GeneBench-Pro Case Study: Transcript-specific lncRNA Dependency versus Local-locus Effects

GeneBench-Pro

June 26, 2026

1 Overview

Long-noncoding-RNA CRISPR interference (CRISPRi) screens often nominate loci before it is clear whether the growth phenotype is caused by the RNA transcript, the promoter, or a nearby coding gene. This case study makes that ambiguity the central bottleneck: the released files contain pooled CRISPRi counts, local-expression readouts, transcript-targeting CasRx follow-up, and a mixed single-guide bridge assay for the nominated lncRNA `LINC473` and nearby coding gene `KIN1`. The reported outputs are two pooled-screen-scale effects and a binary go/no-go decision, `lncrna_specific_lfc`, `neighbor_mediated_lfc`, and `advance_target`. On the released data, the recoverable estimates are a small direct LINC473 transcript effect (-0.048012), a much larger KIN1-mediated effect (-0.638492), and `advance_target`$= 0$, indicating that the apparent pooled-screen phenotype is primarily a local-locus effect rather than evidence for advancing LINC473 as a transcript-directed dependency.

This is a realistic failure mode for lncRNA dependency studies because promoter-targeting CRISPRi can perturb neighboring genes and local regulatory architecture rather than only the RNA product [3, 2, 1, 5], while RNA-targeting CasRx provides an orthogonal follow-up assay whose guide-level activity depends on transcript context and design [6, 7, 4]. The analysis therefore follows the same sequence required by the data: first recover the pooled local CRISPRi model after swap and nuisance correction, then model CasRx along dominant and non-dominant transcript axes, and finally use the bridge assay to put the transcript-targeting follow-up back on the pooled-screen scale. This ordered workflow matters because each shortcut remains scientifically plausible in isolation but changes at least one released answer field enough to fail the public grader contract.

2 Released Prompt and Files

Prompt

```
You are given pooled CRISPRi screening data, guide-level local expression measurements, transcript
    -targeting CasRx follow-up data, and single-guide follow-up growth measurements for a
    nominated lncRNA program (LINC473) and a nearby coding gene (KIN1). The identifiers LINC473,
    KIN1, and ANKRD42 are synthetic benchmark labels; any resemblance to real human genes is
    coincidental.
Estimate the requested quantities.
Definitions:
lncrna_specific_lfc: the pooled-screen matched-control day-10 log2 growth effect expected at 100%
    effective knockdown of the dominant LINC473 transcript, not local DNA-locus effects.
neighbor_mediated_lfc: the pooled-screen matched-control day-10 log2 growth effect expected at
    100% KIN1 repression in the local LINC473-locus model after accounting for concomitant LINC473
     transcript knockdown.
advance_target: 1 if the evidence supports advancing LINC473 as a transcript-directed target, else
     0.
Conventions:
all growth effects are log2(day10/day0) competitive-growth effects relative to matched controls;
more negative numbers indicate stronger loss of fitness;
set advance_target to 1 only if lncrna_specific_lfc <= -0.08 and neighbor_mediated_lfc > -0.25;
    otherwise 0.
These data came from a real experiment; you will be graded not just on numerical correctness but
    the quality of analytical reasoning you exhibit; do not attempt to take any shortcuts.
Return your final answer as exactly one JSON object.
Do not wrap the JSON in markdown.
Do not add prose before or after the JSON.
Do not omit any keys shown in the example.
Return the JSON object in your final answer:
{
  "answer": {
    "advance_target": <int>,
    "lncrna_specific_lfc": <float>,
    "neighbor_mediated_lfc": <float>
  },
  "reasoning": "<description of method and QC>"
}
```

Released data files

File	Format	Contents
`guide_map.tsv.gz`	.tsv.gz	Guide annotations, nominal targets, genomic positions, TSS distances for `LINC473` and `KIN1`, and guide GC fraction.
`crispri_counts.tsv.gz`	.tsv.gz	Day-0 and day-10 pooled CRISPRi guide counts across three replicates.
`local_expression.tsv.gz`	.tsv.gz	Guide-level relative expression of `LINC473`, `KIN1`, `RPLP0`, and `HPRT1`.
`casrx_followup.tsv.gz`	.tsv.gz	CasRx guide efficacy, dominant-isoform overlap, plate identity, and follow-up growth effect.

`guide_followup.tsv.gz`	.tsv.gz	Mixed single-guide bridge assay used to relate pooled-screen and follow-up scales.

3 Answer Fields and Tolerances

The mapping from fitted coefficients to the released answer contract is direct:

$$\texttt{neighbor\_mediated\_lfc} = \widehat{\theta}_{\mathrm{nbr}} = \widehat{\beta}_{\mathrm{nbr}} \tag{1}$$

$$\texttt{lncrna\_specific\_lfc} = \widehat{\theta}_{\mathrm{lnc}} = \widehat{\beta}_{\mathrm{dom}} \tag{2}$$

$$\texttt{advance\_target} = \mathbf{1}\{\widehat{\beta}_{\mathrm{dom}} \le -0.08 \ \wedge \ \widehat{\beta}_{\mathrm{nbr}} > -0.25\}. \tag{3}$$

Answer field	Ground truth	Tolerance / matching rule	Interpretation
`neighbor_mediated_lfc`	-0.638492	Absolute error ≤ 0.030	Stage-1 pooled CRISPRi local-model coefficient for KIN1 repression after swap QC plus GC, promoter-core, and local-expression adjustment.
`lncrna_specific_lfc`	-0.048012	Absolute error ≤ 0.010	lncRNA-specific component after separating transcript and locus effects.
`advance_target`	0	Exact integer match; valid range $\{0, 1\}$	Advancement code from the transcript-specific and neighbor-mediated effect thresholds.

The recoverable reference values are therefore well separated from the main shortcut estimates:

- no swap correction: $(-0.048012,\ -0.497747,\ 0)$;
- no GC adjustment: $(-0.048012,\ -0.715906,\ 0)$;
- no core term: $(-0.048012,\ -0.719684,\ 0)$;
- one-axis CasRx: $(-0.092208,\ 0,\ 1)$;
- global bridge offset: $(-0.118823,\ -0.638492,\ 0)$;
- no plate calibration: $(-0.181311,\ -0.638492,\ 0)$.

4 Structure Diagram

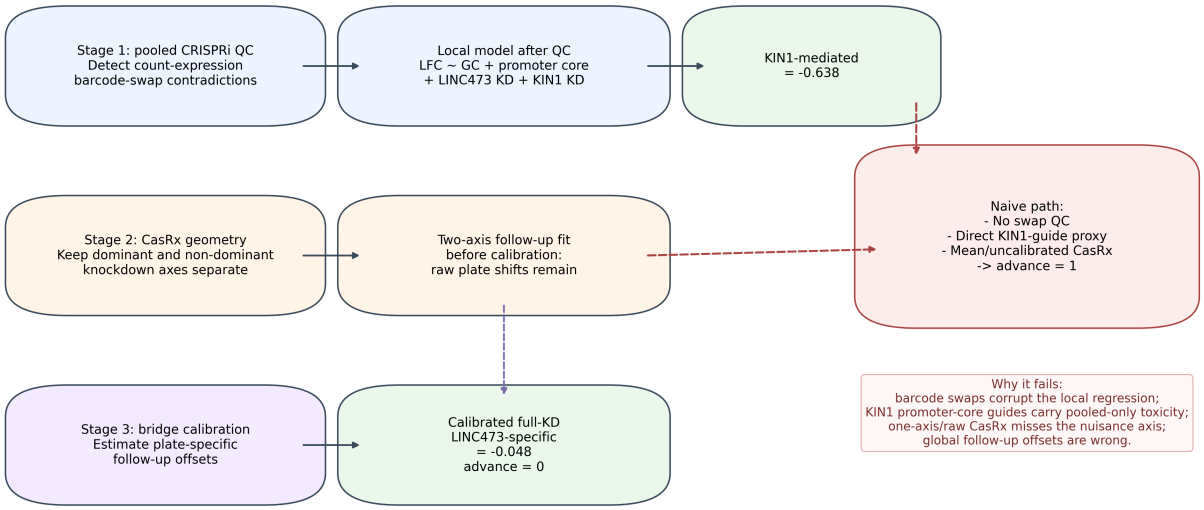

Figure 1: Structure diagram. Blue boxes are Stage 1 pooled CRISPR interference (CRISPRi) quality control and local modeling, orange boxes are Stage 2 RNA-targeting CasRx follow-up modeling, purple is Stage 3 bridge calibration, and green boxes are graded outputs. KD means knockdown and LFC means log2 fold-change. Solid arrows show the reference workflow; the purple dashed arrow shows the uncalibrated CasRx fit entering bridge calibration. Red dashed arrows and the red box mark the tempting but incorrect path: skip swap quality control (QC), use direct KIN1-guide slopes, and treat raw or globally calibrated CasRx as the transcript effect, which would incorrectly set `advance_target` to 1.

5 Variables and Assumptions

Let g index pooled-screen guides, $r \in \{1, 2, 3\}$ index pooled replicates, j index CasRx follow-up guides, and $p \in \{\texttt{plate\_A}, \texttt{plate\_B}, \texttt{plate\_C}\}$ index follow-up plates.

- C_g: guide class. Possible values are `shared_promoter`, `alt_tss_misassigned`, `lnc_exonic`, `neighbor_core`, `neighbor_distal`, `control_nonessential`, and `nontargeting`.
- $\kappa_g^{\mathrm{lnc}} \in [0, 1]$: true LINC473 knockdown fraction for guide g.
- $\kappa_g^{\mathrm{nbr}} \in [0, 1]$: true KIN1 knockdown fraction for guide g.
- $d_g^{\mathrm{lnc}} \in \mathbb{N}$: distance from the guide to the LINC473 TSS in base pairs.
- $d_g^{\mathrm{nbr}} \in \mathbb{N}$: distance from the guide to the KIN1 TSS in base pairs.
- $u_g \in [0, 1]$: guide GC fraction.
- $\Delta_g^{\mathrm{gc}} = \max(u_g - 0.58, 0)$: high-GC excess used in the pooled nuisance model.
- $I_g^{\mathrm{core}} = 1\{d_g^{\mathrm{nbr}} \leq 42\}$: visible KIN1 promoter-core indicator.

- L_g: pooled CRISPRi matched-control guide effect on the pooled-screen scale.
- $N_{0gr}, N_{10gr} \in \mathbb{N}$: day-0 and day-10 pooled CRISPRi counts for guide g in replicate r.
- $E_g^{\mathrm{lnc}}, E_g^{\mathrm{nbr}} \in \mathbb{R}_+$: observed relative expression of LINC473 and KIN1 after guide g.
- $H_g^{\mathrm{RPLP0}}, H_g^{\mathrm{HPRT1}} \in \mathbb{R}_+$: housekeeping expression measurements used for QC only.
- $o_j \in [0,1]$: fraction of the dominant LINC473 isoform overlapped by CasRx guide j.
- $q_j \in [0,1]$: CasRx knockdown efficiency for guide j.
- $D_j = o_j q_j$: effective dominant-isoform knockdown for CasRx guide j.
- $M_j = (1-o_j)q_j$: effective non-dominant knockdown for CasRx guide j.
- Y_j^{cas}: observed CasRx growth effect on the follow-up plate scale.
- Y_g^{bridge}: single-guide follow-up growth effect for bridge guide g.
- δ_p: plate-specific follow-up offset that maps pooled-screen-scale LFCs to the shifted follow-up scale on plate p.

All stochastic components are generated once with RNG seed 1, then held fixed. The public reference values are the recoverable estimates produced by the reference estimator on the released data, not the raw DGP coefficients. In this instance, the recoverable values are very close to the nominal DGP coefficients because the design is constructive and the noise is modest.

6 Data-Generating Process

6.1 Guide classes and latent knockdown structure

The released construction contains 80 pooled guides: 18 shared-promoter LINC473 guides, 8 alternative-TSS-misassigned LINC473 guides, 10 LINC473 exonic guides, 10 KIN1 promoter-core guides, 8 KIN1 promoter-distal guides, 12 ANKRD42 nonessential controls, and 14 nontargeting controls. For each guide g, the latent knockdown fractions and positional annotations are drawn from class-specific uniform ranges:

$$\kappa_g^{\mathrm{lnc}} \sim \mathrm{Unif}(a_{C_g}^{\mathrm{lnc}}, b_{C_g}^{\mathrm{lnc}}) \tag{4}$$

$$\kappa_g^{\mathrm{nbr}} \sim \mathrm{Unif}(a_{C_g}^{\mathrm{nbr}}, b_{C_g}^{\mathrm{nbr}}) \tag{5}$$

$$d_g^{\mathrm{lnc}} \sim \mathrm{DiscreteUnif}(m_{C_g}^{\mathrm{lnc}}, M_{C_g}^{\mathrm{lnc}}) \tag{6}$$

$$d_g^{\mathrm{nbr}} \sim \mathrm{DiscreteUnif}(m_{C_g}^{\mathrm{nbr}}, M_{C_g}^{\mathrm{nbr}}) \tag{7}$$

$$u_g \sim \mathrm{Unif}(a_{C_g}^{\mathrm{gc}}, b_{C_g}^{\mathrm{gc}}). \tag{8}$$

The exact ranges are:

- shared-promoter: $\kappa_g^{\mathrm{lnc}} \in [0.68, 0.95]$, $\kappa_g^{\mathrm{nbr}} \in [0.62, 0.90]$, $d_g^{\mathrm{lnc}} \in [10, 139]$, $d_g^{\mathrm{nbr}} \in [10, 149]$, $u_g \in [0.58, 0.72]$;

- alt-TSS-misassigned: $\kappa_g^{\mathrm{lnc}} \in [0.20, 0.48]$, $\kappa_g^{\mathrm{nbr}} \in [0.52, 0.78]$, $d_g^{\mathrm{lnc}} \in [80, 279]$, $d_g^{\mathrm{nbr}} \in [15, 119]$, $u_g \in [0.54, 0.68]$;
- lnc-exonic: $\kappa_g^{\mathrm{lnc}} \in [0.58, 0.88]$, $\kappa_g^{\mathrm{nbr}} \in [0.08, 0.20]$, $d_g^{\mathrm{lnc}} \in [40, 179]$, $d_g^{\mathrm{nbr}} \in [280, 699]$, $u_g \in [0.36, 0.52]$;
- neighbor-core: $\kappa_g^{\mathrm{lnc}} \in [0.00, 0.08]$, $\kappa_g^{\mathrm{nbr}} \in [0.74, 0.97]$, $d_g^{\mathrm{lnc}} \in [280, 699]$, $d_g^{\mathrm{nbr}} \in [8, 37]$, $u_g \in [0.60, 0.74]$;
- neighbor-distal: $\kappa_g^{\mathrm{lnc}} \in [0.00, 0.10]$, $\kappa_g^{\mathrm{nbr}} \in [0.70, 0.92]$, $d_g^{\mathrm{lnc}} \in [260, 699]$, $d_g^{\mathrm{nbr}} \in [72, 154]$, $u_g \in [0.60, 0.74]$;
- ANKRD42 controls: $\kappa_g^{\mathrm{lnc}}, \kappa_g^{\mathrm{nbr}} \in [0.00, 0.08]$, distances 200 to 699 bp, with six high-GC guides $u_g \in [0.60, 0.72]$ and six low-GC guides $u_g \in [0.36, 0.50]$;
- nontargeting controls: $\kappa_g^{\mathrm{lnc}} = \kappa_g^{\mathrm{nbr}} = 0$, distances 600 to 1199 bp, with seven high-GC guides $u_g \in [0.60, 0.74]$ and seven low-GC guides $u_g \in [0.34, 0.48]$.

6.2 Pooled CRISPRi guide effect

The pooled-screen-scale guide effect is

$$\Delta_g^{\mathrm{gc}} = \max(u_g - 0.58, 0) \tag{9}$$

$$I_g^{\mathrm{core}} = 1\{d_g^{\mathrm{nbr}} \le 42\} \tag{10}$$

$$L_g^* = \beta_{\mathrm{lnc}}\kappa_g^{\mathrm{lnc}} + \beta_{\mathrm{nbr}}\kappa_g^{\mathrm{nbr}} + \beta_{\mathrm{gc}}\Delta_g^{\mathrm{gc}} + \beta_{\mathrm{core}}I_g^{\mathrm{core}}, \tag{11}$$

with $\beta_{\mathrm{lnc}} = -0.048$, $\beta_{\mathrm{nbr}} = -0.684$, $\beta_{\mathrm{gc}} = -0.80$, and $\beta_{\mathrm{core}} = -0.16$. The GC term creates guide-intrinsic toxicity only above GC = 0.58. The core term is the first high-specificity failure mode: tight KIN1 promoter-core guides are more toxic in the pooled CRISPRi screen than their measured KIN1 knockdown alone would predict.

Replicate-specific observed LFCs are

$$\alpha = (0.03,\ 0.11,\ -0.06) \tag{12}$$

$$\varepsilon_{gr} \sim \mathcal{N}(0, 0.035^2) \tag{13}$$

$$L_{gr}^{\mathrm{obs}} = L_g^* + \alpha_r + \varepsilon_{gr}. \tag{14}$$

6.3 Count generation and barcode swaps

Counts are generated from the observed replicate-level LFC:

$$N_{0gr} \sim \max\bigl(250,\ \mathrm{round}\{\mathrm{LogNormal}(\log 1500,\ 0.22^2)\}\bigr) \tag{15}$$

$$N_{10gr} = \max\Bigl(15,\ \mathrm{round}\{N_{0gr}2^{L_{gr}^{\mathrm{obs}}}\}\Bigr). \tag{16}$$

The analysis-normalized matched-control guide effect later uses

$$\widetilde{L}_{gr} = \log_2\left(\frac{N_{10gr}+32}{N_{0gr}+32}\right) - \mathrm{median}\left\{\log_2\left(\frac{N_{10hr}+32}{N_{0hr}+32}\right) : h \in \mathrm{NTC}\right\}. \tag{17}$$

After generating all three replicate count columns, six barcode-swap pairs are applied to the count table only:

$$(\texttt{g003}, \texttt{g067}), (\texttt{g040}, \texttt{g068}), (\texttt{g008}, \texttt{g055}), (\texttt{g015}, \texttt{g069}), (\texttt{g031}, \texttt{g074}), (\texttt{g050}, \texttt{g061}).$$

The local-expression assay is not swapped. This makes count-expression disagreement an explicit asymmetry-breaking signal rather than a hidden trick.

6.4 Local expression assay

For active guides,

$$E_g^{\mathrm{lnc}} = \max(0.03,\ 1 - \kappa_g^{\mathrm{lnc}} + \eta_g^{\mathrm{lnc}}), \quad \eta_g^{\mathrm{lnc}} \sim \mathcal{N}(0, 0.04^2) \tag{18}$$

$$E_g^{\mathrm{nbr}} = \max(0.03,\ 1 - \kappa_g^{\mathrm{nbr}} + \eta_g^{\mathrm{nbr}}), \quad \eta_g^{\mathrm{nbr}} \sim \mathcal{N}(0, 0.04^2). \tag{19}$$

For nontargeting controls, the target-expression readouts are centered near 1 with only measurement noise:

$$E_g^{\mathrm{lnc}} = 1 + \eta_g^{\mathrm{lnc}}, \quad E_g^{\mathrm{nbr}} = 1 + \eta_g^{\mathrm{nbr}}. \tag{20}$$

Housekeeping genes are generated as

$$H_g^{\mathrm{RPLP0}} = \max(0.55,\ 1 + \eta_g^{\mathrm{RPLP0}}), \quad \eta_g^{\mathrm{RPLP0}} \sim \mathcal{N}(0, 0.04^2) \tag{21}$$

$$H_g^{\mathrm{HPRT1}} = \max(0.55,\ 1 + \eta_g^{\mathrm{HPRT1}}), \quad \eta_g^{\mathrm{HPRT1}} \sim \mathcal{N}(0, 0.04^2). \tag{22}$$

These readouts give the correct Stage-1 regressors: observed knockdown fractions $1 - E_g^{\mathrm{lnc}}$ and $1 - E_g^{\mathrm{nbr}}$.

6.5 CasRx follow-up

CasRx is generated constructively as five guide panels with exact efficiencies and overlap values. For each guide j,

$$D_j = o_j q_j \tag{23}$$

$$M_j = (1 - o_j) q_j \tag{24}$$

$$Y_j^{\mathrm{cas},*} = \beta_{\mathrm{lnc}} D_j + \beta_{\mathrm{minor}} M_j + \zeta_j, \tag{25}$$

with $\beta_{\mathrm{minor}} = -0.30$ and panel-specific symmetric deterministic jitter ζ_j of scale 0.002 to 0.003. The visible follow-up value is then shifted by the plate offset:

$$Y_j^{\mathrm{cas}} = Y_j^{\mathrm{cas},*} + \delta_{p(j)}. \tag{26}$$

The exact panel definitions are:

- `plate_A`, near-pure dominant guides `c01--c04`: $q = 0.56$–0.62 and $o = 0.93$–0.96;
- `plate_A`, dominant guides with carryover `c05--c08`: $q = 0.62$–0.71 and $o = 0.72$–0.81;
- `plate_C`, minor-isoform guides `c09--c12`: $q = 0.52$–0.61 and $o = 0.08$–0.14;

- `plate_B`, shared-exon guides `c13--c16`: $q = 0.44$–0.53 and $o = 0.52$–0.64;
- `plate_C`, low-overlap nuisance guides `c17--c20`: $q = 0.21$–0.30 and $o = 0.03$–0.09.

The per-plate offsets are $\delta_{\texttt{A}} = -0.074$, $\delta_{\texttt{B}} = -0.041$, and $\delta_{\texttt{C}} = -0.014$. This is the second high-specificity failure mode: the follow-up plate shift is real and plate-specific, so a single global bridge offset is wrong.

6.6 Single-guide bridge assay

The bridge assay re-measures 18 clean pooled guides on the same shifted follow-up scale as CasRx. For each plate, the bridge deliberately mixes guide classes so it cannot be interpreted as a direct KIN1 or transcript-only assay:

- `plate_A`: one shared-promoter guide, one lnc-exonic guide, one neighbor-core guide, one alt-TSS-misassigned guide, one ANKRD42 control, one NTC;
- `plate_B`: one shared-promoter guide, one lnc-exonic guide, one neighbor-distal guide, one alt-TSS-misassigned guide, one ANKRD42 control, one NTC;
- `plate_C`: one shared-promoter guide, one lnc-exonic guide, one neighbor-core guide, one neighbor-distal guide, one ANKRD42 control, one NTC.

Swapped guide IDs are excluded from bridge selection. The follow-up bridge value for guide g on plate p is

$$Y_g^{\text{bridge}} = \widetilde{L}_g + \delta_p + \xi_g, \tag{27}$$

where $\widetilde{L}_g$ is the observed pooled matched-control guide effect and ξ_g is a small symmetric deterministic jitter of scale 0.003. This bridge is the evidence that lets a careful analyst infer δ_p without being told how.

7 Analyst Walkthrough

7.1 Step 0: Start with the assay geometry

Begin by checking what the released surface actually contains. There are 80 pooled guides, 20 CasRx guides, 18 bridge guides, 14 NTCs, and 12 ANKRD42 nonessential controls. The bridge assay is balanced across plates (6/6/6), and the CasRx panel is visibly structured by both `major_isoform_overlap` and `plate_id`. If this first inventory is skipped and the analysis jumps directly into the nominal `LINC473` guide means, it lands near `nominal_target_mean` $= -0.339356$, which incorrectly triggers `advance_target = 1`.

```
for f in files:
    df = pd.read_csv(f, sep="\t")
    print(f, df.shape, df.columns.tolist())
print(guide_followup["plate_id"].value_counts())
print(casrx_followup["plate_id"].value_counts())
```

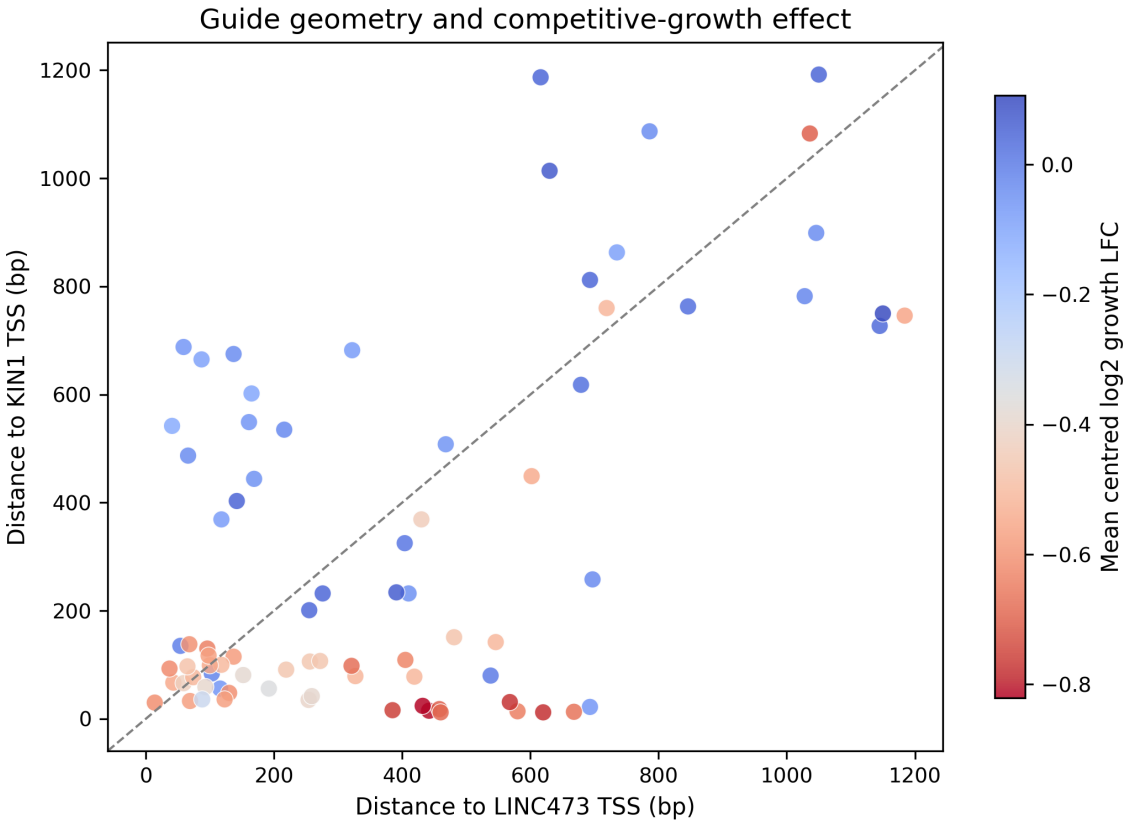

Figure 2: Initial assay overview. Each point is a pooled CRISPRi guide plotted by distance to the LINC473 and KIN1 transcription start sites (TSS; bp, base pairs). Color encodes mean-centered log2 growth fold-change (LFC): redder points are more depleted and bluer points are less depleted or enriched. The gray dashed diagonal is the equal-distance line where a guide is equally far from the two TSSs. The spatial pattern shows why LINC473/KIN1 locus geometry and follow-up plate identity must be checked before using nominal guide means.

7.2 Step 1: Detect the swapped guides before estimating the local model

The first tempting move is to regress pooled growth on measured LINC473 and KIN1 knockdown immediately. That looks reasonable, but on this released dataset it targets the wrong pairing because six count-level barcode-swap pairs break the count-expression relationship. If the Stage-1 model is fit before swap removal, the estimated KIN1-mediated effect is only -0.497747, far from the reference -0.638492.

The diagnostic is visible if you fit a first-pass local model, compute residuals, and look for guides that are impossible under their measured expression. Using the full local model

$$\widetilde{L}_g \sim 1 + \Delta_g^{\mathrm{gc}} + I_g^{\mathrm{core}} + (1 - E_g^{\mathrm{lnc}}) + (1 - E_g^{\mathrm{nbr}}),$$

the residual rule flags eight guide rows: `g003`, `g015`, `g040`, `g050`, `g055`, `g061`, `g068`, and `g069`, all with robust z-scores > 4. These are not all members of all six planted pairs; rather, they are the pair members whose count-expression contradiction is large enough to be detected by the released-data residual rule. The point is not to guess the exact swap mechanism from the prompt; the point is that the released data contain a strong cross-modal contradiction, and a robust residual diagnostic exposes the affected rows [8].

```
beta0 = np.linalg.lstsq(X0, y0, rcond=None)[0]
resid0 = y0 - X0 @ beta0
sigma = median_abs_deviation(resid0) / 0.6745
keep = np.abs(resid0 - np.median(resid0)) < 4 * sigma
```

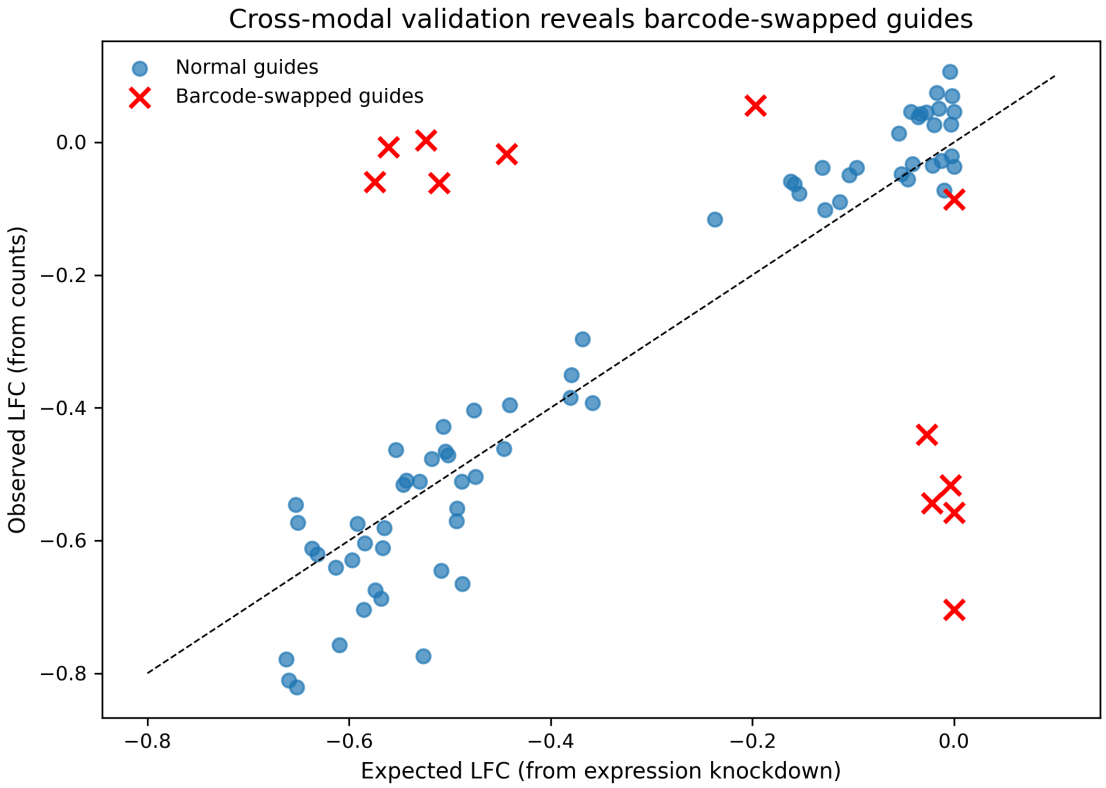

Figure 3: Swap detection. Blue circles are guides consistent with their local-expression measurements, red crosses are barcode-swapped guide rows flagged by the robust residual diagnostic, and the black dashed diagonal is the identity reference where expression-predicted LFC equals count-observed LFC. LFC means log2 fold-change. Points far from the identity line are count-expression contradictions rather than random noise. Removing them is not a cosmetic QC step: it moves the neighbor-mediated estimate from -0.497747 to the correct neighborhood near -0.64.

7.3 Step 2: Fit the pooled-screen local model with both nuisance terms

After removing the swapped guides, you still cannot read the KIN1 effect off a direct slope. Two visible nuisances remain.

First, high-GC controls drift downward even though they do not target either locus. Omitting the $\max(\mathrm{GC} - 0.58, 0)$ term moves the KIN1-mediated estimate to -0.715906, because GC-rich local guides inherit guide-intrinsic toxicity that is not caused by KIN1 repression.

Second, KIN1 promoter-core guides are more depleted than distal KIN1 guides even after accounting for measured KIN1 repression. In the released data, the mean pooled matched-control LFC is -0.6706 for visible core guides and -0.4718 for distal KIN1 guides. If you omit the core indicator, the neighbor estimate shifts to -0.719684.

Figure 4 is the visual check for this Stage-1 model choice. It does not use hidden construction labels: the left panel uses the released `nominal_target` field and the measured local-expression knockdowns to show that nominal LINC473 guides do not define a single transcript-only axis. The right panel uses released guide GC and distance-to-KIN1-TSS fields to show why GC excess and promoter-core status must be carried as nuisance terms rather than folded into the KIN1 coefficient.

At this point, the Stage-1 estimator is a linear guide-level model with the variables needed to make the target identifiable:

$$\widetilde{L}_g = \beta_0 + \beta_{\mathrm{gc}}\Delta_g^{\mathrm{gc}} + \beta_{\mathrm{core}} I_g^{\mathrm{core}} + \beta_{\mathrm{lnc}}(1 - E_g^{\mathrm{lnc}}) + \beta_{\mathrm{nbr}}(1 - E_g^{\mathrm{nbr}}) + \epsilon_g. \tag{28}$$

Refit this model on the non-swapped guides and take $\widehat{\beta}_{\mathrm{nbr}}$ as the pooled-screen-scale neighbor effect.

```
X = np.c_[np.ones(n), gc_excess, is_neighbor_core, kd_lnc, kd_nbr]
beta = np.linalg.lstsq(X[keep], y[keep], rcond=None)[0]
neighbor_mediated_lfc = beta[4]
```

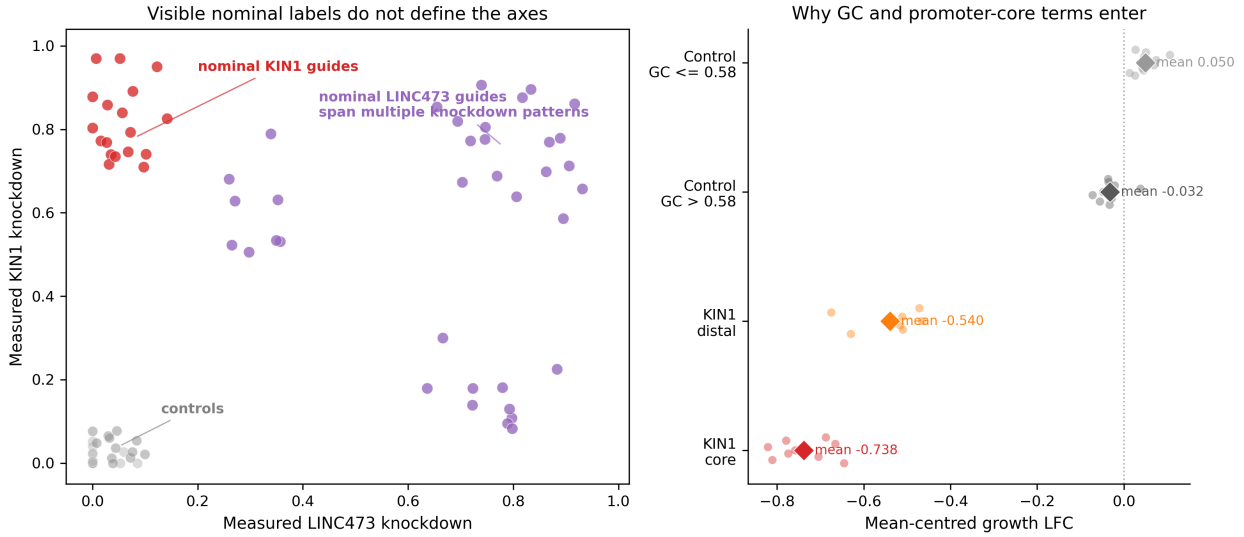

Figure 4: Stage-1 model rationale. Left: each point is a pooled CRISPR interference (CRISPRi) guide plotted by measured LINC473 knockdown and measured KIN1 knockdown from the released local-expression assay; colors encode the released `nominal_target` field. Nominal LINC473 guides span multiple measured knockdown patterns, so the Stage-1 model should use measured LINC473 and KIN1 repression rather than nominal labels alone. Right: the pooled-screen growth LFC summaries use only released guide GC and KIN1-distance fields to show why the Stage-1 local model also needs guide GC excess and KIN1 promoter-core status. High-GC controls drift below low-GC controls, and KIN1 guides within 42 bp of the KIN1 TSS are more depleted than distal KIN1 guides before adjustment.

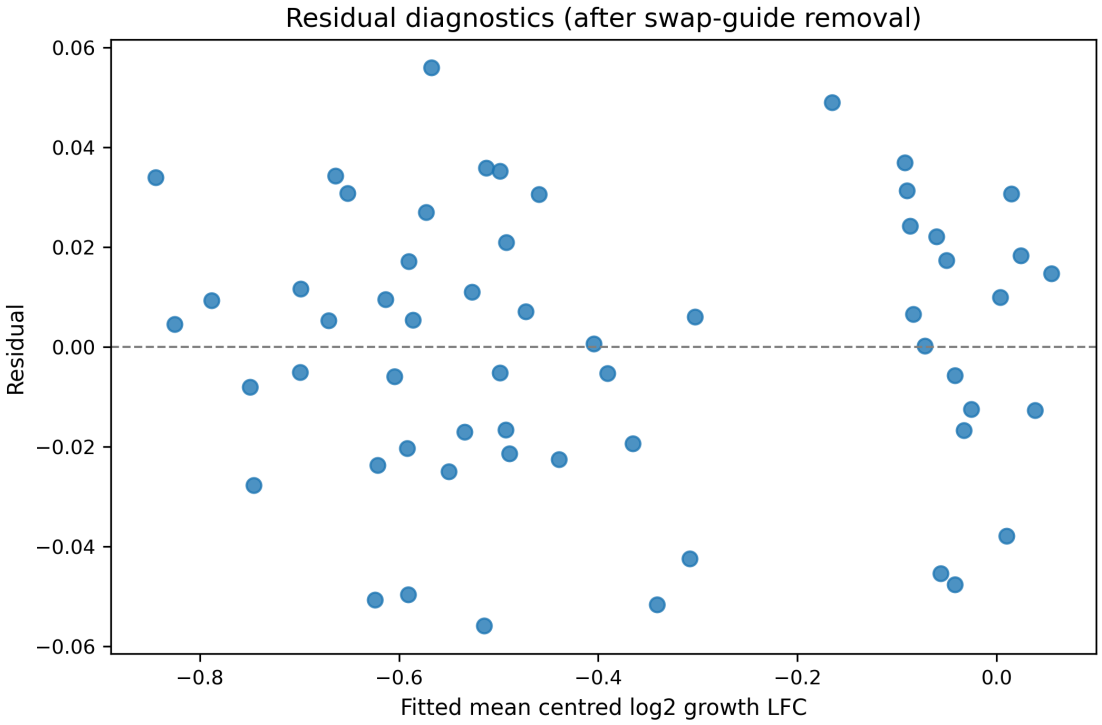

Figure 5: Residual diagnostics. Each point is a non-swapped active pooled CRISPRi guide after fitting the Stage-1 local model with GC excess, KIN1 promoter-core status, LINC473 knockdown, and KIN1 knockdown. The gray dashed horizontal line is zero residual; residuals are observed minus fitted mean-centered log2 growth LFC. Once swaps are removed and the GC/core nuisance structure is included, the remaining residuals are small and the KIN1 coefficient stabilizes at the correct pooled-screen-scale value.

7.4 Step 3: Do not collapse CasRx to a one-axis guide average

Now turn to the transcript-specific effect. The prompt asks for the pooled-screen-scale effect of fully knocking down the *dominant* LINC473 transcript, not the average raw CasRx guide LFC. If you average CasRx guides naively, you get -0.124960, which is much too negative. If you calibrate only roughly but keep a one-axis dominant-only fit, you get -0.092208, still too negative.

The reason is visible in the CasRx design. The plate means are nearly flat: `plate_A` $= -0.1277$, `plate_B` $= -0.1152$, `plate_C` $= -0.1270$, even though the average dominant-overlap is 0.855, 0.580, and 0.085 on those three plates. That is not evidence that dominant-overlap is irrelevant; it is evidence that plate shifts and non-dominant toxicity confound the raw averages.

So define

$$D_j = o_j q_j, \qquad M_j = (1 - o_j) q_j,$$

and keep both. This is the minimal model consistent with the visible assay geometry.

```
cas["eff_dom"] = cas["major_isoform_overlap"] * cas["knockdown_efficiency"]
cas["eff_minor"] = (1 - cas["major_isoform_overlap"]) * cas["knockdown_efficiency"]
```

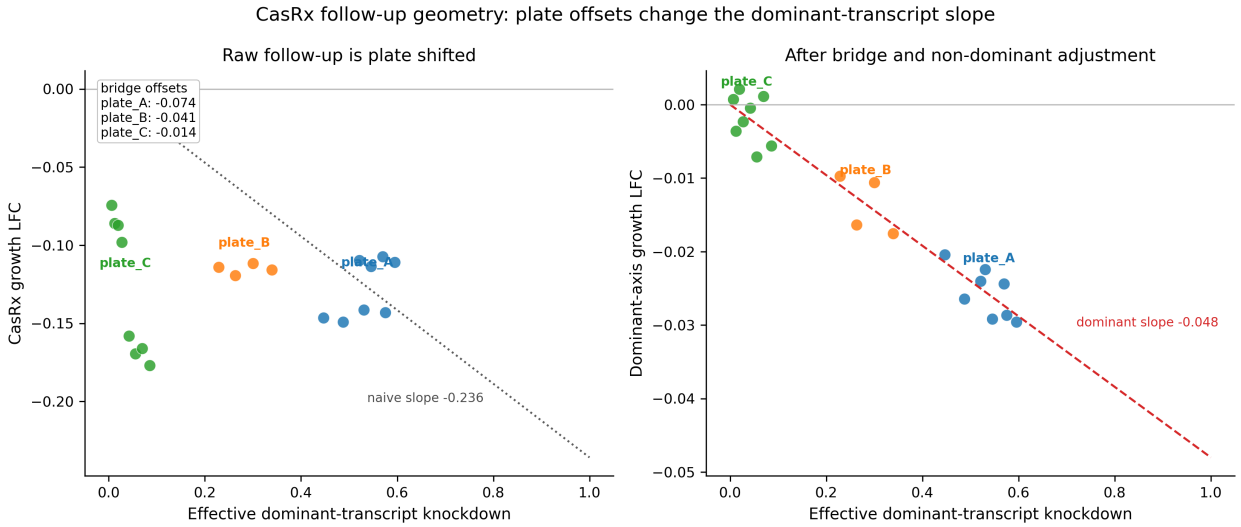

Figure 6: CasRx follow-up geometry. Points are RNA-targeting CasRx guides colored and directly labeled by follow-up plate: blue (`plate_A`), orange (`plate_B`), and green (`plate_C`). The x-axis is effective dominant-transcript knockdown, computed as dominant-isoform overlap times guide knockdown efficiency. The left panel shows raw follow-up growth LFC and the gray dotted naive one-axis slope; the inset lists the bridge-derived plate offsets that make high-overlap guides on `plate_A` incomparable to low-overlap guides on `plate_C` on the raw scale. The right panel subtracts the bridge offset and the fitted non-dominant-transcript component, leaving the dominant-axis growth LFC and the red dashed dominant-transcript slope used for `lncrna_specific_lfc`.

7.5 Step 4: Estimate per-plate offsets from the bridge assay

The bridge assay is the cleanest explicit Stage-3 clue. The same clean pooled guides were re-measured on each follow-up plate, and the bridge follow-up values are almost perfectly rank-preserving relative to pooled-screen LFCs ($r = 0.9966$). But the intercept is not shared across plates. The observed median bridge offsets are

$$\widehat{\delta}_{\mathtt{A}} = -0.073999, \qquad \widehat{\delta}_{\mathtt{B}} = -0.041003, \qquad \widehat{\delta}_{\mathtt{C}} = -0.013987.$$

Each plate mixes controls and active guides, and the within-plate bridge deltas stay tight across those mixed classes, which is the key asymmetry-breaking signal: the plate shift behaves like a genuine follow-up intercept rather than a class-specific biology effect. If calibration is skipped entirely, the dominant CasRx coefficient is -0.181311. If a single global offset is used, it improves to -0.118823, but that is still outside tolerance. This is the same batch-effect logic that appears broadly in high-throughput assays [9]; the bridge design is also conceptually related to bridge-sample calibration strategies [10]. On this released dataset, a two-axis CasRx regression with plate fixed effects also recovers $\widehat{\beta}_{\mathrm{dom}} \approx -0.04808$, so the bridge is not the only algebraic recovery route. Its role is to make those plate terms scientifically interpretable as pooled-to-follow-up scale shifts rather than arbitrary nuisance intercepts.

```
bridge["delta"] = bridge["growth_followup_lfc"] - bridge["pooled_norm_lfc"]
plate_offsets = bridge.groupby("plate_id")["delta"].median()
casrx_cal = casrx["growth_lfc"] - casrx["plate_id"].map(plate_offsets)
```

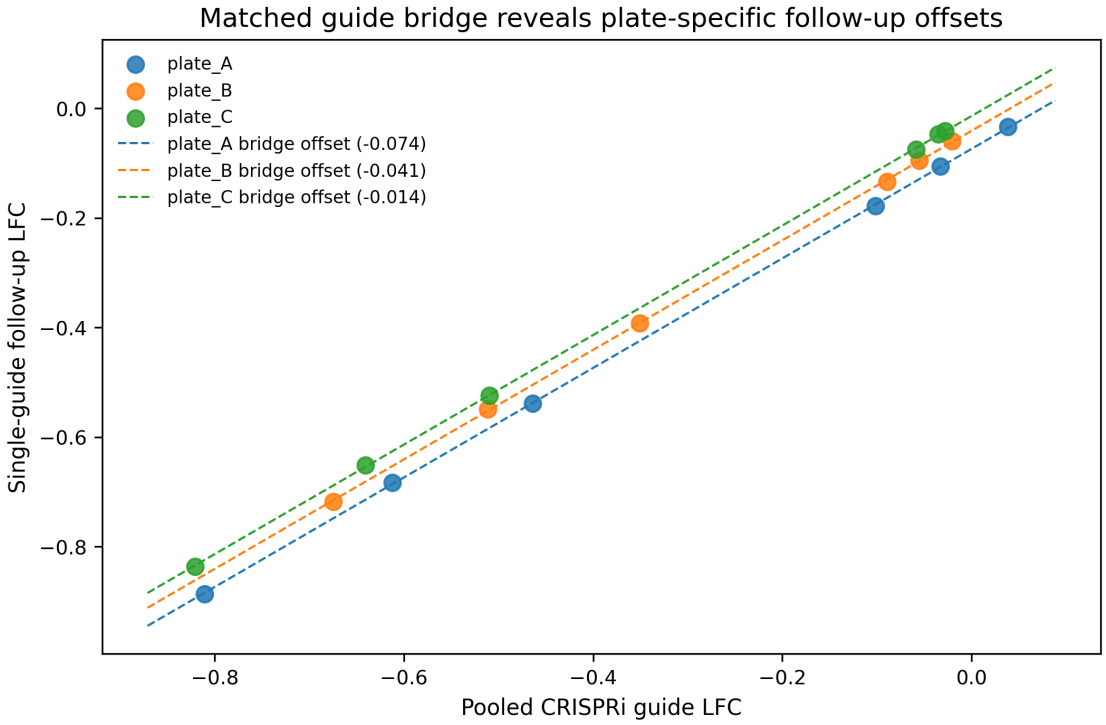

Figure 7: Bridge calibration. Points are bridge guides colored by plate, matching the legend. The x-axis is the pooled CRISPRi guide LFC and the y-axis is the single-guide follow-up LFC for the same guide. Dashed lines are plate-specific calibration lines of the form follow-up LFC = pooled LFC $+\delta_p$, with δ_p shown in the legend for each plate. The vertical separation among dashed lines is the plate offset, which explains why raw or globally calibrated CasRx produces a much too negative transcript effect.

7.6 Step 5: Fit the final transcript model and assemble the answer

After correcting the plate shift, zero dominant and zero non-dominant knockdown should imply zero pooled-screen-scale effect, so fit the calibrated CasRx model through the origin:

$$Y_j^{\text{cas,cal}} = \beta_{\text{dom}} D_j + \beta_{\text{minor}} M_j + \eta_j. \tag{29}$$

In the released data, the raw two-axis fit gives $(\widehat{\beta}_{\text{dom}}, \widehat{\beta}_{\text{minor}}) = (-0.181311, -0.319481)$. After correct per-plate calibration using the bridge-derived offsets, the calibrated dominant coefficient becomes the requested transcript-specific effect.

The final answer fields are therefore assembled as:

$$\texttt{neighbor\_mediated\_lfc} = \widehat{\beta}_{\text{nbr}} \tag{30}$$

$$\texttt{lncrna\_specific\_lfc} = \widehat{\beta}_{\text{dom}} \tag{31}$$

$$\texttt{advance\_target} = \mathbf{1}\{\texttt{lncrna\_specific\_lfc} \le -0.08 \wedge \texttt{neighbor\_mediated\_lfc} > -0.25\}. \tag{32}$$

The realized answer values and tolerance rules are listed in the answer-field table above. The full shortcut outputs are reported in the ablation table.

8 Estimand

The case study has two requested estimands and one deterministic decision rule.

First, the pooled-screen-scale neighbor-mediated effect is the KIN1 coefficient in the correctly specified local pooled CRISPRi model:

$$\theta_{\text{nbr}} = \beta_{\text{nbr}}^{\text{pooled}}, \tag{33}$$

where $\beta_{\text{nbr}}^{\text{pooled}}$ is the coefficient multiplying measured KIN1 knockdown after swap removal and after adjusting for guide GC excess, the promoter-core nuisance term, and concomitant LINC473 knockdown.

Second, the transcript-specific effect is the dominant-axis coefficient in the plate-calibrated CasRx model:

$$\theta_{\text{lnc}} = \beta_{\text{dom}}^{\text{cal}}, \tag{34}$$

where $\beta_{\text{dom}}^{\text{cal}}$ is the pooled-screen-scale effect of one unit of dominant-transcript knockdown after removing plate-specific follow-up offsets using the bridge assay.

The binary decision is

$$\theta_{\text{advance}} = \mathbf{1}\{\theta_{\text{lnc}} \le -0.08 \ \wedge \ \theta_{\text{nbr}} > -0.25\}. \tag{35}$$

These estimands are not the same as the naive nominal-guide means or the raw CasRx averages. Those alternatives fail because they mix transcript and locus effects or compare follow-up measurements across plates without calibration.

9 Estimator

9.1 Stage 1: pooled local model after swap QC

Start from the matched-control pooled guide effect

$$\widetilde{L}_g = \frac{1}{3}\sum_{r=1}^{3}\left[\log_2\left(\frac{N_{10gr}+32}{N_{0gr}+32}\right) - \text{median}_{h\in\text{NTC}}\log_2\left(\frac{N_{10hr}+32}{N_{0hr}+32}\right)\right]. \tag{36}$$

Define observed knockdown fractions

$$\widehat{\kappa}_g^{\text{lnc}} = \max(1-E_g^{\text{lnc}},0), \qquad \widehat{\kappa}_g^{\text{nbr}} = \max(1-E_g^{\text{nbr}},0). \tag{37}$$

Fit the first-pass model

$$\widetilde{L}_g = \beta_0 + \beta_{\text{gc}}\Delta_g^{\text{gc}} + \beta_{\text{core}}I_g^{\text{core}} + \beta_{\text{lnc}}\widehat{\kappa}_g^{\text{lnc}} + \beta_{\text{nbr}}\widehat{\kappa}_g^{\text{nbr}} + \epsilon_g, \tag{38}$$

then compute residuals $\widehat{\epsilon}_g$, the median residual m, MAD = median($|\widehat{\epsilon}_g - m|$), robust scale s = MAD/0.6745, and retain

$$\mathcal{K} = \{g : |\widehat{\epsilon}_g - m| < 4s\}. \tag{39}$$

Refit the same linear model on $g \in \mathcal{K}$. The Stage-1 estimator is

$$\widehat{\theta}_{\text{nbr}} = \widehat{\beta}_{\text{nbr}}. \tag{40}$$

9.2 Stage 2: CasRx effective-knockdown axes

For each CasRx guide j, define

$$D_j = o_j q_j, \qquad M_j = (1 - o_j) q_j. \tag{41}$$

The implementation uses $q_j \geq 0.14$ as a low-activity guardrail; in the realized released data this retains all 20 CasRx guides because the minimum visible knockdown efficiency is 0.21. The important point is not the guardrail itself; it is that the estimator keeps the visible two-axis geometry rather than collapsing everything to a single guide average.

9.3 Stage 3: bridge calibration and calibrated transcript fit

For each bridge guide g on plate $p(g)$, compute

$$\widehat{\delta}_p = \text{median}\left\{ Y_g^{\text{bridge}} - \widetilde{L}_g : p(g) = p \right\}. \tag{42}$$

Then calibrate CasRx back to pooled-screen scale:

$$Y_j^{\text{cas,cal}} = Y_j^{\text{cas}} - \widehat{\delta}_{p(j)}. \tag{43}$$

Finally fit the no-intercept model

$$Y_j^{\text{cas,cal}} = \beta_{\text{dom}} D_j + \beta_{\text{minor}} M_j + \eta_j. \tag{44}$$

The Stage-3 estimator is

$$\widehat{\theta}_{\text{lnc}} = \widehat{\beta}_{\text{dom}}. \tag{45}$$

The no-intercept specification is part of the estimand contract: after bridge calibration, zero effective knockdown on both transcript axes should imply zero pooled-screen-scale effect.

Several numeric choices in this reference estimator are operational conventions supported by the released data rather than universal CRISPRi or CasRx constants. The +32 count pseudocount stabilizes guide-level log-fold changes at the observed count depth; NTC median centering removes replicate-level screen shifts; the 4×MAD residual rule is a robust cross-modal swap diagnostic [8]; the GC excess threshold 0.58 and KIN1 core cutoff 42 bp are visible guide-geometry terms; clipping expression-derived knockdown at zero prevents expression noise from creating negative knockdown; the $q_j \geq 0.14$ CasRx guardrail is inactive in this realized instance because the minimum visible knockdown efficiency is 0.21; and the no-intercept CasRx fit encodes the calibrated zero-knockdown baseline.

10 Decision-Point and Ablation Walkthrough

The table below combines the full ablation suite with the stage at which each shortcut fails. The public target and tolerance windows are listed in the answer-field table above; rows fail when any numeric value exceeds tolerance or when the decision code flips. For compactness, the quantitative-output column abbreviates `lncrna_specific_lfc` as `lnc`, `neighbor_mediated_lfc` as `nbr`, and `advance_target` as `advance`.

Decision point	Analysis / ablation	Quantitative output	Pass?	Failure point	Why the approach is wrong
Reference pipeline	correct_three_stage	lnc -0.048012 (0.00), nbr -0.638492 (0.00), advance 0	yes	none	Reference swap-QC, GC/core-adjusted CRISPRi model, two-axis CasRx fit, and plate bridge calibration.
Target-equivalent check	two-axis CasRx with plate fixed effects	lnc -0.048080 (0.01), nbr -0.638492 (0.00), advance 0	yes	none	Estimates plate shifts directly in the same two-axis CasRx model rather than using the bridge-derived median offsets; the released target is unchanged.
Swap QC	no_swap_correction	lnc -0.048012 (0.00), nbr -0.497747 (4.69), advance 0	no	Stage 1	Leaves barcode-swapped pooled guides in the local model.
Guide nuisance	no_gc_adjustment	lnc -0.048012 (0.00), nbr -0.715906 (2.58), advance 0	no	Stage 1	Confounds guide GC toxicity with neighbor-mediated effect.
Promoter-core nuisance	no_core_term	lnc -0.048012 (0.00), nbr -0.719684 (2.71), advance 0	no	Stage 1	Treats KIN1 promoter-core pooled-only toxicity as ordinary knockdown.
Target definition	nominal_target_mean	lnc -0.339356 (29.13), nbr 0.000000 (21.28), advance 1	no	Stage 1	Uses nominal LINC473 mean rather than matched-control local modeling.
Target definition	top4_nominal_guides	lnc -0.621047 (57.30), nbr 0.000000 (21.28), advance 1	no	Stage 1	Selects the most toxic nominal guides rather than estimating transcript-specific effect.
Local model	lnc_only_regression	lnc 0.065412 (11.34), nbr 0.000000 (21.28), advance 0	no	Stage 1	Ignores neighboring-gene collateral repression.

Decision point	Analysis / ablation	Quantitative output	Pass?	Failure point	Why the approach is wrong
Post-treatment mediator	post_expr_mediation	lnc 0.065379 (11.34), nbr 0.000000 (21.28), advance 0	no	Stage 1	Conditions on LINC473 expression alone and collapses the local model, so KIN1 collateral repression is not estimated.
TSS geometry	nearest_tss_grouping	lnc -0.221817 (17.38), nbr -0.375738 (8.76), advance 0	no	Stage 1	Uses nearest-TSS grouping despite alternative promoter usage.
Neighbor-only shortcut	neighbor_only_plus_nominal_lnc	lnc -0.339356 (29.13), nbr -0.660163 (0.72), advance 0	no	Stage 1	Gets the neighbor scale close while leaving the transcript estimate nominal and wrong.
TSS geometry	proximal_minus_distal_linc473	lnc -0.060638 (1.26), nbr -0.429687 (6.96), advance 0	no	Stage 1	Treats promoter distance as true transcript knockdown.
Target definition	lnc_exonic_only_mean	lnc -0.057644 (0.96), nbr 0.000000 (21.28), advance 0	no	Stage 1	Ignores the neighbor-mediated component of the pooled screen.
Distance proxy	distance_proxy_model	lnc 0.051909 (9.99), nbr -0.605929 (1.09), advance 0	no	Stage 1	Substitutes genomic distance for measured knockdown and collateral structure.
Replicate normalization	pooled_raw_mean_no_center	lnc -0.365769 (31.78), nbr 0.000000 (21.28), advance 1	no	Stage 1	Leaves replicate-specific control baselines uncentered.
CasRx estimator	casrx_unweighted_mean	lnc -0.124960 (7.69), nbr 0.000000 (21.28), advance 1	no	Stage 2	Collapses heterogeneous CasRx guide effects into an unweighted mean.
CasRx estimator	`casrx_calibrated_mean_no_standardize`	lnc -0.036451 (1.16), nbr 0.000000 (21.28), advance 0	no	Stage 2	Calibrates the mean but does not standardize dominant/non-dominant knockdown axes.
CasRx estimator	dominant_only_one_axis	lnc -0.092208 (4.42), nbr 0.000000 (21.28), advance 1	no	Stage 2	Forces dominant and non-dominant transcript effects into one slope.
Bridge calibration	global_bridge_offset	lnc -0.118823 (7.08), nbr -0.638492 (0.00), advance 0	no	Stage 3	Uses one bridge offset despite plate-specific follow-up shifts.
CasRx controls	plate_blind_high_overlap_mean	lnc -0.069447 (2.14), nbr 0.000000 (21.28), advance 0	no	Stage 2	Treats low-overlap guides as plate-blind pseudo-controls.
Transcript geometry	minor_isoform_only	lnc -0.127050 (7.90), nbr 0.000000 (21.28), advance 1	no	Stage 2	Uses the wrong isoform subset for the transcript-specific estimate.

Decision point	Analysis / ablation	Quantitative output	Pass?	Failure point	Why the approach is wrong
CasRx shortcut	top_toxic_casrx_guides	lnc -0.167650 (11.96), nbr 0.000000 (21.28), advance 1	no	Stage 2	Selects toxic CasRx guides instead of modeling effective knockdown.
Promoter-core shortcut	direct_kin1_core_guides	lnc -0.110450 (6.24), nbr -0.670555 (1.07), advance 0	no	Stage 1	Treats KIN1 core-guide toxicity as direct neighbor effect.
Bridge calibration	no_plate_calibration	lnc -0.181311 (13.33), nbr -0.638492 (0.00), advance 0	no	Stage 3	Leaves plate-specific CasRx follow-up offsets uncorrected.
Bridge calibration	casrx_all_mean_no_cal	lnc -0.124960 (7.69), nbr 0.000000 (21.28), advance 1	no	Stages 2–3	Combines CasRx mean collapse with no bridge calibration.
Compound omission	no_swap_no_plate	lnc -0.181311 (13.33), nbr -0.497747 (4.69), advance 0	no	Stages 1 and 3	Leaves both pooled-guide swaps and plate-specific follow-up offsets unresolved.

Table 2: Unified decision-point and ablation walkthrough for the CRISPRi/CasRx transcript-versus-locus problem. Parentheses report tolerance-scaled absolute error for the two numeric fields.

References

[1] Radzisheuskaya A, Shlyueva D, Müller I, Helin K. Optimizing sgRNA position markedly improves the efficiency of CRISPR/dCas9-mediated transcriptional repression. *Nucleic Acids Research*. 2016;44(18):e141. DOI: 10.1093/nar/gkw583.

[2] Liu SJ, Horlbeck MA, Cho SW, et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*. 2017;355(6320):eaah7111. DOI: 10.1126/science.aah7111.

[3] Engreitz JM, Haines JE, Perez EM, et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*. 2016;539:452–455. DOI: 10.1038/nature20149.

[4] Montero JJ, Trozzo R, Sugden M, et al. Genome-scale pan-cancer interrogation of lncRNA dependencies using CasRx. *Nature Methods*. 2024;21(4):584–596. DOI: 10.1038/s41592-024-02190-0.

[5] Replogle JM, Bonnar JL, Pogson AN, et al. Maximizing CRISPRi efficacy and accessibility with dual-sgRNA libraries and optimal effectors. *eLife*. 2022;11:e81856. DOI: 10.7554/eLife.81856.

[6] Konermann S, Lotfy P, Brideau NJ, et al. Transcriptome engineering with RNA-targeting type VI-D CRISPR effectors. *Cell*. 2018;173(3):665–676.e14. DOI: 10.1016/j.cell.2018.02.033.

[7] Wessels HH, Mendez-Mancilla A, Guo X, et al. Massively parallel Cas13 screens reveal principles for guide RNA design. *Nature Biotechnology*. 2020;38:722–727. DOI: 10.1038/s41587-020-0456-9.

[8] Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*. 2013;49(4):764–766. DOI: 10.1016/j.jesp.2013.03.013.

[9] Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*. 2010;11:733–739. DOI: 10.1038/nrg2825.

[10] Xia Q, Thompson JA, Koestler DC. Batch effect reduction of microarray data with dependent samples using an empirical Bayes approach (BRIDGE). *Statistical Applications in Genetics and Molecular Biology*. 2021;20(4–6):101–119. DOI: 10.1515/sagmb-2021-0020.