From a CRISPR Screen to Therapeutic Vulnerabilities: Functional Genomics of Tamoxifen Sensitisation in ER+ Breast Cancer

Almokhtar Aljarodi, Dr Hugo Tavares

almokhtaraljarodi@gmail.com, hugo.tavares@slcu.cam.ac.uk

Department of Genetics, University of Cambridge

KAUST Academy Bioinformatics Research Programme

Abstract

This work connects a genome-wide CRISPR knockout screen to public transcriptomic, dependency and network resources in order to answer two questions fixed in advance: which gene knockouts increase the efficacy of tamoxifen in oestrogen-receptor-positive (ER+) breast cancer cells, and which of those candidates carry the lowest predicted risk of adding to the musculoskeletal symptom burden that drives discontinuation of endocrine therapy. Both halves were written into a dated pre-analysis plan on 5 August 2026, before any result existed. Thirteen of 19,103 fitted genes passed the pre-specified threshold of FDR < 0.10 with a sensitising direction. Four were carried forward under two separate rules: *KDM1A* and *TLK2* on statistical certainty (FDR 0.0004 and 0.0017 against a next-best value of 0.0146), and *USP34* and *VEZF1* as the only candidates that also changed significantly in an independent transcriptomic dataset. Candidate-level signal did not survive replication: of sixteen gene-by-dataset combinations only two reached FDR 0.05, and those two arise in different datasets and different biological contexts, so no candidate is supported by more than one context. Pathway-level signal behaved differently. Oestrogen response (early) and E2F targets were suppressed in all four transcriptomic datasets, and the epithelial-mesenchymal transition programme was enriched in the three long-term resistance contexts while being suppressed under twelve hours of tamoxifen exposure, with twelve of twelve enrichment statistics significant. Baseline dependency profiling across eleven ER+/luminal lines without drug showed *TLK2* required in nine of them, so its screen phenotype cannot be cleanly attributed to tamoxifen. Every remaining candidate carried either a documented tissue liability or an absence of published evidence. The pre-analysis plan anticipated naming a single lowest-risk candidate; the evidence does not support naming one, and that outcome is reported here rather than a ranking built on unequal literature coverage.

Date: August 24, 2026

Code: `https://github.com/almokhtar8-stack/breast_cancer`

Freeze: `science-freeze-2026-08-15`

commit `9a1b7777d6c69c2be44f16f25bc950769dc2ffda`

Contents

1 Introduction **3**
1.1 Clinical Problem and Real-World Impact . . . 3
1.2 Problem Statement . . . 3
1.3 Key Assumptions and Boundaries . . . 3
1.4 Summary of the Approach . . . 4

2 Related Work **4**
2.1 Functional Screens in Endocrine Therapy Response . . . 4
2.2 Transcriptomic Models of Resistance . . . 4
2.3 Resources for Target Assessment . . . 4
2.4 The Gap This Study Addresses . . . 4

3 Methodology **5**
3.1 Overview and Data Sources . . . 5
3.2 CRISPR Screen Reanalysis . . . 5
3.3 Differential Expression . . . 6
3.4 Single-Cell Processing and Pseudobulk . . . 7
3.5 Pathway Analysis . . . 7
3.6 Network Analysis . . . 7
3.7 Safety Assessment . . . 7
3.8 Pre-specification and Reproducibility Controls . . . 7

4 Results **8**
4.1 Thirteen Knockouts Increase Tamoxifen Efficacy . . . 8
4.2 Candidate-Level Signal Does Not Replicate . . . 9
4.3 Pathway-Level Signal Replicates Completely . . . 11
4.4 Network Context . . . 11
4.5 Every Candidate Carries a Liability . . . 12
4.5.1 Baseline Dependency . . . 12
4.5.2 Documented Tissue Liabilities . . . 12
4.5.3 Absence of Evidence . . . 13
4.5.4 Counter-Evidence and Genetic Constraint . . . 14
4.5.5 Conclusion of the Assessment . . . 14

5 Discussion **15**
5.1 The Level at Which the Signal Is Stable . . . 15
5.2 Two States, Not a Trajectory . . . 16
5.3 What the Network Shows and Does Not Show . . . 16
5.4 The Safety Question Cannot Be Answered as Posed . . . 16
5.5 Reproducibility Challenges . . . 17
5.5.1 The Reconstruction Gap . . . 17
5.5.2 Format-Forced Method Divergence . . . 17

6 Limitations and Future Work **17**
6.1 Limitations . . . 17
6.2 Future Work . . . 18

7 Conclusion **18**

8 Data and Code Availability **19**

1 Introduction

1.1 Clinical Problem and Real-World Impact

Roughly 70% of breast cancers are hormone-receptor positive, and endocrine therapy is the backbone of their systemic management. Tamoxifen is a selective oestrogen receptor modulator and competes with oestrogen at the receptor itself. Aromatase inhibitors act further upstream, suppressing oestrogen synthesis and reducing circulating oestrogen by 97 to 99%. Both classes are taken for five to ten years, which turns tolerability from a comfort question into a determinant of outcome.

The tolerability problem is concentrated in the musculoskeletal system. A meta-analysis pooling 21 studies and 13,177 participants estimated arthralgia prevalence on aromatase inhibitors at 46% (pooled estimate 0.459, 95% CI 0.397 to 0.520), with very high between-study heterogeneity ($I^2 = 98\%$) [1]. In a prospective cohort of 503 women receiving aromatase inhibitors, 32.4% discontinued within two years because of adverse effects and 24.3% discontinued specifically because of musculoskeletal symptoms [2]; a separate review reports mean discontinuation attributable to arthralgia of 9.3% [3]. Discontinuation is associated with higher recurrence and mortality, so a drug that is stopped early forfeits the benefit it was prescribed to deliver.

Those figures describe aromatase inhibitors and should not be transferred to tamoxifen. Tamoxifen has partial oestrogen-agonist activity in bone and is generally bone-sparing in postmenopausal women, and its side-effect profile differs accordingly. What generalises across the class is the structural problem: endocrine therapy works only for as long as patients continue to take it, and joint pain, muscle pain and fatigue are the symptoms that most often end that continuation.

1.2 Problem Statement

This creates a specific and unusual design target. A gene whose loss makes tamoxifen more effective offers a route to a better response without escalating dose. But a sensitising target is only useful if inhibiting it does not add to the symptom burden that already limits adherence. The second requirement is rarely asked at the point where candidates are chosen, and it is normally asked too late, after chemistry has been committed. Asking it at the shortlist stage is the distinguishing feature of this project.

The question was therefore written in two halves, both fixed in a dated pre-analysis plan on 5 August 2026 before any result existed:

1. Which gene knockouts make tamoxifen more effective in ER+ breast cancer cells?
2. Which of those candidates carry the lowest predicted risk of contributing to the joint pain, muscle pain and fatigue that drive patients to discontinue endocrine therapy?

1.3 Key Assumptions and Boundaries

Four boundaries constrain every claim that follows, and they are stated at the outset rather than buried in the discussion.

- The screen was run in MCF7-V, a drug-tolerant clone rather than a line with acquired resistance. The knockouts modulate tamoxifen response and increase tamoxifen efficacy. They do not reverse resistance or restore sensitivity, and this report never says that they do.
- The work is entirely computational. There is no laboratory validation, no candidate is a validated target, and no gene is shown to cause resistance.
- No molecular docking or binding-energy calculation was performed as part of the network analysis. Where network structure is discussed, it indicates reachability, not efficacy. Geomet-

ric pocket detection (fpocket) exists elsewhere in the project repository but is outside the scope of this report and supports no claim made here.

- No claim is made that any candidate would reduce the side effects of endocrine therapy. The claim under test is narrower: whether a candidate would avoid adding to a burden that already exists.

1.4 Summary of the Approach

We reanalysed a genome-wide CRISPR knockout screen performed in MCF7-V cells under 4-hydroxytamoxifen, attempted to replicate each candidate in four independent transcriptomic datasets spanning cell-line resistance, independent resistant sublines, recurrent tumours and acute ex vivo treatment, tested whether signal survived at the level of pathways rather than genes, placed the candidates in a functional association network, and assessed each one for tissue liability using dependency data and a structured literature search. The headline outcome is a mismatch of levels: the biology replicates as a coordinated programme while failing to replicate as any individual gene, and no candidate can be named lowest-risk on the available evidence.

2 Related Work

2.1 Functional Screens in Endocrine Therapy Response

Pooled CRISPR knockout screening in a drug-treated versus untreated contrast is the standard way to ask which gene losses change a cell population's response to a compound. The screen reanalysed here follows that design, contrasting oestradiol plus 4-hydroxytamoxifen against oestradiol alone in MCF7-V [4]. The advantage over expression profiling is direct: the screen measures a functional requirement under drug pressure rather than an abundance difference in a resistant state. Its limitation is equally direct, in that a depletion phenotype in a pooled population does not distinguish a drug-specific effect from a general growth requirement, which is why the baseline dependency comparison in Section 4.5 is necessary rather than optional.

2.2 Transcriptomic Models of Resistance

Four public transcriptomic resources are used here, and each was generated to answer a different question. GSE118713 compares an MCF7 parental line with a tamoxifen-resistant derivative. GSE111151 extends the same comparison across four genetic backgrounds, MCF7, T-47D, ZR-75-1 and BT-474, each with resistant sublines. GSE240112 contrasts primary with recurrent tumours. GSE245601 exposes ER+/HER2− tumour tissue to tamoxifen for twelve hours ex vivo [5]. Studies of this kind are commonly pooled to increase power. Section 5.2 shows why pooling the acute dataset with the three resistance datasets would be a mistake in this instance, since one programme moves in opposite directions in the two contexts.

2.3 Resources for Target Assessment

Three community resources supply the assessment layer. DepMap provides genome-wide loss-of-function fitness estimates across cancer cell lines using the Chronos model [6, 7], which allows a sensitisation phenotype to be separated from a baseline growth requirement. STRING supplies scored functional associations between human proteins [8]. The MSigDB Hallmark collection provides a compact set of well-characterised biological programmes suitable for enrichment testing without the redundancy of larger collections [9].

2.4 The Gap This Study Addresses

Candidate selection in functional genomics usually optimises a single axis, namely the strength or the certainty of the phenotype, and defers tolerability questions to preclinical pharmacology.

The consequence is that tissue liability is discovered after a chemistry programme is already committed. The approach taken here inverts the order by attaching a structured tissue-liability assessment to the shortlist itself, using dependency data and a documented literature search with recorded query strings.

A second gap concerns the evidence base for that assessment. An attempt to find tissue-level transcriptomic data from patients receiving endocrine therapy, in muscle, bone or synovium, found none openly available. The closest human muscle studies either used targeted assays without depositing omics data, or withheld raw data for consent reasons. The assessment presented here therefore rests on cell-line dependency and on published gene-level evidence, and its limits are a property of what is public rather than of the method.

3 Methodology

3.1 Overview and Data Sources

Figure 1 gives the shape of the analysis. A genome-wide CRISPR knockout screen supplies candidates under two separate selection rules; four independent transcriptomic datasets test whether those candidates replicate and whether signal replicates at the level of programmes where it does not at the level of genes; network context, baseline dependency and a structured literature search supply the assessment. All datasets are publicly available and none is under restricted access, with the single qualification recorded in Section 5.5.1. Table 1 lists them together with the design feature that governed how each was analysed.

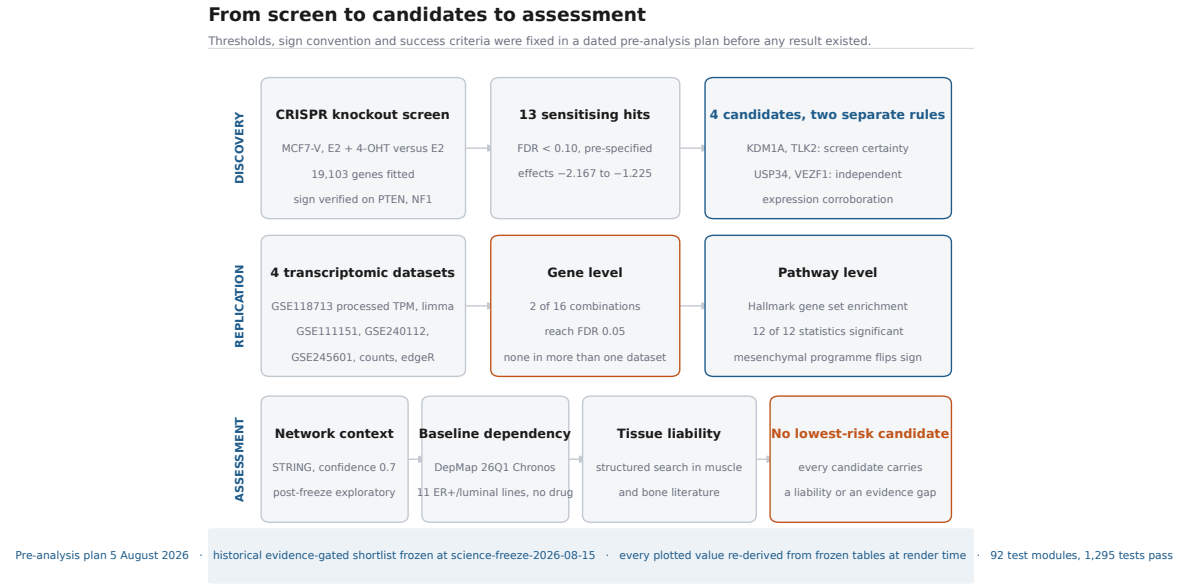


Figure 1 Analysis workflow. Thresholds, sign convention and success criteria were fixed in a dated pre-analysis plan before any of these steps produced a result. This panel is a schematic and is the only figure in the report that does not re-derive its plotted values from the frozen result tables, because it plots none.

3.2 CRISPR Screen Reanalysis

The screen contrast was E2 + 4-OHT against E2 in MCF7-V, and 19,103 genes were fitted. The sign convention was fixed in advance: a negative effect means the knockout increases tamoxifen-induced depletion, which is a sensitising direction. *PTEN* and *NF1* were used as free positive controls to verify that the sign was oriented correctly before any candidate was inspected. The

Table 1 Data sources and the design feature that governed their analysis. Deposited format, not analyst preference, determined the differential expression framework for each dataset.

Source	Type	Design
Hany et al. 2023 [4]	CRISPR screen	Genome-wide knockout, contrast E2 + 4-OHT versus E2, MCF7-V
GSE118713	Bulk RNA	MCF7 parental versus tamoxifen-resistant (TAMR) and fulvestrant-resistant (FASR) derivatives; three biological replicates per condition, all from a single MCF7-based experimental model; processed matrix available as TPM
GSE111151	Bulk RNA	Four cell lines (MCF7, T-47D, ZR-75-1, BT-474), each with resistant sublines; raw counts
GSE240112	scRNA	Primary ($n = 3$) versus recurrent ($n = 3$) tumours; unpaired, different patients, two different tissue banks
GSE245601 [5]	scRNA	Ten paired ER+/HER2− tumours, control versus 12 h 10 μM tamoxifen ex vivo; 3 of the 10 pairs (Tumour_02, Tumour_03, Tumour_07) meet the frozen $\geq$50-malignant-cell-per-arm pseudobulk eligibility rule, so the effective paired pseudobulk sample size is $n = 3$
DepMap 26Q1 [6, 7]	Dependency	Chronos gene effect, 11 evaluable ER+/luminal breast lines
STRING [8]	Network	Human functional associations, confidence 0.7
MSigDB [9]	Gene sets	Hallmark collection

significance threshold, FDR < 0.10, was declared before results existed and was not revised afterwards.

3.3 Differential Expression

The framework used for each dataset was determined by the format of the processed matrix that was analysed. For GSE118713 the processed expression matrix used in this reanalysis was available as TPM. Count-based negative binomial models such as edgeR, and the voom transformation, both require counts and were therefore not applicable to that matrix. Raw sequencing reads are available through SRA and could in principle have been re-quantified; that was not done here. Rather than re-quantifying the deposited raw reads, differential expression was performed using limma [10] on $\log_2(\mathrm{TPM}+1)$: `lmFit`, `contrasts.fit` and `eBayes(trend=FALSE, robust=FALSE)` under the design `~0 + group`.

The other three datasets had counts and were analysed with edgeR [11]: TMM normalisation, `filterByExpr` supplied with the design matrix, `estimateDisp`, and the quasi-likelihood F-test via `glmQLFit` and `glmQLFTest`. The designs were:

- GSE111151: `~cell_line + resistance_status`
- GSE240112: `~group`
- GSE245601: `~patient + treatment`

This divergence is a consequence of deposition practice rather than a methodological choice, and Section 5.5.2 records what it cost.

3.4 Single-Cell Processing and Pseudobulk

Single-cell datasets were processed in Seurat [12]. Doublets were removed with scDblFinder [13] and DoubletFinder [14]. Malignant cells were inferred with inferCNV [15] and cross-checked with CopyKAT [16]. Counts were then summed per patient per condition to give pseudobulk profiles, so that the patient rather than the cell is the unit of replication. GSE245601 is therefore analysed as per-tumour pseudobulk, not as single-cell data, and it measures acute response to twelve hours of tamoxifen rather than resistance. A pre-specified eligibility rule requires at least 50 malignant cells in *both* arms of a tumour pair before that pair contributes to malignant pseudobulk. Three of the ten pairs meet this rule, so the effective paired pseudobulk sample size for GSE245601 is $n = 3$ rather than $n = 10$. GSE240112 is unpaired: the primary and recurrent groups come from different patients and from two different tissue banks.

3.5 Pathway Analysis

Preranked gene set enrichment [17] was run against the MSigDB Hallmark collection [9] using `gseapy.prerank` with the multilevel significance method [18], on genome-wide ranked results, with Benjamini–Hochberg correction applied within each dataset. Ranking used the full genome-wide result, not a filtered candidate list, which is what makes the pathway analysis independent of the candidate selection.

3.6 Network Analysis

Functional associations were taken from STRING [8] for *Homo sapiens* at a confidence threshold of 0.7, fixed before the query was run. One standardised query was applied identically to all four candidates. Shortest paths were computed with NetworkX [19] using `nx.all_shortest_paths` and exported for Cytoscape [20]. A sensitivity check was run at confidence 0.4. This analysis is post-freeze and exploratory, and STRING associations are functional rather than evidence of physical binding, activation or inhibition.

3.7 Safety Assessment

Baseline dependency was taken from DepMap 26Q1 Chronos gene effect [6, 7] across the eleven evaluable ER+/luminal breast lines, with no drug present. This asks whether a gene is required for growth on its own, independently of tamoxifen.

Tissue liability was assessed by a structured literature search in skeletal muscle and in bone, conducted per candidate using current gene symbols together with historical aliases. Within the frozen safety module the query strings and returned counts were recorded alongside the results. That module was built around the historical frozen shortlist; for the two post-audit additions, *KDM1A* and *TLK2*, the muscle and bone evidence reported here was assembled by targeted review of the primary literature rather than by that same recorded-query pipeline, and this difference in provenance is noted rather than smoothed over. Three categories of evidence were kept strictly separate throughout: expression of a gene in a tissue, demonstrated functional requirement in that tissue, and evidence of toxicity. Results that could not be verified against the primary record were excluded, and each exclusion was recorded rather than silently dropped. Where a search returned nothing, this report states that no research was found, which is a statement about the search and its coverage rather than about the world.

3.8 Pre-specification and Reproducibility Controls

The pre-analysis plan is dated 5 August 2026 and fixes the thresholds, the sign convention and the success criteria before any result existed. Amendments to it are logged with dates and commit hashes rather than edited in place. The candidate list was frozen at the tag `science-freeze-2026-08-15` (commit `9a1b7777d6c69c2be44f16f25bc950769dc2ffda`) and veri-

fied unmoved; the SHA-256 of the frozen shortlist was recorded and is unchanged. Every value plotted in every result figure is re-derived from the frozen result tables at render time, and the build fails if a value has moved. The analysis was subjected to two external audits and to continuous independent review by a second automated reviewer, with corrections recorded alongside the analyses they changed. The repository contains 92 pytest modules, one per source module, and the full suite passes (1,295 tests passed, 1 skipped at the time of writing).

One provenance gap is recorded here rather than left implicit. The project environment declares `gseapy` without a version pin, and no lock file, run log or result-table header records the exact version that produced the frozen enrichment tables, so that version cannot now be established from the repository. Because the multilevel significance implementation changed between `gseapy` 1.3.0 and 1.3.1, the twelve Hallmark statistics reported in Section 4.3 were independently recomputed under `gseapy` 1.3.1 from the frozen ranked gene lists, using the same pre-specified parameters (`min_size` 15, `max_size` 500, 1,000 permutations, seed 20260813, `method="multilevel"`). All twelve normalised enrichment scores reproduced to four decimal places and all twelve remained significant at FDR 0.05, so the reported values do not depend on which of those versions was originally used. The missing pin is nonetheless a real reproducibility gap and is listed as such.

The term used throughout is pre-specified rather than pre-registered. A dated plan held under version control with a git tag is a real commitment device, but it is not a public preregistration, and the weaker word is the accurate one.

4 Results

4.1 Thirteen Knockouts Increase Tamoxifen Efficacy

Of 19,103 fitted genes, 13 passed the pre-specified threshold of FDR < 0.10 with a sensitising direction (Table 2, Figure 2). Two of the thirteen sit above the conventional 0.05 line and are retained because 0.10 was the declared threshold, not because they were re-examined afterwards.

Four genes were carried forward, under two rules applied separately rather than combined into a single score. *KDM1A* and *TLK2* were carried on statistical certainty: their FDR values, 0.0004 and 0.0017, are roughly an order of magnitude smaller than anything else in the list, where the next-best value is 0.0146. *USP34* and *VEZF1* were carried on a second and independent criterion, being the only genes in the thirteen that also showed a significant change in an independent transcriptomic dataset.

Two genes with larger effect sizes were deliberately not carried forward. *USP17L29* (effect -2.145, FDR 0.028) had zero baseline expression in every downstream dataset, produced no differential expression fit, and had no entry in the GSE111151 count matrix, so it could not be followed anywhere. *TADA2B* (effect -2.064, FDR 0.0146) had no supporting evidence in any independent dataset, reaching FDR 0.985 in the resistance model.

The selection history is worth stating plainly, because the current four candidates are not the original shortlist. The **historical frozen shortlist** consisted of *USP34*, *VEZF1*, *EML5* and *CITED2*, and entry to it required transcriptomic corroboration. *KDM1A* and *TLK2* were excluded by that gate. Following an external audit that judged the gate scientifically unnecessary, the **post-audit focus set** was defined as *KDM1A*, *TLK2*, *USP34* and *VEZF1*; these are the **current four candidates carried into the final comparative analysis** reported here, and *EML5* and *CITED2* are not carried forward. The audit's reasoning was that a gene whose loss changes the functional response to a drug need not be differentially expressed in resistant cells, and requiring it to be discards exactly the class of functional modifier the screen was designed to find. Failure of expression replication does not invalidate a functional CRISPR hit, because the two measurements address different biological properties. The sequence is recorded here because the correction improved the analysis, and because a shortlist that changed for a stated reason is

more interpretable than one that appears to have been right first time.

Table 2 The thirteen genes passing the pre-specified screen threshold (FDR < 0.10) with a sensitising direction, ordered by effect size. A negative effect means the knockout increases tamoxifen-induced depletion.

Gene	Effect	FDR	Carried forward
KDM1A	-2.167	0.0004	Yes, on statistical certainty
USP17L29	-2.145	0.0280	No, not followable downstream
TADA2B	-2.064	0.0146	No, no independent support
TLK2	-1.848	0.0017	Yes, on statistical certainty
PET117	-1.828	0.0419	No
SUPT4H1	-1.684	0.0373	No
TSR3	-1.619	0.0680	No
VEZF1	-1.602	0.0373	Yes, on independent expression change
EIF4ENIF1	-1.551	0.0373	No
HMGB1	-1.474	0.0373	No
CTDNEP1	-1.393	0.0692	No
USP34	-1.391	0.0417	Yes, on independent expression change
ICK	-1.225	0.0373	No

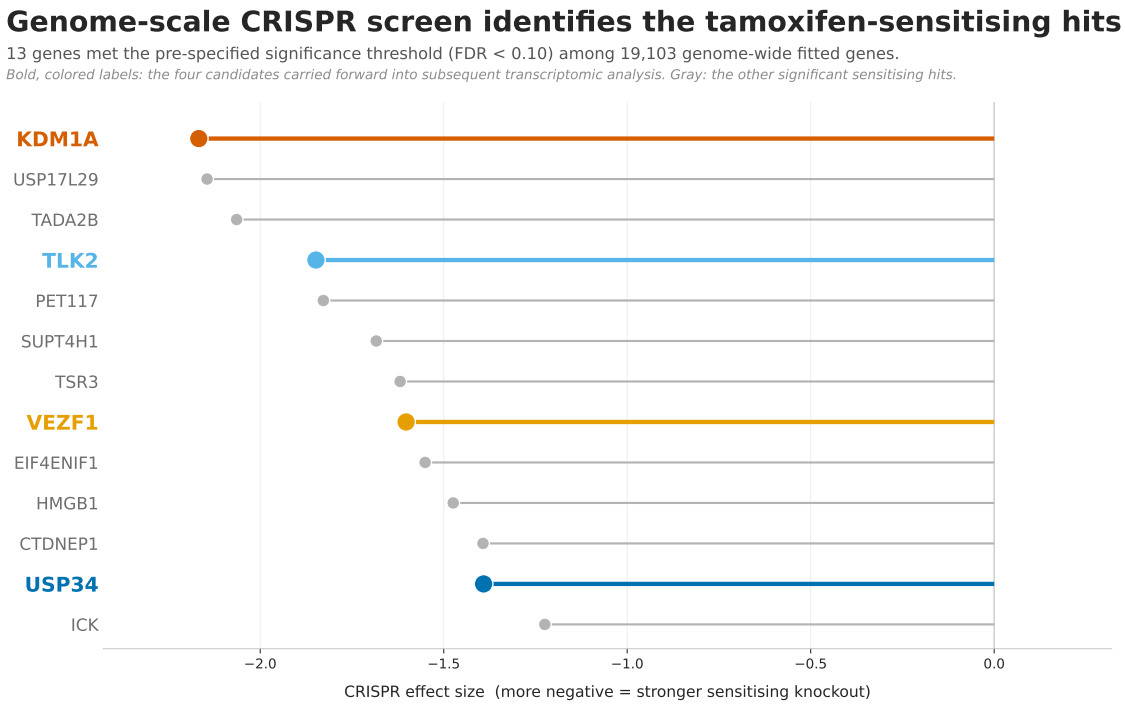


Figure 2 Thirteen of 19,103 fitted genes pass the pre-specified threshold with a sensitising direction. The four genes carried forward are shown in bold colour and the other nine significant hits in grey. Effect sizes span a narrow range, from -2.167 to -1.225, so effect size alone does not separate the list; the false discovery rates in Table 2 do, which is why *KDM1A* and *TLK2* were carried forward on certainty rather than on magnitude.

4.2 Candidate-Level Signal Does Not Replicate

Each of the four candidates was tested in each of the four transcriptomic datasets, giving sixteen gene-by-dataset combinations. Two reached significance (Table 3): *USP34* in GSE118713 at FDR 0.0073, and *VEZF1* in GSE240112 at FDR 0.0195. The remaining fourteen did not. Figure 3 shows

the underlying observations and makes the reason visible: within-group spread is comparable to between-group separation in most gene and dataset combinations.

No candidate reached significance in more than one dataset, and the two significant results arise in different datasets and in different biological contexts: *USP34* in a cell-line resistance model and *VEZF1* in recurrent tumours. The three resistance-context datasets (GSE118713, GSE111151, GSE240112) were not combined into a single formal cross-dataset test, and GSE245601 would in any case have to be excluded from such a combination because it measures acute response rather than resistance (Section 5.2). Counting how many datasets return a significant result is a weak ranking criterion, because it rewards a single strong result in a single context; on the evidence in Table 3 no candidate-level expression result is supported by more than one context, and none is treated here as validation.

Table 3 Replication of the four candidates across four transcriptomic datasets, as false discovery rates. Two of sixteen combinations reach FDR 0.05, shown in bold. They fall in different datasets and different biological contexts, so no candidate is supported in more than one context. Each dataset is treated as its own biological context and no formal cross-dataset pooled test was performed.

Dataset	*KDM1A*	*TLK2*	*USP34*	*VEZF1*
GSE118713 (cell-line resistance)	0.494	0.675	**0.0073**	0.238
GSE111151 (independent sublines)	0.91	0.84	0.63	0.61
GSE240112 (recurrent tumours)	0.59	0.88	0.23	**0.0195**
GSE245601 (acute 12 h)	0.37	0.88	0.90	0.89

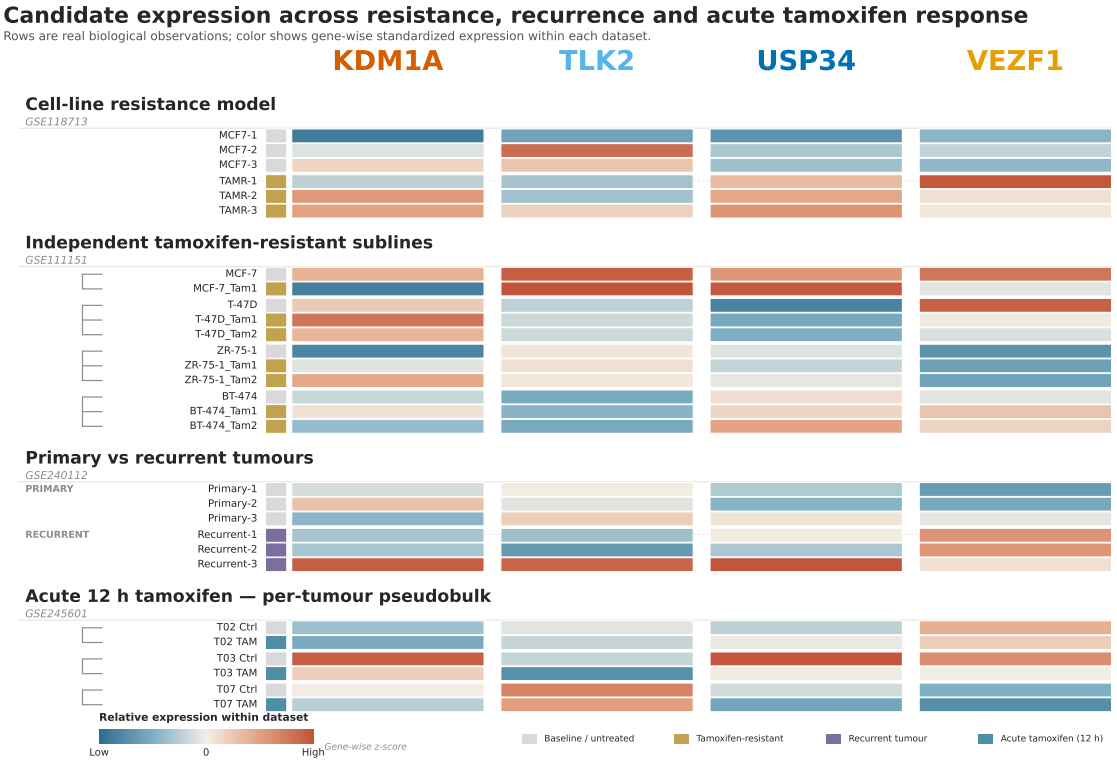


Figure 3 The four candidates across all 29 biological observations in the four transcriptomic datasets, grouped by context. Colour is a gene-wise z-score computed *within* each dataset, so it shows where a sample sits relative to others in the same study and never a cross-dataset absolute expression level. Read against Table 3, the figure shows why the formal tests mostly return nulls.

4.3 Pathway-Level Signal Replicates Completely

Where individual genes failed to replicate, programme-level signal did not. Three Hallmark programmes were examined across the same four datasets, giving twelve enrichment statistics, and all twelve are significant at FDR 0.05 (Table 4, Figure 4).

Oestrogen response (early) and E2F targets are suppressed in all four datasets. The epithelial-mesenchymal transition programme is enriched in the three long-term contexts and suppressed in the twelve-hour context.

Three points govern how this should be read. First, the four datasets are separate studies with no shared timeline, so the contrast between them is a contrast between two distinct biological states and not a temporal progression or trajectory. Second, the consequence of that structure is practical: pooling acute treatment response with acquired resistance would place a positive and a negative enrichment of the same programme into the same average and cancel two genuine signals against each other. Third, none of the four candidate genes appears in the leading edge of any of these programmes, so the pathway result is independent of the candidate selection and is not circular.

Table 4 Normalised enrichment scores for three Hallmark programmes across four datasets. All twelve values are significant at FDR 0.05. The mesenchymal programme is enriched in the three long-term contexts and suppressed under twelve hours of tamoxifen. These are four separate studies with no shared timeline, so the difference is between two states rather than a progression.

Programme	Cell-line resistance	Independent sublines	Recurrent tumours	Acute 12 h
Oestrogen response (early)	−2.88	−1.84	−2.43	−2.07
E2F targets	−2.16	−1.80	−1.41	−1.58
Epithelial-mesenchymal transition	+1.98	+2.17	+1.51	−1.56

4.4 Network Context

At STRING confidence 0.7 the candidate subnetwork contains 47 nodes, 147 edges and three separate components. *KDM1A* has 12 partners and *USP34* has 9, both in the same component. *TLK2* has 5 partners and forms its own separate component. *VEZF1* has zero partners at this threshold.

KDM1A and *USP34* are three edges apart, connected through *DNMT1* and a ubiquitin gene, with edge scores of 0.998, 0.923 and 0.982 along the *UBC* route (Figure 5). All four equally short routes pass through *DNMT1* and differ only in which ubiquitin gene occupies the middle position (*RPS27A*, *UBA52*, *UBB*, *UBC*). This is one bridge described four ways, not four independent connections, and it should not be read as corroboration by repetition.

The caveat needs stating precisely. *DNMT1* has degree 17 in this subnetwork and is a high-degree hub within it. *UBC* has degree 7 here, lower than *KDM1A*'s 12, so it is not a hub by this figure's own data; what is true of *UBC* is that it is a ubiquitin gene that associates promiscuously in STRING overall. The two nodes therefore fail the same test for different reasons, and describing both as high-degree hubs would misstate what this subnetwork shows.

At confidence 0.4 the *TLK2* component merges with the rest of the network through *SRSF1* and *ASF1A*, so its separation at 0.7 is threshold-dependent. *VEZF1* remains isolated at both thresholds, which makes its isolation the better supported of the two observations. The 0.7 threshold was fixed before the query was run.

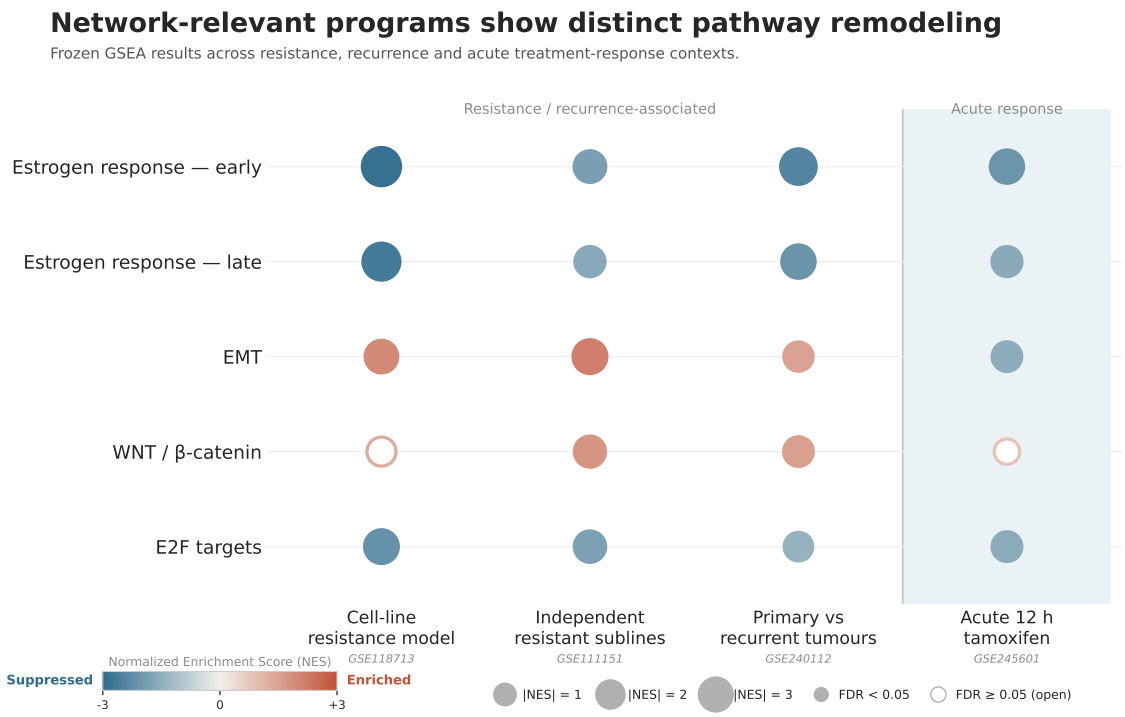


Figure 4 The biology repeats where the individual genes did not. Marker colour gives the direction of the normalised enrichment score and marker size its magnitude; open markers denote FDR ≥ 0.05. Oestrogen response and cell-cycle programmes are suppressed in all four datasets, while the mesenchymal programme separates the three long-term resistance contexts from the acute twelve-hour context, in which it is suppressed. The figure also carries late oestrogen response and WNT signalling, shown for context and not among the three programmes reported in Table 4.

4.5 Every Candidate Carries a Liability

4.5.1 Baseline Dependency

Across the eleven ER+/luminal lines with no drug present, *KDM1A* and *USP34* were required in none, *VEZF1* in three, and *TLK2* in nine (Table 5). *TLK2*'s screen result therefore cannot be cleanly attributed to tamoxifen, because a gene the cells need in order to grow at all will deplete under any additional pressure. This is a specificity problem rather than a safety problem in the tissue sense, and it is the reason *TLK2* is not treated as a leading candidate despite its screen certainty.

Figure 6 places the two axes against each other directly. Sensitisation strength and baseline dependency are separate properties, and the four candidates occupy three different positions in that space: *KDM1A* and *USP34* sensitise without being required at baseline, *VEZF1* sits in between, and *TLK2* sits high on dependency, which is what disqualifies it.

4.5.2 Documented Tissue Liabilities

USP34 has a documented role in bone. Knockdown by siRNA and shRNA in human mesenchymal stromal cells, and Prx1-Cre conditional knockout in mice, showed that *USP34* stabilises Smad1 and RUNX2 by deubiquitination and so counteracts Smurf1, with ectopic bone formation reduced by 35% in *USP34*-depleted cells [21]. A 2020 erratum to that paper withdrew and replaced three figure panels (1F, 6M and 7F), corrected a sequencing dataset that had been linked to a different study by the same authors (the correct data are deposited under GSE115789), and defined previously undefined scale bars; the authors state that the core conclusions are unchanged [22]. The corrected record is the one cited here.

KDM1A has a documented role in skeletal muscle, but the direction is unresolved. Ablation or

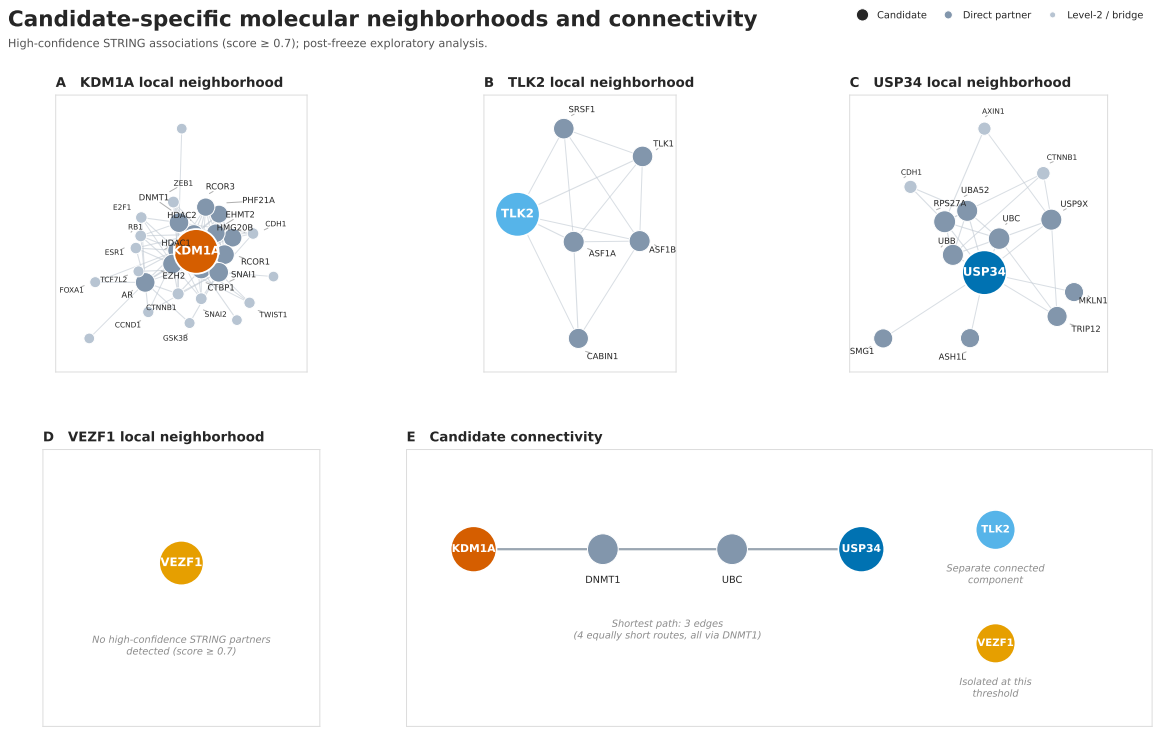


Figure 5 Local STRING neighbourhoods for each candidate at confidence 0.7 (panels A to D) and the connectivity between them (panel E). *VEZF1* has no high-confidence partners at this threshold and *TLK2* forms a separate component. *KDM1A* and *USP34* sit three edges apart, joined through *DNMT1* and a ubiquitin gene; all four equally short routes pass through *DNMT1* and differ only in the ubiquitin gene at the middle position, so this is a single bridge rather than four independent links. This analysis is post-freeze and exploratory.

inhibition of Lsd1 delays muscle regeneration and diverts satellite cells towards brown adipocytes, so loss is harmful in that setting [23]. Conversely, LSD1 inhibition with CC-90011 protects male mice against dexamethasone-induced muscle wasting while preserving the anti-inflammatory effect of the glucocorticoid [24]. A third study, using muscle-specific LSD1-knockout mice, reports that LSD1 loss instead exacerbates dexamethasone-induced atrophy in fast-fibre-dominant muscle while enhancing exercise-induced hypertrophy in slow-fibre-dominant muscle [25]. These two studies conflict specifically on the role of LSD1 in glucocorticoid-induced atrophy; they are not opposite across skeletal muscle biology as a whole, and the third study additionally addresses a fibre-type selectivity and an exercise axis that the second does not examine. Taken with the satellite-cell regeneration setting, which is a different question again, the direction of effect in skeletal muscle is reported here as unresolved.

KDM1A also has a documented role in bone, and here the published evidence does point in opposing directions. Suppression or pharmacological inhibition of LSD1 impairs osteoblast differentiation and reduces bone-nodule formation, with LSD1 occupying RUNX2-binding sites [26]. Conversely, LSD1 knockdown has been reported to promote osteoblast differentiation, with mesenchymal LSD1 deletion increasing bone mass in mice through WNT7B and BMP2 signalling [27]. Because primary studies support effects in both directions, the direction in bone is described here as mixed.

4.5.3 Absence of Evidence

For *USP34* in skeletal muscle, and for *TLK2* and *VEZF1* in both skeletal muscle and bone, no relevant research was identified in the structured search. The searches used current symbols together with historical aliases, and every query string and returned count was recorded. An

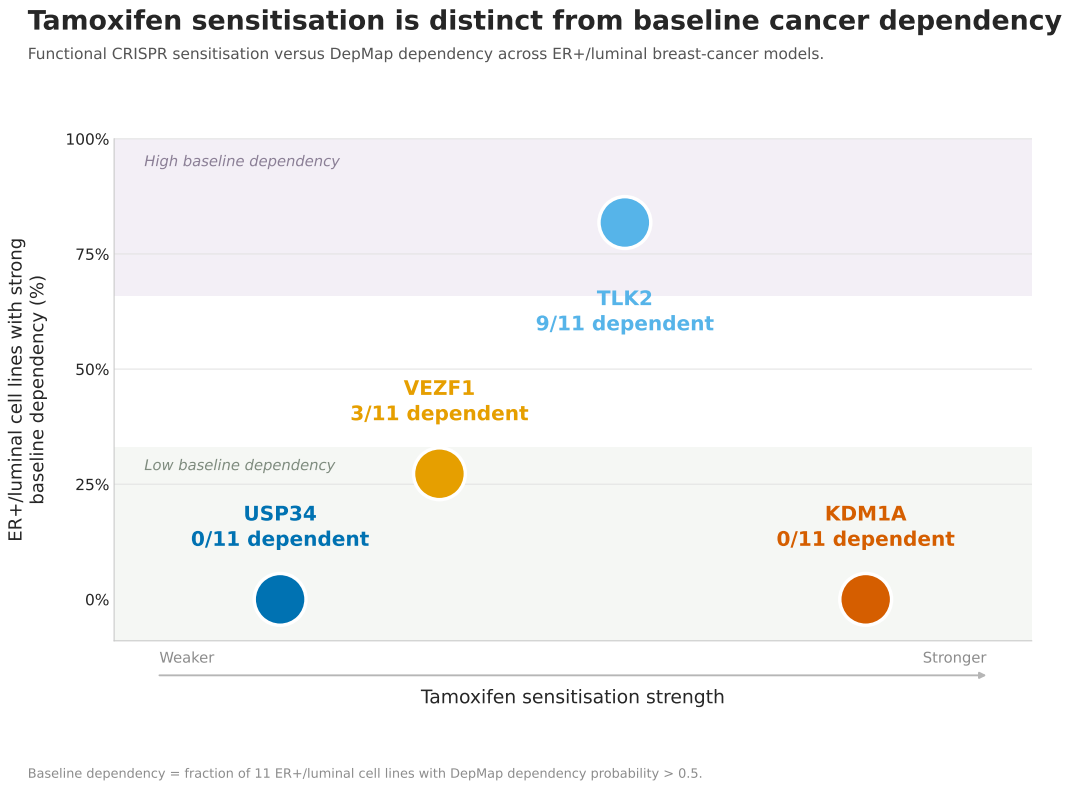


Figure 6 Tamoxifen sensitisation is distinct from baseline cancer-cell dependency. The vertical axis is the fraction of the eleven ER+/luminal lines in which the gene is a strong dependency with no drug present; the horizontal axis is sensitisation strength from the screen. *TLK2* is required in nine of eleven lines, so its screen phenotype cannot be cleanly attributed to tamoxifen. *KDM1A* and *USP34* sensitise while being required in none, which is the combination the screen was intended to find.

absence of findings here is a property of the search and its coverage, not a statement about the world, and it is reported as such.

4.5.4 Counter-Evidence and Genetic Constraint

Two further pieces of evidence about *USP34* are disclosed here because they cut against it. Loss of *USP34* has been reported to promote epithelial-mesenchymal transition and stem-like features in mammary epithelium [28], which is directional counter-evidence in the tissue of interest. Separately, six individuals with heterozygous *USP34* loss of function and a neurodevelopmental phenotype including distal limb anomalies have been described, five of them confirmed de novo [29]. Independently of that cohort report, *USP34* is strongly constrained in human population data: the values used here (LOEUF 0.152, 90% CI 0.125 to 0.186; pLI 1.0; observed/expected loss-of-function 70/460.3) were retrieved from the gnomAD v4.1.1 release on 14 August 2026, and the constraint metrics themselves are defined in [30]. These values are attributed here to gnomAD rather than to the cohort report, whose own printed constraint figure reflects a different gnomAD release. Complete loss of one copy in development is not a prediction of partial pharmacological inhibition in an adult, but a gene under this degree of constraint is not a gene to approach casually.

4.5.5 Conclusion of the Assessment

No candidate can be named lowest-risk. The two carrying the fewest documented liabilities, *TLK2* and *VEZF1*, are also the two least studied, so a ranking built on the absence of findings would reward obscurity rather than safety. The pre-analysis plan anticipated naming one candidate at this point. Reporting that the evidence does not support naming one is the correct handling of

Table 5 Liability summary for the four candidates. Baseline dependency is the number of the eleven ER+/luminal lines requiring the gene with no drug present. No candidate is free of liability, and the two with the fewest documented findings are also the two least studied.

Gene	Dependency	Bone	Skeletal muscle	Other disclosed evidence
KDM1A	0 of 11	Documented, direction mixed	Documented, direction unresolved	None
USP34	0 of 11	Documented, loss reduces bone formation	No relevant research found	EMT and stemness counter-evidence; highly constrained, heterozygous loss of function phenotype
TLK2	9 of 11	No relevant research found	No relevant research found	Required for growth in most lines without drug
VEZF1	3 of 11	No relevant research found	No relevant research found	Cardiac structure and contractility phenotype (zebrafish, with cardiomyocyte corroboration); no skeletal-muscle or bone study identified

an asymmetry in publication coverage, and it is the answer to the second half of the question even though it is not the answer the plan expected.

5 Discussion

5.1 The Level at Which the Signal Is Stable

The most informative comparison in this study is between Table 3 and Table 4. The same four datasets were interrogated in both. At the level of individual genes, two of sixteen combinations reached significance, and no candidate was supported in more than one dataset. At the level of Hallmark programmes, twelve of twelve statistics were significant, with consistent direction for two of the three programmes across every dataset.

The straightforward reading is that transcriptional resistance signal in ER+ breast cancer is reproducible as a coordinated programme and is not reproducible as the expression of any particular gene. The datasets differ in cell line, tissue source, patient population and platform, and a programme composed of many genes tolerates that variation in a way that a single transcript does not. This also explains why the screen and the transcriptomic datasets disagree without either being wrong: the screen measures a functional requirement under drug pressure, whereas the transcriptomic datasets measure abundance in resistant or treated states, and a gene can be functionally required without changing its own expression. That was the substance of the audit finding which reinstated *KDM1A* and *TLK2*.

The consequence for candidate selection is uncomfortable but clear. Requiring transcriptomic corroboration of a functional screen hit filters on a property the screen was not measuring. Doing so is defensible as a conservatism device, but it should be declared as a preference rather than presented as validation, and this report does not treat the two combinations in Table 3 as validation of anything.

5.2 Two States, Not a Trajectory

The epithelial-mesenchymal transition result is the single finding most vulnerable to overreading. The programme is enriched in cell-line resistance, in independent resistant sublines and in recurrent tumours, and suppressed after twelve hours of tamoxifen exposure. The tempting sentence is the one claiming that the programme switches on "over time". Nothing in this design supports it, and this report does not make that claim anywhere. Four separate studies with no shared timeline cannot describe a progression, and the twelve-hour dataset is not an early timepoint of the others; it is a different experiment on different material.

What the data support is that acute drug response and long-term resistance are distinct states with opposite behaviour in this programme. The practical consequence matters more than the interpretive one. Any analysis that pools acute treatment data with acquired resistance data in order to increase sample size would average a positive enrichment against a negative one and report an attenuated or absent effect, discarding two real and opposite signals in the process. The result argues for keeping the two contexts separate as a matter of design.

5.3 What the Network Shows and Does Not Show

The proximity of *KDM1A* and *USP34* is a genuine observation about shared functional neighbourhood, and it is a weaker observation than a three-edge path might suggest. All four shortest routes are the same route with an interchangeable ubiquitin gene in the middle, and the intermediate node *DNMT1* is a high-degree hub in this subnetwork. Paths through hubs are cheap: in a network where a small number of nodes connect to a large fraction of the rest, short paths between arbitrary genes are the expected outcome rather than evidence of a relationship.

The *VEZF1* observation is the better supported network result precisely because it is negative. *VEZF1* has no partners at confidence 0.7 and remains isolated at 0.4, while *TLK2*'s separate component dissolves when the threshold is relaxed. An isolation that survives a threshold change carries more information than one that does not, though isolation in STRING reflects how much a gene has been studied as much as how it behaves. Structural proximity in this analysis indicates reachability, not efficacy, and no molecular docking or binding-energy calculation was performed as part of it.

5.4 The Safety Question Cannot Be Answered as Posed

The second half of the question assumed that candidates could be ranked by predicted tissue risk. In practice the evidence base is too uneven for that ranking to mean anything. *USP34* and *KDM1A* have documented tissue roles because they have been studied; *TLK2* and *VEZF1* appear clean because searches in muscle and bone returned nothing. Ranking on what has been published would place the well-characterised genes last and the uncharacterised genes first, which inverts the relationship between evidence and confidence.

There is an asymmetry worth naming in the other direction as well. *KDM1A* is the only candidate in this set for which a small-molecule inhibitor has been used in vivo in the published evidence assembled here, namely CC-90011 in [24]. Tractability of that kind is an argument for attention, not an argument for safety, and it sits alongside the unresolved skeletal-muscle direction rather than outweighing it.

Two further constraints limit what could be concluded even with a complete literature. Published evidence of complete gene deletion, whether in a conditional mouse or in a human loss-of-function cohort, is not a prediction of partial pharmacological inhibition in an adult; the dose, duration and degree of inhibition all differ. And the tissue-level data that would allow the second half of this question to be answered properly does not currently exist in public form, as noted in Section 2.4.

Nothing in this section should be read as a claim that any candidate would reduce the side effects of endocrine therapy. The claim under test was narrower throughout: whether a candidate

would avoid adding to a burden that already exists.

5.5 Reproducibility Challenges

Two problems arose that could not be solved and are reported rather than smoothed over.

5.5.1 The Reconstruction Gap

The malignant-cell classification used for GSE245601 is a reconstruction of the authors' approach, not their published labels. Our pipeline retained 44,140 cells against the 40,428 shown in the published figure, and classified substantially fewer epithelial cells as malignant. Every parameter in that pipeline is tagged by provenance: [A] stated in the published methods, [B] recovered from author code, [C] our own decision. A search of GEO, of the paper's 76 supplementary worksheets, and of the authors' repository found no per-cell labels, and the processed data sit behind dbGaP controlled access, so a cell-by-cell concordance cannot be computed. The pseudobulk results derived from this dataset should be read with that gap in mind: the aggregation step reduces sensitivity to individual cell assignments, but it does not eliminate it.

5.5.2 Format-Forced Method Divergence

The processed expression matrix used in this analysis for GSE118713 was available as TPM. Count-based models could not be applied directly to that processed matrix: edgeR's negative binomial model requires counts, and voom estimates its mean-variance trend from counts. Although raw sequencing reads are available through SRA and could in principle have been re-quantified, this was outside the scope of the present reanalysis; limma on $\log_2(\mathrm{TPM}+1)$ was therefore used for the matrix analysed. One of four datasets was consequently analysed in a different statistical framework from the others, so the four effect estimates are not directly exchangeable and are not treated as such anywhere in this report. The divergence is a consequence of which processed matrix was taken up rather than of any claim that raw data were unavailable, and re-quantification from SRA remains an available route for a future analysis that needs a single common framework.

6 Limitations and Future Work

6.1 Limitations

- The study is entirely computational. There is no laboratory validation of any result reported here. *Next step:* the experiments in Section 6.2, beginning with EXP-1.
- GSE118713 contains three biological replicates per condition, but all conditions derive from a single MCF7-based experimental model, so inference across independent genetic backgrounds is limited and the dataset cannot speak to whether an effect generalises beyond that lineage. The *USP34* result in this dataset should be read with that scope in mind. *Next step:* require corroboration from a design with multiple independent parental backgrounds.
- GSE240112's groups came from two different tissue banks, so disease state and tissue source are perfectly confounded and cannot be separated by any model that could be fitted to these data. *Next step:* a cohort with primary and recurrent samples drawn from one source.
- GSE111151, the only design with multiple independent genetic backgrounds, returned nulls. A post-hoc power calculation using its own fitted dispersions gives a median minimum detectable effect of 0.627 $\log_2$, while the largest observed candidate effect was 0.24. The null is uninformative rather than negative. This calculation is descriptive and serves only to interpret a specific observed null; it is not a design calculation, and observed p-values were not converted into power. *Next step:* a replication design powered for effects near 0.25 $\log_2$.

- Only 3 of the 10 tumour pairs in GSE245601 meet the pre-specified $\geq$50-malignant-cell-per-arm eligibility rule, so the acute-response arm rests on an effective paired pseudobulk sample size of $n = 3$. This is a small number of biological units and the acute-context results should be read as descriptive. *Next step:* a larger paired ex vivo cohort, or a protocol that recovers more malignant cells per arm.
- The network analysis is post-freeze and exploratory. STRING associations are functional and do not establish physical binding, activation or inhibition. *Next step:* treat the *KDM1A* to *USP34* bridge as a hypothesis for targeted co-perturbation, not as a mechanism.
- DepMap describes cancer cell lines, not patients, and baseline dependency in a cell line is a statement about proliferation in culture. *Next step:* the non-cancer cell contexts in EXP-2A and EXP-2B.
- Published evidence of complete gene deletion is not a prediction of partial pharmacological inhibition in an adult. *Next step:* graded knockdown rather than knockout where the question is dose.
- No public tissue data exists from patients on endocrine therapy in the tissues where the symptom burden is felt. *Next step:* this is a deposition problem rather than an analysis problem, and it will not be closed by reanalysis.

6.2 Future Work

The proposed follow-up is genetic rather than pharmacological, because no selective *USP34* inhibitor is available and the question can be asked without one.

- **EXP-1.** Deplete *USP34* in a tamoxifen-resistant ER+ line and re-challenge with tamoxifen, using a full dose-response with Bliss-independence interaction framing rather than a single-dose comparison. EMT and stemness readouts (CDH1, CDH2, SNAI1, AXIN1, active β-catenin) are mandatory given the counter-evidence in [28].
- **EXP-2A.** The same perturbation in MCF10A, testing whether the effect is selective for the cancer context.
- **EXP-2B.** The same perturbation in primary human mesenchymal stromal cells undergoing osteogenic differentiation, given the documented bone role.
- **EXP-3.** An analogous *VEZF1* perturbation with tamoxifen re-challenge.
- **EXP-4.** The same perturbation of *VEZF1* in primary human vascular endothelial cells, prompted by a reported cardiac structure and contractility phenotype in zebrafish, corroborated in primary cardiomyocytes [31].
- **EXP-5.** A targeted test of the *VEZF1*-TEAD1 hypothesis, designed so that it can reject the hypothesis and not only confirm it.

7 Conclusion

A genome-wide CRISPR screen identified thirteen gene knockouts that increase tamoxifen efficacy in ER+ breast cancer cells under a threshold fixed before any result existed, and four were carried forward under two separate and declared rules. Those four did not replicate: two of sixteen gene-by-dataset combinations reached significance, and no candidate was supported by more than one independent dataset. The same four datasets produced twelve of twelve significant enrichment statistics at the level of Hallmark programmes, with oestrogen response and cell-cycle programmes suppressed everywhere and the mesenchymal programme separating long-term resistance from acute treatment as two distinct states. The safety assessment deprioritised *TLK2*

on specificity grounds, since it is required for growth in nine of eleven lines without drug, and found that every remaining candidate carries either a documented tissue liability or an absence of published evidence in the tissues that matter.

The pre-analysis plan anticipated naming one lowest-risk candidate. The evidence does not support that, and this report says so. The finding that generalises beyond this project is the mismatch of levels: functional screens and expression datasets measure different things, and the signal in ER+ tamoxifen resistance is stable as a coordinated programme while being unstable as any individual gene. A pipeline that requires expression corroboration of a functional screen hit will discard the class of modifier the screen was built to find, and a pipeline that pools acute drug response with acquired resistance will cancel two real signals against each other. Both are avoidable, and both are more useful conclusions than a candidate list would have been.

8 Data and Code Availability

All analysis code, the pre-analysis plan, the amendment log and the test suite are available at `https://github.com/almokhtar8-stack/breast_cancer`. The historical evidence-gated shortlist (*USP34*, *VEZF1*, *EML5*, *CITED2*) is frozen at tag `science-freeze-2026-08-15`, commit `9a1b7777d6c69c2be44f16f25bc950769dc2ffda`, with the SHA-256 of the frozen shortlist file recorded as `b6990d7e20f1191df7ebbca18be7542cc312e6e8489f05af6e4b4d31e66fb9dc` and verified unchanged. The post-audit focus set analysed in this report (*KDM1A*, *TLK2*, *USP34*, *VEZF1*) was defined after that freeze, for the reasons given in Section 4.1, and is not the same set as the historical frozen shortlist. Provenance for every figure, including source data, render module and file hashes, is recorded in `poster/figure_manifest.tsv`.

Transcriptomic datasets are available from the Gene Expression Omnibus under accessions GSE118713, GSE111151, GSE240112 and GSE245601. The per-cell malignant labels underlying GSE245601 are not distributed with the GEO record and the corresponding processed data are held under dbGaP controlled access, which is the gap described in Section 5.5.1. Dependency data are from DepMap release 26Q1 (Chronos gene effect). Network data are from STRING for *Homo sapiens* at confidence 0.7, and gene sets are the MSigDB Hallmark collection. The CRISPR screen data are from the source publication cited in Table 1.

No new data were generated by this study.

References

1. David Beckwée et al. “Prevalence of aromatase inhibitor-induced arthralgia in breast cancer: a systematic review and meta-analysis”. In: *Supportive Care in Cancer* 25.5 (2017), pp. 1673–1686. DOI: 10.1007/s00520-017-3613-z.
2. N. Lynn Henry et al. “Predictors of aromatase inhibitor discontinuation as a result of treatment-emergent symptoms in early-stage breast cancer”. In: *Journal of Clinical Oncology* 30.9 (2012), pp. 936–942. DOI: 10.1200/JCO.2011.38.0261.
3. Sara Kim, Nan Chen, and Pankti Reid. “Current and future advances in practice: aromatase inhibitor-induced arthralgia”. In: *Rheumatology Advances in Practice* 8.2 (2024), rkae024. DOI: 10.1093/rap/rkae024.
4. Dina Hany, Vasiliki Vafeiadou, and Didier Picard. “CRISPR-Cas9 screen reveals a role of purine synthesis for estrogen receptor α activity and tamoxifen resistance of breast cancer cells”. In: *Science Advances* 9.19 (2023), eadd3685. DOI: 10.1126/sciadv.add3685.
5. Hyunsoo Kim et al. “Tamoxifen response at single-cell resolution in estrogen receptor-positive primary human breast tumors”. In: *Clinical Cancer Research* 29.23 (2023). Author list per the correction in *Clinical Cancer Research* 2025;31(13):2839, doi:10.1158/1078-0432.CCR-25-1944, pp. 4894–4907. DOI: 10.1158/1078-0432.CCR-23-1248.

6. Broad Institute Cancer Dependency Map. *DepMap 26Q1 Public*. Broad Institute, Cambridge, MA. https://depmap.org/portal/. 2026.
7. Joshua M. Dempster et al. "Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects". In: *Genome Biology* 22.1 (2021), p. 343. DOI: 10.1186/s13059-021-02540-7.
8. Damian Szklarczyk et al. "The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest". In: *Nucleic Acids Research* 51.D1 (2023), pp. D638–D646. DOI: 10.1093/nar/gkac1000.
9. Arthur Liberzon et al. "The Molecular Signatures Database (MSigDB) hallmark gene set collection". In: *Cell Systems* 1.6 (2015), pp. 417–425. DOI: 10.1016/j.cels.2015.12.004.
10. Matthew E. Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (2015), e47. DOI: 10.1093/nar/gkv007.
11. Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2010), pp. 139–140. DOI: 10.1093/bioinformatics/btp616.
12. Yuhan Hao et al. "Dictionary learning for integrative, multimodal and scalable single-cell analysis". In: *Nature Biotechnology* 42.2 (2024), pp. 293–304. DOI: 10.1038/s41587-023-01767-y.
13. Pierre-Luc Germain et al. "Doublet identification in single-cell sequencing data using scDblFinder". In: *F1000Research* 10 (2022). Version 2, peer-review approved, p. 979. DOI: 10.12688/f1000research.73600.2.
14. Christopher S. McGinnis, Lyndsay M. Murrow, and Zev J. Gartner. "DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors". In: *Cell Systems* 8.4 (2019), 329–337.e4. DOI: 10.1016/j.cels.2019.03.003.
15. Timothy Tickle et al. *inferCNV of the Trinity CTAT Project*. Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA. https://github.com/broadinstitute/inferCNV. 2019.
16. Ruli Gao et al. "Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes". In: *Nature Biotechnology* 39.5 (2021), pp. 599–608. DOI: 10.1038/s41587-020-00795-2.
17. Aravind Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550. DOI: 10.1073/pnas.0506580102.
18. Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. "GSEApy: a comprehensive package for performing gene set enrichment analysis in Python". In: *Bioinformatics* 39.1 (2023), btac757. DOI: 10.1093/bioinformatics/btac757.
19. Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring network structure, dynamics, and function using NetworkX". In: *Proceedings of the 7th Python in Science Conference (SciPy 2008)*. Pasadena, CA, 2008, pp. 11–15.
20. Paul Shannon et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". In: *Genome Research* 13.11 (2003), pp. 2498–2504. DOI: 10.1101/gr.1239303.
21. Yu-Chen Guo et al. "Ubiquitin-specific protease USP34 controls osteogenic differentiation and bone formation by regulating BMP2 signaling". In: *The EMBO Journal* 37.20 (2018), e99398. DOI: 10.15252/embj.201899398.

22. Yu-Chen Guo et al. “Erratum to: Ubiquitin-specific protease USP34 controls osteogenic differentiation and bone formation by regulating BMP2 signaling”. In: *The EMBO Journal* 39.20 (2020), e105578. DOI: 10.15252/embj.2020105578.

23. Milica Tosic et al. “Lsd1 regulates skeletal muscle regeneration and directs the fate of satellite cells”. In: *Nature Communications* 9.1 (2018), p. 366. DOI: 10.1038/s41467-017-02740-5.

24. Qingshuang Cai et al. “LSD1 inhibition circumvents glucocorticoid-induced muscle wasting of male mice”. In: *Nature Communications* 15.1 (2024), p. 3563. DOI: 10.1038/s41467-024-47846-9.

25. Hirotaka Araki et al. “LSD1 defines the fiber type-selective responsiveness to environmental stress in skeletal muscle”. In: *eLife* 12 (2023), e84618. DOI: 10.7554/eLife.84618.

26. Petri Rummukainen et al. “Lysine-specific demethylase 1 (LSD1) epigenetically controls osteoblast differentiation”. In: *PLOS ONE* 17.3 (2022), e0265027. DOI: 10.1371/journal.pone.0265027.

27. Jianquan Sun et al. “Histone demethylase LSD1 regulates bone mass by controlling WNT7B and BMP2 signaling in osteoblasts”. In: *Bone Research* 6 (2018), p. 14. DOI: 10.1038/s41413-018-0015-x.

28. Eunhye Oh et al. “Inhibition of ubiquitin-specific protease 34 (USP34) induces epithelial-mesenchymal transition and promotes stemness in mammary epithelial cells”. In: *Cellular Signalling* 36 (2017), pp. 230–239. DOI: 10.1016/j.cellsig.2017.05.009.

29. H. Wigoda, Bryn D. Webb, et al. “USP34 haploinsufficiency as a cause of neurodevelopmental phenotypes”. In: *Clinical Genetics* 110.3 (2026), pp. 315–324. DOI: 10.1111/cge.70194.

30. Konrad J. Karczewski et al. “The mutational constraint spectrum quantified from variation in 141,456 humans”. In: *Nature* 581.7809 (2020), pp. 434–443. DOI: 10.1038/s41586-020-2308-7.

31. Jere Paavola et al. “Vezf1 regulates cardiac structure and contractile function”. In: *EBioMedicine* 51 (2020), p. 102608. DOI: 10.1016/j.ebiom.2019.102608.