An unbiased reconstruction of the T helper cell type 2 differentiation network

Authors: Johan Henriksson[1,3], Xi Chen[1], Tomás Gomes[1], Kerstin Meyer[1], Ricardo Miragaia[1], Ubaid Ullah[4], Jhuma Pramanik[1], Riita Lahesmaa[4], Kosuke Yusa[1], Sarah A Teichmann[1,2,*]

Affiliations:

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom

[2]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

[3]Karolinska Institutet, Department. of Biosciences and Nutrition, Hälsovägen 7, Novum, SE-141 83, Huddinge, Sweden

[4]Turku Centre for Biotechnology, Tykistokatu 6 FI-20520, Turku, Finland

*To whom correspondence should be addressed: st9@sanger.ac.uk

Tomas: tpcg@sanger.ac.uk
Ricardo: rm13@sanger.ac.uk
Ubaid Ullah: uullah@btk.fi
Jhuma: jp19@sanger.ac.uk
Kerstin: km16@sanger.ac.uk
Kosuke Yusa: ky1@sanger.ac.uk
Riita Lahesmaa: riitta.lahesmaa@utu.fi
Sarah A Teichmann: st9@sanger.ac.uk

keywords: CD4 T helper cell, mouse, Cas9, CRISPR, pooled screen, library, retrovirus,

knock-out, ATAC-seq, ChIP-seq, time-course

Abstract: T helper type 2 (Th2) cells are important regulators of our adaptive immune response, particularly the response against parasites, and have relevance for auto-immunity as well as tumour progression. This classic T helper type has been studied intensively, but not systematically. Using newly developed, genome-wide retroviral CRISPR knock-out (KO) technology, combined with RNA-seq, ATAC-seq and ChIP-seq, we have dissected the regulatory circuitry governing differentiation in these cells. During Th2 activation/differentiation approximately 4000 genes are perturbed, with at least 200 genes specifically associated with the Th2 program in mouse and human. We confirm previously known Th2 driver genes and have discovered several novel genes, including transcription factors, metabolic genes and potential receptors/cytokines, critical for Th2 function. Our study provides an atlas for, but not limited to, the Th2 regulatory network, pinpointing the key players of Th2 differentiation.

Introduction

CD4+ T helper (Th) cells are a central part of the adaptive immune system and play a key role in infections, auto-immunity and tumor-repression. During the immune response, Th cells activate 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 understanding Th2 cells 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], which in turn induces the Th2 master regulator *Gata3*[4–7]. As Th2 cells upregulate their key inducer, a positive feedback loop is formed. Th1 cells possess an equivalent mechanism for their defining transcription factor, Tbx21, which represses Gata3. Gata3 is able to inhibit IFNG, the main cytokine driving Th1 differentiation. Thus, the balance of the two transcription factors Tbx21 and Gata3 define the Th1-Th2 axis[8]. There are however many genes affecting this balance, and alternative Th fates are frequently affected by overlapping sets of regulatory genes. Despite the importance of different T helper subtypes, so far only the Th17 subtype has been examined systematically[9,10].

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 RNA interference screen has been performed *in vivo* on mouse T cells[11]. However, recently-developed CRISPR technology has the advantages of higher specificity and greater flexibility, allowing knock-out[12], repression[13–15] and activation[16–18]. Currently all existing CRISPR libraries are lentiviral-based and therefore unable to infect murine T helper cells[19]. 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. We chose *Il4*, *Il13*, *Gata3*, *Irf4* and *Xbp1* as our primary screen read-outs. *Il4*, *Il13*, *Gata3* are at the core of the Th2 network[8] while *Irf4* and *Xbp1* have been suggested to have supporting roles in keeping the chromatin accessible and for overcoming the stress response associated with rapid protein synthesis during T cell activation[20–22]. Selected genes discovered by the screen were validated in individual knock-outs and assayed by RNAseq. To place the discovered genes into the context of Th2 differentiation, we profiled developing Th2 cells using RNA-seq for gene expression, ATAC-seq for chromatin accessibility and ChIP-seq of three key transcription factors: GATA3, IRF4 and BATF. Here we provide a resource of genes potentially associated with Th2 differentiation and, by integrating a range of datasets and past discoveries, highlight those most likely to be functional.

Results

Retroviral sgRNA delivery enables CRISPR/Cas9 mediated genome editing in primary mouse T cells

Figure 1 depicts an overview of our experimental approach. First, a high complexity retroviral sgRNA library was generated (see online methods) (Fig 1b) and transduced into purified T cells from mouse spleens. We activated naive CD4+ T cells with anti-CD3 and anti-CD28 together with IL4 at day 0, using a newly developed protocol to efficiently culture large numbers of primary cells (see methods). At day 1, T cells were transduced with the retroviral libraries and selected with puromycin from day 3. After dead cell removal the actual screens were carried out on day 4. Our screening strategy used two different approaches. For *Il4*, *Il13* and *Xbp1* and *Gata3*, we used T cells from transgenic mice, carrying a fluorescent reporter driven by the promoter of the respective genes. Cell populations with high versus low fluorescence will be enriched in sgRNAs targeting genes inhibiting or promoting Th2 cell differentiation, respectively. In addition, we carried out two screens in which T cells were stained with antibodies for IRF4, XBP1 or GATA3. Again, high protein expression is the read-out of efficient differentiation. 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 reporter high and low fractions. We will refer to the top enriched gRNA target genes as “hits”.

In our experimental system, viral infection requires proliferation. Therefore, theoretically, genes that are required for differentiation only at a very early time point, but not later on, could be missed by our screen. However, genes that are required for both induction and maintenance of differentiation will be identified. Furthermore, our system allows the large-scale testing of a whole genome sgRNA library. Our approach therefore operates at an entirely different scale from a recently performed CRISPR knock-out (KO) study in mouse T cells that relied on producing a new transgenic mouse lines[23]. The success of our approach is illustrated by the fact

that a screen based on GATA3 antibody staining identified the *Gata3* gene itself as its top hit (see below).

Screen identifies core transcription factors affecting T cell differentiation

In total 10 genetic screens were carried out. As an illustration of the results obtained, Fig 2a shows the hits in a screen using anti-Gata3 antibody staining (i.e. sgRNA for specific target genes), ranked by MAGeCK[24] p-value, against the fold change (Th2, 0h vs 72h, described later) of those sgRNA targeted genes. Notably, *Gata3* is identified as a regulator of itself, as expected. 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. We note that *Gata3* expression changes are relatively small, but still result in a strong phenotype. 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 a screen for Y. If gene X falls within the top 1% of ranked hits we will write $X^{\rightarrow y!}$. Reassuringly, hits that positively regulate *Gata3* include genes already known to be involved in T cells development, e.g. *Gata3* itself, followed by Stat6$^{\rightarrow Gata3}$, F2rl1$^{\rightarrow Gata3}$ (ligand for F2r, whose deficiency has several immunophenotypes[25]), Pxk$^{\rightarrow Gata3}$ (associated with systemic lupus, SLE[26]) and the IL4 receptor itself (IL4 used here to induce Th2). Candidate genes that negatively regulate *Gata3* are harder to interpret but overall we find genes involved in K^+ and Ca^+ activity such as *Slc5a11*$^{\rightarrow Gata3}$ (also related to SLE[27]). Calcium signaling and metabolism are already known to be important for immunity[28]. One of the top enriched genes in the low fluorescent population is the antibacterial Sept6$^{\rightarrow Gata3}$ which is highly expressed in Th1[29].

Next we identified hits that were consistent between screens (see methods). A comprehensive list of all genes is included in Supplementary data and results are summarised in Figure 2b. Notably, for Gata3 screens we found high reproducibility between independent screens, even though Gata3^{hi} populations were selected either by reporter gene expression or by antibody staining. Whilst there was some overlap between candidate regulators of *Gata3* and *Il4* expression, most of the chosen phenotypes generated their own unique set of hits, with only few common shared regulators amongst the most consistent hits. However, the phenotypes for the screens are clearly linked and the same hits are frequently identified in multiple screens (e.g. Pou6f1$^{\rightarrow Il4!,Gata3}$ and Trappc12$^{\rightarrow Il4!,Irf4,Gata3!}$). Examples of candidate regulators identified in screens for 3 or more different targets, include *Ccdc134*, *Il27ra* and *Lag3*. We compared the results our screens to randomised screen scores, suggesting reasonable efficiency of our screens (Fig 2c).

As a further quality control we also compared the screens with our own hit calling model, BaIOPSE (Bayesian Inference Of Pooled Screen Enrichment). The model is described in Fig 2d, and it uses a hit calling approach orthogonal to the RRA (Robust Rank Aggregation) analysis deployed by MAGeCK. In short, the BaIOPSE hierarchical model simultaneously fits size factors, overall screen efficiencies, probe efficiencies and gene KO effect. Qualitatively, we find that there is reasonable overlap with MAGeCK. We further performed a GO analysis of top hits from all the screens (Fig 2e). Out of the identified categories, calcium and MAPK signaling

have the lowest p-values. In addition, many lymphocyte related GO terms are identified. While BaIOPSE allows a more consistent integration of multiple screen replicates than MAGeCK[24], we have however still decided to use MAGeCK for the remainder of this paper because of it's higher community acceptance and the strong priors used by BaIOPSE.

We also compared the sgRNA counts from *Gata3*^{hi} population to the original library in order to identify genes essential for survival (original library count table available in Supplementary material). *Gata3* does not appear to be differentially enriched, but *Stat6* sgRNAs still showed significant enrichment in the Gata3^{hi} population, consistent with its role in activation. Compared to the original sgRNA libraries, the consensus of the top 300 candidate genes from each screen include *Fadd*$^{\rightarrow Il4}$, involved in T cell proliferation[30], *Pim3*$^{\rightarrow Gata3}$, related to cell cycling and apoptosis[31], *Pou6f1*$^{\rightarrow Il4,Il13,Gata3}$ (a transcription factor with few known functions so far) and Srp72$^{\rightarrow Il4}$ (associated with myelodysplasia[32]). Candidate genes with negative effect on chosen targets include Med12, an essential regulator of HSC homeostasis[33].

Finally, we have examined whether genes classically associated with T cell survival were identified as hits in any of the screens, and discovered - amongst others - *Il2*$^{\rightarrow Irf4,Gata3}$ (both *Gata3* expression and proliferation), *Cxcr2*$^{\rightarrow Il4}$ (*Gata3* expression) and *Irf4*$^{\rightarrow Il13}$. We conclude that even if the survival screen GO-terms (not shown) suggests that we can separate genes affecting differentiation and viability, these two processes are clearly linked.

We also compared the hits with the screened immunological phenotypes of the International Mouse Phenotyping Consortium (IMPC)[34]. The full comparison is in the Supplementary Material. Among the transcription factors having potential immune related phenotypes, we noted in particular *Tcf7*$^{\rightarrow Il4}$, *Irf5*$^{\rightarrow (Il4)}$, *Xbp1*$^{\rightarrow Il4}$, *Mbd1*$^{\rightarrow Il13}$, *Arid1b*$^{\rightarrow Irf4}$ and *Zfp408*$^{\rightarrow Gata3}$.

Time-course data and human-mouse comparison reveals the Th2 gene set

To further filter out which genes appear to be of Th2 relevance, we generated time course data on mouse and human T cells during activation/differentiation (Fig. 3a). Both mouse and human primary T cells were isolated and stimulated with anti-CD3 and CD28 for activation. The addition of IL4 to the medium resulted in the formation of Th2 cells, while absence of IL4 resulted in Th0 cells. RNA-seq was performed on Th2 and Th0, and ATAC-seq on Th2. The large number of data points allowed us to reconstruct the time-course trajectory of Th2 differentiation by principle component analysis (PCA), using RNA-seq data or ATAC-seq data alone (Fig 3b, c). When carrying out differential gene expression (DE) analysis between the Th0 and Th2 populations, we split the time-course into the early (0h-6h) and late (6h-72h) phases (Fig 3a). In the mouse, a large number of DE genes were identified at both phases, with approximately 10% of the genes being differentially regulated over time (Fig 3d). Importantly, of the DE genes a sizeable fraction (21%) were also identified in at least one of our genetic screens, providing orthogonal data for their importance. Evolutionary conservation can also suggest functional relevance. We therefore carried out an equivalent analysis in cultured human 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 homologues in the mouse response.

A total of 216 genes are DE in both mouse and human, early or late ($p=10^{-4}$). Of these, genes that also are top hits in our CRISPR screens are shown in Fig 3d. We note the presence of the 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, others are related to metabolism, such as Pparg, which is thought to signal through mTOR and controls fatty acid uptake[35]. Another metabolic gene, related to fatty acid transport[36], that has a strong phenotype in our screen is Abcd$^{\rightarrow Il13,Ir4,Gata3!}$ which has not been studied in T cells. The Th1-repressor *Mapkapk3*$^{\rightarrow Il4,Gata3!,37}$ is also a metabolic gene. Other genes have more diverse function, including the the Stat-inhibitor *Socs1*$^{\rightarrow Irf4,Xbp1}$ (Suppressor of cytokine signaling 1). The Il13 hit *Rasgrp1*$^{\rightarrow Il13,Irf4}$ (RAS guanyl releasing protein 1) is known to be involved in T cell maturation[38] and links Guanyl to the RAS pathway. Interestingly another Guanylate protein, Gbp4$^{\rightarrow Il13}$, is also an Il13 hit (but with higher DE p-value, not in this list). The 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[39]. In conclusion, a human-mouse comparison of differentially expressed genes highlights cytokines and TFs known to be important in Th2 differentiation, and suggests additional hits in our screens that are likely to be of functional importance, in particular genes that act as metabolic regulators.

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.

Chromatin dynamics reveal developmentally important TF

To gain further insight into regulation of gene expression we examined chromatin accessibility using ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing). The chromatin of naive T cells is condensed until activation. It has previously been shown that e.g. *Stat5* cannot access the promoters of its target genes until activation[40]. Th2 differentiation is then 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[41]. Overall there is a massive gain of chromatin accessibility from 0 to 2h (Sup. Fig. 1 and Fig 3f). Interestingly, the chromatin then appears to recondense continuously and this process progresses past 72 hours, as indicated by the 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. A previous model has shown that 80% of T cells die instead of entering mitosis[42]. The simultaneous, rapid activation of many genomic regions may require strong checkpoint control - as evidenced by cell death - to ensure correct operation.

We next compared TF binding between human and mouse. Using FIMO we predicted TF binding sites within peaks. This approach is expected to give many false positives and we therefore concentrated on ATAC peaks that are conserved between mouse and human by calculating the percentage of overlapping peaks between species (10-15%)(Fig 3e). Previous work showed that whilst individual binding sites are often not conserved, regulatory pathways and nodes frequently are[43]. We thus decided to use the conserved binding sites for the rest of the analysis.

For each transcription factor we examined how the ATAC peaks, in which the TF motifs are found, are changing over time (Fig 3f). Batf::jun is one of the (composite) motifs associated with the largest increase in relative peak size. This is consistent with the suggestion that Batf is able to open chromatin[9]. Of the TFs identified in the screen, *Jun*$^{\rightarrow Il13}$, *Fos*$^{\rightarrow Irf4,Xbp1}$, *Fosl2*$^{\rightarrow Gata3!}$ and *Gata3*$^{\rightarrow Xbp1,Gata3}$ are all associated with increasing peak height. However, this needs to be related to their expression levels. Notably, *Fosl2* expression peaks at 1-2h in Th0/2 with largely unchanging levels in Th1/2/17/Treg[29]. Since JUN, FOS and FOSL2 all recognise the AP-1 motif, it is likely that other TFs bind the same motifs at the later time-points. Overexpression of *Fosl2* has been shown to block IL17A production in Th17 by competing for AP1-sites[9], but overall *Fosl2* expression is low in lymphoid cells[44]. At the other extreme, some TFs are dominant in decaying peaks, 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*$^{\rightarrow Il13!}$. *Foxo1*$^{\rightarrow Il13!,Xbp1!}$ has recently been shown to inhibit H3K27me3 deposition at pro-memory T-cell genes[45]. In our Th2/0 time course, Foxo1 is decreasing in both expression level and ATAC peak size. A list of all transcription factors ranked by the height of peaks containing their cognate motif is provided in Supplementary materials.

For some of the most crucial Th2 transcription factors (*Gata3*, *Batf* and *Irf4*) we decided not to rely on ATAC-seq putative binding sites but instead generated ChIP-seq data at 72 hours after Th2 differentiation. *Batf* and *Irf4* have a large overlap as previously reported[9], with the 72h peak overlap shown in Fig 3g. In addition we find that the *Yy1*$^{\rightarrow Il4,Il13,Xbp1,Gata3}$ binding motif is enriched within the *Gata3* peaks (MEME $p=2.5*10^{-58}$). This is consistent with previous reports that *Yy1* overexpression is sufficient to drive Th2 cytokine expression[46] while CD4-dependent *Yy1*-KO mice struggle to produce Th2 cells. However, *Yy1* overexpression does not rescue *Gata3*-deficiency. This study also showed that *Yy1* and *Gata3* are co-located in Th2, but not Th1 cells. The ATAC-seq data time course (Fig 3f) shows that peaks containing *Gata3* motifs increase over time while the *Yy1* peaks are stable or decreases slightly. This finding, together with the strong phenotype in our screen, pinpoints *Yy1* as one of the most important supporting but not driving Th2 genes. Furthermore, *Yy2*$^{\rightarrow Gata3}$ is also a hit in the screens, although its expression is low in these cells.

To conclude, the early change in ATAC-seq peak size over time reflects a rapid increase in accessibility for all TFs. Later, we see evidence of pioneer factors driving the opening of chromatin and find TFs whose binding is reduced with time. These are all likely to be functionally important as either activators or repressors for the specific T helper type.

Time-course regulatory network have screen-enriched regulons

So far, hits have been considered one at a time. The generation of transcriptional networks potentially allows us to integrate the effect of multiple regulators. A regulatory network was created using ARACNe-AP[47] on the time course RNA-seq data (full network in Supplementary materials). In essence, edges are created between genes having a high probability of high mutual information (correlation). Only the edges conserved between human and mouse were kept. In this network the screen hits cluster together, in particular those for *Il4* ($p=10^{-2}$ for combined rank<300), but also *Gata3*, *Il13* and *Irf4*. The screen hits also cluster together with hits from other screens. The upstream genes of *Xbp1*$^{\rightarrow Il4}$ have a similar, but weaker trend.

The full network is too complex to be visualized. However individual regulons with a high number of hits are displayed in Fig 4h, filtered by ATAC conserved binding sites. This filtering also imposes a directionality on the edges. For each gene enriched in a particular screen, a binomial test was performed to see if the connected genes are also enriched in the screen. For the ATAC-filtered time-course network of DE mouse genes, *Ube2m*$^{\rightarrow Il4!,Gata3!}$ has 4 out of 16 genes affecting *Il4* (p=0.5%). Only *Ybx1*$^{\rightarrow Il4}$ (p=1.1%) comes close. *Irf8*$^{\rightarrow Xbp1}$ is the top *Xbp1*-regulon (p=0.1%) followed by *Hk1*$^{\rightarrow Il4,Xbp1}$ (p=0.2%) and *Sqstm1*$^{\rightarrow Xbp1!}$ (p=1%). The canonical TF *Tbx21*$^{\rightarrow Il13,Xbp1}$ is 5th but with p=14%. The top regulon for *Irf4* is the widely affecting Arginyl aminopeptidase, *Rnpep*$^{\rightarrow Il4,Irf4,Gata3}$ (p=4%). This gene has similarities to other aminopeptidases and inhibiting drugs against these have been shown to modulate the immune response[48].

The level of the above mentioned genes were also compared between different Th types[29]. Ube2m stand out in iTreg, Sqstm1 in Treg/Th17, and Irf8 and Tbx21 vary overall. However, for example, Rnpep is not DE. This highlights the orthogonality of the ARACNe-based network approach, and our ability to extract hits which are neither DE nor TFs.

Motif activity analysis verifies known Th2 transcription factors

The ARACNe approach is based on similarities in gene expression between a given TF and its potential targets. However, TF function is frequently determined post-transcriptionally, for example STAT6$^{\rightarrow Gata3}$ activity is regulated by phosphorylation, and TBX21$^{\rightarrow Xbp1,Il13}$ partially operates by sequestering and displacing GATA3$^{\rightarrow Gata3,Xbp1,8}$ once expressed (100-fold up-regulated in Th1[29]). In contrast, the ISMARA algorithm 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. Edges are then curated by searching for co-regulation amongst predicted targets of a given TF (see methods). Interestingly we find that there is very high correlation in the predicted network when comparing results obtained with sites predicted from ATAC-seq peaks to results obtained with measured TF binding (Chip-seq data) (Fig 4b), suggesting that the algorithm performs well on ATAC-seq input data and allowing us to analyse all TFs, not just those with ChIP-seq data.

To obtain an overview of the role of all the transcription factors we have categorized factors according to their activity over time within one differentiation pathway, and whether their activity differs when comparing Th2 to Th0 cells (Fig 4f). In other words two distinct comparisons are made: Firstly t=0h versus t=72h within the Th2 compartment, which we term “activation”, and secondly Th0 versus Th2 cells at t=72h, which we term “differentiation”. Figure 4c 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 *Yy1*$^{\rightarrow Il4,Il13,Xbp1,Gata3}$. The majority of TFs display a behaviour reflecting both activation and differentiation (Fig 4d). The calculated activity score was also related to expression (Fig 4c, e). The Th2-defining TF *Gata3*$^{\rightarrow Gata3!,Xbp1!}$ shows transient activity but its expression increases 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 Il4r$^{\rightarrow Gata3}$. However, our data suggests that Stat6 activity continues to increase throughout differentiation. Interestingly all the STAT proteins map closely together in Fig 4d, affecting both differentiation and activation, possibly all contributing to different extents that may be influenced by their change in expression (Fig 4e). Foxo1$^{\rightarrow Il13,Xbp1}$ and *Xbp1*$^{\rightarrow Il4}$ are also strongly connected to activation and differentiation, but with effects in the opposite direction. Previous work suggested that the primary role of *Batf* is to open the chromatin together with *Irf4*[9], and this is consistent with our analysis in Fig 3f. Here it is one of the strongest differentiators, suggesting that the opening is restricted to sites required for differentiation. The roles of other genes are more diffuse. TFs that were identified as hits include *Atf4*$^{\rightarrow Il4!,}$ [49]and *Yy2*$^{\rightarrow Gata3}$ affecting both activation and differentiation, but with weaker effects. The previously seen *Yy1*$^{\rightarrow Il4,Il13,Xbp1,Gata3}$ is only related to activation. *Id4*$^{\rightarrow Il13,(Il4),Xbp1}$, *Ebf1*$^{\rightarrow Irf4}$, *Foxp2*$^{\rightarrow Gata3}$ and *Fli1*$^{\rightarrow Il4!}$ are further pure activators. The identification of a cluster of E2F-proteins (e.g. *E2f7*$^{\rightarrow Il13}$) as strongly activation-related is consistent with their role in cell cycle control.

The MARA approach allowed us to extract canonical Th2 transcription factors, such as Stat6, Gata3 and Batf, and in addition highlighted newly identified TF-hits (E2f7, Foxo1) likely to be relevant for Th2 differentiation. Since MARA is not directly dependent TF-target gene co-variation, the output is complementary to the previous DE and ARACNe approaches.

Plasticity effect of genes interrogated by individual CRISPR KO

Our generation of ATAC-seq and ChIP-seq data allowed a detailed interrogation of the involvement of TFs in Th2 differentiation. However, most of the hits from the screens are not transcription factors. To study these genes, we performed targeted KO followed by RNA-sequencing. Besides a non-targeting control sgRNA, we also included a palette of known regulators, such as *Gata3*, as reference points. Several of the chosen genes have been studied before though not specifically in T helper cells. The KO gene list is by no means comprehensive and additional genes can be found in the Supplementary material.

While transcription factors were classified into modes simply using MARA, we utilised a different, but related method to to summarise the effects observed in different knock-out RNA-seq data sets Using the *in vitro* differentiation data (Fig 3a,b) we defined an activation axis as the DE genes from Th2(0h) vs Th2(72h) and a differentiation axis as the DE genes from Th0(72h) vs Th2(72h). For each knock-out a DE list vs non-targeting control was derived and its overlap with the differentiation and activation DE lists assessed (Fig 5a). It should be noted that some genes might not be consistently higher or lower in Th2 versus Th0 cells. There are thus many other possible definitions of the differentiation axis. To find if a KO aligns with one of these axes, the DE genes of the KO are compared, by sign only, to the DE genes of the axis. From this analysis, we find that in particular Stat6 stands out on both axes, consistent with the MARA analysis. To give more examples, the most differentiation related genes are B230219D22Rik$^{\rightarrow Xbp1}$, Fam32a$^{\rightarrow Gata3!}$, and Abcg4$^{\rightarrow Irf4}$. The first two have not been studied before while Abcg4 is a cholesterol transporter, as yet unstudied in the T cell context. It is however known that steroids such as cholesterol regulate immune cells[50]. Abcg1$^{\rightarrow Il4,Xbp1}$ is also present in the screen.

The activity/differentiation analysis only considers Th0 vs Th2. However extensive previous work indicates that different Th fates are interrelated and gene KOs may push cells towards different fates. To assess this, first a DE list of KO versus WT are obtained. Next we compare our Th2 gene expression signature to that from other culture conditions[29], which resulted in the formation of Th1, Th17, iTreg or nTreg cells, generating a second DE list. The overlap of the two lists is assessed (Fig 5b) quantitatively (% overlap, shown by colour intensity) and qualitatively (same or opposite direction, depicted in blue and red). A complex pattern emerges that further emphasises the complexity and plasticity within the T helper cell compartment. A simplified overview using a dimensionality reduction is also provided (Fig 5c). In this visualisation, all data points start in the middle and are then moved towards or from different fates by vectors representing the similarity between the KO DE list and the fate DE list, as indicated by colours in Fig 5b. By design, KOs having small or ambiguous effect are placed in the middle. In contrast, high Stat6 expression pushes cells towards a Th2 phenotype in three of the four different comparison, with the effect that Stat6 maps very closely to the Th2 fate in Figure 5c. We note that a central position does not reflect a Th0 phenotype since all comparisons are relative to the IL4 induced Th2 state. It reflects either very small changes from Th2, or opposing effects in different T fates that can cancel each other out in this graphic representation.

Our analysis illustrates that the different T cell fates are connected, but that their relationships are complex. Some genes can favour both Th17 and Th_1 (e.g. *Bcl11b*), or Th17 and iTreg (e.g *Hsph1*), or Treg and iTreg (e.g *Fam32a*) or lastly, Treg and Th1 (e.g. *Cd200*) (Fig 5b). Inspecting individual genes, *Stat6*$^{\rightarrow Gata3}$ is shown to promote Th2, and IL13$^{\rightarrow Il4,Xbp1}$ is identified as pro-Th1. *Il13* is known to be dispensable for Th2 and can in fact be produced by Th1 and Th17[51]. The dimensionality reduction pinpoints additional strong pro-Th2 gene hits: *Nrd1/Nrdc*$^{\rightarrow Irf4}$, which has recently been shown to regulate inflammation in the murine stomach by controlling *Tnfa* shedding[52]and *Slc5a1*$^{\rightarrow Irf4}$ (glucose/galactose transport). Further Th1/2-regulators might include *Slc25a3*$^{\rightarrow Irf4}$ (transporter of phosphate into mitochondria),

Herc6$^{\rightarrow Il4}$ (E3 ligase for ISG15, important for antiviral activity[53]), *Cxcr7/Ackr3*$^{\rightarrow Il13,Xbp1}$ *(chemokine receptor)*, the elusive[54] *Thy1*, *Pgk1*$^{\rightarrow Il4}$ (glycolysis-related enzyme) and the novel *Lrrc40*$^{\rightarrow Gata3}$. The possibly most interesting gene is the secreted *Ccdc134*$^{\rightarrow Il4!,Irf4!,Gata3}$ which has previously been reported as a novel cytokine affecting ERK and JNK, however no follow-up has been carried out[55]. Here it favours Th2 over Treg, but not over Th1/17.

In summary, we have carried out a gene-by-gene validation for a number of selected Th2 regulators. RNA-seq patterns for KOs of previously known regulators, such as Gata3 and Stat6, are consistent with their ascribed role. However, in addition to this, we find that some of the newly identified hits, demonstrate extremely interesting phenotypes with respect to T helper differentiation. These are exemplified by Abcg4, Fam32 and Lrrc40 and these genes deserve further study in the future.

Discussion

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 provide a resource of genes important for T 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 brings the advantages of improved targeting and gene disruption instead of just down-regulation. In addition, we generated RNA-seq data for both mouse and human T cells following a time course to Th2 cell differentiation. We have generated a website (http://data.teichlab.org) that will allow the scientific community to query individual genes with respect Th2 expression dynamics and chromatin accessibility. By combining our CRISPR KO screen with time-course data, we have been able to provide an updated map of the most important genes for Th2 polarization.

Key to our study is a systematic and unbiased approach to interrogate genes that contribute to Th2 differentiation. Whilst our analysis has identified a large number of hits, our ability to represent a differentiation process as a regulatory circuit diagram is still only rudimentary and hence it is difficult to place the identified genes within this context. Key challenges relate to our limited ability to (i) define regulatory networks for different types of proteins and (ii) design read-outs solely associated with the differentiation process in question, as outlined below.

The hits we identify belong to many different classes of proteins, including cytokines, transcription factors, proteins involved in calcium signalling and metabolic genes. The mechanisms by which they impinge on the “regulatory network” is in many cases still unclear and difficult to address experimentally, especially as protein-protein interactions are still hard to profile at scale. An exception are transcription factors: we know these regulate large number of genes by virtue of binding to their promoters and protein-DNA interaction can be profiled genome-wide through the time course of differentiation. We have taken advantage of this in our study and have employed RNA-seq, ATAC-seq and ChIP-seq to dissect the gene regulatory

network. This has allowed to us to identify TFs that play an important role in development and have also been highlighted by our genetic screens.

In our analysis, we have chosen 5 different read-outs (Gata3, Il4, Il13, Xbp1, and Irf4), each known to be associated with Th2 differentiation. However, we do not know exactly at what stage in the developmental trajectory their function is required and their analysis at the end stage (day 4) of differentiation may not truly reflect the role they play. In addition, an *in vitro* differentiated cell may not fully reflect the cell's *in vivo* characteristics, and a particular phenotype may have relevance to more than one cell type. To overcome these limitations, many studies rely on *in vivo* mouse experiments to see if the ability to mount an anti-helminth response is retained. However, such a complex assay is clearly not suitable for high-throughput screening, meaning that less distal phenotypes need to be chosen as proxy. However, the fact that the different screens we have carried out generated many overlapping hits increases our confidence that relevant phenotypes were chosen.

Due to the difficulties of placing other non-TF hits into a regulatory wiring diagram, we have chosen to interpret our findings in the context of "Waddington's landscape" (Figure 6a) which can be applied to all genes. Within the developmental landscape of Th2 differentiation we propose 3 modes of action: Activating, Differentiating and Amplifying. Genes are sorted according to the first two in Fig 6b. Genes related to activation either drive cells towards or inhibit cells from developing toward a mature state, as seen as the difference between early and late gene expression. Genes related to differentiation are selected based on the difference between Th2 and Th0. Not all genes appear to fit the modes of activation versus differentiation. For example, Fli1 has previously been shown to be important for Th2, but not Th1, but does not come out as a differentiation related in our analysis. To resolve this, we propose the mode of Amplification, which refers to how the Waddington valleys are constructed. While a gene or pathway might not be driving the system, it might be rate-limiting for the differentiation to proceed in a certain direction. Seeing that metabolic genes affect differentiation, but possibly do not drive it, these could be seen as amplifiers. Further, some transcription factors with neutral impact may function as amplifiers to the system. This can occur through stabilizing DNA binding by acting cooperatively, such as among the many AP-1 factors, or in the case of Yy1 interacting with GATA3.

Our analysis of Th2 differentiation has allowed us to re-discover many 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, Bcl11b, and Bhlhe40 (Fig 6b). Additional genes that deserve further future study include Cd200, Fam32b, Abcg4, B23219D22Rik and Fam71b, as well as the cytokine Ccdc134.

Many questions remain - most centrally, what is a Th2 cell? Here we have used, among other, *Il4* and *Gata3* expression as proxies, but both of these measures have caveats. If a KO of a gene important for secretion blocks IL4-release, should such a gene be considered a differentiation gene? Similarly, the localization of GATA3 has been shown to be controlled by TBX21[8] and be crucial for the Th1/2 axis. Neither of these effects are captured by our read-out.

Many studies rely instead on *in vivo* mouse experiments to see if the capability of anti-helminth response is retained. However, a Th2 response is not equivalent to simply the presence of IL4-secreting *Gata3*-positive Th2 cells. The greatest challenge in basic T cell research is thus to clarify the cell type definitions. This requires a paradigm shift in which qualitative models such as the one in Fig 6 are replaced with quantitative models. In particular Waddington's landscape needs be explicitly mapped for all T helper cell types together, which is now within reach using methods such as CROP-seq[56]. In the context of our model we would expect that "activating" factors might affect several T cell fates, whilst "differentiating" factors may be subtype specific. However, this will have to await future experiments. At this stage, our paper provides both a list of the most crucial genes that deserve further analysis, and a working protocol for CRISPR-mediated KO, and is thus a major step towards a complete model of T helper cell biology.

Author contributions

R.M. helped with cell culturing and early trials of virus production and infections. X.C. and U.U. performed the time course RNA-seq, ATAC-seq and ChIP-seq. X.C. further helped with with western blots, antibody staining and DNA isolation from fixed cells. T.G. performed the time course data spline DE and PCA. K.M. advised on the ARACNe analysis and helped with the manuscript. K.Y. provided the Cas9 mice, the sequencing protocol and advised on cloning. 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, 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 Kirton and Jennie Graham helped with cell sorting.

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.T. by the European Research Council.

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 and the R code used for the analysis is available on Github (https://github.com/mahogny/th2crispr).

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. 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).
5. 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).
6. 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).
7. 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).
8. 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).

9. Ciofani, M. *et al.* A validated regulatory network for Th17 cell specification. *Cell* **151,** 289–303 (2012).

10. Yosef, N. *et al.* Dynamic regulatory network controlling TH17 cell differentiation. *Nature* **496,** 461–468 (2013).

11. Chen, R. *et al.* In Vivo RNA Interference Screens Identify Regulators of Antiviral CD4+ and CD8+ T Cell Differentiation. *Immunity* **41,** 325–338 (2014).

12. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–826 (2013).

13. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154,** 442–451 (2013).

14. Kearns, N. A. *et al.* Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat. Methods* **12,** 401–403 (2015).

15. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159,** 647–661 (2014).

16. Maeder, M. L. *et al.* CRISPR RNA-guided activation of endogenous human genes. *Nat. Methods* **10,** 977–979 (2013).

17. Perez-Pinera, P. *et al.* RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat. Methods* **10,** 973–976 (2013).

18. 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).

19. Baumann, J. G. *et al.* Murine T cells potently restrict human immunodeficiency virus infection. *J. Virol.* **78,** 12537–12547 (2004).

20. 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).

21. Li, P. *et al.* BATF-JUN is critical for IRF4-mediated transcription in T cells. *Nature* **490,** 543–546 (2012).

22. 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).

23. Beil-Wagner, J. *et al.* T cell-specific inactivation of mouse CD2 by CRISPR/Cas9. *Sci. Rep.* **6,** 21377 (2016).

24. Li, W. *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15,** 554 (2014).

25. Blake, J. A. *et al.* Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res.* **45,** D723–D729 (2017).

26. Vaughn, S. E. *et al.* Lupus risk variants in the PXK locus alter B-cell receptor internalization. *Front. Genet.* **5,** 450 (2014).

27. Tsai, L.-J. *et al.* The sodium-dependent glucose cotransporter SLC5A11 as an autoimmune modifier gene in SLE. *Tissue Antigens* **71,** 114–126 (2007).

28. Feske, S. Calcium signalling in lymphocyte activation and disease. *Nat. Rev. Immunol.* **7,** 690–702 (2007).

29. Stubbington, M. J. T. *et al.* An atlas of mouse CD4+ T cell transcriptomes. *Biol. Direct* **10,** 14 (2015).

30. Marín-Rubio, J. L. *et al.* Deregulated FADD expression and phosphorylation in T-cell lymphoblastic lymphoma. *Oncotarget* **7,** 61485–61499 (2016).

31. Quan, J., Zhou, L. & Qu, J. Knockdown of Pim-3 suppresses the tumorigenicity of glioblastoma by regulating cell cycle and apoptosis. *Cell. Mol. Biol.* **61,** 42–50 (2015).

32. Király, A. P. *et al.* Familial Acute Myeloid Leukemia and Myelodysplasia in Hungary. *Pathol. Oncol. Res.* (2017). doi:10.1007/s12253-017-0216-4

33. Aranda-Orgilles, B. *et al.* MED12 Regulates HSC-Specific Enhancers Independently of Mediator Kinase Activity to Control Hematopoiesis. *Cell Stem Cell* **19,** 784–799 (2016).
34. Meehan, T. F. *et al.* Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nat. Genet.* (2017). doi:10.1038/ng.3901
35. 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).
36. Dean, M., Rzhetsky, A. & Allikmets, R. The human ATP-binding cassette (ABC) transporter superfamily. *Genome Res.* **11,** 1156–1166 (2001).
37. 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).
38. 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).
39. Obata, Y. *et al.* The epigenetic regulator Uhrf1 facilitates the proliferation and maturation of colonic regulatory T cells. *Nat. Immunol.* **15,** 571–579 (2014).
40. 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).
41. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
42. Proserpio, V. *et al.* Single-cell analysis of CD4+ T-cell differentiation reveals three major cell states and progressive acceleration of proliferation. *Genome Biol.* **17,** 103 (2016).
43. Stefflova, K. *et al.* Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* **154,** 530–540 (2013).

44. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science* **347,** 1260419–1260419 (2015).

45. 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).

46. 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).

47. Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32,** 2233–2235 (2016).

48. Noma, T., Klein, B., Cupissol, D., Yata, J. & Serrou, B. Increased sensitivity of IL2-dependent cultured T cells and enhancement of in vitro IL2 production by human lymphocytes treated with Bestatin. *Int. J. Immunopharmacol.* **6,** 87–92 (1984).

49. 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).

50. Mahata, B. *et al.* Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep.* **7,** 1130–1142 (2014).

51. Gallo, E., Katzman, S. & Villarino, A. V. IL-13-producing Th1 and Th17 cells characterize adaptive responses to both self and foreign antigens. *Eur. J. Immunol.* **42,** 2322–2328 (2012).

52. Kimura, Y. *et al.* Nardilysin regulates inflammation, metaplasia, and tumors in murine stomach. *Sci. Rep.* **7,** 43052 (2017).

53. Oudshoorn, D. *et al.* HERC6 is the main E3 ligase for global ISG15 conjugation in mouse

cells. *PLoS One* **7,** e29870 (2012).

54. Haeryfar, S. M. M. & Hoskin, D. W. Thy-1: more than a mouse pan-T cell marker. *J. Immunol.* **173,** 3581–3588 (2004).
55. Huang, J. *et al.* CCDC134, a novel secretory protein, inhibits activation of ERK and JNK, but not p38 MAPK. *Cell. Mol. Life Sci.* **65,** 338–349 (2008).
56. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* (2017). doi:10.1038/nmeth.4177
57. Tzelepis, K. *et al.* A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep.* **17,** 1193–1205 (2016).
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. Quail, M. A. *et al.* SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. *BMC Genomics* **15,** 110 (2014).
63. Heng, T. S. P. *et al.* The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9,** 1091–1094 (2008).
64. Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *Journal of Statistical*

Software, Articles **76,** 1–32 (2017).

65. Carlson, M. GO.db: A set of annotation maps describing the entire Gene Ontology. (2016).
66. 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).
67. 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).
68. 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).
69. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
70. 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).
71. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
72. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** 202–208 (2009).
73. 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
74. Pimentel, H. J., Bray, N., Puente, S., Melsted, P. & Pachter, L. Differential analysis of

RNA-Seq incorporating quantification uncertainty. *bioRxiv* 058164 (2016).

75. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
76. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–1018 (2011).
77. 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).
78. Schmitges, F. W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.* **26,** 1742–1752 (2016).
79. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9,** e1003118 (2013).
80. Balwierz, P. J. *et al.* ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.* **24,** 869–884 (2014).
81. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9,** 171–181 (2014).
82. 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).
83. 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).
84. 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).

Figures

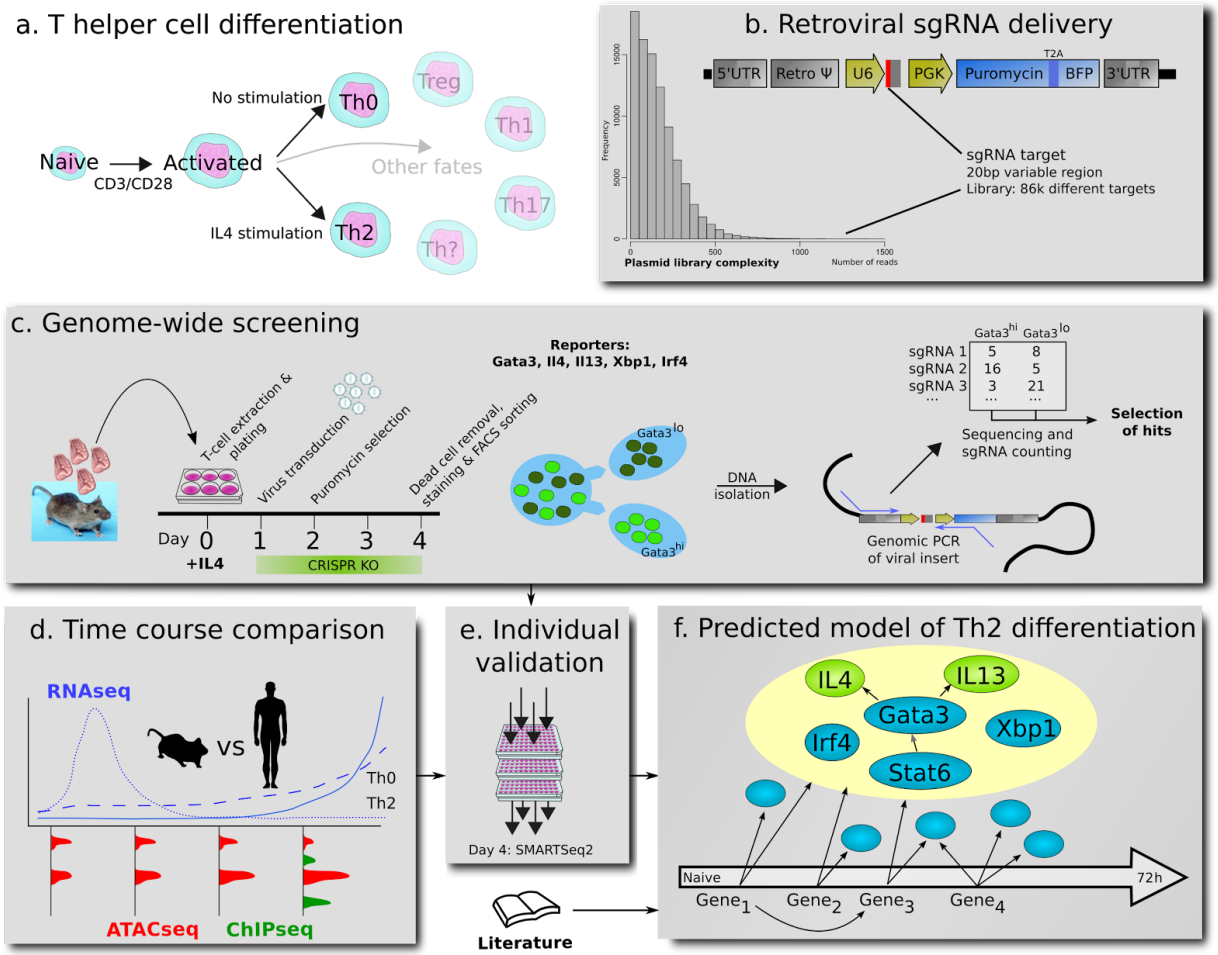

Figure 1: Overview of the experimental KO screening strategy. (a) In our culture system, naive, *ex vivo* T cells are differentiated into Th2 cells by IL4. Potential alternative T cell fates that may be open to genetically perturbed cells are highlighted. *In vivo*, T cells develop into different subtypes dependent on stimuli. The stimuli are created by other immune cells, including T cells themselves. **(b)** The retrovirus is based on murine stem cell virus (MSCV) and holds one sgRNA against a gene, and allows for BFP and puro selection. For the screening we have a mixture (pool) of plasmids, containing over 86 000 sgRNAs, from all of which we produce 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. The sgRNA affecting gene expression is 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 genes ("hits") were analyzed based on their dynamics according 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.

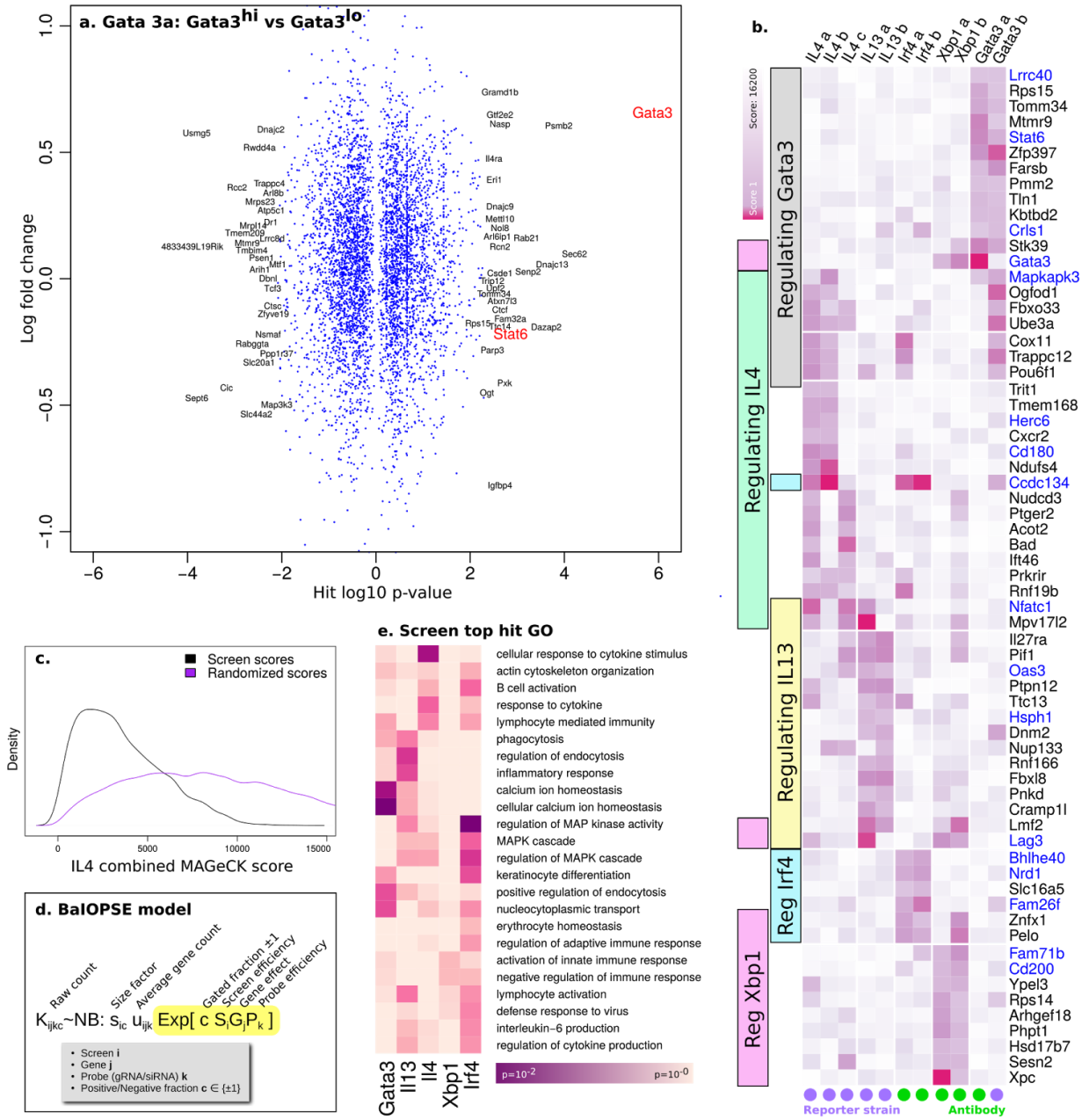

Figure 2: Results from genome-wide Th2 differentiation screen. (a) Hits from screen for Gata3 expression measured by antibody staining. The x-axis denotes the p-value for differential expression obtained by MAGeCK, the y-axis shows the fold change in expression between *Gata3*^{hi} and *Gata3*^{lo} sorted cells. Highlighted in red are *Gata3* and *Stat6*, since these are known to control *Gata3* expression. **(b)** Summary results of all 10 screens carried out. Genes that were consistent hits in multiple screens are shown (see methods for gene selection). The purple

colour show 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. **(c)** Combined screen scores for the IL4 screens vs randomized score distribution **(d)** The alternative BaIOPSE hit calling model **(e)** GO annotation of top hits for each screen as defined by BaIOPSE. The color represents Log_{10} p-value.

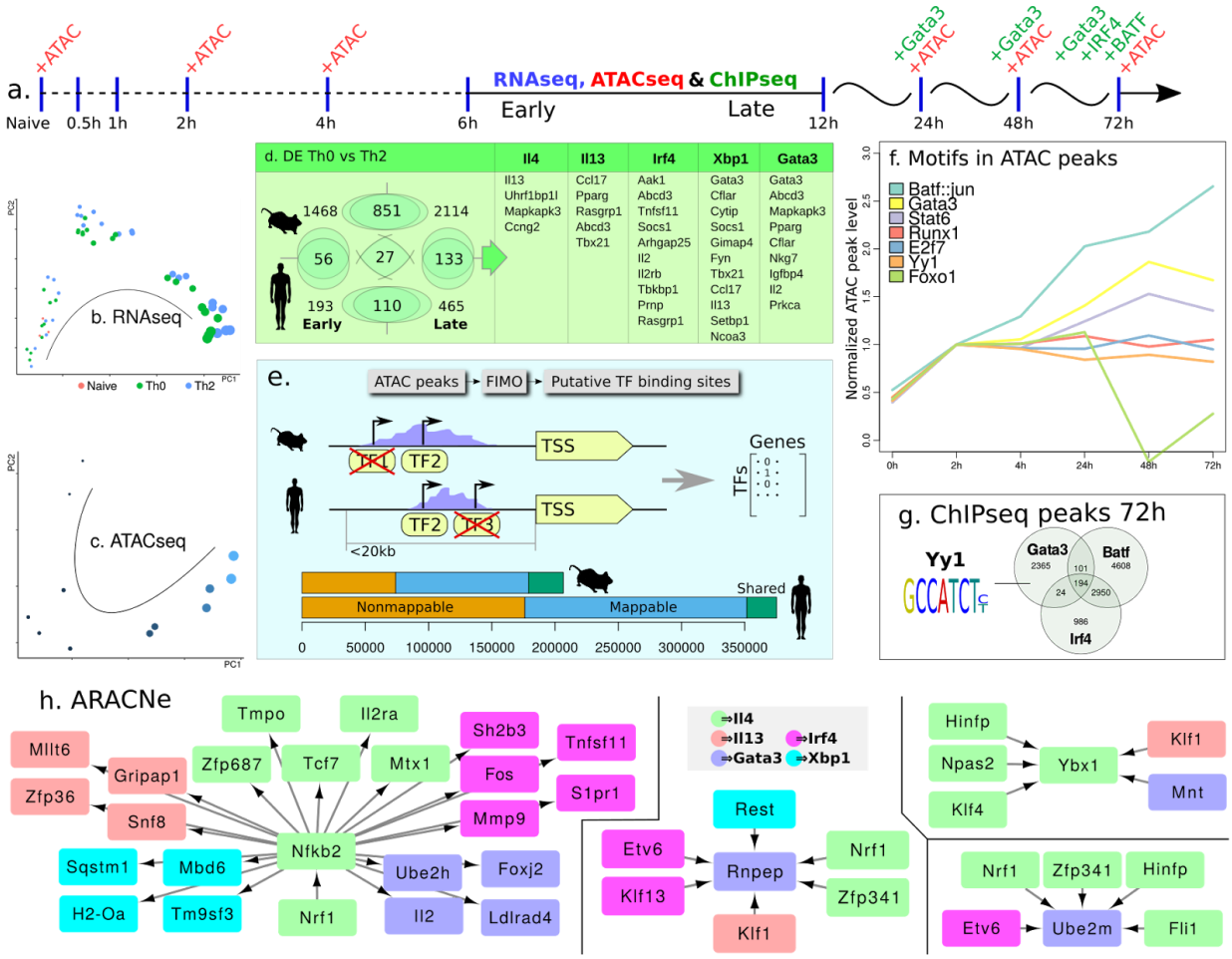

Figure 3. Assessment of hits using time course data (a) The chosen time-points for RNA-seq, ATAC-seq and ChIP-seq. **(b)** PCA projection of bulk RNA-seq and **(c)** ATAC-seq samples. The naive samples separate in the third principal component not shown. **(d)** Number of differentially expressed genes 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)** ChIP-seq overlap and the YY1 motif found within the GATA3 peaks. **(h)** Examples of regulons from the ARACNe network of time-course RNA-seq with high enrichment in hits. Hits from different genetic screens are shown in colour as indicated.

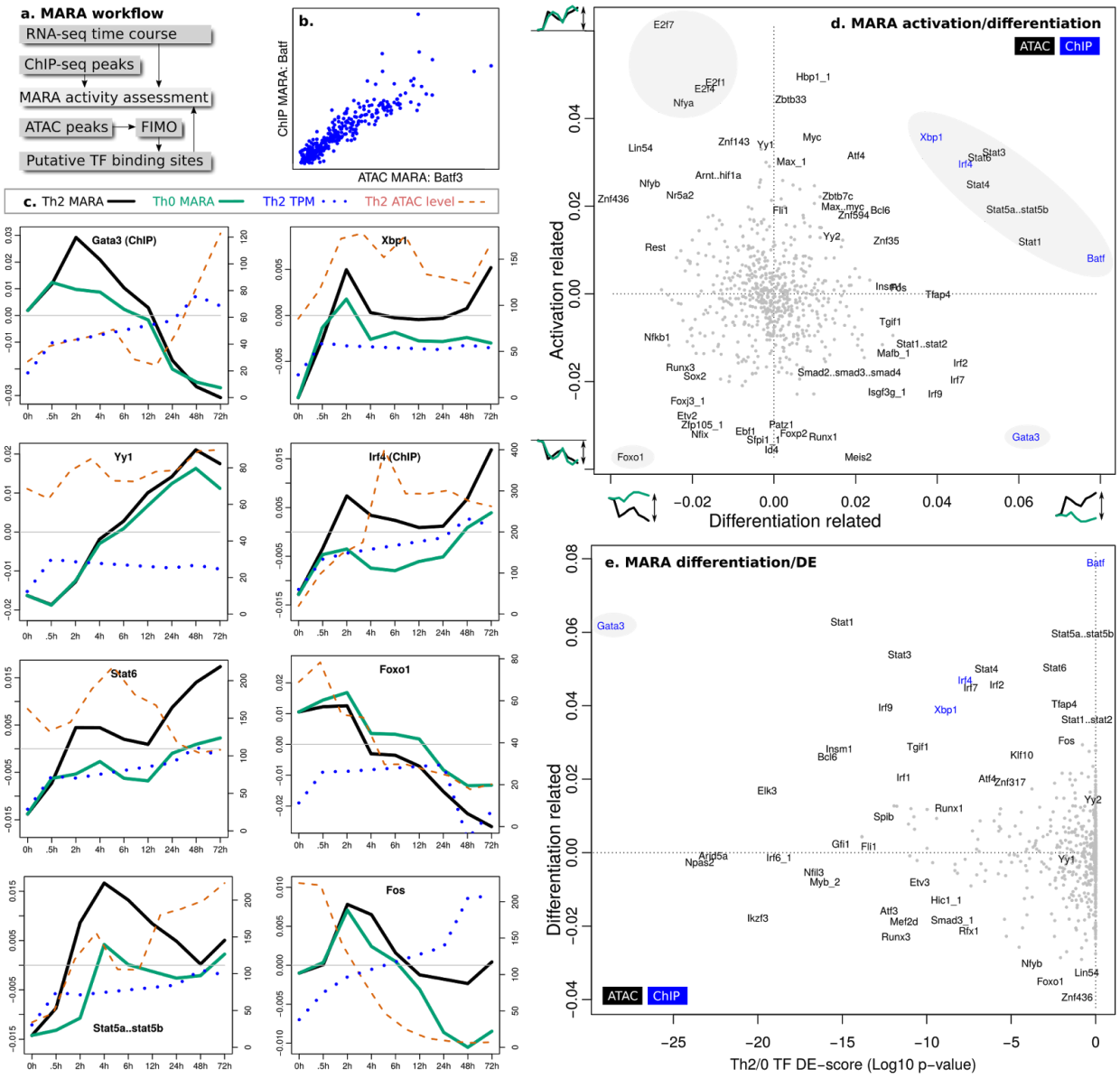

Figure 4: Analysis of transcription factor activity (a) Workflow for combining putative binding sites with time-course RNA-seq. **(b)** Comparison of BATF activity predictions for individual genes, by ATAC-seq predicted binding sites and ChIP-seq peaks. **(c)** Dynamics of selected transcription factors, comparing their expression level, activity in Th2 (black line) and Th0 (green line) and chromatin openness. **(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.

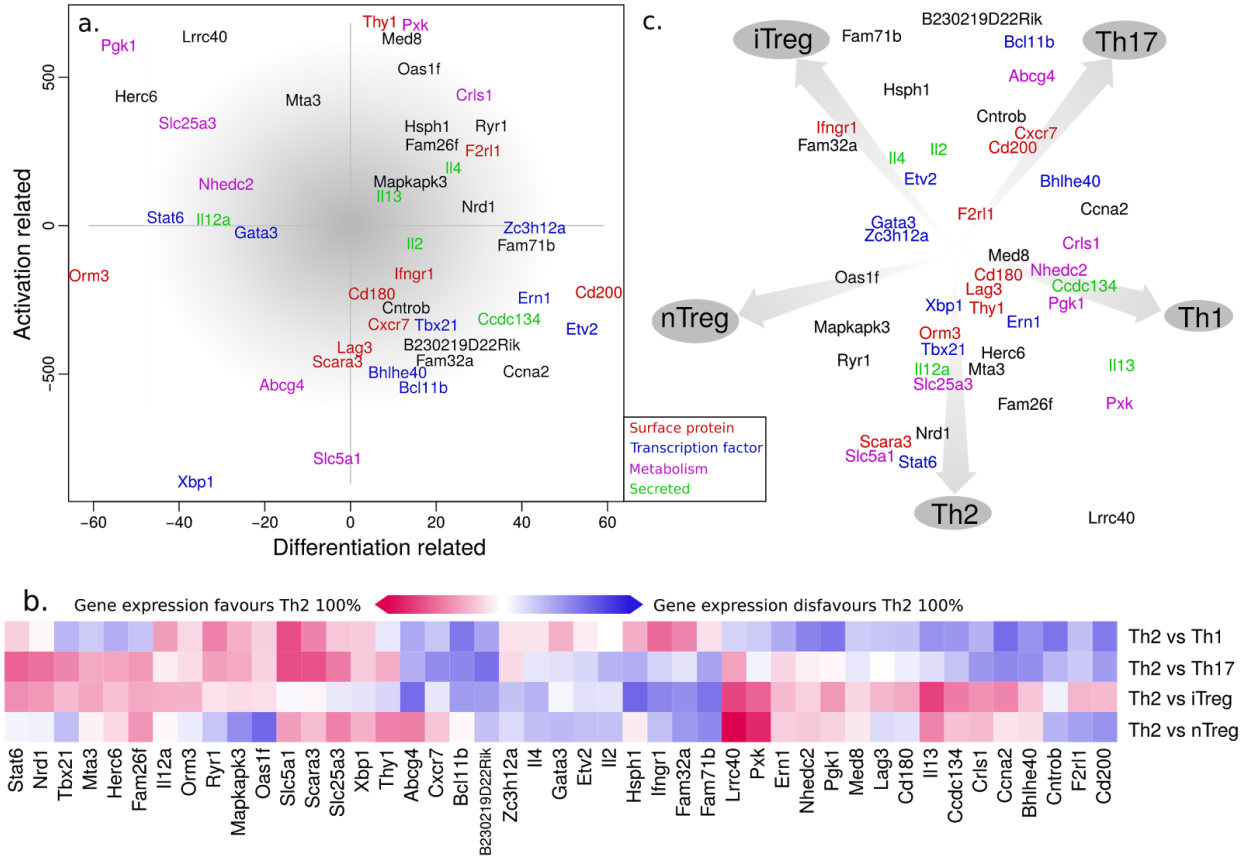

Figure 5: KO effect on T helper cell identity (a) The effect on differentiation and activation of gene KOs. DE genes are in KO to WT contrasts and compared to DE genes lists defined for activation Th2 (t=0) vs (t=72h) and Th2 vs Th0 (t=72h) (further defined in text). Genes/KOs toward the middle of the plot have the least effect. **(b)** Gene influence on the Th identity. The KO DE gene lists are compared with fate DE list, both quantitatively (% overlap, shown by colour intensity) and qualitatively (same or opposite direction, depicted in blue and red). A dimensionality reduced summary is also shown in **(c)**, where the middle point means neutral/weak phenotype.

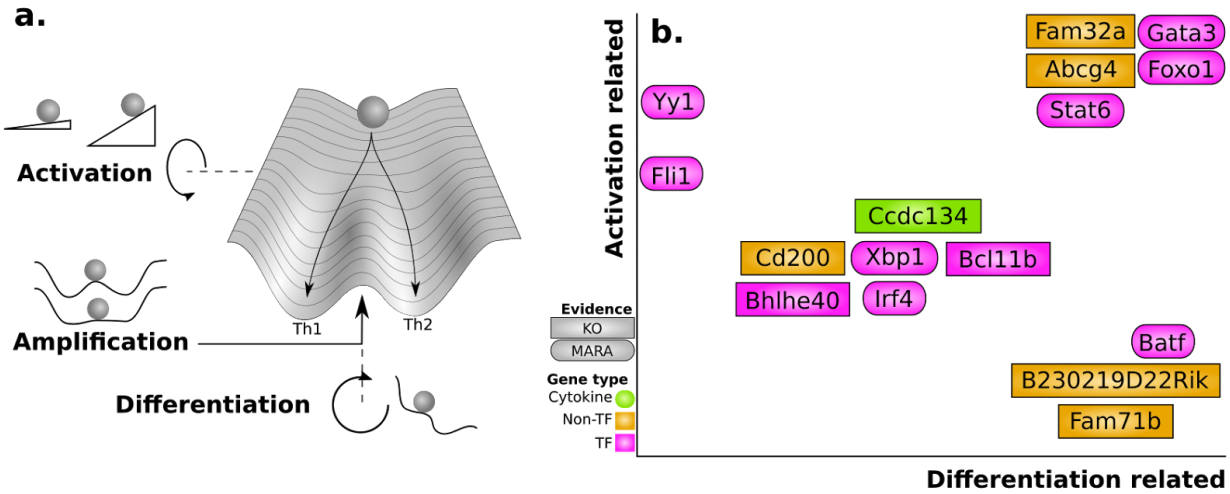

Figure 6: A summary of Th2 differentiation (a) While the genes controlling Th2 fate are from a wide range of programs, their behavior can be categorized into modes **(b)** Some of the genes according to these categories, including several unstudied genes as reported from our screens.

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.

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, the sgRNA part of a previous mouse KO sgRNA pooled library[57] (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 NEB BamHI-HF 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 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).

Validation of *Cas9* mouse

Rosa26$^{Cas9/+}$ mice[57] 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.

Expression of *Cas9* was confirmed by western blot (anti-Cas9, BioLegend 7A9, #844301) as well as 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 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 in 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” (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 over-night (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, catalog #562642) and Alexa Fluor 488 Mouse anti-GATA3 (BD Biosciences, catalog #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 (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) in a reaction with Q5 Hot Start High-Fidelity 2× Master Mix (NEB #M0494L). In each 50ul reaction we never loaded more than 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[62]. After Ampure XP bead purification (beads made up 70% of the solution) and Bioanalyzer QC, the libraries were pooled and sequenced using the iPCRtagseq custom sequencing primer on an Illumina HiSeq 2500. 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 make up 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[24] 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 $(r_1 r_2 r_3 ... r_n)$^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[63].

The BaIOPSE model was implemented in STAN[64] 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[65] 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), 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 Hiseq2000 using the v4 chemistry 125bp PE.

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[66] 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). The human RNA was then extracted and treated analogously to the mouse RNA.

Time-course ATAC-seq data generation

Experiments were done according to the published protocol[67] 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 on an Illumina Hiseq2500 platform, 75 bp, PE.

ChIP-seq data generation

ChIPmentation[68] was used to investigate the transcription factor 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.

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-Lauryolsarcosine). Chromatin was sonicated using Bioruptor Pico (Diagenode) with 30 seconds ON/30 seconds OFF for 10 cycles. 30 ul 10% Triton X-100 were 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 was 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 Hiseq2000 platform using the v4 chemistry (75 bp PE).

ChIP-seq peak analysis

The reads were first trimmed using Trimmomatic 0.36[69] with settings ILLUMINACLIP:NexteraPE-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:30. Peaks were then called using MACS2[41], merged over time, and annotated using HOMER[70].

The quality of the peaks were 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”[71], for 200, 300, 400, 500bp regions around the peaks. These were fed into MEME[72] for additional motif discovery.

Time-course RNA-seq differential expression

Gene expression from RNA-seq data was quantified in TPM using Salmon v0.6.0[73], 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[74], 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

ATAC-seq motif extraction

ATAC-seq reads were aligned using Bowtie 2[75] with the parameter –X 2000 and the mouse genome mm10. This was followed by peak calling on each replicate individually using MACS2[41] 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[76], and motifs from the JASPAR 2016 database[77] considering only those starting in MA or PB. In addition, we supplemented with a more recent list of C2H2 motifs[78]. To make the analysis more targeted, only motifs from transcription factors differentially expressed 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[79]. 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[71]. 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.

ARACNe analysis

The ARACNe-AP[47] package was used as follows. For a given network, the RNA-seq data is stored in files net.exp. The list of core genes (TFs and cytokines) is stored in net.tf. To initialize the algorithm, this command is used: “java -Xmx5G -jar Aracne.jar -e net.exp -o net.out --tfs net.tf --pvalue 1E-5 --seed 1 --calculateThreshold --nodpi”. Next, the MI calculation is done using “java -Xmx7G -jar Aracne.jar -e net.exp -o net.out --tfs net.tf --pvalue 1E-5 --seed …. --nodpi”. Finally, the data is summarized using “java -Xmx120000M -jar Aracne.jar -o net.out --tfs net.genes --consolidate --nodpi”.

Analysis of the network was performed in R. To validate the network the clustering of hits in a certain screen Z was assessed based on the number of neighbours X, Y, both being Z-hits. Using a bootstrap test, the number of such neighbours was compared to randomized networks where the identify of the vertices was swapped.

To check if a gene X is especially important for Z, enrichment for nearby Z-hits was tested. For each gene X (as a vertex), the other genes (vertices) one edge apart were checked for being Z-hits. The gene X is considered enriched for hits if it has more hits than average, tested using a binomial test.

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[80] 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.

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 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). Virus was subsequently made in 293t-cells, in 24w format. 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[81]. Libraries were made using Nextera XT and all 96*3 libraries sequenced on 2 lanes HiSeq 2500, 150PE.

Follow-up knock-out RNA-seq analysis

Reads were filtered using Cutadapt for the Smartseq-2 TSO and mapped using GSNAP[82]. The software featureCounts was then used to produce a final count table[83]. The effects of the KO was studied using an EdgeR[84] 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[29] with $p<10^{-2}$. For Fig 5ab we then calculate the similarity score as $s_i=\sum_{i\in g} ref_i\, ko_i/|g|$, where g is the set of genes which in the KO have a 2-fold change.

The plasticity dimension reduction Fig 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 scramble 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 and not absolute coordinates in the plot.