Publication details:	Chan WF, Coughlan HD, Iannarella N, Smyth GK, Johanson TM, Keenan CR, Allan RS. Identification and characterization of the long non-coding RNA Gm13218 as a novel regulator of Gata3. *Immunology and Cell Biology.* 2021 99(3):323-332
Published version Is available at:	https://doi.org/10.1111/imcb.12408

Terms of use:	This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

DR CHRISTINE KEENAN (Orcid ID : 0000-0002-6057-1855)
ASSOCIATE PROFESSOR RHYS ALLAN (Orcid ID : 0000-0003-0906-2980)

Article type : Short Communication

Identification and characterization of the long non-coding RNA Gm13218 as a novel regulator of Gata3

Wing Fuk Chan[1,2], Hannah D Coughlan[1,2], Nadia Iannarella[1], Gordon K Smyth[1,3], Timothy M Johanson[1,2]*, Christine R Keenan[1,2]*, Rhys S Allan[1,2]*

[1]The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia.
[2]Department of Medical Biology, The University of Melbourne, Parkville, VIC 3010, Australia.
[3]School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia.
*These authors contributed equally

Correspondence:
Rhys S Allan (rallan@wehi.edu.au).
Christine R Keenan (keenan.c@wehi.edu.au).
Immunology Division, Walter and Eliza Hall Institute of Medical Research
1G Royal Parade, Parkville
Victoria, Australia 3052
Ph: +61 3 9345 2999 fax: +61 3 9347 0852

Running Heading: Gm13218 as a novel regulator of Gata3

Keywords:
Genome organization

Enhancer

Long non-coding RNA

T cell

Th2 cell

Gata3

Abstract

The eukaryotic genome is three-dimensionally segregated into discrete globules of topologically associating domains (TADs), within which numerous cis-regulatory elements such as enhancers and promoters interact to regulate gene expression. In this study, we identify a T cell-specific sub-TAD containing the *Gata3* locus, and reveal a previously uncharacterised long non-coding RNA (lncRNA) Gm13218 within a distant enhancer lying approximately 280 kilobases downstream of *Gata3*. Gm13218 expression is highly correlated with that of Gata3 during early immune system development and T helper 2 (T_H2) cell differentiation. Inhibition and overexpression of *Gm13218* suggest that it may be involved in the establishment, but not the maintenance of *Gata3* expression. Overall, we propose that *Gm13218* is a novel regulator of *Gata3* and may inform therapeutic strategies in diseases such allergy and lymphoma where *Gata3* has a pathological role.

Introduction

Recent genome-wide studies have shown that the eukaryotic genome is physically segregated into topologically associating domains (TADs), within which promoters and other cis-regulatory elements such as enhancers preferentially interact with each other [1]. TADs may contain smaller nested sub-TADs; however, whether there is a functional distinction between these remains controversial [1, 2]. Regardless, elements across TADs and sub-TADs are largely shielded from crosstalk by boundaries enriched with CTCF sites and transcriptionally active housekeeping genes to ensure proper insulation. As such, these preferentially self-interacting domains form gene regulatory units to contain appropriate enhancer-promoter interactions. Although TADs appear to be largely conserved between cell types, certain TADs display lineage specificity [3]. These lineage-restricted TADs may harbour distinct cis-elements that are critical in driving the transcription of the underlying genes, and thereby influence the development and divergence of different cell lineages.

The immune system develops from multi-potent progenitor cells into a variety of functionally distinct cell types. T lymphocytes engage in the clearance of virally infected cells, or release different cytokines to coordinate proper and integrative immune functions. This functional diversity arises from the divergence of transcriptional programmes which are intricately linked with the underlying chromatin topology. Previous work from our lab and others have constructed Hi-C interaction profiles from various immune cell types, and demonstrated that cells of different immune lineages exhibit distinct genome organisation [3]. We have found thousands of differentially interacting regions (DIs) between distinct lineages, for example 15,152 DIs between naïve CD4$^+$ T and B cells [4]. Many of these DIs cluster within TADs and contain lineage-specific genes thus allowing potential identification novel cis-regulatory elements.

Here, we explore T cell-specific genome organisation and reveal a long non-coding RNA (lncRNA) Gm13218 in an enhancer of the *Gata3* gene. We find the expression of this lncRNA to be highly coordinated with Gata3. Although knockdown of the RNA transcript and CRISPR silencing had minimal impact on Gata3 expression in T cells, ectopic overexpression of Gm13218 in B cells induced Gata3 transcription. This lncRNA therefore presents as a novel regulator in the establishment but not maintenance of *Gata3* expression.

Results

Lineage-restricted interactions enable discoveries of novel cis-elements

In order to uncover T cell-specific cis-regulatory elements from our Hi-C data, we identified and ranked genomic regions bearing the most extensive clustering of DIs between CD4$^+$ T cells and B cells [4]. We focused on differential interactions where one genomic anchor bears a differentially expressed genes[5] in these populations, and found the top 5 differentially interacting regions to contain the *Bcl11b*, *Gata3*, *Gm12289*, *Ppm1h* and *Themis* loci (Figure 1a). These interactions are absent or grossly reduced in the stem and progenitor enriched Lin$^-$Sca1$^+$cKit$^+$ (LSK) population (Supp Fig 1a) as well as the mature granulocyte population (Supp Fig 1b), but are conserved in CD8$^+$ T cells (Supp Fig 1c). Given its importance in T cell differentiation and dysregulation in disease[6] we decided to focus on understanding the regulation of the *Gata3* gene. We noted that the T cell-specific interactions around the *Gata3* locus occur in a self-interacting region bounded by CTCF in T cells of approximately 430 kb of genetic space, and contains no other protein-coding genes. This strongly self-interacting region occurs within a larger regulatory landscape of nearly 2 Mb which is broadly conserved in B cells (Supp Fig 2) suggesting the T cell-specific structure is a sub-TAD [1] .

We next overlayed our Hi-C data with ATAC-seq data from T and B cells [8] to reveal putative cis-regulatory regions noting several T-cell specific accessible regions between 200 kb and 385 kb downstream of the *Gata3* transcription start site (TSS) (Figure 1b). We then used CRISPR/Cas9 mediated deletion of three putative long-range regulatory elements annotated G1, G2 and G3 (Figure 1b) in the EL4 T cell line to interrogate the functional importance of these regions. Deletion of region G1 increased Gata3 expression by 50%, whereas G2 deletion caused an 80% loss of transcription (Figure 1c). In contrast, removal of element G3 had no effect. Of note, G2 overlaps with an enhancer previously identified to be important for Gata3 expression and T cell development[9, 10], but not previously shown to physically interact with the Gata3 locus.

A long non-coding RNA Gm13218 is located in an enhancer of Gata3

Upon closer inspection within region G2, we found a transcription unit of a poorly-annotated non-coding RNA, Gm13218. Long non-coding RNAs (lncRNAs) and enhancer RNAs (eRNAs) can modulate gene transcription or mediate 3D chromatin looping, therefore we next embarked to characterise Gm13218. To determine the full-length sequence of Gm13218 we conducted 5' and 3' RACE on RNA from EL4 T cells. We determined Gm13218 to be a lncRNA with variable transcript lengths ranging from 630 to 690 nucleotides (Figure 2a, sequence of three variant transcripts deposited in Genbank: accession numbers MT791034, MT791035, MT791036). It consists of three exons with alternative splicing at the first intron-exon junction, it bears a 5' cap, and is polyadenylated. Furthermore, the promoter or TSS of Gm13218 coincides with the most prominent T cell specific accessible region in region G2 (Figure 2b).

Gm13218 is co-regulated with Gata3

Next, we performed RT-qPCR to investigate the expression pattern of Gm13218 in relation to Gata3 mRNA. Gata3 transcripts were abundant in the thymus and also present at lower levels in lymph node, spleen and kidney tissues, but barely detectable in bone marrow (Figure 2c). Similarly, Gm13218 is mostly expressed in thymus with limited expression in the lymph node and undetectable in spleen, bone marrow and kidney tissues (Figure 2d). These expression patterns were recapitulated in the T and B cell lines EL4 and A20 whereby EL4 T cells showed high levels of both Gata3 and Gm13218, whereas the A20 B cell line showed no expression of either RNA species. Among thymocytes of different T cell developmental transitions, both Gata3 and Gm13218 expressions peak at the DN2 and DN3 stages (Figure 2c, d). After the transition to DN4, both transcripts decrease and remain at low level in the later DP, $CD4^+$ and $CD8^+$ stages. Notably, the expression patterns of Gata3 and Gm13218 in different thymocyte subsets were found to be positively correlated (Figure 2e).

Given that Gata3 is also critical for T_H2 differentiation[6] we examined Gata3 and Gm13218 expression across 7 days of T_H1 and T_H2 polarised $CD4^+$ T cell differentiation. While expression of both Gata3 and Gm13218 remained low in unpolarised activated $CD4^+$ T cells and throughout T_H1 polarisation (Figure 2f, g), both transcripts were significantly upregulated in the first 24 hours of T_H2 culture, remaining high across the 7 days. Interestingly, Gata3 transcription increased and stayed at the peak level (5-fold) until the experimental end point, whereas Gm13218 transcripts peaked at days 2-4, reaching about 70-fold, and declined to 28-fold on day 7. Nonetheless, the expression patterns across the polarisation period are highly positively correlated (Figure 1h).

Gata3 regulatory elements are conserved between mouse and human

We noticed that the sequence of accessible regions within region G2 is conserved with a non-coding DNA region near GATA3 in human chromosome 10. We therefore next explored whether the regulatory interactions identified in mouse cells are conserved in human cells utilizing our previously published human Hi-C data from naïve B and $CD4^+$ T cells [11]. We found the conserved sequences to be both within the same self-associating structure as GATA3 and also located roughly 280 kb downstream of the GATA3 TSS in human $CD4^+$ T cells (Figure 3a), suggesting that the synteny and genetic distance between are also conserved. Like in mouse cells, these interactions are completely absent in human naïve B cells (Figure 3a), are conserved in human naïve $CD8^+$ T cells (Supplementary figure 3a), and occur within a broader self-interacting structure of nearly 2 Mb in size (Supplementary figure 3b). We then exploited the FANTOM database [12] to investigate the occurrence of non-coding transcription within the GATA3-containing sub-TAD. Intriguingly, bidirectional transcription appears to initiate from the sequence conserved cis-element from which Gm13218 is transcribed in mouse cells. This bidirectional transcription generates two lncRNAs, CAT00000105356.1 and CAT00000117261.1 (Figure 3b), which both show lineage specificity in T cells (Figure 3c, d). Thus overall, the T cell specific sub-TAD, the synteny, the interaction between GATA3 and a downstream lncRNA initiating cis-element, and the coordinated expression of both in T cells, are well conserved between mouse and human.

Functional interrogation of Gm13218

Finally, we assessed the function of the lncRNA Gm13218 in regulating Gata3 expression through a variety of approaches. We knocked down endogenous Gm13218 in EL4 T cells using a pool of four antisense oligonucleotides (ASOs) to target all exons simultaneously. While ASO delivery markedly downregulated Gm13218 to 20%, Gata3 transcription was surprisingly unaffected (Figure 4a). We also silenced transcription of Gm13218 via CRISPR interference using catalytically dead Cas9 (dCas9) fused to a KRAB domain. Stable transfectants showed the transcription of Gm13218 was greatly suppressed to only 3% of original levels, but Gata3 expression was again unaltered (Figure 4b), suggesting that Gm13218 has minimal influence on steady state Gata3 expression. We next attempted to assess if it has a functional role during the establishment of the 3D chromatin environment. To do this, we ectopically expressed Gm13218 in the B cell line A20 that express negligible levels of *Gata3* using a *PiggyBac* transposon vector. Stable transfectants of A20 revealed a successful overexpression of Gm13218, which led to a modest 1.7-fold upregulation of Gata3 levels (Figure 4c), suggesting that Gm13218 might play a role in the establishment of Gata3 expression.

Discussion

The advent of chromosome conformation capture techniques have allowed three-dimensional chromatin structures to be revealed[13-15]. Here, we have identified the uncharacterised lncRNA Gm13218 within a broader region previously identified to regulate Gata3 expression[9, 10], but not previously shown to physically interact with the Gata3 locus. Interestingly, our results suggest that Gm13218 might have a role in the establishment phase in initiating Gata3 expression, rather than a role in steady-state transcription.

LncRNAs or eRNAs are well known to exhibit very high tissue- or developmental stage-specificity [16], with functions only being critical in a limited developmental transitions or timeframe. However, addressing this in the appropriate setting such as early T cell development or T_H2 differentiation has thus far proven technically challenging. For example the rapid upregulation of Gm13218 and Gata3 within 24 hours in T_H2 conditions provides only a very limited timeframe and thereby makes it difficult to use the ASO or dCas9-KRAB based strategies that we readily employed in cell lines.

Gm13218 defies some other general properties of enhancer RNAs (eRNA) [17]. Transcription at enhancers usually generates bidirectional transcripts (2D-eRNAs), and at a lesser extent, unidirectional transcripts (1D-eRNAs). While 2D-eRNAs are generally non-polyadenylated and short with length less than 2 kb, 1D-eRNAs are polyadenylated and long, usually over 4 kb in length. Gm13218 is the predominant RNA species near the Gata3 enhancer and is polyadenylated, apparently fitting the properties of a 1D-eRNA. However, it is relatively short with a length of about 690 nt. More importantly, RNA splicing rarely occurs in eRNAs [18], yet Gm13218 is alternatively spliced, generating multiple RNA species with slight variations at the first exon-intron junction. On the other hand, a sub-group of lncRNAs, termed activating ncRNA (ncRNA-a), possess enhancer-like properties [17]. These ncRNA-a contrast with eRNAs in that ncRNA-a are spliced and polyadenylated RNA species, and are unidirectionally transcribed from promoter-like regions. As such, it overall appears that Gm13218 is an activating lncRNA transcribed from the Gata3 enhancer.

The 3-dimensional genome structure around the Gata3 locus was found to be conserved in human T cells and absent in human B cells. Furthermore, the TAD structure around Gata3 and the synteny between it and conserved cis-elements were also found to be conserved in human T cells. Through whole genome analyses of different vertebrate species, it has been demonstrated that the evolutionary DNA rearrangement events occurred most frequently at TAD boundaries [19]. Conversely, rearrangement that would break the synteny within a TAD is very rare. This indicates TADs and sub-TADs are critical regulatory units in which the underlying 3D interactions and hence gene expression patterns are crucial and therefore well preserved through evolution. As such, the conservation of synteny and the long-range contact between Gata3 and lncRNA initiating cis-elements indicates these elements are likely important to Gata3 expression across species.

Gata3 is implicated as a mediator driving allergic asthma and T cell lymphoma[6]. Strategies have therefore been developed to target Gata3 protein as well as coding mRNA. However, Gata3 is also expressed in other non-haematopoietic tissues of neural, urogenital and cardiac origins, and thus interference may have negative consequences. Targeting cis-elements may allow greater specificity than targeting the transcription factor as different tissues appear to utilise distinct individual cis-regulatory elements (e.g. Gata3 utilizes an enhancer 571 kb downstream of the TSS in inner-ear cells[20], and 113 kb upstream of its TSS in kidney tissue[21]). As such, with the ever-growing toolbox of genome editing techniques, the deciphering of cis-regulome in a three-dimensional context could pave the way for a finely tuned stage- or lineage-specific adjustment of transcription.

Methods

Mice and cell isolation

Primary cells and tissues were obtained from male C57BL/6 mice between 6-12 weeks old unless otherwise specified. All mice were maintained at The Walter and Eliza Hall Institute Animal Facility under specific-pathogen-free conditions and were randomly chosen for each experiment. Murine and human splenic CD4$^+$ and mature B cell isolations have been described previously [4, 11]. Thymic cells were harvested from male C57BL/6 mice into single-cell suspension with red cell lysis and stained with CD19-PacB (ID3, in-house, WEHI, VIC, Australia), CD8-APC (53.6.7, in-house, WEHI), CD4-PE (GK1.5, in-house, WEHI), CD25-PE/Cy7 (PC61.5, eBioscience, Scoresby, VIC, Australia) and CD44-FITC (IM7, in-house, WEHI) antibodies. Thymic CD4$^+$ cells were obtained as CD19$^-$ CD4$^+$ CD8$^-$, whereas thymic CD8$^+$ cells as CD19$^-$ CD4$^-$ CD8$^+$. DP cells were sorted as CD19$^-$ CD4$^+$ CD8$^+$. DN1 population was obtained from CD19$^-$ CD4$^-$ CD8$^-$ CD44$^+$ CD25$^-$ population. DN2 was sorted as CD19$^-$ CD4$^-$ CD8$^-$ CD44$^+$ CD25$^+$, DN3 as CD19$^-$ CD4$^-$ CD8$^-$ CD44$^-$ CD25$^+$. Finally, DN4 was acquired from CD19$^-$ CD4$^-$ CD8$^-$ CD44$^-$ CD25$^-$ population.

Cell culture

All immune cells were cultured in RPMI 1640 with 2 mM GlutaMAX (Life technologies, Mulgrave, VIC, Australia), 50 µM β-mercaptoethanol (Sigma-Aldrich, North Ryde, NSW, Australia) and 10% heat-inactivated foetal calf serum (FCS; Sigma-Aldrich) unless specified otherwise. HEK 293T cells were cultured in DMEM with 2 mM GlutaMAX and 10% heat-inactivated FCS without antibiotics. Naïve CD4$^+$ T lymphocytes were activated and polarised as previously described [22, 23]. Briefly, a flat-bottom 96-well plate was coated with 10 µg mL^{-1} anti-CD3 (145-2C11, Becton Dickinson, Macquarie Park, NSW, Australia) and 5 µg mL^{-1} anti-CD28 (37.51, Becton Dickinson) in PBS overnight at 4°C. Splenic cells were harvested from male C57BL/6 mice into single-cell suspension with red cell lysis. Naïve CD4$^+$ T lymphocytes were then purified via negative selection (Miltenyi Biotec, Macquarie Park, NSW, Australia). Cells were then resuspended with cell media with corresponding polarising cytokines at a density of 1x10^6 cells mL^{-1} and seeded into the antibody-coated plate at 100,000 cells/well. T$_H$1 culture medium consisted of 5 ng mL^{-1} murine IL-12 (Peprotech, Cranbury, NJ, USA) and 10 µg mL^{-1} anti-IL-4 (eBioscience). T$_H$2 culture medium contained 50 ng mL^{-1} murine IL-4 (Peprotech), 10 µg mL^{-1} anti-IL-12 (eBioscience) and 10 µg mL^{-1} anti-IFN-γ (eBioscience). Cells were removed from anti-CD3 and anti-CD28 stimulation 48 hours later and replated in the same T$_H$1 or T$_H$2 conditions with addition of human recombinant IL-2 (Peprotech) at 30 U mL^{-1}.

In situ Hi-C

We analysed *In situ* Hi-C data from mouse and human libraries. The mouse data was from our previous publications [4, 11], deposited in GSE99151 and GSE105918. Human *in situ* Hi-C data was from our previous publication [11], deposited in GSE105776**.** Libraries were preprocessed and analysed as in [4, 11] with parameter changes for the differential interaction (DI) analysis.

DIs between the cell types were detected using the diffHic package [24]. Read pairs were counted into 20 kbp bin pairs. Bins with counts less than 5, on the first diagonal of the interaction space or containing blacklisted genomic regions as defined by ENCODE for mm10 or hg39 were removed. For filtering, bin pairs were retained if they had average interaction intensities more than 5-fold higher than the background ligation frequency. The ligation frequency was estimated from the inter-chromosomal bin pairs from a 1 Mbp bin-pair count matrix. Counts were normalized between libraries using a LOESS-based method with bin pairs less than 80 kbp from the diagonal normalized separately from other bin pairs. Tests for differential interactions were performed using the quasi-likelihood (QL) framework and a robust empirical Bayes strategy from the edgeR package. For the mouse analysis, the design matrix was constructed using a one-way layout that specified the cell type. For the human analysis, the design matrix was constructed using a layout that specified the cell type and the human donor. A generalized linear model (GLM) was fitted to the counts for each bin pair [25]. For each bin pair, a *P*-value was computed and adjusted for multiple testing using the Benjamini-Hochberg method. If a bin pair had a FDR below 5%, it was defined as a DI. To reduce redundancy in the results, DIs adjacent in the interaction space were aggregated into clusters. DIs were merged into a cluster if they overlapped in the interaction space, to a maximum cluster size of 2 Mbp. The significance threshold for each bin pair was defined such that the cluster-level FDR was controlled at 5%. Cluster statistics were computed using csaw package v1.12.0 [26]. Clustered DIs were ranked by number of contributing bin pairs with a positive logFC.

Contact matrices were created from the libraries using the inflate function in diffHic for various bin sizes with no filtering. Contact matrices from biological replicates were summed. Plaid plots were constructed using the contact matrices and the plotHic function from the Sushi R package. The inferno color palette from the *viridis* package (https://CRAN.R-project.org/package=viridis) was used and the range of color intensities in each plot was scaled according to the library size of the sample, to facilitate comparisons between plots from different samples. DI arcs were plotted with the plotBedpe function of the Sushi package. The z-score shown on the vertical axis was calculated as $-\log_{10}$ (*P*-value). ATAC-seq and CTCF ChIP-seq coverage was plotted with the plotBedgraph function of the Sushi package. Processed ATAC-seq profiles were obtained from GEO GSE100738 [8] and converted from BigWig to bedGraph files with bamCoverage from Deeptools v2.5.3 [27]. CTCF ChIP-seq fastq files were downloaded from GSE60482 for the CD4$^+$ T cells and GSE44637 for B cells. Files were aligned with Rsubread to the mm10 reference genome. Duplicate reads were removed with Picard tools (https://broadinstitute.github.io/picard/). BedGraphs were created with bamCoverage from Deeptools.

dCas9-KRAB

pHR-SFFV-dCas9-BFP-KRAB (#46911) was obtained from Addgene (Watertown, MA, USA). gRNA vector was constructed from pLH-sgRNA1-2XPP7 (Addgene# 75390). In brief, a modified version of sgRNA scaffold, with an A-U flip and hairpin extension as to confer better stability [28], was ordered from IDT as gBlocks gene fragment and amplified with a forward primer containing an AgeI site and a BbsI cloning cassette, and a reverse primer containing EcoRI site (Supp Table 1). The PCR product was subsequently cloned into the corresponding site in pLH-sgRNA1-2XPP7. The hygromycin resistance gene was replaced with an mKO2 gene that has the BbsI site removed by site-directed mutagenesis (Supp Table 1). The target sites for Cas9 were designed either by CHOPCHOP [29] or through the IDT online design tool (Supp Table 1). For cloning target sequence into the corresponding gRNA vector, the 20 bp protospacer sequence, ordered as a pair of complementary oligos with additional nucleotides ACCG- and AAAC- at the 5' end of the sense and antisense oligos, respectively, were annealed by heating at 95°C for 5 minutes and subsequent cooling to room temperature at a rate of -0.1°C/s. The annealed oligos were ligated to the BbsI cut site of the gRNA vector.

Lentivirus production and transduction

One day prior to transfection, HEK293T cells were seeded at a density of $1.2x10^6$ cells/well in a 6-well plate in 2 ml Opti-MEM I (Invitrogen, Scoresby, VIC, Australia), 2mM GlutaMAX, 1mM Sodium Pyruvate (Sigma-Aldrich) and 5% FCS. Transfection of HEK293T was performed using Lipofectamine 3000 (Invitrogen) as per the manufacturers' instructions. Cells were co-transfected with packaging plasmids (pCMV-VSV-g and psPAX2) at 0.17 pmol each and about 0.23 pmol transfer construct to make up a final mass of 3.3 µg. Virus was harvested 24- and 52-hour post-transfection. Transduction was performed in a 12-well plate, with 500,000 cells resuspended in 1 mL viral supernatant supplemented with 8 µg mL^{-1} polybrene (Merck-Millipore, North Ryde, NSW, Australia). Cells were spun at 2500 rpm at 32°C for 90 minutes. Stable dCas9-KRAB transfectants were sorted by FACS and allowed to grow for two weeks before analysis.

In vitro transcription of sgRNA

Transcription template was generated by PCR using Q5 high fidelity DNA polymerase (New England Biolabs, Ipswich, MA, USA) with dNTP (Promega, Madison, WI, USA) and an annealing temperature of 58°C. Universal primers as well as a specific primer bearing the sgRNA flanked by T7 promoter sequence and scaffold were used (Supplementary table 1). Deletional sgRNAs are also detailed in Supplementary table 1. PCR products were purified by DNA clean & concentrator-25 (Zymo Research, Irvine, CA, USA). *In vitro* transcription was performed by incubating 5 µg of transcription template with NTP, pyrophosphatase (ThermoFisher, Scoresby, VIC, Australia), RNase Inhibitor (Lucigen, Middleton, WI, USA) and NxGen T7 RNA polymerase (Lucigen) at 37 °C for 18 hours. Transcription template was then digested by TURBO DNase (Invitrogen) at 37 °C for 45 minutes. The remaining sgRNA was then purified by RNA clean & concentrator-25 (Zymo Research).

Ribonucleoprotein (RNP) assembly and delivery

For Cas9 RNP, 150 pmol of *in vitro* transcribed sgRNA was incubated with 100 pmol of recombinant Cas9 nuclease (Integrated DNA Technologies, Coralville, IA, USA) at room temperature for 15 minutes. RNP with 100 pmol of electroporation enhancer (Integrated DNA Technologies) were subsequently transfected into cells via electroporation. Delivery into EL4 was performed via 4D-Nucleofector (Lonza, Basel, Switzerland) with buffer SF and pulse code CM120.

RACE (Rapid amplification of cDNA ends)

RNA of EL4 cells was extracted using NucleoSpin RNA Plus (Macherey-Nagel, Düren, Germany) with gDNA removal, and was dephosphorylated by calf intestinal alkaline phosphatase (New England Biolabs) at 37°C for 90 minutes. Decapping of RNA was then conducted by RNA 5' Pyrophosphohydrolase (New England Biolabs) with incubation at 37°C for 2 hours. RACE DNA-RNA chimeric adapter (Supplementary table 1) was ligated to the decapped RNA by T4 RNA Ligase 1 (New England Biolabs) through an incubation at 17°C for 16 hours. The adapter-ligated RNA was reverse transcribed using Superscript IV (Invitrogen) with anchored oligo dT primer as per manufacturer’s instruction. In 3' RACE an overhang oligo dT primer (Supplementary table 1) was used to generate cDNA from the unmodified EL4 RNA with Superscript IV. Gm13218 5' sequence was amplified by an overhang primer specific to the RACE adapter, together with an overhang-specific primer and a gene specific primer (Supplementary table 1). PCR was performed using Phusion HS (ThermoFisher) with thermocycle of 98°C for 30s, followed by 35 cycles of 98°C for 10s and 72°C for 1 minute. Nested PCR using a second gene specific primer (Supplementary table 1) was performed on the diluted PCR product from the first round. Gm13218 3' sequence was amplified with gene specific primers as detailed in Supplementary table 1. RACE products were cloned and amplified with CloneJET PCR cloning kit (ThermoFisher) and sequence verified. Three variants are deposited in Genbank under accession numbers MT791034, MT791035 and MT791036.

Antisense oligo (ASO) knockdown

Antisense oligos were designed to target different exons, together with negative control ASO they were all ordered as LNA (locked nucleic acid) GapmeRs (Qiagen, Hilden, Germany). Experimental ASO sequences are detailed in Supplementary table 1. ASOs were reconstituted at a concentration of 50 µM and the 4 ASOs targeting Gm13218 were pooled together at a final reaction concentration of 6 µM, whereas negative control ASO was used alone at 6 µM. Electroporation was performed on EL4 cells (buffer SF, pulse code CM120) using the 4D-Nucleofector system (Lonza) on day 0 and day 2 with RNA extracted on day 3.

PiggyBac non-coding RNA overexpression

A hyperactive version of *piggyBac* transposase has been characterised [30] and the plasmid clone was a kind gift from Wellcome Sanger Institute, UK. Vector with *piggyBac* 5'- and 3'-LTR (long terminal repeat) was obtained from Addgene (#84241). Non-coding RNA Gm13218 driven by CMV promoter, a bGH polyadenylation signal and TagBFP driven by PGK promoter (Supplementary table 1) were assembled into *piggyBac* transposon using NEBuilder HiFi DNA Assembly (New England Biolabs). To generate a stable cell line expressing Gm13218, 1 µg of transposon plasmid together with 1 µg *piggyBac* transposase plasmid were co-transfected into A20 using 4D-Nucleofector in buffer SF with pulse code FF113. Transposon with TagBFP alone was used as the control.

Quantitative reverse transcription PCR (RT-qPCR)

RNA was extracted using NucleoSpin RNA Plus (Macherey-Nagel) with gDNA removal. One step RT-qPCR was performed using 20 ng RNA with iTaq Universal probe supermix (Bio-Rad, Hercules, CA, USA) and β-actin acting as a reference. Primer and probe sequences are detailed in Supplementary table 1. Gene expression was normalised to the endogenous control (ΔCT) and relative expression was evaluated as 2^-ΔCT.

Acknowledgments

We thank the staff of the core facilities at the Walter and Eliza Hall Institute, particularly the Flow Cytometry facility (WEHIFACS) and the Bioservices department. This work was supported by grants and fellowships from the National Health and Medical Research Council of Australia (CRK #1125436, TMJ #1124081, RSA #1100451, GKS & RSA #1158531), and the Australian Research Council (RSA # 130100541). This study was made possible through Victorian State Government Operational Infrastructure Support and Australian Government NHMRC Independent Research Institute Infrastructure Support scheme and the Australian Cancer Research Fund.

Author contributions

Investigation: WFC, NI, TMJ. Methodology: WFC. Formal Analysis: HDC, GKS. Visualisation: WFC, HDC, CRK. Writing – original draft: WFC. Writing – review & editing: CRK, RSA. Supervision: TMJ, CRK, RSA. Funding acquisition: GKS, TMJ, CRK, RSA. Conceptualization: RSA.

Declaration of interests

The authors declare no competing interests.

References

1. Bonev B, Cavalli G. Organization and function of the 3D genome. *Nat Rev Genet* 2016; **17:** 661-678.

2. Dixon JR, Gorkin DU, Ren B. Chromatin Domains: The Unit of Chromosome Organization. *Mol Cell* 2016; **62:** 668-680.

3. Stadhouders R, Filion GJ, Graf T. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* 2019; **569** (7756)**:** 345-354.

4. Johanson TM, Lun ATL, Coughlan HD, *et al.* Transcription-factor-mediated supervision of global genome architecture maintains B cell identity. *Nat Immunol* 2018; **19:** 1257-1264.

5. Heng TS, Painter MW, Immunological Genome Project C. The Immunological Genome Project: networks of gene expression in immune cells. *Nat Immunol* 2008; **9:** 1091-1094.

6. Zaidan N, Ottersbach K. The multi-faceted role of Gata3 in developmental haematopoiesis. *Open Biol* 2018; **8**: 180152.

7. Phillips-Cremins JE, Sauria ME, Sanyal A, *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* 2013; **153:** 1281-1295.

8. Yoshida H, Lareau CA, Ramirez RN, *et al.* The cis-Regulatory Atlas of the Mouse Immune System. *Cell* 2019; **176:** 897-912 e20.

9. Hosoya-Ohmura S, Lin YH, Herrmann M, *et al.* An NK and T cell enhancer lies 280 kilobase pairs 3' to the gata3 structural gene. *Mol Cell Biol* 2011; **31:** 1894-1904.

10. Ohmura S, Mizuno S, Oishi H, *et al.* Lineage-affiliated transcription factors bind the Gata3 Tce1 enhancer to mediate lineage-specific programs. *J Clin Invest* 2016; **126:** 865-878.

11. Johanson TM, Coughlan HD, Lun ATL, *et al.* Genome-wide analysis reveals no evidence of trans chromosomal regulation of mammalian immune development. *PLoS Genet* 2018; **14:** e1007431.

12. Hon CC, Ramilowski JA, Harshbarger J, *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* 2017; **543** (7644)**:** 199-204.

13. Lieberman-Aiden E, van Berkum NL, Williams L, *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009; **326** (5950)**:** 289-293.

14. Nora EP, Lajoie BR, Schulz EG, *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 2012; **485** (7398)**:** 381-385.

15. Dixon JR, Selvaraj S, Yue F, *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012; **485** (7398)**:** 376-80.

16. Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* 2014; **15:** 7-21.

17. Orom UA, Shiekhattar R. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell* 2013; **154:** 1190-1193.

18. Koch F, Fenouil R, Gut M, *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat Struct Mol Biol* 2011; **18:** 956-963.

19. Krefting J, Andrade-Navarro MA, Ibn-Salem J. Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biol* 2018; **16:** 87.

20. Moriguchi T, Hoshino T, Rao A, *et al.* A *Gata3* 3' Distal Otic Vesicle Enhancer Directs Inner Ear-Specific *Gata3* Expression. *Mol Cell Biol* 2018; **38**: e00302-18.

21. Hasegawa SL, Moriguchi T, Rao A, Kuroha T, Engel JD, Lim KC. Dosage-dependent rescue of definitive nephrogenesis by a distant Gata3 enhancer. *Dev Biol* 2007; **301:** 568-577.

22. Allan RS, Zueva E, Cammas F, *et al.* An epigenetic silencing pathway controlling T helper 2 cell lineage commitment. *Nature* 2012; **487**(7406)**:** 249-253.

23. Keenan CR, Iannarella N, Garnham AL, *et al.* Polycomb repressive complex 2 is a critical mediator of allergic inflammation. *Jci Insight* 2019; **4**: e127745.

24. Lun AT, Smyth GK. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *Bmc Bioinformatics* 2015; **16:** 258.

25. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research* 2012; **40:** 4288-97.

26. Lun AT, Smyth GK. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic acids research* 2016; **44:** e45.

27. Ramirez F, Ryan DP, Gruning B, *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research* 2016; **44**(W1)**:** W160-165.

28. Chen B, Gilbert LA, Cimini BA, *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* 2013; **155:** 1479-1491.

29. Labun K, Montague TG, Krause M, Torres Cleuren YN, Tjeldnes H, Valen E. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic acids research* 2019; **47**(W1)**:** W171-W174.

30. Yusa K, Zhou L, Li MA, Bradley A, Craig NL. A hyperactive piggyBac transposase for mammalian applications. *Proc Natl Acad Sci USA* 2011; **108:** 1531-1536.

Figure Captions:

Figure 1. Identification of T cell-specific cis-regulatory elements

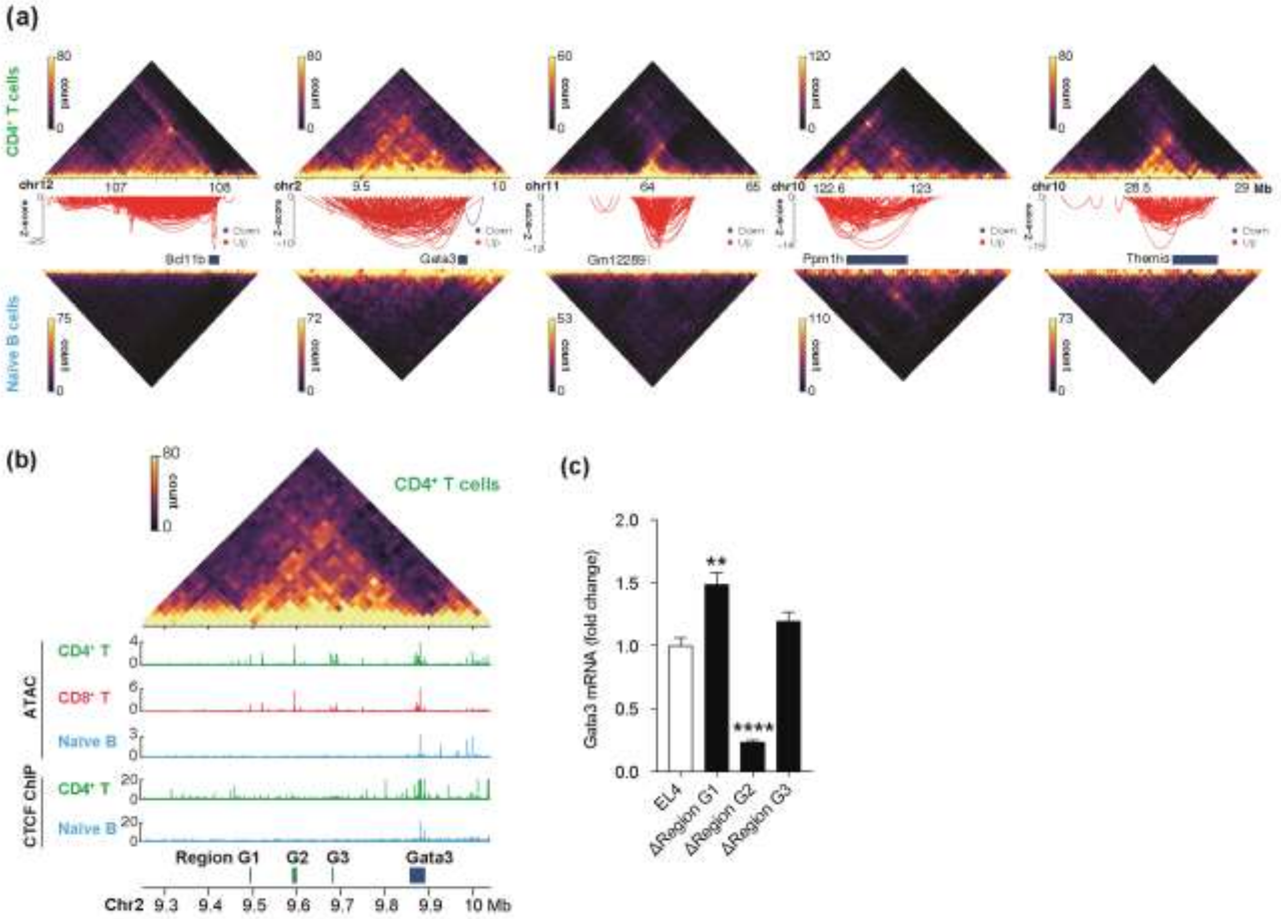
(a) Hi-C contact matrices of the top 5 genes showing T cell-specific chromatin organisation around Bcl11b, Gata3, Gm12289, Ppm1h and Themis loci between murine splenic naïve CD4$^+$ T cells and splenic naïve B cells. Differential interactions are shown as arcs with Z-score as -log$_{10}$ (*P*-value). Up indicates interactions that are stronger in CD4$^+$ T cells compared to B cells, whereas down indicates weaker interactions. **(b)** Hi-C contact map of Gata3 locus in naïve CD4$^+$ T cells overlaid with ATAC-seq chromatin accessibility profiles of splenic naïve B, CD4$^+$ and CD8$^+$ T cells and CTCF ChIP-seq from naïve B and CD4$^+$ T cells. Putative cis-elements G1, G2 and G3 are annotated. **(c)** Transcription levels of Gata3 in different cis-element knockout mutants compared to unmodified EL4 cell line. Data shown as mean and SEM from 4–10 independent biological replicates. ***P* < 0.01, *****P* < 0.0001 from a two-tailed Student's *t*-test comparing to the control.

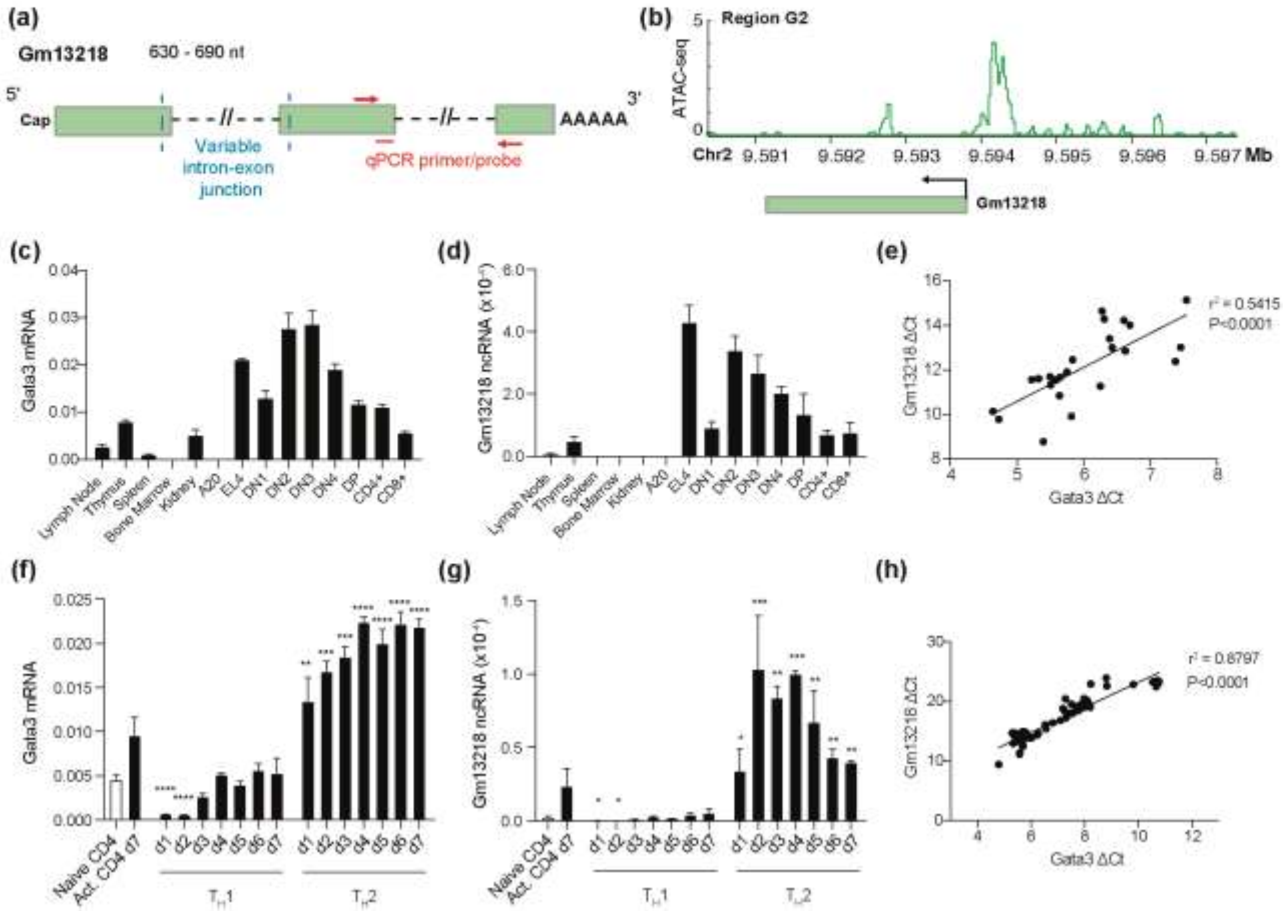
Figure 2. The lncRNA Gm13218 from a distal enhancer is co-expressed with Gata3 during T cell development and Th2 differentiation. (a) Schematic of the Gm13218 RNA transcript structure. **(b)** Chromatin accessibility profile of splenic naïve CD4$^+$ T cells within region G2 aligned with the Gm13218 TSS. Blue boxes indicate exons, whereas red arrows and bar represent primers and probe used for rt-qPCR. **(c,d)** Gata3 and Gm13218 expression level across tissues and various thymic T cell subsets. Data obtained from three independent biological replicates and shown as mean and SEM normalised to Actb (2^-ΔCT). **(e)** Correlation of expression between Gata3 and Gm13218 across different thymic T cell subsets, plotted as ΔCt of each gene to Actb. **(f, g)** Gata3 and Gm13218 expression level across 7 days of non-biased, T$_H$1 and T$_H$2 CD4$^+$ T cell activation shown as mean and SEM normalised to Actb expression (2^-ΔCT) from three independent experiments. **P* < 0.05, ***P* < 0.01, ****P* < 0.001, *****P* < 0.0001 from one-way ANOVA with Dunnett's post-hoc test of ΔCt values comparing to the naïve CD4 control group. **(h)** Correlation of expression between Gata3 and Gm13218 across the 7 days activation period for T$_H$1, T$_H$2 and non-biased activation, plotted as ΔCt of each gene to Actb.

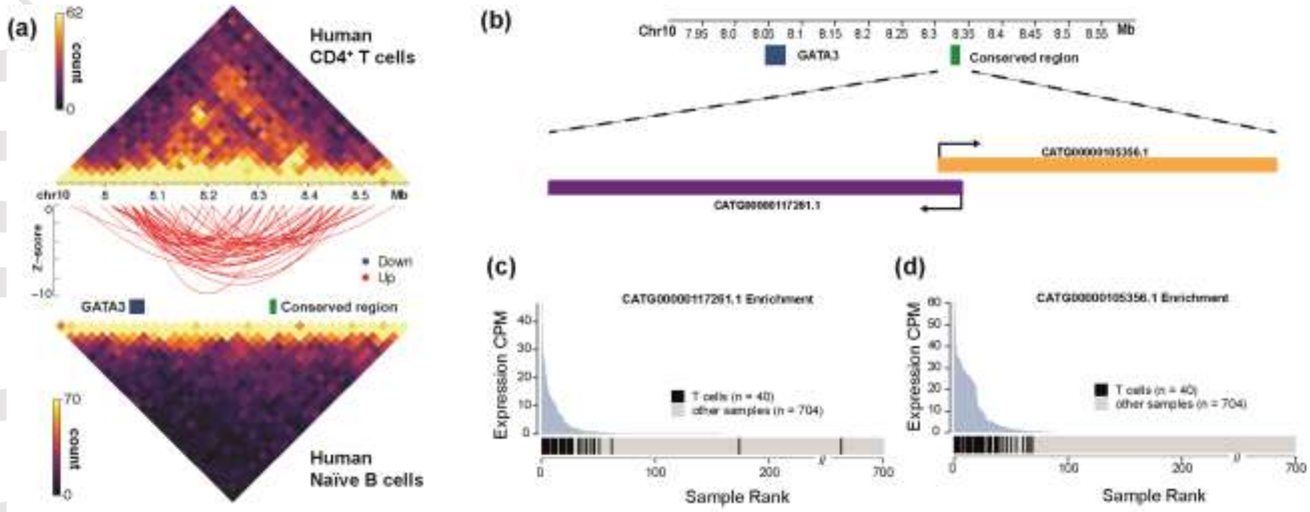
Figure 3. Synteny of the distal enhancer and Gata3 loci is conserved in human T cells. (a) Hi-C contact matrices around the GATA3 locus in human naïve B and CD4$^+$ T cells. Blue bar indicates GATA3 locus, whereas green bar represents the sequence conserved

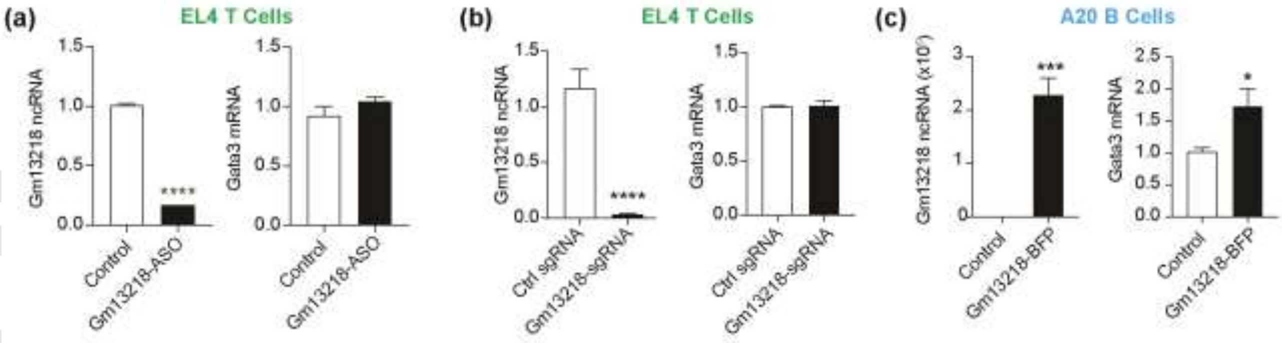
cis-elements. Differential interactions are shown as arcs with Z-score as -$\log_{10}$ (*P*-value). Up indicates interactions that are stronger in $CD4^+$ T cells compared to B cells, whereas down indicates weaker interactions. **(b)** LncRNA species transcribed from the conserved cis-elements, identified by FANTOM database. Length of bars for lncRNAs not in scale. **(c, d)** Expression enrichment profiles of lncRNAs CAT00000105356.1 and CAT00000117261.1 across different cell types. Data retrieved from FANTOM. Black vertical lines indicate the T cell samples, with expression ranked in a descending order from left to right.

Figure 4. Functional interrogation of Gm13218. (a) Gata3 and Gm13218 expression upon delivery of ASOs. Four ASOs targeting different regions of Gm13218 were pooled together during transfection. Data shown as mean and SEM from 2 independent experiments. **(b)** Gata3 and Gm13218 transcript levels after CRISPR silencing of the Gm13218 locus mediated by dCas9-KRAB. Data shown as mean and SEM from three independent experiments. **(c)** Gata3 and Gm13218 levels upon ectopic and stable overexpression of Gm13218. Data shown as mean and SEM from six biological replicates. *$P < 0.05$, ***$P < 0.001$, ****$P < 0.0001$ from a two-tailed *t*-test comparing to the control. Transcript levels for all the RT-qPCR experiments were normalised to the Actb endogenous control.


(a)
CD4⁺ T cells
Naive B cells
80
count
0
chr12
107
108
Z-score
Down
Up
Bcl11b
75
80
chr2
9.5
10
-10
Gata3
72
60
chr11
64
65
-12
Gm12289
53
120
chr10
122.6
123
-14
Ppm1h
110
80
chr10
28.5
29 Mb
-15
Themis
73
(b)
80
count
0
CD4⁺ T cells
ATAC
CD4⁺ T
4
0
CD8⁺ T
6
0
Naive B
3
0
CTCF ChIP
CD4⁺ T
20
0
Naive B
20
0
Region G1
G2
G3
Gata3
Chr2
9.3
9.4
9.5
9.6
9.7
9.8
9.9
10 Mb
(c)
Gata3 mRNA (fold change)
2.0
1.5
1.0
0.5
0.0
**
****
EL4
ΔRegion G1
ΔRegion G2
ΔRegion G3


(a)
Gm13218
630 - 690 nt
5'
Cap
3'
AAAAA
Variable intron-exon junction
qPCR primer/probe
(b)
Region G2
ATAC-seq
5
0
Chr2 9.591 9.592 9.593 9.594 9.595 9.596 9.597 Mb
Gm13218
(c)
Gata3 mRNA
0.04
0.03
0.02
0.01
0.00
Lymph Node
Thymus
Spleen
Bone Marrow
Kidney
A20
EL4
DN1
DN2
DN3
DN4
DP
CD4+
CD8+
(d)
Gm13218 ncRNA (x10⁻³)
6.0
4.0
2.0
0.0
(e)
Gm13218 ΔCt
Gata3 ΔCt
r² = 0.5415
P<0.0001
16
14
12
10
8
4 5 6 7 8
(f)
Gata3 mRNA
0.025
0.020
0.015
0.010
0.005
0.000
Naive CD4
Act. CD4 d7
d1 d2 d3 d4 d5 d6 d7
TH1
TH2
(g)
Gm13218 ncRNA (x10⁻³)
1.5
1.0
0.5
0.0
(h)
Gm13218 ΔCt
Gata3 ΔCt
r² = 0.8797
P<0.0001
30
20
10
0
4 6 8 10 12


(a)
62
count
0
Human CD4+ T cells
chr10 8 8.1 8.2 8.3 8.4 8.5 Mb
Z-score
0
-10
Down
Up
GATA3
Conserved region
70
count
0
Human Naïve B cells
(b)
Chr10 7.95 8.0 8.05 8.1 8.15 8.2 8.25 8.3 8.35 8.4 8.45 8.5 8.55 Mb
GATA3
Conserved region
CATG00000105356.1
CATG00000117261.1
(c)
CATG00000117261.1 Enrichment
Expression CPM
40
30
20
10
0
T cells (n = 40)
other samples (n = 704)
0 100 200 700
Sample Rank
(d)
CATG00000105356.1 Enrichment
Expression CPM
60
50
40
30
20
10
0
T cells (n = 40)
other samples (n = 704)
0 100 200 700
Sample Rank


(a)
EL4 T Cells
Gm13218 ncRNA
1.5
1.0
0.5
0.0
****
Control
Gm13218-ASO
Gata3 mRNA
1.5
1.0
0.5
0.0
Control
Gm13218-ASO
(b)
EL4 T Cells
Gm13218 ncRNA
1.5
1.0
0.5
0.0
****
Ctrl sgRNA
Gm13218-sgRNA
Gata3 mRNA
1.5
1.0
0.5
0.0
Ctrl sgRNA
Gm13218-sgRNA
(c)
A20 B Cells
Gm13218 ncRNA (x10)
3
2
1
0
***
Control
Gm13218-BFP
Gata3 mRNA
2.5
2.0
1.5
1.0
0.5
0.0
*
Control
Gm13218-BFP