RESEARCH ARTICLE

HUMAN IMMUNOLOGY

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

Mika Gustafsson,[1,2]*† Danuta R. Gawel,[1]† Lars Alfredsson,[3] Sergio Baranzini,[4] Janne Björkander,[5] Robert Blomgran,[6] Sandra Hellberg,[7] Daniel Eklund,[8] Jan Ernerudh,[7,8] Ingrid Kockum,[9] Aelita Konstantinell,[1,10] Riita Lahesmaa,[11] Antonio Lentini,[1] H. Robert I. Liljenström,[1] Lina Mattson,[1] Andreas Matussek,[5] Johan Mellergård,[12] Melissa Mendez,[13] Tomas Olsson,[9] Miguel A. Pujana,[14] Omid Rasool,[11] Jordi Serra-Musach,[14] Margaretha Stenmarker,[5] Subhash Tripathi,[11] Miro Viitala,[11] Hui Wang,[1,15] Huan Zhang,[1] Colm E. Nestor,[1]† Mikael Benson[1]*†

Early regulators of disease may increase understanding of disease mechanisms and serve as markers for presymptomatic 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 (chromatin immunoprecipitation sequencing) and small interfering RNA knockdowns. Differential mRNA expression of the TFs and their targets in T cell–associated diseases supports 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 GRNs with genome-wide association studies may be generally applicable for functional and clinical studies of early disease development.

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 to follow 100,000 subjects for 20 to 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. Because there is currently limited understanding of evolving disease processes, we will henceforth refer to any regulatory gene that occurs before symptomatic stages as "early."

Here, we hypothesized that early regulators with diagnostic potential in T cell–associated diseases can be systematically inferred by (i) constructing a gene regulatory network (GRN) of T cell differentiation and (ii) prioritizing hub transcription factors (TFs) in that GRN, which are enriched for disease-associated single-nucleotide polymorphisms (SNPs) identified by genome-wide association studies (GWAS).

The background to the hypotheses is previous studies showing that early TFs can be inferred on the basis of 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 presymptomatic patients, a GRN should ideally be constructed on the basis of time series analyses of a cellular model of an evolving disease process. Here, we focused on CD4$^+$ T cell–associated diseases and upstream 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*).

[1]The Centre for Individualised Medicine, Department of Clinical and Experimental Medicine, Division of Pediatrics, Linköping University, SE-581 83 Linköping, Sweden. [2]Bioinformatics, Department of Physics, Chemistry, and Biology, Linköping University, SE-581 83 Linköping, Sweden. [3]Institute of Environmental Medicine, Karolinska Institutet, SE-171 77 Solna, Sweden. [4]Department of Neurology, University of California, San Francisco, CA 94158, USA. [5]Futurum-Academy for Health and Care, County Council of Jönköping, SE-551 85 Jönköping, Sweden. [6]Department of Clinical and Experimental Medicine, Division of Microbiology and Molecular Medicine, Linköping University, SE-581 83 Linköping, Sweden. [7]Department of Clinical and Experimental Medicine, Division of Clinical Immunology, Unit of Autoimmunity and Immune Regulation, Linköping University, SE-581 83 Linköping, Sweden. [8]Department of Clinical Immunology and Transfusion Medicine, Linköping University, SE-581 83 Linköping, Sweden. [9]Department of Clinical Neurosciences, Karolinska Institutet and Centrum for Molecular Medicine, SE-171 77 Stockholm, Sweden. [10]Department of Medical Biology, The Arctic University of Norway, NO-9037 Tromsø, Norway. [11]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, FI-20520 Turku, Finland. [12]Department of Neurology and Department of Clinical and Experimental Medicine, Linköping University, SE-581 83 Linköping, Sweden. [13]Laboratorio de Investigación en Enfermedades Infecciosas, LID, Universidad Peruana Cayetano Heredia, Lima PE-15102, Peru. [14]Program Against Cancer Therapeutic Resistance (ProCURE), Cancer and Systems Biology Unit, Catalan Institute of Oncology, IDIBELL, L'Hospitalet del Llobregat, ES-08908 Barcelona, Spain. [15]Department of Immunology, MD Anderson Cancer Center, Houston, TX 77030, USA.
*Corresponding author. E-mail: mika.gustafsson@liu.se (M.G.); mikael.benson@liu.se (M.B.)
†These authors contributed equally to this work and should be regarded as shared first or last authors, respectively.

Thus, we hypothesized that a GRN of T cell differentiation could serve as a model of evolving disease processes in T cell–associated disease to infer early TFs with diagnostic potential for early disease detection (see Fig. 1 for an outline of the study).

RESULTS

Early $\mathrm{T_H}1/\mathrm{T_H}2$ TFs were enriched for disease-associated SNPs

To identify early 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 ($\mathrm{T_H}$) subsets, namely, $\mathrm{T_H}1$, $\mathrm{T_H}2$, $\mathrm{T_H}17$, and $\mathrm{T_{regs}}$ (T regulatory 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 data set of human T cell differentiation generated to date. Each time point displayed a distinct expression profile (Fig. 2A and fig. S1), and appropriate differentiation of each subset was confirmed by measurement of signature gene expression in each subset by quantitative real-time polymerase chain reaction (qPCR) (fig. S1) (*9*). Having obtained a list of all predicted human TFs (n = 1750) from the animal transcription factor database (AnimalTFDB) (*10*), we identified TFs that were differentially expressed [t $\text{test}_{\text{limma}}$ Benjamini-Hochberg false discovery rate (FDR) < 0.1] between subsets at 6 or 24 hours, 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 Materials and Methods and table S1). We found that the only TF set with statistically significant enrichment for disease-associated SNPs, defined as those with a GWA of $P < 1 \times 10^{-5}$, was the upstream $\mathrm{T_H}1/\mathrm{T_H}2$ TFs; odds ratio (OR) = 2.7, Fisher exact test $P = 1.0 \times 10^{-7}$ [Figs. 1 (step 2) and 2B, figs. S2 and S3, and table S7].

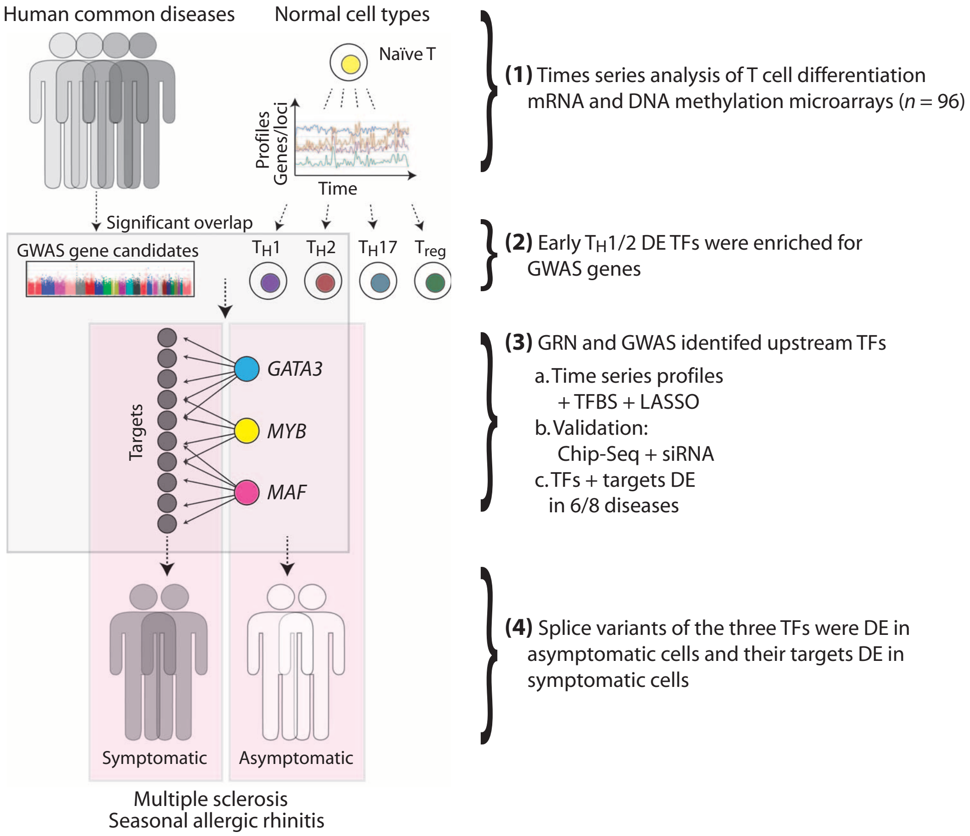

Human common diseases
Normal cell types
Naïve T
Profiles
Genes/loci
Time
Significant overlap
GWAS gene candidates
$\mathrm{T_H}1$
$\mathrm{T_H}2$
$\mathrm{T_H}17$
$\mathrm{T_{reg}}$
Targets
GATA3
MYB
MAF
Symptomatic
Asymptomatic
Multiple sclerosis
Seasonal allergic rhinitis
(1) Times series analysis of T cell differentiation mRNA and DNA methylation microarrays (n = 96)
(2) Early $\mathrm{T_H}1/2$ DE TFs were enriched for GWAS genes
(3) GRN and GWAS identifed 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

Fig. 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 TFs that were differentially expressed (DE) at 6 and 24 hours between $\mathrm{T_H}1$ and $\mathrm{T_H}2$ subsets were enriched for disease-associated SNPs from GWAS. (**3**) A GRN for $\mathrm{T_H}1$ and $\mathrm{T_H}2$ polarization was inferred by combining the microarray data from (1) with predicted TF binding sites (TFBSs) using sparse penalized regression [least absolute shrinkage and selection operator (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 chromatin immunoprecipitation sequencing (ChIP-seq) and small interfering RNA (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 three 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 three TFs were differentially expressed in asymptomatic cells and their targets differentially expressed in symptomatic cells.

Early TFs in T cell–associated diseases were identified by combining a GRN with GWAS data

To prioritize upstream hub TFs for further investigation, we proceeded to construct the first large-scale $\mathrm{T_H}1/\mathrm{T_H}2$ GRN. This was constructed from the $\mathrm{T_H}1/\mathrm{T_H}2$ time series expression profiling data and predicted TFBSs (*11*), using the LASSO algorithm (Fig. 1, step 3) (*12*). In addition, we performed DNA methylation profiling of the same cells 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 6112 predicted interactions from 378 TFs onto 1563 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 $P < 3.2 \times 10^{-12}$). Remarkably, 10 GWAS TFs were among the 11 TFs with most targets (Fig. 2C). To prioritize among those 10 GWAS TFs for further study, we repeated the analyses using all nominally associated disease SNPs ($P < 1 \times 10^{-3}$). We found that 3 TFs, namely, *GATA3*, *MAF*, and *MYB*, were associated with 91% of the disease SNPs among the 10 TFs identified above (OR = 9.22, bootstrap $P = 6.6 \times 10^{-3}$; Fig. 2D and fig. S5).

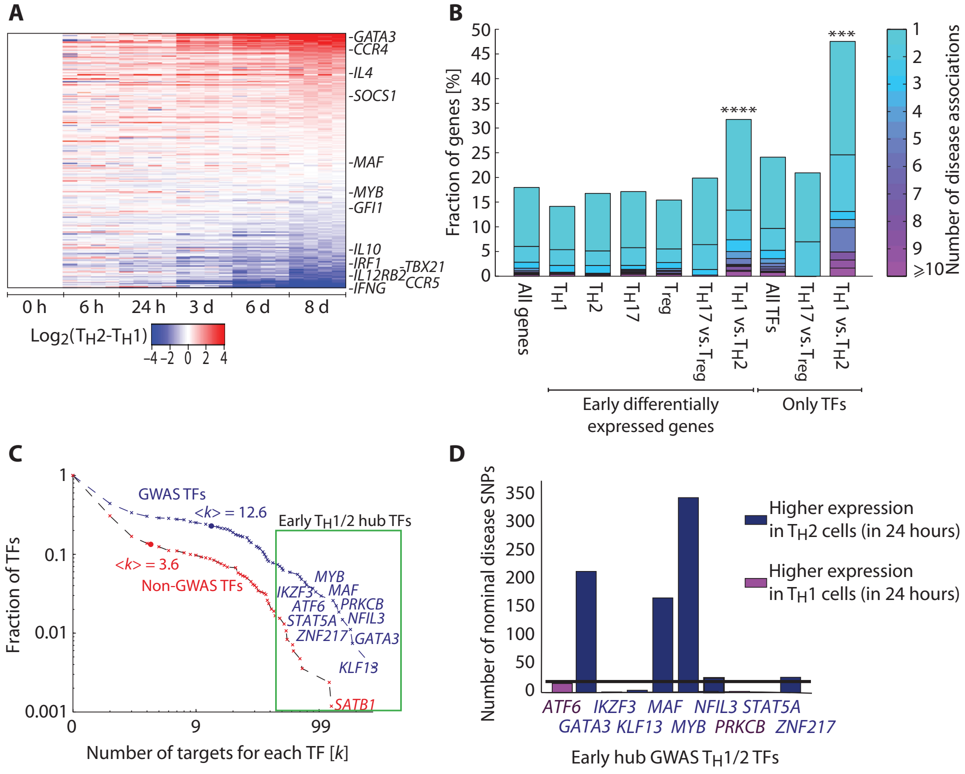


Fig. 2. Time series analysis of T cell differentiation identified early hub TFs enriched for disease-associated SNPs identified by GWAS. (**A**) Heat map of the expression of subset-specific genes from (*9*) reveals appropriate expression of T_H lineage markers in polarized cells. (**B**) Enrichment of genes with disease-associated SNPs identified by GWAS from differential expression analysis of time series microarrays ($P < 0.05$ and Benjamini-Hochberg FDR < 0.1). The subsets derived from individual T cell subsets ("T_H1", "T_H2", "T_H17", and "T_{reg}") are based on the differentially expressed genes between the 6 and 24 hours within each cell polarization. The P values were calculated using Fisher's exact test relative to the background level (****$P < 1.0 \times 10^{-12}$, ***$P < 1.0 \times 10^{-6}$). 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 ($P < 3.2 \times 10^{-12}$). Gene names are displayed for the first 11 upstream TFs in T_H1/T_H2 differentiation. (**D**) *GATA3*, *MAF*, and *MYB*, which were differentially expressed in T_H1/T_H2, 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.

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

To directly validate the GWAS-GRN, we performed ChIP of GATA3, MAF, and MYB in differentiated human T_H1 and T_H2 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 TFs, both MAF and MYB showed narrow and pronounced binding at the transcription start sites 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 T_H1. Over 70% of peaks of MYB binding in replicate 1 were also present in replicate 2 (OR = 36.4, bootstrap $P < 1 \times 10^{-5}$), whereas 79% of peaks of MAF binding identified in replicate 1 were also present in replicate 2 (OR = 39.9, bootstrap $P < 1 \times 10^{-5}$), verifying the reproducibility of the ChIP-seq assays (Fig. 3C).

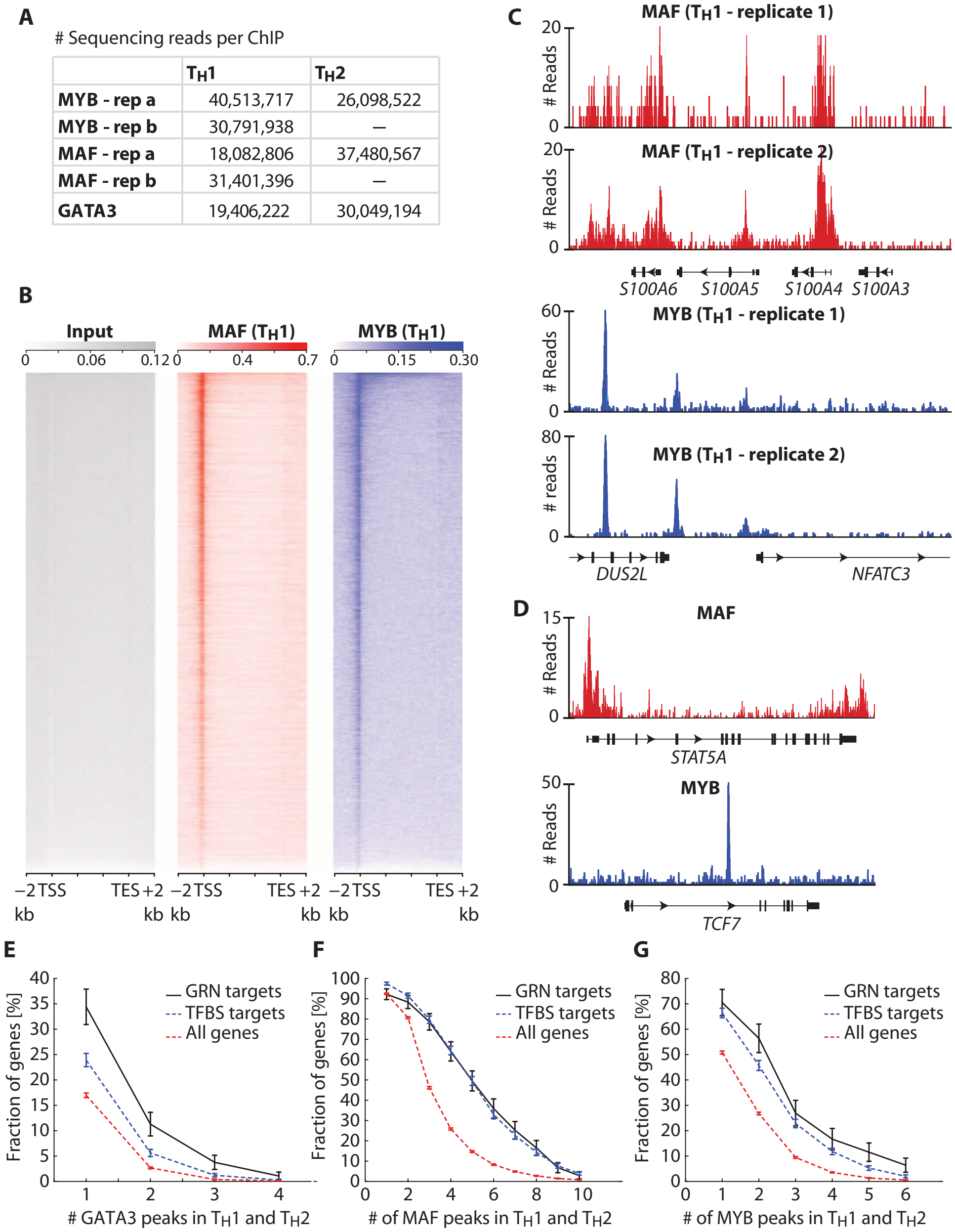
GATA3, MAF, and MYB bindings with higher confidence (lower peak P value) were significantly enriched [OR_{GATA3} = 7.30, weighted bootstrap (Supplementary Materials and Methods), $P_{GATA3} < 1 \times 10^{-81}$; OR_{MAF} = 1.92, $P_{MAF} < 1 \times 10^{-22}$; OR_{MYB} = 3.17, $P_{MYB} < 1 \times 10^{-27}$] for the GRN-predicted targets of the three TFs (Fig. 3, D to G). This finding was also confirmed by comparisons with previously published independent GATA3 ChIP-seq data (*13*) (OR = 2.91, bootstrap $P < 1 \times 10^{-21}$; fig. S18D). We then tested if the GRN predictions that used both TFBS and time series data yielded higher overlap than those 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 $P < 1 \times 10^{-5}$; Supplementary Materials and Methods, Fig. 3, E to G, fig. S6, and table S12). In further support of the accuracy of the GRN, comparisons with siRNA-mediated knockdowns of *GATA3* and *MAF* in T_H2, 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. S18, B 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 and 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, RA [Gene Expression Omnibus (GEO) accession GSE4588], MS (*15*), SLE (GEO accession GSE4588), HES (*16*), CLL (*17*), ATL (*18*), AML (*19*), and SAR (*20*). We found that the three TFs were differentially expressed in six of the diseases, and most of their predicted targets are in all eight diseases (Fig. 4B, figs. S16 and 17, and table S11). We also made similar observations in original expression profiling studies of $CD4^+$ T cells from patients with malignant (breast cancer) and infectious diseases (tuberculosis and influenza) (Fig. 4B).

Several disease-associated SNPs were in linkage disequilibrium with splice affecting regions (table S5), and eQTL analyses supported that



# Sequencing reads per ChIP

	T_H1	T_H2
MYB - rep a	40,513,717	26,098,522
MYB - rep b	30,791,938	—
MAF - rep a	18,082,806	37,480,567
MAF - rep b	31,401,396	—
GATA3	19,406,222	30,049,194

Fig. 3. ChIP-seq of *MAF*, *MYB*, and *GATA3* in differentiated human T_H1 and T_H2 validates the GWAS-GRN. (**A**) The number of 50-base pair single-end reads successfully aligned to the human genome (hg19) for each ChIP assay. Matching input samples were sequenced for each IP. (**B**) Heat maps showing enrichment of *MYB* and *MAF* across gene bodies. Each row represents a gene, ordered from top to bottom by total enrichment levels. TSS, transcription start site; TES, transcription end sites. (**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 GRN approach. Input (gray), *MAF* (red), and *MYB* (blue). (**E** to **G**) Enriched binding for *GATA3* (E), *MAF* (F), and *MYB* (G) predicted targets in T_H1 and T_H2 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 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 ($P_{GATA3} = 1.80 \times 10^{-6}$, $OR_{GATA3} = 1.62$; $P_{MYB} = 0.01$, $OR_{MYB} = 1.29$).

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 (Fig. 4C and Supplementary Materials and 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 the identification of early regulators of disease, direct detection of such regulators would require studies of numerous molecular layers, in numerous cells types, at numerous time points, in thousands of healthy individuals over many years. Instead, we used the asymptomatic stage of 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 because of 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 unstimulated $CD4^+$ T cells outside of the pollen season, when patients are asymptomatic. In this model system, we propose that unstimulated 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) (figs. S7 and 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), and all measured splice variants of *MYB* were differentially expressed, of which variants 4 (NM_001161656.1) and 5 (NM_001161657.1) were most significant (Fig. 5A). The microarray expression levels were validated by qPCR (fig. S8, A to C). The mean expression levels of these splice variants not only separated patients and controls with high accuracy [area under the precision recall curve (AUC-PR) = 0.83, $P = 2.9 \times 10^{-3}$; Fig. 5B] but also significantly correlated with the symptom

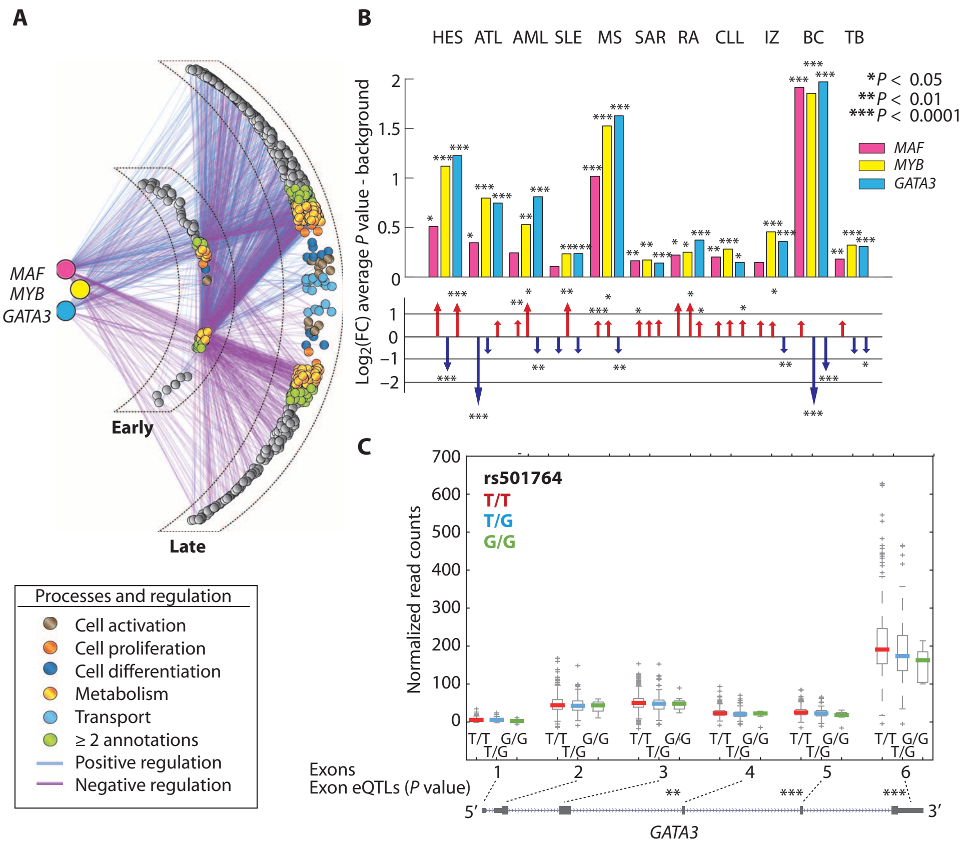


Fig. 4. Predicted targets of *GATA3*, *MAF*, and *MYB* are differentially expressed in T cell–associated diseases. (**A**) Gene ontology enrichment analysis of predicted targets of *GATA3*, *MAF*, and *MYB* from the T_H1/T_H2 GRN at the early-stage and all late-stage genes. (**B**) Inferred targets of *GATA3*, *MAF*, and *MYB* have in general lower *P* values than nontarget genes in nine T cell–related diseases: hyper eosinophilic syndrome (HES), adult T cell leukemia/lymphoma (ATL), acute myeloid leukemia (AML), systemic lupus erythematosus (SLE), MS, SAR, influenza (IZ), breast cancer (BC), and tuberculosis (TB). (Lower) Arrows depicting $\log_2$ fold changes [$\log_2$(FC)] of the expression of each TF in patients compared to controls. Magnitudes of the arrows are depicted as follows: largest [abs $\log_2(FC) > 2$], middle [abs $1 < \log_2(FC) < 2$], smallest [abs $\log_2(FC) < 1$]. Red up-facing and blue down-facing arrows depict $\log_2(FC) > 0$ and $\log_2(FC) < 0$, 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**) Expression quantitative trait loci (eQTL) analysis of the exons of *GATA3*. Normalized RNA sequencing (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 ($**P < 0.01$, $***P < 0.001$). RA, rheumatoid arthritis; CLL, chronic lymphocytic leukemia.

scores in patients during the pollen season (Pearson's correlation coefficient = 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, $P = 0.039$).

To dissect the roles of the variants, we measured the splice variants in T_H1/T_H2 differentiation by qPCR and analyzed the structure of the differences of corresponding proteins (figs. S9 and S10). We then aimed at functionally validating the disease relevance of the splice variants. siRNA mediated knockdown of *MAF* splice variant 2 followed by T_H2 polarization and exon array analysis (figs. S11 and S12, table S10). *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). These genes were also the *MAF* targets with the lowest *P* values in allergen-challenged T cells from SAR patients (bootstrap $P < 1 \times 10^{-5}$; fig. S13 and 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-year treatment, both compared to the expression before and during the treatment, AUC-PR = 0.85 ($P < 2.9 \times 10^{-3}$) and AUC-PR = 0.95 ($P < 1.5 \times 10^{-4}$), 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 unstimulated $CD4^+$ T cell from MS patients in remission showed a 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 Materials and Methods description) showed a significant decrease of the *GATA3* gene during remission compared to controls (t $\text{test}_{\text{limma}}$ $P = 0.033$), as well as a decrease in remission compared to relapse (paired t $\text{test}_{\text{limma}}$ $P = 1.4 \times 10^{-3}$), and significant differential expression of its predicted targets during relapse compared to controls (bootstrap $P < 1 \times 10^{-12}$). The predicted targets of *MAF* and *MYB* were also differentially expressed during relapse ($P = 5.3 \times 10^{-3}$ and $P = 1.6 \times 10^{-4}$, 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 ~25,000 patients and controls (*21*). For *MAF*, *MYB*, and *GATA3*, we found significant SNPs ($P_{MAF} = 1.9 \times 10^{-5}$, $P_{MYB} = 5.8 \times 10^{-6}$, $P_{GATA3} = 6.6 \times 10^{-5}$). Indeed, the three TFs were among the 1% most enriched for disease-associated SNPs (Fisher exact test $P = 3 \times 10^{-6}$). We further asked if the number of disease-associated SNPs correlated with disease severity in a cohort

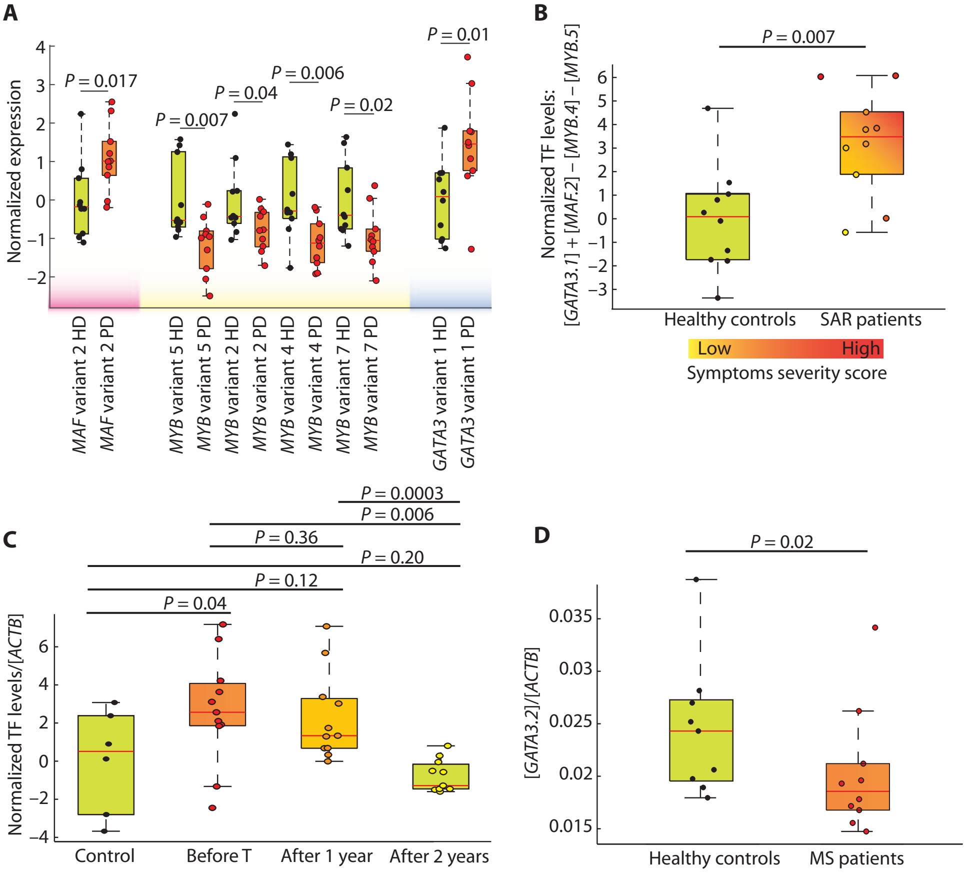


Fig. 5. Splice variants of *GATA3*, *MAF*, and *MYB* are differentially expressed in asymptomatic patients. (**A**) Splice variant expression in unchallenged cells of 10 SAR patients (SAR) and 10 healthy controls measured by exon microarrays. PD, patient unstimulated cells; HD, healthy control unstimulated cells (table S13). (**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 and healthy subjects are marked with green color; patients are marked to indicate symptom severity. Yellow color indicates low, orange indicates medium, and red indicates high symptom severity (table S14). (**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 1 and 2 years of immunotherapy (T) (table S15). (**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 and D), nonparametric Wilcoxon test was used. The box plots depict median (horizontal line), interquartile range (boxes), and whiskers (dashed line), and circles represent individuals (table S16).

consisting of 2085 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 Materials and 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$_H$ 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 GRNs. We present an analytical strategy to identify early 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 the asymptomatic stages of the two diseases would serve as models of early, presymptomatic stages. Using exon profiling and qPCR in SAR and MS, respectively, we revealed differential expression of splice variants of *GATA3*, *MAF*, and *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 *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 such as RA (*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*).

The specificity might be increased by extending the GRN to include other potential early genetic or epigenetic regulators, like signaling molecules, noncoding RNAs, or histone modifiers, all of which can be prioritized on the basis of 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 on the basis of in vitro derivation of epithelial cells and validated by functional experiments in primary or transformed epithelial cell lines. Potential early regulators can be identified on the basis of 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 downstream 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.

MATERIALS AND METHODS

Study design

The overall objective of this study was to identify early regulators of T cell–associated diseases by identifying upstream TFs in T cell differentiation and by prioritizing hub TFs that were enriched for disease-associated polymorphisms. Upstream TFs were identified by constructing a 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 T$_H$ 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, MS and SAR, were analyzed during asymptomatic and symptomatic stages of both diseases. Sample sizes were determined on the basis of our previous studies (*26*) 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 the 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 (*27*), and for comparisons between two states by using *t* test from the LIMMA package in R, with 10% FDR according to the Benjamini-Hochberg method ($P < 0.05$). The *P* values for qPCR were calculated using one-sided nonparametric Wilcoxon test. In several instances, we tested the enrichment of low *P* values within a set of genes (for example, target genes of a TF). This was performed using a bootstrap test on the average $\log_{10} P$ value of the target set, where *P* values were estimated from 1×10^6 randomizations using (*28*). For set enrichment analysis (for example, GWAS genes), *P* values were calculated using one-sided Fisher exact test using all genes as background ($n = 22{,}500$).

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/7/313/313ra178/DC1
Materials and Methods
Fig. S1. Expression of signature genes in T cell subsets.
Fig. S2. Comparison of GWAS enrichment for gene sets defined from alternative strategies defined by the T cell differentiation microarray data.
Fig. S3. GWAS SNPs in different populations.
Fig. S4. Methylation probe levels of selected housekeeping and tissue-specific genes.
Fig. S5. Enrichment in GWAS SNPs in GATA3, MAF, and MYB in T cell diseases, infectious diseases, and malignancies.
Fig. S6. Overlap between different methods to define GATA3 bindings.
Fig. S7. GRN-predicted targets of GATA3, MAF, and MYB are differentially expressed in six T cell–related diseases.
Fig. S8. Validation of the differentially expressed splice variants in SAR.
Fig. S9. Time series analyses of GATA3, MAF, and MYB splice variant expression during T$_H$1 and T$_H$2 differentiation.
Fig. S10. Predicted effects of alternative splicing on the protein structures of GATA3, MAF, and MYB.
Fig. S11. Heat maps of the top 50 most differentially expressed genes in siRNA-mediated MAF knockdown using four different vectors.
Fig. S12. Effect of MAF knockdown on T$_H$1, T$_H$2, T$_H$17, and T$_{reg}$ signature genes.
Fig. S13. MAF siRNA splice–specific targets.
Fig. S14. Expression profiles of SAR patients and controls show a consisting overlap even for the most significantly differentially expressed genes.
Fig. S15. Expression profiles of MS patients and controls show a consisting overlap even for the most significantly differentially expressed genes.
Fig. S16. Expression overlap between patients and controls in many T cell–associated diseases.
Fig. S17. Expression overlap between patients and controls in many T cell–associated diseases.
Fig. S18. GRN validation.
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 Ingenuity Pathway Analysis.
Table S3. All pathways for the early predicted targets of all early TFs.
Table S4. Statistics of all eight 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 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 naïve questions "how many protein-protein interactions (PPIs) have the T$_H$ differentially expressed genes, globally or just early," and "would the TFs be identified 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* (NM_001031804 and NM_005360) 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 Fig. 5A.

Table S14. Source data values for Fig. 5B.
Table S15. Source data values for Fig. 5C.
Table S16. Source data values for Fig. 5D.
Data S1. Five methylation arrays covering all known enhancers based on previous publications using DNase I hypersensitive sites sequencing.
Data S2. A validated network, Cytoscape file.
References (*29–47*)

REFERENCES AND NOTES

1. L. Hood, N. D. Price, Demystifying disease, democratizing health care. *Sci. Transl. Med.* **6**, 225ed5 (2014).
2. A. Aytes, A. Mitrofanova, C. Lefebvre, M. J. Alvarez, M. Castillo-Martin, T. Zheng, J. A. Eastham, A. Gopalan, K. J. Pienta, M. M. Shen, A. Califano, C. Abate-Shen, Cross-species regulatory network analysis identifies a synergistic interaction between *FOXM1* and *CENPF* that drives prostate cancer malignancy. *Cancer Cell* **25**, 638–651 (2014).
3. P. Sumazin, X. Yang, H.-S. Chiu, W.-J. Chung, A. Iyer, D. Llobet-Navas, P. Rajbhandari, M. Bansal, P. Guarnieri, J. Silva, A. Califano, An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* **147**, 370–381 (2011).
4. J. C. Chen, M. J. Alvarez, F. Talos, H. Dhruv, G. E. Rieckhof, A. Iyer, K. L. Diefes, K. Aldape, M. Berens, M. M. Shen, A. Califano, Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell* **159**, 402–414 (2014).
5. M. Gustafsson, M. Edström, D. Gawel, C. E. Nestor, H. Wang, H. Zhang, F. Barrenäs, J. Tojo, I. Kockum, T. Olsson, J. Serra-Musach, N. Bonifaci, M. A. Pujana, J. Ernerudh, M. Benson, Integrated genomic and prospective clinical studies show the importance of modular pleiotropy for disease susceptibility, diagnosis and treatment. *Genome Med.* **6**, 17 (2014).
6. D. Engelbertsen, L. Andersson, I. Ljungcrantz, M. Wigren, B. Hedblad, J. Nilsson, H. Björkbacka, T-helper 2 immunity is associated with reduced risk of myocardial infarction and stroke. *Arterioscler. Thromb. Vasc. Biol.* **33**, 637–644 (2013).
7. K. Hung, R. Hayashi, A. Lafond-Walker, C. Lowenstein, D. Pardoll, H. Levitsky, The central role of CD4$^+$ T cells in the antitumor immune response. *J. Exp. Med.* **188**, 2357–2368 (1998).
8. X. Cheng, J. Wang, N. Xia, X.-X. Yan, T.-T. Tang, H. Chen, H.-J. Zhang, J. Liu, W. Kong, S. Sjöberg, E. Folco, P. Libby, Y.-H. Liao, G.-P. Shi, A guanidine-rich regulatory oligodeoxynucleotide improves type-2 diabetes in obese mice by blocking T-cell differentiation. *EMBO Mol. Med.* **4**, 1112–1125 (2012).
9. H. Zhang, C. E. Nestor, S. Zhao, A. Lentini, B. Bohle, M. Benson, H. Wang, Profiling of human CD4$^+$ T-cell subsets identifies the T_H2-specific noncoding RNA *GATA3-AS1*. *J. Allergy Clin. Immunol.* **132**, 1005–1008 (2013).
10. H.-M. Zhang, H. Chen, W. Liu, H. Liu, J. Gong, H. Wang, A.-Y. Guo, AnimalTFDB: A comprehensive animal transcription factor database. *Nucleic Acids Res.* **40**, D144–D149 (2012).
11. J. Ernst, H. L. Plasterer, I. Simon, Z. Bar-Joseph, Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.* **20**, 526–536 (2010).
12. J. H. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
13. A. Kanhere, A. Hertweck, U. Bhatia, M. R. Gökmen, E. Perucha, I. Jackson, G. M. Lord, R. G. Jenner, T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements. *Nat. Commun.* **3**, 1268 (2012).
14. S. Bruhn, Y. Fang, F. Barrenäs, M. Gustafsson, H. Zhang, A. Konstantinell, A. Krönke, B. Sönnichsen, A. Bresnick, N. Dulyaninova, H. Wang, Y. Zhao, J. Klingelhöfer, N. Ambartsumian, M. K. Beck, C. Nestor, E. Bona, Z. Xiang, M. Benson, A generally applicable translational strategy identifies S100A4 as a candidate gene in allergy. *Sci. Transl. Med.* **6**, 218ra4 (2014).
15. J.-C. Corvol, D. Pelletier, R. G. Henry, S. J. Caillier, J. Wang, D. Pappas, S. Casazza, D. T. Okuda, S. L. Hauser, J. R. Oksenberg, S. E. Baranzini, Abrogation of T cell quiescence characterizes patients at high risk for multiple sclerosis after the initial neurological event. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 11839–11844 (2008).
16. M. Ravoet, C. Sibille, C. Gu, M. Libin, B. Haibe-Kains, C. Sotiriou, M. Goldman, F. Roufosse, K. Willard-Gallo, Molecular profiling of CD3$^-$CD4$^+$ T cells from patients with the lymphocytic variant of hypereosinophilic syndrome reveals targeting of growth control pathways. *Blood* **114**, 2969–2983 (2009).
17. G. Görgün, T. A. W. Holderried, D. Zahrieh, D. Neuberg, J. G. Gribben, Chronic lymphocytic leukemia cells induce changes in gene expression of CD4 and CD8 T cells. *J. Clin. Invest.* **115**, 1797–1805 (2005).
18. C. A. Pise-Masison, M. Radonovich, K. Dohoney, J. C. Morris, D. O'Mahony, M.-J. Lee, J. Trepel, T. A. Waldmann, J. E. Janik, J. N. Brady, Gene expression profiling of ATL patients: Compilation of disease-related genes and evidence for TCF4 involvement in BIRC5 gene expression and cell viability. *Blood* **113**, 4016–4026 (2009).
19. R. Le Dieu, D. C. Taussig, A. G. Ramsay, R. Mitter, F. Miraki-Moud, R. Fatah, A. M. Lee, T. A. Lister, J. G. Gribben, Peripheral blood T cells in acute myeloid leukemia (AML) patients at diagnosis have abnormal phenotype and genotype and form defective immune synapses with AML blasts. *Blood* **114**, 3909–3916 (2009).
20. C. E. Nestor, F. Barrenäs, H. Wang, A. Lentini, H. Zhang, S. Bruhn, R. Jörnsten, M. A. Langston, G. Rogers, M. Gustafsson, M. Benson, DNA methylation changes separate allergic patients from healthy controls and may reflect altered CD4$^+$ T-cell population structure. *PLOS Genet.* **10**, e1004059 (2014).
21. S. Sawcer, G. Hellenthal, M. Pirinen, C. C. Spencer, N. A. Patsopoulos, L. Moutsianas, A. Dilthey, Z. Su, C. Freeman, S. E. Hunt, S. Edkins, E. Gray, D. R. Booth, S. C. Potter, A. Goris, G. Band, A. B. Oturai, A. Strange, J. Saarela, C. Bellenguez, B. Fontaine, M. Gillman, B. Hemmer, R. Gwilliam, F. Zipp, A. Jayakumar, R. Martin, S. Leslie, S. Hawkins, E. Giannoulatou, S. D'alfonso, H. Blackburn, F. Martinelli Boneschi, J. Liddle, H. F. Harbo, M. L. Perez, A. Spurkland, M. J. Waller, M. P. Mycko, M. Ricketts, M. Comabella, N. Hammond, I. Kockum, O. T. McCann, M. Ban, P. Whittaker, A. Kemppinen, P. Weston, C. Hawkins, S. Widaa, J. Zajicek, S. Dronov, N. Robertson, S. J. Bumpstead, L. F. Barcellos, R. Ravindrarajah, R. Abraham, L. Alfredsson, K. Ardlie, C. Aubin, A. Baker, K. Baker, S. E. Baranzini, L. Bergamaschi, R. Bergamaschi, A. Bernstein, A. Berthele, M. Boggild, J. P. Bradfield, D. Brassat, S. A. Broadley, D. Buck, H. Butzkueven, R. Capra, W. M. Carroll, P. Cavalla, E. G. Celius, S. Cepok, R. Chiavacci, F. Clerget-Darpoux, K. Clysters, G. Comi, M. Cossburn, I. Cournu-Rebeix, M. B. Cox, W. Cozen, B. A. Cree, A. H. Cross, D. Cusi, M. J. Daly, E. Davis, P. I. de Bakker, M. Debouverie, M. B. D'hooghe, K. Dixon, R. Dobosi, B. Dubois, D. Ellinghaus, I. Elovaara, F. Esposito, C. Fontenille, S. Foote, A. Franke, D. Galimberti, A. Ghezzi, J. Glessner, R. Gomez, O. Gout, C. Graham, S. F. Grant, F. R. Guerini, H. Hakonarson, P. Hall, A. Hamsten, H. P. Hartung, R. N. Heard, S. Heath, J. Hobart, M. Hoshi, C. Infante-Duarte, G. Ingram, W. Ingram, T. Islam, M. Jagodic, M. Kabesch, A. G. Kermode, T. J. Kilpatrick, C. Kim, N. Klopp, K. Koivisto, M. Larsson, M. Lathrop, J. S. Lechner-Scott, M. A. Leone, V. Leppä, U. Liljedahl, I. L. Bomfim, R. R. Lincoln, J. Link, J. Liu, A. R. Lorentzen, S. Lupoli, F. Macciardi, T. Mack, M. Marriott, V. Martinelli, D. Mason, J. L. McCauley, F. Mentch, I.-L. Mero, T. Mihalova, X. Montalban, J. Mottershead, K.-M. Myhr, P. Naldi, W. Ollier, A. Page, A. Palotie, J. Pelletier, L. Piccio, T. Pickersgill, F. Piehl, S. Pobywajlo, H. L. Quach, P. P. Ramsay, M. Reunanen, R. Reynolds, J. D. Rioux, M. Rodegher, S. Roesner, J. P. Rubio, I. M. Rückert, M. Salvetti, E. Salvi, A. Santaniello, C. A. Schaefer, S. Schreiber, C. Schulze, R. J. Scott, F. Sellebjerg, K. W. Selmaj, D. Sexton, L. Shen, B. Simms-Acuna, S. Skidmore, P. M. Sleiman, C. Smestad, P. S. Sørensen, H. B. Søndergaard, J. Stankovich, R. C. Strange, A. M. Sulonen, E. Sundqvist, A. C. Syvänen, F. Taddeo, B. Taylor, J. M. Blackwell, P. Tienari, E. Bramon, A. Tourbah, M. A. Brown, E. Tronczynska, J. P. Casas, N. Tubridy, A. Corvin, J. Vickery, J. Jankowski, P. Villoslada, H. S. Markus, K. Wang, C. G. Mathew, J. Wason, C. N. Palmer, H. E. Wichmann, R. Plomin, E. Willoughby, A. Rautanen, J. Winkelmann, M. Wittig, R. C. Trembath, J. Yaouanq, A. C. Viswanathan, H. Zhang, N. W. Wood, R. Zuvich, P. Deloukas, C. Langford, A. Duncanson, J. R. Oksenberg, M. A. Pericak-Vance, J. L. Haines, T. Olsson, J. Hillert, A. J. Ivinson, P. L. De Jager, L. Peltonen, G. J. Stewart, D. A. Hafler, S. L. Hauser, G. McVean, P. Donnelly, A. Compston; International Multiple Sclerosis Genetics Consortium; Wellcome Trust Case Control Consortium 2, Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
22. T. Takeuchi, K. Suzuki, *CD247* variants and single-nucleotide polymorphisms observed in systemic lupus erythematosus patients. *Rheumatology* **52**, 1551–1555 (2013).
23. M. J. H. Coenen, J. M. Hofstra, H. Debiec, H. C. Stanescu, A. J. Medlar, B. Stengel, A. Boland-Augé, J. M. Groothuismink, D. Bockenhauer, S. H. Powis, P. W. Mathieson, P. E. Brenchley, R. Kleta, J. F. M. Wetzels, P. Ronco, Phospholipase A2 receptor (*PLA2R1*) sequence variants in idiopathic membranous nephropathy. *J. Am. Soc. Nephrol.* **24**, 677–683 (2013).
24. H. Zhang, M. Gustafsson, C. Nestor, K. F. Chung, M. Benson, Targeted omics and systems medicine: Personalising care. *Lancet Respir. Med.* **2**, 785–787 (2014).
25. F. Barrenäs, S. Chavali, A. C. Alves, L. Coin, M.-R. Jarvelin, R. Jörnsten, M. A. Langston, A. Ramasamy, G. Rogers, H. Wang, M. Benson, Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. *Genome Biol.* **13**, R46 (2012).
26. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
27. A. Conesa, M. J. Nueda, A. Ferrer, M. Talón, maSigPro: A method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* **22**, 1096–1102 (2006).
28. T. A. Knijnenburg, L. F. A. Wessels, M. J. T. Reinders, I. Shmulevich, Fewer permutations, more accurate *P*-values. *Bioinformatics* **25**, i161–i168 (2009).
29. H. Wang, F. Barrenäs, S. Bruhn, R. Mobini, M. Benson, Increased IFN-γ activity in seasonal allergic rhinitis is decreased by corticosteroid treatment. *J. Allergy Clin. Immunol.* **124**, 1360–1362 (2009).
30. M. Benson, M. A. Langston, M. Adner, B. Andersson, Å. Torinssson-Naluai, L. O. Cardell, A network-based analysis of the late-phase reaction of the skin. *J. Allergy Clin. Immunol.* **118**, 220–225 (2006).
31. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
32. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, X. S. Liu, Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
33. M. J. Li, P. Wang, X. Liu, E. L. Lim, Z. Wang, M. Yeager, M. P. Wong, P. C. Sham, S. J. Chanock, J. Wang, GWASdb: A database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **40**, D1047–D1054 (2012).

34. T.-H. Chang, H.-Y. Huang, J. B.-K. Hsu, S.-L. Weng, J.-T. Horng, H.-D. Huang, An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. *BMC Bioinformatics* **14**, S4 (2013).
35. J. Qian, T. Hastie, J. Friedman, R. Tibshirani, N. Simon, Glmnet for Matlab; www.stanford.edu/~hastie/glmnet_matlab/.
36. A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, A. Califano, ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**, S7 (2006).
37. R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga, V. Thorsson, The Inferelator: An algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.* **7**, R36 (2006).
38. I. V. Kapitannikov, A. A. Popov, E. V. Shimbarevich, L. D. Rumsh, V. K. Antonov, [C-terminal amidation of acylamino acids and peptides using a transpeptidation method catalyzed by carboxypeptidase Y]. *Bioorg. Khim.* **14**, 797–801 (1988).
39. M. Gustafsson, M. Hörnquist, A. Lombardi, Constructing and analyzing a large-scale gene-to-gene regulatory network Lasso-constrained inference and biological validation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2**, 254–261 (2005).
40. M. Gustafsson, M. Hörnquist, J. Lundström, J. Björkegren, J. Tegnér, Reverse engineering of gene networks with LASSO and nonlinear basis functions. *Ann. N.Y. Acad. Sci.* **1158**, 265–275 (2009).
41. A. Greenfield, C. Hafemeister, R. Bonneau, Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics* **29**, 1060–1067 (2013).
42. M. E. Studham, A. Tjärnberg, T. E. M. Nordling, S. Nelander, E. L. L. Sonnhammer, Functional association networks as priors for gene regulatory network inference. *Bioinformatics* **30**, i130–i138 (2014).
43. R. Jörnsten, T. Abenius, T. Kling, L. Schmidt, E. Johansson, T. E. M. Nordling, B. Nordlander, C. Sander, P. Gennemark, K. Funa, B. Nilsson, L. Lindahl, S. Nelander, Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol. Syst. Biol.* **7**, 486 (2011).
44. A. H. Beecham, N. A. Patsopoulos, D. K. Xifara, M. F. Davis, A. Kemppinen, C. Cotsapas, T. S. Shah, C. Spencer, D. Booth, A. Goris, A. Oturai, J. Saarela, B. Fontaine, B. Hemmer, C. Martin, F. Zipp, S. D'Alfonso, F. Martinelli-Boneschi, B. Taylor, H. F. Harbo, I. Kockum, J. Hillert, T. Olsson, M. Ban, J. R. Oksenberg, R. Hintzen, L. F. Barcellos; Wellcome Trust Case Control Consortium 2 (WTCCC2), International IBD Genetics Consortium (IIBDGC), C. Agliardi, L. Alfredsson, M. Alizadeh, C. Anderson, R. Andrews, H. B. Søndergaard, A. Baker, G. Band, S. E. Baranzini, N. Barizzone, J. Barrett, C. Bellenguez, L. Bergamaschi, L. Bernardinelli, A. Berthele, V. Biberacher, T. M. C. Binder, H. Blackburn, I. L. Bomfim, P. Brambilla, S. Broadley, B. Brochet, L. Brundin, D. Buck, H. Butzkueven, S. J. Caillier, W. Camu, W. Carpentier, P. Cavalla, E. G. Celius, I. Coman, G. Comi, L. Corrado, L. Cosemans, I. Cournu-Rebeix, B. A. C. Cree, D. Cusi, V. Damotte, G. Defer, S. R. Delgado, P. Deloukas, A. di Sapio, A. T. Dilthey, P. Donnelly, B. Dubois, M. Duddy, S. Edkins, I. 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. Lie, C. M. Lill, M. Lindén, J. Link, F. Luessi, J. Lycke, F. Macciardi, S. Männistö, C. P. 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. Rodegher, D. Rog, M. Salvetti, N. C. Schnetz-Boutaud, F. Sellebjerg, R. C. Selter, C. 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. Tienari, C. van Duijn, E. M. Visser, S. Vucic, H. Westerlind, J. S. Wiley, A. Wilkins, J. F. Wilson, J. Winkelmann, J. Zajicek, E. Zindler, J. L. Haines, M. A. Pericak-Vance, A. J. Ivinson, G. Stewart, D. Hafler, S. L. Hauser, A. Compston, G. McVean, P. De Jager, S. J. Sawcer, J. L. McCauley; International Multiple Sclerosis Genetics Consortium (IMSGC), Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
45. R. H. Roxburgh, S. R. Seaman, T. Masterman, A. E. Hensiek, S. J. Sawcer, S. Vukusic, I. Achiti, C. Confavreux, M. Coustans, E. le Page, G. Edan, G. V. McDonnell, S. A. Hawkins, M. Trojano, M. Liguori, E. Cocco, M. G. Marrosu, F. Tesser, M. Leone, A. Weber, F. Zipp, B. Miterski, J. T. Epplen, A. Oturai, P. S. Sørensen, E. G. Celius, N. T. Lara, X. Montalban, P. Villoslada, A. M. Silva, M. Marta, I. Leite, B. Dubois, J. P. Rubio, H. Butzkueven, T. Kilpatrick, M. P. Mycko, K. W. Selmaj, M. E. Rio, M. J. Sá, G. Salemi, G. Savettieri, J. Hillert, D. A. S. Compston, Multiple sclerosis severity score: Using disability and disease duration to rate disease severity. *Neurology* **64**, 1144–1151 (2005).
46. Y. Chen, D. L. Bates, R. Dey, P.-H. Chen, A. C. D. Machado, I. A. Laird-Offringa, R. Rohs, L. Chen, DNA binding by GATA transcription factor suggests mechanisms of DNA looping and long-range gene regulation. *Cell Rep.* **2**, 1197–1206 (2012).
47. L. Guo, Y. Du, S. Chang, K. Zhang, J. Wang, rSNPBase: A database for curated regulatory SNPs. *Nucleic Acids Res.* **42**, D1033–D1039 (2014).

Acknowledgments: We thank the International Multiple Sclerosis Genetics Consortium for providing GWAS data and J. Nedergaard Larsen (ALK-Abello) for providing allergen extract. **Funding:** This work was supported by the Cancer fund, Swedish Medical Research Council (K2013-61X-22310-01-04 to J.E. and 2012-3168 to M.B.), the Academy of Finland Centre of Excellence in Molecular Systems Immunology and Physiology Research 2012–2017 (Decision No. 250114 to R.L.), the Sigrid Jusélius Foundation (R.L.), the Generalitat de Catalunya AGAUR (2014-SGR364), the Spanish Association Against Cancer (AECC 2010), and the Spanish Ministry of Health ISCIII FIS (PI12/01528) and RTICC (RD12/0036/0008). **Author contributions:** M.G. and D.R.G. conceived and performed bioinformatics analyses; L.A., S.B., S.H., D.E., J.E., I.K., R.L., J.M., L.M., H.R.I.L., T.O., J.B., R.B., A.M., M.M., and M.S. collected and analyzed clinical materials; A.K., R.L., O.R., S.T., and M.V. performed and analyzed ChIP-seq data; A.K., A.L., H.W., and H.Z. performed the experiments; M.A.P. and J.S.-M. contributed to the study design and bioinformatics analyses; and C.E.N. and M.B supervised the study. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The data sets for this study are available at data S1 and S2 and at the GEO data repository under super series accession GSE60680.

Submitted 20 August 2015
Accepted 1 October 2015
Published 11 November 2015
10.1126/scitranslmed.aad2722

Citation: M. Gustafsson, D. R. Gawel, L. Alfredsson, S. Baranzini, J. Björkander, R. Blomgran, S. Hellberg, D. Eklund, J. Ernerudh, I. Kockum, A. Konstantinell, R. Lahesmaa, A. Lentini, H. R. I. Liljenström, L. Mattson, A. Matussek, J. Mellergård, M. Mendez, T. Olsson, M. A. Pujana, O. Rasool, J. Serra-Musach, M. Stenmarker, S. Tripathi, M. Viitala, H. Wang, H. Zhang, C. E. Nestor, M. Benson, A validated gene regulatory network and GWAS identifies early regulators of T cell–associated diseases. *Sci. Transl. Med.* **7**, 313ra178 (2015).

Science Translational Medicine AAAS

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

Mika Gustafsson, Danuta R. Gawel, Lars Alfredsson, Sergio Baranzini, Janne Björkander, Robert Blomgran, Sandra Hellberg, Daniel Eklund, Jan Ernerudh, Ingrid Kockum, Aelita Konstantinell, Riita Lahesmaa, Antonio Lentini, H. Robert I. Liljenström, Lina Mattson, Andreas Matussek, Johan Mellergård, Melissa Mendez, Tomas Olsson, Miguel A. Pujana, Omid Rasool, Jordi Serra-Musach, Margaretha Stenmarker, Subhash Tripathi, Miro Viitala, Hui Wang, Huan Zhang, Colm E. Nestor and Mikael Benson (November 11, 2015)

Science Translational Medicine **7** (313), 313ra178. [doi: 10.1126/scitranslmed.aad2722]

Editor's Summary

Identifying disease before it starts

Diseases may be easier to treat if caught early. However, means of identifying early disease–– especially before symptoms appear––are in short supply. Now, Gustafsson *et al.* identify early regulators of T cell–mediated disease by finding transcription factors involved in T cell differentiation that are enriched in disease-associated polymorphisms. Three such experimentally validated transcription factors––*GATA3*, *MAF*, and *MYB*––and their targets were found to be differentially expressed in asymptomatic stages of two different T cell–mediated diseases––multiple sclerosis and seasonal allergic rhinitis. These data not only provide potential markers of disease development but also shed light on the mechanistic underpinning of T cell –mediated disease.

The following resources related to this article are available online at http://stm.sciencemag.org.
This information is current as of April 16, 2016.

Article Tools	Visit the online version of this article to access the personalization and article tools: http://stm.sciencemag.org/content/7/313/313ra178
Supplemental Materials	*"Supplementary Materials"* http://stm.sciencemag.org/content/suppl/2015/11/09/7.313.313ra178.DC1
Related Content	The editors suggest related resources on *Science*'s sites: http://stm.sciencemag.org/content/scitransmed/5/189/189sr4.full http://stm.sciencemag.org/content/scitransmed/6/218/218ec9.full
Permissions	Obtain information about reproducing this article: http://www.sciencemag.org/about/permissions.dtl