GUIDE-SPECIFIC VARIANCE AND SHARED CONTROLS OBSCURE GENE-LEVEL RELIABILITY IN HUMAN BETA-CELL CRISPR SCREENS

OpenAI 5.6 Sol
GloGlo Research Team

Abstract. Pooled CRISPR screens are natural benchmarks for computational models of beta-cell stress and function, but only if their gene-level phenotypes reproduce across reagents and controls. We reanalyzed two public genome-scale EndoC-βH1 screens. In the cytokine-relative recovery screen (`GSE205853`), gene-level reliability is 0.017–0.029, with a 95% upper bound near 0.056; 99.07% of single-guide replicate variance is noise. Reusing one untreated or day-zero denominator raises apparent reliability from 0.018 to 0.671 or 0.526 while leaving the treated measurements unchanged. In the insulin-content screen (`PRJEB44712`), the same guide agrees across replicates at Spearman $\rho = 0.334$, whereas independently recomputed disjoint-guide gene scores correlate between -0.0102 and 0.0026. Repeated guide identity and shared references can therefore create apparent reproducibility without a reliable gene-level outcome. Beta-cell model evaluation should begin with reagent-disjoint or biological-unit-independent outcome qualification.

1 Introduction

Genome-scale perturbation screens offer a direct route from a genetic intervention to a cellular phenotype. In type 1 diabetes research, human beta-cell screens have been used to study relative survival under inflammatory stress and genes associated with intracellular insulin content [1, 2]. These screens also appear to provide natural targets for computational prediction: a model can rank thousands of genes, and its ranking can be compared with the deposited screen statistic.

That comparison is informative only if the outcome contains reproducible gene-level variance. Replicate agreement alone does not establish this condition. Replicates may contain the same guides, and guide abundance, cutting efficiency, sequence-specific toxicity, and off-target activity can therefore reproduce without representing a gene effect. Two contrasts may also use the same input or control sample. Noise in that shared denominator then enters both contrasts with the same sign and induces correlation.

The distinction is especially important when the number of scored genes is large. Thousands of genes can make a model comparison numerically precise even when the biological outcome has little capacity to distinguish one predictor from another. A near-zero model score can then mean either that the model is uninformative or that the outcome is not an adjudicating instrument. These explanations cannot be separated by applying another model to the same response.

Date: August 30, 2026.
Key words and phrases. pancreatic beta cell, type 1 diabetes, CRISPR screen, measurement reliability, shared denominator, guide-specific variance.
OpenAI 5.6 Sol performed the computational investigation and prepared the argument and manuscript. The GloGlo research team posed the question, set the research direction, and guided the investigation.

We therefore treat outcome reliability as a property to be estimated before model evaluation. We examine two public EndoC-βH1 CRISPR screens with complementary designs. The cytokine screen in `GSE205853` contains two disjoint GeCKOv2 guide pools, three treated and three untreated culture replicates per pool, and a day-zero representation sample per pool [1]. The insulin-content screen in `PRJEB44712` contains 71,090 base guides recorded under two replicate suffixes and makes it possible to compare same-guide replication with gene scores formed from disjoint guide sets [2]. Together, the datasets expose two failure modes: common denominators can manufacture apparent reliability, and same-guide replication can measure stable reagent behavior rather than gene biology.

2 Reliability estimands

Let an observed gene score be

$$X_g = T_g + E_g, \tag{1}$$

where T_g is stable gene-level variation and E_g contains measurement error and reagent-specific variation not attributable to the gene. Under the classical independent-error model, reliability is

$$R = \frac{\mathrm{Var}(T)}{\mathrm{Var}(X)}. \tag{2}$$

If a predictor is independent of E conditional on T, its maximum Pearson correlation with X is bounded by $\sqrt{R}$ [3, 4]. This familiar attenuation relation is useful only when the repeat measurements isolate T.

For a gene-level target, two admissible repeats must change either the molecular reagents or the independent biological units. Repeating the same guide measures $T_g + W_{ug}$, where W_{ug} is a stable guide-specific effect. Likewise, if

$$X_g^{(1)} = N_g^{(1)} - D_g, \qquad X_g^{(2)} = N_g^{(2)} - D_g, \tag{3}$$

share a noisy log-scale denominator D_g, then $\mathrm{Cov}(X^{(1)}, X^{(2)})$ contains $\mathrm{Var}(D)$ even when the numerators have no common biological signal. Equations (1)–(3) motivate the two empirical tests below.

3 Data and methods

3.1 Cytokine-relative guide-recovery screen

We reconstructed guide counts for `GSE205853` from the public GEO/SRA deposit. The experiment uses GeCKOv2 pools A and B. Each pool contains three cytokine-treated cultures, three matched untreated cultures, and one day-zero representation sample. Counts were normalized within pool by the median-of-ratios method. Guide effects were $\log_2$ ratios with a fixed pseudocount, then averaged by gene. Filters used only day-zero abundance and never the cytokine outcome.

We estimated gene-level reliability by two structures. First, disjoint replicate pairs compared treated replicate i with its own untreated replicate i and treated replicate j with untreated replicate j; no sample was shared between halves. Second, the two GeCKOv2 pools supplied disjoint guide panels, samples, and pool-specific representation controls. The pool-A–pool-B correlation was projected to the combined statistic by the Spearman–Brown formula

$$R_k = \frac{kr}{1 + (k-1)r}. \tag{4}$$

Gene-bootstrap intervals quantify numerical stability across the measured gene universe. They do not turn genes into independent experimental replications.

We also fitted the method-of-moments model

$$V_{ugr} = \mu + G_g + W_{ug} + \varepsilon_{ugr}, \tag{5}$$

where G_g is stable gene variation, W_{ug} is stable variation of guide u nested in gene g, and ε_{ugr} is replicate noise. Gene-level covariance was estimated across the two guide pools; same-guide covariance across culture replicates estimated the sum of gene and guide components. The components were bootstrapped by resampling genes 2,000 times.

To isolate denominator-induced agreement, the treated numerators from replicates 1 and 2 were held fixed under three regimes: separate untreated denominators, one common untreated denominator, or a common day-zero denominator. A deliberately invalid sensitivity analysis selected genes using the deposited false-discovery rate; it was retained as a diagnostic for selection-induced correlation, not as a reliability estimate. Gene-label permutation and a contrast without shared biology were used as empirical nulls.

3.2 Intracellular-insulin-content screen

The `PRJEB44712` supplementary guide table has 142,180 rows. Removing the terminal `_r0` and `_r1` suffixes gives 71,090 base guides measured twice across 18,056 genes; 17,445 genes have four base guides. The outcome is relative guide recovery after fluorescence sorting on intracellular insulin content.

We computed four correlations. The same-guide comparison paired each base guide across replicate suffixes. The same-guide-set gene comparison averaged the same four guides separately within each replicate. Disjoint-guide comparisons used stable-hash 2+2 splits of base guides and averaged both replicate measurements within each half. Fully disjoint comparisons separated both guide identity and replicate suffix. Five stable splits were recomputed independently. The published hit subset was evaluated only to measure the effect of conditioning on the reported outcome.

4 Results

4.1 Gene-level variance in the cytokine screen is small and unresolved

Two independently implemented routes place gene-level reliability between 0.017 and 0.029, with lower confidence limits at or below zero and upper limits of approximately 0.054–0.056. The corresponding point ceilings for a gene-level predictor are Spearman $\rho = 0.125$–0.161, with a 95% upper bound of 0.226. The disjoint-pool result remains small across abundance and pseudocount choices. A separate audit spanning 18 preprocessing combinations obtained pool-A–pool-B Spearman correlations from 0.0079 to 0.0272.

The variance decomposition explains why ordinary replicate agreement can overstate the gene-level quantity. For one guide in one replicate, the estimated variance fractions were 99.07% replicate noise, 0.83% stable guide-specific variation, and 0.10% stable gene variation. The bootstrap interval for the gene component crossed zero. Of the approximately 0.93% that reproduced, about nine parts in ten were guide-specific and one part was gene-level. The model predicted a one-replicate-pair correlation of 0.0141, close to the observed 0.0178, providing an internal check on the decomposition.

The screen's disjoint-pool essentiality contrast was also weak: Pearson $r = 0.0495$ and Spearman $\rho = 0.0350$. The data therefore do not distinguish a globally weak pooled screen from a cytokine

Table 1: Reliability estimates for GSE205853. Intervals are gene-bootstrap intervals and describe numerical stability over the measured gene universe.

Estimator	Gene-level reliability	95% interval	Implied maximum Spearman ρ
Variance components	0.0170	−0.0212 to 0.0541	0.1247
Disjoint guide pools	0.0285	0.0001 to 0.0558	0.1614
Same-guide replicate split	0.0515	0.0269 to 0.0752	0.2171; includes guide effects

Table 2: Shared-denominator sensitivity in GSE205853. Both rows in each comparison use the same treated numerators; only the reference structure changes.

Denominator regime	Pearson r	95% interval	Projected reliability
Separate untreated samples	0.0062	−0.0086 to 0.0208	0.018
One shared untreated sample	0.4042	0.3922 to 0.4163	0.671
One shared day-zero sample	0.2702	0.2563 to 0.2839	0.526

phenotype with particularly little stable gene-level variation. What is identified is the measurement bound for the deposited design, not the biological reason for that bound.

4.2 A shared denominator manufactures agreement

Changing only the denominator altered the correlation by almost two orders of magnitude (Table 2). With separate untreated controls, the two gene contrasts correlated at Pearson $r = 0.0062$ and projected reliability was 0.018. Reusing one untreated sample raised r to 0.4042 and reliability to 0.671. Reusing day zero raised r to 0.2702 and reliability to 0.526. The apparent reliability was therefore inflated 66-fold or 44-fold while the treated numerators were unchanged.

Filtering on the deposited false-discovery rate produced a related artifact. The split-half correlation rose from approximately 0.006 to 0.458, a 75-fold change. Because the filter was computed from the combined outcome, it selected genes whose halves happened to agree and cannot estimate out-of-sample reliability.

4.3 Same-guide replication does not reproduce gene effects

In PRJEB44712, the same base guide correlated across suffix replicates at Spearman $\rho = 0.334$, and gene scores formed from the same four guides correlated at $\rho = 0.271$ (Table 3). Those values collapse when guide identity is removed. Across five independently recomputed stable 2+2 guide partitions, disjoint-guide gene correlations ranged from −0.0102 to 0.0026. Splits that also separated replicate suffixes ranged from −0.0066 to 0.0031.

The control counts themselves correlated at $\rho = 0.971$ across replicate suffixes. Moreover, when analysis was restricted to the published hit subset, all five disjoint-guide correlations were negative, ranging from −0.214 to −0.123. This reversal is consistent with conditioning on apparent replication that was driven by the same guides and nearly identical reference structure. The complete guide universe contains no demonstrable positive gene-level reliability under reagent-disjoint scoring.

Table 3: Replication structures in PRJEB44712.

Comparison	Shared structure	Spearman ρ
Same base guide, r0 versus r1	Guide identity	0.3344
Gene score by replicate	Same four guides	0.2709
Gene score, disjoint 2+2 guide splits	None at guide level	−0.0102 to 0.0026
Disjoint guides and replicate suffixes	Neither guide nor replicate	−0.0066 to 0.0031

5 Discussion

The two screens reveal the same principle through different experimental structures. A repeat measurement must remove the feature whose reliability is being estimated. To estimate a gene effect, it is not sufficient to repeat the same guide. To estimate a ratio or contrast, it is not sufficient to repeat the numerator while keeping the same noisy reference. Same-guide and common-reference comparisons answer legitimate questions about technical repeatability, but they do not estimate gene-level reliability.

This distinction changes how pooled screens should be used as computational benchmarks. A correlation between a predictor and a screen cannot be interpreted without first establishing a nonzero measurement scale for the target. Likewise, a low model correlation against an outcome whose reliability interval includes zero does not compare model quality. The appropriate order is to qualify the outcome, freeze the model and comparator only after the qualification gate is passed, and then score once on a reagent- and biological-unit-independent test.

The results also clarify what deeper screens must add. More genes do not resolve the problem: genome coverage increases the number of rows without creating independent measurements of each gene effect. The informative axes are disjoint reagents, independent donors or differentiation campaigns, and sample-specific input or control measurements. A useful release should expose the complete guide universe, the guide-to-gene map, all sample-level counts, and the biological-unit crosswalk. Reliability should be prespecified as an experimental acceptance criterion rather than reported only after model performance is known.

The biological endpoints remain equally important. The GSE205853 outcome is relative guide representation after cytokine challenge in an immortalized beta-cell line. It does not separate absolute death from proliferation or recovery. The PRJEB44712 outcome is intracellular insulin content, not dynamic glucose-stimulated secretion. A screen intended to nominate cell products or interventions would need separate measurements of absolute live beta-cell number, early death, identity, secretory dynamics, insulin content, and the relevant immune challenge. These measurements address different biological questions and cannot substitute for one another.

5.1 Limitations

The conclusions concern two specific EndoC-βH1 screen designs. They do not imply that cytokine-response or insulin-regulatory gene effects are absent in primary islets or stem-cell-derived islets. There is only one pair of disjoint GeCKOv2 panels in GSE205853; the gene-bootstrap intervals therefore describe stability across genes rather than population variation across independent screens. The variance-component point estimate for the gene term is close enough to zero that its sign is unresolved. The firm result is the small upper bound and its sensitivity to reagent and denominator sharing. In PRJEB44712, four guides per gene permit only 2+2 partitions, so each reagent-disjoint

half is shallow. The consistent collapse across stable partitions establishes that the attractive same-guide number is not a gene-level reliability estimate; it does not estimate the reliability of a future deeper-reagent experiment.

6 A minimum design for an adjudicating beta-cell screen

The failure modes suggest a compact prospective specification.

1. Use enough reagents per gene to form at least two disjoint, adequately powered panels; preserve reagent identities through analysis.
2. Repeat the experiment across independent donors or independently matured differentiation campaigns. Cells, wells, guides, and sequencing reads remain nested observations rather than biological replicates.
3. Give each contrast its own input or control measurement. A bridge standard may connect batches, but it must not be the only denominator shared by the reliability halves.
4. Freeze reagent filters using pre-outcome quantities, and report disjoint-reagent, disjoint-biological-unit, and empirical-null correlations before examining selected genes.
5. Separate relative guide recovery from absolute survival, and separate intracellular content from dynamic secretion and beta-cell identity.

Only after these gates pass does the screen become an instrument for comparing computational predictors. Candidate validation remains a subsequent experiment with independent reagents, rescue, direct fate, function, and identity measurements.

7 Conclusion

The public beta-cell screens examined here reproduce guides and controls more strongly than they reproduce gene effects. In the cytokine screen, gene-level reliability is bounded near zero and can be inflated 44–66-fold by a shared denominator. In the insulin-content screen, a same-guide correlation of 0.334 collapses to approximately zero when guides are disjoint. These are measurement results with a direct consequence for T1D computational research: outcome qualification must precede model evaluation. The next useful dataset is not merely larger. It is designed so that gene-level reliability can be measured independently of guide identity, shared controls, and nested technical observations.

Data and code availability

The source data are public under GEO accession `GSE205853` and ENA project `PRJEB44712`, with the latter's complete guide table provided in the source article supplement. The independent `GSE205853` re-derivation, machine-readable outputs, and analysis scripts are contained in the accompanying GloGlo research release under `validation/outcome_reliability_independent_replication_v1/`. The independent cross-screen adequacy re-derivation is under `validation/independent_data_adequacy_audit_2026_08/`.

References

[1] Benaglio P, et al. Type 1 diabetes risk genes mediate pancreatic beta cell survival in response to proinflammatory cytokines. *Cell Genomics*. 2022;2:100214. doi:10.1016/j.xgen.2022.100214.

[2] Rottner AK, et al. A genome-wide CRISPR screen identifies CALCOCO2 as a regulator of beta-cell function influencing type 2 diabetes risk. *Nature Genetics*. 2023;55:54–65. doi:10.1038/s41588-022-01261-2.

[3] Spearman C. The proof and measurement of association between two things. *American Journal of Psychology*. 1904;15:72–101.

[4] Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison–Wesley; 1968.