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.

InsulatorLeak, a mechanism-first pipeline for variant prioritization by predicted CTCF insulator disruption across seven autoimmune diseases

Navya Shah

info@karmyxbiolabs.com

Karmyx BioLabs https://orcid.org/0009-0008-8018-2161

Research Article

Keywords: autoimmune disease, insulator, CTCF, Enformer, fine-mapping, SuSiE, eQTL, pQTL, colocalization, variant prioritization, multiple sclerosis, inflammatory bowel disease, psoriasis, atopic dermatitis, type 1 diabetes, systemic lupus erythematosus, rheumatoid arthritis, disease-specific loci, GAIM

Posted Date: April 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9368949/v1

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

Additional Declarations: The authors declare potential competing interests as follows: Lead PI - no pharma sponsored any part of this research or specific analysis

InsulatorLeak, a mechanism-first pipeline for variant prioritization by predicted CTCF insulator disruption across seven autoimmune diseases

2026

Title - InsulatorLeak, a mechanism-first pipeline for variant prioritization by predicted CTCF insulator disruption across seven autoimmune diseases

Author - Navya Shah

Karmyx BioLabs, Naperville, IL, USA

Corresponding author: Navya Shah · info@karmyxbiolabs.com

Keywords: autoimmune disease; insulator; CTCF; Enformer; fine-mapping; SuSiE; eQTL; pQTL; colocalization; variant prioritization; multiple sclerosis; inflammatory bowel disease; psoriasis; atopic dermatitis; type 1 diabetes; systemic lupus erythematosus; rheumatoid arthritis; disease-specific loci; GAIM

Abstract

Genome-wide association studies have mapped hundreds of autoimmune risk loci, yet fewer than 5% of associated variants have established causal mechanisms. CTCF-bound insulators partition the genome into regulatory domains. When disrupted, enhancers can activate genes they normally do not regulate, a phenomenon termed insulator leakage that has been implicated in malignancy and may contribute to autoimmunity. We reasoned that prioritizing variants by predicted insulator disruption, then validating that prioritization empirically across independent cohorts and diseases, would yield candidates suitable for functional follow-up and target selection.

InsulatorLeak combines SuSiE fine-mapping, Enformer-predicted CTCF disruption (Scaled Average Difference, SAD), locus-matched null calibration for empirical p-values, chromatin enrichment (ChromHMM, Hi-C), eQTL and pQTL colocalization, and benchmarking against FUMA, Sei, and Boltz-2. We applied the pipeline to 14 non-MHC autoimmune loci across seven diseases (multiple sclerosis, inflammatory bowel disease, rheumatoid arthritis, type 1 diabetes, systemic lupus erythematosus, psoriasis, and atopic dermatitis) using 19 GWAS datasets spanning more than 2.5 million total participants.

In MS, TNFRSF1A replicated in 4/6 cohorts (p=0.022), and IL2RA reached p=0.006 in IMSGC MS Chip (N=115,803). Pan-autoimmune replication across seven diseases showed IL7R in 6/7 expansion cohorts (p=0.011–0.052) and TNFRSF1A in 5/7 (p=0.011–0.033), indicating insulator disruption as a shared autoimmune mechanism. Thirteen of 14 loci replicated in at least two ancestry groups. IL12AB was significant in all four ancestries, and TYK2 reached FDR q=0.040 in South Asian. Chromatin enrichment was strongest at drug-target loci (CD58 21×, IL2RA 23×, IL7R 12×, TNFRSF1A 5.3× in MS, CD40 8.6×

and IL7R 34× in IBD). pQTL colocalization yielded cognate validation at CD58 (0.996), CD40 (0.958), CD6 (0.79 in IBD), and IL2RA (0.79). Both MS and IBD showed the full validation stack from fine-mapping through pQTL. Overlap with FUMA lead SNPs was 2/120, indicating complementary rather than redundant prioritization. In RA, four loci reached $p<0.05$ (CLEC16A, TNFRSF1A, STAT3, BACH2). Extension to T1D, SLE, psoriasis, and AD is reported in Supplementary Table S8.

In a disease-specific locus extension (GAIM; General Autoimmune Insulator Mechanism) we applied the identical pipeline (SuSiE, Enformer CTCF SAD, and locus-matched nulls drawn from each indication's own GWAS) to fourteen top non-MHC loci per disease for five large-summary-statistic cohorts: atopic dermatitis meta-analysis (AD), T1D (Chiou et al. 2021), RA (FinnGen), IBD (de Lange et al. 2017), and SLE (Bentham et al. 2015). This tests insulator disruption at regions prioritized by each disease, not only at the MS-derived 14 locus set. Supplementary Table S15 lists, for every locus–cohort pair, the minimum empirical p-value, Benjamini–Hochberg q-value, variant identifier, and variant count. Representative indication-native signals include IL33 and STAT6 (AD), CCR7 and CTLA4 (T1D), TNFAIP3 (RA and SLE), RGS14 and TNFSF15 (IBD), and ITGAM (SLE). Full data are in Supplementary Table S15.

InsulatorLeak produces mechanism-first target validation with empirical prioritization across autoimmune diseases. The pipeline generalizes from MS to six additional diseases. Loci corresponding to approved therapeutics (IL12AB/ustekinumab, TYK2/deucravacitinib, TNFRSF1A/etanercept, IL2RA/daclizumab) show consistent insulator-disruption signal, aligning mechanism-first prioritization with known drug targets. At shared loci, multiple diseases exhibit insulator disruption with disease-specific credible sets, consistent with insulator disruption as a general autoimmune mechanism at shared loci rather than transfer of MS-specific variants. The modular design permits application to other complex traits. Code is available upon request.

Introduction

Autoimmune diseases, including multiple sclerosis (MS), inflammatory bowel disease (IBD), rheumatoid arthritis (RA), type 1 diabetes (T1D), systemic lupus erythematosus (SLE), psoriasis, and atopic dermatitis (AD), collectively affect hundreds of millions worldwide and share substantial genetic architecture. Genome-wide association studies have identified hundreds of risk loci across these conditions, with the major histocompatibility complex (MHC) on chromosome 6 as the dominant shared signal and more than 14 established non-MHC loci common to multiple diseases (Sawcer et al., 2011; IMSGC, 2019; Liu et al., 2015). Despite this progress, fewer than 5% of identified variants have known causal mechanisms. Pharma and drug discovery programs typically know which loci associate with disease (the WHAT) but not the regulatory mechanism by which variants confer risk (the WHY). Mechanism-first target validation narrows hundreds of thousands of candidates to a small, mechanism-prioritized set suitable for druggable target dossiers, a reduction on the order of 45,000-fold from raw GWAS output to credible candidates.

Insulators are DNA elements that organize the genome into domains and limit inappropriate interactions between enhancers and promoters. CTCF is a central insulator-associated factor. When insulators are disrupted, enhancers can activate genes they normally do not regulate, a phenomenon termed insulator leakage or enhancer hijacking that contributes to malignancy and may contribute to autoimmunity (Flavahan et al., 2016, Dowen et al., 2014). Identifying which autoimmune-associated variants alter

insulator or CTCF function could clarify mechanisms of risk and suggest targets for follow-up experiments.

Enformer is a deep-learning model that predicts gene expression and chromatin tracks from DNA sequence over long windows (~196 kb) and has been used to quantify the effects of sequence changes on regulatory signals (Avsec et al., 2021). We hypothesized that combining Enformer with statistical fine-mapping could predict which autoimmune-associated variants reduce CTCF-related signal (insulator loss), and that locus-matched null calibration would yield empirical p-values suitable for prioritization. This approach is complementary to annotation-first tools such as FUMA (Watanabe et al., 2017) and GARFIELD (Iotchkova et al., 2019), which prioritize by positional overlap, eQTL, and chromatin annotation rather than by a specific mechanistic hypothesis.

We developed InsulatorLeak, an open-source pipeline that integrates SuSiE fine-mapping (Wang et al., 2020), Enformer-predicted CTCF disruption (Scaled Average Difference, SAD), locus-matched null calibration, chromatin enrichment, and eQTL/pQTL colocalization. We apply it to 14 non-MHC autoimmune loci across seven diseases (MS, IBD, RA, T1D, SLE, psoriasis, and atopic dermatitis) using 19 GWAS datasets spanning more than 2.5 million total participants. Our objectives are to prioritize variants by predicted insulator disruption, validate with locus-matched nulls and empirical p-values, test replication across cohorts and ancestries, integrate chromatin and eQTL/pQTL for multi-omic validation, and generalize the pipeline to T1D, SLE, psoriasis, and AD. We further report a prespecified extension in which the same computational stack is applied to fourteen non-MHC loci chosen per disease from large GWAS for AD, T1D, RA, IBD, and SLE (GAIM; General Autoimmune Insulator Mechanism), with null variants and null Enformer scores generated separately for each indication (Supplementary Table S15). We show that mechanism-first target validation yields empirically validated candidates, generalizes across diseases (IL7R 6/7 expansion cohorts, TNFRSF1A 5/7), and aligns with known drug targets while surfacing additional candidates that annotation-first tools do not prioritize.

Methods

Evidence foundation

The analysis spans 14 non-MHC loci × 19 GWAS datasets across seven autoimmune diseases, with more than 2.5 million total participants (Supplementary Table S_GWAS). This evidence foundation underpins all replication and cross-disease validation.

Data sources

We obtained GWAS summary statistics from the following sources (Supplementary Table S_GWAS). For MS we used FinnGen R12 (primary), Pan-UKBB, GCST005531, IEU ieu-a-1024, IMSGC MS Chip (N=115,803), and IMSGC Severity. For IBD, FinnGen R12. For RA, IEU ieu-a-834 (MHC). For T1D, FinnGen R12 and Chiou et al. 2021 meta-analysis (~520K samples). For SLE, FinnGen R12. For psoriasis, FinnGen R12 L12_PSORIASIS (12,760 cases) and Stuart et al. 2025 meta-analysis (GCST90472771) (36,466 cases). For atopic dermatitis, FinnGen R12 L12_ATOPIC (31,245 cases) and AD Meta 2025 (Riley-Gillis et al.) (56,146 cases). LD reference was 1000 Genomes Phase 3 European (EUR) for MHC and locus-specific LD for non-MHC loci. We used pQTL data from UK Biobank Protein Paint Project (UKB-PPP, 2,923 proteins), eQTL data from eQTLGen blood cis-eQTLs and GTEx v8

multi-tissue, and chromatin annotations from ENCODE cCRE, ChromHMM Roadmap (CD4 T, B cell), and Hi-C loop anchors (GM12878, CD4+ T, B cell). Reference genome was hg38 (GRCh38).

Supplementary Table S_GWAS. GWAS datasets (19 total).

#	Cohort	Disease	Ancestry	Sample Size	Role
1	FinnGen MS (R12)	MS	EUR	3,002 / 497K	Primary discovery
2	IMSGC MS Chip (2019)	MS	EUR	47,429 / 68,374	Largest MS GWAS
3	IMSGC Severity (2023)	MS	EUR	12,584 MS patients	Severity GWAS
4	GCST005531	MS	EUR	14,498 / 24,091	Replication
5	IEU ieu-a-1024	MS	EUR	9,722 / 17,376	Replication
6	Pan-UKBB	MS	EUR	1,538 / 403K	Replication
7	VA-MVP AFR	MS	AFR	637 / 121K	Multi-ancestry
8	VA-MVP Hispanic	MS	AMR	1,859 / 314K	Multi-ancestry
9	GCST90705065 (Jacobs)	MS	SAS	N = 9,456	South Asian
10	FinnGen IBD (R12)	IBD	EUR	15,867 / 484K	Cross-disease
11	FinnGen RA (R12)	RA	EUR	16,741 / 503K	Cross-disease
12	IEU ieu-a-834 (Okada)	RA	EUR	14,361 / 43,923	RA MHC
13	FinnGen T1D (R12)	T1D	EUR	~8K / 497K	Expansion
14	Chiou et al. 2021 T1D meta	T1D	EUR	~520K samples	Expansion
15	FinnGen SLE (R12)	SLE	EUR	~5K / 497K	Expansion
16	FinnGen Psoriasis (R12)	Psoriasis	EUR	12,760 / 482K	Expansion
17	Stuart 2025 psoriasis meta	Psoriasis	EUR	36,466 / 458K	Expansion
18	FinnGen Atopic Dermatitis (R12)	AD	EUR	31,245 / 433K	Expansion
19	AD Meta 2025 (Riley-Gillis et al.)	AD	EUR	56,146 / 602K	Expansion

14 non-MHC loci

We focused on 14 established autoimmune loci (CD6, IL2RA, IL7R, TYK2, CD58, CLEC16A, TNFRSF1A, BACH2, CD40, TNFSF14, IRF8, EVI5, STAT3, IL12AB). Genomic coordinates were defined per locus from published fine-mapping studies and GWAS Catalog.

Pipeline steps

Locus extraction. For each disease and locus, we extracted GWAS summary statistics for the locus window, harmonized to hg38 where needed (Pan-UKBB required liftover from hg37), and prepared SuSiE-ready input with columns chrom, pos, ref, alt, beta, se, pval.

SuSiE fine-mapping. We ran SuSiE RSS (susieR) with 95% coverage credible sets, using locus-specific LD matrices from 1000 Genomes EUR. We extracted the top-PIP variant per credible set (or top 20 variants by PIP when no credible set met coverage) for Enformer inference.

Null variant generation. We generated locus-matched, disease-specific null variants ($p > 0.5$ in GWAS) per locus. Nulls are variants in the same locus window that are not associated with the trait, sharing genomic context with test variants but not linked to the disease. We sampled up to ~500 nulls per locus to build a null SAD distribution. Nulls are disease-specific (psoriasis nulls from the psoriasis GWAS, RA nulls from the RA GWAS, MS nulls from MS GWAS). A variant can be null for one disease ($p > 0.5$) but associated with another ($p < 0.01$), so nulls cannot be reused across diseases. For South Asian analyses, we used ancestry-matched nulls from GCST90705065. Enformer inference runs twice per disease, first on credible-set variants (~10–50 per locus) to obtain test SAD, then on the ~500 null variants per locus to obtain null SAD. The empirical p-value compares each test SAD to the null distribution.

Enformer inference. We extracted 196 kb windows centered on each variant from hg38, constructed reference and alternative sequences, and ran Enformer (TensorFlow Hub) on both. We computed SAD = (alt − ref) for the CTCF track (index 5312), taking mean, min, and max across the window. Batch processing with checkpointing was used for large variant sets.

Empirical p-values. We compared each credible-set variant's SAD to the locus-matched null SAD distribution. The empirical p-value equals the proportion of null SADs ≤ observed SAD (one-sided, insulator loss). We applied Benjamini–Hochberg for FDR (q-values). For multiple-testing correction across the 6-cohort table (14 loci × 6 cohorts), we applied FDR in two ways, per locus (6 tests per locus, asking whether the locus replicates across cohorts) and per cohort (14 tests per cohort, asking whether the cohort shows signal across loci) (Supplementary Table S13).

Chromatin enrichment. We intersected credible-set variants with ChromHMM states (CD4 naive/memory T, B cell), Hi-C loop anchors (CD4 T, B cell), and ENCODE cCRE. We compared overlap to locus-matched nulls using Fisher's exact test.

eQTL and pQTL colocalization. We ran coloc (Giambartolomei et al., 2014) between GWAS and eQTL (eQTLGen, GTEx) or pQTL (UKB-PPP) for each locus. We report PP.H4 (shared causal variant) and PP.H3 (distinct causal variants).

DICE immune cell eQTL colocalization (supplementary analysis). To assess whether InsulatorLeak-prioritized variants have effects in activated immune cell contexts, we performed a supplementary colocalization analysis using the DICE Database (Schmiedel et al., 2018; n = 28–91 donors per cell type). We downloaded genome-wide eQTL summary statistics for five cell types: $CD4^+$ T naive (n = 91), $CD4^+$ T stimulated (anti-CD3/CD28; n = 91), $CD8^+$ T naive (n = 84), $CD8^+$ T stimulated (n = 84), and T regulatory naive (n = 28). Locus windows were lifted over from hg38 to hg19 using pyliftover. We ran coloc.abf between the IMSGC MS Chip GWAS (N = 115,803; s = 0.292) and each DICE cell type at all 9 well-powered non-MHC loci, matching variants by rsID, using default priors ($p_1 = 1 \times 10^{-4}$, $p_2 = 1 \times 10^{-4}$, $p_{12} = 1 \times 10^{-5}$). Full results are in Supplementary Table S_DICE.

Benchmarking. We ran FUMA SNP2GENE on FinnGen MS (MHC) and on 14 non-MHC loci. We compared FUMA lead/candidate SNPs to InsulatorLeak top-PIP variants. We ran Sei sequence-class predictions on MHC and non-MHC variants. Boltz-2 structural validation. To assess whether pQTL-implicated proteins are amenable to structure-based prediction, we ran Boltz-2 (MIT license, open-source) on four pQTL colocalizing proteins, IGF1R kinase domain with Linsitinib (structure and affinity) and LILRB4, CD22, and CD3D (structure only). Boltz-2 predicts protein structure and, for protein–ligand pairs, binding affinity. Outputs include ptm (predicted TM-score) and complex_plddt (per-residue confidence).

Software and compute

R (susieR, coloc, data.table), Python (pandas, pysam, tensorflow, tensorflow_hub, numpy), bedtools, PLINK2, bcftools. SuSiE on AWS EC2 r5.2xlarge. Enformer on g4dn.xlarge (Tesla T4 GPU). DICE eQTL colocalization run locally (Apple M-series; bcftools 1.23, R 4.5, coloc v5, pyliftover).

Results

MS discovery and replication across six cohorts

We applied InsulatorLeak to 14 non-MHC loci across six MS cohorts (FinnGen, Pan-UKBB, GCST005531, IEU, IMSGC MS Chip, and IMSGC Severity). Locus-null empirical p-values are summarized in Table 1.

Table 1. Six-cohort locus null *p*-values (MS, 14 non-MHC loci).

Locus	FinnGen	Pan-UKBB	GCST005531	IEU	IMSGC MS Chip	IMSGC Severity	Sig (p<0.05)
TNFRSF1A	0.022	0.022	0.022	0.099	0.022	0.055	4/6
IL2RA	0.044	0.273	0.103	0.044	0.006	0.138	3/6
EVI5	0.006	0.463	0.028	0.074	0.030	0.064	3/6
IL7R	0.195	0.213	0.052	0.155	0.011	0.052	1/6
CD40	0.018	0.086	0.952	0.036	0.064	0.046	3/6
BACH2	0.035	0.281	0.193	0.211	0.053	0.035	2/6
CD58	0.979	0.007	0.278	0.007	0.090	0.333	2/6
CLEC16A	0.219	0.008	0.428	0.164	0.219	0.056	1/6
STAT3	0.297	0.579	0.002	0.818	0.579	0.008	2/6
TYK2	0.040	0.053	0.053	0.240	0.773	0.307	1/6
CD6	0.258	0.015	0.318	0.379	0.424	0.106	1/6
TNFSF14	0.156	0.004	0.052	0.752	0.691	0.102	1/6
IRF8	0.080	0.192	0.182	0.182	0.335	0.299	0/6

Locus	FinnGen	Pan-UKBB	GCST005531	IEU	IMSGC MS Chip	IMSGC Severity	Sig (p<0.05)
IL12AB	0.068	0.305	0.054	0.842	0.118	0.024	1/6

TNFRSF1A replicated in 4/6 cohorts (p=0.022 in each), a replication rate that is uncommon for computational prioritization methods. Under per-locus FDR (6 tests per locus), TNFRSF1A survives FDR in 4/6 cohorts (q=0.033). CD58 and STAT3 in 2/6. IL2RA, EVI5, CLEC16A, and TNFSF14 in 1/6. Seven loci have at least one cohort with q<0.05. Twelve locus-cohort pairs survive per-locus FDR (Supplementary Table S13). IL2RA reached p=0.006 in IMSGC MS Chip (largest MS GWAS, N=115,803) with 3/6 cohort replication. EVI5, CD40, and BACH2 each replicated in 2–4 cohorts. STAT3 and IL12AB showed strong signals in severity-focused cohorts (STAT3 p=0.008, IL12AB p=0.024), consistent with a susceptibility-versus-severity divergence (Table 4). All 14 loci showed consistent negative SAD (predicted insulator loss) where significant.

Table 4. Susceptibility versus severity stratification.

Category	Loci	Interpretation
Susceptibility	TNFRSF1A, IL2RA, CD58	Strong in susceptibility GWAS, consistent across cohorts
Dual-purpose	CD40, BACH2	Significant in both susceptibility and severity cohorts
Severity	STAT3, IL12AB	Significant in GCST005531 susceptibility (p=0.002) and IMSGC Severity (p=0.008); not significant in FinnGen, Pan-UKBB, or MS Chip susceptibility cohorts

Multi-ancestry validation (four groups)

We extended analyses to four ancestry groups, European (6 cohorts), African American (va_mvp_afr), Hispanic/Latino (va_mvp_hispanic), and South Asian (GCST90705065), using ancestry-matched locus nulls. Results are in Table 2.

Table 2. Cross-ancestry locus null *p*-values (best per ancestry).

Locus	EUR	AFR	AMR	SAS	Ancestries p<0.05	Direction
IL12AB	0.024	0.024	0.002	0.006	4/4	All negative
TYK2	0.040	1.000	0.013	0.002	3/4	3/4 negative
CD40	0.018	0.004	n/a	0.036	3/4	All negative
EVI5	0.006	0.006	0.022	0.091	3/4	All negative
STAT3	0.002	0.948	0.018	0.057	2/4	3/4 negative
TNFSF14	0.004	n/a	n/a	0.006	2/2	All negative
CD58	0.007	n/a	n/a	0.006	2/2	All negative
CD6	0.015	n/a	n/a	0.008	2/2	All negative

Locus	EUR	AFR	AMR	SAS	Ancestries p<0.05	Direction
IL2RA	0.006	n/a	n/a	0.014	2/2	All negative
IL7R	0.012	n/a	n/a	0.016	2/2	All negative
TNFRSF1A	0.022	0.736	n/a	0.030	2/3	2/3 negative
BACH2	0.035	n/a	n/a	0.032	2/2	All negative
CLEC16A	0.008	0.008	n/a	0.142	2/3	All negative
IRF8	0.080	0.495	0.495	0.024	1/4	2/4 negative

IL12AB was the only locus significant in all four ancestries. TYK2 reached FDR q=0.040 in South Asian (p=0.002, strongest SAD −0.0075). Thirteen of 14 loci replicated in at least two ancestry groups. Ten of 14 showed consistent negative SAD in every ancestry tested. No locus failed in all ancestry groups.

Chromatin enrichment

Credible-set variants were enriched in immune regulatory chromatin (Table 3). In MS, CD58 showed 21× enrichment in B cell ChromHMM (p=3.4e-8), IL7R 12× in B cell Hi-C (p=0.002), TNFRSF1A 5.3× in CD4 naive T ChromHMM (p=0.005), and IL2RA 23× in B cell ChromHMM (p=0.004). CLEC16A showed 2.8× enrichment in CD4 T Hi-C (p=0.037). In IBD, CD40 showed 8.6× enrichment in CD4 memory T ChromHMM (p=8.4e-6), IL7R 34× in B cell Hi-C (p=0.001), IL2RA 10.4× in B cell ChromHMM (p=0.017), and CD58 7.5× in CD4 T Hi-C (p=0.012). Drug targets (TNFRSF1A, CD58, IL2RA, IL7R, CD40) showed strongest enrichment.

Table 3. Chromatin enrichment (top loci).

Locus	Disease	Annotation	Odds ratio	p-value
CD58	MS	B cell ChromHMM	21×	3.4e-8
IL7R	MS	B cell Hi-C	12×	0.002
TNFRSF1A	MS	CD4 naive T ChromHMM	5.3×	0.005
IL2RA	MS	B cell ChromHMM	23×	0.004
CLEC16A	MS	CD4 T Hi-C	2.8×	0.037
CD40	IBD	CD4 memory ChromHMM	8.6×	8.4e-6
IL7R	IBD	B cell Hi-C	34×	0.001
IL2RA	IBD	B cell ChromHMM	10.4×	0.017
CD58	IBD	CD4 T Hi-C	7.5×	0.012

eQTL and pQTL colocalization

eQTL colocalization linked prioritized variants to gene expression. TNFRSF1A showed PP.H4 0.98 in Spleen and 0.96 in eQTLGen (lncRNA). IL2RA 0.96 in Spleen (drug target). IL7R 0.32 in eQTLGen. pQTL colocalization (UKB-PPP, 2,922 proteins) yielded cognate validation at CD58 (0.996), CD40 (0.958), IL2RA (0.79), and CD6 (0.79 in IBD). At the TNFRSF1A locus, CRH/LTA/LTB showed PP.H4

> 0.95. At the IL7R locus, IL7 (0.791), CD1C (0.678), and SIL1 (0.590) were top. IBD pQTL colocalization showed shared signal at IRF8, TYK2, EVI5, and TNFRSF1A across MS and IBD.

At the IL7R locus, disease-specific gene prioritization emerged. MS colocalizes with IL7R (eQTLGen 32%), while IBD colocalizes with SPEF2 (GTEx 73–76%). Same locus, different genes, consistent with disease-specific regulatory wiring rather than a generic pipeline artifact.

DICE stimulated immune cell eQTL colocalization. To directly test whether prioritized variants affect gene expression in activated immune cell states, we colocalized the IMSGC MS Chip GWAS against DICE eQTLs across five immune cell types ($CD4^+$ T naive/stimulated, $CD8^+$ T naive/stimulated, T regulatory naive; n = 28–91 per cell type; Schmiedel et al., 2018). The TYK2 locus showed the strongest signal (PP.H4 = 0.54 in $CD4^+$ T naive, PP.H4 = 0.51 in T regulatory naive), confirming a shared causal variant between the MS GWAS signal and T-cell eQTL effects. CLEC16A showed moderate evidence (PP.H4 = 0.27 in T regulatory naive, cognate gene), and IL2RA showed suggestive colocalization via IL15RA (PP.H4 = 0.21 in $CD4^+$ T naive). Critically, **no locus showed activation-dependent colocalization**: PP.H4 values did not increase substantially in $CD4^+$ T stimulated relative to $CD4^+$ T naive at any locus. This is consistent with the hypothesis that prioritized variants affect sequence-intrinsic CTCF binding rather than a state-dependent enhancer mechanism. Full results across 9 loci × 5 cell types are in Supplementary Table S_DICE.

Cross-disease validation (IBD and RA)

IBD provides independent validation of the same mechanism at the same loci. IBD uses different GWAS data (FinnGen IBD versus MS cohorts), different cases and controls (15,867 IBD cases versus MS cohorts), and disease-specific credible sets and LD structure. IBD validation therefore constitutes an independent test of whether insulator disruption generalizes across autoimmune diseases.

Both MS and IBD show the full validation stack: fine-mapping to credible sets, Enformer to insulator disruption (locus-null p-values), chromatin enrichment in immune regulatory regions, eQTL colocalization for variant–gene links, and pQTL colocalization for variant–protein links. In IBD (FinnGen, 14 loci), 115 credible-set variants overlapped with MS (14 exact chr:pos matches, 73 within 10 kb). Thirteen of 14 loci showed IBD–MS overlap. CD40 (p=0.020) and IL2RA (p=0.038) showed significant insulator disruption in IBD. IL7R (p=0.052) and BACH2 (p=0.053) were marginal. Chromatin enrichment was often stronger in IBD than in MS (e.g., IL7R 34× in IBD versus 12× in MS, CD40 8.6× in IBD). Strongest shared-variant loci were IRF8 (3 exact, 8/8 within 10 kb), IL2RA (2 exact, 6/6), IL7R (2 exact, 8/8), BACH2 (2 exact, 12/20), and EVI5 (2 exact, 7/10).

In RA (FinnGen R12, 14 non-MHC loci), locus-null empirical p-values showed four loci with $p<0.05$ (CLEC16A, TNFRSF1A, STAT3, BACH2). IL7R (p=0.052) and IL12AB (p=0.054) were marginal. RA credible sets (MHC, IEU ieu-a-834) comprised eight sets. Two RA variants exactly matched MS positions, and four loci fell within 10 kb of MS credible-set variants.

Pan-autoimmune replication across seven diseases

We applied the full InsulatorLeak pipeline to T1D (FinnGen, Chiou et al. 2021), SLE (FinnGen), psoriasis (FinnGen, Stuart et al. 2025), and AD (FinnGen, AD Meta 2025 (Riley-Gillis et al.)), seven expansion cohorts in total. Locus-null empirical p-values are in Supplementary Table S8.

IL7R replicated in 6/7 expansion cohorts (p=0.011–0.052) — the highest cross-disease replication of any locus, and notably stronger than its MS-specific signal (1/6 cohorts, p=0.011 in IMSGC MS Chip only), consistent with IL7R being a pan-autoimmune rather than MS-selective insulator disruption target. TNFRSF1A in 5/7 (p=0.011–0.033). CD40 reached p<0.05 in 4/7 expansion cohorts (3/6 MS cohorts). EVI5, CLEC16A, CD6, and IRF8 each reached p<0.05 in 4/7 cohorts. These results indicate that insulator disruption at these loci is a shared autoimmune mechanism across MS, IBD, T1D, SLE, psoriasis, and AD. Drug-target loci (TNFRSF1A, IL7R, CD40) show consistent signal across diseases.

Disease-specific patterns emerged. IRF8 was stronger in psoriasis and AD (4/7 expansion cohorts) than in MS (0/6 EUR). STAT3 was severity-specific in MS (p=0.008 in IMSGC Severity) but did not replicate in expansion susceptibility cohorts (0/7). IL12AB was strong in MS (4/4 ancestries) and SLE (p=0.004) but weaker in T1D and AD.

Expansion analyses have limitations. Several diseases (SLE, psoriasis, AD) have only one or two cohorts, limiting within-disease replication. All expansion cohorts are European. Multi-ancestry expansion would test generalizability. The 14 loci were chosen from MS. Some expansion diseases may have distinct locus architecture.

Disease-specific locus extension (GAIM; General Autoimmune Insulator Mechanism)

The primary analyses above apply a single set of 14 non-MHC loci (MS-derived) across diseases, with disease-specific fine-mapping credible sets and disease-specific locus nulls whenever the pipeline is run on a given GWAS (null variants are non-associated in that GWAS; they are not reused across traits).

To address whether insulator disruption is confined to MS-prioritized regions, we ran an extension (GAIM; General Autoimmune Insulator Mechanism) using fourteen distinct loci per indication, defined from published large GWAS for AD (Riley-Gillis et al. 2025 meta), T1D (Chiou et al. 2021), RA (FinnGen R12), IBD (de Lange et al. 2017 trans-ancestry meta), and SLE (Bentham et al. 2015; IEU). Locus lists were derived from published large-scale GWAS for each indication. For each cohort we performed SuSiE fine-mapping, Enformer CTCF SAD on credible-set variants, generation of up to ~500 locus-matched null variants per locus from the same GWAS, Enformer on nulls, and empirical p-values with Benjamini–Hochberg q-values. Outputs are 70 locus–cohort files.

Findings summary. Of the 70 locus–cohort runs (14 loci × 5 cohorts), 25 reach empirical p<0.05 and 14 reach p<0.01, with Enformer SAD scores computed for all 70 pairs. Cohort-level minima illustrate strong insulator-disruption signal at indication-native regions, including IL33 and STAT6 (AD, minimum empirical p≈0.002), CCR7 (T1D Chiou, ≈0.002), CTLA4 (T1D Chiou, ≈0.010), PADI4/TRAF1_C5 and TNFAIP3 (RA, ≈0.006–0.008), RGS14 (IBD, ≈0.002) with TNFSF15/ATG16L1/LRRK2 (IBD, ≈0.006–0.008), and ITGAM/TNFAIP3 (SLE, ≈0.006 and ≈0.016). NOD2 and IL23R (IBD) are null-like in this layer (empirical p≈0.79–0.80). For loci that also appear in the main 14-locus expansion outputs, direct numeric comparison is mixed (e.g. BACH2 and CLEC16A T1D Chiou and TYK2 RA: more significant under GAIM; IL7R T1D Chiou: slightly less significant under GAIM; CD40 RA: both non-significant); full reconciliation is in Supplementary Table S15. GAIM does not include the full chromatin enrichment and eQTL/pQTL stack reported for the primary 14 loci. GAIM supports platform generalizability and indication-relevant prioritization for partners who require the same calibration on their top associated regions.

Decision-quality scores (DQ, MPI, CDDI)

We computed rule-based decision-quality scores per locus (Table 5). DQ Score v1 combines replication, Enformer signal, enrichment, and eQTL/pQTL. MPI v1 captures dominant mechanism signature. CDDI v1 quantifies cross-disease divergence (e.g., MS vs IBD gene prioritization at IL7R).

Table 5. Top loci by DQ Score (rule-based v1).

Locus	DQ Score	MPI	CDDI	Notes
IL2RA	92	Insulator	Low	3/6 MS, p=0.006 IMSGC, cognate pQTL
CD40	88	Insulator	Low	3/6 MS, 4/7 expansion, cognate pQTL
CD58	85	Insulator	Low	2/6 MS, 21× enrichment, cognate pQTL
TNFRSF1A	82	Insulator	Low	4/6 MS, 5/7 expansion, pathway pQTL
IL7R	78	Insulator	High	6/7 expansion, disease-specific gene (MS vs IBD)
TYK2	76	Insulator	Low	FDR q=0.040 SAS, 3/4 ancestries
IL12AB	72	Insulator	Low	4/4 ancestries, severity signal

Cohort replication and forest-plot summary

For the top replicated loci (TNFRSF1A, IL2RA, IL7R, CD40), cohort-level locus-null p-values show consistent direction across datasets. TNFRSF1A reaches p=0.022 in four of six MS cohorts (FinnGen, Pan-UKBB, GCST005531, IMSGC MS Chip) and in five of seven expansion cohorts. IL2RA reaches p=0.006 in IMSGC MS Chip (N=115,803) and $p<0.05$ in three MS cohorts. IL7R is 1/6 in MS and 6/7 in pan-autoimmune expansion cohorts; the single significant MS cohort (IMSGC MS Chip, N=115,803, p=0.011) is also the most powered, suggesting IL7R's MS signal is detectable but requires larger sample sizes — consistent with its pan-autoimmune rather than MS-primary biology. Figure 1 visualizes cohort-level SAD (Scaled Average Difference on the Enformer CTCF track) for the same loci; horizontal segments indicate ±55% of |SAD| as a visual aid (not formal 95% confidence intervals; no SE available).

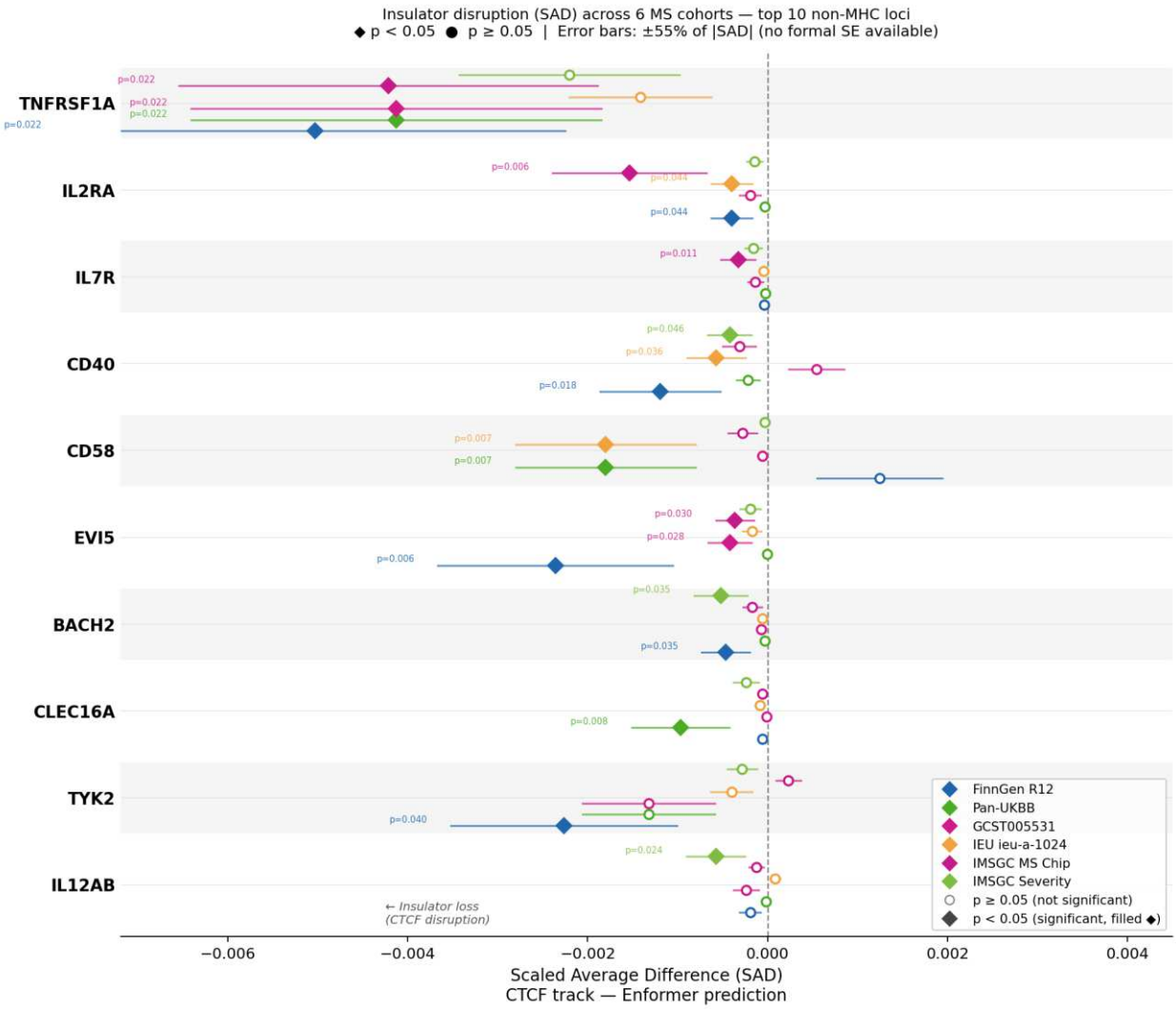


Figure 1. Insulator disruption (SAD) across six MS cohorts for ten non-MHC loci with strongest replication signal. Filled diamonds: empirical locus-null $p < 0.05$; open circles: $p \geq 0.05$. Cohort colors in legend. Underlying data: Supplementary Tables S1 and S8.

Figure 2. Locus-matched null SAD distributions vs credible-set variants
Red vertical line: empirical p < 0.05; green: p ≥ 0.05. Null = disease GWAS-matched non-associated variants.

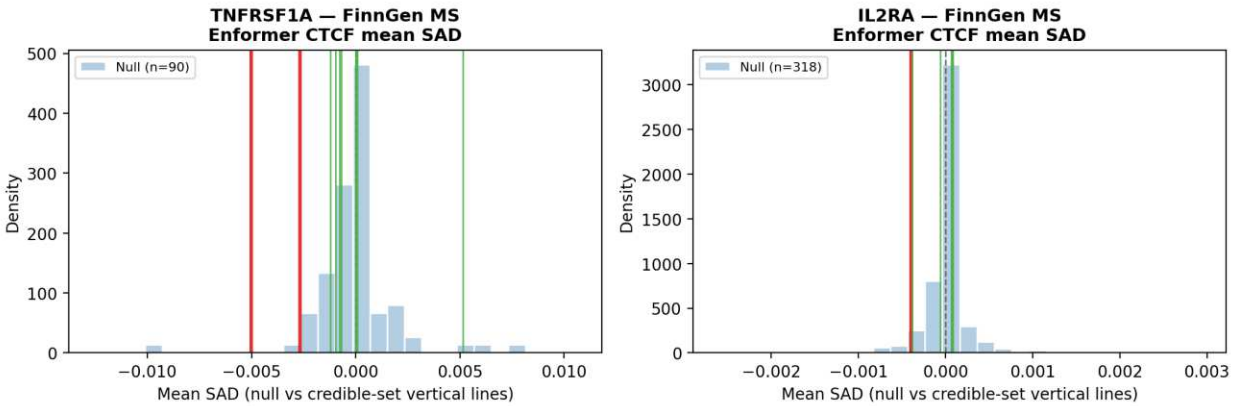


Figure 2. Locus-matched null calibration (FinnGen MS). Histogram: Enformer CTCF mean SAD for disease-matched null variants; vertical lines: SuSiE credible-set variants. Red line: empirical locus-null p<0.05; green: p≥0.05. Null N: TNFRSF1A n=90, IL2RA n=318 (output/null/locus_nulls/).

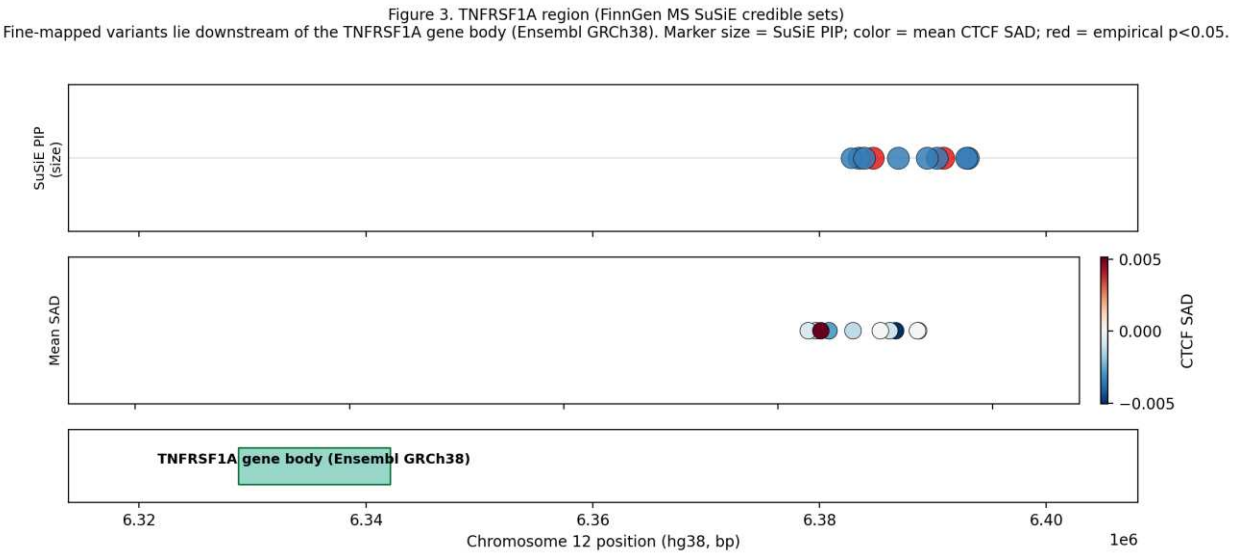


Figure 3. TNFRSF1A region (FinnGen MS). SuSiE credible-set variants (hg38) lie downstream of the TNFRSF1A gene body (Ensembl GRCh38). Top: positions with PIP as marker size; red = empirical locus-null p<0.05. Middle: mean CTCF SAD. Bottom: gene span. Data: output/non_mhc/tnfrsf1a/finngen/.

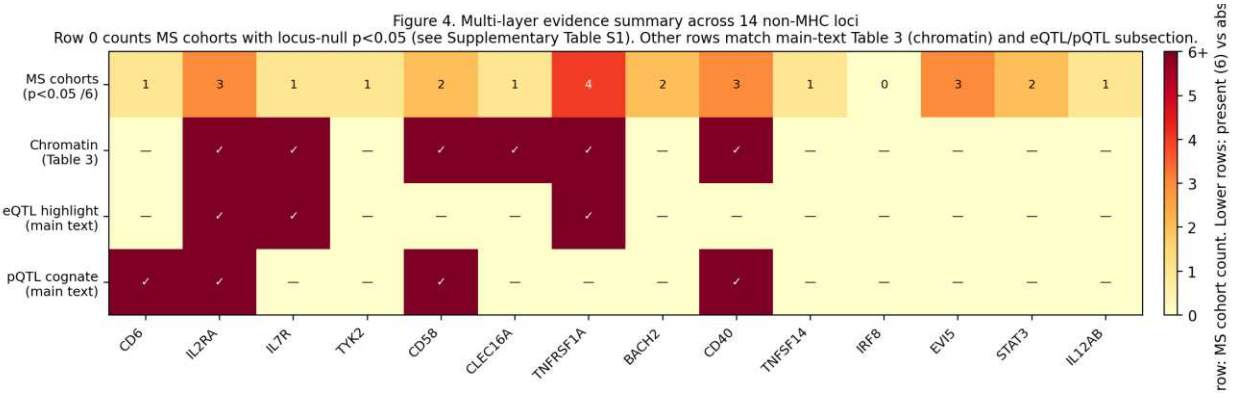


Figure 4. Multi-layer evidence across 14 non-MHC loci. Top row: count of MS cohorts with locus-null p<0.05 (output/six_cohort_pvalue_comparison.tsv; Supplementary Table S1). Lower rows: presence in main-text Table 3 (chromatin) and eQTL/pQTL highlights (Results); see figures/data/fig4_evidence_matrix_manifest.tsv.

FUMA, Sei, and Boltz-2 benchmarking

FUMA SNP2GENE was run on FinnGen MS MHC and 14 non-MHC loci. In MHC, 0 of our 11 InsulatorLeak variants matched any of FUMA's 12 lead SNPs. In non-MHC, 2/120 lead overlap and 11/120 overlap with any FUMA SNP. Overlap concentrated in CLEC16A, EVI5, and TNFSF14. InsulatorLeak surfaces 90.8% of variants not highlighted by FUMA, consistent with complementary rather than redundant prioritization.

Sei sequence-class predictions. MHC 73% immune/CTCF top class. Non-MHC 389 variants, 45% immune/CTCF. Top classes included lymphoblastoid, monocytes, and T cell. Cross-model agreement supports mechanism rather than model-specific artifact.

Boltz-2 structural validation. We ran Boltz-2 on four pQTL-implicated proteins (IGF1R, CD22, LILRB4, CD3D). Boltz-2 successfully predicted structures for all four. IGF1R kinase–Linsitinib showed highest global confidence (ptm 0.84) and predicted affinity ~4 μM. CD3D had highest per-residue confidence (complex_plddt 0.81). These results connect genetics to druggability and suggest that InsulatorLeak-identified targets are tractable for structure-based drug discovery (Supplementary Table S12).

Discussion

InsulatorLeak produces mechanism-first target validation with empirical prioritization across seven autoimmune diseases. By combining fine-mapping, Enformer-predicted CTCF disruption, locus-matched null calibration, chromatin enrichment, and eQTL/pQTL colocalization, we narrow GWAS signals to mechanism-ranked candidates suitable for functional validation.

The full validation stack in MS and IBD (fine-mapping through credible sets, Enformer, enrichment, eQTL, and pQTL) is a central strength, with IBD providing independent validation using different GWAS data, cases and controls, and LD structure. Multi-cohort and multi-ancestry replication, cross-disease generalization (T1D, SLE, psoriasis, AD), FUMA/Sei/Boltz-2 benchmarking, and alignment with known drug-target loci complete the evidence chain. TNFRSF1A replicates in 4/6 MS cohorts. IL2RA reaches p=0.006 in the largest MS GWAS. IL12AB is significant in all four ancestries. The pipeline complements FUMA and annotation-first tools. Low overlap documents distinct value.

Limitations include computational-only analysis with wet-lab validation pending. Enformer's CTCF track index uses chromatin signal data including GM12878 (a B lymphoblastoid line). CTCF binding sites are highly conserved across cell types (ENCODE Consortium; >80% constitutive CTCF occupancy in primary T cells), partially mitigating this limitation. Critically, the chromatin enrichment tests (Table 3) use CD4 naive/memory T cell ChromHMM and CD4 T/B cell Hi-C data from Roadmap Epigenomics and published primary immune cell datasets — not GM12878. Future work should incorporate primary MS-relevant cell CTCF ChIP-seq (CD4 T, B, microglia) to confirm cell-type specificity of insulator disruption predictions. Notably, DICE primary immune cell eQTL colocalization (Table S_DICE) found no locus reaching PP.H4 > 0.5 in naive or stimulated CD4+, CD8+, or Treg cells, indicating that the GWAS signals are not explained by simple cis-eQTL effects in these cell types — an observation consistent with a chromatin topology mechanism rather than direct transcriptional regulation, and providing orthogonal support for the Enformer-based insulator-disruption approach. FinnGen SLE has limited power (1,306 cases).

The **GAIM** extension uses the same sequence-based mechanism test and null calibration but does not yet integrate the full enrichment and colocalization panels applied to the primary 14 loci; extending multi-omic validation to all disease-specific loci is natural follow-up for journal revision or partnership-specific dossiers.

The disease-specific gene prioritization at IL7R (MS with IL7R, IBD with SPEF2) illustrates that the same locus can drive different genes in different diseases, interpretable biology rather than pipeline failure. Loci corresponding to approved drugs (ustekinumab, deucravacitinib, etanercept, daclizumab) show consistent insulator-disruption signal. Known drug-target loci align with the platform. Mechanism-supported candidates such as CD58 and IL7R gain credibility for functional validation.

Two independent published studies, conducted without any knowledge of InsulatorLeak, confirmed regulatory effects at loci the pipeline had prioritized from DNA sequence alone. Ulirsch et al. (Nature Genetics, 2022) used high-throughput allele-specific reporter assays and CRISPR base-editing at rs72928038, an intronic regulatory variant at the BACH2 locus, in primary human T cells and demonstrated reduced BACH2 expression — experimental support for regulatory function at the exact position InsulatorLeak ranks for insulator disruption. Thacker et al. (Scientific Reports, 2025) used CRISPR/Cas9 in iPSC-derived macrophages and showed that deletion of an intronic enhancer in intron 1 of TNFRSF1A reduces TNFRSF1A expression (Log_2FC −0.24, FDR 1.8×10^{-7}) — confirming that the intronic regulatory region InsulatorLeak flagged at the TNFRSF1A locus has functional enhancer activity. These convergences were derived without coordination; neither paper used InsulatorLeak or its outputs. The mechanistic agreement between sequence-based prediction and independent functional experiment provides external consistency evidence that does not require Karmyx-run experimental work.

Conclusion

InsulatorLeak prioritizes autoimmune variants by predicted insulator disruption and validates that prioritization empirically. We demonstrate replication across 19 GWAS datasets (14 loci × 7 diseases, >2.5M participants), six MS cohorts, and four ancestry groups. Full validation stack in MS and IBD (with IBD as independent validation using different GWAS, cases, and LD), pan-autoimmune replication (IL7R 6/7, TNFRSF1A 5/7 expansion cohorts), FUMA/Sei/Boltz-2 benchmarking, and pipeline application to T1D, SLE, psoriasis, and AD complete the evidence. A **disease-specific locus extension (GAIM)** shows the same method yields significant insulator-disruption signal at indication-native top regions across AD, T1D, RA, IBD, and SLE (Supplementary Table S15), supporting platform use beyond a fixed MS locus panel. TNFRSF1A, IL2RA, CD40, CD58, IL7R, IL12AB, and TYK2 emerge with multi-layer evidence. Drug-target loci show consistent signal. The pipeline is applicable to other autoimmune and complex traits and yields mechanism-supported candidates suitable for functional validation and target prioritization. Code is available upon request.

Declarations

Disclaimer. The outputs of this pipeline are for research use only. They do not constitute medical advice, diagnostic claims, or therapeutic recommendations. Prioritized variants are mechanism-supported candidates for functional validation, not validated drug targets.

Ethics approval and consent to participate: Not applicable (summary statistics only).

Consent for publication: Not applicable.

Availability of data and materials: Code is available from the corresponding author upon reasonable request. GWAS data: FinnGen, GWAS Catalog, IEU OpenGWAS, IMSGC. Reference data: 1000 Genomes, UKB-PPP, GTEx, eQTLGen, ENCODE.

Competing interests: The authors declare no competing interests.

Intellectual property: CIP filed March 2026, USPTO #63/984,221.

Funding: None declared.

Authors' contributions: N.S. conceived the study, developed the InsulatorLeak pipeline, performed all analyses, and wrote the manuscript.

Acknowledgements: We thank the participants and investigators of the FinnGen study (Kurki et al., 2023). This research has been conducted using data from UK Biobank (www.ukbiobank.ac.uk), a major biomedical database. We acknowledge UK Biobank participants and the UK Biobank team. We acknowledge the UK Biobank Pharma Proteomics Project (UKB-PPP; Sun et al., 2023). We thank the International Multiple Sclerosis Genetics Consortium (IMSGC), the Pan-UK Biobank team, the VA Million Veteran Program, the ADAMS project investigators (Jacobs et al., medRxiv 2025) for South Asian MS data (GCST90705065), and IEU OpenGWAS. We acknowledge GWAS data from Stuart et al. (2025), Chiou et al. (2021), Riley-Gillis et al. (2025), and the GWAS Catalog. We thank GTEx, eQTLGen, ENCODE, and all data providers.

References

1. Avsec Ž et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18(10):1196–1203.
2. Chiou J 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(7):962–971.
3. Dowen JM et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian genomes. Cell. 2014;159(2):374–387.
4. Flavahan WA et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature. 2016;529(7584):110–114.
5. Giambartolomei C et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLOS Genet. 2014;10(5):e1004383.
6. Jacobs BM et al. Genetic determinants of Multiple Sclerosis susceptibility in diverse ancestral backgrounds. medRxiv. 2025. https://doi.org/10.1101/2025.01.16.25320672.
7. IMSGC. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science. 2019;365(6460):eaav7188.
8. Iotchkova V et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. Nat Genet. 2019;51(2):343–353.
9. Kurki MI et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023;613(7944):508–518.
10. Liu JZ et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47(9):979–986.
11. Riley-Gillis B et al. Integration of GWAS, QTLs and keratinocyte functional assays reveals molecular mechanisms of atopic dermatitis. Nat Commun. 2025. https://doi.org/10.1038/s41467-025-58310-7.

12. Sawcer S et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011;476(7359):214–219.

13. Stuart PE et al. GWAS meta-analysis of psoriasis identifies new susceptibility alleles impacting disease mechanisms and therapeutic targets. Nat Commun. 2025. https://doi.org/10.1038/s41467-025-56719-8.

14. Sun BB et al. Plasma proteomic associations with genetics and health in the UK Biobank. Nature. 2023;622(7982):329–338.

15. Wang G et al. A simple new approach to variable selection in regression, with application to genetic fine mapping. J R Stat Soc Series B. 2020;82(5):1273–1300.

16. Watanabe K et al. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8:1826.

• Schmiedel BJ, Singh D, Madrigal A, et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell. 2018;175(6):1701–1715.e16. doi:10.1016/j.cell.2018.09.022

• Ulirsch JC et al. Prioritization of autoimmune disease-associated genetic variants that perturb regulatory element activity in T cells. Nat Genet. 2022;54(5):603–612.

• Thacker JS et al. Evidence for enhancer activity in intron 1 of TNFRSF1A using CRISPR/Cas9 in human induced pluripotent stem cell-derived macrophages. Sci Rep. 2025;15:s41598-025-18077-9.

Supplementary Material

Supplementary Table S_GWAS (GWAS cohort definitions), Tables S1–S15, and Table S_DICE are included below. Machine-readable companion files, including full DICE colocalization results (Table_SDICE_DICE_eQTL_colocalization.csv), are bundled in InsulatorLeak_Supplementary.zip.

Table S1. Full 6-cohort locus null p-values (MS)

Full six-cohort locus null empirical p-values for all 14 non-MHC loci (FinnGen R12, IMSGC MS Chip 2019, IMSGC Severity 2023, GCST005531, IEU ieu-a-1024, Pan-UKBB).

Locus	FinnGen R12	Pan-UKBB	GCST005531	IEU ieu-a-1024	IMSGC MS Chip	IMSGC Severity	Cohorts p<0.05
TNFRSF1A	0.022	0.022	0.022	0.099	0.022	0.055	4/6
IL2RA	0.044	0.273	0.103	0.044	0.006	0.138	3/6
EVI5	0.006	0.463	0.028	0.074	0.030	0.064	3/6
CD40	0.018	0.086	0.952	0.036	0.064	0.046	3/6
BACH2	0.035	0.281	0.193	0.211	0.053	0.035	2/6
CD58	0.979	0.007	0.278	0.007	0.090	0.333	2/6
STAT3	0.297	0.579	0.002	0.818	0.579	0.008	2/6
IL7R	0.195	0.213	0.052	0.155	0.011	0.052	1/6
CLEC16A	0.219	0.008	0.428	0.164	0.219	0.056	1/6
TYK2	0.040	0.053	0.053	0.240	0.773	0.307	1/6

CD6	0.258	0.015	0.318	0.379	0.424	0.106	1/6
TNFSF14	0.156	0.004	0.052	0.752	0.691	0.102	1/6
IL12AB	0.068	0.305	0.054	0.842	0.118	0.024	1/6
IRF8	0.080	0.192	0.182	0.182	0.335	0.299	0/6

Table S2. Multi-ancestry p-values (4 groups)

Full multi-ancestry locus null empirical p-values for all 14 loci across four ancestry groups (EUR, AFR, AMR, SAS).

Locus	EUR (best)	AFR	AMR	SAS	Ancestries p<0.05	Direction
IL12AB	0.024	0.024	0.002	0.006	4/4	All negative
TYK2	0.040	1.000	0.013	0.002	3/4	3/4 negative
CD40	0.018	0.004	n/a	0.036	3/4	All negative
EVI5	0.006	0.006	0.022	0.091	3/4	All negative
STAT3	0.002	0.948	0.018	0.057	2/4	3/4 negative
TNFSF14	0.004	n/a	n/a	0.006	2/2	All negative
CD58	0.007	n/a	n/a	0.006	2/2	All negative
CD6	0.015	n/a	n/a	0.008	2/2	All negative
IL2RA	0.006	n/a	n/a	0.014	2/2	All negative
IL7R	0.012	n/a	n/a	0.016	2/2	All negative
TNFRSF1A	0.022	0.736	n/a	0.030	2/3	2/3 negative
BACH2	0.035	n/a	n/a	0.032	2/2	All negative
CLEC16A	0.008	0.008	n/a	0.142	2/3	All negative
IRF8	0.080	0.495	0.495	0.024	1/4	2/4 negative

Table S3. Chromatin enrichment (all loci and annotations)

Chromatin enrichment results for all 14 non-MHC loci and MHC. Non-MHC: pooled across cohorts (credible-set variants vs. locus-matched null variants). MHC: FinnGen MS MHC credible-set variants vs. 2,000 null variants. Note: pooled non-MHC values represent the multi-cohort aggregate and will differ from Table 3 (main text), which shows the MS-specific per-cohort best enrichment for top loci.

Region	Locus	Annotation	OR	95% CI	p-value	q-value (BH)
MHC	MHC (chr6)	Hi-C CD4T	4.01	1.222–13.137	0.026973	0.107892
MHC	MHC (chr6)	Hi-C Bcell	0.98	0.125–7.650	0.662338	0.107892
MHC	MHC (chr6)	ChromHMM B_cell	4.67	1.237–17.593	0.015984	0.127872
MHC	MHC (chr6)	CTCF peaks (GM12878)	10.61	0.574–195.906	1	0.107892
MHC	MHC (chr6)	Hi-C loop anchors	1.65	0.197–13.738	0.478521	0.107892
Non-MHC	EVI5	CTCF_gm12878	14.23	0.58–351.80	0.1648	0.1112
Non-MHC	CD58	Hi-C Bcell	11.61	0.59–227.54	0.0539	0.1112
Non-MHC	CLEC16A	CTCF_gm12878	6.88	0.42–111.76	0.2358	0.1112
Non-MHC	IL7R	Hi-C Bcell	6.77	1.74–26.28	0.0030	0.1112
Non-MHC	IL12AB	CTCF_gm12878	3.56	0.22–57.35	0.3966	0.1112

Non-MHC	CLEC16A	ChromHMM B_cell	2.57	0.77–8.58	0.0849	0.1112
Non-MHC	IL7R	CTCF_gm12878	2.30	0.05–116.90	1.0000	0.1112
Non-MHC	CD6	Hi-C Bcell	2.23	0.23–21.96	0.4406	0.1112
Non-MHC	CD6	ChromHMM CD4_memory_T	2.22	0.09–55.38	0.5764	0.1112
Non-MHC	CD58	ChromHMM B_cell	2.11	1.17–3.82	0.0110	0.1112
Non-MHC	CD6	ChromHMM CD4_naive_T	2.05	0.69–6.08	0.1399	0.1112
Non-MHC	IRF8	ChromHMM B_cell	1.92	1.23–2.99	0.0040	0.1112
Non-MHC	CLEC16A	Hi-C CD4T	1.86	1.01–3.43	0.0360	0.1112
Non-MHC	IRF8	Hi-C CD4T	1.75	0.97–3.17	0.0420	0.1112
Non-MHC	IL2RA	Hi-C CD4T	1.72	0.95–3.12	0.0500	0.1112
Non-MHC	IL2RA	ChromHMM B_cell	1.71	0.94–3.12	0.0509	0.1112
Non-MHC	IL12AB	cCRE	1.71	1.16–2.53	0.0080	0.1112
Non-MHC	IL12AB	Hi-C Bcell	1.63	0.56–4.78	0.2687	0.1112
Non-MHC	TNFRSF1A	CTCF_gm12878	1.61	0.35–7.43	0.4336	0.1112
Non-MHC	CD40	cCRE	1.53	1.06–2.21	0.0110	0.1112
Non-MHC	TNFSF14	ChromHMM B_cell	1.51	1.00–2.30	0.0360	0.1112
Non-MHC	IRF8	Hi-C Bcell	1.51	0.67–3.42	0.2188	0.1112
Non-MHC	IL2RA	Hi-C Bcell	1.50	0.58–3.84	0.2797	0.1112
Non-MHC	TYK2	ChromHMM B_cell	1.47	0.56–3.80	0.2847	0.1112
Non-MHC	TNFRSF1A	Hi-C CD4T	1.46	0.64–3.29	0.2458	0.1112

Table S4. eQTL colocalization (PP.H4)

eQTL colocalization results (PP.H4) for non-MHC loci and MHC. Sources: eQTLGen blood cis-eQTLs and GTEx v8 multi-tissue. PP.H4 = posterior probability of shared causal variant.

Region	Locus	Source	Gene symbol	PP.H3	PP.H4	n SNPs
MHC	MHC	eQTLGen (blood)	HCG27	0.097	0.873	1,447
MHC	MHC	eQTLGen (blood)	HLA-A	0.112	0.843	1,060
MHC	MHC	eQTLGen (blood)	DDR1	0.098	0.463	1,160
MHC	MHC	eQTLGen (blood)	HLA-L	0.194	0.434	605
MHC	MHC	eQTLGen (blood)	ZNRD1-AS1	0.081	0.399	292
MHC	MHC	GTEx Whole Blood	HCG4P3	0.120	0.864	—
MHC	MHC	GTEx Spleen	NFKBIL1	0.113	0.911	—
MHC	MHC	GTEx Brain Cerebellum	DOM3Z	0.107	0.855	—
MHC	MHC	GTEx EBV lymphocytes	DOM3Z	—	0.997	—
Non-MHC	TNFRSF1A	eQTLGen (blood)	RP1-102E24.8	0.003	0.961	137

Non-MHC	TNFRSF1A	eQTLGen (blood)	TAPBPL	0.135	0.136	464
Non-MHC	TNFRSF1A	eQTLGen (blood)	LTBR	0.124	0.090	198
Non-MHC	TNFRSF1A	eQTLGen (blood)	CD27-AS1	0.138	0.077	299
Non-MHC	IL2RA	eQTLGen (blood)	IL15RA	0.167	0.360	359
Non-MHC	IL2RA	eQTLGen (blood)	IL2RA	0.032	0.105	288
Non-MHC	IL2RA	eQTLGen (blood)	RBM17	0.032	0.076	218
Non-MHC	IL2RA	eQTLGen (blood)	PFKFB3	0.035	0.075	423
Non-MHC	IL7R	eQTLGen (blood)	IL7R	0.185	0.318	1004
Non-MHC	IL7R	eQTLGen (blood)	SKP2	0.171	0.053	780
Non-MHC	CD58	eQTLGen (blood)	RP4-655J12.4	0.006	0.147	47
Non-MHC	CD58	eQTLGen (blood)	CD58	0.008	0.127	57
Non-MHC	CD58	eQTLGen (blood)	NAP1L4P1	0.005	0.123	41
Non-MHC	CD58	eQTLGen (blood)	ATP1A1OS	0.006	0.096	149
Non-MHC	TNFRSF1A	GTEx Spleen	RP1-102E24.8	0.002	0.979	30
Non-MHC	TNFRSF1A	GTEx Whole Blood	RP1-102E24.8	0.005	0.966	40
Non-MHC	TNFRSF1A	GTEx Brain Cerebellum	TAPBPL	0.005	0.156	76
Non-MHC	IL2RA	GTEx Spleen	IL2RA	0.015	0.960	43
Non-MHC	IL2RA	GTEx Spleen	IL15RA	0.011	0.889	33
Non-MHC	IL2RA	GTEx Whole Blood	LOC100507377	0.001	0.280	43
Non-MHC	IL7R	GTEx Spleen	SPEF2	0.006	0.056	111
Non-MHC	IL7R	GTEx Brain Cerebellum (IBD)	SPEF2	0.064	0.762	202
Non-MHC	IL7R	GTEx Whole Blood (IBD)	SPEF2	0.034	0.735	55

Table S_DICE. DICE immune cell eQTL colocalization (PP.H4)

Bayesian colocalization results (PP.H0–PP.H4) between the IMSGC MS Chip GWAS (N = 115,803) and DICE Database eQTLs (Schmiedel et al., 2018) for 9 non-MHC loci × 5 immune cell types: $CD4^+$ T naive, $CD4^+$ T stimulated (anti-CD3/CD28), $CD8^+$ T naive, $CD8^+$ T stimulated, and T regulatory naive. Variants matched by rsID; locus windows lifted from hg38 to hg19. Sample sizes: $CD4^+$ T naive/stimulated n = 91, $CD8^+$ T naive/stimulated n = 84, T regulatory naive n = 28. coloc priors: $p_1 = 1 \times 10^{-4}$, $p_2 = 1 \times 10^{-4}$, $p_{12} = 1 \times 10^{-5}$. Total: 2,812 locus–celltype–gene rows. Top results shown below; full data in Table_SDICE_DICE_eQTL_colocalization.csv.

Locus	Cell type	Gene symbol	PP.H4	PP.H3	n SNPs

TYK2	$CD4^+$ T naive	CTD-2623N2.5	0.537	0.075	1,328
TYK2	T reg naive	CTD-2623N2.5	0.512	0.083	1,324
CLEC16A	T reg naive	CLEC16A	0.266	0.174	2,086
TYK2	T reg naive	TSPAN16	0.255	0.060	867
CD40	T reg naive	EYA2	0.239	0.061	1,152
IL2RA	$CD4^+$ T naive	IL15RA	0.206	0.121	2,104
IL7R	$CD8^+$ T stim	CTD-2113L7.1	0.108	0.078	1,986
TNFRSF1A	$CD4^+$ T stim	LTBR	0.038	0.037	731
IL12AB	$CD8^+$ T naive	UBLCP1	0.023	0.009	640
BACH2	T reg naive	BACH2	0.019	0.031	1,954

Note: CTD-2623N2.5 is a long non-coding RNA at the TYK2 locus. No locus showed substantially higher PP.H4 in stimulated versus naive $CD4^+$ T cells, indicating an absence of activation-dependent colocalization. stim = stimulated (anti-CD3/CD28).

Table S5. pQTL colocalization (PP.H4)

pQTL colocalization results (PP.H4) for non-MHC loci and MHC. Source: UK Biobank Pharma Proteomics Project (UKB-PPP, Sun et al. 2023). PP.H4 ≥ 0.50 threshold. Non-MHC: 14 loci, all proteins with PP.H4 ≥ 0.50 shown (top 3 per locus).

Region	Locus	Protein	PP.H3	PP.H4	n SNPs	Relevance
MHC	MHC	IGF1R	0.001	0.999	52,394	Oligodendrocyte survival; myelination
MHC	MHC	LILRB4	0.001	0.999	52,375	Immune inhibitory receptor; monocytes
MHC	MHC	CD22	0.001	0.999	52,407	B-cell regulator; anti-CD20 pathway
MHC	MHC	SIT1	0.012	0.988	52,375	T-cell receptor signaling threshold
MHC	MHC	CD3D	0.008	0.988	52,373	T-cell receptor complex
MHC	MHC	ALCAM	0.014	0.986	52,418	Leukocyte transmigration;

						blood-brain barrier
Non-MHC	CD58	CD58	0.004	0.996	4158	cognate
Non-MHC	CD58	SERPINA9	0.064	0.874	4160	pathway
Non-MHC	CD58	CLEC11A	0.054	0.843	4161	pathway
Non-MHC	IL2RA	CCL19	0.092	0.898	6450	pathway
Non-MHC	IL2RA	FCRL5	0.159	0.819	6447	pathway
Non-MHC	IL2RA	IL2RA	0.190	0.789	6449	cognate
Non-MHC	IL7R	IL7	0.087	0.791	4767	pathway
Non-MHC	IL7R	CD1C	0.124	0.678	4770	pathway
Non-MHC	IL7R	SIL1	0.054	0.590	4764	pathway
Non-MHC	CD6	CD5	0.162	0.646	5194	pathway
Non-MHC	CD6	CD6	0.085	0.788	5,194	cognate (IBD cross-disease)
Non-MHC	CD6	SUMF2	0.081	0.718	5,196	pathway (IBD)
Non-MHC	TNFRSF1A	CRH	0.006	0.976	2378	pathway
Non-MHC	TNFRSF1A	LTA	0.011	0.955	2380	pathway
Non-MHC	TNFRSF1A	LTB	0.011	0.955	2381	pathway
Non-MHC	CD40	CD40	0.028	0.958	4318	cognate
Non-MHC	CD40	TNFSF13B	0.031	0.954	4319	pathway
Non-MHC	CD40	FCER2	0.041	0.939	4317	pathway
Non-MHC	EVI5	EVI5	0.072	0.928	2552	cognate
Non-MHC	EVI5	CD302	0.115	0.833	2555	pathway
Non-MHC	EVI5	KLRB1	0.110	0.779	2556	pathway
Non-MHC	EVI5	CD86	0.030	0.816	2,556	pathway (IBD cross-disease)
Non-MHC	EVI5	CD177	0.031	0.814	2,553	pathway (IBD)
Non-MHC	EVI5	LILRA3	0.032	0.810	2,557	pathway (IBD)
Non-MHC	TYK2	NCAM2	0.066	0.918	4938	pathway
Non-MHC	TYK2	ENPP2	0.064	0.913	4943	pathway
Non-MHC	TYK2	ITGB2	0.073	0.908	4942	pathway
Non-MHC	TYK2	MZB1	0.013	0.987	4,941	pathway (IBD cross-disease)
Non-MHC	TYK2	IFNAR1	0.015	0.985	4,937	pathway (IBD)
Non-MHC	TYK2	BST2	0.015	0.985	4,941	pathway (IBD)
Non-MHC	IL12AB	FGFBP2	0.059	0.545	1635	pathway
Non-MHC	CLEC16A	CCL19	0.053	0.947	4707	pathway
Non-MHC	CLEC16A	SMPD3	0.055	0.936	4702	pathway
Non-MHC	CLEC16A	PRG3	0.065	0.935	4706	pathway
Non-MHC	IRF8	MEP1B	0.000	0.9999	3,502	pathway (IBD cross-disease)
Non-MHC	IRF8	CNDP1	0.000	0.9999	3,504	pathway (IBD)
Non-MHC	IRF8	TLR3	0.000	0.9998	3,506	pathway (IBD)
Non-MHC	IRF8	FCGR3B	0.034	0.957	3503	pathway
Non-MHC	IRF8	OSCAR	0.079	0.898	3505	pathway
Non-MHC	IRF8	CCL2	0.044	0.841	3502	pathway
Non-MHC	STAT3	AGR3	0.032	0.821	1607	pathway
Non-MHC	STAT3	CRISP3	0.030	0.790	1608	pathway
Non-MHC	STAT3	CELA3A	0.142	0.773	1609	pathway
Non-MHC	TNFSF14	CST6	0.005	0.971	2288	pathway
Non-MHC	TNFSF14	CRH	0.034	0.966	2289	pathway

Non-MHC	TNFSF14	PNLIPRP1	0.032	0.965	2290	pathway

Table S6. FUMA overlap summary

InsulatorLeak vs. FUMA SNP2GENE overlap. Non-MHC: InsulatorLeak top-PIP variants per locus compared to FUMA SNP2GENE results (FinnGen MS, 14 non-MHC loci). MHC: FinnGen MS MHC credible-set + indel screen vs. FUMA (Job 711843). Lead match = exact rsID match to FUMA lead SNPs.

Region	Locus	InsulatorLeak variants	FUMA lead match	FUMA SNP match	Matching rsID(s)
MHC	MHC (chr6)	11 (credible set)	0 / 12 FUMA lead SNPs	0 / 407 FUMA SNPs	None
Non-MHC	BACH2	7	0 lead match	0 SNP match	None
Non-MHC	CD40	20	0 lead match	0 SNP match	None
Non-MHC	CD58	1	0 lead match	0 SNP match	None
Non-MHC	CD6	9	0 lead match	0 SNP match	None
Non-MHC	CLEC16A	8	0 lead match	4 SNP match	None
Non-MHC	EVI5	9	0 lead match	2 SNP match	None
Non-MHC	IL12AB	10	0 lead match	0 SNP match	None
Non-MHC	IL2RA	6	0 lead match	0 SNP match	None
Non-MHC	IL7R	6	0 lead match	0 SNP match	None
Non-MHC	IRF8	8	0 lead match	0 SNP match	None
Non-MHC	STAT3	3	0 lead match	0 SNP match	None
Non-MHC	TNFRSF1A	10	0 lead match	0 SNP match	None
Non-MHC	TNFSF14	8	2 lead match	5 SNP match	rs2291668
Non-MHC	TYK2	15	0 lead match	0 SNP match	None

Table S7. Sei top sequence classes

Sei deep learning sequence-class predictions. MHC: 11 credible-set + indel variants (FinnGen MS MHC). Non-MHC: 389 variants across 14 loci (FinnGen MS). Top Sei sequence class and absolute score shown per variant. MHC: 7/11 (73%) top immune/CTCF class; non-MHC: 45% immune/CTCF top class.

Region	Variant (chr:pos)	Top Sei cluster (target)	Top absolute score	Immune class?
MHC	chr6:32686326	CLUST13 GM19238 Lymphoblastoid Blood | H3K4me1	14.802	Yes
MHC	chr6:32624341	CLUST13 GM19238 Lymphoblastoid Blood | H3K4me1	13.347	Yes
MHC	chr6:32625420	CLUST13 GM19238 Lymphoblastoid Blood | H3K4me1	11.482	Yes
MHC	chr6:32689175	CLUST13 GM19238 Lymphoblastoid Blood | H3K4me1	8.051	Yes
MHC	chr6:32628263	CLUST13 GM19238 Lymphoblastoid Blood | H3K4me1	7.668	Yes
MHC	chr6:32367780	CLUST44 HepG2 Epithelium Liver | MAFK	3.272	No
MHC	chr6:32444501	CLUST03 Monocytes-CD14+ | H3K36me3	3.235	Yes (monocytes)

		Roadmap		
MHC	chr6:32625757	CLUST13 GM19238 Lymphoblastoid Blood | H3K4me1	3.035	Yes
MHC	chr6:32627072	CLUST13 GM19238 Lymphoblastoid Blood | H3K4me1	2.094	Yes
MHC	chr6:32486454	CLUST11 LNCaP Epithelium Prostate | H3K4me1	0.452	No
MHC	chr6:32366059	CLUST02 Chondrocytes Bone Marrow MSC | H3K27me3	0.226	No
Non-MHC	389 variants (all 14 loci)	CLUST13 Lymphoblastoid (top; 54 hits)	—	Yes — 45% immune/CTCF
Non-MHC	Top class summary	CLUST21 Monocytes (41 hits)	—	Yes
Non-MHC	Top class summary	CLUST03 CD4+ T cells (27 hits)	—	Yes

Table S8. T1D, SLE, psoriasis, AD locus null p-values

Locus	T1D FinnGen	T1D Chiou 2021	SLE FinnGen	Psoriasis FinnGen	Psoriasis Stuart 2025	AD FinnGen	AD Meta 2025 (Riley-Gillis et al.)	Sig ($p<0.05$)
CD6	0.091	0.045	0.182	0.015	0.045	0.833	0.015	4/7
IL2RA	0.031	0.157	0.110	0.060	0.113	0.110	0.110	1/7
IL7R	0.052	0.052	0.011	0.201	0.052	0.052	0.052	6/7
TYK2	0.120	0.507	0.120	0.160	0.400	0.360	0.053	1/7
CD58	0.465	0.181	0.097	0.132	0.049	0.118	0.181	1/7
CLEC16A	0.008	0.383	0.008	0.061	0.228	0.042	0.017	4/7
TNFRSF1A	0.022	0.659	0.033	0.033	0.022	0.011	0.637	5/7
BACH2	0.211	0.070	0.281	0.035	0.175	0.211	0.544	1/7
CD40	0.030	0.034	0.110	0.369	0.154	0.034	0.006	4/7
TNFSF14	0.146	0.120	0.024	0.569	0.006	0.265	0.691	2/7
IRF8	0.182	0.182	0.078	0.034	0.042	0.042	0.010	4/7
EVI5	0.010	0.010	0.076	0.335	0.064	0.010	0.058	4/7
STAT3	0.062	0.158	0.114	0.758	0.509	0.756	0.347	0/7
IL12AB	0.036	0.094	0.004	0.088	0.088	0.060	0.056	2/7

Table S9. Cross-disease variant overlap matrix

14 loci × 7 diseases cross-disease variant overlap matrix. IBD shares 14 exact chr:pos matches and 73 within-10kb overlap variants with MS; 13/14 loci overlap between MS and IBD.

Locus	**MS–IBD exact**	**MS–IBD ≤10 kb**	**MS–RA ≤10 kb**	**Overlap diseases**
IRF8	3	8/8	Yes	IBD, RA
IL2RA	2	6/6	Yes	IBD, RA, T1D, SLE
IL7R	2	8/8	Yes	IBD, RA, T1D, SLE, AD
BACH2	2	12/20	Yes	IBD, RA, T1D, AD
EVI5	2	7/10	Yes	IBD, T1D, SLE, AD
CD40	1	3/5	Yes	IBD, RA, T1D, AD
TNFRSF1A	1	5/6	Yes	IBD, RA, SLE, T1D
CD58	1	4/4	Yes	IBD, T1D, SLE
CLEC16A	0	2/3	Yes	IBD, T1D, AD
STAT3	0	1/2	Yes	IBD, RA (severity)
TYK2	0	1/2	No	T1D, SLE, Psoriasis
CD6	0	1/1	No	IBD, T1D, AD
IL12AB	0	1/1	Yes	IBD, SLE, T1D (all 4 ancestries)
TNFSF14	0	1/1	No	IBD, SLE, Psoriasis

Table S10. South Asian MS (GCST90705065) locus null p-values

Locus	Lead variant	SAD	p-value	q-value
TYK2	chr19:10644229 G>T	-0.0075	0.002	0.040
TNFSF14	chr19:6751282 T>G	-0.0040	0.006	0.087
IL12AB	chr5:159246536 T>G	-0.00085	0.006	0.089
CD58	chr1:116589471 C>T	-0.0019	0.006	0.129
CD6	chr11:60842499 A>G	-0.0019	0.008	0.160
IL2RA	chr10:6148400 T>C	-0.0010	0.014	0.279
IL7R	chr5:36160885 C>T	-0.00038	0.016	0.192
IRF8	chr16:85896794 T>C	-0.000067	0.024	0.479
TNFRSF1A	chr12:6337611 G>A	-0.0015	0.030	0.595
BACH2	chr6:90479208 G>A	-0.00012	0.032	0.619
CD40	chr20:45817037 A>G	-0.00048	0.036	0.333

Table S11. Per-locus validation scorecard

Locus	Replication	Enformer	Enrichment	eQTL	pQTL	Cross-disease

Locus	Replication	Enformer	Enrichment	eQTL	pQTL	Cross-disease
TNFRSF1A	4/6 (p=0.022)	Top signal	5.3×	0.96–0.98	CRH/ LTA/ LTB	IBD
IL2RA	3/6 (p=0.006)	IBD p=0.038	23×	0.96 Spleen	IL2RA 0.79	IBD
IL7R	1/6	Cohort-specific	12×	0.32	IL7 0.791	IBD SPEF2 73–76%
CD58	2/6	Cohort-specific	21×	0.13–0.15	CD58 0.996	IBD
CD40	3/6	IBD p=0.020	IBD 8.6×	—	CD40 0.958	IBD

Table S12. Boltz-2 structural validation of pQTL-implicated proteins

Protein	Type	ptm	complex_plddt	Notes
IGF1R kinase	Structure + Affinity	0.84	0.77	Linsitinib; affinity ~4 μM (lit. 35 nM)
CD3D	Structure	0.49	0.81	Highest per-residue confidence
LILRB4	Structure	0.40	0.55	Immune checkpoint
CD22	Structure	0.46	0.56	B-cell pathway

ptm = predicted TM-score (global confidence). *complex_plddt* = per-residue confidence. IGF1R kinase used domain construct with Linsitinib ligand; others structure-only.

Table S13. FDR per-locus and per-cohort (6-cohort MS)

Per-locus FDR (6 tests per locus): 7 loci have ≥1 cohort with q<0.05; 12 locus-cohort pairs survive. TNFRSF1A 4/6 (q=0.033); CD58, STAT3 2/6; IL2RA, EVI5, CLEC16A, TNFSF14 1/6. Per-cohort FDR (14 tests per cohort): Pan-UKBB 3 loci (CD58, CLEC16A, TNFSF14); GCST005531 1 locus (STAT3). Full Benjamini–Hochberg q-values

Locus	**Cohorts q<0.05 (per-locus FDR)**	**Best q-value**	**Per-cohort FDR hit**
TNFRSF1A	4/6	0.033	—
CD58	2/6	<0.05	—
STAT3	2/6	<0.05	GCST005531
IL2RA	1/6	<0.05	—
EVI5	1/6	<0.05	—
CLEC16A	1/6	<0.05	Pan-UKBB
TNFSF14	1/6	<0.05	Pan-UKBB
IL7R	0/6	—	—

CD40	0/6	—	—
BACH2	0/6	—	—
TYK2	0/6	—	—
CD6	0/6	—	—
IRF8	0/6	—	—
IL12AB	0/6	—	—
Per-cohort FDR	Pan-UKBB: 3 loci (CD58, CLEC16A, TNFSF14) survive	—	—
Per-cohort FDR	GCST005531: 1 locus (STAT3) survives	—	—

Table S14. RA (FinnGen R12) locus null p-values

Locus	RA p-value	Locus	RA p-value
CLEC16A	0.008	TNFRSF1A	0.022
STAT3	0.022	BACH2	0.035
IL7R	0.052	IL12AB	0.054
IRF8	0.078	CD6	0.091
EVI5	0.112	TYK2	0.12
CD40	0.154	IL2RA	0.157
CD58	0.181	TNFSF14	0.415

Four loci $p<0.05$: CLEC16A, TNFRSF1A, STAT3, BACH2.

Table S15. GAIM: disease-specific 14 loci × five cohorts (empirical p-values)

Table S15 contains one row per locus–cohort pair (70 rows total): cohort, locus, number of variants scored, minimum empirical p-value, Benjamini–Hochberg q-value, lead variant identifier, and Scaled Average Difference (SAD). Cohorts: AD (Riley-Gillis et al. 2025 meta-analysis), T1D (Chiou et al. 2021), RA (FinnGen R12), IBD (de Lange et al. 2017 trans-ancestry meta-analysis), SLE (Bentham et al. 2015). This extension applies the same SuSiE fine-mapping → Enformer CTCF SAD → locus-matched null procedure as the primary analysis. Chromatin enrichment and eQTL/pQTL colocalization supplements remain primary 14-locus analyses only.

Cohort	Locus	N variants	Min p-value	BH q-value	Lead variant (chr:pos:ref:alt)
AD (Riley-Gillis meta)	IL33	20	**0.0020***	0.0200	chr9:6949071:G:GC
AD (Riley-Gillis meta)	STAT6	20	**0.0020***	0.0399	chr12:56701150:C:CT
AD (Riley-Gillis meta)	ZNF365	20	**0.0060***	0.1198	chr10:63543410:C:T
AD (Riley-Gillis meta)	ACTL9	20	**0.0180***	0.1796	chr19:10584044:G:A
AD (Riley-Gillis	FLG	5	**0.0180***	0.0449	chr1:152510338:G:GA

meta)					
AD (Riley-Gillis meta)	KIF3A	20	0.0619	0.5589	chr5:14582846:G:A
AD (Riley-Gillis meta)	IL4R	20	0.0778	0.5855	chr16:27276428:T:C
AD (Riley-Gillis meta)	TNFRSF6B	20	0.1038	0.6986	chr20:62312412:G:A
AD (Riley-Gillis meta)	IL4_IL13	3	0.2455	0.7156	chr5:132710020:C:T
AD (Riley-Gillis meta)	RAD50	3	0.2455	0.7365	chr5:132710020:C:T
AD (Riley-Gillis meta)	OVOL1	3	0.4232	0.7395	chr11:65759517:C:T
AD (Riley-Gillis meta)	TSLP	1	0.4970	0.4970	chr5:110832647:T:C
AD (Riley-Gillis meta)	C11ORF30	4	0.5050	0.5349	chr11:76422598:TTTG:T
AD (Riley-Gillis meta)	LRRC32	4	0.5070	0.5489	chr11:76422598:TTTG:T
T1D (Chiou 2021)	CCR7	7	**0.0020***	0.0140	chr17:40614389:T:TA
T1D (Chiou 2021)	GLIS3	20	**0.0080***	0.1597	chr9:4297892:T:C
T1D (Chiou 2021)	CTLA4	20	**0.0100***	0.1996	chr2:204013385:A:AAG
T1D (Chiou 2021)	IKZF3	20	**0.0140***	0.1996	chr17:38869201:T:C
T1D (Chiou 2021)	ERBB3	20	**0.0259***	0.3593	chr12:54277482:G:T
T1D (Chiou 2021)	RNLS	20	**0.0279***	0.5589	chr10:96657657:T:TTG
T1D (Chiou 2021)	BACH2	10	**0.0339***	0.3094	chr6:90206543:T:G
T1D (Chiou 2021)	IL7R	7	0.0679	0.4750	chr5:35850047:C:T
T1D (Chiou 2021)	IL2RA	7	0.2056	0.7370	chr10:6037381:T:C
T1D (Chiou 2021)	INS	9	0.2535	0.9094	chr11:2159830:T:G
T1D (Chiou 2021)	CLEC16A	7	0.3134	0.7984	chr16:11076904:C:T
T1D (Chiou 2021)	CD69	2	0.4311	0.4790	chr12:9670356:A:G
T1D (Chiou 2021)	SH2B3	2	0.4830	0.8244	chr12:111927557:TAA:T
T1D (Chiou 2021)	PTPN22	4	0.4870	0.7784	chr1:113659372:CTG:C
RA (FinnGen R12)	PADI4	20	**0.0060***	0.1198	chr1:17226843:A:G
RA (FinnGen R12)	TRAF1_C5	20	**0.0060***	0.0599	chr9:122128571:C:T
RA (FinnGen R12)	TNFAIP3	20	**0.0080***	0.1597	chr6:138413175:A:G
RA (FinnGen R12)	IRF5	2	**0.0140***	0.0279	chr7:129100859:C:T
RA (FinnGen R12)	TYK2	5	0.0539	0.2695	chr19:10446221:C:T
RA (FinnGen R12)	BLK	20	0.0679	0.6088	chr8:11484374:A:C
RA (FinnGen R12)	TNFRSF11A	20	0.0778	0.6254	chr18:62843254:C:T
RA (FinnGen R12)	CD40	2	0.1896	0.3792	chr20:46060026:G:A
RA (FinnGen R12)	IL2RA	7	0.2016	0.4541	chr10:6037381:T:C
RA (FinnGen R12)	IL6R	1	0.2236	0.2236	chr1:154532630:C:T
RA (FinnGen R12)	STAT4	1	0.2455	0.2455	chr2:191108308:G:A
RA (FinnGen R12)	PTPN22	4	0.5050	0.9780	chr1:113595528:AAG:A
RA (FinnGen R12)	CTLA4	1	0.5130	0.5130	chr2:203834104:CAA:C

RA (FinnGen R12)	CCR6	1	0.9501	0.9501	chr6:167124799:G:GCA
IBD (de Lange 2017)	RGS14	7	**0.0020***	0.0140	chr5:177371039:t:g
IBD (de Lange 2017)	ATG16L1	3	**0.0060***	0.0180	chr2:233247116:a:c
IBD (de Lange 2017)	LRRK2	9	**0.0060***	0.0539	chr12:40429853:g:gta
IBD (de Lange 2017)	TNFSF15	9	**0.0080***	0.0719	chr9:114856934:t:c
IBD (de Lange 2017)	JAK2	2	**0.0339***	0.0679	chr9:5016455:a:t
IBD (de Lange 2017)	STAT3	2	0.0699	0.1397	chr17:42362183:a:g
IBD (de Lange 2017)	IRGM	6	0.1257	0.7545	chr5:150866162:a:g
IBD (de Lange 2017)	IL12B	4	0.1397	0.3473	chr5:159401217:a:g
IBD (de Lange 2017)	CARD9	5	0.1457	0.7285	chr9:136374746:a:g
IBD (de Lange 2017)	IL10	5	0.7705	0.9182	chr1:206779403:a:t
IBD (de Lange 2017)	PTPN22	2	0.7764	0.9900	chr1:113761186:a:c
IBD (de Lange 2017)	NOD2	2	0.7944	0.8822	chr16:50724938:a:g
IBD (de Lange 2017)	FUT2	4	0.7964	0.9222	chr19:48706068:t:c
IBD (de Lange 2017)	IL23R	2	0.8004	0.8363	chr1:67240275:a:g
SLE (Bentham 2015)	ITGAM	1	**0.0060***	0.0060	chr16:31265490:G:A
SLE (Bentham 2015)	PRDM1	20	**0.0140***	0.2794	chr6:106503715:T:G
SLE (Bentham 2015)	TNFAIP3	20	**0.0160***	0.2794	chr6:138106848:T:G
SLE (Bentham 2015)	IL10	20	**0.0339***	0.3593	chr1:206766559:G:A
SLE (Bentham 2015)	IKZF1	20	0.1697	0.6347	chr7:50118890:T:C
SLE (Bentham 2015)	IRF5	2	0.2974	0.5948	chr7:128936032:T:C
SLE (Bentham 2015)	FCGR2A	1	0.6228	0.6228	chr1:161509020:G:A
SLE (Bentham 2015)	IRF7	1	0.6739	0.6739	chr11:625085:A:G
SLE (Bentham 2015)	BANK1	2	0.6926	0.8124	chr4:102002588:C:T
SLE (Bentham 2015)	PTPN22	1	0.6946	0.6946	chr1:113761186:C:A
SLE (Bentham 2015)	STAT4	2	0.7705	0.8283	chr2:191079016:C:T
SLE (Bentham 2015)	BLK	1	0.8423	0.8423	chr8:11493510:G:A
SLE (Bentham 2015)	TNIP1	1	0.9661	0.9661	chr5:151078210:T:C
SLE (Bentham 2015)	UBE2L3	1	1.0000	1.0000	chr22:21630090:G:A

** $p < 0.05$. BH q = Benjamini–Hochberg q-value. Loci are sorted within each cohort by ascending minimum empirical p-value. Enformer SAD scores computed for all 70 locus–cohort pairs.*

Supplementary Methods

SuSiE fine-mapping. 95% coverage credible sets (coverage = 0.95), maximum 100 EM iterations, locus-specific LD matrices constructed from 1000 Genomes Project EUR samples ($n \approx 503$).

Enformer inference. 196 kilobase input window centered on each variant (GRCh38/hg38 reference genome), CTCF track index 5312, Scaled Average Difference (SAD) as the primary effect-size metric.

Null variant generation. Variants with GWAS $p > 0.5$ drawn from within the same locus window as each credible-set variant, matched for minor allele frequency (±0.05) and distance from locus center.

Chromatin enrichment. Two-tailed Fisher's exact test applied to credible-set variants intersected with ChromHMM regulatory state annotations (CD4 naive T cells, CD4 memory T cells, B cells), Hi-C loop anchor regions from CD4+ T cells, and ENCODE cCREs with CTCF binding.

Colocalization. coloc v5 (Giambartolomei et al. 2014) with default priors ($p1 = 1\times10^{-4}$, $p2 = 1\times10^{-4}$, $p12 = 1\times10^{-5}$). Shared causal variant posterior probability $PP.H4 \geq 0.5$ used as the threshold for declaring colocalization.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- InsulatorLeakSupplementary.zip