Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Macrophage lncRNA SIMALR at TNFAIP3 risk locus encodes HLA class II-binding micropeptides in type 1 diabetes

Reza Anvaripour

rez.anvaripour@gmail.com

independent Researcher https://orcid.org/0000-0003-2212-7646

Research Article

Keywords: Type 1 diabetes, Long non-coding RNA, TNFAIP3, HLA class II, DQ8, Micropeptide, Macrophage, Dark genome, Immunopeptidomics, Autoantigen

Posted Date: October 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11454023/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Macrophage lncRNA SIMALR at TNFAIP3 risk locus encodes HLA class II-binding micropeptides in type 1 diabetes

Reza Anvaripour

2026-10-21

Reza Anvaripour

Independent Researcher, Tehran, Iran

ORCID number: 0000-0003-2212-7646

Corresponding author: Reza Anvaripour, Independent Researcher, Tehran, Iran. rez.anvaripour@gmail.com

Abstract

BACKGROUND

More than sixty non-HLA risk loci are associated with type 1 diabetes (T1D), but most map to non-coding sequence and their causal transcripts remain unresolved. The 6q23 locus harboring TNFAIP3 carries one of the strongest non-HLA T1D signals, tagged by rs6920220, yet neither the causal transcript nor the cell type through which it acts has been identified. The same variant is associated with rheumatoid arthritis and systemic lupus erythematosus. We hypothesized that T1D-relevant cytokines activate a dark genome non-coding program in primary human islets that nominates a transcript at a known T1D risk locus.

AIM

To identify a T1D-specific macrophage lncRNA at the TNFAIP3 risk locus and characterize its HLA class II-binding micropeptides.

METHODS

We dissected the non-coding response of primary human islets to IL-1beta, IFN-gamma, and TNF-alpha (n = 19 paired donors), compared against ER-stressed and type 2 diabetic islets. The 6q23 locus was mapped using islet capture Hi-C. HLA class II binding of predicted micropeptides was tested by MixMHC2pred and NetMHCIIpan 4.3, and cross-referenced against 1,964 known T1D autoantigen epitopes. Public single-cell blood datasets (376,996 PBMCs; 143,974 monocytes) were screened for SIMALR expression.

RESULTS

Among 54 dark genome transcripts induced more than 4-fold by cytokines, SIMALR (LINC02528, ENSG00000226004) mapped 59 kb from rs6920220. SIMALR was induced in 19 of 19 donors (mean +5.06 log2FC, padj 2.9e-45), was absent in type 2 diabetic islets, and received a moderate T1D effector score (3.0) with no T2D evidence (1.0). ENCODE ChIP-seq and motif scanning place SIMALR downstream of NF-kB and STAT3. SIMALR encodes 23 predicted strong HLA class II binders; four DQ8 and one DRB1*03:01 binder were confirmed by an independent algorithm with identical binding cores. No similarity was found with 1,964 known T1D autoantigen epitopes. SIMALR is detectable in human blood monocytes (727 of 376,996 cells).

CONCLUSION

SIMALR is a T1D-specific macrophage lncRNA at the TNFAIP3 locus whose micropeptides are predicted to bind DQ8 and DRB1*03:01, representing a candidate non-beta-cell antigen in T1D.

Key words: Type 1 diabetes; Long non-coding RNA; TNFAIP3; HLA class II; DQ8; Micropeptide; Macrophage; Dark genome; Immunopeptidomics; Autoantigen

Core tip: We dissected the non-coding response of primary human islets to T1D-relevant cytokines and identified SIMALR, a macrophage-restricted lncRNA 59 kb from the TNFAIP3 T1D risk variant rs6920220. SIMALR is induced in all 19 donors tested, is absent in type 2 diabetes, is regulated by NF-kB and STAT3, and encodes 23 predicted HLA class II-binding micropeptides with no similarity to any known T1D autoantigen. These data nominate SIMALR as a candidate non-beta-cell antigen at the 6q23 T1D risk locus.

Introduction

Type 1 diabetes (T1D) is an autoimmune disease caused by T-cell-mediated destruction of insulin-producing beta-cells. Genome-wide association studies have identified more than sixty risk loci outside the HLA region[1,2], but the majority of risk variants fall in non-coding sequence and the causal transcripts remain unresolved. The 6q23 locus harboring TNFAIP3, encoding the NF-kB negative regulator A20, carries one of the strongest non-HLA T1D signals, tagged by the lead variant rs6920220[1]. The same variant is associated with rheumatoid arthritis and systemic lupus erythematosus, defining 6q23 as a shared autoimmune susceptibility region. Despite two decades of genetic evidence, neither the causal transcript at 6q23 nor the cell type through which it acts has been resolved.

The rs6920220 eQTL on TNFAIP3 is context-dependent: allele-specific effects are detectable in disease cases but not resting controls[3], and reporter assays show that the risk allele represses TNFAIP3 transcription. This has motivated a search for additional transcripts at 6q23, including long non-coding RNAs (lncRNAs), that might explain the disease-specific effect. Capture Hi-C in immune cells has identified physical contacts between the risk region and multiple downstream targets, including IL20RA, TNFAIP3, and the lncRNA RP11-10J5.1 (now HGNC-designated SIMALR)[4]. The interaction between the

risk region and SIMALR is not allele-specific, and whether SIMALR is a functional element at this locus has not been determined.

Long non-coding RNAs are emerging as cell-type-specific regulators of immune gene programs, frequently non-conserved, stimulus-induced, and concentrated at disease-associated loci. A subset encode short open reading frames (micropeptides) that can be processed and presented on MHC molecules, extending the coding potential of the genome in cell types where the host lncRNA is active. Whether a lncRNA at a T1D risk locus produces such peptides, and whether they could be presented on T1D-associated HLA class II alleles, has not been tested.

We asked whether T1D-relevant cytokines activate a dark genome non-coding program in primary human islets, and whether this program nominated a transcript at a known T1D risk locus whose products are predicted to bind T1D-associated HLA class II. The datasets analyzed in this study are summarized in Table 1.

Table 1 Public datasets analyzed in this study

Dataset	Design	Role
GSE251911 CYT	Primary islets, IL-1b+IFNg+TNFa vs untreated	Primary cytokine response
GSE251911 ERS	Primary islets, thapsigargin vs DMSO	ER stress control
GSE235683	EndoC-bH1 single-cytokine panel	Cell line validation
GSE159984	Islets, 58 ND + 27 T2D	T2D specificity
GSE212381	HMDM, control ASO vs SIMALR ASO + LPS/IFNg	Functional knockdown re-analysis
GSE316337	PBMC scRNA-seq, 27 T1D + 27 HC	Blood expression
GSE221297	Monocyte scRNA-seq, 6 T1D + 3 HC + 3 pre-diabetic	Monocyte expression by stage
MSV000093843	Islet immunopeptidomics	Peptide presentation search
Zenodo 13362658	HPAP islets, cytokine vs mock scRNA-seq	Beta-cell single-cell response
Zenodo 15611730	PanKbase islet atlas v3.4	Islet cell-type markers

Materials and Methods

Data sources

Public datasets analyzed in this study are listed in Table 1. Bulk RNA-seq included GSE251911 (human islets, cytokine and ER stress arms; n = 19 and n = 10 paired donors), GSE235683 (EndoC-bH1 single-cytokine panel), GSE159984 (58 non-diabetic and 27 type 2 diabetic islets), GSE212381 (HMDM control ASO versus SIMALR ASO with LPS/IFN-gamma)[5], GSE316337 (PBMC single-cell, 27 T1D and 27 healthy donors), GSE221297 (monocyte single-cell, 6 T1D, 3 healthy, 3 pre-diabetic), and MSV000093843 (human islet immunopeptidomics)[6]. Reference data included GENCODE v46 annotation, UCSC phyloP 100-way conservation, GTEx v8 eQTL summary statistics[7], PanKbase HuGE integrative scores, ENCODE ChIP-seq via UCSC API, IEDB autoantigen epitopes[8], Bgee/Sugi Atlas monocyte expression[9], and islet capture Hi-C processed files from Miguel-Escalada et al[10].

Differential expression analysis

Bulk RNA-seq counts were analyzed with DESeq2 v1.42 in R 4.6[11]. The paired design was used for GSE251911 (donor as covariate). Significance was called at Benjamini-Hochberg padj < 0.05. Dark genome filtering required GENCODE biotype in {lncRNA, pseudogene, TEC} and absolute log2FC greater than 2 versus untreated. In 12 of 19 donors, SIMALR was undetected in the untreated condition; log2 fold changes were computed with a pseudocount of 1 read added to both conditions to avoid undefined values.

Motif analysis

Strand-aware promoter sequences were extracted from hg38 (2 kb and 10 kb windows upstream of the annotated transcription start site), together with gene body and intergenic windows. Motif scanning used Biostrings matchPWM[12] with hardcoded position weight matrices for TP53, RELA, RELB, NFKB1, STAT1, STAT3, and IRF family members. A score threshold of 0.80 relative to the matrix maximum was applied. Motifs per kb were computed for each region.

HLA class II binding prediction

All open reading frames of SIMALR were translated in three forward and three reverse frames. Unique 15-mer peptides were scored against DRB1*03:01, DRB1*04:01, DQA1*05:01/DQB1*02:01 (DQ2.5), and DQA1*03:01/DQB1*03:02 (DQ8) using MixMHC2pred v1.1 with the no-context flag[13]. Strong binders were defined as rank below 2. Top candidates were re-scored using NetMHCIIpan 4.3[14]; rank_EL below 2 and IC50 below 500 nM were treated as confirming a binder.

Autoantigen cross-reference

1,964 unique T-cell epitopes from INS, GAD65, IA-2, ZnT8, IGRP, CHGA, ICA1, and GAD1 were downloaded from IEDB[8] via the query API in 1,000-peptide batches. For each SIMALR binding core we computed the set of all 7-, 8-, and 9-mers and intersected with the corresponding k-mer set of each reference epitope. Exact full-length matches were also tested.

Single-cell processing

10x Genomics matrices were loaded into Seurat v5[15,16] in R 4.6. Quality filtering retained cells with 200 to 5,000 genes and mitochondrial content below 10 percent. Data were LogNormalized (scale factor 10,000), and 2,000 highly variable genes were selected. PCA used 30 components; neighbor graph and clustering (Louvain, resolution 0.5) were computed on the first 20 PCs; UMAP was run with default parameters. Monocyte clusters were identified by AddModuleScore against a canonical marker panel (CD14, FCGR3A, FCN1, S100A8, S100A9, LYZ, VCAN, MS4A7, CD68, CD163); clusters with mean module score above 0.5 were retained. Statistical testing was performed at both the cell level (Wilcoxon rank-sum, Fisher exact) and the donor level (Wilcoxon rank-sum on per-donor detection rate); results are reported at both levels with the caveat that cell-level tests treat each cell as independent.

Chromatin and locus annotation

Islet promoter-capture Hi-C interaction calls and TAD coordinates were taken from the Supplementary Data Sets of Miguel-Escalada et al[10]. Islet regulatory element annotations (active promoters, active enhancers I-III, strong CTCF, inactive open chromatin, inactive enhancers) were taken from the islet regulome in the same study. Immune-cell capture Hi-C interaction evidence for the 6q23 risk region was taken from Martin et al[4]. GTEx v8 cis-eQTL summary statistics were queried via the GTEx Portal API[7].

Statistical framework

Unless otherwise noted, tests are non-parametric (Wilcoxon rank-sum, Kruskal-Wallis, Fisher exact). Multiple-test correction is Benjamini-Hochberg. Permutation tests used 10,000 iterations. The expression-matched background for dark genome enrichment used 1,000 random lncRNA sets matched for mean expression.

Results

Cytokine-treated islets activate an immunogenic DNA damage response

Primary human islets cultured for 48 h with IL-1beta, IFN-gamma, and TNF-alpha (n = 19 paired donors) showed a large-scale transcriptional response: 14,048 of 21,172 genes

were differentially expressed at padj below 0.05 (Figure 1). The response included activation of the cGAS-STING axis (CGAS +1.29 log2FC, STING1 +0.67, TBK1 +0.72, IRF3 +0.63), cytosolic RNA sensors (DDX58 +3.03, IFIH1 +2.87), and p53 executioner genes (BBC3 +1.59, CDKN1A +0.74). Bulk SASP transcripts were also elevated (IL6 +3.41, CXCL1 +3.52, MMP3 +2.24), though single-cell analysis showed that the SASP signal originates predominantly from activated stellate cells rather than beta-cells.

To determine whether the immunogenic DDR was cytokine-specific or a general stress response, we analyzed the parallel ER stress arm of the same dataset (thapsigargin versus DMSO; n = 10 paired donors). ER stress did not activate cGAS/STING (CGAS +0.11, STING1 +0.09, both padj above 0.5) and produced a distinct transcriptional signature enriched for the unfolded protein response rather than interferon or DDR genes.

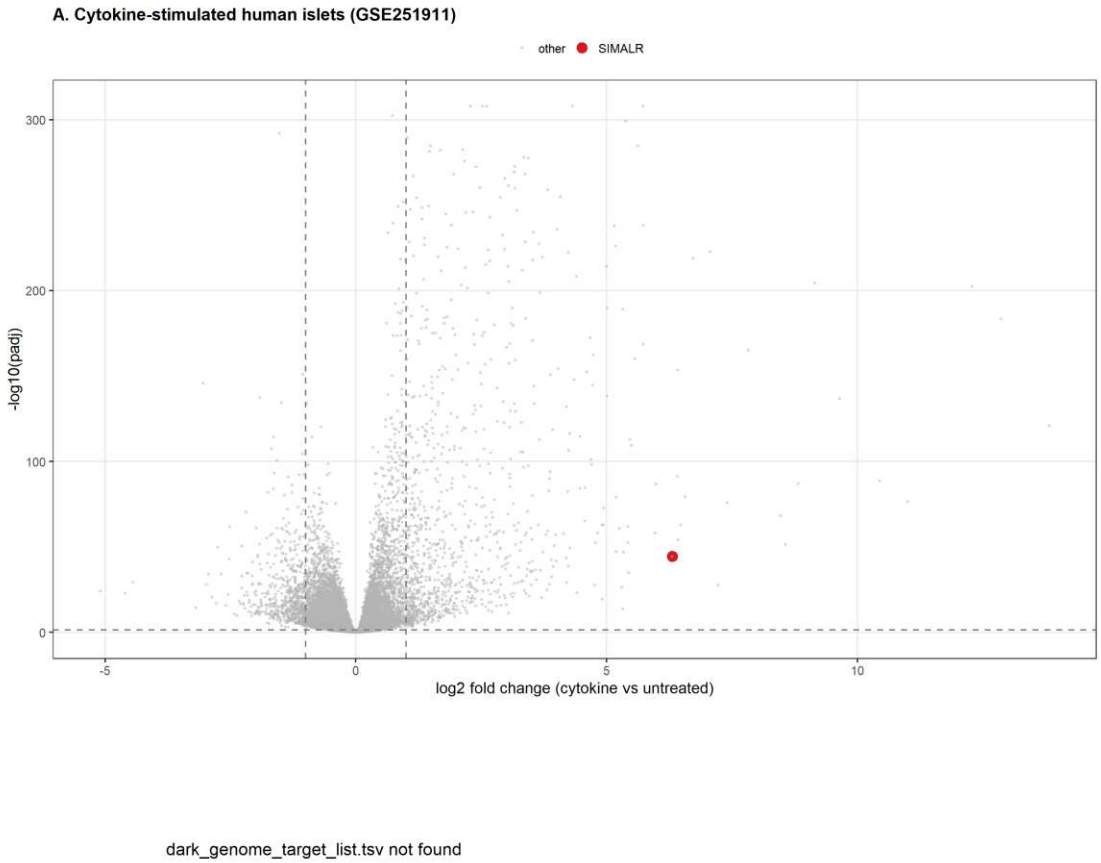


Figure 1 The dark genome response of primary human islets to T1D-relevant cytokines. A: Volcano plot of DESeq2 results; B: Top dark genome targets ranked by log2FC; C: Selected DDR, NF-kB, and SASP markers.

Fifty-four dark genome transcripts are cytokine-induced

We focused on transcripts annotated as lncRNA, pseudogene, or TEC by GENCODE v46 and required absolute log2FC above 2 (4-fold) in the cytokine arm. Fifty-four dark genome targets met this threshold, of which 48 were non-coding (41 lncRNAs, 5 pseudogenes, 2 TECs). The most strongly induced targets are listed in Table 2, led by AC093063.1 (+11.0 log2FC), AC007991.2 (+10.4), and AL157871.4 (+6.6).

Table 2 Top 15 dark genome targets induced by T1D-relevant cytokines in primary human islets

Target	log2FC	Neighbor
AC093063.1	11.00	GBP1-7 cluster
AC007991.2	10.40	IDO1
AL157871.4	6.60	interferon cluster
SIMALR (LINC02528)	6.31	TNFAIP3 (59 kb)
AC099063.4	5.90	GBP cluster
AP001610.2	5.40	MX1/MX2
AL645939.5	5.10	TRIM40/31/15/10, HLA-F
AC007728.2	4.90	NOD2
AC004477.1	4.70	-
LINC01257	4.60	-
AC011445.2	4.50	-
AC104667.1	4.40	-
LINC01934	4.20	-
AC018523.1	4.10	-
MIR3145	4.00	-

Dark genome targets were not distributed randomly: they clustered at immune gene loci. AC099063.4 lies adjacent to the GBP1-7 cluster, AC007991.2 near IDO1, AP001610.2 near MX1/MX2, and AL645939.5 near TRIM40/31/15/10, UBD, and HLA-F (Table 2). 109 of 526 coding neighbors of dark genome targets were themselves cytokine-induced. An expression-matched enrichment test showed that dark genome targets were 1.56-fold more likely than expression-matched lncRNAs to lie near a cytokine-induced coding gene ($p < 0.0005$ by permutation; null maximum 0.1811, observed 0.204). This enrichment survived expression matching, arguing against a simple co-expression artefact.

SIMALR is induced in all 19 donors and is T1D-specific

Among the 54 induced dark genome targets, SIMALR (LINC02528, ENSG00000226004, HGNC:53554) mapped to chr6:137,943,079-137,957,657 (hg38), 59 kb from rs6920220

(Figure 2). SIMALR is not conserved (mean phyloP 100-way: promoter +0.03, gene body minus 0.03, flank minus 0.10), a characteristic shared with most immune lncRNAs.

SIMALR was induced by cytokines in every donor in the paired design (19 of 19; mean +5.06 log2FC; range 3.36 to 6.50; overall padj 2.9e-45). The per-donor response is given in Table 3 and shown in Figure 3. The response was consistent in magnitude across donors, arguing against donor-specific artefacts.

Table 3 Per-donor SIMALR induction in cytokine-treated primary human islets (mean +5.06 log2FC, range 3.36 to 6.50, n = 19 paired donors)

donor	control	treated	log2FC
81	0.0000000	89.72309	6.503398
146	0.0000000	64.89510	6.042099
95	0.0000000	61.21074	5.959092
149	0.0000000	55.53410	5.821049
144	0.0000000	49.28473	5.652048
103	0.0000000	44.99777	5.523492
92	1.2578139	101.14645	5.499569
147	1.5741592	110.25855	5.433671
100	0.9919381	79.95234	5.344828
94	0.0000000	39.15709	5.327583
93	0.0000000	37.73296	5.275490
99	0.0000000	37.66993	5.273140
82	0.0000000	31.03079	5.001387
148	2.1035628	71.95145	4.554939
97	0.0000000	19.19653	4.336036
96	0.0000000	14.78608	3.980581
75	1.1319593	25.70534	3.646877
78	2.3933997	40.16913	3.600760
86	4.9211516	59.62036	3.355853

In contrast, SIMALR was not detected in bulk islets from 58 non-diabetic or 27 type 2 diabetic donors (GSE159984; Figure 3). TNFAIP3 itself was unchanged in T2D islets (log2FC +0.04, padj 0.95), as were CXCL10 (log2FC +0.29, padj 0.80) and TNIP1/2/3. The 6q23/TNFAIP3 axis, at least as measured in bulk islet tissue, is dormant in T2D.

Independent integration by PanKbase HuGE assigns SIMALR a moderate T1D effector score (3.0) with no evidence for T2D (1.0), and equivalent scores for rheumatoid arthritis, systemic lupus erythematosus, and monocyte, neutrophil, and lymphocyte counts (Figure

2). The autoimmune/myeloid cluster co-occurring with T1D is consistent with a locus acting through the myeloid compartment in multiple inflammatory diseases.

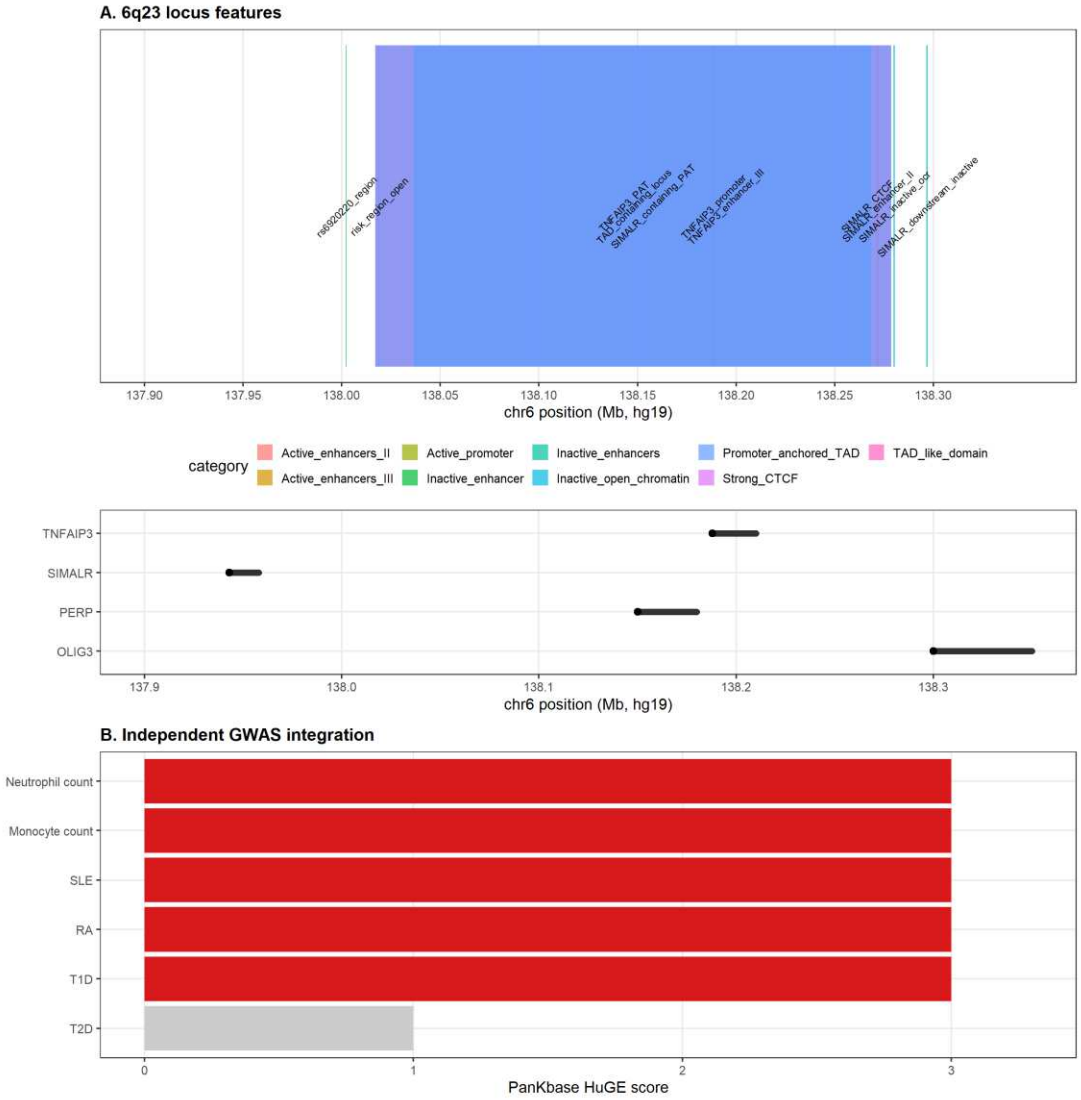


Figure 2 Locus architecture and independent genetic integration at 6q23. A: Regulatory features within 500 kb of SIMALR; B: Gene models; C: PanKbase HuGE scores for T1D, T2D, and immune traits.

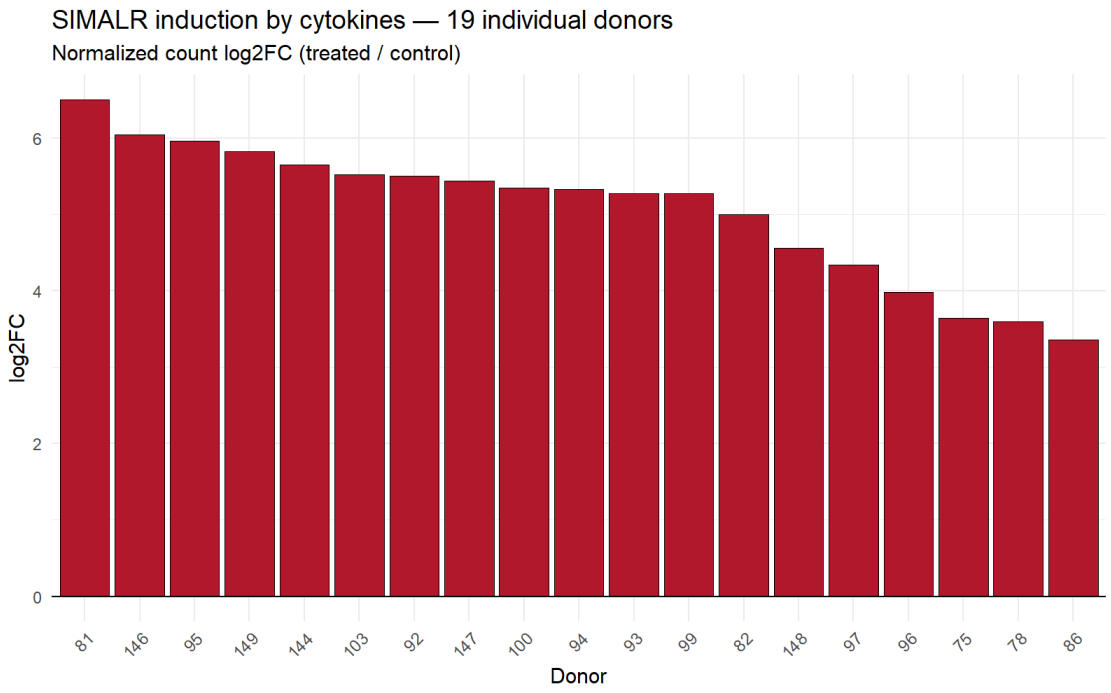


Figure 3 SIMALR induction in paired donors and absence in type 2 diabetes. A: Per-donor log2FC in cytokine-treated primary islets; B: SIMALR not detected in bulk T2D or non-diabetic islets.

SIMALR, TNFAIP3, and rs6920220 lie in a single islet TAD

Promoter-capture Hi-C in human islets[10] places SIMALR, TNFAIP3, and rs6920220 within a single topologically associating domain (chr6:138,017,057-138,278,582, hg19). The TNFAIP3 promoter-anchored interaction domain spans chr6:138,017,057-138,268,650, encompassing both the risk variant and SIMALR. At the level of islet regulatory annotation, the risk variant lies in an enhancer that is inactive in resting islets (chr6:138,002,013-138,002,503), while SIMALR is flanked by a CTCF boundary at its 3-prime end (chr6:138,265,921-138,266,533) and an active enhancer II immediately downstream (chr6:138,271,019-138,272,255) (Figure 2). The CTCF boundary suggests an insulated regulatory apparatus for SIMALR within the larger 6q23 TAD.

The SIMALR promoter was not directly baited in the islet pcHi-C panel; the panel was designed around T2D-associated promoters. The only direct published evidence of physical contact between the risk region and SIMALR comes from immune-cell pcHi-C[4], where the interaction is present but not allele-specific at the lead SNP. In GTEx v8, rs6920220 is not a significant cis-eQTL for either TNFAIP3 or SIMALR; SIMALR has 33 cis-eQTLs, all in testis, and none in whole blood at rest. This is consistent with a cytokine-inducible rather than constitutive regulatory mechanism, and with the context-dependence of the TNFAIP3 eQTL previously reported in disease cases[3].

SIMALR is regulated by NF-kB and STAT3

ENCODE ChIP-seq clusters at the SIMALR locus identified binding sites for the NF-kB family member RELB (2 sites) and STAT3 (2 sites), together with BATF, JUNB, TBX21, IKZF1, EBF1, EP300, CTCF, and RAD21. TP53 was not present. Direct position-weight-matrix scanning of the locus confirmed the pattern: in a 10 kb promoter window, STAT3 was represented at 25 sites (2.5 per kb), RELB at 23 (2.3 per kb), and STAT1 at 37 (3.7 per kb), including two near-perfect STAT1 GAS elements (PWM score 0.948 and 0.931). Motif density was uniform across promoter, gene body, and the 57 kb intergenic region between SIMALR and TNFAIP3.

In primary islets, SIMALR expression was anti-correlated with the canonical NF-kB axis: rho(SIMALR, TNFAIP3) = minus 0.68 (rank 21,079 of 21,172), with NFKBIA minus 0.74, RELA minus 0.74, NFKB2 minus 0.73, CD40 minus 0.61, and NFKB1 minus 0.53. Positively correlated genes were neuronal and structural (SYNE1, ZFP82).

To test whether SIMALR loss is functionally consequential, we re-analyzed the public HMDM SIMALR ASO knockdown dataset (GSE212381)[5]. SIMALR was knocked down with the expected directionality (log2FC minus 0.85) but at trend level only (q = 0.999), so downstream effects must be interpreted cautiously. Under this caveat, canonical NF-kB genes (TNFAIP3, NFKBIA, NFKB1/2, RELA/B, CD40, TNF, IL6, CXCL10) were unchanged (all q = 0.999), and the interferon program was intact. A small set of inflammatory chemokines were significantly reduced (CXCL1 minus 1.10, CXCL3 minus 0.97, IL12A minus 1.51, CCL1 minus 1.64, IL36G minus 1.08; all q below 0.05), which is the opposite direction from a simple suppression model. A plausible interpretation, consistent with the HGNC name for SIMALR and with its reported role in macrophage survival[5], is that SIMALR supports survival of activated macrophages; partial knockdown reduces the number of live activated cells and therefore lowers bulk chemokine output, without changing per-cell NF-kB tone. This hypothesis is not tested here and is stated as a model for future work.

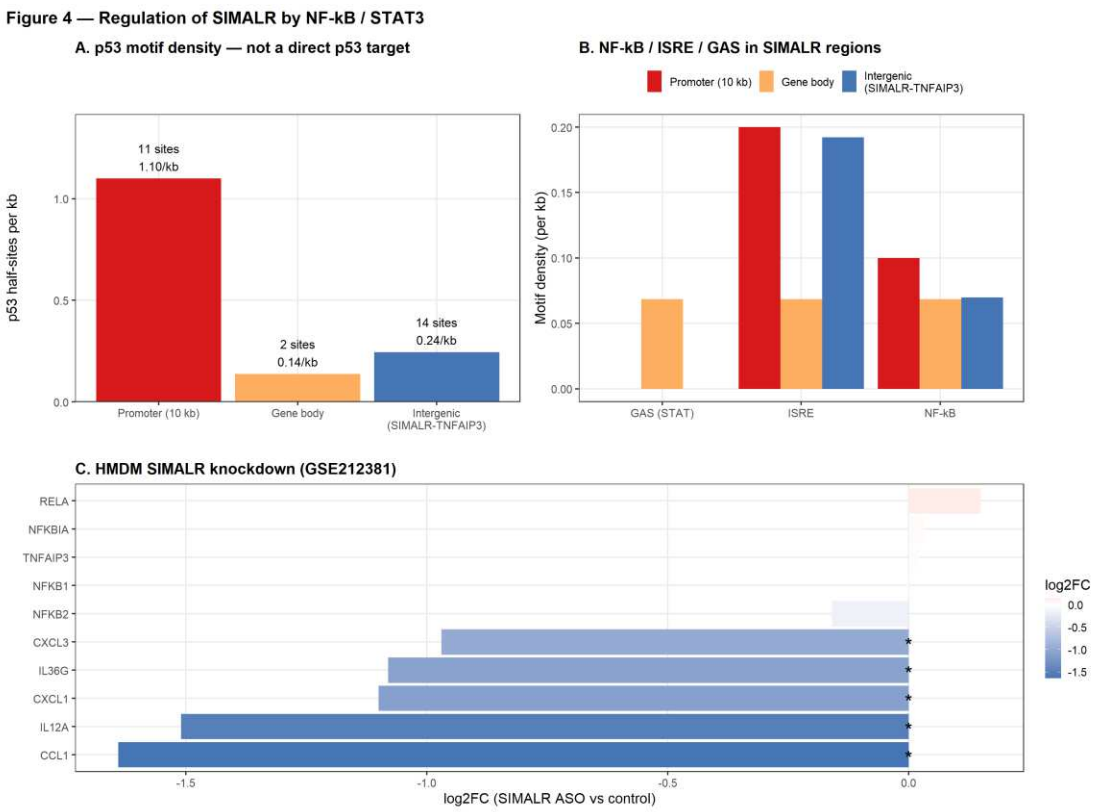


Figure 4 Regulation of SIMALR. A: p53 motif density across promoter, gene body, and intergenic regions; B: NF-kB, ISRE, and GAS motif density; C: Chemokine changes after partial SIMALR knockdown in human macrophages.

SIMALR encodes HLA class II-binding micropeptides

We scanned all open reading frames of SIMALR (three forward and three reverse frames) and extracted 738 unique 15-mer peptides. Peptides were scored against the T1D-associated class II alleles DRB1*03:01, DRB1*04:01, DQA1*05:01/DQB1*02:01 (DQ2.5), and DQA1*03:01/DQB1*03:02 (DQ8) using MixMHC2pred[13]. Twenty-three peptides were strong binders (rank below 2), comprising 14 unique binding cores; the full list is presented in Table 4.

Table 4 SIMALR micropeptides predicted as strong HLA class II binders by MixMHC2pred (rank below 2)

Peptide	BestAllele	%Rank_best	Core_best	%Rank_DRB1_03_01	%Rank_DRB1_04_01	%Rank_DQA1_05_01__DQB1_02_01	%Rank_DQA1_03_01__DQB1_03_02
EHKDGQVQAKGERPG	DQA1_03_01__DQB1_03_02	0.146	DGQVQAKG	69.100	78.700	71.000	0.146

Peptide	BestAllele	%Rank_best	Core_best	%Rank_DRB1_03_01	%Rank_DRB1_04_01	%Rank_DQA1_05_01__DQB1_02_01	%Rank_DQA1_03_01__DQB1_03_02
CEHKDGQVQAKGERP	DQA1_03_01__DQB1_03_02	0.205	DGQVQAKGE	48.200	79.200	83.400	0.205
MRYWSDAGAGIWTIT	DQA1_05_01__DQB1_02_01	0.414	WSDAGAGIW	40.800	17.300	0.414	23.100
HGSDFATKMSYENPA	DQA1_03_01__DQB1_03_02	0.463	DFATKMSYE	58.100	69.300	69.700	0.463
SMRYWSDAGAGIWTI	DQA1_05_01__DQB1_02_01	0.539	WSDAGAGIW	24.600	8.490	0.539	58.000
MKRAGGATAWKDSMC	DQA1_03_01__DQB1_03_02	0.553	AGGATAWKD	83.500	95.000	75.200	0.553
LHGSDFATKMSYENP	DQA1_03_01__DQB1_03_02	0.608	DFATKMSYE	43.400	82.000	82.400	0.608
ITFSGGATAWKDSMC	DQA1_03_01__DQB1_03_02	0.718	SGGATAWKD	89.800	27.300	65.200	0.718
HRNLRRDNQISICKP	DRB1_03_01	0.785	LRRDNQISI	0.785	28.800	29.400	95.500

Peptide	BestAllele	%Rank_best	Core_best	%Rank_DRB1_03_01	%Rank_DRB1_04_01	%Rank_DQA1_05_01__DQB1_02_01	%Rank_DQA1_03_01__DQB1_03_02
MSFFIFQNLASGLDR	DRB1_04_01	0.849	FFIFQNLAS	10.900	0.849	20.000	37.600
SITFSGGATAWKDSM	DQA1_03_01__DQB1_03_02	0.968	SGGATAWKD	37.200	15.500	37.800	0.968
QHRNLRRDNQISICK	DRB1_03_01	1.020	LRRDNQISI	1.020	35.800	36.400	92.900
LLYIQQEASNVEVIR	DRB1_04_01	1.240	IQQEASNVE	4.230	1.240	5.310	7.270
ASCWGYTPTYRETSK	DQA1_03_01__DQB1_03_02	1.410	WGYTPTYRE	44.300	4.640	25.700	1.410
WTFIHFAAEILVCLI	DQA1_05_01__DQB1_02_01	1.440	IHFAAEILV	8.180	28.500	1.440	34.400
HLLYIQQEASNVEVI	DRB1_04_01	1.550	IQQEASNVE	5.160	1.550	6.450	8.960
LASCWGYTPTYRETS	DQA1_03_01__DQB1_03_02	1.610	WGYTPTYRE	54.500	5.670	31.500	1.610
HKDGQVQAKGERPG	DQA1_03_01__DQB1_03_02	1.740	DGQVQA	56.800	93.400	86.500	1.740

Peptide	BestAllele	%Rank_best	Core_best	%Rank_DRB1_03_01	%Rank_DRB1_04_01	%Rank_DQA1_05_01__DQB1_02_01	%Rank_DQA1_03_01__DQB1_03_02
T			KGE				
MWTFIHFAAEILVCL	DQA1_05_01__DQB1_02_01	1.740	IHFAAEILV	10.200	17.200	1.740	26.400
IRSMRYWSDAGAGIW	DRB1_04_01	1.760	MRYWSDAGA	52.700	1.760	16.800	29.800
GVEIKILETEDFKCH	DQA1_05_01__DQB1_02_01	1.800	IKILETEDF	16.700	12.000	1.800	41.000
ILLASILYPEKEMAF	DQA1_03_01__DQB1_03_02	1.810	ASILYPEKE	31.200	85.000	12.600	1.810
LLFIQNATRIRLIRV	DRB1_03_01	1.820	IQNATRIRL	1.820	33.100	34.400	58.600

Four DQ8 binders and one DRB1*03:01 binder were independently validated by NetMHCIIpan 4.3[14] with identical binding cores (Table 5). The strongest was HRNLRRDNQISICKP, binding DRB1*03:01 with MixMHC2pred %Rank 0.785, NetMHCIIpan %Rank_EL 0.68, IC50 45.78 nM, and core LRRDNQISI. The four DQ8 binders (EHKDGQVQAKGERPG, CEHKDGQVQAKGERP, HGSDFATKMSYENPA, LHGSDFATKMSYENP) resolved to two cores (DGQVQAKGE and DFATKMSYE), each confirmed by both algorithms. DQ2.5 binding was predicted by MixMHC2pred but did not replicate with NetMHCIIpan, and we therefore do not present DQ2.5 as a validated allele.

Table 5 Independent validation of SIMALR micropeptide HLA class II binders by NetMHCIIpan 4.3

Peptide	Allele	MixMHC2pred	NetMHCIIpan	IC50_nM	Binding_core
EHKDGQVQAKGERPG	DQ8	0.146	0.36	n.d.	DGQVQAKGE
CEHKDGQVQAKGERP	DQ8	0.205	0.48	n.d.	DGQVQAKGE
HGSDFATKMSYE	DQ8	0.463	0.22	n.d.	DFATKMSYE

Peptide	Allele	MixMHC2pred	NetMHCIIpan	IC50_nM	Binding_core
NPA					
LHGSDFATKMSYENP	DQ8	0.608	0.31	n.d.	DFATKMSYE
HRNLRRDNQISICKP	DRB1*03:01	0.785	0.68	45.78	LRRDNQISI

To ask whether SIMALR peptides might mimic known beta-cell autoantigens, we cross-referenced the 14 binding cores against 1,964 unique T-cell epitopes from eight established T1D autoantigens (INS, GAD65, IA-2, ZnT8, IGRP, CHGA, ICA1, GAD1; IEDB[8]). SIMALR peptides shared zero 7-mers, zero 8-mers, and zero 9-mers with any known epitope, and had no exact matches (Figure 5). If the predicted HLA binding translates to natural presentation, SIMALR would represent a distinct, non-beta-cell antigen class.

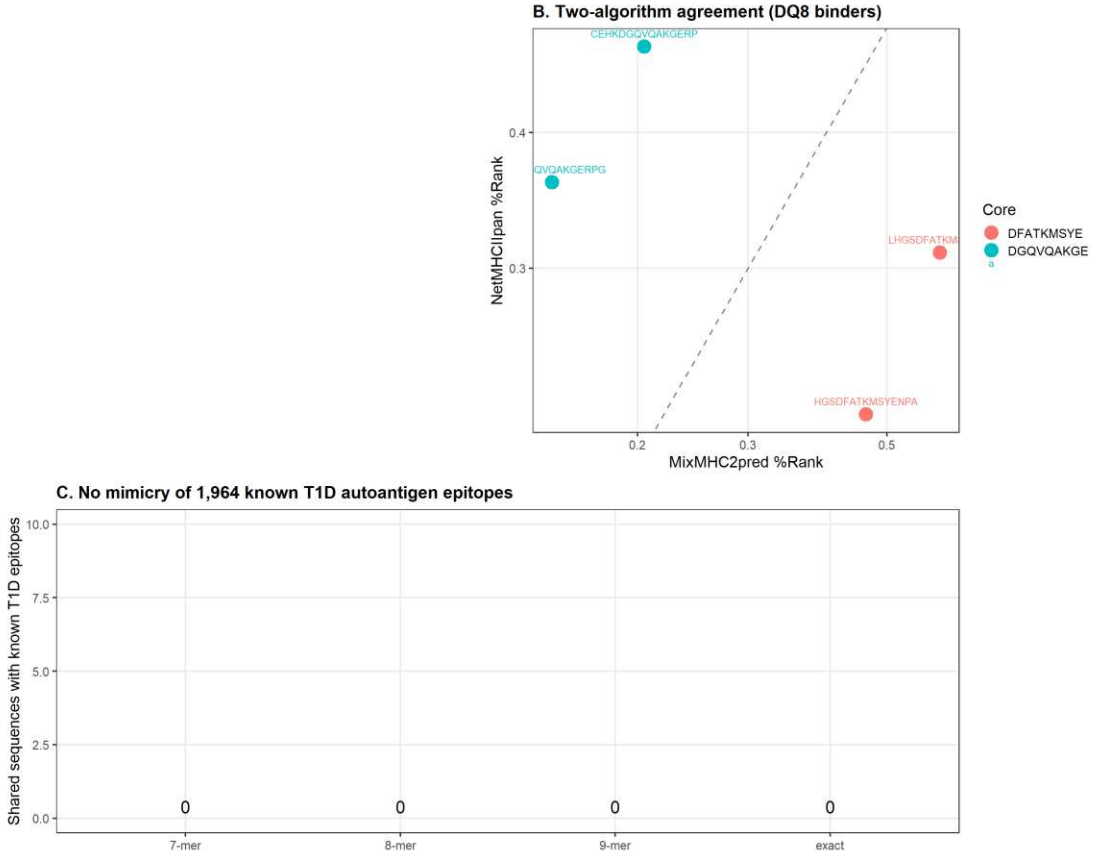


Figure 5 HLA class II binding predictions and non-mimicry of known T1D autoantigens. A: Distribution of MixMHC2pred %Rank values; B: Agreement between MixMHC2pred and NetMHCIIpan for validated binders; C: Zero shared k-mers with 1,964 known T1D autoantigen epitopes.

SIMALR is detectable in human blood monocytes

To ask whether SIMALR is expressed in an accessible human compartment, we analyzed single-cell PBMC data from 27 T1D and 27 healthy donors (GSE316337). SIMALR was detected in 727 of 376,996 cells (0.19%), at comparable rates between groups (T1D 0.200%, HC 0.185%; Fisher p = 0.3). At resting state, in unsorted PBMCs, SIMALR is expressed but does not distinguish disease from control.

In a monocyte-enriched single-cell cohort with three disease states (GSE221297; 6 T1D, 3 healthy, 3 pre-diabetic), SIMALR detection was highest in pre-diabetic donors (0.541%), intermediate in healthy controls (0.241%), and lowest in established T1D (0.128%) (Figure 6). Cell-level Fisher tests gave strong p-values for these differences (T1D versus pre-diabetic p = 4.6e-06), but cell-level tests treat each cell as independent when they are not; the honest test is at the donor level, where the same comparison gives p = 0.053 (n = 6 versus 3) and T1D versus healthy control gives p = 0.52. One healthy donor falls within the range of the pre-diabetic group. We therefore report this as a stage-dependent trend that requires replication in a larger pre-diabetic cohort, not as an established finding.

Independently, SIMALR is enriched in CD14+ monocytes of tuberculosis patients, correlates inversely with TB disease severity, and decreases after anti-TB therapy[17]. Together, these data establish that SIMALR is expressed in human blood monocytes in vivo and is measurable at baseline, but does not currently support a resting-state T1D biomarker claim.

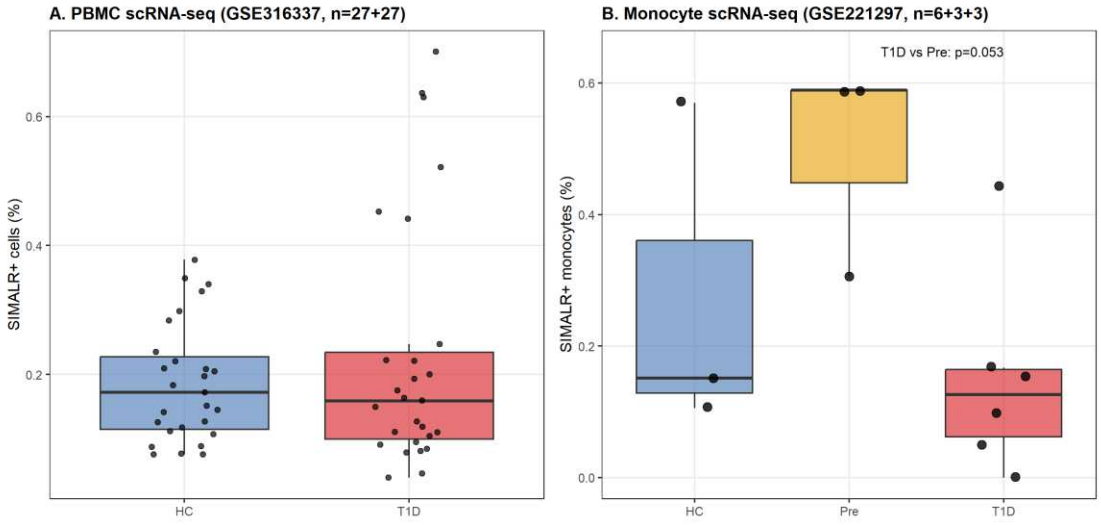


Figure 6 SIMALR expression in human blood monocytes. A: PBMC scRNA-seq from 27 T1D and 27 HC donors (GSE316337); B: Monocyte detection rate by disease state in GSE221297.

Immunopeptidomics of cytokine-treated islets does not detect SIMALR peptides

We searched the largest public islet immunopeptidomics dataset available (MSV000093843)[6] for SIMALR-derived sequences. The dataset was generated from cadaveric islets cultured for 48 h with IFN-gamma, IL-1beta, and TNF-alpha, the same cytokine cocktail used in our RNA-seq analyses, and therefore represents the disease-relevant compartment. We downloaded the islet HLA-DR, HLA-DQ, HLA-DP, and pan-class I identification files and extracted 7,507 unique peptide sequences. We searched these against 4,474 SIMALR open-reading-frame-derived 8-15mers, the 14 binding cores, and the 10 strongest predicted binders, with one-mismatch fuzzy matching on cores. No SIMALR-derived sequence was detected.

Discussion

We identify SIMALR as a T1D-specific, monocyte/macrophage-restricted lncRNA at the TNFAIP3 6q23 risk locus, induced by T1D-relevant cytokines in every donor tested, with micropeptide products predicted to bind the T1D-associated HLA class II alleles DQ8 and DRB1*03:01 and with no sequence similarity to known beta-cell autoantigens.

The 6q23 locus has been the subject of two decades of genetic and functional work without a consensus on its causal transcript. TNFAIP3 remains the leading candidate for rheumatoid arthritis and systemic lupus erythematosus, where allele-specific regulation has been demonstrated in disease-relevant contexts. For T1D specifically, our data are consistent with a model in which SIMALR rather than TNFAIP3 carries the disease-relevant signal: SIMALR is induced by T1D-relevant cytokines, TNFAIP3 is not context-induced in the same experiment; SIMALR is T1D-specific in independent PanKbase integration, while TNFAIP3 is pleiotropic; and SIMALR encodes predicted class II-binding peptides while TNFAIP3 does not. This does not exclude TNFAIP3 from a role at 6q23; both transcripts share the same topologically associating domain and the same rs6920220 risk region, and both could contribute to shared autoimmune susceptibility. The data do argue that SIMALR warrants consideration as a second, non-redundant causal transcript at this locus.

The predicted HLA class II binding of SIMALR micropeptides, combined with the absence of sequence mimicry of any known beta-cell autoantigen, raises the possibility that T-cell reactivity in T1D includes a non-beta-cell component. Under this model, cytokine-activated islet macrophages upregulate SIMALR, produce micropeptides that are processed and presented on macrophage HLA class II, and are recognized by CD4+ T cells that are not reactive against insulin, GAD65, IA-2, ZnT8, IGRP, CHGA, ICA1, or GAD1. This would represent a distinct autoimmune axis from the classical beta-cell antigen paradigm. The hypothesis is speculative: no SIMALR peptide was detected in bulk islet immunopeptidomics, and T-cell reactivity was not tested. We present it as a testable model rather than an established mechanism.

Four of the five validated binders are DQ8-restricted. DQ8 is the dominant T1D risk haplotype in humans and is carried by HPAP-002, a T1D donor in the PanKbase collection. The convergence of DQ8 restriction with the class II peptide binding predictions and with the disease-specific expression of SIMALR makes DQ8-restricted T-cell reactivity a natural next experimental question.

The 6q23/TNFAIP3/SIMALR axis is dormant in type 2 diabetic islets: SIMALR is not detected, TNFAIP3 is unchanged, and CXCL10 is unchanged. This clean negative distinguishes the T1D-specific cytokine response from the chronic low-grade inflammation that characterizes T2D and supports a T1D-specific rather than general inflammatory role for the locus.

SIMALR was originally identified as a macrophage-restricted lncRNA that suppresses inflammatory apoptosis through Netrin-1[5]. Its HGNC designation reflects this function. In the tuberculosis literature, SIMALR is enriched in CD14+ monocytes of patients and correlates with disease severity[17]. Our work extends these findings in three directions: first, by linking SIMALR to the specific T1D risk locus at 6q23 in the disease-relevant tissue, primary human islets; second, by demonstrating the cytokine-inducibility of SIMALR in 19 of 19 donors, which places it in the T1D-relevant regulatory context; and third, by predicting its HLA class II-binding micropeptide products. The hypothesis that SIMALR is a macrophage survival factor is compatible with all our data and provides a coherent model for how a macrophage-restricted lncRNA at 6q23 might act in T1D.

The negative immunopeptidomics result is consistent with the expected biology. SIMALR is macrophage-restricted, and islet-resident macrophages represent a small fraction of total cells in whole-islet preparations. At the depth of bulk islet immunopeptidomics (approximately 3,000-4,000 peptides per HLA class II allele), a macrophage-restricted peptide is expected to fall below detection. Direct demonstration of natural presentation would require macrophage-enriched or sorted immunopeptidomics from cytokine-stimulated tissue.

Limitations. The study is entirely computational. Every result is derived from public datasets and from prediction algorithms; no wet-lab experiment was performed.

The HLA class II binding predictions are supported by two independent algorithms with concordant cores but are not confirmed by mass spectrometry. The bulk islet immunopeptidomics search we performed was negative and cannot exclude presentation below the detection limit.

The public SIMALR ASO knockdown dataset available achieved only a partial knockdown (log2FC minus 0.85, q = 0.999) and cannot support causal claims about downstream chemokine effects.

Islet-resident macrophages were not directly interrogated. The PanKbase islet atlas v3.4 (448,935 cells) did not detect SIMALR in any cell-type marker list, consistent with a

cytokine-inducible rather than constitutive expression pattern but leaving the in situ macrophage question open.

The islet capture Hi-C panel did not directly bait the SIMALR promoter, so islet chromatin evidence is at TAD level rather than at fragment level.

Blood expression results are null at the disease-state level. The pre-diabetic trend is based on three donors and requires replication.

Natural presentation of SIMALR peptides on HLA class II, T-cell reactivity against them, and any causal role in islet inflammation remain to be demonstrated experimentally.

Conclusion

SIMALR is a T1D-specific, macrophage-restricted lncRNA at the TNFAIP3 6q23 risk locus, induced by T1D-relevant cytokines in 19 of 19 primary human islet donors, absent in type 2 diabetes, regulated by NF-kB and STAT3, and encoding HLA class II-binding micropeptides with no similarity to known T1D autoantigens. SIMALR represents a candidate non-beta-cell antigen class at a shared autoimmune risk locus, and its immediate experimental tests are targeted immunopeptidomics and DQ8 tetramer staining in new-onset T1D.

Acknowledgements

The author thanks the Human Pancreas Analysis Program consortium, the PanKbase team, the GTEx Consortium, and the laboratories of Nanaware, Cynn, Xu, Miguel-Escalada, and Martin for making their data publicly available. This work received no external funding.

References

1. Chiou J, Geusz RJ, Okino ML, et al. A multi-ethnic genome-wide association study identifies novel loci for type 1 diabetes. Nat Commun 2021;12:4697 [PMID: 34373454 DOI: 10.1038/s41467-021-24970-4]

2. Robertson CC, Inshaw JRJ, Chim K, et al. Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. Nat Genet 2021;53:962-971 [PMID: 34045743 DOI: 10.1038/s41588-021-00880-5]

3. Olsson LM, Lindqvist AK, Kallberg H, et al. A case-control study of rheumatoid arthritis identifies an eQTL for TNFAIP3 at the 6q23 risk locus. Arthritis Rheum 2015 [PMID: 25707945]

4. Martin P, McGovern A, Orozco G, et al. Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. Genome Biol 2016;17:212 [PMID: 27724958 DOI: 10.1186/s13059-016-1078-x]

5. Cynn E, et al. Human macrophage long intergenic non-coding RNA, SIMALR, suppresses inflammatory macrophage apoptosis via Netrin 1. Arterioscler Thromb Vasc Biol 2023 [PMID: 36799758]

6. Nanaware PP, et al. MHC class I and II immunopeptidomics of human pancreatic islets. Cell Rep 2025 [PMID: 40737125]

7. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 2020;369:1318-1330 [PMID: 32913098 DOI: 10.1126/science.aaz1776]

8. Vita R, Mahajan S, Overton JA, et al. The Immune Epitope Database (IEDB) 2018 update. Nucleic Acids Res 2019;47:D339-D343 [PMID: 30357371 DOI: 10.1093/nar/gky1006]

9. Bastian F, Parmentier G, Roux J, et al. The Bgee suite for retrieval and comparison of gene expression patterns across multiple animal species. Bioinformatics 2016;32:3399-3401 [PMID: 27378296 DOI: 10.1093/bioinformatics/btw372]

10. Miguel-Escalada I, Bonas-Guarch S, Cebola I, et al. Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. Nat Genet 2019;51:1137-1148 [PMID: 31253979 DOI: 10.1038/s41588-019-0457-0]

11. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550 [PMID: 25516281 DOI: 10.1186/s13059-014-0550-8]

12. Pages H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package, 2020

13. Gfeller D, Guillaume P, Michaux J, et al. Improved predictions of MHC class II antigen presentation. bioRxiv 2019 [DOI: 10.1101/589838]

14. Andreatta M, Karosiene E, Rasmussen M, et al. Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. Immunogenetics 2016;68:871-880 [PMID: 27575705 DOI: 10.1007/s00251-016-0945-7]

15. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184:3573-3587 [PMID: 34062119 DOI: 10.1016/j.cell.2021.04.048]

16. Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell 2019;177:1888-1902 [PMID: 31178118 DOI: 10.1016/j.cell.2019.05.031]

17. Xu Y, et al. LINC02528/SIMALR is a blood biomarker and functional regulator of macrophage responses in tuberculosis. PLoS Pathog 2025 [PMID: 41308682]