GeneBench-Pro Case Study: Ambient-corrected State-restricted scRNA eQTL Estimation

GeneBench-Pro

June 26, 2026

1 Overview

This case study asks an analyst to recover a state-restricted single-cell eQTL effect from a small panel of five genes measured in 588 cells across 24 donors. The released target is a single number: the per-allele log rate ratio for `CXCL10` expression in the activated monocyte subpopulation, with the realized value and tolerance summarized in the answer-field table below. A naive model fit to the raw counts gives a biased estimate outside the released tolerance, because ambient RNA changes both the expression outcome and the recovered cell state.

The case study is a three-stage cascade. Stage 1 is a nuisance-estimation problem: the target gene and the activation markers are all contaminated by ambient RNA from the droplet background, and the analyst must discover this from the empty-droplet file, estimate per-cell contamination fractions using `HBB` as a proxy, and subtract the ambient contribution from `CXCL10`; the reference implementation also applies the same correction to the marker genes. Stage 2A is a state-recovery problem: the analyst must identify which cells are "activated" using the `IFI6`, `ISG15`, and `LST1` marker panel. Corrected-marker scores have a clean threshold at 1.2, while raw-marker scores live on a different numerical scale and are target-equivalent when thresholded at the raw-score gap near 2.5. The failure mode is not raw markers per se, but reusing the corrected-score cutoff on raw scores. Stage 2B is a visible nuisance-recovery problem: the analyst must infer the donor-level technical contamination group from per-donor mean $\hat{\rho}$. This public estimator covariate, $\widehat{B}_i$, is a contamination-derived proxy recovered from released cells and empty droplets; it is distinct from the latent simulated batch label B_i used below to describe the data-generating process. Stage 3 is the count-regression problem: pseudobulk corrected `CXCL10` within the activated state by donor, then fit a Poisson generalized linear model (GLM) with a log-exposure offset and the inferred technical-group covariate. The latent batch label is not provided in the donor metadata, but the adjustment is not a hidden batch convention: the required public covariate is recoverable from the released contamination patterns because per-donor mean $\hat{\rho}$ separates into two non-overlapping groups.

The design is grounded in recent single-cell eQTL literature showing that regulatory effects are strongly state-dependent [1–3], in ambient-RNA work showing that background counts can distort cell-type classification and downstream analysis if not modeled [4–7], and in single-cell analysis guidance that treats the donor rather than the cell as the independent experimental unit for donor-level inference [8].

2 Released Prompt and Files

Prompt

```
Estimate the per-allele log rate ratio for CXCL10 expression in the activated monocyte
    subpopulation from the provided single-cell RNA-seq data.
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": {
    "beta_activated": <float>
  },
  "reasoning": "<description of method and QC>"
}
```

Released data files

File	Format	Contents
`cells.csv.gz`	.csv.gz	One row per cell, with `cell_id`, `donor`, total UMI count, and observed counts for `HBB`, `IFI6`, `ISG15`, `LST1`, and `CXCL10`.
`donors.csv.gz`	.csv.gz	Donor metadata: `donor`, cis genotype dosage `g`, `sex`, `age`, and `bmi`. The technical batch label is intentionally absent.
`empty_drops.csv.gz`	.csv.gz	Empty-droplet barcodes with total UMIs and the same five gene counts, used to estimate the ambient RNA profile.

3 Answer Fields and Tolerances

The released answer contract contains one public field:

$$\texttt{beta\_activated} = \hat{\beta}. \tag{1}$$

The data-generating coefficient is -0.65; the finite-sample maximum-likelihood estimate (MLE) on the released data is the public target.

Answer field	Ground truth	Tolerance / matching rule	Interpretation
`beta_activated`	-0.599956	Absolute error ≤ 0.050	Per-allele log rate ratio for `CXCL10` expression in the recovered activated monocyte subpopulation.

4 Structure Diagram

Figure 1 is the report-level map for the cascaded inference problem. It previews which intermediate decisions are answer-changing: estimating ambient RNA from empty droplets, correcting the `CXCL10` target, using a scale-consistent marker rule for state recovery, recovering the contamination-defined technical group, and fitting the donor-level generalized linear model with both offset and the recovered technical covariate.

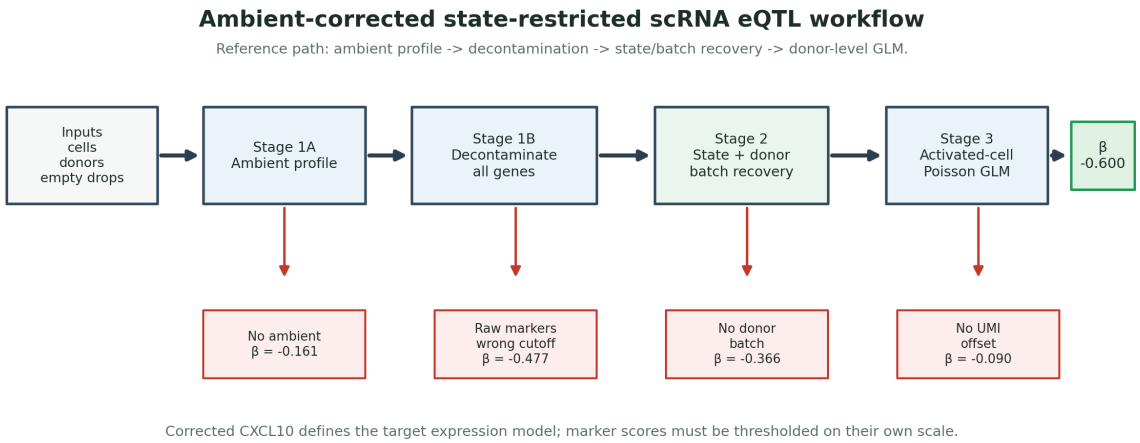

Figure 1: Pipeline view of the ambient-corrected state-restricted single-cell eQTL analysis. Blue boxes and black arrows show the reference workflow from input files through ambient-profile estimation, per-cell decontamination, marker-based state scoring, donor-level recovery of the contamination-defined technical group, activated-cell pseudobulk, and a Poisson generalized linear model (GLM) with offset and recovered technical-group covariate. Red callout boxes and red downward arrows show shortcut analyses and their biased effect estimates. The green output box reports the final per-allele log rate ratio, $\hat{\beta} = -0.600$, for `CXCL10` in activated monocytes. The Stage 2A failure shown here is the scale-mismatch shortcut that applies the corrected-score cutoff to raw-marker scores; raw-marker scores thresholded at their own gap are target-equivalent in this realized data.

5 Variables and Assumptions

- $G_i \in \{0, 1, 2\}$: donor-level cis genotype, with exactly eight donors in each genotype class.
- $B_i \in \{A, B\}$: latent simulated donor batch label used in the data-generating process. All $G_i = 2$ donors and donors `D09`, `D11`, `D13`, and `D15` are in batch B, creating deliberate batch–genotype confounding.
- $u_i \sim \mathcal{N}(0, 1)$: donor-level latent effect that perturbs activation probability and expression rates.
- $S_{ij} \in \{0, 1\}$: latent activation state for cell j from donor i, where 1 denotes activated.
- T_{ij}: total UMI count for cell ij.
- $\rho_{ij} \in [0.02, 0.45]$: cell-specific ambient contamination fraction, higher in batch B and in resting cells.

- π_g^{amb}: ambient per-UMI fraction for gene g, estimated from empty droplets.
- Y_{ijg}: observed count for gene g in cell ij.
- $\widehat{Y}_{ij,g}^{\text{native}}$: estimated native count after subtracting the ambient contribution.
- $\widehat{B}_i = I(\bar{\rho}_i > 0.18)$: public estimator's recovered high-contamination technical-group proxy, computed from released per-donor mean contamination estimates. This proxy is used for adjustment in place of the unreleased latent simulated batch label B_i.

Assumptions visible to the analyst are standard for droplet-based single-cell count modeling: counts are non-negative integers, library size varies across cells, ambient signal can be estimated from empty droplets, and donor-level genotype effects are most stable after pseudobulking [1–6].

6 Data-Generating Process

Donor structure

$$G_i \in \{0, 1, 2\}, \qquad n_{G=0} = n_{G=1} = n_{G=2} = 8. \tag{2}$$

$$B_i = \begin{cases} B, & G_i = 2 \\ B, & \text{donor} \in \{\texttt{D09}, \texttt{D11}, \texttt{D13}, \texttt{D15}\} \\ A, & \text{otherwise} \end{cases} \tag{3}$$

This induces a deliberate batch–genotype correlation: batch B is enriched for high genotypes.

$$u_i \sim \mathcal{N}(0, 1), \qquad n_i^{\text{cells}} \sim \text{Unif}\{18, \ldots, 31\}. \tag{4}$$

Cell-level quantities

$$\text{logit}\{P(S_{ij} = 1)\} = -1.25 + 1.2\, G_i + 0.55\, I(B_i = B) + 0.3\, u_i. \tag{5}$$

Activated cells are more common at higher genotype and in batch B.

$$T_{ij} \sim \text{Poisson}\big(1050 + 130\, S_{ij} - 140\, I(B_i = B) + 50\, u_i\big), \qquad T_{ij} \geq 350. \tag{6}$$

$$\rho_{ij} = \text{clip}\big[\text{logit}^{-1}\big(-2.5 + 1.6\, I(B_i = B) + 0.4\,(1 - S_{ij}) + 0.35\, \varepsilon_{ij}\big),\ 0.02,\ 0.45\big], \tag{7}$$

where $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$. Contamination is heavier in batch B and in resting cells.

Ambient profile

$$(\pi_{\texttt{HBB}}^{\text{amb}},\ \pi_{\texttt{IFI6}}^{\text{amb}},\ \pi_{\texttt{ISG15}}^{\text{amb}},\ \pi_{\texttt{LST1}}^{\text{amb}},\ \pi_{\texttt{CXCL10}}^{\text{amb}}) = (0.070,\ 0.045,\ 0.035,\ 0.009,\ 0.050). \tag{8}$$

The elevated `IFI6` and `ISG15` in the ambient pool reflect lysed activated cells contributing interferon-stimulated transcripts to the droplet background. This is the mechanism that shifts raw marker scores onto a different threshold scale.

Native expression fractions

In activated cells:

$$f^{\text{act}}_{\texttt{IFI6}} = 0.055, \qquad f^{\text{act}}_{\texttt{ISG15}} = 0.045, \quad f^{\text{act}}_{\texttt{LST1}} = 0.003, \tag{9}$$

$$f^{\text{act}}_{\texttt{CXCL10}} = \exp\big(-6.0 + 2.0 - 0.65\, G_i + 0.60\, I(B_i = B) + 0.12\, u_i\big). \tag{10}$$

In resting cells:

$$f^{\text{rest}}_{\texttt{IFI6}} = 0.0005, \qquad f^{\text{rest}}_{\texttt{ISG15}} = 0.0004, \quad f^{\text{rest}}_{\texttt{LST1}} = 0.065, \tag{11}$$

$$f^{\text{rest}}_{\texttt{CXCL10}} = \exp\big(-6.0 + 0.20\, G_i + 0.60\, I(B_i = B) + 0.12\, u_i\big). \tag{12}$$

Note the opposite-sign genotype effect: the activated coefficient is -0.65 (strong negative) while the resting coefficient is $+0.20$ (weak positive). This amplifies the bias when resting cells are misclassified as activated. For `HBB`, native expression is set to zero in every cell,

$$f^{\text{native}}_{ij,\texttt{HBB}} = 0 \quad \text{for all donors, cells, and activation states.} \tag{13}$$

Thus observed `HBB` counts arise from the ambient term only, making `HBB` an ambient-contamination proxy rather than an endogenous expression target.

Count model

$$\mu_{ijg} = T_{ij} \left[(1 - \rho_{ij})\, f^{\text{native}}_{ijg} + \rho_{ij}\, \pi^{\text{amb}}_{g} \right]. \tag{14}$$

$$\lambda_{ijg} \sim \text{Gamma}(k = 8,\ \theta = \mu_{ijg}/8), \qquad Y_{ijg} \sim \text{Poisson}(\lambda_{ijg}). \tag{15}$$

Empty droplets:

$$T^{\text{empty}}_{m} \sim \text{Poisson}(180), \qquad Y^{\text{empty}}_{mg} \sim \text{Poisson}(T^{\text{empty}}_{m}\, \pi^{\text{amb}}_{g}), \tag{16}$$

for $m = 1, \ldots, 120$.

7 Analyst Walkthrough

Domain primer. In modern single-cell eQTL analyses, a "cell type" is often too coarse. The same locus can have a strong effect in one activation state and little or no effect in another, so state recovery can be part of the estimator rather than a convenience [1–3]. At the same time, droplet-based counts are vulnerable to ambient RNA from lysed cells, and that contamination can distort both expression estimates and cell-type classification if you ignore empty droplets [4–6].

Step 1: Notice that multiple genes are contaminated

A naive analysis would sum raw `CXCL10` across all cells from each donor and fit a Poisson model with genotype and batch. That gives a biased estimate, which may look stable but is far from the activated-state target. The first data signature appears before any model fitting: empty droplets

show a large HBB component (7.0% of UMIs), but also substantial IFI6 (4.5%) and ISG15 (3.5%). These are activation markers in the ambient pool, indicating that the droplet background carries an interferon signature from lysed activated cells. CXCL10 is also present at 5.0%.

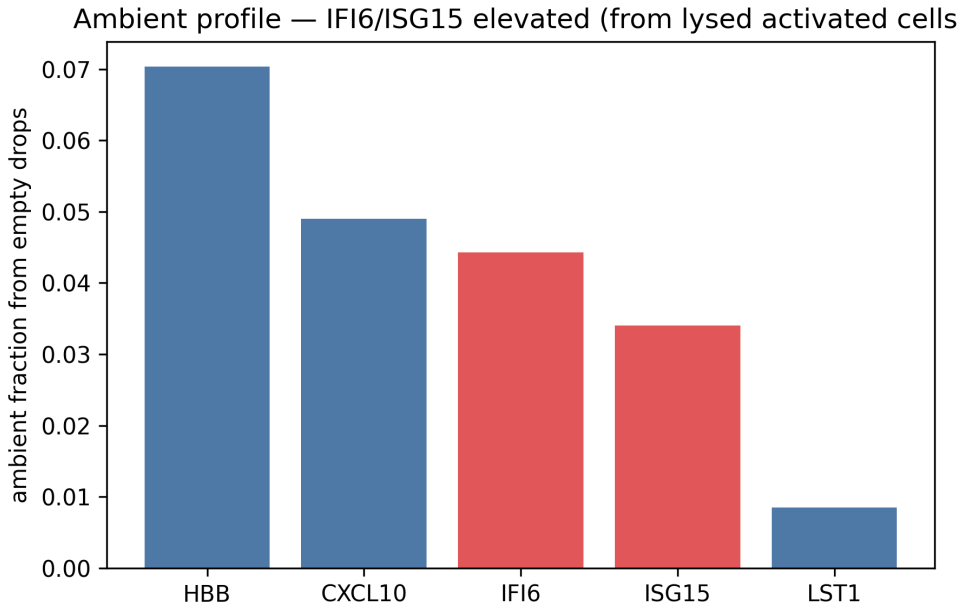

Figure 2: Ambient RNA profile estimated from empty droplets. Bar height is the fraction of empty-droplet UMIs assigned to each gene. Blue bars mark the ambient proxy or target/context genes (HBB, CXCL10, and LST1); red bars mark activation markers (IFI6 and ISG15) whose ambient signal can distort state calling. HBB dominates and provides a contamination proxy, while elevated IFI6/ISG15 show that the activation markers themselves are contaminated.

A careful analyst would next check whether cells with more apparent contamination also have higher raw CXCL10. They do:

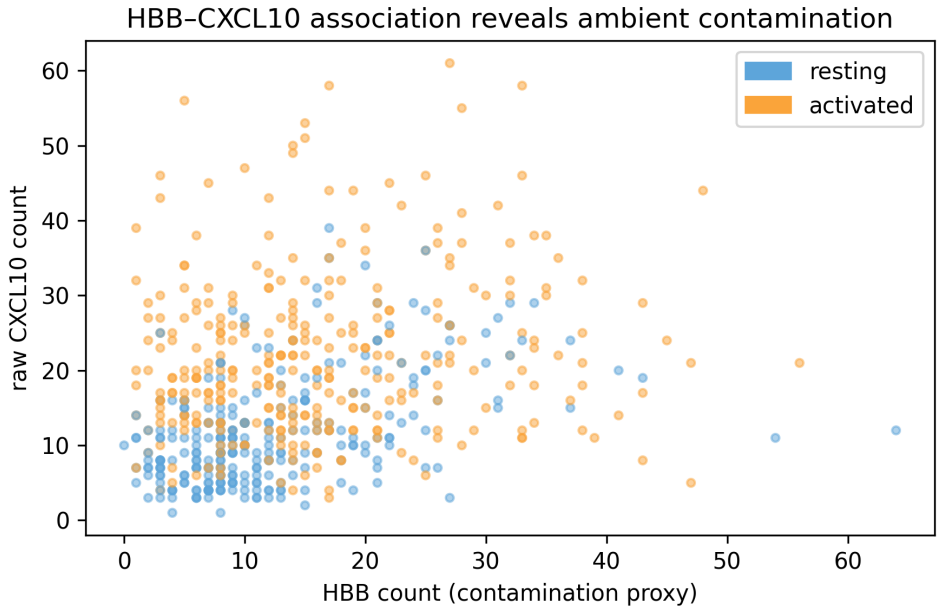

Figure 3: Raw `HBB` and raw `CXCL10` counts are positively associated across cells. Each point is one cell; the x-axis is raw `HBB` count, used as an ambient-contamination proxy, and the y-axis is raw `CXCL10` count, the target gene. Blue and orange points denote latent resting and activated states for expository validation only; the analyst-visible diagnostic is the count association between the ambient marker and target gene. Resting cells with high `HBB` have inflated `CXCL10` from ambient RNA rather than native expression.

The corresponding calculation is:

```
prof = empty[genes].sum() / empty["total_umi"].sum()
rho_hat = clip(HBB / (total_umi * prof["HBB"]), 0, 0.5)
```

Step 2: Estimate ambient contamination and correct the target

The answer-changing ambient correction is the subtraction of ambient contribution from `CXCL10`, the target gene. The reference implementation also subtracts ambient contribution from the marker genes before computing the activation score, which puts the marker score on a cleaner and more interpretable scale. The ambient pool has 4.5% `IFI6` and 3.5% `ISG15`, both activation markers. For contaminated resting cells in batch B ($\rho \approx 0.31$, $T \approx 900$), the ambient contribution to raw `IFI6` is roughly $900 \times 0.31 \times 0.045 \approx 12.6$ counts, dwarfing the native resting expression of $900 \times 0.69 \times 0.0005 \approx 0.3$ counts. The practical consequence is that raw and corrected marker scores require different thresholds; it is not evidence that marker correction itself is required to recover the final `beta_activated` answer.

```
for gene in ["CXCL10", "IFI6", "ISG15", "LST1"]:
    corr = max(raw - total_umi * rho_hat * prof[gene], 0)
```

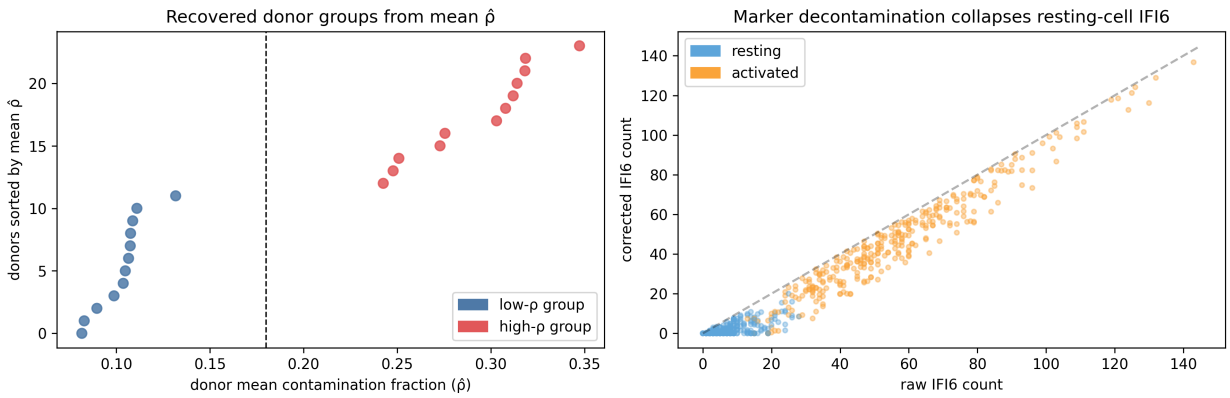

Figure 4: Left: donor mean contamination estimates, sorted by $\hat{\rho}$. Blue points denote the low-contamination donor group, red points denote the high-contamination donor group, and the black dashed vertical line marks the threshold $\hat{\rho} = 0.18$ lying in the empty interval between groups. Right: raw versus corrected `IFI6` counts for individual cells; the gray dashed diagonal is the identity line where correction would not change the count. Blue and orange points denote latent resting and activated states for expository validation only. The analyst-visible patterns are the donor-level contamination gap and the shift from raw to corrected marker scale after ambient subtraction.

Step 3: Recover the activated state from marker scores

The reference implementation computes the activation score from the corrected markers:

$$A_{ij} = \log(1 + \texttt{IFI6}_{ij}^{\text{corr}}) + \log(1 + \texttt{ISG15}_{ij}^{\text{corr}}) - \log(1 + \texttt{LST1}_{ij}^{\text{corr}}).$$

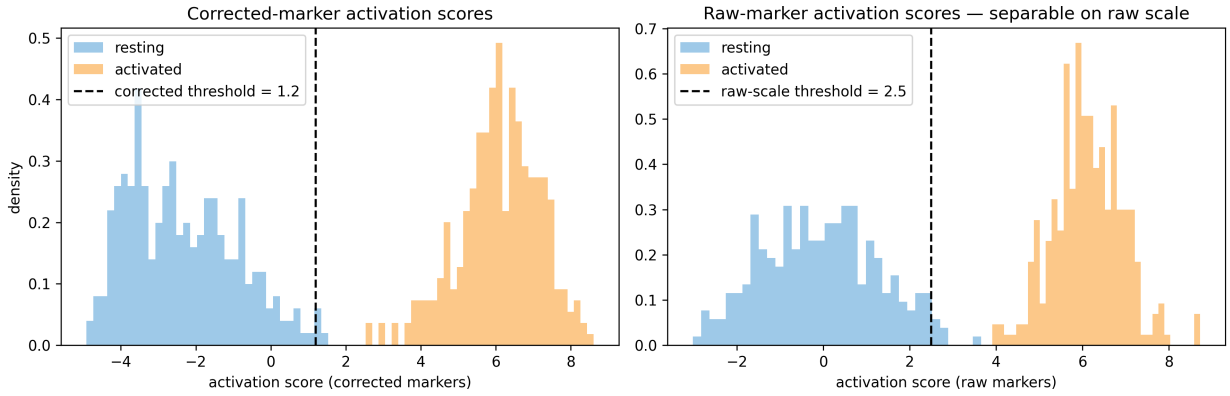

Figure 5: Activation-score distributions from marker counts. Blue and orange histograms denote latent resting and activated cells for expository validation only. Left: corrected-marker scores use the black dashed threshold $A_{ij} = 1.2$, which lies in the corrected-score gap. Right: raw-marker scores use the black dashed threshold $A_{ij}^{\text{raw}} = 2.5$, which lies in the raw-score gap. Both marker-score scales separate the activated mode well enough for the final donor model when `CXCL10` is ambient-corrected; the invalid shortcut is applying the corrected-score threshold of 1.2 to raw-marker scores.

The corrected activation scores have a clean threshold gap: the largest score below the 1.2 cutoff is 1.164, the smallest score above it is 1.238, and thresholds from 1.0 through 1.5 recover estimates within tolerance. Raw marker scores are shifted upward by ambient `IFI6`/`ISG15`, so they should

not reuse the corrected-score threshold. Thresholding the raw score at the visible raw-scale gap, $A_{ij}^{\text{raw}} > 2.5$, gives $\hat{\beta} = -0.590$ (error = 0.010), and a threshold of 3.0 gives $\hat{\beta} = -0.602$ (error = 0.003). In contrast, applying the corrected-score cutoff of 1.2 to raw-marker scores gives $\hat{\beta} = -0.477$ (error = 0.123, or 2.5× the tolerance) because it overcalls contaminated resting cells as activated. The necessary step is therefore activated-state recovery with a scale-consistent marker rule, not marker decontamination as a private threshold convention.

Using only `IFI6` to define activation gives a worse estimate, because a single corrected marker is noisier than the composite score.

Step 4: Recover the contamination-defined technical group

The donor metadata provides genotype, sex, age, and BMI, but not the latent simulated batch label. The analyst does not need that unreleased label directly. Instead, the public estimator recovers a visible technical-group proxy from contamination: compute per-donor mean $\hat{\rho}$ (from Step 1B) and plot it. The 24 donors split into two cleanly separated groups: 12 low-contamination donors with $\hat{\rho} \approx 0.08$–0.13 and 12 high-contamination donors with $\hat{\rho} \approx 0.24$–0.35, with no overlap. In the data-generating process these groups correspond to batch A and batch B, respectively; alternatively, clustering donors by mean `HBB` reveals the same two groups.

This recovered grouping corresponds to a strong technical effect: high-contamination donors have ∼4× higher contamination, lower library sizes, and higher activation probability. Crucially, all $G = 2$ donors and half of $G = 1$ donors are in the high-contamination group, creating technical-confounding of genotype that biases the estimate if not adjusted for. The adjustment is therefore an observed-data technical nuisance correction, not a request to infer unreleased metadata. The left panel of Figure 4 is the visible diagnostic: the two donor groups are separated before any genotype model is fit.

The donor table includes uninformative covariates (`age`, `bmi`, and `sex`). Analyses that include only these covariates without discovering the contamination-defined technical group get a biased answer ($\hat{\beta} \approx -0.37$, error = 0.23). Including `sex`, `age`, and `bmi` in addition to the recovered technical group gives $\hat{\beta} = -0.558$, still within the released tolerance; the critical decision is recovering the technical group, not excluding ordinary covariates.

```
donor_rho = cells.groupby("donor")["rho_hat"].mean()
batch = (donor_rho > 0.18).astype(int) # clean gap at ~0.18
```

Step 5: Fit the state-restricted donor model with offset and batch

After ambient subtraction, state recovery, and batch discovery, aggregate corrected `CXCL10` and total UMIs across activated cells within each donor, then fit the Poisson rate model:

$$\log E(Y_i) = \alpha + \beta\, G_i + \gamma\, I(B_i = B) + \log(T_i).$$

If the offset is omitted, the estimate shifts substantially. If the batch covariate is dropped, $\hat{\beta}$ also shifts because batch B is enriched for high genotype. With no batch and offset only, $\hat{\beta} = -0.366$ (error = 0.234). Using an approximate corrected/native exposure offset rather than total activated-cell UMIs gives $\hat{\beta} = -0.611$, and fitting the Poisson estimating equation to fractional corrected counts rather than rounded corrected counts gives $\hat{\beta} = -0.600$; both are target-equivalent in this instance.

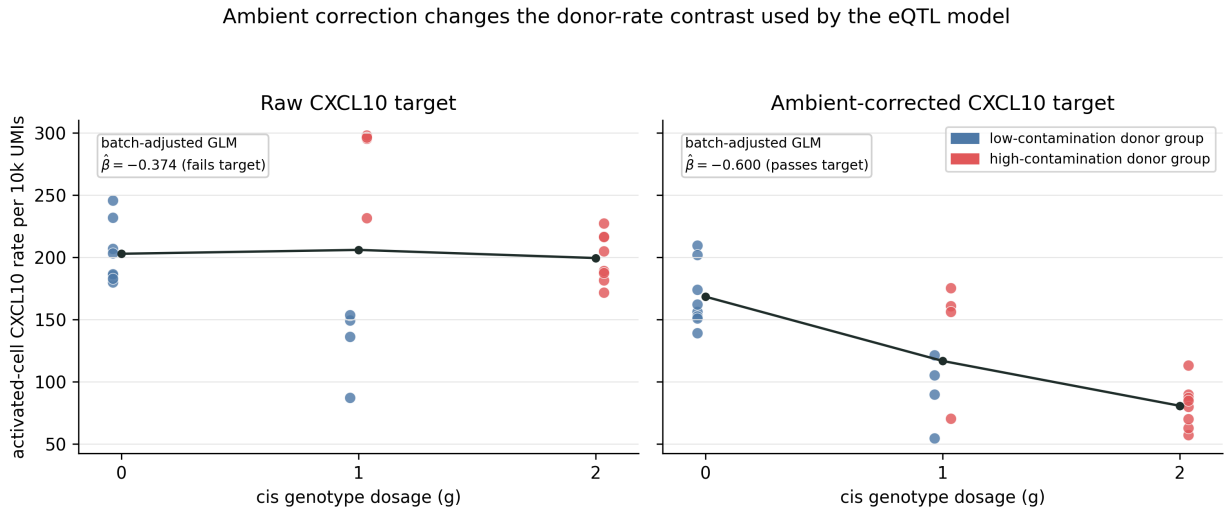

Figure 6: Donor-level activated-cell `CXCL10` rates before and after ambient correction. Each point is one donor after corrected-marker state recovery; the y-axis is the activated-cell `CXCL10` rate per 10,000 UMIs, and the x-axis is cis genotype dosage g. Blue points denote the low-contamination donor group and red points denote the high-contamination donor group recovered from mean $\hat{\rho}$ in Figure 4; the black line connects genotype-specific mean rates within each panel. The left panel uses raw `CXCL10` counts and gives a batch-adjusted donor GLM estimate $\hat{\beta} = -0.374$, outside tolerance. The right panel uses ambient-corrected `CXCL10` counts and gives $\hat{\beta} = -0.600$, the public answer field `beta_activated`.

Reference implementation sketch:

```
pb = activated.groupby("donor").agg(
    cxcl10_native=("CXCL10_corr", "sum"),
    total_umi=("total_umi", "sum"))
glm = Poisson(cxcl10_native ~ g + batchB,
              offset=log(total_umi))
```

Step 6: Check the model fit

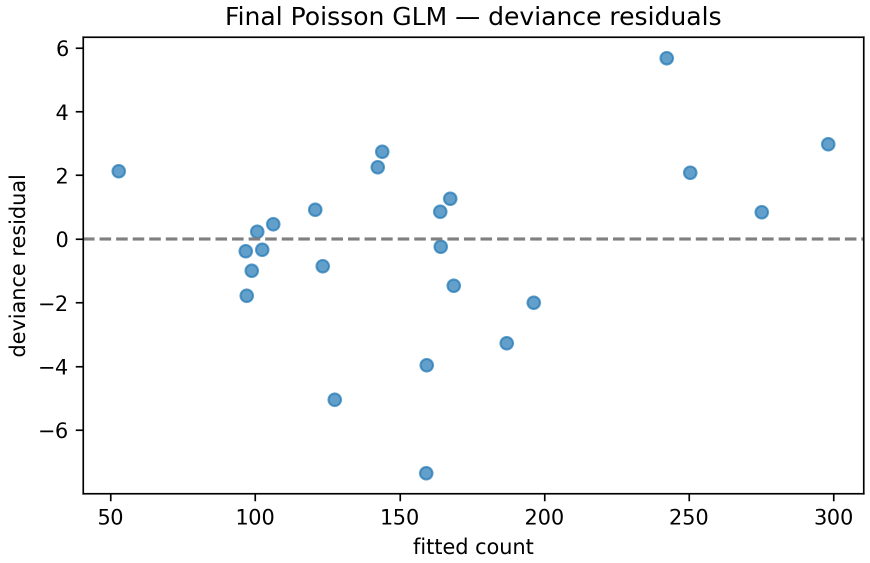

Figure 7: Deviance residuals from the final Poisson generalized linear model (GLM). Each point is one donor after activated-cell pseudobulking; the x-axis is fitted `CXCL10` count and the y-axis is the deviance residual. The gray dashed horizontal line marks zero residual. The plot is used as a visual model-fit diagnostic and shows no visually obvious single-donor outlier in the fitted-count range.

This residual check is not an additional estimator choice; it is the final diagnostic that the donor-level Poisson model used for `beta_activated` does not show an obvious fitted-count pattern or single-donor residual that would force a different public analysis.

Final answer assembly. The final model contributes three fitted terms: the intercept $\hat{\alpha} = -4.079$, the per-allele genotype coefficient $\hat{\beta} = -0.5999557$, and the high-contamination technical-batch coefficient $\hat{\gamma} = 0.471$. For donor i, the fitted activated-cell native `CXCL10` mean is

$$\widehat{\mu}_i = T_i^{\text{act}} \exp(-4.079 - 0.5999557\, G_i + 0.471\, I(\widehat{B}_i = B)).$$

The reported answer is the genotype component, `beta_activated` $= -0.5999557$, rounded in prose to -0.600.

8 Estimand

The requested estimand is the activated-cell donor-level log rate ratio

$$\beta^\star = \arg\min_{\beta}\ \mathrm{KL}\Big(Y_i^{\text{act,corr}} \,\Big\|\, \mathrm{Poisson}\big(T_i^{\text{act}} \exp(\alpha + \beta\, G_i + \gamma\, I(B_i = B))\big)\Big), \tag{17}$$

where $Y_i^{\text{act,corr}}$ is the donor-level sum of ambient-corrected `CXCL10` counts over activated cells and T_i^{act} is the corresponding activated-cell total UMI exposure. In the reference implementation, activated cells are identified by corrected-marker scores; raw-marker scores thresholded at their

own gap define a target-equivalent sensitivity in this realized data. This is a statistical estimand on the released data, not the raw data-generating coefficient -0.65.

The naive all-cell model fails because it targets a mixture estimand that blends activated and resting states and leaves ambient contamination inside the outcome. A raw-marker model fails only when the raw score is judged against the corrected-score cutoff, because that scale mismatch conditions on contaminated resting cells with the wrong genotype–expression relationship.

9 Estimator

Stage 1A: ambient profile. Estimate per-UMI ambient fractions from empty droplets:

$$\widehat{\pi}_g^{\text{amb}} = \frac{\sum_m Y_{mg}^{\text{empty}}}{\sum_m T_m^{\text{empty}}}. \tag{18}$$

Stage 1B: per-cell contamination. Use `HBB` as an ambient-dominated marker:

$$\widehat{\rho}_{ij} = \min\left\{0.5,\ \max\left(0,\ \frac{Y_{ij,\texttt{HBB}}}{T_{ij}\,\widehat{\pi}_{\texttt{HBB}}^{\text{amb}}}\right)\right\}. \tag{19}$$

Subtract the expected ambient contribution from all non-`HBB` genes:

$$\widehat{Y}_{ijg}^{\text{native}} = \max\left\{Y_{ijg} - T_{ij}\,\widehat{\rho}_{ij}\,\widehat{\pi}_g^{\text{amb}},\ 0\right\}, \qquad g \in \{\texttt{IFI6}, \texttt{ISG15}, \texttt{LST1}, \texttt{CXCL10}\}. \tag{20}$$

Stage 2A: state recovery from marker scores. Define the reference corrected-marker score

$$\widehat{A}_{ij} = \log(1 + \widehat{Y}_{ij,\texttt{IFI6}}^{\text{native}}) + \log(1 + \widehat{Y}_{ij,\texttt{ISG15}}^{\text{native}}) - \log(1 + \widehat{Y}_{ij,\texttt{LST1}}^{\text{native}}), \tag{21}$$

and call cell ij activated if $\widehat{A}_{ij} > 1.2$. A raw-marker score using the same three genes is an accepted sensitivity when it is thresholded on the raw scale, for example at the visible gap near $A_{ij}^{\text{raw}} > 2.5$; what fails is reusing the corrected-score cutoff on raw scores.

Stage 2B: observed technical-batch recovery. Aggregate contamination estimates by donor:

$$\bar{\rho}_i = \frac{1}{n_i}\sum_j \widehat{\rho}_{ij}. \tag{22}$$

The released data have a non-overlapping two-cluster structure in $\bar{\rho}_i$, with low-contamination donors in $[0.082, 0.132]$ and high-contamination donors in $[0.243, 0.347]$. The batch covariate used in the estimator is therefore

$$\widehat{B}_i = I(\bar{\rho}_i > 0.18), \tag{23}$$

where any threshold in the empty interval between the two groups yields the same donor assignments. This makes the nuisance adjustment recoverable from visible data: the analyst does not need a hidden batch label, because $\widehat{B}_i$ is a deterministic summary of the released cells and empty droplets.

Stage 3: donor-level Poisson GLM. Aggregate over activated cells:

$$\widehat{Y}_i = \sum_{j:\widehat{A}_{ij}>1.2} \widehat{Y}_{ij,\texttt{CXCL10}}^{\text{native}}, \qquad \widehat{T}_i = \sum_{j:\widehat{A}_{ij}>1.2} T_{ij}. \tag{24}$$

Fit

$$\widehat{Y}_i \sim \text{Poisson}\Big(\widehat{T}_i \exp(\alpha + \beta\, G_i + \gamma\, \widehat{B}_i)\Big)\,, \tag{25}$$

and report $\hat{\beta}$.

The estimator is intentionally built from transparent, visible-data ingredients rather than from a single packaged single-cell workflow: empty-droplet ambient-profile estimation, `HBB`-derived per-cell contamination estimation, marker-panel state recovery, and donor-level Poisson count regression. Because ambient subtraction can produce non-integer corrected counts, the Poisson GLM should be read as a log-link estimating equation on ambient-corrected native signal; rounding the donor-level corrected counts is a deterministic reporting choice and is target-equivalent to fitting the same equation on fractional corrected counts in this instance. The literature anchors the biological and statistical ingredients, while the released diagnostics and ablations show that the answer changes when target-gene ambient correction, state recovery, donor aggregation, or technical-batch adjustment is omitted.

10 Decision-Point and Ablation Walkthrough

The table below combines the full ablation set with the stage at which each shortcut fails. The public target and tolerance are listed in the answer-field table above; rows labeled "yes" recover the same realized-data target.

Decision point	Analysis / ablation	Quantitative output	Pass?	Failure point	Why the approach is wrong
Reference pipeline	Correct donor pseudobulk	$\hat{\beta} = -0.599956$, error 0.000000, 0.0× tol	yes	none	Reference donor pseudobulk after corrected-marker state recovery, corrected target expression, offset, and recovered batch adjustment.
Target-equivalent check	Correct cell-level Poisson	$\hat{\beta} = -0.599695$, error 0.000261, 0.0× tol	yes	none	Acceptable point-estimate sensitivity using the same corrected state, corrected target, offset, and batch.
Target-equivalent check	Reference + sex/age/BMI	$\hat{\beta} = -0.558408$, error 0.041548, 0.8× tol	yes	none	Ordinary donor covariates do not replace the recovered technical group, but adding them to the reference model remains within tolerance.
Target-equivalent check	Corrected/native exposure offset	$\hat{\beta} = -0.611306$, error 0.011350, 0.2× tol	yes	none	Uses an approximate native-UMI exposure after ambient subtraction rather than total activated-cell UMIs; the target conclusion is unchanged.
Target-equivalent check	Fractional corrected counts	$\hat{\beta} = -0.600178$, error 0.000222, 0.0× tol	yes	none	Treats the Poisson GLM as a log-link estimating equation on fractional corrected counts instead of rounded donor-level counts.
Target-equivalent check	Raw-marker state call at raw-scale gap	$\hat{\beta} = -0.590095$, error 0.009861, 0.2× tol	yes	none	Uses the same marker panel on raw counts with threshold $A^{\text{raw}} > 2.5$; because `CXCL10` is still corrected and state recovery is preserved, the target conclusion is unchanged.
Scale mismatch	Raw markers with corrected cutoff	$\hat{\beta} = -0.476520$, error 0.123436, 2.5× tol	no	Stage 2A state recovery	Reuses the corrected-score cutoff $A > 1.2$ on the raw-marker score scale, overcalling contaminated resting cells as activated.
Ambient target	All cells, raw target	$\hat{\beta} = -0.161412$, error 0.438543, 8.8× tol	no	Stage 1 ambient correction / Stage 2A state recovery	Models raw CXCL10 across all cells, so ambient background and state mixture dominate the estimand.
Ambient target/state	Activated cells, raw target	$\hat{\beta} = -0.373918$, error 0.226038, 4.5× tol	no	Stage 1 ambient correction	Restricts to activated cells but leaves ambient CXCL10 in the outcome.
State restriction	All cells, corrected target	$\hat{\beta} = -0.221104$, error 0.378852, 7.6× tol	no	Stage 2A state recovery	Corrects counts but estimates a mixture of activated and resting states.

Decision point	Analysis / ablation	Quantitative output	Pass?	Failure point	Why the approach is wrong
Marker panel	IFI6-only state call	$\hat{\beta} = -0.420124$, error 0.179832, 3.6× tol	no	Stage 2A state recovery	Uses one noisy marker and misses the antagonist LST1 component of the state score.
Selection bias	Selection on target gene	$\hat{\beta} = -0.295804$, error 0.304152, 6.1× tol	no	Stage 2A state recovery	Selects cells on contaminated CXCL10, conditioning on the outcome.
Exposure offset	No exposure offset	$\hat{\beta} = -0.089944$, error 0.510011, 10.2× tol	no	Stage 3 donor GLM/offset	Omits the library-size exposure offset in the count model.
Compound omission	Raw, no recovered group, no offset	$\hat{\beta} = 0.497241$, error 1.097197, 21.9× tol	no	Stage 1 ambient correction / Stage 2B recovered technical group / Stage 3 donor GLM/offset	Leaves raw counts, the contamination-derived technical group, and exposure offset unresolved.
Compound omission	Corrected, no recovered group or offset	$\hat{\beta} = 0.167083$, error 0.767039, 15.3× tol	no	Stage 2B recovered technical group / Stage 3 donor GLM/offset	Corrects counts but omits both the contamination-derived technical group and exposure offset.
Compound omission	All-cell raw, no adjustment	$\hat{\beta} = 0.265017$, error 0.864973, 17.3× tol	no	Stage 1 ambient correction / Stage 2A state recovery / Stage 2B recovered technical group / Stage 3 donor GLM/offset	Combines raw target counts, all-cell mixture, no batch, and no offset.
Ambient target/state	Raw-scale state and raw target	$\hat{\beta} = -0.374966$, error 0.224990, 4.5× tol	no	Stage 1 ambient correction	Uses a plausible raw-marker state threshold but leaves ambient `CXCL10` in the outcome.
Recovered technical group	No recovered group	$\hat{\beta} = -0.366414$, error 0.233542, 4.7× tol	no	Stage 2B recovered technical group	Omits the contamination-derived technical-group proxy $\widehat{B}_i$.
HBB shortcut	HBB filter ≤ 2	$\hat{\beta} = -0.221855$, error 0.378101, 7.6× tol	no	Stage 1 ambient correction	Filters on HBB but does not perform ambient decontamination or the activated-state model.
HBB shortcut	HBB filter ≤ 3	$\hat{\beta} = -0.191040$, error 0.408916, 8.2× tol	no	Stage 1 ambient correction	Uses a looser HBB filter as a substitute for decontamination.
HBB shortcut	HBB filter ≤ 5	$\hat{\beta} = -0.041434$, error 0.558521, 11.2× tol	no	Stage 1 ambient correction	Lets more ambient-contaminated cells into a raw-expression OLS shortcut.

Table 2: Unified decision-point and ablation walkthrough for the ambient-state single-cell eQTL problem.

11 References

1. Nathan A, Asgari S, Ishigaki K, et al. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature*. 2022;606:120–128. DOI: 10.1038/s41586-022-04713-1.

2. Yazar S, Alquicira-Hernandez J, Wing K, et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science*. 2022;376(6589):eabf3041. DOI: 10.1126/science.abf3041.

3. Kang JB, Raveane A, Nathan A, Soranzo N, Raychaudhuri S. Methods and insights from single-cell expression quantitative trait loci. *Annual Review of Genomics and Human Genetics*. 2023;24:277–303. DOI: 10.1146/annurev-genom-101422-100437.

4. Caglayan E, Liu Y, Konopka G. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron*. 2022;110:4043–4056.e5. DOI: 10.1016/j.neuron.2022.09.010.

5. Floriddia E. The impact of ambient RNA. *Nature Neuroscience*. 2022;25:1583. DOI: 10.1038/s41593-022-01232-0.

6. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*. 2020;9:giaa151. DOI: 10.1093/gigascience/giaa151.

7. Arora JK, James LK, Charoensawan V. Understanding and mitigating the impact of ambient mRNA contamination in single-cell RNA-sequencing analysis. *PLOS ONE*. 2025;20(9):e0332440. DOI: 10.1371/journal.pone.0332440.

8. Zimmerman KD, Espeland MA, Langefeld CD. A practical solution to pseudoreplication bias in single-cell studies. *Nature Communications*. 2021;12:738. DOI: 10.1038/s41467-021-21038-1.