive T cells were purified, cultured *in vitro* with IL4, and transduced as described before. The original pMSCV backbone was used as a negative control. On day 5, 20k BFP+ cells were FACS sorted, spun down and lysed in RLT. RNA-seq libraries were generated as described before. The remaining cells were used for ChIPmentation, with the IP performed against

3xFLAG (antibody ref here). RNA-seq read and ChIPmentaton reads were processed as before. RNA-seq and ChIP-seq libraries were sequenced on a HiSeq 2500, 50BP SE.

Presence of the overexpressed gene was verified by manual inspection of RNA-seq reads as well as *p*-value in DE gene list. Constructs failing this test were excluded. DESeq2 was used for the analysis and we tested 3 linear models: ~treatment, ~treatment + mouse, and treatment *vs* all samples. They all give consistent results. Here we report using the model ~treatment + mouse and do not report less consistent DE genes. Interestingly we obtained two clones of *Lrrc40*, where one was truncated, but they qualitatively yield the same DE genes during overexpression, and cluster together using tSNE (data not shown). This suggests that 5' We have deposited both versions to Addgene, and here report the genes for the full-length *Lrrc40*.

Clustering and analysis of ChIP-seq datasets

Additional BED-files of ChIP-seq datasets were downloaded from Cistrome[87]. The following external datasets were included, in addition to the 3xFLAG BHLHE40 & PPARG, and our endogenous BATF, IRF4 and GATA3 ChIP-seq: ASCL2 Th0 overexpression[88] (GSM1276938), BACH2 Th2[89] (GSM1547779), CDK9 Th2[90] (GSM1527704), E2F4 DC cells 120 min post LPS-stimulation[91] (GSM881061), E2F1 DC cells 120 min post LPS-stimulation[91] (GSM881057), ELF1 Th0 FOXP3-[92] (GSM999185), ETS1 Th2[48] (GSM654875), ETV6 Th2[93] (GSM2634697), FLI1 Th2[48] (GSM654872), FOLS2 Th0[94] (GSM1004808), FOXO1 Foxp3+ CD4[94] (GSM1480611), FOXP3 Treg[95] (GSM2387501), GTF2B CD4/CD8[96] (GSM727002), GTF2F1 CD4/CD8[96] (GSM727004), IKZF2 Treg[97] (GSM1876372), IRF8 Th2[93] (GSM2634696), JARID2 Th17[98] (GSM1151625), JUN CD4[26] (GSM978754), JUNB CD4 resting and activated[99] (GSM1646847 and GSM1646848), MAF Th17[98] (GSM1151623), NIPBL CD4/CD8[100] (GSM1184315), PARP14 Th2[101] (GSM1242997), PRDM1 Treg[102] (GSM1964752), RAD21 CD4/CD8[100] (GSM1184316), RARA Th1[103] (GSM1474186), RORC Th0[11] (GSM1004853), RUNX1 CD4[99] (GSM1646844), RUNX1 CD4 + PMA[99] (GSM1646846), SATB1 peripheral CD4[104] (GSM2409720), SATB1 thymus CD4[104] (GSM2409719), SMAD4 Th17[105] (GSM2706519 and GSM2706520), SMARCA4 Th2 resting and stimulated[106] (GSM585295 and GSM585297), SMC1A CD4/CD8[107] (GSM1504389, GSM1504390), SPNS1 Th2[108] (GSM550319), STAT1[109] (GSM994528), STAT1 Th0[110] (GSM1601720), STAT3 Th0[110] (GSM1601721), STAT2 Th0[111] (GSM2538951), STAT3 Th0[11] (GSM1004857), STAT4 Th1[108] (GSM550303), STAT5 Th0[112] (GSM2055717 and GSM2055711), STAT6 Th2[108] (SRR054675), TBX21 Th1[113] (GSM836124), XBP1 Th2 (manuscript under review)[69].

The Rtsne package[114] was used on a matrix consisting of 1 wherever a gene had a close ChIP peak according to annotatePeaks.pl[74], otherwise 0. To generate the network diagrams, the output of annotatePeaks.pl was processed with R to select peaks near the TSS of chosen genes. The network was written to an output file and rendered using Graphviz (https://www.graphviz.org).