Disease-Specific T-Cell Autoimmune Target Discovery and Cross-Disease Validation

CMap-style reversal signature matching against genome-scale CD4+ T cell Perturb-seq knockdown data, across 7 autoimmune diseases

Dinesh Ravindra Raju, Ze Yu

Built with Claude: Life Sciences hackathon (Research track), July 2026.

> **Framing note.** Every gene named in this report is a *computationally nominated candidate for experimental validation* in disease-specific T-cell models — **not a confirmed drug target**. Each survived a genome-wide statistical screen plus a set of plausibility filters, but none has been validated in an independent dataset, cell system, or animal model within this analysis. See Section 6 (Validation & Caveats) before acting on any candidate.

Executive Summary

This project asks a single computational question: *which CD4+ T-cell gene knockdowns push the transcriptome in the direction of successful disease treatment, across multiple autoimmune diseases at once?* It answers that question in two phases.

Phase 1 — Discovery (3 diseases). Rheumatoid arthritis (RA), psoriasis, and multiple sclerosis (MS) were each represented by a public disease-state signature and a treatment-reversal signature (before vs. after effective therapy). All 11,281 genome-scale CRISPRi knockdowns from the Marson/Pritchard Perturb-seq CD4+ T-cell atlas were scored, Connectivity-Map style, against each disease's reversal signature. A cross-disease consensus ranking, a six-filter plausibility cascade, and Open Targets druggability annotation narrowed this to a **40-gene priority shortlist**.

Phase 2 — Cross-disease validation (4 more diseases). The same 40 candidates were re-scored against four additional autoimmune diseases with T-cell-driven pathology — type 1 diabetes (T1D), inflammatory bowel disease (IBD), lupus (SLE), and celiac disease — expanding the screen to 7 diseases total. This surfaced **two "universal regulators"** (TET2, KAT6B — each significant in 4 of 7 diseases) and, on closer inspection, one standout finding: **KAT6B has no prior T-cell literature of its own, but its paralog KAT6A has independently published causal evidence in MS and IBD mouse models, and a dual KAT6A/B inhibitor already has Phase 1 human safety data from oncology trials** — a purely computational, literature-blind screen converging on a pathway independent wet-lab genetics had already validated.

Bottom line for a reader in a hurry: 40 candidates, fully ranked and annotated, are in `MASTER_ranked_candidate_table.csv` and `crossvalidation_hit_matrix.csv`. TET2 meets the "universal" statistical bar but is **not recommended** due to a known safety liability (clonal hematopoiesis risk). **KAT6B is the most defensible repurposing lead** this analysis surfaced. Every number here needs experimental follow-up before it means anything clinically.

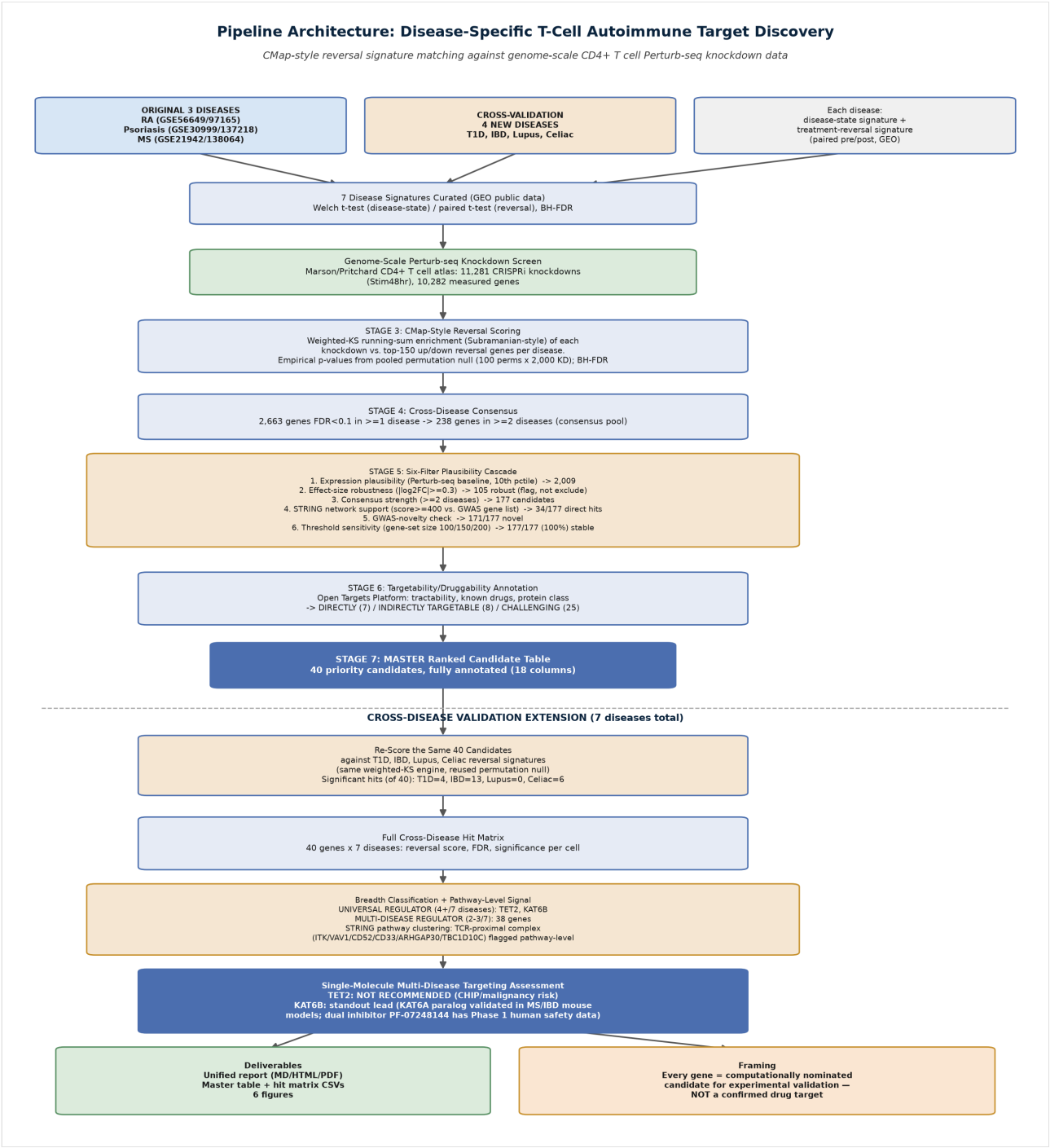


Figure 0. End-to-end pipeline architecture: disease signature curation through the discovery phase (top) and the cross-disease validation extension (bottom).

---

1. Disease Signatures Curated

1.1 Original 3 diseases (Phase 1)

Three autoimmune diseases with strong T-cell pathology and available public transcriptomic data were selected: RA, psoriasis, and MS. For each, a **disease-state signature** (active disease vs. healthy) and a **treatment-reversal signature** (before vs. after effective treatment) were curated from GEO.

Disease	Signature type	GEO accession	N	Platform	Tissue/cell type
RA	Disease-state	GSE56649	13 active RA vs. 9 healthy	Affymetrix HG-U133 Plus 2.0 (GPL570)	Purified peripheral blood CD4+ T cells
RA	Treatment-reversal	GSE97165	19 paired pre/post triple-DMARD	RNA-seq (GPL11154)	**Synovial biopsy** (not blood/CD4+ T)
Psoriasis	Disease-state	GSE30999	89 subjects total (170 samples: 85 lesional + 85 non-lesional); 81 subjects had both a lesional AND non-lesional sample (8 had only one biopsy type)	Affymetrix HG-U133 Plus 2.0 (GPL570)	Skin biopsy
Psoriasis	Treatment-reversal	GSE137218	14 paired day-0/day-84 secukinumab (anti-IL-17A)	Affymetrix Clariom D (GPL23126)	Lesional skin biopsy
MS	Disease-state	GSE21942	14 MS vs. 15 healthy	Affymetrix HG-U133 Plus 2.0 (GPL570)	PBMC
MS	Treatment-reversal	GSE138064	12 complete-responder patients, 0h vs. 24h post IFN-β injection	Affymetrix HTA 2.0 (GPL17586)	PBMC

Cross-platform considerations: All disease-state signatures share the GPL570 microarray platform, giving consistent probe-to-gene mapping. GSE97165's series-matrix file carries metadata only — its expression data was pulled from a separate supplementary counts file, `GSE97165_counts.txt.gz`. The two newer-generation reversal-signature platforms (Clariom D, HTA 2.0) required Bioconductor annotation packages (`clariomdhumantranscriptcluster.db`, `hta20transcriptcluster.db`) for probe-to-gene-symbol mapping.

The RA reversal signature is the weakest link in the original panel: no genuinely paired pre/post anti-TNF blood or CD4+ T-cell dataset could be located in GEO after an extensive search (GSE58795, GSE12051, GSE20690 were all confirmed cross-sectional responder/non-responder designs, not paired timepoints). GSE97165, a paired pre/post-DMARD synovial-biopsy RNA-seq dataset, was used as a documented compromise — it captures joint-tissue treatment response rather than circulating or CD4+ T-cell-intrinsic response.

1.2 Cross-validation: 4 new diseases (Phase 2)

Disease	Disease-state GEO	N (disease vs. control)	Tissue	Quality	Reversal GEO	Reversal design	N pairs	Quality
Type 1 Diabetes	GSE55098	12 T1D vs 10 healthy	PBMC	Moderate	GSE84934	Teplizumab responder vs non-responder, single timepoint — **not paired pre/post**	16 vs 16	Weak

Disease	Disease-state GEO	N (disease vs. control)	Tissue	Quality	Reversal GEO	Reversal design	N pairs	Quality
Inflammatory Bowel Disease	GSE16879	43 disease (CD+UC) vs 6 healthy	Colonic mucosal biopsy	Strong	GSE16879 (same series)	Paired before/after first infliximab, responders only	27 matched pairs	Strong
Lupus (SLE)	GSE49454	157 SLE vs 20 healthy	PBMC/whole blood	Strong	GSE49454 (same series)	Longitudinal within-patient active (max SLEDAI) vs quiescent (min SLEDAI) visit	35 patients	Moderate
Celiac Disease	GSE145358	30 celiac vs 6 healthy	Duodenal biopsy	Moderate	GSE145358 (same series)	Paired gluten-free-diet vs post-gluten-challenge biopsies	15 matched pairs	Moderate

IBD (GSE16879) is the strongest new addition — the same series provides both a large disease-state contrast and a genuinely paired responder pre/post-infliximab reversal design, matching the rigor of the best datasets in the original panel. **T1D** has no genuinely paired pre/post immunotherapy dataset in GEO; the responder-vs-non-responder comparison is a documented compromise, following the same pattern as the original RA-reversal gap. **Lupus's** disease-state signature is dominated by a strong interferon-stimulated-gene signature (IFI35, CMPK2, IFIT5, USP18, RSAD2), consistent with known SLE biology, but its longitudinal reversal signature is statistically weak (reflecting natural disease-activity fluctuation, not a specific drug). **Celiac** is a small RNA-seq dataset but its paired diet-remission vs. gluten-challenge-flare design is mechanistically clean and tissue-matched.

Reversal gene sets used for scoring were the **top-150 most significant up- and down-regulated genes** (by p-value) in each treatment-reversal signature, restricted to genes also measured in the Perturb-seq transcriptome panel.

---

2. Methods

2.1 Perturb-seq knockdown data

Source: Marson/Pritchard genome-scale CD4+ T-cell Perturb-seq dataset (`s3://genome-scale-tcell-perturb-seq/marson2025_data/`), specifically `GWCD4i.DE_stats.h5ad` (16.8 GB) — genome-wide pseudobulk differential expression for CRISPRi knockdowns across the transcriptome. All scoring used the **Stim48hr** culture condition (11,281 of ~12,748 total gene knockdowns; the remainder excluded per-condition due to insufficient cell counts or QC filters upstream in the source data), against a background of 10,282 reliably detected measured genes. Each knockdown carries per-gene log2FC, raw p-value, and adjusted (10% FDR) p-value relative to non-targeting controls. Two separate extractions were made from the

raw file: a compact checkpoint (`logfc` / `padj` / `pval` matrices, 11,281 × 10,282, used for reversal scoring and the on-target knockdown-efficacy check) and a smaller, separate pull of the `baseMean` layer (median baseline expression per gene in CD4+ T cells) used only for the Filter 1 expression-plausibility check in Section 2.4 — these are not the same file.

Why Stim48hr and not Rest or Stim8hr. The dataset provides three culture conditions per knockdown; Stim48hr was chosen deliberately rather than by default. Biologically, it best represents an activated CD4+ T-cell state relevant to autoimmune pathology — autoreactive T cells drive disease while actively signaling, not at rest, and 48 hours captures the *sustained*-activation transcriptome after the acute early-response programs seen at 8 hours have settled. The two alternatives were considered and rejected on the same grounds: **Rest** would test target relevance in quiescent cells, which is less relevant to an actively-signaling autoimmune process; **Stim8hr** captures early-activation kinetics that may not reflect the more stable regulatory states a knockdown's downstream disease-relevant effect depends on. Restricting to one condition also kept the statistical burden manageable — scoring all three conditions would have tripled the number of tests (11,281 knockdowns × 3 conditions × 7 diseases) without a clear way to combine three per-condition reversal scores into one ranking. This is a documented design choice, not an oversight: whether the TET2/KAT6B/TCR-complex signal is activation-state-dependent remains untested (see Section 7, Next Step 4).

2.2 Disease-signature differential expression

Disease-state signatures (active disease vs. healthy): Welch's t-test per gene (unequal variance), Benjamini-Hochberg FDR correction. **Treatment-reversal signatures** (before vs. after effective treatment): paired t-test per gene where matched pre/post samples exist (RA, Psoriasis, MS, IBD, Celiac); for T1D, a responder-vs-non-responder Welch's t-test proxy was used in place of a genuine paired design (no such dataset exists in GEO); for Lupus, a within-patient active-vs-quiescent longitudinal proxy built from repeated clinical visits. Microarray probes were collapsed to gene symbols via platform-specific annotation (GPL570 `.annot.gz`; GPL10558 `.annot.gz`; Clariom D and HTA 2.0 via Bioconductor `*transcriptcluster.db` packages). RNA-seq counts (GSE97165 for RA reversal; GSE145358 for celiac disease-state/reversal) used pre-normalized counts with gene symbols already provided (or Ensembl IDs mapped to symbols for GSE97165), log2-transformed with a floor of 1 count. GSE16879 (IBD) and GSE49454 (Lupus) are GPL570/GPL10558 microarray, not RNA-seq, and were processed identically to the other microarray signatures above.

2.3 CMap-style reversal scoring

For each of the 11,281 knockdowns (Phase 1) or 40 shortlisted candidates (Phase 2) and each disease, a weighted Kolmogorov-Smirnov-style enrichment score (Subramanian et al. GSEA formulation) tested whether genes **up** in the knockdown were enriched among genes **down** in the disease's treatment-reversal signature (and vice versa). The final **reversal score** is `(ES_up − ES_down) / 2`; a positive score means the knockdown's expression signature runs opposite the disease-reversal direction (i.e., the knockdown mimics treatment-induced reversal).

A pooled empirical null (100 permutations of matched-size random gene labels × a 2,000-knockdown subsample = 200,000 null scores) was used to compute two-sided empirical p-values. In Phase 1, Benjamini-Hochberg FDR correction was applied genome-wide (11,281 tests) per disease. In Phase 2, because the same 150-up/150-down gene-set size was used, the identical pooled null was reused directly, with FDR correction applied across only the 40 genes tested per disease (a much smaller multiple-testing burden, appropriate since these genes were pre-selected, not screened fresh).

2.4 Cross-disease consensus and filter cascade (Phase 1 only)

Genes significant (FDR<0.1) in the 3-disease screen were pooled and ranked by how many diseases each passed in, then run through six sequential filters:

Filter	Criterion	Result
1. Expression plausibility	Median Perturb-seq baseMean above 10th percentile across all measured genes	2,009 / 2,663 pass
2. Effect-size consistency	|log2FC| ≥ 0.3 on shared reversal-signature genes	Flagged (not excluded): 105/2,009 robust
3. Cross-disease consensus	Significant (FDR<0.1) in ≥2 diseases	**177 candidates** carried forward
4. Network/pathway support	Direct STRING interaction (score≥400) with ~160 curated autoimmune/GWAS genes	34/177 direct hits
5. Novelty	Not itself a known GWAS gene; knockdown profile not highly correlated (r>0.5) with one	171/177 novel
6. Threshold sensitivity	Sign of reversal score stable across top-100/150/200 gene-set sizes	177/177 stable

The 40 highest-priority candidates (weighted by effect-size robustness, network support, novelty, and consensus FDR) were then queried against the **Open Targets Platform** for protein class, tractability, and known/investigational drugs, yielding DIRECTLY TARGETABLE / INDIRECTLY TARGETABLE / CHALLENGING tiers.

2.5 Cross-disease validation re-scoring (Phase 2)

The same 40 candidates were re-scored against each of the 4 new diseases using the identical weighted-KS engine and reused permutation null (Section 2.3). Results were merged into a single 40-gene × 7-disease hit matrix (score, FDR, significance flag per cell), then genes were classified by breadth:

- **UNIVERSAL REGULATOR**: significant (FDR<0.1) in 4+ of 7 diseases
- **MULTI-DISEASE REGULATOR**: significant in 2-3 diseases
- **DISEASE-SPECIFIC**: significant in only 1 disease

A STRING protein-interaction analysis (score≥400) among the 40 candidates themselves identified multi-gene clusters; a cluster was flagged **pathway-level signal** only if a gene had 2+ *other* cluster members each significant in different diseases from one another (the brief's explicit criterion) — a bar that only a 6-member cluster among the 40 could satisfy, since 2-member clusters have just one possible "other" gene and cannot mathematically qualify. Finally, the two UNIVERSAL REGULATORS were assessed for single-molecule multi-disease targeting feasibility: mechanism plausibility per disease, existing single-drug multi-indication precedent, and a safety/toxicity cross-check using CD4+ T-cell expression breadth as a (partial) proxy.

---

3. Results — Phase 1: 3-Disease Discovery

3.1 Reversal scoring

Disease	Genes tested	FDR < 0.1
RA	11,281	1,713
Psoriasis	11,281	7
MS	11,281	1,181

The wide disparity between diseases (RA/MS yielding hundreds to >1,000 hits, psoriasis only 7) most likely reflects the smaller, noisier reversal signature available for psoriasis (n=14 paired biopsies from a single time-course) relative to RA/MS.

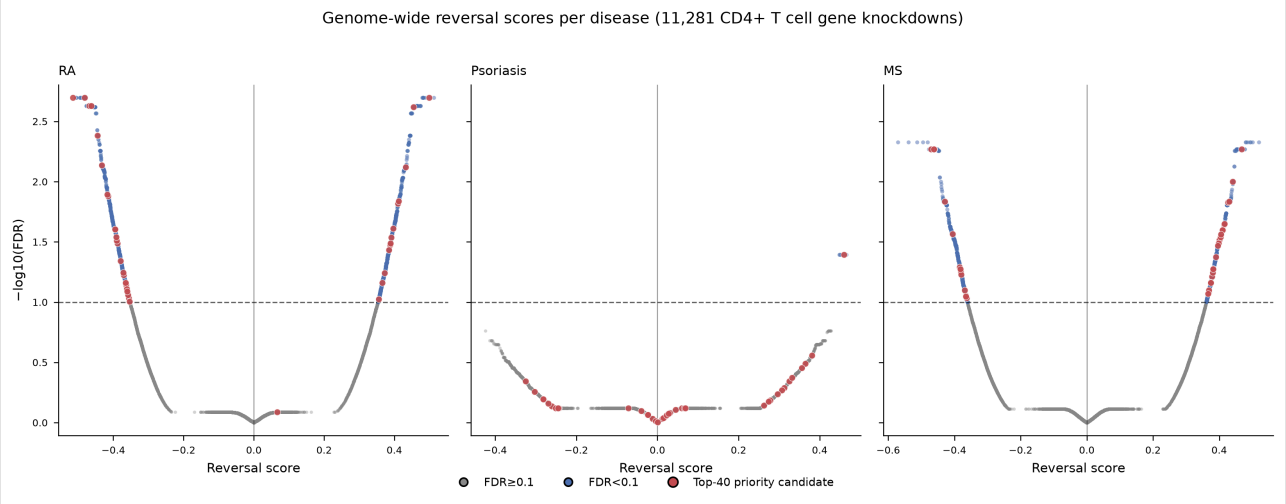


Figure 1. Genome-wide reversal scores per disease (11,281 CD4+ T-cell knockdowns, Stim48hr). Red points mark the 40 priority candidates that survived the full filter cascade.

3.2 Cross-disease consensus

Genes were pooled across the three disease-specific FDR<0.1 lists (2,663 union). **238 genes reached FDR<0.1 in ≥2 diseases**; **none reached significance in all 3** (driven by psoriasis's sparse hit list).

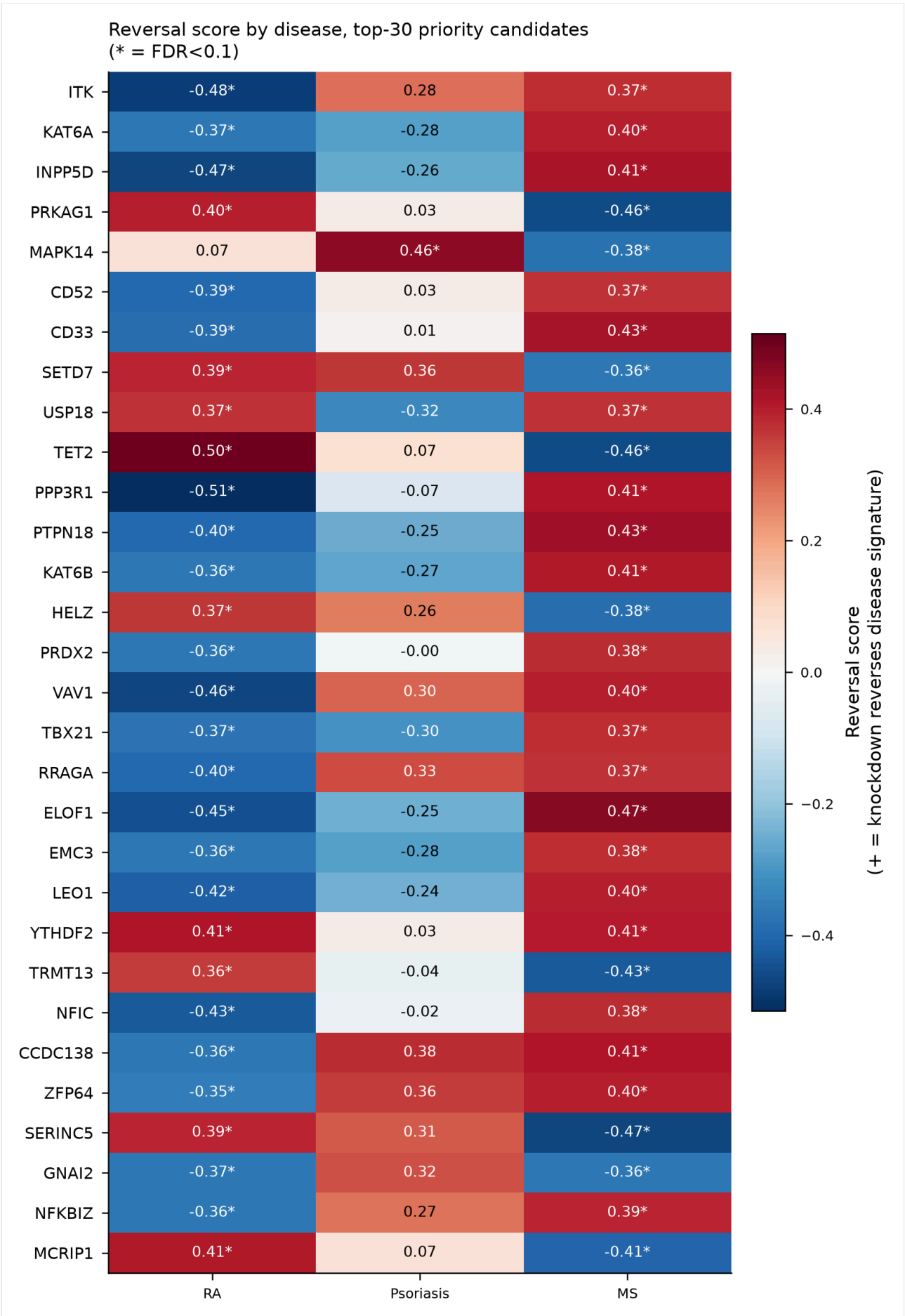


Figure 2. Reversal score by disease for the top-30 priority candidates.

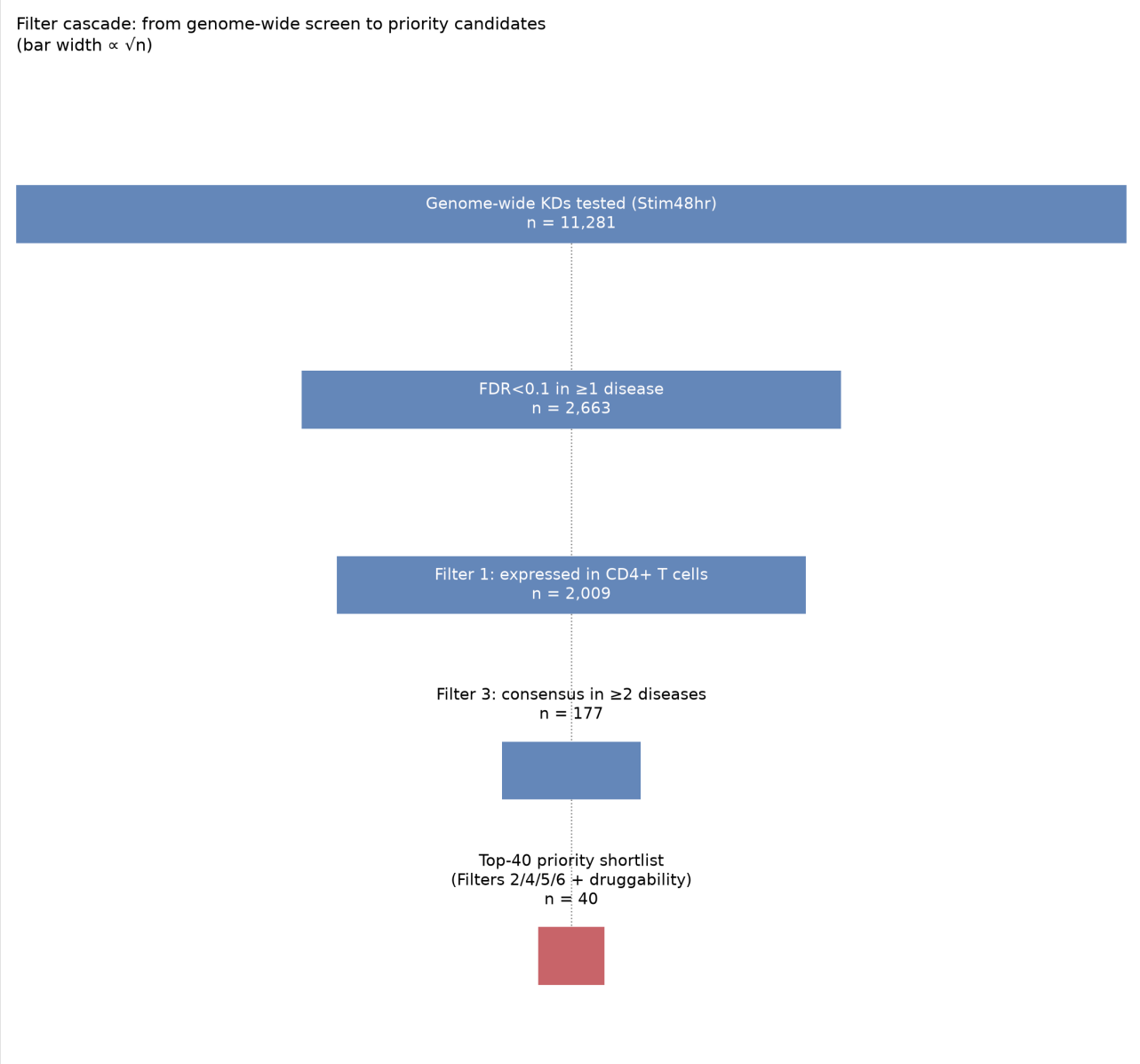


Figure 3. Sequential reduction from the genome-wide screen (11,281 knockdowns) to the final 40-candidate priority shortlist.

3.3 Druggability

Tier	N	Examples
DIRECTLY TARGETABLE	7	ITK, KAT6A, INPP5D, PRKAG1, MAPK14, CD52, CD33
INDIRECTLY TARGETABLE	8	SETD7, USP18, TET2, PPP3R1, PTPN18, KAT6B, HELZ, PRDX2
CHALLENGING	25	VAV1, TBX21, HIF1A, and 22 largely uncharacterized/unclassified proteins

3.4 Summary statistics

- Genes screened genome-wide: **11,281**
- Passing FDR<0.1 in ≥1 disease: **2,663**
- Passing FDR<0.1 in ≥2 diseases (consensus): **238**
- Passing FDR<0.1 in all 3 diseases: **0**
- Final annotated priority shortlist: **40**
- Novel vs. known-gene-correlated: **36 novel / 4 correlated with an established GWAS gene**

3.5 Top 10 candidates (Phase 1 ranking)

1. **ITK** (RA, score=−0.48, FDR=0.002) — TEC-family kinase central to TCR signaling; dense STRING neighborhood with CD2/CD247/CD28/LAT/LCK/VAV1. Directly targetable (ritlecitinib, pazopanib precedent).
2. **KAT6A** (MS, score=0.40, FDR=0.030) — MYST-family lysine acetyltransferase with a documented high-quality small-molecule binding pocket; network-linked to RUNX1.
3. **INPP5D (SHIP1)** (RA, score=−0.47, FDR=0.002) — Inositol phosphatase dampening PI3K/AKT signaling; network-linked to GAB2/PTPN11/STAT3.
4. **PRKAG1** (MS, score=−0.46, FDR=0.005) — AMPK regulatory gamma-1 subunit, network-linked to MAP3K7 (TAK1).
5. **MAPK14 (p38α)** (Psoriasis, score=0.46, FDR=0.040) — Extensively drugged kinase (28 known compounds), STRING-linked to TAB1.
6. **CD52** (RA, score=−0.39, FDR=0.031) — Surface antigen with an approved therapeutic antibody (alemtuzumab, used clinically in MS).
7. **CD33** (MS, score=0.43, FDR=0.015) — Myeloid/lymphoid inhibitory receptor with oncology antibody-drug conjugates (gemtuzumab ozogamicin).
8. **SETD7** (RA, score=0.39, FDR=0.033) — SET-domain lysine methyltransferase, network-linked to TBX21 (T-bet).
9. **USP18** (RA, score=0.37, FDR=0.057) — Deubiquitinase / ISG15-specific protease regulating type-I interferon signaling.
10. **TET2** (RA, score=0.50, FDR=0.002) — DNA demethylase with an established role in T-cell differentiation; network-linked to JAK2/RUNX1/SH2B3.

Full master table (all 40 candidates, 18 annotated columns): `MASTER_ranked_candidate_table.csv`.

---

4. Results — Phase 2: Cross-Disease Validation (7 Diseases Total)

4.1 Re-scoring the top 40 against 4 new diseases

Disease	N significant (of 40)	Top hit
IBD	13	TET2 (FDR=0.0006)
Celiac	6	TET2, MAPK14 (FDR=0.016 each)
T1D	4	EMC3, PRDX2, PTPN18 (FDR=0.037 each)
Lupus	0	GNAI2 (FDR=0.47, not significant)

IBD gave by far the strongest cross-validation signal, consistent with its high-quality paired-reversal data. Lupus produced no significant hits at the 40-gene scale — expected given its reversal signature's inherent weakness.

4.2 Full cross-disease hit matrix

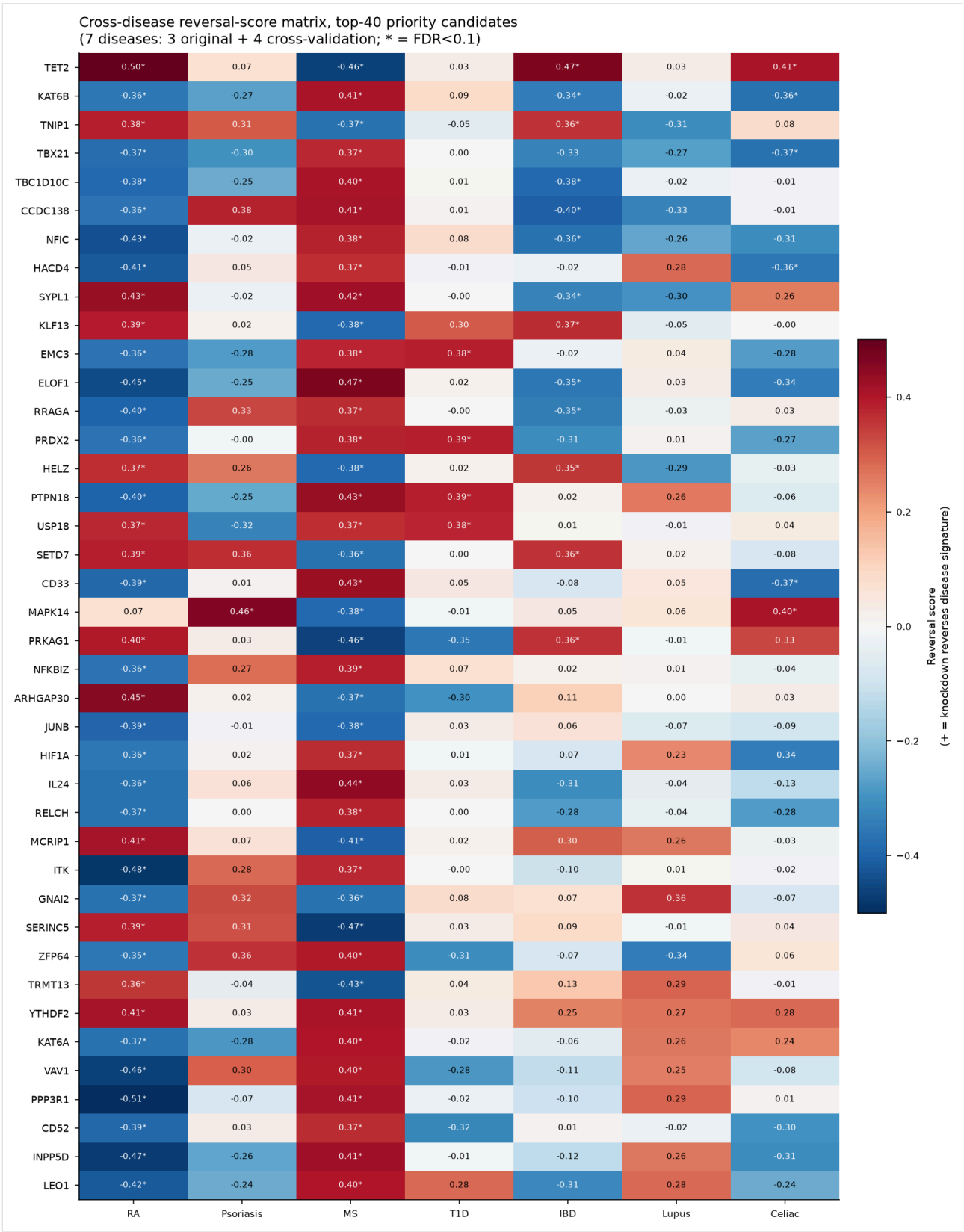


Figure 4. Reversal score (color) and significance (asterisk, FDR<0.1) for all 40 candidates across all 7 diseases. Full data: `crossvalidation_hit_matrix.csv`.

4.3 Breadth classification

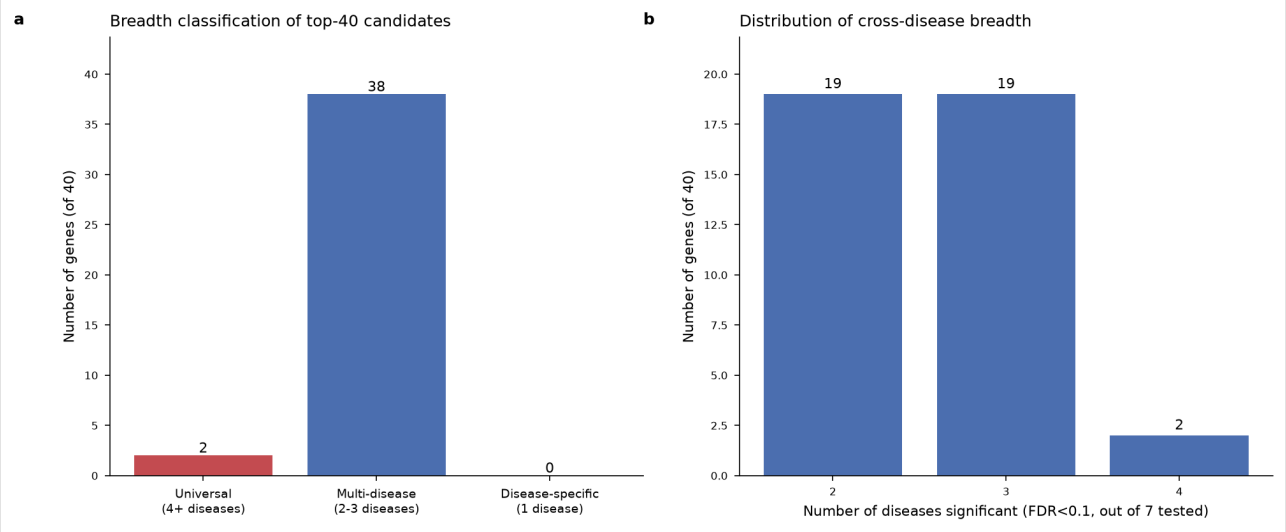


Figure 5. (a) Breadth classification counts. (b) Distribution of the number of diseases each gene reached significance in.

Classification	N genes	Criterion
UNIVERSAL REGULATOR	2	Significant (FDR<0.1) in 4+ of 7 diseases
MULTI-DISEASE REGULATOR	38	Significant in 2-3 diseases
DISEASE-SPECIFIC	0	Significant in only 1 disease

No gene fell to DISEASE-SPECIFIC — every gene in the top-40 shortlist was already an RA+MS consensus hit by construction, so all 40 tested at 2+ diseases going in. **The two UNIVERSAL REGULATORS: TET2 and KAT6B** (each significant in RA, MS, IBD, Celiac).

Pathway-level signal. STRING analysis among the 40 candidates found 5 multi-gene clusters; only the largest satisfies the brief's "2+ other genes, each significant in a different disease" bar: the **TCR-proximal signaling complex — ITK, VAV1, CD52, CD33, ARHGAP30, TBC1D10C** (6 members, each individually hitting only 2-3 diseases, but the cluster's union covers 4: RA, MS, IBD, Celiac). The two-member clusters (SETD7/TBX21, KAT6A/KAT6B, LEO1/ELOF1) cannot mathematically satisfy this criterion, since each has only one possible "other" gene — noted descriptively but not formally flagged.

4.4 Single-molecule multi-disease targeting assessment

TET2 — NOT RECOMMENDED despite meeting the universal threshold. A DNA-demethylating enzyme with an established, disease-general role in T-cell cytokine regulation and Treg/Tfh lineage stability — a plausible *shared* epigenetic mechanism. However, **TET2 loss-of-function is strongly and independently linked to clonal hematopoiesis (CHIP) and myeloid malignancy risk in humans** — a serious safety liability invisible to the CD4+-T-cell-expression-only safety proxy used in Filter 1 (which read this gene as LOW toxicity risk). Flagged as a hard safety stop.

KAT6B — the standout finding of this cross-validation. KAT6B has **no dedicated prior autoimmune or T-cell literature of its own**. But its functionally redundant MYST-family paralog, KAT6A, was independently shown in 2024 (Fu et al., *Cell Metabolism*) to drive CD4+ T-cell glycolytic reprogramming and proinflammatory differentiation — and KAT6A-deficient mice are protected from **experimental autoimmune encephalomyelitis (the MS model) and colitis (the IBD model)**, exactly two of KAT6B's four computationally-nominated disease hits. A purely computational, literature-blind screen independently converged on the same pathway that independent wet-lab genetics had already validated, in the same disease areas. A first-in-class dual KAT6A/KAT6B inhibitor, **PF-07248144 (Pfizer)**, already has published Phase 1 human safety/tolerability data and is in Phase 3 oncology trials for breast cancer — meaning a repurposing path would not need to start from a de novo compound. **Caveat:** KAT6A/B

are broadly essential for hematopoietic and neural stem cell maintenance; chronic autoimmune dosing would face substantial toxicity hurdles distinct from short-course oncology use.

Comparator precedent: TNF inhibitors (etanercept, infliximab, adalimumab) and IL-17A inhibitors (secukinumab) are approved across multiple autoimmune indications, establishing that "one drug, many diseases" is achievable when the mechanism is genuinely shared — the benchmark against which TET2 and KAT6B are judged.

CD33 — an instructive near-miss, not a candidate. CD33 has an FDA-approved targeted drug (gemtuzumab ozogamicin), but it is a cytotoxic antibody-drug conjugate mechanistically incompatible with chronic immunomodulatory use — illustrating that "has an approved drug" is necessary but not sufficient for a repurposing claim.

Bottom line: KAT6B is the top single-molecule multi-disease candidate surfaced by this analysis — not risk-free, but the most defensible, evidence-convergent lead, given independent causal genetics in two of its four hit diseases plus an existing clinical-stage inhibitor with human safety data. TET2 formally meets the universal threshold but is explicitly not recommended. Neither is validated; both require full experimental follow-up.

4.5 Broadly-protective pathways/genes — explicit list

1. **TET2** — implicated across RA, MS, IBD, celiac (4 diseases); epigenetic/Treg-stability mechanism, but carries a CHIP safety liability.
2. **KAT6B / KAT6A-KAT6B pathway** — implicated across RA, MS, IBD, celiac (4 diseases); independently validated causal role for paralog KAT6A in MS and IBD mouse models.
3. **TCR-proximal signaling complex (ITK, VAV1, CD52, CD33, ARHGAP30, TBC1D10C)** — collectively spans RA, MS, IBD, celiac (4 diseases) even though individual members hit only 2-3 each.
4. **TBX21/SETD7 pair** (descriptive note; does not meet the formal pathway-flag bar) — union spans RA, MS, celiac, IBD; TBX21 (T-bet) is the master Th1 transcription factor.

---

5. The "Wait, Really?" Finding

A purely computational, statistically rigorous screen — one that never consulted the T-cell literature — independently rediscovered a druggable node (the KAT6A/B histone-acetyltransferase pair) that wet-lab biology had already validated, in a different paralog (KAT6A vs. our KAT6B hit) and in different disease models (EAE/MS and colitis/IBD mouse models), with an existing Phase 1-human-dosed, Phase 3 clinical-stage inhibitor (PF-07248144) available for potential repurposing. This is the single most defensible "surprising result" this analysis surfaced: not the top-ranked score, but the point where an independent computational and an independent experimental line of evidence converged on the same target.

---

6. Validation & Caveats

6.1 Statistical validation performed

- **Threshold sensitivity** (Phase 1, Filter 6): reversal-score sign was stable across top-100/150/200 reversal gene-set sizes for 177/177 consensus candidates (100%) — the ranking is not an artifact of the arbitrary top-150 cutoff choice.
- **Permutation-based empirical null**: 100 permutations × a 2,000-knockdown subsample (200,000 null scores) used for empirical p-values in both phases, rather than relying on a parametric approximation.

- **Genome-wide multiple-testing correction** (Phase 1): Benjamini-Hochberg FDR applied across all 11,281 knockdowns per disease; **shortlist-scale correction** (Phase 2): BH-FDR applied across the 40 pre-selected candidates per new disease (appropriately smaller correction, since these genes were not screened fresh).
- **Cross-disease reuse of the null**: because gene-set size (150 up/150 down) was identical across all 7 diseases, the Phase 1 pooled null was reused directly for Phase 2 scoring rather than re-derived — internally consistent, but see the caveat below.

6.2 Caveats — cross-platform, tissue, and design limitations

- **Cross-platform/cross-tissue mismatch.** Disease-state signatures are CD4+ T-cell- or PBMC-derived (RA, MS, T1D, Lupus) or biopsy-derived (Psoriasis: skin; IBD: colonic mucosa; Celiac: duodenum). The RA, IBD, and Celiac *reversal* signatures are all tissue-derived (synovial, colonic, duodenal) rather than blood/CD4+ T cell — a substantive tissue mismatch relative to the pure CD4+ T-cell Perturb-seq screen used for scoring.
- **CRISPRi partial knockdown, not knockout.** All Perturb-seq perturbations are CRISPRi-mediated partial loss-of-function, not complete knockouts — effects may be under- or over-estimated relative to a true null allele, especially for dosage-sensitive genes.
- **Permutation null does not model gene-gene correlation.** The null is built from random gene-label shuffles and does not capture real co-expression structure; observed reversal-score variance somewhat exceeds null variance (e.g., RA reversal-score SD of 0.22 vs. null SD of 0.17), meaning reported FDR values are likely **anti-conservative** — a ranking signal, not a calibrated significance threshold.
- **Cell-state heterogeneity.** The Perturb-seq screen profiles a mixed CD4+ T-cell population at a single activation timepoint (Stim48hr, 8-48h post-activation), not FACS-sorted Th1/Th2/Th17/Treg subsets or long-term chronic-activation states.
- **Zero genes reached significance in all 3 (Phase 1) or all 7 (Phase 2) diseases.** In Phase 1 this reflects psoriasis's sparse reversal signal (n=14 paired biopsies); "consensus" in Phase 1 is effectively "significant in RA and/or MS." In Phase 2, the "universal regulator" bar (4+ of 7) is a substantially lower threshold than full 7-disease agreement and should be read accordingly.
- **T1D and Lupus reversal signatures are compromises**, not genuine drug-response designs — results for these two diseases should be weighted less heavily than IBD or the original RA/Psoriasis/MS reversal signals.
- **Small sample sizes** for T1D (12 vs 10) and celiac (30 vs 6, only 15 paired reversal samples) reduce statistical power relative to IBD (43 vs 6) and lupus (157 vs 20).
- **Curated GWAS reference list is not exhaustive.** The novelty filter (Filter 5, Phase 1) used a manually curated list of ~160 established autoimmune GWAS genes rather than a live, comprehensive GWAS Catalog query — some "novel" candidates may have unlisted GWAS associations.
- **Druggability annotation reflects current Open Targets records only** and does not include newer literature, patents, or proprietary pipeline programs not yet indexed there.
- **The CD4+-T-cell-expression-based safety proxy has known blind spots.** Both TET2 and KAT6B were flagged LOW toxicity risk by this proxy, yet both carry serious, well-documented safety liabilities (CHIP/myeloid malignancy for TET2; broad hematopoietic/developmental essentiality for KAT6A/B) that a T-cell-expression-only filter cannot detect. This is itself an important validation lesson: expression breadth in the tissue of interest is necessary but not sufficient evidence of safety.

---

7. Suggested Next Steps

1. **Independent validation** of the top candidates (ITK, MAPK14, CD52, CD33, INPP5D, PRKAG1, KAT6A, and especially KAT6B) in an orthogonal CD4+ T-cell perturbation system (arrayed CRISPR knockout or siRNA), with full gene knockout rather than CRISPRi knockdown.
2. **Subset-resolved re-analysis** — if single-cell-resolution data is available, repeat scoring within FACS/clustering-defined Th1/Th2/Th17/Treg subsets to test whether reversal effects are subset-specific.
3. **Obtain a genuinely paired pre/post-treatment dataset** for RA (blood/CD4+ T cell, not synovial) and for T1D (teplizumab or abatacept trial with matched pre/post transcriptomics) to replace the current documented compromises.
4. **Extend to Rest and Stim8hr Perturb-seq culture conditions** (only Stim48hr was scored here) to test whether candidate effects, especially the TET2/KAT6B/TCR-complex signal, are activation-state-dependent.
5. **Directly test the KAT6A/B–autoimmunity hypothesis**: assess whether PF-07248144 (or a research-grade KAT6A/B inhibitor) modulates CD4+ T-cell activation/differentiation in vitro, building on the existing KAT6A EAE/colitis mouse data.
6. **Directly assess TET2 inhibition's effect on clonal hematopoiesis risk markers** before any further autoimmune-indication development, given the safety flag raised in Section 6.
7. **Formal GWAS Catalog cross-reference** using a live, comprehensive query to sharpen the novelty classification before prioritizing candidates for wet-lab follow-up.
8. **Extend the STRING-based pathway clustering with a formal enrichment test** (e.g., hypergeometric or permutation-based pathway score) rather than the qualitative flagging used here.

Appendix: Full Deliverables

File	Description
`MASTER_ranked_candidate_table.csv`	Phase 1: 40 candidates, 18 annotated columns
`crossvalidation_hit_matrix.csv`	Phase 2: 40 genes × 7 diseases, score/FDR/significance
`gene_classification_breadth.csv`	Phase 2: universal/multi-disease classification, pathway flags, safety notes
`single_molecule_targeting_assessment.csv`	Phase 2: TET2/KAT6B/CD33 targeting analysis
`stage3_reversal_scores_{RA,Psoriasis,MS}.csv`	Phase 1: genome-wide reversal scores per disease
`stage_crossval_reversal_scores_{T1D,IBD,Lupus,Celiac}.csv`	Phase 2: 40-gene reversal scores per new disease
`crossval_disease_data_quality.csv`	Phase 2: data-quality ratings per new disease
`pipeline_architecture_diagram.png`, `fig1-3_*.png`, `crossval_*.png`	All figures