A validated gene regulatory network and GWAS identifies early regulators of T cell-associated diseases


Linköping University Post Print

N.B.: When citing this work, cite the original article.

Original Publication:
http://dx.doi.org/10.1126/scitranslmed.aad2722
Copyright: American Association for the Advancement of Science
http://www.aaas.org/

Postprint available at: Linköping University Electronic Press
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-123522
A validated gene regulatory network and GWAS identifies early regulators of T-cell associated diseases

Mika Gustafsson1,2†, Danuta R. Gawel1†, Lars Alfredsson3, Sergio Baranzini4, Janne Björkander5, Robert Blomgran6, Sandra Hellberg7, Daniel Eklund8, Jan Ernerudh7,8, Ingrid Kockum9, Aelita Konstantinell1,15, Riita Lahesmaa10, Antonio Lentini1, Robert Liljenström1, Lina Mattson1, Andreas Matussek5, Johan Mellergård11, Melissa Mendez12, Tomas Olsson9, Miguel A. Pujana13, Omid Rasool10, Jordi Serra-Musach13, Margaretha Stenmarker5, Subhash Tripathi10, Miro Viitala10, Hui Wang1,14, Huan Zhang1, Colm E. Nestor1†, Mikael Benson1†*

1The Centre for Individualised Medicine, Department of Pediatrics, Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

2Bioinformatics, Department of Physics, Chemistry, and Biology, Linköping University, Linköping, Sweden

3Institute of Environmental Medicine, Karolinska Institutet, Solna, Sweden

4Department of Neurology, University of California San Francisco San Francisco, CA, USA

5Futurum-Academy for Health and Care, County Council of Jönköping, Jönköping, Sweden

6Division of Microbiology and Molecular Medicine, Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

7Department Clinical and Experimental Medicine, Division of Clinical Immunology, Unit of Autoimmunity and Immune Regulation, Linköping University, Sweden

8Department of Clinical Immunology and Transfusion Medicine, Linköping University, Linköping, Sweden

9Department of Clinical Neurosciences, Karolinska Institutet and Centrum for Molecular Medicine, Stockholm, Sweden

10Turku Centre of Biotechnology, University of Turku and Åbo Akademi University, Turku, Finland

11Department of Neurology and Department of Clinical and Experimental Medicine, Linköping University, Sweden
One Sentence Summary: Combining a gene regulatory network and disease-association data identified early regulators of T-cell associated diseases.
Abstract

Early regulators of disease may increase understanding disease mechanisms, and serve as markers for pre-symptomatic diagnosis and treatment. However, early regulators are difficult to identify because patients generally present after they are symptomatic. We hypothesized that early regulators of T-cell associated diseases could be found by identifying upstream transcription factors (TFs) in T-cell differentiation, and by prioritizing hub TFs that were enriched for disease associated polymorphisms. A gene regulatory network (GRN) was constructed by time-series profiling of the transcriptomes and methylomes of human CD4⁺ T-cells during in vitro differentiation into four helper T-cell lineages, in combination with sequence-based TF-binding predictions. The TFs GATA3, MAF and MYB, were identified as early regulators and validated by ChIP-Seq and siRNA knockdowns. Differential mRNA expression of the TFs and their targets in T-cell associated diseases support their clinical relevance. To directly test if the TFs were altered early in disease, T cells from patients with two T-cell mediated diseases, multiple sclerosis and seasonal allergic rhinitis, were analyzed. Strikingly, the TFs were differentially expressed during asymptomatic stages of both diseases, whereas their targets showed altered expression during symptomatic stages. This analytical strategy to identify early regulators of disease by combining gene regulatory networks with GWAS may be generally applicable for functional and clinical studies of early disease development.

Data access: The microarray data from this study has been deposited under the Gene Expression Omnibus super series accession, GSE60680:

Introduction

The ability to predict and prevent disease before it becomes symptomatic could open avenues to more preventative therapies (1). Such a change would require identification of diagnostic markers for early disease detection. This is a substantial challenge in common diseases like allergy, obesity, cancer or diabetes, which evolve over years or even decades. To address this challenge, prospective studies of very large cohorts of initially healthy subjects over decades would be needed. Remarkably, such a study was recently started, with the aim is to follow one hundred thousand subjects for 20-30 years, and repeatedly analyze multiple potential diagnostic markers to predict disease (1). However, the identification of such markers is complicated by the involvement of thousands of genes in multiple cell types in different tissues and organs, which may change at different time points of the disease process. Since there is currently limited understanding of evolving disease processes, we will henceforth refer to any regulatory gene that occurs before symptomatic stages as “early”.

In this study, we hypothesized that early regulators with diagnostic potential in T-cell associated diseases can be systematically inferred by 1) constructing a gene regulatory network (GRN) of T cell differentiation, and 2) prioritizing hub transcription factors (TFs) in that GRN, which are enriched for disease-associated SNPs identified by GWAS.

The background to the hypotheses is previous studies showing that early transcription factors (TFs) can be inferred based on their predicted binding sites in the promoter regions of known disease-associated genes. Such approaches have successfully identified driver mutations in cancer and other diseases (2-4).

These studies were based on cells from patients with established diseases, or cell lines, which are
not representative of early disease processes. In the absence of samples from pre-symptomatic patients, a GRN should ideally be constructed based on time series analyses of a cellular model of an evolving disease process. Here, we focused on CD4$^+$ T-cell associated diseases and upstream transcription factors (TFs) in T-cell differentiation. T-cell differentiation has a key role in orchestrating immune responses in multiple, highly diverse diseases, including autoimmune and allergic diseases (5), atherosclerosis (6), cancer (7), and obesity (8). Thus, we hypothesized that a GRN of T cell differentiation could serve as a model of evolving disease processes in T-cell associated diseases, in order to infer early TFs with diagnostic potential for early disease detection (see Figure 1 for an outline of the study).
Results

**Early Th1/Th2 transcription factors were enriched for disease-associated SNPs**

To identify early transcription factors (TFs) of human T-cell differentiation, we performed time series expression profiling of naïve CD4+ T-cells during differentiation into four major T helper cell subsets, namely Th1, Th2, Th17 and T regulatory (Treg) cells (Fig. 1: step 1). Samples from four replicate experiments were subjected to gene expression microarray analysis at 0, 6, 24, 72, 144, and 192 hours of *in vitro* differentiation resulting in a total of 84 transcriptome profiles. This represents the most comprehensive expression dataset of human T-cell differentiation generated to date. Each time-point displayed a distinct expression profile (Fig. 2A, fig. S1) and appropriate differentiation of each subset was confirmed by measurement of signature gene expression in each subset by qPCR (fig. S1)(9). Having obtained a list of all predicted human TFs (N = 1,750) from the animal transcription factor database (AnimalTFBD) (10), we identified TFs that were differentially expressed (t-test, Benjamini Hochberg false discovery rate < 0.1) between subsets at 6 h or 24 h, hereafter referred to as *upstream* TFs (10).

We next tested for enrichment for SNPs that were identified from published GWAS data, within TFs that were differentially expressed in the T helper subsets at different time points (see supplementary methods, table S1). Interestingly, we found that the only TF set with statistically significant enrichment for disease-associated SNPs, defined as those with a genome-wide association of \( P < 10^{-5} \), were the upstream Th1/Th2 TFs; odds ratio (OR) = 2.7, Fisher exact test \( P = 1.0 \times 10^{-7} \) (Fig. 1, step 2 and Fig. 2B, fig. S2-S3).
Early TFs in T-cell associated diseases were identified by combining a GRN with GWAS data.

In order to prioritize upstream hub TFs for further investigation, we proceeded to construct the first large-scale Th1/Th2 gene regulatory network (GRN). This was constructed from the Th1/Th2 time-series expression profiling data and predicted TF binding sites (11), using the LASSO algorithm (Fig. 1: step 3) (12). In addition, we performed DNA methylation profiling of the same cells in order to exclude potential target genes whose upstream regulatory regions were methylated, and therefore less likely to be regulated by the TFs (fig. S4). The resulting GRN consisted of 6,112 predicted interactions from 378 TFs onto 1,563 mRNA targets.

The distribution of the number of TF targets in the GRN was skewed such that a few hub TFs regulated a large number of genes (Fig. 2C). Moreover, hub TFs containing at least one disease associated SNP (GWAS-TFs) had on average 3.7 times more targets than the non-GWAS TFs (Fig. 2C, bootstrap, P<3.2x10^{-12}). Remarkably, ten GWAS TFs were among the 11 TFs with most targets (Fig. 2C). In order to prioritize among those ten GWAS TFs for further study, we repeated the analyses using all nominally associated disease SNPs (P < 10^{-3}). We found that three TFs, namely GATA3, MAF and MYB, were associated with 91% of the disease SNPs among the ten TFs identified above (OR = 9.22, bootstrap P = 6.6 x 10^{-3}, Fig. 2D, fig. S5).

The T-cell GWAS-GRN was validated by combining ChIP-seq, siRNA and expression profiling

In order to directly validate the GWAS-GRN we performed chromatin immunoprecipitation of GATA3, MAF and MYB in differentiated human Th1 and Th2 cells followed by massively parallel sequencing (ChIP-Seq). These represent the first genome-wide maps of MYB and MAF binding in human cells. A minimum of 18 million aligned reads were generated per ChIP and matching control
(INPUT) samples (Fig. 3A). In complete agreement with their role as transcription factors, both MAF and MYB showed narrow and pronounced binding at the transcription start sites (TSSs) of genes (Fig. 3B). This pattern was not observed in input samples. To further validate the accuracy of the ChIP assays, biological replicates of MYB and MAF ChIP-Seq were performed in Th1 cells. Over 70% of peaks of MYB binding in replicate 1 were also present in replicate 2 (OR = 36.4, bootstrap P<10^{-5}) whereas 79% of peaks of MAF binding identified in replicate 1 were also present in replicate 2 (OR = 39.9, bootstrap P<10^{-5}), verifying the reproducibility of the ChIP-Seq assays (Fig. 3C).

Importantly, GATA3, MAF and MYB bindings with higher confidence (lower peak p-value) were significantly enriched (OR_{GATA3} = 7.30, weighted bootstrap – supplementary methods, P_{GATA3} < 10^{-81}, OR_{MAF} = 1.92, P_{MAF} < 10^{-22}, OR_{MYB} = 3.17, P_{MYB} < 10^{-27}) for the GRN predicted targets of the 3 TFs (Fig. 3D-G). This finding was also confirmed by comparisons with previously published independent GATA3 ChIP-Seq data (13) (OR = 2.91, bootstrap P< 10^{-21}, fig. S18D). We then tested if the GRN predictions which used both TFBS and time-series data yielded higher overlap than using only the predictions from TFBS source. Indeed, we found that the GRN predictions yielded generally significantly higher ChIP-Seq overlap (average OR = 1.48, Fisher combined P< 10^{-5}, supplementary methods, Fig. 3E-G, fig. S6, table S12). In further support of the accuracy of the GRN, comparisons with siRNA mediated knockdowns of GATA3 and MAF in Th2 cells followed by expression profiling (14), showed significant enrichment of GRN predicted early targets both compared to all genes (OR_{GATA3} = 2.22, bootstrap P_{GATA3} = 2.9 \times 10^{-3}, OR_{MAF} = 3.45, P_{MAF} = 4.5 \times 10^{-3}) and to TFBS (fig. S18B and C).
**GATA3, MAF, MYB and their predicted targets were differentially expressed in T-cell associated diseases**

We next sought to determine the potential clinical relevance of *GATA3*, *MAF*, and *MYB* in T-cell associated diseases. First, pathway analysis of the predicted targets of the three TFs (*GATA3*, *MAF* and *MYB*) revealed significant enrichment for several disease relevant pathways, including cell activation and differentiation (Tables S2-S3). Significant enrichment for such terms was observed for both early and late targets of the three TFs (Fig. 4A).

We directly tested if the three TFs and their predicted targets from the GRN were differentially expressed in disease using expression profiling data of CD4+ T-cells from eight T-cell associated diseases (Table S4), namely rheumatoid arthritis (RA; GEO Accession: GSE4588), multiple sclerosis (MS) (15), systemic lupus erythematosus (SLE; GEO Accession: GSE4588), lymphocytic variant of hyper eosinophilic syndrome (HES) (16), chronic lymphocytic leukemia (CLL) (17), adult T-cell leukemia/lymphoma (ATL) (18), acute myeloid leukemia (AML) (19), and seasonal allergic rhinitis (SAR) (20). We found that the three TFs were differentially expressed in six of the diseases, and most of their predicted targets in all eight diseases (Fig. 4B, supplementary methods).

We also made similar observations in original expression profiling studies of CD4+ T-cells from patients with malignant (breast cancer) and infectious diseases (tuberculosis, influenza) (Fig. 4B).

Interestingly, several disease-associated SNPs were in linkage disequilibrium with splice affecting regions (Table S5), and eQTL analyses supported that these SNPs induced alternative splicing of the three TFs (Table S1). Using RNA-Seq data we identified disease associated SNPs that correlated with differential splicing (Figure 4C, supplementary methods). Thus, in the next section, we characterized splice variant expression of *GATA3, MAF* and *MYB* in two relapsing diseases.
Splice variants of *GATA3*, *MAF* and *MYB* were differentially expressed in asymptomatic patients with SAR

Although the overall goal of our research is identification of early regulators of disease, direct detection of such regulators would require studies of numerous molecular layers, in numerous cell types, at numerous time-points in 1,000’s of healthy individuals over many years. Instead, we used the asymptomatic stage of seasonal allergic rhinitis (SAR) as a proxy for the early disease state. SAR is an optimal model of relapsing diseases, because it occurs at a defined time point each year, and due to a known external trigger (14). It is possible to model gene expression changes associated with asymptomatic and symptomatic stages, by comparing *in vitro* allergen-challenged with un-stimulated CD4+ T-cells outside of the pollen season, when patients are asymptomatic. In this model system, we propose that un-stimulated T-cells can serve as a model of the early disease state, whereas allergen-challenged cells can be used to examine if their predicted targets change in expression during the symptomatic stage (Fig. 1: step 4). We performed such analyses based on prospective clinical studies of 10 patients and 10 controls. In summary, exon profiling by microarray identified differential expression of splice variants of each of the three TFs in the early, asymptomatic stage (unstimulated T-cells), and differential expression of their predicted targets in the symptomatic stage (allergen-challenged T-cells) (fig. S7-S8). Many of the differentially expressed splice variants of *GATA3*, *MAF* and *MYB* have not been previously described in SAR. During the asymptomatic stage, *MAF* splice variant 2 (NM_001031804.2), *GATA3* splice variant 1 (NM_001002295.1), all measured splice variants of *MYB* were differentially expressed, of which variants 4 (NM_001161656.1) and 5 (NM_001161657.1) were most significant (Figure 5A). The microarray expression levels were validated by qPCR (fig. S8A - C). Interestingly, the mean expression levels of these splice variants separated patients and controls with high accuracy (area
under the precision recall curve (AUC-PR = 0.83, P=2.9 x 10^{-3}, Fig. 5B), but also significantly correlated with the symptom scores in patients during the pollen season (PCC=0.67, P=0.019, Fig. 5B). Furthermore, we replicated our findings in an independent material consisting of 14 patients and 6 controls, in which their mean expression levels also separated patients and controls with high accuracy (Fig. 5C, AUC-PR= 0.77, P=0.039).

In order to dissect the roles of the variants we measured the splice variants in Th1/Th2 differentiation by qRT-PCR and analyzed the structure of the differences of corresponding proteins (fig. S9 and S10). We then aimed at functionally validating the disease relevance of the splice variants. siRNA mediated knock-down of MAF splice variant 2, followed by Th2 polarization and exon array analysis (fig. S11-S12). MAF splice variant 2 was chosen because it was the only variant for which we could design computationally predicted and experimentally verified siRNA. We showed that 65 of the predicted 103 targets of MAF were affected by MAF splice variant 2 siRNA (FDR<0.1). Interestingly, these genes were also the MAF targets with lowest P-values in allergen-challenged T-cells from SAR patients (bootstrap P<10^{-5}, fig. S13, Table S6).

A particular advantage of SAR as a disease model is that patients can be cured by specific immunotherapy, in which a small dose of allergen is repeatedly administered. We followed 14 patients before, during (after 1 year) and after 2 years of sublingual immunotherapy. We found that this therapy resulted in reversal of expression of the analyzed splice variants to levels observed in healthy controls. We found a marginal change in the patients during treatment compared to before, AUC-PR=0.62 (P=0.18), which became significant after the 2 years treatment, both compared to the expression before and during the treatment, AUC-PR 0.85, P<2.9 x 10^{-3} and 0.95, P<1.5 x 10^{-4}, respectively (Fig. 5C).
*GATA3*, *MAF* and *MYB* were associated with MS through enrichment of SNPs, and target genes increased in expression during relapse

qRT-PCR analysis of splice variants of *GATA3*, *MAF* and *MYB* in MS in 10 un-stimulated CD4$^+$ T cell from MS patients in remission showed significant decrease of *GATA3* splice variant 2 (NM_002051.2), compared to 10 healthy controls (Wilcoxon test $P = 0.019$, Fig. 5D). Further analysis of a prospective gene expression microarray study of MS patients seen both during relapse and remission using previously unpublished information about the disease status of the patients (15) (see supplementary methods description) showed a significant decrease of the *GATA3* gene during remission compared to controls ($t$-test$_{LIMMA}$ $P=0.033$), as well as a decrease in remission compared to relapse (paired $t$-test$_{LIMMA}$ $P=1.4 \times 10^{-3}$), and significant differential expression of its predicted targets during relapse compared to controls (bootstrap $P < 10^{-12}$). The predicted targets of *MAF* and *MYB* were also differentially expressed during relapse ($P=5.3 \times 10^{-3}$ and $1.6 \times 10^{-4}$, respectively). This led us to search for MS-associated genetic variants in these two TFs, as well as in *GATA3*. We analyzed original data from the MS consortium GWAS study of $\sim$25,000 patients and controls (21). For *MAF*, *MYB*, and *GATA3* we found significant SNPs ($P_{MAF}= 1.9 \times 10^{-5}$, $P_{MYB} = 5.8 \times 10^{-6}$, and $P_{GATA3} = 6.6 \times 10^{-5}$). Indeed, the three TFs were among the 1% most enriched for disease-associated SNPs (Fisher exact $P = 3 \times 10^{-6}$). We further asked if the number of disease associated SNPs correlated with disease severity in a cohort consisting of 2,085 patients for whom previously unpublished data about disease severity was available (21). We found a marginal (7.1%), but statistically robust, enrichment (bootstrap $P=0.015$) of SNPs among the patients with severe disease (see supplementary methods for details).
Discussion

The identification of early disease regulators is important to understand disease mechanisms, as well as to find candidates for early diagnosis and treatment. Here, we characterized changes in expression and DNA methylation in CD4+ T-cells at multiple time-points during differentiation into four T-helper cell subsets. This is the most detailed and large-scale study of transcription dynamics during human T-cell differentiation to date, and an ideal system in which to generate and test gene regulatory networks (GRNs). We present an analytical strategy to identify early transcription factors (TFs) in T-cell associated diseases using the GRN of T-cell differentiation in combination with GWAS data. Briefly, the strategy identified three early hub TFs, GATA3, MAF and MYB, which were highly enriched for disease-associated polymorphisms. Both the TFs and their predicted targets were differentially expressed in T-cells from patients with symptomatic T-cell associated diseases. Exon profiling showed that splice variants of these TFs were differentially expressed during asymptomatic stages of MS and SAR, and their predicted targets during symptomatic stages. The results were validated in independent materials, as well as by ChIP-Seq and siRNA knockdowns. As further discussed below, we propose that the analytical strategy and our data can be used to systematically identify early regulators of disease.

Limitations of the study include that we only focused on TFs as early regulators. As discussed below, other types of regulators may also have pathogenic and diagnostic relevance. The construction of a TF-based GRN may be confounded by variable knowledge about which genes are regulated by different TFs. Another important limitation is that the identification of early regulators of complex disease should ideally be based on long-term prospective studies of very large cohorts, similar to the one referred to in the introduction (1). As a proxy for early disease we instead performed clinical studies of two relapsing diseases, MS and SAR. The rationale was that
asymptomatic stages of the two diseases would serve as models of early, pre-symptomatic stages. Using exon profiling and qPCR in SAR and MS, respectively, we revealed differential expression of splice variants of \textit{GATA3}, \textit{MAF} and \textit{MYB} that had not been previously described in either disease. Crystallographic and computational studies indicate that the splice variants may affect DNA binding (fig. S10 shows structures). Knockdown studies of \textit{MAF} variant 2, followed by expression profiling, revealed a significant overlap with differentially expressed genes in the symptomatic stage of T-cells from patients with SAR. Moreover, in combination, splice variants separated patients from controls with high accuracy, and correlated with patient symptom scores. Our findings are consistent with the increasingly recognized importance of splice variants in diseases; rheumatoid arthritis (22), nephropathy (23), and MS. From a diagnostic perspective, our findings highlight the importance of searching for combinations of different types of variables. For example, our analyses of large-scale GWAS of MS showed highly significant enrichment of disease-associated SNPs in the three TFs, which in combination were associated with disease severity. From the perspective of predictive and preventative medicine, these findings suggest an important direction for future research: to examine if combinations of different variables, such as splice variants and SNPs, can be used to predict specific diseases (24).

That specificity might be increased by extending the GRN to include other potential early genetic or epigenetic regulators, like signaling molecules, non-coding RNAs or histone modifiers, all of which can be prioritized based on their number of predicted targets and GWAS. The power of combining disease-associated data from different genomic layers has recently been described (25). T cells may be ideal for such studies because they constantly patrol all parts of the body. They are, either primarily or secondarily, involved in allergic, autoimmune, infectious or malignant diseases. From a translational perspective, early regulators and their targets in T cells could therefore have
substantial diagnostic and therapeutic potential. Moreover, our strategy can be applied to any disease where the affected cell type is known and for which a GRN can be constructed. For example, a GRN can potentially be inferred based on *in vitro* derivation of epithelial cells, and validated by functional experiments in primary or transformed epithelial cell lines. Potential early regulators can be identified based on the prioritization principles described above, and examined in epithelial cells from patients with, or at risk for, dermatological diseases, like eczema or psoriasis. From a clinical perspective, identification of early regulators, or their down-stream targets, whose protein products can be measured in body fluids would be optimal. If so, combinations of proteins that reflect different organs and diseases could be repeatedly monitored, or used for differential diagnosis of identified disease processes.

In summary, we propose that our strategy and GRN can be applied to identify early regulators in T-cell associated diseases, and have made the GRN and analytical tools available for this purpose. Moreover, further studies are warranted to test if the principles are generally applicable to other cell types and diseases.
Material and Methods

Study design

The overall objective of this study was to identify early regulators of T cell-associated diseases by identifying upstream transcription factors (TFs) in T-cell differentiation, and by prioritizing hub TFs that were enriched for disease associated polymorphisms. Upstream TFs were identified by constructing a gene regulatory network (GRN) of T cell differentiation, based on time-series profiling of the transcriptomes and methylomes of human CD4+ T-cells during in vitro differentiation into four helper T-cell lineages, in combination with sequence-based TF-binding predictions. The GRN was validated by ChIP-Seq and siRNA knockdowns. The TFs were validated by differential mRNA expression of the TFs and their targets in active states of several T-cell associated diseases. To directly test if the TFs were altered early in disease, T cells from patients with two T-cell mediated diseases, multiple sclerosis (MS) and seasonal allergic rhinitis (SAR), were analyzed during asymptomatic and symptomatic stages of both diseases. Sample sizes were determined based on our previous studies (32), and replicated in independent studies of MS and SAR, as detailed below. The analyses were not blinded, nor randomized. All studies were approved by the ethics board of Universities of Gothenburg, Lima and Linköping.

Statistical Analysis

Gene expression microarray data were quantile-normalized, and differentially expressed genes were for time-series data determined using maSigPro (28), and for comparisons between two states by using t-test from the LIMMA package in R, with 10% false discovery rate according to the Benjamini Hochberg method (P<0.05). The P-values for qPCR was calculated using one-sided non-parametric Wilcoxon test. In several instances we tested the enrichment of low P-values within a set of genes (e.g. target genes of a TF). This was performed using a bootstrap test
on the average log10 P-value of the target set, where P-values were estimated from 10^6 randomizations using (29). For set enrichment analysis (e.g., GWAS-genes) P-values were calculated using one-sided Fisher exact test using all genes as background (n=22,500).
Disclosure declaration

The authors declare no conflicts of interest.

List of Supplementary Materials

Supplementary methods:

- Study subjects
- In vitro polarization of human Th1, Th2, Th17 and Treg cells
- Culturing of peripheral blood mononuclear cells (PBMCs) from patients with seasonal allergic rhinitis (SAR)
- Sublingual immunotherapy (SLIT)
- siRNA mediated gene knock down
- Th1/Th2 regulatory region analysis
- Quantitative PCR (qPCR)
- Culturing of CD4+ T-cells from patients with MS
- Public T-cell disease data
- Th1/Th2 regulatory region analysis
- Bioinformatics analyses
- Chip-Seq analyses
- SNP analyses
- Construction of the Th1/Th2 GRN
- Genome-wide network inference.
- Network reconstruction details.
- MS gene expression in relapse and remission.
• GWAS analysis of multiple sclerosis.
• Classification of patients/controls based on machine learning.
• Gene Regulatory Network and TFBS comparison using ChIP-Seq peak counts

**Supplementary figures:**

• Figure S1. Expression of signature genes in T cell subsets.
• Figure S2. Comparison of GWAS enrichment for gene sets defined from alternative strategies defined by the T-cell differentiation microarray data.
• Figure S3. GWAS SNPs in different populations.
• Figure S4. Methylation probe levels of selected house-keeping and tissue specific genes.
• Figure S5. Enrichment in GWAS SNPs in GATA3, MAF and MYB in T cell diseases, infectious diseases and malignancies.
• Figure S6. Overlap between different methods to define GATA3 bindings.
• Figure S7. GRN-predicted targets of GATA3, MAF and MYB are differentially expressed in 6 T cell related diseases.
• Figure S8. Validation of the differentially expressed splice variants in SAR
• Figure S9. Time series analyses of GATA3, MAF and MYB splice variant expression during Th1 and Th2 cell differentiation.
• Figure S10. Predicted effects of alternative splicing on the protein structures of GATA3, MAF and MYB.
• Figure S11. Heatmaps of the top 50 most differentially expressed genes in siRNA mediated MAF knockdown using 4 different vectors.
• Figure S12. Effect of MAF knockdown on Th1, Th2, Th17 and Treg signature genes.
• Figure S13. MAF siRNA splice specific targets.
• Figure S14. Expression profiles of SAR patients and controls show a consisting overlap even for the most significantly differentially expressed genes.
• Figure S15. Expression profiles of MS patients and controls show a consisting overlap even for the most significantly differentially expressed genes.
• Figure S16. Expression overlap between patients and controls in many T cell associated diseases.
• Figure S17. Expression overlap between patients and controls in many T cell associated diseases.
• Figure S18. GRN validation.

Supplementary tables

• Table S1. Predicted regulatory SNPs for GATA3, MAF and MYB using rSNPBase.
• Table S2. All pathways for the targets of GATA3, MAF and MYB respectively reported by IPA.
• Table S3. All pathways for the early predicted targets of all early TFs.
• Table S4. Statistics of all 8 T cell diseases.
• Table S5. SNPs that might affect splicing processes were identified by mapping all disease-associated SNPs (GWAS) to the splice regulatory sites (SRS) of MAF, MYB and GATA3.
• Table S6. Gene ontology enrichment analysis of biological processes of the MAF splice variant 2 specific targets.
• Table S7. Results for the naive questions "how many PPIs have the Th differentially expressed genes, globally or just early", and "would had been identified the TFs through this PPI analysis?"

• Table S8. Splice variant qPCR details (reagents and company).

• Table S9. The list of all genes with at least one disease associated SNP, identified by GWAS.

• Table S10. MAF knockdown efficiency.

• Table S11. Results of the patient control classification using GATA3, MAF and/or MYB or their targets expression values.

• Table S12. GRN predictions compared with TFBS using ChIP-Seq peak counts.

• Table S13. Source data values for figure 5A.

• Table S14. Source data values for figure 5B.

• Table S15. Source data values for figure 5C.

• Table S16. Source data values for figure 5D.

Supplementary data

• Data S1. Five methylation arrays covering all known enhancers based on previous publications using DNase-seq.

• Data S2. A validated network, Cytoscape file
References


Elovaara, F. Esposito, N. Evangelou, B. Fiddes, J. Field, A. Franke, C. Freeman, I. Y.
Frohlich, D. Galimberti, C. Gieger, P. A. Gourraud, C. Graetz, A. Graham, V. Grummel,
C. Guaschino, A. Hadjixenofontos, H. Hakonarson, C. Halfpenny, G. Hall, P. Hall, A.
Hamsten, J. Harley, T. Harrower, C. Hawkins, G. Hellenthal, C. Hillier, J. Hobart, M.
Hoshi, S. E. Hunt, M. Jagodic, I. Jelčić, A. Jochim, B. Kendall, A. Kermode, T. Kilpatrick,
K. Koivisto, I. Konidari, T. Korn, H. Kronsbein, C. Langford, M. Larsson, M. Lathrop, C.
Lebrun-Frenay, J. Lechner-Scott, M. H. Lee, M. A. Leone, V. Leppä, G. Liberatore, B. A.
Manrique, R. Martin, V. Martinelli, D. Mason, G. Mazibrada, C. McCabe, I. L. Mero, J.
Mescheriakova, L. Moutsianas, K. M. Myhr, G. Nagels, R. Nicholas, P. Nilsson, F. Piehl,
M. Pirinen, S. E. Price, H. Quach, M. Reunanen, W. Robberecht, N. P. Robertson, M.
Schaefer, S. Shaunak, L. Shen, S. Shields, V. Siffrin, M. Slee, P. S. Sorensen, M. Sorosina,
M. Sospedra, A. Spurkland, A. Strange, E. Sundqvist, V. Thijs, J. Thorpe, A. Ticca, P.
Wilson, J. Winkelmann, J. Zajicek, E. Zindler, J. L. Haines, M. A. Pericak-Vance, A. J.
(IIBDGC), Analysis of immune-related loci identifies 48 new susceptibility variants for

Achiti, C. Confavreux, M. Coustans, E. le Page, G. Edan, G. V. McDonnell, S. Hawkins,
Zipp, B. Mitterski, J. T. Epplen, A. Oturai, P. S. Sorensen, E. G. Celius, N. T. Lara, X.


Acknowledgements: We thank the International Multiple Sclerosis Genetics Consortium for providing GWAS data, and J. Nedergaard Larsen (ALK-Abello) for providing allergen extract.

Figures

Figure 1. Schematic illustration of the experimental approach. (1) Naïve T-cells were polarized into four different subsets and analyzed with mRNA- and DNA methylation microarrays at different time-points. (2) Upstream transcription factors (TFs) that were differentially expressed (DE) at 6h and 24h between Th1 and Th2 subsets were enriched for disease associated SNPs from genome-wide association studies (GWAS). (3) A gene regulatory network (GRN) for Th1 and Th2 polarization was inferred by combining the microarray data from (1) with predicted transcription factor bindings (TFBS) using sparse penalized regression (LASSO), (3a) Using GWAS and the GRN, GATA3, MAF, and MYB were identified as putative early regulators in T-cell diseases. (3b) The inferred edges from GATA3, MAF, and MYB in the GRN were validated by ChIP-Seq and siRNA. (3c) Analysis of T-cells from T-cell associated diseases showed that both the TFs and their targets were differentially expressed. (4) The potential of the 3 TFs to predict disease was demonstrated by analyzing asymptomatic patients with two relapsing diseases, multiple sclerosis (MS) and seasonal allergic rhinitis (SAR). Splice variants of the 3 TFs were DE in asymptomatic cells and their targets DE in symptomatic cells.
Figure 2. Time-series analysis of T-cell differentiation identified early hub transcription factors (TFs) enriched for disease associated SNPs identified by GWAS. (A) Heatmap of the expression of subset-specific genes from (9), reveals appropriate expression of Th lineage-markers in polarized cells. (B) Enrichment of genes with disease associated SNPs identified by GWAS from differential expression (DE) analysis of time-series microarrays (P<0.05 and Benjamini Hochberg FDR<0.1). The subsets derived from individual T-cell subsets (‘Th1’, ‘Th2’, ‘Th17’, and ‘Treg’) are based on the DE genes between the 6h and 24h within each cell polarization. The P-values were calculated using Fisher’s exact test relative to the background level. (**** P < 1.0 x10^{-12}, *** P < 1.0 x10^{-6}). Color labeling denotes the number of disease associations for a gene. (C) Hub TFs with the highest number of targets were highly enriched for disease associated SNPs identified by GWAS. Distribution (reverse cumulative) of the number of targets for each TF, divided into TFs that were GWAS (blue), and non-GWAS (red). The GWAS TFs had 12.6 mean number of targets (<k>), and the non-GWAS had 3.6 targets on average (P<3.2x10^{-12}). Gene names are displayed for the first 11 upstream TFs in Th1/2 differentiation. (D) GATA3, MAF, and MYB, which were differentially expressed in Th1/Th2 cells were most enriched for nominal disease associated SNPs identified by GWAS, (OR= 9.22, P= 0.0066). The black solid line represents the background level of all disease-associated genes.
Figure 3. ChIP-Sequencing of MAF, MYB, and GATA3 in differentiated human Th1 and Th2 cells validates the GWAS-GRN. (A) The number of 50 bp single end reads successfully aligned to the human genome (hg19) for each ChIP assay. Matching INPUT samples were sequenced for each IP. (B) Heatmaps showing enrichment of MYB and MAF across gene bodies. Each row represents a gene, ordered from top to bottom by total enrichment levels. (C) Genomic profiles of MAF (red) and MYB (blue) enrichment reveal high levels of concordance between biological replicates. (D) MAF (red) and MYB (blue) are enriched at genes predicted to be their targets by our gene regulatory network approach. Input (grey), MAF (red) and MYB (blue). (E-G) Enriched binding for GATA3 (E), MAF (F) and MYB (G) predicted targets in Th1 and Th2 cells from ChIP-Seq analysis. The vertical axis represents the fraction of targets with a minimum ChIP-Seq count using the inferred targets (upper black curve), the predicted TF-DNA bindings (TFBS, middle blue curve), and all genes (lower red curve). The inferred targets were enriched for ChIP-Seq counts and the mean ChIP-Seq counts were highest for the inferred targets in genes ($P_{GATA3}=1.80 \times 10^{-6}$, $OR_{GATA3}=1.62$, $P_{MYB}=0.01$, $OR_{MYB}=1.29$).
Figure 4. Predicted targets of *GATA3*, *MAF*, and *MYB* are differentially expressed in T-cell associated diseases. (A) Gene ontology (GO) enrichment analysis of predicted targets of *GATA3*, *MAF*, and *MYB* from the Th1/Th2 cell GRN at the early stage and all late stage genes. (B) Inferred targets of *GATA3*, *MAF* and *MYB* have in general lower P-values than non-target genes in 9 T-cell related diseases: lymphocytic variant of hyper eosinophilic syndrome (HES), adult T-cell leukemia/lymphoma (ATL), acute myeloid leukemia (AML), systemic lupus erythematosus (SLE), multiple sclerosis (MS), seasonal allergic rhinitis (SAR), influenza (IZ), breast cancer (BC), tuberculosis (TB). (Lower) Arrows depicting log2 fold changes (lg2FC) of the expression of each TF in patients compared to controls. Magnitudes of the arrows are depicted as: largest (abs lg2FC>2), middle (abs 1<lg2FC<2), smallest (abs lg2FC<1). Red up-facing, and blue down-facing arrows depict lg2FC>0 and lg2FC<0 respectively. (Upper) Bars mark the difference in the mean p-values for the targets of each TF compared to all genes (* P<0.05, ** P < 0.01; *** P < 0.0001 from bootstrap test). (C) eQTL analysis of the exons of *GATA3*. Normalized RNA-Seq counts across *GATA3* exons. Counts of each exon are subdivided according the genotypes of the variant rs501764. Exons showing significant differential expression across these genotypes are marked by asterisks (** P < 0.01; *** P < 0.001).
Figure 5. Splice variants of *GATA3*, *MAF*, and *MYB* are differentially expressed in asymptomatic patients. (A) Splice variant expression in unchallenged cells of 10 SAR patients (SAR) and 10 healthy controls (HC) measured by exon microarrays. (B) Normalized sum of the differentially expressed variants of *GATA3*, *MAF*, and *MYB*. This separated patients and controls with high accuracy (P = 0.014), and correlated highly with the symptom severity score. Circles represent individuals: healthy subjects are marked with green color, patients are marked to indicate symptom severity. Yellow color indicates low, orange - medium and read - high symptom severity. (C) Normalized average splice expression of *GATA3.1, MAF.2, MYB.4*, and *MYB.5*, which was differentially expressed in an independent material of SAR, and gradually decreased after one and two years of immunotherapy (T). (D) Expression of *GATA3.2* in asymptomatic MS patients and healthy controls. P-values for microarrays were calculated using limma t-test (A), and for qPCR measurements (C-D) using non-parametric Wilcoxon test. The box-plots depict median (horizontal line), inner quartile range (boxes), whiskers (dashed line), circles represent individuals.
(1) Times-series analysis of T cell differentiation mRNA and DNA methylation microarrays (n=96)

(2) Early Th1/2 DE TFs were enriched for GWAS-genes

(3) GRN and GWAS identified upstream TFs
   a. time-series profiles + TFBS + LASSO
   b. validation: Chip-seq + siRNA
   c. TFs + targets DE in 6/8 diseases

(4) Splice variants of the three TFs were DE in asymptomatic cells, and their targets DE in symptomatic cells
# sequencing reads per ChIP

<table>
<thead>
<tr>
<th>Gene</th>
<th>Th1</th>
<th>Th2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYB</td>
<td>40,513,717</td>
<td>26,098,522</td>
</tr>
<tr>
<td>MYB</td>
<td>30,791,938</td>
<td>-</td>
</tr>
<tr>
<td>MAF</td>
<td>18,082,806</td>
<td>37,480,567</td>
</tr>
<tr>
<td>MAF</td>
<td>31,401,396</td>
<td>-</td>
</tr>
<tr>
<td>GATA3</td>
<td>19,406,222</td>
<td>30,049,194</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>Input</th>
<th>MAF (Th1)</th>
<th>MYB (Th1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.06</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.7</td>
</tr>
</tbody>
</table>

C

<table>
<thead>
<tr>
<th>MAF (Th1 - replicate 1)</th>
<th>MAF (Th1 - replicate 2)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>MYB (Th1 - replicate 1)</th>
<th>MYB (Th1 - replicate 2)</th>
</tr>
</thead>
</table>

D

<table>
<thead>
<tr>
<th>MAF</th>
<th>MYB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

E

- GRN targets
- TFBS targets
- All genes

F

- GRN targets
- TFBS targets
- All genes

G

- GRN targets
- TFBS targets
- All genes
Supplementary materials

Supplementary methods

Study subjects
The study comprised seven independent materials: 1) Th1, Th2, Th17 and regulatory T polarized cells derived from freshly isolated naïve CD4+ T-cells collected from four healthy donors; 2) CD4+ T-cells from 10 SAR patients (six women, mean age 24.5 ± SEM 2.2 years) and 10 healthy donors (five men, mean age 21.7 ± 1.9 years); 3) CD4+ T-cells from 14 SAR patients (11 women, mean age 39.1 ± SEM 2.9 years) before and during two years of treatment with immunotherapy, and 6 healthy donors (three men, mean age 35.0 ± 0.9 years); 4) Eight patients with symptomatic influenza A or B (four women, mean age 66.9 years. 5) Ten patients with active pulmonary tuberculosis (6 women, mean age 32.5), and six, matched healthy controls (3 women, mean age 36.3) from Lima, Peru. 6) Eight breast cancer patients. The mean age of the patients was 74.4 years, range 67 to 82 years. 7) CD4+ T-cells from 10 MS patients (all women, mean age 42.7 ± SEM 2.7 years) and 10 healthy donors (all women, mean age 42.6 ± 2.8 years). The SAR patients and healthy donors were of Swedish origin and recruited at The Queen Silvia Children’s Hospital, Gothenburg, or Ryhov Hospital, Jönköping. SAR was defined by a positive SAR history and a positive skin prick test or by a positive ImmunoCap Rapid (Phadia, Uppsala, Sweden) to birch and/or grass pollen. Patients with perennial symptoms or asthma were not included. The healthy subjects did not have any history of SAR and had negative ImmunoCap Rapid tests. Influenza A or B was verified by PCR analysis of nasal secretions. Tuberculosis was diagnosed by sputum acid-fast bacilli smear-positivity and confirmed by Microscopic-Observation Drug-Susceptibility (MODS) assay. Breast cancer patients were sampled before radical surgery, i.e. every participating patient had an untreated malignant disease. All MS patients were diagnosed with definite relapsing-
remitting MS according to the McDonald criteria and recruited at the Department of Neurology at Linköping University Hospital. The median Expanded Disability Status Scale score was 3 (range 0-4) and the median Multiple Sclerosis Severity Score was 2.7 (range 0.21-4.95). None of the MS patients had received corticosteroids or immuno-modulatory treatment for at least two months prior to sampling and had been relapse-free for at least three months at the time of blood collection. The healthy donors were recruited amongst blood donors visiting the Blood Centre at Linköping University Hospital. The microarray data from this study has been deposited under the Gene Expression Omnibus SuperSeries GSE60680.

**In vitro polarization of human Th1, Th2, Th17 and Treg cells**

Naïve CD4+ T-cells were isolated from healthy donor buffy coats (n=4) using a naïve CD4+ T-cell isolation kit (Miltenyi). Naïve CD4+ T-cells were stimulated with plate-bound anti-CD3 (500 ng/mL) and soluble anti-CD28 (500 ng/mL) in the presence of IL-12 (5 ng/mL), IL-2 (10 ng/mL) and anti–IL-4 (5 µg/mL) for Th1, IL-4 (10 ng/mL), IL-2 (10 ng/mL), anti-IL-12 (5 µg/mL) and anti–IFN-g (5 µg/mL) for Th2, TGF-β (1 ng/mL), IL-1β (10 ng/mL), IL-6 (25 ng/mL), IL-21 (25 ng/mL) and IL-23 (25 ng/mL) for Th17, and TGF-β (10 ng/mL) and IL-2 (10 ng/mL) for Treg. Cells were cultured for six days in Iscove’s modified Dulbecco medium (IMDM) supplemented with 2 mM L-glutamine (PAA Laboratories), 10% heat-inactivated FCS (PAA Laboratories), 5 µM β–mercaptoethanol (Sigma-Aldrich) and 50 µg/mL gentamicin (Sigma-Aldrich) then re-stimulated with plate-bound anti-CD3 and soluble anti-CD28 in the presence of corresponding polarizing cytokines and antibodies for another 2 days (9). RNA was extracted using a miRneasy Mini Kit (Qiagen) and RNA concentrations were quantified with a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies). For the gene expression microarray analysis, cRNA was prepared using a Low Input QuickAmp Labeling Kit (Agilent Technologies). For Th1 and Th2 cells the gene
expression microarray analysis was performed using SurePrint G3 Human Gene Expression 8x60K microarray kit and for Th17 and Treg cells the gene expression microarray analysis was performed using SurePrint G3 Human Gene Expression 8x60K v2 microarray kit, according to the manufacture’s instruction (Agilent Technologies).

**Culturing of peripheral blood mononuclear cells (PBMCs) from patients with seasonal allergic rhinitis (SAR)**

Challenges with allergen were performed as previously described (26, 27). Briefly, PBMCs were prepared from fresh blood from 10 patients with SAR and 10 healthy controls using Lymphoprep (Axis-Shield PoC) according to the manufacturer’s protocol. PBMCs were stimulated with allergen extract (ALK-Abelló A/S; 100 μg/ml) or diluent (PBS) in RPMI 1640 supplemented with 2 mM L-glutamine (PAA Laboratories), 5% human AB serum (Lonza), 5 μM β–mercaptoethanol (Sigma-Aldrich) and 50 μg/mL gentamicin (Sigma-Aldrich). After 17 hours of incubation, total CD4+ T-cells were enriched from PBMCs by magnetic sorting (MACS). Total RNA was extracted using a miRneasy Mini Kit (Qiagen), cRNA was prepared using a Low Input QuickAmp Labeling Kit and the expression microarrays were performed using Agilent SurePrint G3 Human Exon 4x180K Microarrays according to the manufacturer's instructions.

**Sublingual immunotherapy (SLIT)**

Pre-treatment samples were collected in the first year, immediately prior to initiation of SLIT. The treatment for birch allergy included a 11-day progression of doses phase for Staloral® (Stallergenes SA), during which the extract was given from 10 to finally 300 IR/ml. Patients were instructed to keep the drops under the tongue for 2 min before swallowing them. When the 240-IR dose was reached (8 drops from the 300-IR/ml bottle), the patients continued on this level for 5-6 months,
followed by 6-months with no treatment. The post-treatment samples were collected after one and two years of treatment. All patients responded favorably to treatment.

**siRNA mediated gene knock down**

CD4+ naïve T (NT) cells were isolated as defined above (n=2) then transfected with 600 nM siRNA targeting MAF (J-003746-07, J-003746-08, J-003746-09 or J-003746-10) or non-targeting siRNA (ThermoFisherScientific Inc) by electroporation with the Amaxa nucleofection program U-014 using the Human T-cell Nuclefector Kit (Lonza). Five hours after transfection, cells were washed, activated and polarized towards Th2 for 24 hours (see above). RNA was extracted and gene expression microarray analyses were performed using Agilent SurePrint G3 Human Exon 4x180K Microarrays according to the manufacturer's instructions. Knockdown efficiency of MAF was calculated for all 4 vectors and for both splice variants. The best results were achieved with vector J-003746-09. Knockdown efficiency of MAF splice variant 2 is 38.5%, MAF splice variant 1 is 1.7%. Therefore for further analyses we have selected the experiment where the knockdown was performed with J-003746-09 vector. We performed correlation analysis from all variants of the 103 targets from the GRN and MAF variant 2 over all the 6 exon arrays, and associated the targets with at least one significant correlation with MAF variant 2 (P<0.05, FDR<0.1).

**Th1/Th2 regulatory region analysis**

We first downloaded all any DNase peak in Th1 and Th2 cells from Encode database. We then designed methylation microarrays (Agilent Technologies) within the accessible DNA regions (in total 79,789 enhancers) and analyzed the methylation expression of naïve T-cells (NT), and 7 days polarized Th1, and Th2 cells. The data were quantile normalized, each regulatory were mapped the upstream gene (n=16,969). Then the distribution of all active regulatory regions was used to define if a regulatory region was active: an enhancer was non-active if its methylation probe level
was more than two standard deviations below the mean probe level (Z<-2), see Figure S4 for more information. We defined a gene as possibly Th1/Th2 accessible if it had any active regulatory region in NT, Th1 or Th2 cells. To validate GRN the GATA3 ChIP-Seq binding sties in polarized Th1 and Th2 cells was analyzed from publicly available Supplementary material (13).

**Quantitative PCR (qPCR)**

For quantification of mRNA levels, total RNA was extracted (see above) and cDNA synthesis was performed using the High Capacity cDNA Reverse Transcriptase kit (Applied Biosystems, Life Technologies), according to the manufacturer’s instructions. Real-time PCR was performed on an ABI 7900HT Real-Time PCR cycler using the Taqman Gene Expression Master Mix and Taqman Universal Master Mix II, no UNG (Applied Biosystems). Probes were obtained from (Applied Biosystems) and (Integrated DNA Technologies) (Table S8). qPCR data was analyzed using the comparative ΔCt method with ACTB as an internal control.

**Culturing of CD4+ T-cells from patients with MS**

PBMCs were enriched from venous blood from MS patients and healthy controls using Lymphoprep (Axis-Shield). PBMCs were washed, resuspended in freezing medium consisting of 10% dimethyl sulfoxide (DMSO; Sigma-Aldrich), 50% heat-inactivated fetal bovine serum (FBS; Sigma-Aldrich) and 40% IMDM (Invitrogen; Paisley) supplemented with L-glutamine (292 mg/l; Sigma-Aldrich), sodium bicarbonate (3.024 g/l; Sigma-Aldrich), penicillin (50 IE/ml; Cambrex), streptomycin (50 µg/ml; Cambrex), and 100× non-essential amino acids (10 ml/l; Gibco BRL) and placed in a freezing container at -70°C and subsequently transferred to liquid nitrogen before further separation. The mononuclear cells were thawed with a mean viability of 95% ±0.03 SD as assessed by trypan blue staining and CD4+ T-cells were isolated by magnetic separation (Miltenyi Biotec) and cultured in IMDM with 5% FBS (HyClone®, Thermo Scientific) for 24 hours at 37°C.
with 5% CO$_2$. Total RNA was extracted using the Allprep DNA/RNA Micro kit (Qiagen) according to the manufacturer’s instructions. RNA concentration and integrity was evaluated using the RNA 6000 Nano Labchip® kit (Agilent Technologies) on an Agilent 2100 Bioanalyzer (Agilent Technologies).

**Public T-cell disease data**

Public data were downloaded from the Gene Expression Omnibus that met the following criteria on 31 December 2013 (Table S4): i) expression profiling of unstimulated CD4$^+$ T-cells from healthy controls and patients with T-cell-related non-virus diseases; ii) at least five samples per disease and controls; and iii) the patients were not drug treated and symptomatic, iv) samples come from in vivo activated (GSE4588, GSE8835, GSE12079, GSE14317, GSE14924, GSE13732, GSE50101).

**Th1/Th2 regulatory region analysis**

DNase I hypersensitive sites (DHS) in Th1 and Th2 cells were downloaded from ENCODE. We then designed methylation microarrays (Agilent Technologies) within the accessible DNA regions (in total 79,789 enhancers) and analyzed the methylation levels of naïve T-cells (NT), and NT polarized towards Th1, and Th2 for 7 days. Data was quantile normalized, each regulatory region was mapped to the upstream gene and the distribution of all active regulatory regions was used to define if a regulatory region was active: an enhancer was non-active if its methylation probe level was more than two standard deviations below the mean probe level (Z<-2), see Figure S4 for more information. We defined a gene as possibly Th1/Th2 accessible if it had any active regulatory region in NT, Th1 or Th2 cells. This resulted in 16,969 out of 22,628 genes to be identified as DNA accessible. Data is available through Supplementary data 1.
**Bioinformatics analyses**

Probes with a signal below the detection limit in more than half of the samples were removed from analysis and data were quantile normalized. We considered mean values over the probes mapped to the same mRNA molecule; differential expression of cases and controls was computed using the LIMMA package in R. To find differentially expressed (DE) genes between the time-series measurements we used the maSigPro algorithm implemented in R by default parameters (28). For the Th1/Th2 genes, this tested all the Th1 with Th2 time-points (and similarly for Th17/Treg). From the maSigPro genes we defined the subset of *upstream* genes if it were DE at 6h or 24h in either Tippets or Fishers method for combining P-values. Upstream genes for each specific T-cell subset were defined by ranking the combined P-values from the 6h and 24h time-points, and selecting the same number of genes as for the upstream Th1/Th2 genes as differentially expressed (n=489). Individual genes were considered differentially expressed if the double-sided P<0.05 using 10% false discovery rate by the Benjamini Hochberg procedure.

The P-values for the target genes in diseases was calculated separately for each disease and TF by first forming the average of the log10 P-values for the predicted target genes, and forming a bootstrap P-value using the log10 P-values of the particular disease as background. The correlation analysis was carried out from geometric mean of the P-values of its predicted targets for each of the eight diseases and TFs.

To determine if a set of genes was differentially expressed we analyzed the average of the -log10 P-values of the group as a test-statistic and formed a bootstrap P-value using the -log10 P-values of the particular data set as background, by resampling the original data $10^6$ times. This test is a weighted version of the established GSEA approach and is more sensitive for the purpose of testing if a single gene set is enriched for genes with high scores (e.g. low P-values). Estimation of small $P$-
values from the bootstrap test were approximated using the algorithm of (29). Effect size of enrichment analyses are throughout displayed as odds ratios, i.e. the average value of the test statistic in the test group divided by the average value in is complementary group. The correlation analysis for the TF-target relationships for the eight public diseases analysis was carried out from geometric mean of the P-values of its predicted targets for each of the eight diseases, and the corresponding P-value of the TFs (from Limma).

The P-values for qPCR was calculated using one-sided non-parametric Wilcoxon test. For gene set enrichment analysis (e.g., GWAS-genes) P-values were calculated using one-sided Fisher exact test using all genes as background (n=22,500).

**Chip-Seq analyses**

Chromatin-immunoprecipitation sequencing (Chip-Seq) was performed as previously described by us. Human naïve CD4+ T-cells were differentiated toward Th1 and Th2 until day 6 as described above. To fix the DNA-protein complex, cells were cross-linked with 1/10 volume of cross-linking solution (50 mM HEPES-KOH [pH 7.5], 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, and 11% formaldehyde) for 10 min followed by quenching with 1/20 volume of 2.5 mM glycine solution for 10 min. Cells were harvested and washed twice with ice-cold PBS and pelleted, and nuclear fractions were prepared. The cell pellet was dissolved in 20 ml lysis buffer 1 (50 mM HEPES KOH [pH 7.5], 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% Nonidet P-40, and 0.25% Triton X-100 with protease inhibitor Complete Mini [Roche]) and incubated for 10min at 4°C followed by centrifugation at 2000 × g for 10 min at 4°C. The cell pellet was resuspended in 10 ml lysis buffer 2 (200 mM NaCl, 1 mM EDTA [pH 8], 0.5 mM EGTA, and 10 mM Tris [pH 8], with protease inhibitor) at room temperature for 10 min. Nuclei were pelleted at 2000 × g for 10 min at 4°C and resuspended in 1ml lysis buffer 3 (1 mM EDTA [pH 8], 0.5 mM EGTA [pH 8], and 10 mM Tris
[pH 8], with protease inhibitor). The chromatin was sonicated for DNA fragmentation and analyzed with agarose gel electrophoresis. The majority of the DNA fragments were in the 500–100 bps range. Cell debris was pelleted by spinning at 10,000 × g for 10 min, and the supernatant was used for chromatin immunoprecipitation (ChIP). 20 μl chromatin supernatant was retained as the DNA input control. Sheep anti-rabbit IgG-conjugated Dynabeads (Dynal) were mixed with antibodies from commercial sources (α-GATA3, Santa Cruz sc-22206; α-MAF, Santa Cruz sc-10017X; α-MYB, Abcam ab45150) in blocking solution (PBS containing 5 mg/ml BSA) o/n at 4°C. Beads were collected and washed three times with PBS/BSA. Chromatin was equally divided into sample and control, adjusted with Triton-X-solution (0.1% Triton X-100, 0.1% sodium deoxycholate, and 1 mM PMSF) and mixed with magnetic beads pre-coated antibodies and incubated at 4°C o/n. Next day, magnetic beads were collected, put on magnetic stand and the supernatant was removed. Then beads were washed with 1 ml RIPA buffer (50 mM HEPES [pH 7.6], 1 mM EDTA, 0.7% sodium deoxycholate, 1% Nonidet P-40, and 0.5 M LiCl, with protease inhibitor mixture) at 4°C to remove nonspecific binding. The beads were collected and washed with RIPA buffer a total of 8-10 times. After washing once with 1 ml TE buffer, the beads were collected by centrifugation at 2000 × g for 2 min and resuspended in 50 μl elution buffer (50 mM Tris [pH 8], 10 mM EDTA, and 1% SDS). Chromatin was eluted from the beads by incubation at 65° for 30 min in a thermo-mixer followed by centrifugation for 1 min at 2000 × g. A total of 50 μl eluted supernatants were removed and mixed with 120 μl TE buffer. Similarly, 20 μl chromatin input control was mixed with 150 μl TE buffer. Decrosslinking was performed by incubation at 65°C overnight. Proteins in the DNA sample were digested in 120 μl proteinase K solution (2% glycogen, 5% proteinase K stock solution, and TE) for 2 h at 37°C. The DNA was then extracted twice with phenol and once with 24:1 chloroform/isoamyl alcohol. The sample was adjusted to 200 mM NaCl. After ethanol precipitation, the DNA was dissolved in 30 μl TE containing 10 μg DNase-free RNase A and
incubated for 2 h at 37°C. The DNA was further purified with a Qiagen PCR kit (Qiagen). Libraries for sequencing were generated as previously described. Single-end, 50bp reads were sequenced on an Illumina HiSeq 2000 sequencer.

Technical replicates were merged at the alignment step. Reads were aligned to the genome hg19 using Bowtie2 (default settings) (30). Peaks were called using MACS14 (31), with the p-value cutoff of 10^{-4}. Since the control sample for GATA3 has failed, during the peak calling input was ignored. For MAF and MYB appropriate controls were used. Peaks were mapped to all genes with TSS in the range of 4kb upstream and downstream from the peaks using BEDTools (32). In order to validate whether GRN GATA3, MAF and MYB predicted targets are enriched for ChIP-Seq peaks with higher confidence (lower peak p-value) we have performed weighted bootstrap test. We have checked if weighted average number of peaks mapped to TF-targets is higher than for 10 000 sets of random genes. Weights were calculated as a 1- geometric mean of the peaks p-values.

**SNP analyses**

We downloaded GWAS data compiled by the National Human Genome Research Institute (NHGRI) (2014/03/06). We first manually classified the 421 of the 995 traits as disease traits and removed the remaining traits from the analysis. We then mapped the SNPs to the nearest upstream and downstream gene and referred to the genes with at least one mapped disease SNP mapping as GWAS (Table S9). For the SNP frequency enrichment study we downloaded and mapped all nominally disease associated SNPs with P<10^{-3} (33) version 3 (2014/03/06). This data includes in the disease SNPs from NHGRI and in total 193,209 disease associations.

In order to analyze if SNPs were associated with the splicing process we first mapped the disease associated SNPs from NHGRI to the splice regulatory sites (SRS) of GATA3, MAF, and MYB (34).
SRS (defined as exon splicing silencer, exon splicing enhancer, intron splicing silencer, intron splicing enhancer and splicing sites) were predicted with RegRNA2.0 online tool, and identified 81 such predicted SRSs (Table S5). To test this more realistically we performed exon eQTL analyses for the GATA3 exons. We choose GATA3 as the most interesting example since it exists in only two variants, which differ only in exon number 4, which therefore can be used to understand the effect of splicing. Similarly, MAF exists in two variants but were excluded from the analysis since we could not have RNA-seq data for it, while MYB was excluded since it had exists in many splice variants. Briefly, data from genotypes and normalized exon counts were downloaded from the Geuvadis repository. The association between genotypes and exon counts was computed by fitting linear equations and p values obtained based on the F-test. Only samples with a European origin were used in this analysis. Exon annotations were based on Ensembl GRCh37.

Construction of the Th1/Th2 GRN
From the list of DE genes we removed the genes networks, which were not Th1/Th2 accessible form the methylation analysis. We then defined genes as transcription factors (TFs) from the list of 1750 human transcription factors, and co-factors from the Animal TF database (10) (2013/01/01). For sequenced based TF-target predictions we used the predictions by (11). From this data a confidence interaction matrix $P_{ij}$ was formed, representing our prior confidence of a regulation of TF $i$ onto target gene $j$, $P_{ij} = \frac{k_i}{n}$ if TF $i$ binds to target gene $j$, and $P_{ij} = 1 + \frac{k_i}{n}$ otherwise ($k_i$ is the number of targets of TF $i$). Next, for all genes $i$ and each time-point $k$ we computed the difference ($y_i$) between the expression levels of the Th1 ($x_{TH1}^{Th1}(t_k)$) and Th2 ($x_{TH2}^{Th2}(t_k)$) polarizations, respectively: $y_i(t_k) = x_{TH1}^{Th1}(t_k) - x_{TH2}^{Th2}(t_k)$. Subsequently, the rate of change $\dot{y}(t_k)$ by the mean of forward and backward differences at the interior points and the corresponding
differences end points was estimated. A linear model was used for the time evolution
\[ \dot{y}_i(t) = \sum_j w_{ij} y_j(t) \]
of all the Th1/2 genes and inferred the \( w_{ij} \) by the LASSO (12). The LASSO inference takes the following form:

\[
\min_{\mathbf{w}} \sum_k \left( \dot{y}_i(t_k) - \sum_{j \in \text{TF}} w_{ij} y_j(t_k) \right)^2 + \lambda_i \sum_{j \in \text{TF}} \left| w_{ij} \right| \quad \forall i
\]

We solved this problem for each gene individually using the glmnet implementation in Matlab with default parameter settings (12, 35), i.e. determining \( \lambda_i \) from 10-fold cross-validation using the one standard error criteria. For a description of LASSO and network inference alternatives see below.

**Genome-wide network inference**

As a brief background, two main alternatives exist for genome-scale modeling, namely information theory-based methods (e.g. ARACNE (36)) and regression based approaches using Ordinary Differential Equations (ODEs). Several comparisons between the two have been made, where some show regression-based approaches to be superior for dynamics. Regression based approaches approximate the ODEs that describes the kinetics by differential equations. This enables construction of parameters describing how each gene is regulated to be performed in parallel, which allows the approach to scale linearly in time for each gene. The most popular regression based approach is called Least Absolute Shrinkage and Selection Operator (LASSO). This has been used within the network inference tools Inferrelator (37), TILAR (38) and by us (39, 40). For our purpose the LASSO has two key features: 1) The LASSO penalizes complex models with many interactions, and can also successfully “softly” incorporate multiple data types in the network identification processes (41, 42), which means that each data type contributes to the network inference process. 2) Apart from performing well in, benchmark tests, the LASSO approach has
identified topological network properties that are biologically relevant, such as the identification of hubs, and key regulators of gene modules (39, 43).

**Network reconstruction details**

Th1 and Th2 cells are considered to counter-regulate each other. The expression signals we are considering in this manuscript describe the difference between the two cellular types. For the following reasons we find that the corresponding GRN to describe the Th1/Th2 switch. First, we identified hub TFs such as GATA3 which are known to be involved in the transcriptional regulation of this switch between the T-cell subsets. Second, we found enrichment of GWAS-genes within the early Th1-Th2 TFs. Third, we found a 1.9-fold enrichment (Fisher exact $P=10^{-19}$) for targets for all TFs with predicted binding sites when using LASSO on the Th1/Th2 data. Neither of those facts were seen when repeating the analysis for the individual subsets. Furthermore, the overlap analysis motivated our integration of TF bindings to the LASSO model, since the data-sets then supports each other. Our subsequent analysis of Chip-seq and siRNA of GATA3, MAF and MYB showed further enrichments, thus confirming that the integrated approach filtered out functional and direct physical TF-target interactions in both Th1 and Th2 cells. We believe that this kind of cautious integration of multiple data sources testing for overlap between the data sets is generally needed in the context of network construction. We believe that this kind of cautious integration of multiple data sources testing for overlap between the data sets is generally needed in the context of network construction.

**MS gene expression in relapse and remission**

We analyzed gene expression data followed patients with a clinically isolated syndrome at their first visit after discovery, and at a subsequent follow up study one year after. The full data set consists of 113 samples of patients and controls where most of the material consists of paired
samples (15). In this study we used additional records of these patients, such as the time-point for their remissions, the status of their recovery, and their treatments. We used this information to define patients which were in a relapse at their first visit and in remission at their second visit. We defined a relapse as within 90 days from their first clinical symptoms. We tested also more stringent criteria for the relapse definition, since the relapse may decline faster. However, we found the largest separation to controls for a relapse definition of 90 days, which suggests that immunological effects of the relapse are detectable in blood at this time. We defined a remission as more than a year since their last symptoms for and they had also clinically recovered as judged by their physician. By these definition we found that 30 patients were sampled in a remission for which we had 10 controls, which was used for the relapse vs. healthy comparison. Furthermore we found that 11 of these patients and 10 of the controls were also sampled during relapse. This set of patients were used for the remission vs. control comparison and for the remission vs. relapse comparison in the manuscript.

**GWAS analysis of multiple sclerosis**

We analyzed the genotype array of all common SNPs from original data from the MS consortium study of ~25,000 (21) individuals. We found that one or more disease-associated SNPs mapped to each of the 3 TFs (as described above): GATA3 (P = 6.6 x 10^{-5}), MAF (P= 1.9 x 10^{-5}), MYB (5.8 x 10^{-6}). These 3 TFs ranked among the top 1% genes that were enriched for disease-associated SNPs (P=3 x 10^{-6} using Bootstrap permutation test). A total of nine SNPs mapped to the three TFs with P-value<0.001. These SNPs were queried for predicted regulatory effects using the regulome database (http://www.regulomedb.org/ by 18/6 2015), and we found seven to have any predicted regulatory effect (score 1-6). We found 4 MYB SNPs (rs9321490, rs9399141, rs2050019, rs1569534), which defined two regions of regulatory interest suggested by enhancer and DNAase data in T cells. For GATA3 we found one predicted regulatory SNP in the intron of GATA3
(rs1399180), and two about 300 kb upstream of GATA3 (rs7096374, rs10905371). For MAF one predicted regulatory SNP (rs386965) was found. In order to test if the seven SNPs were associated with disease severity, we analyzed an independent study of 2,085 patients with MS. These patients were genotyped with the Immunocip array, as previously described (44). However, our analyses were based on previously unpublished data about disease severity. We identified 252 of the 2,085 the patients that had a severe variant of the disease. These were patients that had a) started second line treatment of Tyasbri, Gilyenia or Mabtera, and b) MSSS value above the median MSSS value of all MS patients. MSSS uses the disability combined with disease duration to rate disease severity (45).

We found that three of the seven SNPs were assayed on the Immunochip (rs386965, rs9321490, rs9399141), and for the GATA3 intronic SNP we found rs3781094 as being located 2708 base pairs away and be in linkage disequilibrium (D’>0.9). We found that the MS patients with severe disease had 7.1% more of the disease-related SNPs than the remaining MS patients (P=0.015, Bootstrap test).

**Classification of patients/controls based on machine learning**

To investigate possible clinical implication of the use of the upstream transcription factors and their targets we constructed a Support Vector Machine (SVM) classifier. The SVM classifier was created with the MATLAB function `svmtrain()`. To assess the quality of classification we have used a leave-one-out method. Area under ROC curve (AUC) was estimated using the `trapz()` MATLAB function for trapezoidal numerical integration.

We tested whether the 3 TFs and/or their targets might be used to discriminate between patients and controls. Our results showed that expression of GATA3, MAF and MYB allowed good
classification in 6 out of 8 diseases used in manuscript \((AUC > 0.85)\), while the predicted targets of the upstream TFs allowed even better classification for all 8 studied diseases \((AUC > 0.95)\).

**Gene Regulatory Network and TFBS comparison using ChIP-Seq peak counts**

In order to investigate if GRN predictions have more ChIP-Seq counts than TFBS predictions we have performed bootstrap test using all genes from the network as a background. We have calculated the ratio of the average number of peaks mapped to the GRN-predicted TF-targets divided by the average number of peaks mapped to TFBS-predicted TF-targets. Next we have calculated same ratio for randomly selected genes according to the bootstrap methodology. We have performed 10 000 permutations. Average odds ratio and Fisher combined P-value of all three tests (one for each TF; \(GATA3\), \(MAF\) and \(MYB\)) are reported in the main text. For detailed information please see Table S12.
Supplementary figures

**Figure S1.** Expression of signature genes in T cell subsets. qPCR bar plots show mean expression values of signature TFs for the different T cell subsets after 6 days of polarization in four biological replicates, normalized against the *ACTB* levels. The error bars represent one standard deviation.
Figure S2. Comparison of GWAS enrichment for gene sets defined from alternative strategies defined by the T-cell differentiation microarray data. (A) Enrichment of genes that were mapped to at least one disease associated SNP (GWAS) in many gene subsets. Early Th2 (Th1) genes are defined as differentially expressed between Th1 and Th2 cells in 6h or 24h of polarization assay that had higher expression in Th2(Th1). Late Th2 (Th1) genes are defined as differentially expressed between Th1 and Th2 cells in 3d, 6d or 8d of polarization assay that had higher expression in Th2(Th1). Same principal was used to identify early/late Th17/Treg genes. Activation Th2(Th1) genes are defined as differentially expressed between Naïve T-cells and Th2(Th1) cells in 6h or 24h of polarization assay. (B) Enrichment of genes that were mapped to at least one disease SNP (GWAS) depending on the gene p-value cutoff. Th1 (Th2, Th17, Treg) – p-values calculated with t-test\textsubscript{LIMMA} between Th1 (Th2, Th17, Treg) in 6h and 24h of polarization. ‘Th2 vs Th2 ‘ (‘Th17 vs Treg ‘) combined p-values calculated between Th1 (Th17) and Th2 (Treg) cells in 6h and/or 24h of polarization (see supplementary methods), ‘Th2 vs Th2 TFs‘ (‘Th17 vs Treg TFs‘) – selected combined p-values of transcription factors calculated between Th1 (Th17) and Th2 (Treg) cells in 6h and/or 24h of polarization (see supplementary methods). (C) Enrichment of genes that were mapped to at least one disease SNP (GWAS) depending on the SNP-disease association p-value cutoff. Th1 (Th2, Th17, Treg) – p-values calculated with t-test\textsubscript{LIMMA} between Th1 (Th2, Th17, Treg) in 6h and 24h of polarization. ‘Th2 vs Th2 ‘ (‘Th17 vs Treg ‘) combined p-values calculated between Th1 (Th17) and Th2 (Treg) cells in 6h and/or 24h of polarization (see supplementary methods), ‘Th2 vs Th2 TFs‘ (‘Th17 vs Treg TFs‘) – selected combined p-values of transcription factors calculated between Th1 (Th17) and Th2 (Treg) cells in 6h and/or 24h of polarization (see supplementary methods). ‘all genes’ is a group of all genes that were measured on the microarray and ‘all TFs’ is a group of all transcription factors measured on the microarray.
**Figure S3.** GWAS SNPs in different populations. The Y axis shows the percentage of GWAS SNPs present in multiple populations according to the information collected by the HapMap project (data downloaded via HapMart on 01/04/2015).
Figure S4. Methylation probe levels of selected house-keeping (A) and tissue specific genes (B). The levels were transformed to standard z-scores using the standard deviation from the full distribution of all probes.
Figure S5. Enrichment in GWAS SNPs in GATA3, MAF and MYB in T cell-associated diseases, infectious diseases and malignancies. T cell-association was defined based on MedLine. P-values were calculated using a permutation test (10 000 permutations).
Figure S6. Overlap between different methods to define GATA3 bindings: (A) Chip-Seq (‘Chip-Seq (new)’) experiment, siRNA mediated knockdown (‘knockdown’) and transcription factor binding sites sequence-based predictions (‘TFBS’). (B) Venn diagrams of the overlap between GRN predicted (‘LASSO), Chip-Seq and public Chip-Seq (‘public Chip-Seq’) experiments. P-values were calculated using Fisher Exact test. Network inferred GATA3 targets yielded significantly higher ChIP-seq overlap compared to TFBS (OR=1.62, P=2.41 x 10^{-6}). This finding was also confirmed by comparisons with previously published independent ChIP-seq data (OR=1.55, P=3.4 x 10^{-5}).
Figure S7. GRN-predicted targets of GATA3, MAF and MYB are differentially expressed in 6 T cell related diseases: Lymphocytic variant of hypereosinophilic syndrome (HES), adult T-cell leukaemia/lymphoma (ATL), acute myeloid leukaemia (AML), systemic lupus erythematosus (SLE), multiple sclerosis (MS) and seasonal allergic rhinitis (SAR). In SAR both symptomatic (public material and exon arrays) and asymptomatic patients (exon arrays) were analyzed. Bars mark the difference in the mean p-values for the targets of each TF compared to all genes (* P<0.05, ** P < 0.01; *** P < 0.0001, ****<0.0000001 from permutation test).
**Figure S8.** Validation of the differentially expressed splice variants in SAR. The expression levels measured with microarray significantly correlated with the expression values obtained with qPCR for each of (A) *GATA3.1*, (B) *MAF.2* and (C) *MYB*. PCC=Pearson correlation coefficient.
Figure S9

- Individual subject’s gene expression normalized by ACTB in Th1
- Average normalized by ACTB gene expression in Th1
- Individual subject’s gene expression normalized by ACTB in Th2
- Average normalized by ACTB gene expression in Th2

Gene expression normalized by ACTB

GATA3 (NM_001002295)

GATA3 (NM_002051)

MAF (NM_001031804)

MAF (NM_001031804, NM_005360)

MYB (NM_001161659, NM_005735, NM_001161658, NM_001130173)

MYB (NM_001161656, NM_001130173)
Figure S9. Time series analyses of *GATA3, MAF* and *MYB* splice variant expression during Th1 and Th2 cell differentiation. The splice variant expression levels were measured with qPCR relative to *ACTB* in Naïve T cells, as well as Th1 and Th2 cells after 1 and 5 days of polarization. The cells were derived from three healthy individuals. Blue/red squares/dots represent splice variant expression in each individual, while blue dashed lines show average splice variant expression in Th1/Th2 cells.
Figure S10. Predicted effects of alternative splicing on the protein structures of GATA3, MAF and MYB. (A) Structure of the GATA3 protein as determined by crystallography (46). The blue circle shows the position that differs between the two isoforms. (B) Predicted protein structure of the MAF proteins. The blue circle shows the predicted location of the difference between the two isoforms. (C) Differences between MYB splice variants. Blue rectangles denotes MYB exons.
Figure S11. Heatmaps of the top 50 most differentially expressed genes in siRNA mediated MAF knockdown using 4 different vectors. The genes were analyzed with gene expression profiling.
Figure S12 Effect of MAF knockdown on Th1, Th2, Th17 and Treg signature genes.

Knockdown was performed with 4 different siRNA’s: J-003746-07 (vector v1), J-003746-08 (vector v2), J-003746-09 (vector v3), and J-003746-10 (vector v4).
Figure S13. *MAF* siRNA splice specific targets. (A) Stars denotes network inferred predicted *MAF* targets. In order to identify *MAF* splice variant 2 and *MAF* splice variant 1 specific targets we have calculated the correlation between GRN predicted targets and *MAF* specific variants in *MAF* knockdown experiment. (B) Cumulative distribution of specific *MAF* splice variant targets P-values in allergen challenged T cells from SAR patients.
Figure S14. Expression profiles of SAR patients and controls show a consisting overlap even for the most significantly differentially expressed genes. In figure (A) the expression values of the top 5 most significantly differentially expressed splice-variants between SAR patients and controls are presented (resting cells - left panel, allergen stimulated cells - right panel). Red and blue dots denote the expression values in patients and controls respectively. (B) Percentage of the patients and controls with overlapping expression in resting cells (left panel) and allergen stimulated cells (right panel). For example in resting cells (left panel) for all genes with p-value ≤ 0.05 (x-axis equal to 10^{-1.3}) on average ~65% of patients and controls have overlapping gene expression. The overlap is measured as the percentage of samples for which sensitivity and specificity are both <1 when assessing classification quality with ROC. For instance the most differentially expressed splice-variant in resting cells (NM_001009608, fig. S14A) shows zero overlap, since we can separate patients and controls based on the expression values of this gene only. However, for the second most differentially expressed splice-variant (NM_032782) the overlap would be 20% as four samples fall into the uncertain area where sensitivity and specificity are both <1 and there are 20 samples (4/20*100%) in total. (C) The plot presents relationship between the percentage of the genes and the specific overlap between patients and controls. For instance in resting cells 10% of all the measured genes (X axis equal to 10^{-1}) show overlap between patients and controls of approximately 65% or less. Gene expression overlap was calculated as explained above.
Figure S15. Expression profiles of MS patients and controls show a consisting overlap even for the most significantly differentially expressed genes. In figure (A) the expression values of the top 5 most significantly differentially expressed splice-variants between MS patients and controls are presented (non-symptomatic disease stage - left panel, active disease - right panel). Red and blue dots denote the expression values in patients and controls respectively. (B) Percentage of the patients and controls with overlapping expression in non-symptomatic disease stage (left panel) and active disease (right panel). For example in non-symptomatic disease stage (left panel) for all genes with p-value ≤ 0.05 (x-axis equal to $10^{-1.3}$) on average ~62% of patients and controls have overlapping gene expression. The overlap is measured as the percentage of samples for which sensitivity and specificity are both <1 when assessing classification quality with ROC. For instance the most differentially expressed splice-variant in non-symptomatic disease stage (HBB, fig. S15A) shows zero overlap, since we can separate patients and controls based on the expression values of this gene only. However, for the second most differentially expressed splice-variant (HBA2/1) the overlap would be 9.52% as two samples fall into the uncertain area where sensitivity and specificity are both <1 and there are 21 samples (2/21*100%) in total. (C) The plot presents relationship between the percentage of the genes and the specific overlap between patients and controls. For instance in non-symptomatic disease stage 10% of all the measured genes (X axis equal to $10^{-1}$) show overlap between patients and controls of approximately 52% or less. Gene expression overlap was calculated as explained above.
Figure S16

average overlap between expression of patients and controls [%]

gene p-value
Figure S16 Expression overlap between patients and controls in many T cell associated diseases. Percentage of the patients and controls with overlapping expression in multiple diseases: lymphocytic variant of hyper eosinophilic syndrome (HES), adult T-cell leukemia/lymphoma (ATL), acute myeloid leukemia (AML), multiple sclerosis (MS), systemic lupus erythematosus (SLE), seasonal allergic rhinitis (SAR), rheumatoid arthritis (RA), chronic lymphocytic leukemia (CLL), tonsillitis (TO), tuberculosis (TB), breast cancer (BC) and influenza (IZ). The overlap is measured as explained in the figures S14 and S15.
**Figure S17 Expression overlap between patients and controls in many T cell associated diseases.** The plot presents relationship between the percentage of the genes and the specific overlap between patients and controls in multiple diseases: lymphocytic variant of hyper eosinophilic syndrome (HES), adult T-cell leukemia/lymphoma (ATL), acute myeloid leukemia (AML), multiple sclerosis (MS), systemic lupus erythematosus (SLE), seasonal allergic rhinitis (SAR), rheumatoid arthritis (RA), chronic lymphocytic leukemia (CLL), tonsillitis (TO), tuberculosis (TB), breast cancer (BC) and influenza (IZ). For instance in lymphocytic variant of hyper eosinophilic syndrome (HES) 10% of all the measured genes (X axis equal to $10^{-1}$) show overlap between patients and controls of approximately 10% or less. Gene expression overlap was calculated as explained in figures S14-S15.
Figure S18. GRN validation. (A) Inferred combined targets of NFATC3 and NFKB1 were enriched for differentially expressed genes in siRNA experiments. NFATC3 and NFKB1, knockdowns when analyzed separately showed no significant enrichment of differentially expressed genes among the GRN predicted early targets (OR = 1.90, P= 0.14 and OR = 2.11, P= 0.11 respectively). However when analyzed the predicted targets of the two TFs together, the combined predicted early targets of NFATC3 and NFKB1 were significantly enriched for differentially expressed genes compared to controls (P = 0.0076, OR =4.12). Plot presents
cumulative distribution of Fisher combined P-values of \textit{NAFTC3} and \textit{NFKB1} in siRNA for inferred combined early target genes of \textit{NAFTC3} and \textit{NFKB}, TFBS and all early genes. (B & C) Inferred targets of \textit{GATA3} and \textit{MAF} were enriched for differentially expressed genes in siRNA experiments. Cumulative distribution of P-values of calculated by siRNA for inferred early target genes of \textit{GATA3} (B) and \textit{MAF} (C) respectively (upper curve), TFBS (dashed lower curve) and all early genes (solid lower curve). In both cases the GRN-predicted targets of \textit{GATA3} and \textit{MAF} showed significantly lower average P-values ($P_{GATA3}=0.0029$, $P_{MAF}=0.0045$). (D) Enriched binding for \textit{GATA3} predicted targets in Th1 and Th2 cells from public ChIP-seq analysis. The vertical axis represents the fraction of targets with a minimum ChIP-seq count using the inferred targets (upper curve), the predicted TF-DNA bindings (TFBS, middle curve), and all genes (lower curve). The inferred targets were enriched for ChIP-seq counts and the mean ChIP-seq counts were highest for the inferred targets in genes ($P=3.4 \times 10^{-5}$)
Supplementary tables

Table S1. Predicted regulatory SNPs for GATA3, MAF and MYB using rSNPBase (47). Regulatory SNPs (rSNPs) are the SNPs in the regulatory regions (for example proximal and distal transcriptional regulation sites inside elements associated with DNA accessibility, in the RNA binding proteins-associated regions, or SNPs affecting microRNA-mRNA interference). We found that 14 disease associated SNPs have a predicted regulatory effect and 252 which are in LD block with rSNP (106 for GATA3, 61 for MAF and 85 for MYB). We therefore also analyzed the positions of all known SNPs downloaded from USCS database and found many in the splice regulatory motif regions to be in linkage disequilibrium (LD) with disease-associated SNPs. Splice regulatory motifs are exonic and intronic splicing enhancer and silencers, splicing sites or open reading frames. For GATA3 we have found 129 SNPs in the splice regulatory motif regions identified with RegRNA 2.0. Twenty nine of those SNPs are common SNPs (allele frequency > 0.01). All of those SNPs are in LD with 15 disease associated SNPs. For MAF we have identified 6 SNPs in the splice regulatory motifs in LD with 5 disease associated SNPs, 3 SNPs with allele frequency greater than 0.83 (Those SNPs are associated with Nasopharyngeal carcinoma, Type II Diabetes and triglyceride levels, Renal function-related traits BUN, Multiple sclerosis and Rheumatoid arthritis).

Table S2. All pathways for the targets of GATA3, MAF and MYB respectively reported by IPA (Ingenuity® Systems, www.ingenuity.com) by 2013-11-19. Column A lists the g pathways, column B the P-value (Fisher exact test), column C the fraction of the genes in the pathway that was in the list, and column D the names of all the genes in the list that were included in the pathway.

Table S3. All pathways for the early predicted targets of all early TFs. All pathways for the predicted targets of each of the TFs that are differentially expressed early in the Th1 and Th2 polarization assay, and that are differentially expressed in at least 6 T cell associated diseases. Gene
Ontology biological process terms were clustered together by the similarity of the genes involved in those terms with k-means algorithm. P-values were calculated using one-sided Fisher exact test using all genes as background (n=15398).

Table S4. Statistics of all 8 T cell diseases. Column A is the name of the disease. Column B the Gene Expression Series Number. Column C and D are the number of patients and controls in the study.

<table>
<thead>
<tr>
<th>disease names</th>
<th>GSE numbers</th>
<th>number of patients</th>
<th>number of controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rheumatoid arthritis</td>
<td>GSE4588</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Chronic lymphatic leukemia</td>
<td>GSE8835</td>
<td>42</td>
<td>24</td>
</tr>
<tr>
<td>Hypereosinophilic Syndrome</td>
<td>GSE12079</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>Acute T cell Leukemia</td>
<td>GSE14317</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>Acute Myeloid Leukemia</td>
<td>GSE14924</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Systemic Lupus Erythematosus</td>
<td>GSE4588</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>CIS-Multiple Sclerosis</td>
<td>GSE13732</td>
<td>39</td>
<td>30</td>
</tr>
<tr>
<td>Seasonal allergic rhinitis</td>
<td>GSE50101</td>
<td>30</td>
<td>15</td>
</tr>
</tbody>
</table>

Table S5. SNPs that might affect splicing processes were identified by mapping all disease-associated SNPs (GWAS) to the splice regulatory sites (SRS) of MAF, MYB and GATA3. SRS (defined as exon splicing silencer, exon splicing enhancer, intron splicing silencer, intron splicing enhancer and splicing sites) were predicted with RegRNA2.0 online tool (34).

Table S6. Gene ontology enrichment analysis of biological processes of the MAF splice variant 2 specific targets (see Figure S13). MAF splice variant 2 specific targets are network inferred MAF targets that are also significantly correlated (Pearson correlation P<0.05, FDR<0.1) with siRNA of MAF variant 2. P-values were calculated with one sided Fisher exact test using all genes as background n = 15,398.
Table S7. Results for the naive questions "how many PPIs have the Th differentially expressed genes, globally or just early", and "would had been identified the TFs through this PPI analysis?" The table shows results for HPRD and for Y2H-only interactions and, in each sheet, you have the proteins ranked according to the ratio: \( \frac{n=DEG \ partners}{N=total \ partners} \) so from this analysis you would had never identified GATA3, MAF, MYB: they rank >300 all-.

Moreover, we asked the following question is "would had been identified the TFs if we extract ppi modules from those networks?" So, first we identified "modules" (MCode-based) in the HPRD or Y2H networks; second, we have defined which modules show a significant enrichment in all-DEG or only early-DEG gene sets; and third, within each enriched module, we have ranked the proteins according to its "centrality betweenness". Again the results showed that you would had never identified GATA3, MYB, MAF: only GATA3 and MAF are included in "early DEG" enriched modules but they are ranked >400 according to their centrality betweenness.

Release #9 of the Human Protein Reference Database (HPRD) was downloaded from the corresponding repository (www.hprd.org). The main network component contained 9,267 nodes or proteins and 36,893 edges or interactions (excluding homodimers). From this dataset, Y2H-based interactions generated a main network component containing 6,208 nodes and 13,700 edges. Direct protein-protein interactions were only computed in the main component and modules were identified by applying an implementation of the MCODE algorithm (default parameters) within the Cytoscape software. The significantly enriched modules for predicted differentially expressed gene products were defined by applying a Fisher exact test (sets compared to total number of proteins in the main component), and the P-values adjusted by the false discovery rate (FDR) approach (cut-off 5%). Modules with less than 10 nodes and with odds ratios < 1 were excluded from this analysis. The betweenness centrality was computed by applying a previously described algorithm implemented within the NetworkX software package.
Table S8. Splice variant qPCR details (reagents and company).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Interrogated Sequence</th>
<th>Exon/Amplicon</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAF Hs.PT.58.1954484.g</td>
<td>NM_001031804</td>
<td>1b-1b/101</td>
<td>IDT</td>
</tr>
<tr>
<td>MAF Hs.PT.58.872936</td>
<td>NM_005360</td>
<td>1a-1c/141</td>
<td>IDT</td>
</tr>
<tr>
<td>MAF Hs04185012_s1</td>
<td>NM_001031804.2 NM_005360.4</td>
<td>1-1/101</td>
<td>Applied Biosystems</td>
</tr>
<tr>
<td>MYB Hs.PT.58.2815709</td>
<td>NM_001161657 NM_001161659 NM_005375</td>
<td>8b-11/113</td>
<td>IDT</td>
</tr>
<tr>
<td>MYB Hs.PT.58.653127</td>
<td>NM_001130173 NM_001161656</td>
<td>9b-10/123</td>
<td>IDT</td>
</tr>
<tr>
<td>MYB Hs.PT.58.14697969</td>
<td>NM_001161660</td>
<td>7-9b/131</td>
<td>IDT</td>
</tr>
<tr>
<td>MYB Hs00920556_m1</td>
<td>NM_001130172.1 NM_001130173.1 NM_001161656.1 NM_001161657.1 NM_001161658.1 NM_001161659.1 NM_005375.2</td>
<td>3-4/76</td>
<td>Applied Biosystems</td>
</tr>
<tr>
<td>GATA3 Hs.PT.58.25659812</td>
<td>NM_001002295</td>
<td>3-4a</td>
<td>IDT</td>
</tr>
<tr>
<td>GATA3 Hs00922327_m1</td>
<td>NM_002051.2</td>
<td>3-4/73</td>
<td>Applied Biosystems</td>
</tr>
<tr>
<td>GATA3 Hs00231122_m1</td>
<td>NM_001002295.1 NM_002051.2</td>
<td>1-1/101</td>
<td>Applied Biosystems</td>
</tr>
<tr>
<td>RPS18 Hs.PT.58.14390640</td>
<td>NM_022551</td>
<td>2-3/130</td>
<td>IDT</td>
</tr>
<tr>
<td>ACTB Hs.PT.39a.22214847</td>
<td>NM_001101</td>
<td>1-2/110</td>
<td>IDT</td>
</tr>
</tbody>
</table>

Table S9. The list of all genes with at least one disease associated SNP, identified by GWAS.
Table S10. **MAF knockdown efficiency** measured in 2 individuals. MAF expression values were normalized by ACTB, non-targeting siRNA was used as a control.

<table>
<thead>
<tr>
<th>Table S10</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>J-003746-07</td>
<td>-97%</td>
</tr>
<tr>
<td>J-003746-08</td>
<td>-248%</td>
</tr>
<tr>
<td>J-003746-09</td>
<td>24%</td>
</tr>
<tr>
<td>J-003746-10</td>
<td>51%</td>
</tr>
</tbody>
</table>

Table S11. **Results of the patient control classification using GATA3, MAF and/or MYB or their targets expression values.** In order to investigate the potential clinical applicability of measuring the upstream transcription factors or their targets we have built a Support Vector Machine classifier, and assessed its performance using leave-one-out cross-validation. We have tested whether the 3 TFs and/or their targets might be used to discriminate between patients and controls. In the table Area under ROC curve is presented.

Table S12. **GRN predictions compared with TFBS using ChIP-Seq peak counts.** In order to investigate if the GRN predictions have on average more ChIP-Seq peak counts than TFBS predictions we have performed bootstrap test (see supplementary methods).

<table>
<thead>
<tr>
<th>TF: GATA3</th>
<th>MAF</th>
<th>MYB</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>1.03x10^{-5}</td>
<td>0.48</td>
</tr>
<tr>
<td>OR</td>
<td>1.61</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Table S13. Source data values for figure 5A. Splice variant expression in unchallenged cells of 10 SAR patients (SAR) and 10 healthy controls (HC) measured by exon microarrays. Expression profiles normalized by the average expression in controls.

<table>
<thead>
<tr>
<th></th>
<th>GATA3_variant1</th>
<th>MAF_variant2</th>
<th>MYB_variant5</th>
<th>MYB_varinat2</th>
<th>MYB_varinat4</th>
<th>MYB_varinat7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>0.7629</td>
<td>2.5526</td>
<td>-2.4963</td>
<td>-1.6990</td>
<td>-0.9652</td>
<td>-2.1038</td>
</tr>
<tr>
<td>Patient</td>
<td>1.7989</td>
<td>0.8646</td>
<td>-2.0595</td>
<td>-1.3399</td>
<td>-1.6302</td>
<td>-1.2243</td>
</tr>
<tr>
<td>Control</td>
<td>1.8695</td>
<td>-1.1080</td>
<td>-0.5848</td>
<td>-0.8251</td>
<td>-0.4831</td>
<td>-1.1992</td>
</tr>
<tr>
<td>Patient</td>
<td>1.4127</td>
<td>1.0011</td>
<td>-0.8050</td>
<td>-0.2239</td>
<td>-0.2655</td>
<td>-1.6141</td>
</tr>
<tr>
<td>Control</td>
<td>0.5195</td>
<td>-0.9651</td>
<td>-0.9574</td>
<td>-1.0294</td>
<td>-0.3312</td>
<td>-0.7538</td>
</tr>
<tr>
<td>Control</td>
<td>0.7051</td>
<td>0.5631</td>
<td>-0.5677</td>
<td>-0.4466</td>
<td>-0.2444</td>
<td>-0.7773</td>
</tr>
<tr>
<td>Control</td>
<td>-1.0591</td>
<td>-0.0552</td>
<td>-0.4956</td>
<td>-0.4136</td>
<td>-0.7566</td>
<td>-0.3782</td>
</tr>
<tr>
<td>Control</td>
<td>0.7249</td>
<td>2.2338</td>
<td>-0.7037</td>
<td>-0.6089</td>
<td>-0.3334</td>
<td>-0.4193</td>
</tr>
<tr>
<td>Patient</td>
<td>-1.2625</td>
<td>-0.2228</td>
<td>-0.1368</td>
<td>0.2397</td>
<td>0.1608</td>
<td>0.2461</td>
</tr>
<tr>
<td>Patient</td>
<td>-1.2755</td>
<td>1.0235</td>
<td>-0.1130</td>
<td>-0.7690</td>
<td>-1.1978</td>
<td>0.3680</td>
</tr>
<tr>
<td>Control</td>
<td>-1.0178</td>
<td>0.7965</td>
<td>-0.7912</td>
<td>-0.4597</td>
<td>-1.7688</td>
<td>-0.6677</td>
</tr>
<tr>
<td>Control</td>
<td>-0.6503</td>
<td>-0.1322</td>
<td>1.2553</td>
<td>0.2167</td>
<td>1.4388</td>
<td>1.4714</td>
</tr>
<tr>
<td>Patient</td>
<td>0.6292</td>
<td>-0.1999</td>
<td>-0.4414</td>
<td>0.0183</td>
<td>-0.1840</td>
<td>0.0416</td>
</tr>
<tr>
<td>Patient</td>
<td>1.0696</td>
<td>1.2133</td>
<td>-1.7877</td>
<td>-1.2197</td>
<td>-1.8944</td>
<td>-1.0842</td>
</tr>
<tr>
<td>Patient</td>
<td>1.4988</td>
<td>0.0350</td>
<td>-0.9360</td>
<td>-0.4340</td>
<td>-1.4696</td>
<td>-0.9385</td>
</tr>
<tr>
<td>Patient</td>
<td>3.7076</td>
<td>0.6389</td>
<td>-1.0082</td>
<td>-0.7910</td>
<td>-1.9281</td>
<td>-1.0204</td>
</tr>
<tr>
<td>Patient</td>
<td>1.7792</td>
<td>1.5243</td>
<td>-1.3092</td>
<td>-1.0193</td>
<td>-1.0475</td>
<td>-1.3325</td>
</tr>
<tr>
<td>Control</td>
<td>-0.0228</td>
<td>-0.2286</td>
<td>1.4132</td>
<td>2.2415</td>
<td>1.1205</td>
<td>1.6445</td>
</tr>
<tr>
<td>Patient</td>
<td>3.0305</td>
<td>2.3082</td>
<td>-0.9293</td>
<td>-0.3220</td>
<td>-0.6195</td>
<td>-0.7562</td>
</tr>
<tr>
<td>Control</td>
<td>0.1934</td>
<td>-0.8816</td>
<td>1.5687</td>
<td>1.0854</td>
<td>1.1973</td>
<td>0.8334</td>
</tr>
</tbody>
</table>
Table S14. Source data values for figure 5B. Normalized average splice expression of GATA3.1, MAF.2, MYB.4, and MYB.5 in unchallenged cells of 10 SAR patients and 10 healthy controls.

<table>
<thead>
<tr>
<th>scores controls</th>
<th>scores patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,533</td>
<td>3,1747</td>
</tr>
<tr>
<td>1,047</td>
<td>3,8459</td>
</tr>
<tr>
<td>-1,7388</td>
<td>4,5263</td>
</tr>
<tr>
<td>-0,092</td>
<td>0,0184</td>
</tr>
<tr>
<td>4,6866</td>
<td>6,0721</td>
</tr>
<tr>
<td>0,2642</td>
<td>3,7861</td>
</tr>
<tr>
<td>-1,3507</td>
<td>1,8726</td>
</tr>
<tr>
<td>-3,3629</td>
<td>-0,5771</td>
</tr>
<tr>
<td>0,8005</td>
<td>3,0061</td>
</tr>
<tr>
<td>-1,7869</td>
<td>6,0817</td>
</tr>
</tbody>
</table>

Table S15. Source data values for figure 5C. Normalized average splice expression of GATA3.1, MAF.2, MYB.4, and MYB.5 measured in Healthy controls and patients before, after one and two years of immunotherapy.

<table>
<thead>
<tr>
<th>Control</th>
<th>Before T</th>
<th>After 1 year</th>
<th>After 2 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3,6753</td>
<td>2,5649</td>
<td>1,7406</td>
<td>-1,5019</td>
</tr>
<tr>
<td>0,1169</td>
<td>1,9124</td>
<td>1,2953</td>
<td>-1,4576</td>
</tr>
<tr>
<td>-2,8022</td>
<td>6,4029</td>
<td>3,024</td>
<td>0,8027</td>
</tr>
<tr>
<td>2,384</td>
<td>3,1094</td>
<td>0,6774</td>
<td>0,2847</td>
</tr>
<tr>
<td>3,075</td>
<td>7,1726</td>
<td>7,0727</td>
<td>-0,5115</td>
</tr>
<tr>
<td>0,9017</td>
<td>3,6313</td>
<td>5,6806</td>
<td>-0,0468</td>
</tr>
<tr>
<td>-3,1052</td>
<td>3,3716</td>
<td>-1,5986</td>
<td></td>
</tr>
<tr>
<td>4,2194</td>
<td>0,3347</td>
<td>-1,3199</td>
<td></td>
</tr>
<tr>
<td>1,846</td>
<td>0,685</td>
<td>-1,4578</td>
<td></td>
</tr>
<tr>
<td>-2,4534</td>
<td>1,3356</td>
<td>-0,574</td>
<td></td>
</tr>
<tr>
<td>-1,3275</td>
<td>-0,0099</td>
<td>-1,2825</td>
<td></td>
</tr>
</tbody>
</table>
Table S16. Source data values for figure 5D. Expression of GATA3.2 in asymptomatic MS patients and healthy controls.

<table>
<thead>
<tr>
<th>MS patient</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0262</td>
<td>0.027</td>
</tr>
<tr>
<td>0.0147</td>
<td>0.0387</td>
</tr>
<tr>
<td>0.0196</td>
<td>0.0282</td>
</tr>
<tr>
<td>0.0172</td>
<td>0.0206</td>
</tr>
<tr>
<td>0.0193</td>
<td>0.0198</td>
</tr>
<tr>
<td>0.0342</td>
<td>0.018</td>
</tr>
<tr>
<td>0.0156</td>
<td>0.0189</td>
</tr>
<tr>
<td>0.0178</td>
<td>0.0243</td>
</tr>
<tr>
<td>0.0212</td>
<td>0.0252</td>
</tr>
<tr>
<td>0.0168</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary data

Data S1. Five methylation arrays covering all known enhancers based on previous publications using DNase-seq. DNA from naïve T-cells, Th0, Th1, Th2 and Th17 cells was collected after 7 days. Arrays were then quantile normalized to correct for inter-array differences in intensity. As only one replicate of each sample was assayed, statistic cannot be performed between samples. On a gene by gene basis, known positive (methylated) controls such as brain and testis-specific genes were methylated, whereas known negative (unmethylated) controls such as the GAPDH promoter were unmethylated confirming the robustness of the remaining data.

Data S2. A validated network, Cytoscape file.