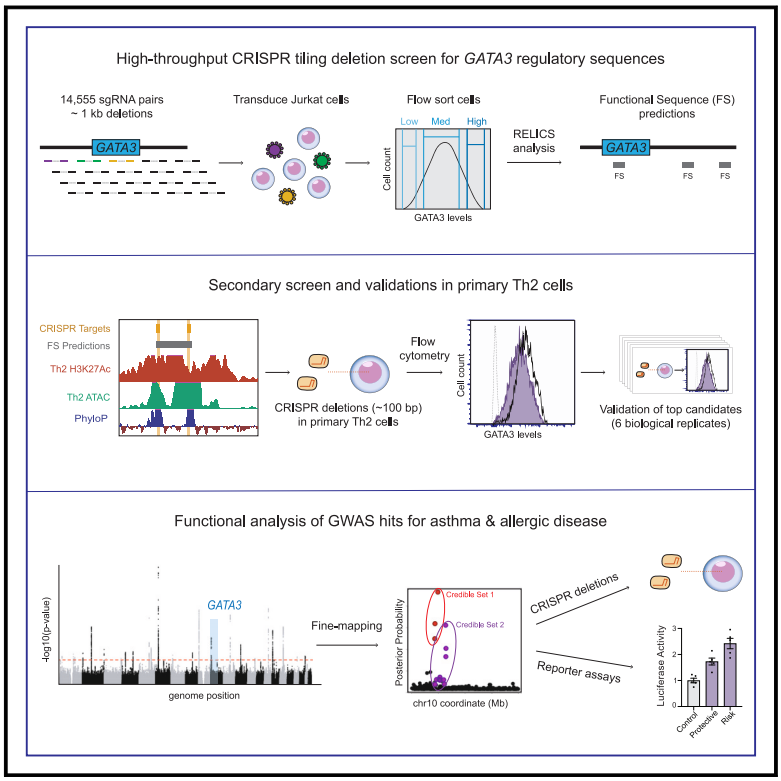
Deletion mapping of regulatory elements for *GATA3* in T cells reveals a distal enhancer involved in allergic diseases

Graphical abstract



Authors

Hsiuyi V. Chen, Michael H. Lorenzini, Shanna N. Lavalle, ..., Karthik Guruvayurappan, Carolyn O'Connor, Graham McVicker

Correspondence

chen_hsiu-yi@idlabs.a-star.edu.sg (H.V.C.), gmcvicker@salk.edu (G.M.)

To map regulatory elements for *GATA3*, we deleted genome sequences in human T cells. We discovered a regulatory sequence 1 Mb downstream of *GATA3* that contains variants associated with allergic disease, demonstrating how deletions can identify regulatory sequences and help interpret hits from genome-wide association studies.

Chen et al., 2023, The American Journal of Human Genetics *110*, 703–714
April 6, 2023

https://doi.org/10.1016/j.ajhg.2023.03.008

Deletion mapping of regulatory elements for *GATA3* in T cells reveals a distal enhancer involved in allergic diseases

Hsiuyi V. Chen,[1,6,10,*] Michael H. Lorenzini,[1,2,10] Shanna N. Lavalle,[1,10] Karthyayani Sajeev,[1,3] Ariana Fonseca,[1] Patrick C. Fiaux,[1,4,7] Arko Sen,[1] Ishika Luthra,[1,8] Aaron J. Ho,[1] Aaron R. Chen,[1,9] Karthik Guruvayurappan,[1,3] Carolyn O'Connor,[5] and Graham McVicker[1,*]

Summary

GATA3 is essential for T cell differentiation and is surrounded by genome-wide association study (GWAS) hits for immune traits. Interpretation of these GWAS hits is challenging because gene expression quantitative trait locus (eQTL) studies lack power to detect variants with small effects on gene expression in specific cell types and the genome region containing *GATA3* contains dozens of potential regulatory sequences. To map regulatory sequences for *GATA3*, we performed a high-throughput tiling deletion screen of a 2 Mb genome region in Jurkat T cells. This revealed 23 candidate regulatory sequences, all but one of which is within the same topological-associating domain (TAD) as *GATA3*. We then performed a lower-throughput deletion screen to precisely map regulatory sequences in primary T helper 2 (Th2) cells. We tested 25 sequences with ~100 bp deletions and validated five of the strongest hits with independent deletion experiments. Additionally, we fine-mapped GWAS hits for allergic diseases in a distal regulatory element, 1 Mb downstream of *GATA3*, and identified 14 candidate causal variants. Small deletions spanning the candidate variant rs725861 decreased GATA3 levels in Th2 cells, and luciferase reporter assays showed regulatory differences between its two alleles, suggesting a causal mechanism for this variant in allergic diseases. Our study demonstrates the power of integrating GWAS signals with deletion mapping and identifies critical regulatory sequences for *GATA3*.

T cells orchestrate adaptive immune responses by differentiating into distinct subsets of effector and regulatory T cells. The GATA3 transcription factor (TF) is central to this process and participates in the differentiation of virtually all T cell subsets. For example, high expression of *GATA3* drives T helper 2 (Th2) cell differentiation,[1] maintains the identity of regulatory T (Treg) cells,[2,3] and disrupts differentiation of T helper 1 and T helper 9 cells.[4,5] Genome-wide association studies (GWASs) have uncovered many genetic variants near to *GATA3* (within 1 Mb) that are significantly associated with immune-related traits, including rheumatoid arthritis,[6,7] multiple sclerosis,[8] type 1 diabetes,[9] asthma, and allergic diseases.[10,11] The risk variants for these traits are non-coding and may affect *GATA3* expression, however their interpretation is challenging because we lack a deep understanding of *GATA3*'s *cis*-regulatory landscape. Because of the high density of GWAS hits near *GATA3* and its cell-type-specific regulation, the *GATA3* locus is an excellent model system for interpreting trait-associated human genetic variation. Here, we integrate functional genomic data with CRISPR-mediated genome deletions to identify regulatory elements for *GATA3* and to illuminate the function of genetic variants associated with allergic diseases.

To confirm that human T cells are an appropriate model for studying *GATA3* regulation, we profiled *GATA3* expression by using a published dataset of sorted immune cells.[12] *GATA3* expression is absent in B cells and monocytes, is moderate in naive $CD4^+$ T cells, and is highest in memory Treg cells and Th2 cells, consistent with its established role in Th2 cell differentiation and maintenance[1,13] (Figures 1A and 1B). Therefore, we decided to focus our study on the regulation of *GATA3* in T cells.

To select an informative genome region for our study, we surveyed trait-associated genetic variation surrounding *GATA3* on chromosome 10. We examined lead single-nucleotide polymorphisms (SNPs) from the GWAS catalog,[17] classifying them into three categories based on immune traits that potentially involve T cell dysfunction: autoimmune and allergic diseases, leukemia and lymphoma, and blood cell traits (Figure 1C). The lead SNPs for these trait categories are distributed across a 2 Mb region surrounding *GATA3*. Lead SNPs for allergic and autoimmune diseases are located predominantly at *GATA3* and downstream with a major cluster 1 Mb away. Most lead SNPs associated with leukemia and lymphoma fall in the gene body of *GATA3*. Finally, lead SNPs for blood

[1]Integrative Biology Laboratory, Salk Institute for Biological Studies, La Jolla, CA, USA; [2]Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA, USA; [3]School of Biological Sciences, University of California San Diego, La Jolla, CA, USA; [4]Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA, USA; [5]Flow Cytometry Core Facility, Salk Institute for Biological Studies, La Jolla, CA, USA
[6]Present address: ASTAR Infectious Diseases Labs, Agency for Science, Technology and Research (ASTAR), Singapore, Singapore
[7]Present address: PetDx, La Jolla, CA, USA
[8]Present address: Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC, Canada
[9]Present address: Renaissance School of Medicine at Stony Brook University, Stony Brook, NY, USA
[10]These authors contributed equally
*Correspondence: chen_hsiu-yi@idlabs.a-star.edu.sg (H.V.C.), gmcvicker@salk.edu (G.M.)
https://doi.org/10.1016/j.ajhg.2023.03.008.

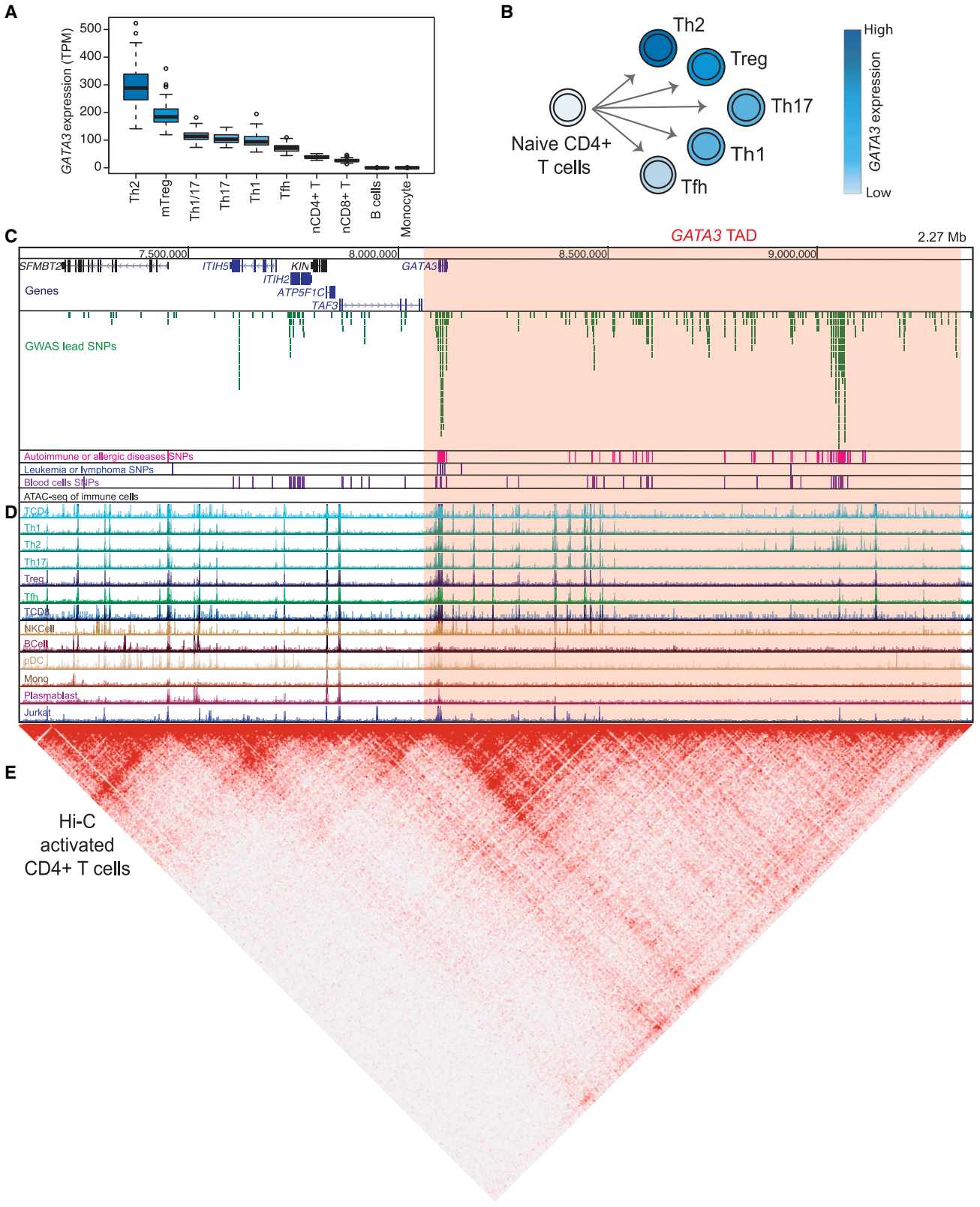


Figure 1. Overview of *GATA3* expression, chromatin accessibility, and nearby trait associations

(A) *GATA3* expression in transcripts-per-million (TPM) in different immune cells from the database of immune cell expression (DICE). Boxplots summarize expression distributions across samples. Center lines, hinges, and whiskers indicate the median, interquartile range, and min/max points within 1.5 times the interquartile range, respectively[12]

(B) Differentiation of naive $CD4^+$ T cells into effector and regulatory T cells; shading indicates *GATA3* expression level.

(C) Lead single-nucleotide polymorphisms (SNPs) from the GWAS catalog in a 2.3 Mb window around *GATA3*. Lead SNPs belonging in three categories (autoimmune or allergic diseases, leukemia or lymphoma, and blood cells) are indicated below.

(D) Chromatin accessibility surrounding *GATA3* in immune cells and Jurkat T cells from published ATAC-seq datasets.[14,15]

(E) Chromatin contact map of the region surrounding *GATA3* from published Hi-C data from $CD4^+$ T cells that were activated for 48 h.[16]

cell traits are scattered across the entire region. The varied locations of lead SNPs for different traits suggest that many different regulatory sequences may be involved in the different trait categories. We therefore considered this 2 Mb region around *GATA3* for our investigation.

Given that active regulatory sequences may physically contact their target genes and are typically in open chromatin, we next determined the locations of topological-associating domains (TADs) and chromatin accessibility at this 2 Mb region. We examined published data from Hi-C performed in activated $CD4^+$ T cells[16] and from the assay for transposase accessible chromatin sequencing (ATAC-seq) performed in sorted immune cell subsets and the Jurkat T cell line.[14,15] These data reveal that *GATA3* is located near the boundary of a large (~1.3 Mb) TAD that extends downstream of the gene (Figures 1C–1E). Throughout this TAD are dozens of accessible chromatin regions that could be considered candidate regulatory elements (Figure 1D), which motivates further functional testing of this region to determine which sequences control *GATA3* expression.

To seek further evidence that this region contains regulatory sequences and variants that affect *GATA3* expression, we asked whether lead SNPs for several immune traits are associated with the expression of *GATA3*, using data from gene expression quantitative trait locus (eQTL) studies.[12,18,19] However, the eQTL data provide only weak evidence that GWAS hits affect *GATA3* expression in T cells (Table S1). Specifically, out of seven GWAS traits and 13 cell types examined, there were only three nominally significant associations with p values between 0.01 and 0.04 (Table S1). One plausible explanation for the lack of eQTL associations is that power to detect eQTLs is limited by the modest effects of common variants on gene expression and by the relatively small sample sizes of eQTL studies in the relevant cell type (T cells).[20]

We reasoned that genomic deletions could determine which sequences regulate *GATA3* expression and could overcome some of the limitations of eQTL studies. Deletions can directly test the effect of sequences on *GATA3* expression (overcoming ambiguity caused by linkage disequilibrium) and are more likely to have large effects on gene expression than common variants used in eQTL studies. We therefore performed a paired-guide tiling deletion CRISPR-Cas9 screen[21,22] to discover regulatory sequences that control *GATA3* expression in T cells. With sufficient deletion efficiency, paired-guide screens can perturb larger genome regions more effectively than single guide screens or base-editor screens, which have previously been used to screen for regulatory elements.[23–26] Deletion mapping is complementary to regulatory mapping with CRISPR interference or CRISPR activation[27–32] and differs in that it tests the effect of deleting a sequence rather than the effect of epigenetic silencing or activation of a sequence.

To screen for regulatory sequences, we designed 14,769 pairs of single guide RNAs (sgRNAs) to tile across a 2 Mb genome region centered on *GATA3*. The guide pairs were designed to target genome sites separated by a median distance of 1,043 bp, with a median step size of 96 bp, such that each base in the screened region would be covered by a median of 8 intended deletions (Figure S1 and Table S4). We note, however, that as a result of variation in guide efficiency and non-homologous end-joining, paired guides generate not only deletions spanning the two target sites but also small insertions/deletions (indels) at each of the target sites.[33,34] To estimate the deletion efficiency of our system, we performed paired-guide deletions of three 1–2 kb sequences in the 2 Mb survey region in Jurkat T cells and measured dropout of the targeted regions by quantitative PCR (qPCR) (Figure S2). We estimated the spanning deletion efficiency to be 20%–25% and account for this in our analysis described below.

We performed our paired guide screen in Jurkat T cells. While Jurkat is a leukemia cell line, it is a useful model system because the chromatin landscape surrounding *GATA3* resembles that of primary T cells (Figures 1D and S3) and it has been previously used to discover disease-relevant T cell regulatory elements.[30] We synthesized an oligo pool encoding sgRNA guide pairs, cloned these into an spCas9 lentiviral vector, and then generated and transduced this lentiviral library into Jurkat T cells (Figure 2A). We performed antibiotic selection to enrich for cells with viral genome integration and flow sorted them into pools based on GATA3 levels by fluorescence-activated cell sorting (Figure S4). Finally, we performed deep sequencing of the sgRNA pairs in each pool (Figure 2A and Table S6). We conducted four biological replicates of the screen, sorting the first two replicates into three pools and the second two replicates into seven pools (Figure 2B).

To determine the effect of guide RNA pairs on GATA3 levels, we estimated the proportion of sgRNA pair counts in each pool and compared the frequency of sgRNA pairs targeting *GATA3* exons to that of non-targeting control sgRNAs (NTCs) included in the screening library. As expected, sgRNA pairs targeting *GATA3* exons are depleted in the high pools and enriched in the low pools (Figure 2C). In contrast, sgRNAs targeting sequences outside of the gene are only slightly enriched in the low pool compared to NTCs, suggesting that a relatively small fraction of the sgRNAs targets regulatory sequences for the gene (Figure S5). These results indicate that the replicates are consistent and that the sgRNA counts from our screen can be used to detect sequences that affect GATA3 levels.

To discover sequences that affect GATA3 levels, we jointly analyzed the screen data across all of the pools and replicates by using RELICS.[37] RELICS is designed to discover functional sequences (FSs) from tiling CRISPR screens and includes features for modeling programmed deletions. RELICS can also leverage data from multiple pools to detect FSs with smaller effect sizes. RELICS has been extensively validated on experimental data and outperforms other tiling CRISPR screen analysis methods.[37] When running RELICS, we used an area of effect model

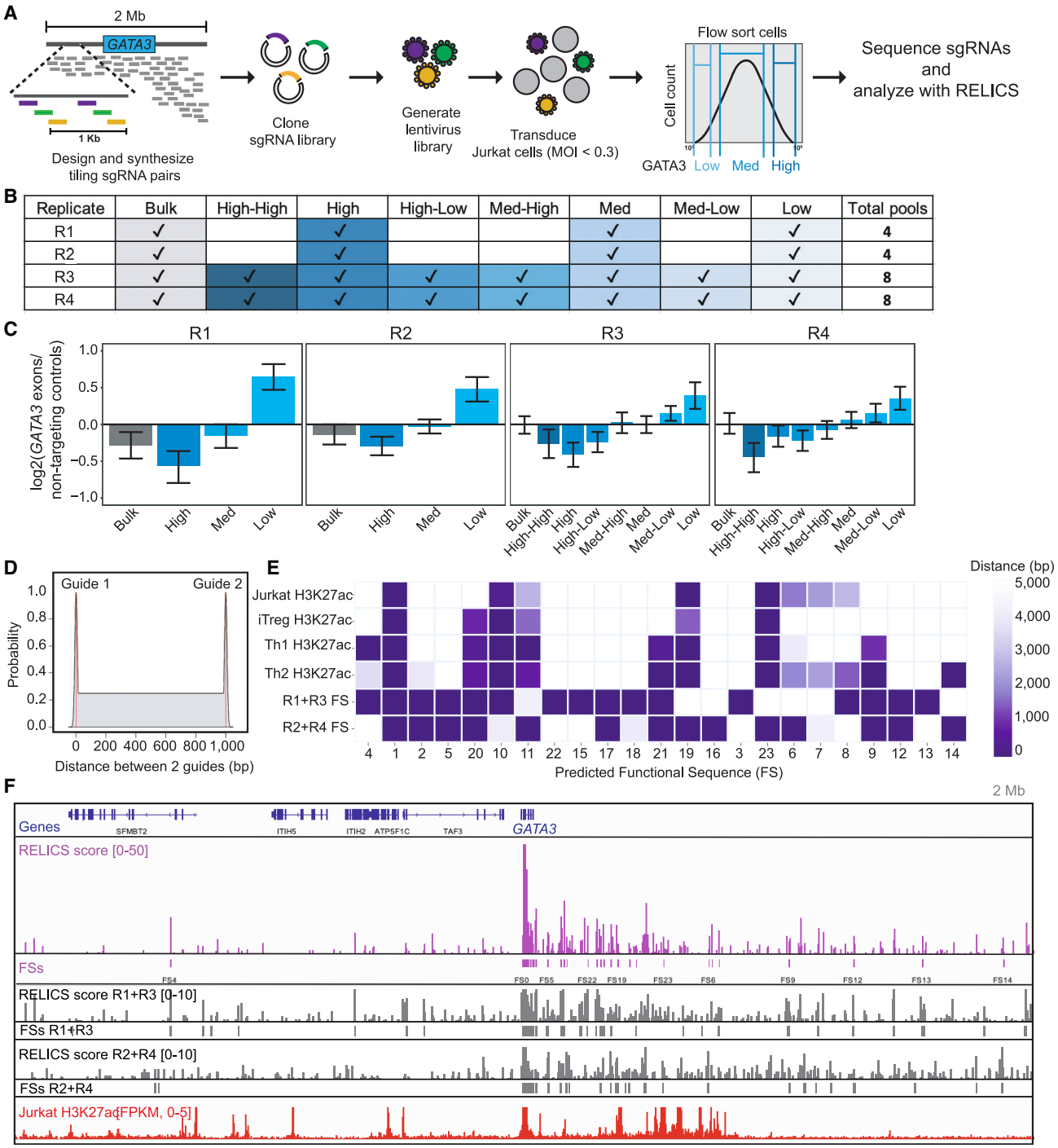


Figure 2. Tiling deletion screen for *GATA3* regulatory elements

(A) Schematic of CRISPR-Cas9 tiling deletion screen performed in Jurkat T cells.

(B) Four biological replicates of the screen were performed, and cells were sorted into three or seven pools based on GATA3 levels.

(C) Guide pairs targeting *GATA3* exons are depleted in low pools and enriched in high pools compared to non-targeting control guides. Plots for each replicate show the log2 ratio of the estimated guide pair proportions in each pool. Proportions were estimated by maximum likelihood by RELICS. Whiskers are 95% confidence intervals estimated from 1,000 bootstrap iterations.

(D) Area of effect model used by the RELICS analysis, which allows for indels at each guide RNA target site and lower-frequency spanning deletions between target sites.

(E) Functional sequences (FSs) predicted by RELICS and colored according to distance to the closest H3K27ac ChIP-seq peak for Jurkat cells, activated induced T regulatory cells (iTregs), activated Th2 cells, and activated Th1 cells from published studies.[35,36] The R1+R3 and R2+R4 rows show distances to FSs predicted from running RELICS on half of the four replicates.

(F) Genome tracks across the 2 Mb screened region showing protein-coding genes, scores, and FSs obtained by running RELICS on all replicates or half of the replicates (R1+R3 or R2+R4) from the tiling deletion screen and H3K27ac ChIP-seq data from Jurkat cells in fragments per kilobase per million mapped reads (FPKM). RELICS scores are log likelihood ratios computed for individual 100 bp windows. RELICS allows FSs to span multiple 100 bp windows and, for this reason, there is not a perfect correspondence between the significant FSs and the highest-scoring 100 bp windows.

that assumes a spanning deletion efficiency of 25% to account for the variable mutation events generated by nonhomologous end-joining (Figure 2D).

In total, RELICS predicted 23 FSs that affect GATA3 levels in Jurkat cells, under a log likelihood ratio threshold of 6 ($p = 5e{-}4$ by likelihood ratio test) (Figures 2E, 2F, and S6). These FSs are distributed asymmetrically; all but one are located within the TAD that contains *GATA3*. Most (18/23 = 78%) of the FSs are located within 0.5 Mb of *GATA3*, and 14/23 (61%) of them overlap with or are near to (within 5 kb) peaks of histone H3 lysine 27 acetylation (H3K27ac) in Jurkat cells or primary T cells (Figure 2E). To quantify the enrichment of FS overlaps with H3K27ac peaks within the *GATA3* TAD, we performed permutations in which we shifted the genome locations of FSs 100,000 times and observed significant enrichments within Jurkat cells, induced T regulatory cells (iTregs), and activated Th1 cells ($p = 0.046$, $p = 0.034$, and $p = 0.0076$ by permutation test) (Figure S7). To evaluate concordance between replicates, we also ran RELICS separately on two groups of replicates: R1+R3 and R2+R4 (Figures 2E and 2F). The overlap between FSs predicted from each of these groups was highly significant ($p = 0.0003$ by genome perturbations). In summary, our deletion screen revealed 23 distinct FSs that may contain regulatory elements for *GATA3* in Jurkat cells.

While the high-throughput deletion screen described above illuminates the regulatory landscape of *GATA3*, a limitation is that it was performed in the Jurkat cell line, which is less physiologically relevant to human traits than primary T cells. In addition, the low efficiency of deletions and the possibility of larger deletion events[38,39] reduce the resolution and accuracy of the FSs identified from the screen (the mean size of our predicted FSs is 1,267 bp). To address these limitations and to identify more precise genomic regions that affect GATA3 levels in primary cells, we performed a secondary lower-throughput screen by introducing ~100 bp CRISPR deletions into primary Th2-polarized human T cells. We selected Th2 cells for this secondary screen because *GATA3* expression is highest in Th2 cells (Figure 1A).

As candidate sequences for the secondary screen, we selected 25 sequences that were marked with varying levels of H3K27ac, accessible chromatin, and evolutionary sequence conservation and that were near to or overlapping FSs from the high-throughput screen (Figures 3A, 3B, and S9). We named each sequence targeted for deletion in the secondary screen with the name of the nearest FS and a unique number. For example, deletions near FS10 were named FS10-1, FS10-2, FS10-3, and FS10-4. As a negative control, we targeted a "safe harbor" (SH) sequence downstream of *GATA3* with no predicted regulatory function (Figures S9A and S9B), and as a positive control, we targeted a *GATA3* exon with a single sgRNA. We designed sgRNA pairs to introduce small deletions of these sequences and transfected them as Cas9 ribonucleoprotein complexes into Th2 cells[40,41] (Figure 3C). We verified the products generated by each deletion experiment by using Sanger sequencing and tracking of indels by decomposition (TIDE) (Figure S9C and Table S8).[42]

For 9/25 of the deleted sequences in the secondary screen, GATA3 levels were reduced by at least 20% on both days compared to the SH deletion (Figure 3E). From these, five deletions were selected for further validation (FS1-1, FS6-5, FS10-3, FS19-3, and FS23-5). We performed six replicate experiments for each of these candidate sequences, where each replicate consisted of an independent Th2 polarization and CRISPR deletion. Deletion efficiencies were estimated to be 50%–100% by TIDE (Figure S9D and Table S8). Deletions of all five regions significantly decreased GATA3 levels at 3 days after transfection as did deletions of three of the regions (FS1-1, FS19-3, and FS23-5) at 4 days after transfection (Figure 3F).

We examined FS1-1 in more detail because deletion of this sequence reduced GATA3 levels by 28% (log2 MFI ratio = −0.478) at day 3 after transfection in Th2 cells (Figures 3D–3F) and this sequence is directly within FS1, the top-ranked prediction from the high-throughput Jurkat screen. FS1-1 is located within the third intron of *GATA3* and overlaps strong peaks of H3K27ac and chromatin accessibility (Figure 3B). Given the overlap with H3K27ac and chromatin accessibility, FS1 most likely acts as an enhancer, although because of its location within an intron, it could also be a splicing regulatory element.

We next examined FS23 because this sequence is orthologous to a mouse genome sequence that was previously identified as a strong enhancer for *Gata3* in mouse T cells.[43] In addition, this sequence overlaps with ATAC-seq peaks in all T cell subsets and has strong H3K27ac signals in Jurkat cells, naive $CD4^+$ T cells, and Th2 cells (Figure 3B). In our secondary Th2 cell screen, deletions of two conserved sequences that directly overlap FS23 (FS23-4 and FS23-5; Figure 3B,E) both reduced GATA3 levels, whereas deletions of the FS23-1, FS23-2, and FS23-3 sequences, which are located ~30 kb downstream and which do not overlap a predicted FS, had minimal effects on GATA3 levels (Figure 3E). Validation experiments targeting FS23-5 confirmed that deletion of this sequence reduces GATA3 levels by 20% (log2 MFI ratio = −0.314) on day 3 and 9.7% (log2 MFI ratio = −0.148) on day 4 after transfection (Figure 3F).

As a further validation, and to confirm that *GATA3* mRNA abundance is the mechanism of diminished GATA3 levels, we quantified *GATA3* mRNA in FS1-1 and FS19-3 deletion replicates by reverse transcription qPCR (RT-qPCR). These deletions decreased *GATA3* mRNA by 35%–44% (Figure 3G and Table S10). To verify that *GATA3* is the primary target gene of these regulatory sequences, we quantified mRNA of the next nearest gene, *TAF3*, in the same total RNA samples. *TAF3* mRNA expression did not significantly change, indicating that these regulatory sequences are likely to be *GATA3* specific, although it is possible that these (or some of the other regulatory

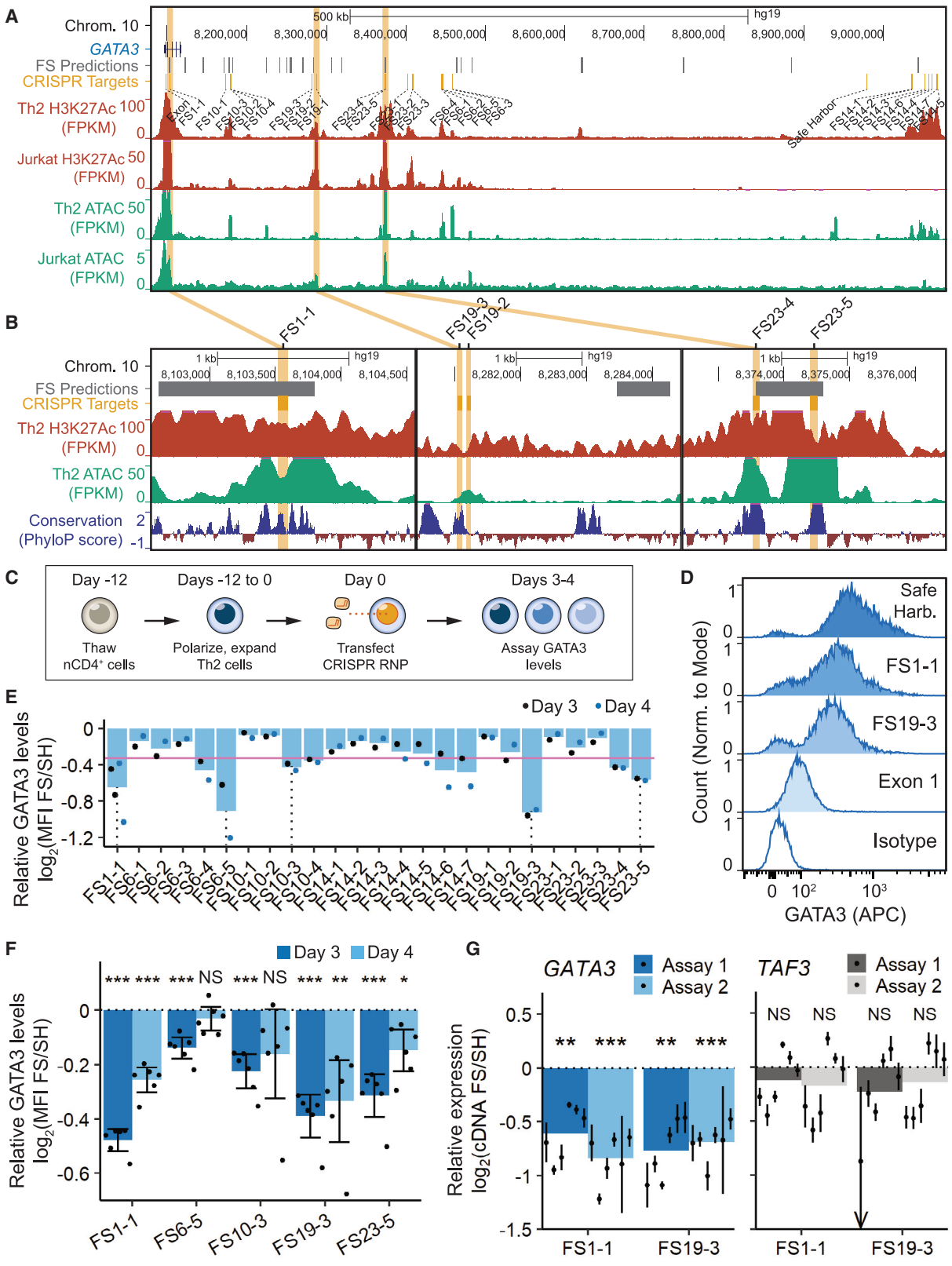


Figure 3. CRISPR deletions identify *GATA3* regulatory sequences in primary Th2 cells

(A) Genome tracks spanning the region downstream of *GATA3*, where we performed targeted deletions to screen for regulatory elements in primary Th2 cells. Tracks are genes, functional sequences (FSs) predicted by RELICS from the high-throughput Jurkat screen, regions targeted for deletion in the secondary Th2 cell screen, and H3K27ac ChIP-seq and ATAC-seq for activated Th2 cells and Jurkat cells from published studies.[14,15,35,36]

(B) Zoom in of genome tracks for deleted sequences FS1-1, FS19-2, FS19-3, FS23-4, and FS23-5 (see Figure S9 for SH, FS6-3, FS6-5, FS10-2, FS10-3, and FS10-4 sequences) with additional track showing 100 vertebrate PhyloP conservation scores.

(legend continued on next page)

sequences) affect other genes (Figure 3G). In summary, we identified and validated five *GATA3* regulatory sequences that are active in Th2 cells, one of which is within a *GATA3* intron and four of which are within the gene desert region downstream of the gene. A limitation of our screening approach is that we prioritized sequences that were near to FSs predicted from the Jurkat screen and we therefore may have missed some Th2 cell regulatory elements that are inactive in Jurkat cells.

Deletion mapping can potentially help determine the molecular function of GWAS hits. To test this idea, we intersected the FSs identified by our high-throughput screen with GWAS hits and observed that FS1 and FS14 are located within two distinct clusters of risk variants associated with autoimmune and allergic diseases (Figure S8). We examined the region surrounding FS14, which is almost 1 Mb downstream of *GATA3* and which has a high density of lead SNPs (Figure S8). This region is contained within a broad 44 kb H3K27ac domain, which is present in Th2 cells but not in other T cell subsets, suggesting that it may be a distal Th2 enhancer (Figure S8). Because Th2 cells have an established role in allergic diseases[44] and high *GATA3* expression is required for differentiation and maintenance of Th2 cells,[1] we decided to (1) analyze recent GWASs for asthma and allergic diseases (allergic rhinitis or eczema)[10] and (2) test the function of sequences containing risk variants in this distal region.

A single GWAS hit for asthma is situated ~1 Mb downstream of *GATA3*, and a pair of independent GWAS hits for allergic diseases are located at ~400 kb and ~1 Mb downstream of *GATA3* (Figures 4A–4D). We named these hits risk region 1 and risk region 2 and focused on risk region 1 because it is associated with both traits and overlaps the large Th2 enhancer-like sequence described above. We performed fine-mapping to identify 95% credible sets (CSs) containing candidate causal variants by using SuSiE.[45] For asthma, we identified two CSs: CS1, which contains three candidate causal SNPs, and CS2, which contains 11 candidate SNPs (Figures 4E and 4F). For allergic diseases, we identified a single CS containing the same three candidate SNPs as CS1 (Figures 4G, 4H, and 5A). This yielded a total of 14 candidate SNPs for further study (Table S14).

We examined the worldwide allele frequencies of the SNPs with the highest posterior inclusion probabilities (PIPs) in CS1 and CS2. For the lead SNP in CS1, rs12413578 (g.9007290C>T [GenBank: NC_000010.11]), the risk allele is the major allele and the protective allele has the highest allele frequencies in populations with European ancestry and the Indian subcontinent (Figure 5B). In contrast, for the lead SNP in CS2, rs725861 (g.9021813A>G [GenBank: NC_000010.11]), the risk allele is the minor allele and has the highest frequencies in Western Africa and the Indian subcontinent (Figure 5B). Neither of the lead SNPs directly overlap with the FS14 element identified in the Jurkat screen.

To test the function of candidate SNPs in the two CSs, we designed pairs of guides to delete ~100 bp sequences surrounding six of the 14 candidates (we were unable to design uniquely targeting guides for the remaining candidates because of the presence of repeat sequences). Out of the six SNPs, only guide pairs targeting two of the SNPs yielded high-efficiency deletions. These two SNPs are the lead SNPs in the two CSs (i.e., those with the highest PIPs in each CS) (Figure 5C). We deleted 115 bp and 83 bp sequences surrounding rs12413578, the highest PIP SNP in CS1, in *in vitro* differentiated Th2 cells. Neither of these deletions affected GATA3 levels in Th2 cells (Figures 5D and 5E), suggesting that the sequence containing rs12413578 may not regulate *GATA3* expression in primary Th2 cells *in vitro*. In contrast, deletion of a 37 bp sequence surrounding rs725861, the highest PIP SNP in CS2, decreased GATA3 levels in Th2 cells from two different donors (Figures 5D and 5E).

To test the regulatory activity of the sequences containing candidate SNPs, we transfected reporter plasmids containing both alleles of six candidate SNPs into HEK293 cells and assessed their ability to drive luciferase activity compared to an SH control sequence (Figure 5F). We performed these assays in HEK293 cells because we could obtain efficient and reproducible transfections of the reporter plasmids in this cell type. Sequences containing three of the candidate SNPs showed lower luciferase activity in HEK293 cells compared to the SH control: rs12413578, which is the lead SNP in CS1; rs144536148, which overlaps with FS14 (g.9001864A>G [GenBank: NC_000010.11]); and rs1444788 (g.9022157T>C [GenBank: NC_000010.11]) (Figure 5F). This may indicate either that these candidate sequences reduce transcription below basal levels or that the SH control sequence itself drives a modest level of transcription when the influence of chromatin structure is removed. The sequences containing the other three SNPs all had at least one allele with higher luciferase activity compared to the SH control.

(C) Experimental workflow for CRISPR deletions in primary Th2 cells.
(D) Representative GATA3 fluorescence intensity from intranuclear immunostaining and flow cytometry from the Th2 cell CRISPR screen, 3 days after transfection. (See Figures S10 and S11 for all histograms).
(E) GATA3 median fluorescence intensity (MFI) in the Th2 cell CRISPR screen, normalized to the safe harbor (SH) sample MFI. The pink horizontal line indicates a 20% reduction in GATA3 levels compared to the SH control.
(F) GATA3 MFI in CRISPR-edited cells from six replicate Th2 polarizations, normalized to MFI of the corresponding SH replicate.
(G) Gene expression of *GATA3* and *TAF3* in replicate FS1-1-deleted or FS19-3-deleted cells (six biological replicates), each quantified by two RT-qPCR assays, and normalized to expression of SH samples and reference genes *GUSB* and *PPIA*.
(E–G) Bars indicate the mean of each sample group. Whiskers are 95% confidence intervals estimated as 2 × SEM. In (G), confidence intervals are computed from three technical replicates for each biological replicate. To compute p values, we performed two-sided paired t tests using the log2 transformed observations and matched SH controls; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

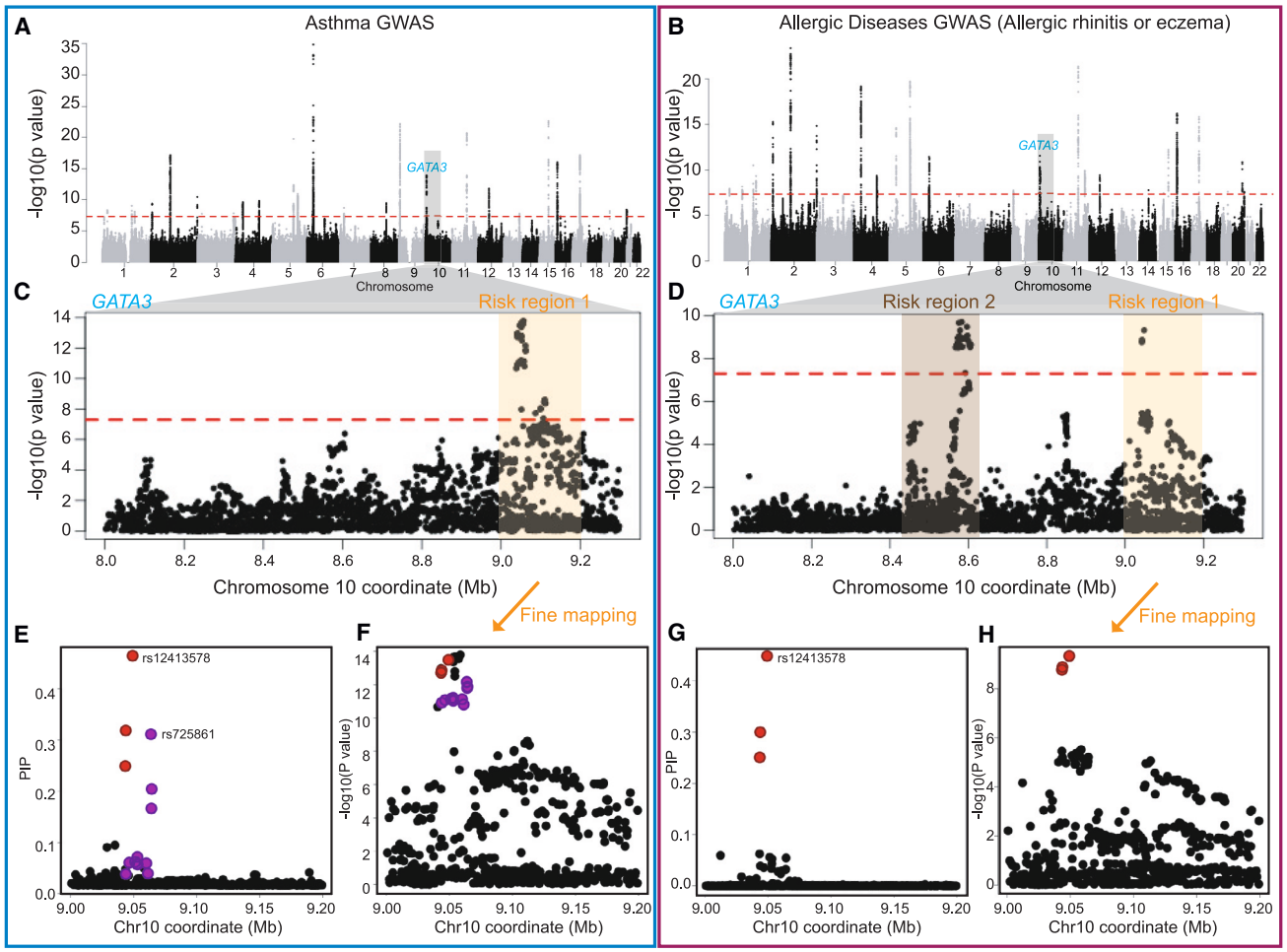


Figure 4. Fine-mapping of genome-wide association study hits at the *GATA3* locus

(A) Manhattan plot for a published GWAS for asthma.[10]

(B) Manhattan plot for a published GWAS of allergic diseases.[10]

(C) Zoom-in of the *GATA3* locus showing a single region of association for asthma, "risk region 1," located ~1 Mb downstream of the gene.

(D) Zoom-in for the *GATA3* locus showing two regions of association for allergic disease downstream of *GATA3*: "risk region 1" and "risk region 2."

(E) Asthma posterior inclusion probabilities (PIPs) for variants in risk region 1 obtained by fine-mapping with SuSIE.[45] The two credible sets (CSs) are indicated with red and purple.

(F) −Log10 p values for the same variants shown in (E).

(G) Allergic disease posterior inclusion probabilities (PIP) for variants in risk region 1 obtained by fine-mapping. The credible set, which is highlighted in red contains the same variants as CS1 for asthma.

(H) −Log10 p values for the same variants shown in (G).

Most notably, sequences spanning the lead SNP for CS2, rs725861, increased luciferase activity, and the risk allele, G, caused significantly stronger induction (2.4-fold) than the protective allele (1.7-fold; $p < 0.05$ by one-way ANOVA) (Figure 5F). Because the reporter assays were carried out in HEK293 cells, a limitation is that the regulatory activity of the tested sequences may not be the same in primary T cells.

In combination, the deletion (Figure 5D) and luciferase experiments (Figure 5F) suggest that the distal sequence containing rs725861 has enhancer activity and that risk variants for allergic diseases and asthma located within this sequence affect *GATA3* expression. We did not test every candidate causal SNP with deletion or luciferase assays, and it is possible that other candidate SNPs also affect gene regulation. Furthermore, given the high number of functional sequences and the two credible sets, it is possible that there are multiple causal variants or haplotypes that affect allergic disease and asthma risk. The sequence containing rs725861 and other candidate SNPs has high levels of H3K27ac that is specific to Th2 cells in which our deletion experiments were performed (Figure S8), and it may have a key regulatory role specific to this cell type. However, our data do not rule out that this sequence may have regulatory activity in other cell types besides Th2 cells. Consistent with this, FS14 was identified in Jurkat cells, and our reporter assays showed regulatory activity in HEK293 cells.

A

Credible Set	GWAS	Candidate causal SNPs
1	Asthma, allergic diseases	rs12413578, rs186856025, rs144536148
2	Asthma	rs11256017, rs2589561, rs2440781, rs1775550, rs1775551, rs2797288, rs957349, rs2589559, rs725861, rs1444788, rs1444789

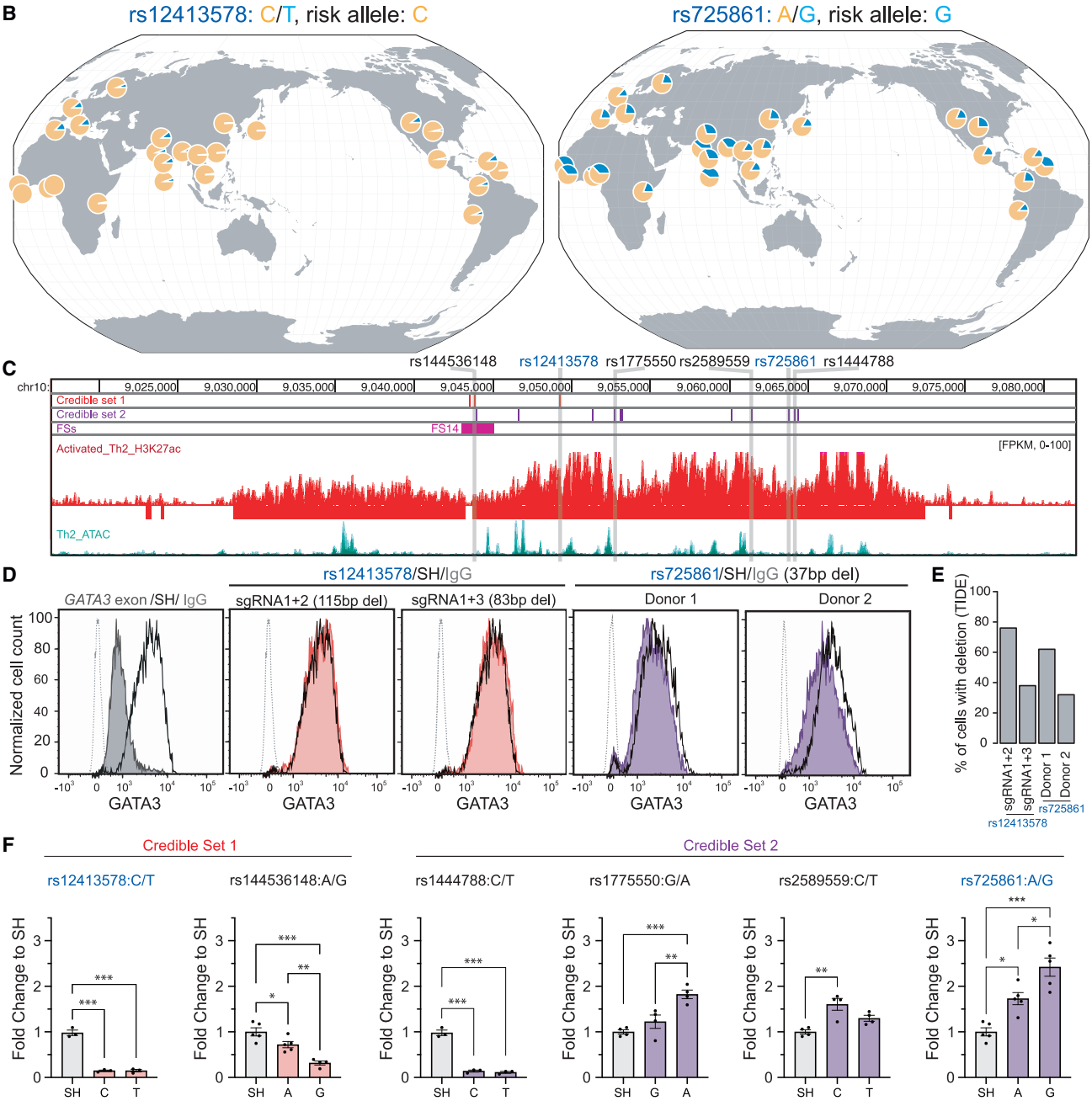


Figure 5. Candidate variants for allergic diseases and asthma in a distal Th2 enhancer

(A) Candidate causal SNPs for asthma and allergic diseases that make up credible sets (CSs) CS1 and CS2.

(B) Worldwide population allele frequencies for the SNPs with the highest PIP in each CS. Allele frequency estimates are from the 1000 Genomes Project[46] and plotted with the Geography of Genetic Variants Browser.[47]

(C) Genome tracks showing the locations of the candidate variants, predicted FSs, H3K27ac ChIP-seq for activated Th2 cells, and ATAC-seq for Th2 cells. SNPs tested with CRISPR deletion or luciferase assays are indicated with gray highlights. Identifiers for the SNPs with the highest PIPs in CS1 and CS2 (rs12413578 and rs725861) are indicated with blue text.

(D) Flow cytometry results following electroporation of Cas9 ribonucleoproteins (RNPs) targeting an exon of *GATA3* or small deletions spanning candidate SNPs rs12413578 or rs725861 in *in vitro* differentiated Th2 cells. Two pairs of gRNAs that differ in one gRNA were used to delete 115 bp or 83 bp spanning rs12413578. A 37 bp sequence spanning rs725861 was deleted in Th2 cells from two different donors.

(E) Estimated percentage of cells carrying targeted deletions for samples in (D) (quantified by Tracking of Indels by Decomposition[42]).

(F) Relative luciferase activity of sequences containing candidate SNPs from CS1 and CS2 following transfection in HEK293 cells. Data are represented as fold change to a 300 bp safe harbor (SH) control. Bars represent mean ± SE, N = 3–5. Data were analyzed by one-way ANOVA. Significance is indicated by *p < 0.05, **p < 0.01, ***p < 0.001.

In response to pathogens, naive $CD4^+$ T cells become activated and differentiate into effector and regulatory T cell types to mount appropriate immune responses. However, dysregulated immune responses lead to autoimmune and allergic diseases. Our results show that expression of *GATA3*, a key regulator of T cell differentiation, is coordinated by a pool of downstream regulatory sequences. Furthermore, our results demonstrate the power of coarse regulatory sequence mapping with a high-throughput deletion screen followed by more precise mapping with small deletions in primary cells. Sequences detected by this approach can be remarkably useful for the interpretation of trait-associated genetic variation, and we provide evidence that risk variants for allergic diseases affect the function of a distal *GATA3* enhancer.

Data and code availability

The published article includes all datasets generated or analyzed during this study. The raw sequencing data for the tiling deletion screen have been deposited in GEO under accession GEO: GSE190860.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.03.008.

Acknowledgments

We thank B. Ren and Y. Diao for assistance with the planning of our tiling deletion screen, Y. Zheng and Z. Liu for helpful discussions, P. Hsu and S. Konermann for advice about CRISPR screens, MaxCyte for their assistance with electroporation, and N. Hah and the Salk Next Generation Sequencing Core for their technical support with sequencing. This study was supported by NIH/NHGRI grant HG011315 to G.M.; NIH/NIDDK grant DK122607; NIH/NIAID grant AI107027; the National Cancer Institute-funded Salk Institute Cancer Center (NIH/NCI CCSG: 2 P30 014195); the 2020 Salk Women & Science Award to H.V.C.; the 2020 Salk Alumni Fellowship Award to H.V.C.; the H.A. and Mary K. Chapman Charitable Trust Fellowship to P.C.F.; the Jesse and Caryl Philips Foundation Fellowship to P.C.F.; the Pioneer Fund Postdoctoral Scholar Award to A.S.; and the Frederick B. Rentschler Developmental Chair to G.M. Sequencing and flow cytometry were carried out by the Next Generation Sequencing Core (RRID: SCR_014846) and Flow Cytometry Core (RRID: SCR_014839) facilities of the Salk Institute with funding from NIH/NCI CCSG (P30 014195), the Chapman Foundation, the Helmsley Charitable Trust, and NIH/OD Shared Instrumentation Grant S10-OD023689 (Aria Fusion cell sorter).

Author contributions

H.V.C. and G.M. conceived of the project. H.V.C. performed the tiling deletion screen in Jurkat cells, analyzed GWAS data, performed CRISPR deletion experiments, performed flow cytometry experiments, made initial versions of figures and tables, and drafted the initial version of the manuscript. M.H.L. performed the secondary Th2 cell screen in primary cells, performed CRISPR validation experiments, performed flow cytometry and qPCR experiments, made related figures and tables, and drafted related sections of the paper. S.N.L. performed luciferase reporter assays, made related figures and tables, and drafted related sections of the paper. K.S. performed CRISPR, flow cytometry, PCR, and TIDE experiments under the guidance of M.H.L. A.F. performed luciferase reporter experiments under the guidance of S.N.L. P.C.F. developed RELICS and utilized it to analyze the screening. A.S. performed analysis of ATAC-seq data and sequence data and provided comments on the manuscript. I.L. wrote scripts to design the guide RNAs used in the screen. A.J.H. performed computational analyses of functional sequences and bioinformatics processing of ChIP-seq and ATAC-seq data. A.R.C. assisted with experiments under the direction of H.V.C. K.G. applied and updated RELICS to analyze the screening data and generated some of the screening figures under the guidance of P.C.F. and H.V.C. C.C. assisted with flow cytometry experiments. G.M. supervised the project, acquired funding for the project, and edited and wrote the manuscript with H.V.C., with input from M.H.L. and S.N.L.

Declaration of interests

Electroporation experiments were performed with an ExPERT MaxCyte ATx instrument that was provided to the McVicker laboratory by MaxCyte, Inc. for technology development and evaluation purposes.

Received: June 10, 2022
Accepted: March 8, 2023
Published: March 28, 2023

References

1. Zheng, W., and Flavell, R.A. (1997). The transcription factor GATA-3 is necessary and sufficient for Th2 cytokine gene expression in CD4 T cells. Cell *89*, 587–596.
2. Wang, Y., Su, M.A., and Wan, Y.Y. (2011). An essential role of the transcription factor GATA-3 for the function of regulatory T cells. Immunity *35*, 337–348.
3. Wohlfert, E.A., Grainger, J.R., Bouladoux, N., Konkel, J.E., Oldenhove, G., Ribeiro, C.H., Hall, J.A., Yagi, R., Naik, S., Bhairavabhotla, R., et al. (2011). GATA3 controls $Foxp3^+$ regulatory T cell fate during inflammation in mice. J. Clin. Invest. *121*, 4503–4515.
4. Lee, H.J., Takemoto, N., Kurata, H., Kamogawa, Y., Miyatake, S., O'Garra, A., and Arai, N. (2000). GATA-3 induces T helper cell type 2 (Th2) cytokine expression and chromatin remodeling in committed Th1 cells. J. Exp. Med. *192*, 105–115.
5. Goswami, R., Jabeen, R., Yagi, R., Pham, D., Zhu, J., Goenka, S., and Kaplan, M.H. (2012). STAT6-dependent regulation of Th9 development. J. Immunol. *188*, 968–975.
6. Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P., Zhernakova, A., Stahl, E., Viatte, S., McAllister, K., et al. (2012). High density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat. Genet. *44*, 1336–1340.
7. Ha, E., Bae, S.-C., and Kim, K. (2021). Large-scale meta-analysis across East Asian and European populations updated genetic architecture and variant-driven biology of rheumatoid arthritis, identifying 11 novel susceptibility loci. Ann. Rheum. Dis. *80*, 558–565.
8. International Multiple Sclerosis Genetics Consortium (2019). Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science *365*, eaav7188.

9. Chiou, J., Geusz, R.J., Okino, M.-L., Han, J.Y., Miller, M., Melton, R., Beebe, E., Benaglio, P., Huang, S., Korgaonkar, K., et al. (2021). Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. Nature *594*, 398–402.
10. Zhu, Z., Lee, P.H., Chaffin, M.D., Chung, W., Loh, P.-R., Lu, Q., Christiani, D.C., and Liang, L. (2018). A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. Nat. Genet. *50*, 857–864.
11. Johansson, Å., Rask-Andersen, M., Karlsson, T., and Ek, W.E. (2019). Genome-wide association analysis of 350 000 Caucasians from the UK Biobank identifies novel loci for asthma, hay fever and eczema. Hum. Mol. Genet. *28*, 4022–4041.
12. Schmiedel, B.J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A.G., White, B.M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J.A., McVicker, G., et al. (2018). Impact of genetic polymorphisms on human immune cell gene expression. Cell *175*, 1701–1715.e16.
13. Zhu, J., Min, B., Hu-Li, J., Watson, C.J., Grinberg, A., Wang, Q., Killeen, N., Urban, J.F., Guo, L., and Paul, W.E. (2004). Conditional deletion of Gata3 shows its essential function in TH1-TH2 responses. Nat. Immunol. *5*, 1157–1165.
14. Calderon, D., Nguyen, M.L.T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J.V., et al. (2019). Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat. Genet. *51*, 1494–1505.
15. Massarat, A.R., Sen, A., Jaureguy, J., Tyndale, S.T., Fu, Y., Erikson, G., and McVicker, G. (2021). Discovering single nucleotide variants and indels from bulk and single-cell ATAC-seq. Nucleic Acids Res. *49*, 7986–7994.
16. Gate, R.E., Cheng, C.S., Aiden, A.P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M.G., Subramaniam, M., Shamim, M., et al. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat. Genet. *50*, 1140–1150.
17. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901.
18. Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell *167*, 1398–1414.e24.
19. Kerimov, N., Hayhurst, J.D., Peikova, K., Manning, J.R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M.P., Kuzmin, I., Trevanion, S.J., et al. (2021). A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat. Genet. *53*, 1290–1299.
20. Mostafavi, H., Spence, J.P., Naqvi, S., and Pritchard, J.K. (2022). Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. Preprint at bioRxiv. https://doi.org/10.1101/2022.05.07.491045.
21. Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K.C., Huang, H., Liu, T., Marina, R.J., et al. (2017). A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat. Methods *14*, 629–635.
22. Gasperini, M., Findlay, G.M., McKenna, A., Milbank, J.H., Lee, C., Zhang, M.D., Cusanovich, D.A., and Shendure, J. (2017). CRISPR/Cas9-mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. Am. J. Hum. Genet. *101*, 192–205.
23. Canver, M.C., Smith, E.C., Sher, F., Pinello, L., Sanjana, N.E., Shalem, O., Chen, D.D., Schupp, P.G., Vinjamur, D.S., Garcia, S.P., et al. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature *527*, 192–197.
24. Canver, M.C., Lessard, S., Pinello, L., Wu, Y., Ilboudo, Y., Stern, E.N., Needleman, A.J., Galactéros, F., Brugnara, C., Kutlar, A., et al. (2017). Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. Nat. Genet. *49*, 625–634.
25. Korkmaz, G., Lopes, R., Ugalde, A.P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R., and Agami, R. (2016). Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat. Biotechnol. *34*, 192–198.
26. Cheng, L., Li, Y., Qi, Q., Xu, P., Feng, R., Palmer, L., Chen, J., Wu, R., Yee, T., Zhang, J., et al. (2021). Single-nucleotide-level mapping of DNA regulatory elements that control fetal hemoglobin expression. Nat. Genet. *53*, 869–880.
27. Klann, T.S., Black, J.B., Chellappan, M., Safi, A., Song, L., Hilton, I.B., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2017). CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. Nat. Biotechnol. *35*, 561–568.
28. Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M.D., Banerjee, B., Syed, T., Emons, B.J.M., Gifford, D.K., and Sherwood, R.I. (2016). High-throughput mapping of regulatory DNA. Nat. Biotechnol. *34*, 167–174.
29. Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., et al. (2019). Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat. Genet. *51*, 1664–1669.
30. Simeonov, D.R., Gowen, B.G., Boontanrart, M., Roth, T.L., Gagnon, J.D., Mumbach, M.R., Satpathy, A.T., Lee, Y., Bray, N.L., Chan, A.Y., et al. (2017). Discovery of stimulation-responsive immune enhancers with CRISPR activation. Nature *549*, 111–115.
31. Ray, J.P., de Boer, C.G., Fulco, C.P., Lareau, C.A., Kanai, M., Ulirsch, J.C., Tewhey, R., Ludwig, L.S., Reilly, S.K., Bergman, D.T., et al. (2020). Prioritizing disease and trait causal variants at the TNFAIP3 locus using functional and genomic features. Nat. Commun. *11*, 1237.
32. Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F., et al. (2021). Genome-wide enhancer maps link risk variants to disease genes. Nature *593*, 238–243.
33. Canver, M.C., Bauer, D.E., Dass, A., Yien, Y.Y., Chung, J., Masuda, T., Maeda, T., Paw, B.H., and Orkin, S.H. (2014). Characterization of genomic deletion efficiency mediated by clustered regularly interspaced palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells. J. Biol. Chem. *289*, 21312–21324.
34. Bosch-Guiteras, N., Uroda, T., Guillen-Ramirez, H.A., Riedo, R., Gazdhar, A., Esposito, R., Pulido-Quetglas, C., Zimmer, Y., Medová, M., and Johnson, R. (2021). Enhancing CRISPR deletion via pharmacological delay of DNA-PKcs. Genome Res. *31*, 461–471.
35. Soskic, B., Cano-Gamez, E., Smyth, D.J., Rowan, W.C., Nakic, N., Esparza-Gordillo, J., Bossini-Castillo, L., Tough, D.F., Larminie, C.G.C., Bronson, P.G., et al. (2019). Chromatin activity

at GWAS loci identifies T cell states driving complex immune diseases. Nat. Genet. *51*, 1486–1493.
36. Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.-L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.
37. Fiaux, P.C., Chen, H.V., Chen, P.B., Chen, A.R., and McVicker, G. (2020). Discovering functional sequences with RELICS, an analysis method for CRISPR screens. PLoS Comput. Biol. *16*, e1008194.
38. Kosicki, M., Tomberg, K., and Bradley, A. (2018). Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. Nat. Biotechnol. *36*, 765–771.
39. Owens, D.D.G., Caulder, A., Frontera, V., Harman, J.R., Allan, A.J., Bucakci, A., Greder, L., Codner, G.F., Hublitz, P., McHugh, P.J., et al. (2019). Microhomologies are prevalent at Cas9-induced larger deletions. Nucleic Acids Res. *47*, 7402–7417.
40. Schumann, K., Lin, S., Boyer, E., Simeonov, D.R., Subramaniam, M., Gate, R.E., Haliburton, G.E., Ye, C.J., Bluestone, J.A., Doudna, J.A., and Marson, A. (2015). Generation of knock-in primary human T cells using Cas9 ribonucleoproteins. Proc. Natl. Acad. Sci. USA *112*, 10437–10442.
41. Hendel, A., Bak, R.O., Clark, J.T., Kennedy, A.B., Ryan, D.E., Roy, S., Steinfeld, I., Lunstad, B.D., Kaiser, R.J., Wilkens, A.B., et al. (2015). Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. Nat. Biotechnol. *33*, 985–989.
42. Brinkman, E.K., Chen, T., Amendola, M., and van Steensel, B. (2014). Easy quantitative assessment of genome editing by sequence trace decomposition. Nucleic Acids Res. *42*, e168.
43. Hosoya-Ohmura, S., Lin, Y.-H., Herrmann, M., Kuroha, T., Rao, A., Moriguchi, T., Lim, K.-C., Hosoya, T., and Engel, J.D. (2011). An NK and T cell enhancer lies 280 Kilobase Pairs 3′ to the Gata3 structural gene. Mol. Cell Biol. *31*, 1894–1904.
44. Nakayama, T., Hirahara, K., Onodera, A., Endo, Y., Hosokawa, H., Shinoda, K., Tumes, D.J., and Okamoto, Y. (2017). Th2 Cells in Health and Disease. Annu. Rev. Immunol. *35*, 53–84.
45. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J. Roy. Stat. Soc. B Stat. Methodol. *82*, 1273–1300.
46. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.
47. Marcus, J.H., and Novembre, J. (2017). Visualizing the geography of genetic variants. Bioinformatics *33*, 594–595.