## 

## Chromatin loops and expression QTL colocalization reveal novel gene targets for T1D-associated GWAS variants in immune cells

­Team: Ay Lab

Contact: Ferhat Ay Ph.D - [ferhatay@lji.org](mailto:ferhatay@lji.org)

Contact: Joaquin Reyna - [jreyna@lji.org](mailto:jreyna@lji.org)

## 

## Section A: Hypothesis Proposal

### Abstract

Type 1 diabetes (T1D) is a disease characterized by the destruction of β cell populations in the pancreas. Immune cells, and specifically T cells, have been implicated to play a key role in destroying insulin producing β cells by infiltrating the pancreas. To better understand the role of immune regulation in T1D, we colocalized the gene expression quantitative trait loci (eQTL) signals from 18 different immune cell populations (15 from DICE, 3 from BLUEPRINT) with T1D GWAS signals to gather non-coding variants that are likely causal for both the gene expression and the disease association. We further overlapped these variants with chromatin loops mapped in a subset of these immune cell populations to identify potential target genes of the significant non-coding SNPs. Aside from well-studied genes such as *BACH2, UBASH3A, PTPN22* and *SIRPG,* we identified *AP003774.1*, a long non-coding RNA, that is looping to a ~15kb away regulatory element overlapping a colocalized SNP (rs479777) in various T cell subsets. The looped region overlaps the promoter of another gene (promoter-promoter loop), *CCDC88B*, however, the eQTL association for this SNP is specific to *AP003774.1* and is remarkably strong for resting T cell subsets, NK cells and naïve B cells. The same SNP creates strong binding sites for multiple important transcription factors in donors with the non-reference allele leading to higher expression of *AP003774.1.* We hypothesize that the overexpression of *AP003774.1* lncRNA mediated through specific non-coding variants in different immune cell populations play a role in immune-related aspects of T1D.

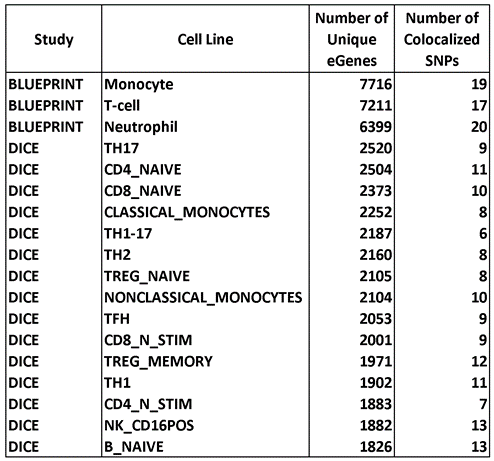
### Brief Statement

Our initial hypothesis was that the genetic variants identified through eQTL colocalization analysis and linked to their target genes using cell-type-specific chromatin loops are highly likely to play important roles in immune-related aspects of T1D. In support of this hypothesis, we identified well-known genes in T1D through these analyses but we also found novel genes including a particular lncRNA that is a potential target for further T1D research. *AP003774.1* is a long non-coding RNA, the expression of which is under control of genetic variants including rs479777 in different immune cell populations including CD4+ T cell subsets. This particular variant sits on critical positions for the binding of multiple different transcription factors and acts on *AP003774.1* through a promoter-promoter loop identified in the same immune cells. Our emerging hypothesis is that the dysregulation of *AP003774.1* expression in immune cells is associated with T1D risk and pathogenesis, which warrants further study.

### Hypothesis

We integrated several datasets including T1D GWAS, eQTL studies in several immune cells, HiChIP in the same cell lines, single-cell ATAC-seq data, transcription factor binding motifs and several literature-based sources to prioritize a small subset of candidate T1D genes. As part of this effort we will explain each dataset, our pipeline and then dive into the *AP003774.1* locus, which is a novel finding from our approach in the context of T1D.

Type 1 diabetes has been shown to include a strong genetic component. Recently, Chiou et al., 2021 performed a T1D-GWAS study using 18,942 T1D patients and 501,638 controls revealing a total of 92 loci with 136 independent association signals. In addition, they performed single-nucleus ATAC-seq on peripheral blood and pancreas followed by clustering, which provided open chromatin information for each distinct cell type in those tissues. The large sample size and results



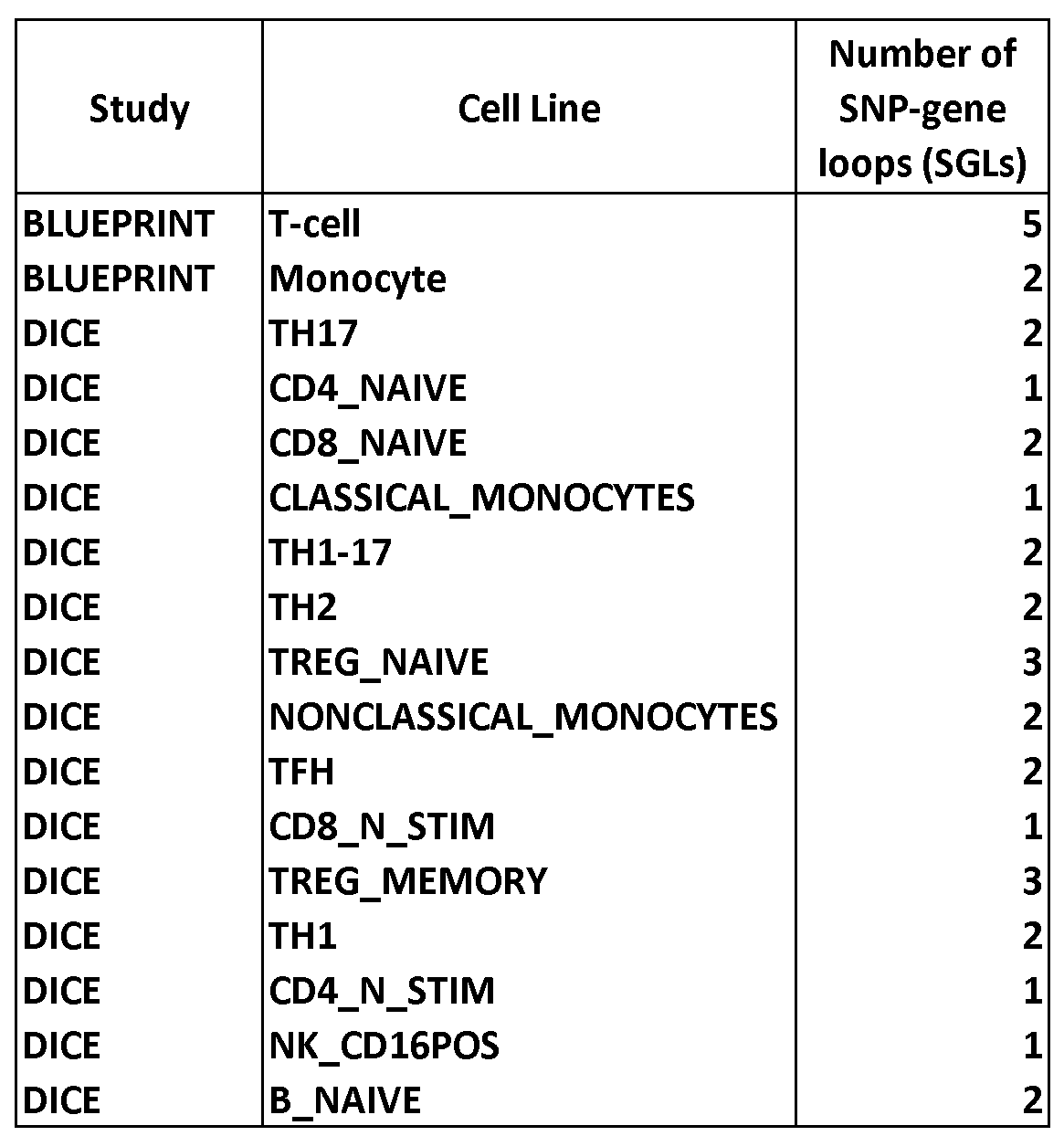
**Table 1: Summary of eQTL associations and colocalization analyses broken down by study and cell type.**

from this paper highlighted the need for further analysis of the identified GWAS loci in each relevant cell type. As part of our analyses we took the larger set of SNPs prior to fine-mapping which includes 45,994 SNPs with genome-wide significance (p-value ≤ 5e-8).

As part of our effort we utilized eQTL summary statistics from multiple different immune cell types. More specially, we included eQTLs from BLUEPRINT which includes monocytes, neutrophils and T-cells as well as eQTLs from DICE which include gene expression from naive B cells (NB), classical monocytes (CM), non-classical monocytes (NCM), T follicular helper cells (Tfh), Th1, Th2, Th17 and Th1/17 cells, Natural killer cells (NK), naive CD4 (CD4N), and CD8 (CD8N) T cells (Mu et al., 2021; Chandra et al., 2020; Schmiedel et al., 2018). BLUEPRINT reports more eGenes (genes with at least one eQTL) per cell type with a median of 7211 eGenes compared to 2104 eGenes for DICE, a difference driven mainly by sample size. However, DICE project provides a finer level of detail and, in general, higher effect size eQTLs by studying well-characterized cell subsets separately especially for T cells **(Table 1)**.

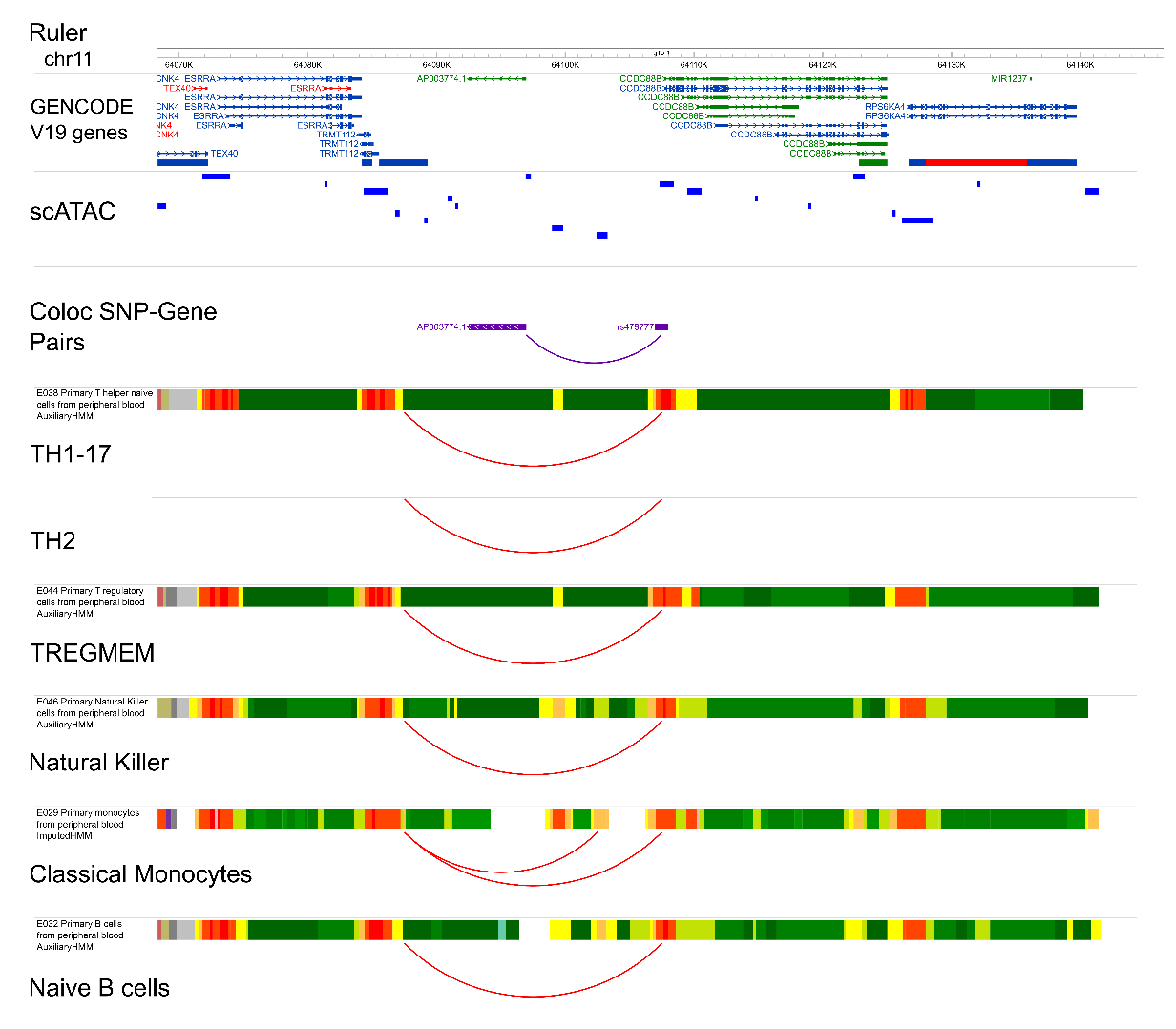
Several papers have coined the term post-GWAS to signify the need to solve the how, where and when of GWAS-SNP function (Pierce et al., 2020). In alignment with this need colocalization analyses have followed which calculate the posterior probability that a specific genetic variant explains two separate signals, in our case we analyzed the colocalization of a GWAS and eQTL signal (Wallace et al., 2021; Wallace et al., 2020; Giambartolomei et al., 2014). We performed this analysis between the (Chiou et al., 2021) GWAS summary statistics and eQTL summary statistics for each immune cell type resulting in 6 to 20 colocalized SNPs per cell type. The number of colocalized SNPs for T-cell subsets ranged from 12 for Treg memory cells to 6 in T helper 1-17 cells, suggesting subset-specific association signals.

Colocalization gave us confidence that a given SNP was potentially causal through modulating the expression of a gene linked to T1D risk or pathogenesis. Given our expertise in the 3D genome organization and mapping chromatin loops in immune cells, we hypothesized that some of these colocalized SNP-gene pairs could be connected via 3D interactions. To that effect, we integrated our recently generated HiChIP data which extracted loops enriched for the active histone mark H3K27ac. For easier reference we denote SNP-gene pairs with loops as SGLs. We found an average of 2 SGLs per cell line and combined all such SNP-gene pairs across all cell lines to establish a set of candidate genes **(Table 2)**. In total we would have 14 genes including *BACH2, UBASH3A, PTPN22, SIRPG* which have been previously associated with type 1 diabetes. Given their previous significance we focused on genes which are less known in literature including *AP003774.1*.



**Table 2: Number of SGLs per study and cell type.**

At this point we had systematically analyzed the genome and for the next half of our analyses, we relied on several genome annotations to study interesting additional parameters such as overlap of the SNPs with transcription factor binding sites and ChromHMM states which may indicate regulatory role for the SGL SNPs on their target genes. We first observed that several cell types (Tregs, Th2, Th1/17, NK, non-classical monocytes and naive B cells) contained the rs479777 SNP within one anchor and the *AP003774.1* gene within the loop body **(Fig. 1)**. By adding the cell type specific chromHMM annotations we found that these loops connected two nearby transcription start sites **(Fig. 1)**. Then, we verified the expression of this gene in the DICE database which shows that *AP003774.1* is indeed expressed at detectable levels in all immune cell types and mainly in lymphocytes **(Fig. 2a)**. We further investigated this locus by adding JASPAR motif tracks which located 2 transcription factors overlapping rs479777, SOX12 and ZNF341 **(Fig. 2b)**. By looking at the JASPAR website we found that rs479777T/C actually swaps a highly conserved C for a T nucleotide **(Fig. 2c,d)** at both TF binding sites and ablates these binding motifs when the reference allele (T) is present. Between these two TFs, ZNF341 regulates the expression of STAT3 and its deficiency has been linked to defective Th17 differentiation in humans (August, 2018). *Sox12*, on the other hand, has been shown to promote Treg differentiation in the periphery in a mouse colitis model (Tanaka et al., 2018). Given that JASPAR relies completely on motif occurrences, we also queried the ADASTRA database for allele-specific binding preferences for ChIP-seq derived transcription factor binding sites (Abramov et al., 2021). ADASTRA did not have any ChIP-seq data for ZNF341 or SOX12, however we found that HTF4, encoded by *TCF12*, shows allele-specific binding preference in GM12878 cells, which are derived from B cells. *TCF12* is a basic helix-loop-helix family transcription factor that is well expressed in T and B cells and is implicated in their early development (Liao & Wang, 2021). Lastly, we went back and utilized the scATAC-seq from (Chiou et al., 2021) where we found an ATAC-seq peak overlapping the rs479777 SNP in activated CD4 T cells, adaptive NK cells, classical monocytes, conventional dendritic cells, cytotoxic CD8 T, cytotoxic NK, megakaryocyte, memory B, memory CD8 T, naive B, naïve T, non-classical monocytes in peripheral blood as well as Tregs, CD8 T cells, macrophages and plasmacytoid dendritic cells in pancreas **(Fig. 1).**

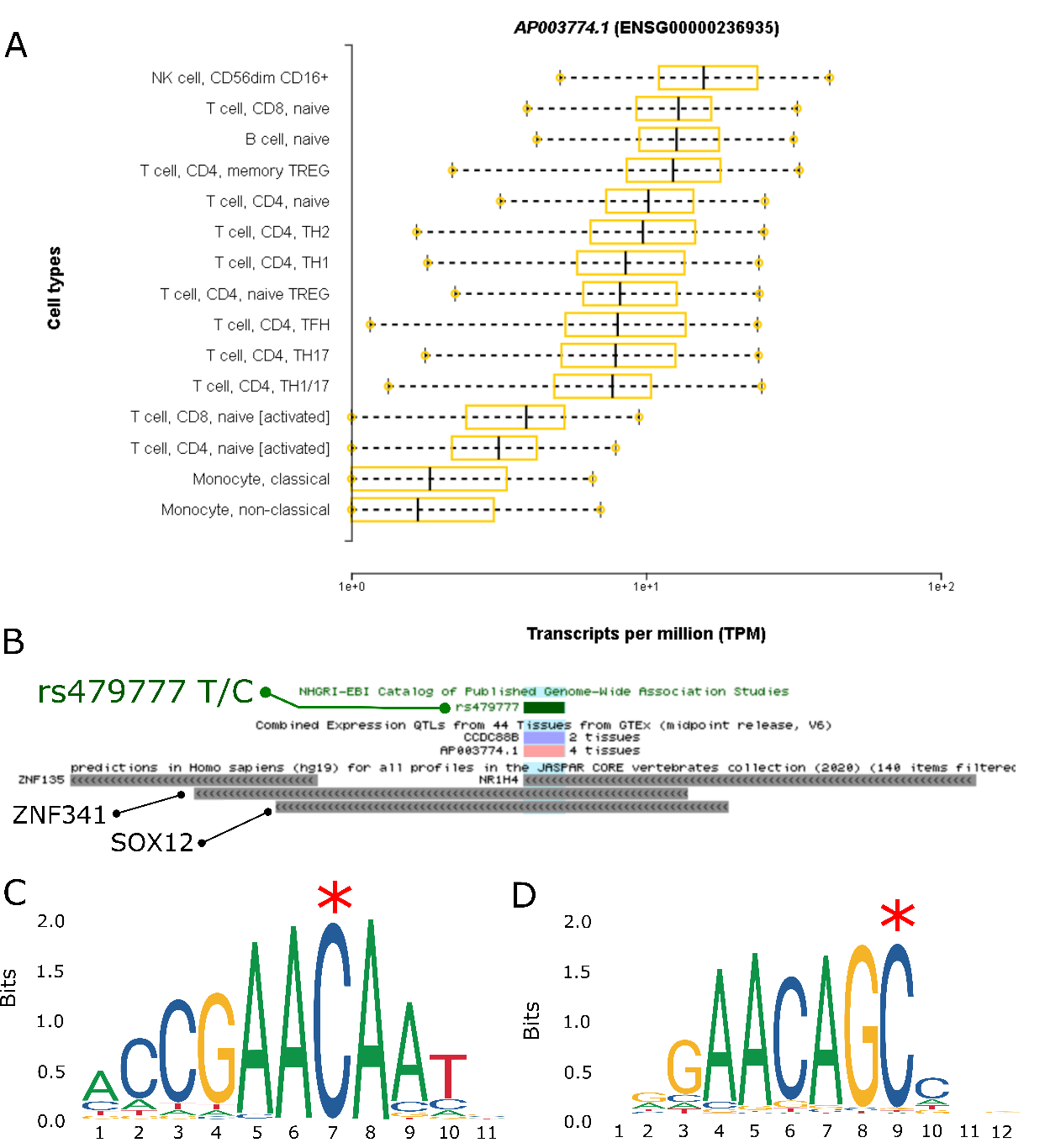


**Figure 1: Investigating the rs479444/AP003774.1 locus for loops. Visualization of the AP003774.1 locus with tracks for gene annotations (Gencode V19), scATAC-seq tracks from Chiou et al., 2021 (bright blue), colocalized SNP-gene pair (purple), and several pairs of chromHMM and HiChIP loops tracks in the corresponding cell types.**

In addition to T1D we also examined the association of rs479777 to other autoimmune diseases. We found associations to sarcoidosis (inflammation and granuloma formation in affected organs) (Fischer et al., 2012), Crohn’s disease (inflammatory bowel disease) (Crohn's disease., 2020) and alopecia areata (reversible hair loss) (Alopecia areata., 2017). Interestingly, we found that rs479777 was significant in a GWAS for thyroid preparations meaning individuals with this variant often require some form of medication to make up for an improper thyroid function (Sakaue et al., 2021). Having evaluated the importance of rs479777, we focused the remainder of this hypothesis on studying *AP003774.1.*

*AP003774.1* is a long noncoding RNA with little functional annotation which continues to mask its priority as a research topic. Initially we found that the rs479777 SNP is likely at the promoter region of *CCDC88B* **(Fig. 1)**. However, the lack of eQTL association between rs479777 and *CCDC88B* in immune cells as well as in T1D related tissue types in GTEx (<https://gtexportal.org>) led us towards *AP003774.1,* specifically due to a chromatin loop that puts the SNP and this gene’s promoter in close proximity.

There are many potential mechanisms through which lncRNAs function including regulation of gene expression in *cis* (Gil & Ulitsky, 2019) or in *trans* or through chromatin looping or post transcriptionally (Statello et al., 2021). Teimuri et al., 2018 have shown that lncRNAs in Th17 could be potential therapeutic drug targets in autoimmune diseases. Almo et al., 2018 also showed that noncoding RNAs play an important role in human CD3+ cells and a much longer list of papers describe the role the lncRNAs can have on T-cell development, differentiation, stability of pathogenic mRNAs and several other autoimmune diseases (Dieter et al., 2021; Plasek & Valadkhan 2020; Gonzalez-Moro et al., 2020; Roy & Awasthi, 2019; Gao et al., 2018; Guo et al., 2019). All of which make it very difficult to study such novel lncRNAs, however, given the strong signals from multiple different complementary analyses, we consider AP003774.1 an important target for future T1D studies.



**Figure 2: Investigating the rs479444/AP003774.1 locus for transcription factor binding sites. A) Gene expression in the DICE database for AP003774.1, B) JASPAR tracks for transcription factor binding sites. The track includes rs497777 (green), eQTL summary from GTEx (purple and pink), and JASPAR 1D tracks (grey), C) SOX12 motif and E) ZNF341 motif with red astrix indicating the location where rs479777 hits.**

### Potential Study Design

Our lab has generated Jurkat cell lines, a ixture of different T-cell subtypes, that express the KRAB-dCas9 construct. Using this CRISPR inhibition (CRISPRi) system we can target the SNP we found associated with AP003774.1 expression by designing guide RNAs that will inhibit the enhancer (open chromatin region) overlapping this SNP. We will quantify the expression level of AP003774.1 as well as other nearby genes in cells with the target gRNA vs controls to test our hypothesis that the specific enhancer region we found is controlling AP003774.1 gene expression in multiple cell types where the SNP showed colocalization.

We have previously used this model in (Chandra et al., 2021) to show that the FHIT gene is regulated by rs11130745 via a looping mechanism. Like in the AP003774.1 context, the FHIT gene is located within the loop body, in addition, this SNP overlaps an enhancer within one anchor and a promoter on the second anchor. A tiling approach was then used to silence nearby regions and resulted in lower FHIT gene expression.

We used a second recombination based CRISPR approach to change the rs1130745 site from a G/G to an A/A and again saw lower expression. By using this system we can better understand the role AP003774.1 plays in T-cells development and T1D pathology.

### **Author list, affiliations, funding acknowledgements**

Joaquin Reyna1,2,\*

Sourya Bhattacharyya1

Nikhil Rao1,4

Abhijit Chakraborty1

Ferhat Ay1,3,\*

1La Jolla Institute for Immunology, La Jolla, CA, USA.

2Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA

3Department of Pediatrics, University of California San Diego, La Jolla, CA, USA.

4Department of Bioengineering, University of California San Diego, La Jolla, CA, USA.

\* Contact: [ferhatay@lji.org](mailto:ferhatay@lji.org) & [jreyna@lji.org](mailto:jreyna@lji.org)

## B: Protocol from Protocols.io (PDF)

<https://www.protocols.io/private/B58E0187427B11ECAA1D0A58A9FEAC02>

## Section C: Table of resources used

|  |  |  |  |
| --- | --- | --- | --- |
| Name | RRID | URL | Usage |
| GWAS: Catalog of Published Genome-Wide Association Studies | SCR\_012745 | <http://www.ebi.ac.uk/gwas/> | * Downloaded T1D-GWAS summary statistics from (Chiou et al., 2021) * Downloaded scATAC-seq to investigate open chromatin within the AP003774.1 locus |
| Blueprint Epigenome | SCR\_003844 | <http://www.blueprint-epigenome.eu/> | * Downloaded eQTL summary statistics |
| JASPAR | SCR\_003030 | <http://jaspar.genereg.net/> | * Investigated transcription factor binding motifs for several SNPs |
| Roadmap Epigenomics Project | SCR\_008924 | <http://roadmapepigenomics.org/> | * Annotated the WashU Epigenome Browser with ChromHMM annotations |
| WashU Epigenome Browser | SCR\_006208 | <http://epigenomegateway.wustl.edu/> | * Visualized loops, SNPs, genes, ChromHMM and several other tracks |
| UCSC Genome Browser | SCR\_005780 | <http://genome.ucsc.edu/> | * Visualized additional tracks which were not easily accessible via the WashU Epigenome Browser |
| DICE | nan | <https://dice-database.org/> | * Downloaded eQTL summary statistics |
| FIVEx | nan | <https://fivex.sph.umich.edu/> | * Queried candidate SNP to locate possible SNPs in LD |
| ADASTRA | nan | <https://adastra.autosome.ru/zanthar> | * Investigated allele specific transcription factor binding |

## Section D: References

* Abramov et al., Landscape of allele-specific transcription factor binding in the human genome. Nat Commun. 2021. DOI: <https://doi.org/10.1038/s41467-021-23007-0>.
* Alopecia areata. Nat Rev Dis Primers 3, 17012 (2017). DOI: <https://doi.org/10.1038/nrdp.2017.12>.
* Almo et al., The role of long noncoding RNAs in human T CD3+ cells. J Immunol Sci. 2018. DOI: <https://doi.org/10.29245/2578-3009/2018/1.1109>.
* August, Who regulates whom: ZNF341 is an additional player in the STAT3/TH17 song. Sci Immunol. 2018.DOI: <https://doi.org/10.1126/sciimmunol.aat9779>.
* Chiou et al., Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. Nature. 2021. DOI: <https://doi.org/10.1038/s41586-021-03552-w>.
* Chandra et al., Promoter-interacting expression quantitative trait loci are enriched for functional genetic variants. Nat Genet. 2021. DOI: <https://doi.org/10.1038/s41588-020-00745-3>.
* Crohn's disease. *Nat Rev Dis Primers* 6, 23 (2020). DOI: <https://doi.org/10.1038/s41572-020-0166-0>.
* Dieter et al., The Impact of lncRNAs in Diabetes Mellitus: A Systematic Review and In Silico Analyses. Front Endocrinol (Lausanne). 2021. DOI:https://doi.org/10.3389/fendo.2021.602597.
* Fischer et al., A novel sarcoidosis risk locus for Europeans on chromosome 11q13.1. Am J Respir Crit Care Med. 2012. DOI: <https://doi.org/10.1164/rccm.201204-0708OC>.
* Gao et al., The Role of Long Non-coding RNAs in the Pathogenesis of RA, SLE, and SS. Front Med (Lausanne). 2018. DOI: <https://doi.org/10.3389/fmed.2018.00193>.
* Giambartolomei et al., Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014. DOI: <https://doi.org/10.1371/journal.pgen.1004383>.
* Gil & Ulitsky, Regulation of gene expression by cis-acting long non-coding RNAs. Nat Rev Genet. 2020. DOI: <https://doi.org/10.1038/s41576-019-0184-5>
* Gonzalez-Moro et al., The T1D-associated lncRNA *Lnc13* modulates human pancreatic β cell inflammation by allele-specific stabilization of *STAT1* mRNA. Proc Natl Acad Sci U S A. 2020 DOI: <https://doi.org/10.1073/pnas.1914353117>.
* Guo et al.,. Long noncoding RNA: an emerging player in diabetes and diabetic kidney disease. Clin Sci (Lond). 2019. DOI: <https://doi.org/10.1042/CS20190372>.
* Katsarou et al., Type 1 diabetes mellitus. *Nat Rev Dis Primers* 3, 17016 (2017). DOI: <https://doi.org/10.1038/nrdp.2017.16>
* Liao & Wang, Tcf12 balances the reconstitution and differentiation capacity of hematopoietic stem cell. Blood Science. 2021. DOI: <https://doi.org/10.1097/BS9.0000000000000059>.
* Imagawa et al., A proposal of three distinct subtypes of type 1 diabetes mellitus based on clinical and pathological evidence. Ann Med. 2000. DOI: <https://doi.org/10.3109/07853890008998833>.

## Mountjoy et al., An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet.* 2021. DOI: <https://doi.org/10.1038/s41588-021-00945-5>

* Mu et al., The impact of cell type and context-dependent regulatory variants on human immune traits. *Genome Biol*. 2021. DOI: <https://doi.org/10.1186/s13059-021-02334-x>
* Pierce et al., Post-GWAS knowledge gap: the how, where, and when. NPJ Parkinsons Dis. 2020. DOI: <https://doi.org/10.1038/s41531-020-00125-y>.
* Plasek & Valadkhan, lncRNAs in T lymphocytes: RNA regulation at the heart of the immune response. Am J Physiol Cell Physiol. 2021 DOI: <https://doi.org/10.1152/ajpcell.00069.2020>.
* Roy S, Awasthi A. Emerging roles of noncoding RNAs in T cell differentiation and functions in autoimmune diseases. Int Rev Immunol. 2019. DOI: <https://doi.org/10.1080/08830185.2019.1648454>.
* Sakaue et al., A cross-population atlas of genetic associations for 220 human phenotypes. Nat Genet. 2021. DOI: <https://doi.org/10.1038/s41588-021-00931-x>.
* Schmiedel et al., Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell. 2018 DOI: <https://doi.org/10.1016/j.cell.2018.10.022>..
* Skyler, Characterizing subgroups of type 1 diabetes. Diabetes. 2014. DOI: <https://doi.org/10.2337/db14-1160>.
* Statello et al., Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol. 2021. DOI: <https://doi.org/10.1038/s41580-020-00315-9>.
* Tanaka et al., Sox12 promotes T reg differentiation in the periphery during colitis. J Exp Med. 2018. DOI: <https://doi.org/10.1084/jem.20172082>.
* Teimuri et al., Integrative Analysis of lncRNAs in Th17 Cell Lineage to Discover New Potential Biomarkers and Therapeutic Targets in Autoimmune Diseases. Mol Ther Nucleic Acids. 2018. DOI: <https://doi.org/10.1016/j.omtn.2018.05.022>.
* Wallace, A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genet. 2021. DOI: <https://doi.org/10.1371/journal.pgen.1009440>.
* Wallace, Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLoS Genet. 2020. DOI: <https://doi.org/10.1371/journal.pgen.1008720>.

## Section E: Conflict of Interest

No conflict to declare.

## Other - Thorough Description of CRISPR

**CRISPRi targeting of enhancers using KRAB-dCas9.**

At 3 d before CRISPRi assay, KRAB-dCas9-expressing cells (mCherry positive) were sorted again to ensure that all cells expressed KRAB-dCas9. Then, 44 μM crRNA and tracrRNA (from IDT) duplex specific for each target was prepared by mixing the two RNA oligos in equimolar concentrations in a sterile microcentrifuge and heating at 95 °C followed by cooling at room temperature. Cells were transfected with 3.6 μM crRNA and tracrRNA duplex specific for the target enhancer or for the nontargeting region (from IDT) using the Neon Transfection System (Thermo Fisher Scientific) according to the manufacturer’s protocol (settings: 1,600 V, 10 ms, 3 pulses); see Supplementary Table 1e for crRNA sequences. Fresh medium (as described above) was then added and cells were maintained for 48 h. After 48 h cells were collected and knockdown efficiency for the target gene was analyzed by real-time PCR for transcript levels.

**Inhibition of targeted region using dCas9.**

At 3 d before CRISPRi assay, dCas9-expressing cells (EGFP positive) were sorted to ensure that all cells expressed dCas9. Then, 44 μM crRNA and tracrRNA (IDT) duplex specific for each target was prepared by mixing the two RNA oligonucleotides in equimolar concentrations in a sterile microcentrifuge, and heating at 95 °C followed by cooling at room temperature. RNP complexes were prepared by incubating dCas9 and crRNA–tracrRNA duplex specific for target regions for 20 min at room temperature. Cells were transfected with RNP complex using the Neon Transfection System (Thermo Fisher Scientific) according to the manufacturer’s protocol (settings: 1,600 V, 10 ms, 3 pulses); see Supplementary Table 1e for crRNA sequences. Fresh medium (as described above) was then added and cells were maintained for 48 h. After 48 h cells were collected and knockdown efficiency was analyzed by real-time PCR for transcript levels of target genes.

**Genome editing in activated CD4+ T cells.**

Naïve CD4+ T cells purified using a

magnetic-activated cell-sorting column were resuspended at a concentration of 2.5 × 105

per ml in 1 ml of prewarmed IMDM medium, supplemented with 5%

(vol/vol) heat-inactivated FBS and 2% (vol/vol) human AB serum (CellGro) and activated ex vivo with Dynabeads Human T-Activator CD3/CD28 (Thermo Fisher Scientific) at a bead-to-cell ratio of 2:1 for the indicated duration at 37 °C. Three different activation conditions were used to determine the activation condition that leads to efficient HDR efficiency as well as reproducibility of effect. After activation, Dynabeads were removed and cells were cultured in fresh medium with IL-2 for the indicated number of days. RNP complexes were prepared by incubating dCas9 with either single guide RNA (sgRNA) alone or crRNA– tracrRNA duplex specific for the target region for 20 min at room temperature. Then, 2.0 × 105

activated CD4+ T cells were washed two times with PBS before

resuspension in 8 μl of buffer T, and 80 pmol of HDR template was added to the cell suspension along with RNP complex. Cells were transfected using the Neon Transfection System (Thermo Fisher Scientific) according to the manufacturer’s protocol (settings: 1,600 V, 10 ms, 3 pulses); see Supplementary Table 1e for sgRNA and crRNA sequences. Fresh medium (as described above) was then added and cells were maintained for the indicated duration. Cells were collected, and DNA and RNA were isolated for downstream analysis. Genome editing was verified by Sanger sequencing and effects on gene expression assessed by real-time PCR for transcript levels.