Locat: Joint enrichment and depletion testing identifies localized marker genes in single-cell transcriptomics

Wesley Lewis[1,4], Yariv Aizenbud[5], Francesco Strino[3], Yuval Kluger[1,2,4], Fabio Parisi[3*]

[1]Interdepartmental Program in Computational Biology and Biomedical Informatics, Yale University, 433 Temple Street New Haven, New Haven, 06511, CT, USA.
[2]Department of Applied Math, Yale University, New Haven, 06511, CT, USA.
[3]PCMGF Limited, 51 Clarendon Road, Watford, WD17 1HP, UK, https://pcmgf.com/.
[4]Department of Pathology, Yale School of Medicine, 333 Cedar Street New Haven, New Haven, 06511, CT, USA.
[5]School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel.

*Corresponding author(s). E-mail(s): fabio.parisi@pcmgf.com;
Contributing authors: wes.lewis@yale.edu; aizeny@gmail.com; francesco.strino@pcmgf.com; yuval.kluger@yale.edu;

Abstract

Several methods have been developed to identify marker genes that delineate cell populations in single-cell transcriptomic data, yet most emphasize enrichment within candidate populations without testing whether expression is significantly reduced outside those populations. We present Locat, a framework for identifying highly specific localized genes by testing whether expression is concentrated within compact regions of the cellular embedding and depleted elsewhere. For each gene, Locat fits weighted Gaussian mixture models to gene-specific and background densities, computes test statistics for concentration within compact regions and depletion outside those regions, and integrates the results into a unified localization score.

Across synthetic benchmarks with controlled ground truth, Locat detects localized genes spanning uni-modal, multi-modal, and sparse expression patterns, and

appropriately loses significance when simulated expression becomes indistinguishable from background structure. In biological datasets spanning developmental, perturbation, and differentiation contexts, Locat identifies compact marker sets that capture lineage organization, condition-specific programs, and temporal regulatory dynamics. Localized gene sets are often smaller than conventional feature selections such as highly variable genes, and embeddings constructed from localized gene sets tend to preserve separation of major cell populations and developmental programs. In murine dermis, embeddings computed using localized genes preserve differentiation and cell–cycle trajectories observed in the full dataset. In interferon–$\boldsymbol{\beta}$–treated PBMCs, independent localization analysis of control and stimulated samples reveals stimulus-responsive programs and markers of shared immune populations without requiring batch correction or data integration. In retinoic acid–induced embryonic stem cell differentiation, localized genes exhibit reproducible stage-specific patterns across time points.
Together, these results demonstrate that jointly assessing concentration and depletion yields specific, interpretable marker genes that enable direct cross-condition and multi-sample comparisons of marker genes across diverse biological settings.

1 Introduction

A central goal in single-cell transcriptomics is to identify marker genes whose expression serves as a reliable indicator of biologically meaningful cell populations and states. These markers serve multiple purposes. They can distinguish discrete cell types such as T cells and B cells, trace continuous developmental trajectories such as hematopoietic differentiation, reveal functional heterogeneity within apparently uniform populations such as activated and quiescent states, and define specialized subsets within major lineages, such as naive, regulatory, and memory CD4+ T cells. In many cases, these subsets occupy contiguous regions in transcriptional space despite sharing broader lineage identity.

In this context, localized marker genes exhibit two defining properties: expression that is enriched within a compact subset of cells, together with marked depletion elsewhere. This depletion, particularly in substantial regions outside the target population, is essential for biological specificity and reliable cell type identification. Consistent with this view, a recent large-scale benchmarking study evaluating marker genes across diverse biological contexts shows that marker utility depends critically on expression specificity relative to the full cellular background[1]. In particular, genes with expression confined to restricted regions or cell states, and little to no expression elsewhere, are favored over broadly or diffusely expressed genes that exhibit local enrichment[1].

In practice, most marker discovery methods prioritize genes based on enrichment criteria, favoring expression patterns that are concentrated, smooth, or compact on a cell–cell graph, or enriched within clusters of similar cells. These approaches typically assess enrichment without explicitly testing for the absence of expression outside candidate regions. As a result, diffuse expression patterns cannot be readily distinguished from truly localized ones. Consequently, genes with high baseline expression

may rank highly under enrichment-based scoring schemes despite lacking a well-defined localized domain, conflating statistical enrichment with biological specificity. These limitations motivate the need for approaches that jointly assess enrichment within candidate regions and exclusion outside them, enabling the identification of genes that demarcate cellular populations rather than merely correlate with them.

We present Locat, a statistical framework that formalizes this dual requirement by testing for both concentrated enrichment (which we term *concentration*) and significant exclusion from the remaining cellular background (which we term *depletion*). Rather than relying on predefined cell groupings or discrete partitions, Locat models gene expression patterns as continuous probability densities within user-supplied embeddings, such as PCA or Uniform Manifold Approximation and Projection (UMAP) representations. For each gene g, Locat fits a gene-specific weighted Gaussian mixture model (WGMM) to estimate the density of cells expressing that gene and contrasts this with a background WGMM describing the overall distribution of all cells in the embedding. Concentration and depletion are then assessed by comparing these two density models: the concentration test quantifies the extent to which the gene-specific density aggregates into compact regions of the embedding, while the depletion test evaluates whether expression is significantly under-represented across a substantial amount of the background density outside these concentrated regions. The concentration and depletion p-values are first combined using Cauchy aggregation. This Cauchy-combined value is then adjusted to yield the final localization score used for thresholding and cross-dataset comparisons. By construction, the adjusted score is conservative relative to the Cauchy-combined value (see Methods). Because gene-level tests share a common embedding and background model and differ substantially in prevalence, we assess calibration empirically using simulation-based null analyses across the range of prevalences encountered in the data. We therefore do not rely on a single genome-wide correction as the primary calibration device (Supplementary Methods S1.3; Supplementary Fig. S17).

Conceptually, Locat belongs to a broad class of marker discovery methods that seek genes whose expression is concentrated within restricted regions of transcriptional space, whether those regions are defined by clusters, cell–cell graphs, or continuous manifolds. Several existing tools formalize this idea using different statistical criteria. singleCellHaystack identifies genes with non-uniform expression distributions by comparing observed and expected densities using Kullback–Leibler divergence [2]. Hotspot identifies genes whose expression exhibits significant local autocorrelation on a cell–cell similarity graph, highlighting genes whose values are predictable from their transcriptional neighborhood rather than randomly distributed [3]. GiniClust uses inequality metrics to prioritize genes enriched in rare populations [4]. Graph-diffusion–based methods such as GSPA and LMD operate directly on cell–cell manifolds to capture smooth, coherent, or compact expression patterns [5, 6]. While these approaches differ in how regions are represented and scored, they are unified by an emphasis on enrichment or structured variability as the primary criterion for marker selection.

We apply Locat across diverse biological systems, including embryonic dermis differentiation, perturbation experiments, and temporal developmental series. Across

these settings, the dual assessment of enrichment and depletion identifies highly specific markers that capture both discrete cell types and continuous biological processes, reveal condition-specific localization patterns, and track the temporal emergence of localized expression during development. In embryonic differentiation and perturbation data, embeddings constructed using only top localized genes recover the major cell types and continuous trajectories observed in highly-variable gene embeddings, while relying on a smaller and more interpretable set of markers. In comparisons with existing clustering-free marker selection methods, Locat identifies gene sets that are notably distinct from those prioritized by enrichment-only approaches, with discrepancies driven by broadly expressed background genes filtered by Locat and by selective markers not emphasized by enrichment-based criteria. In time-series data, Locat detects genes whose localization emerges dynamically over time, and localization patterns are consistent across replicates at the same time point, supporting comparisons of localized programs across samples and conditions.

Together, these results show that Locat provides a general and reliable framework for discovering highly specific localized markers. By directly testing both enrichment within a region and depletion outside it, Locat produces interpretable, high-specificity gene sets that support meaningful comparisons across samples, conditions, and datasets.

2 Results

2.1 Localization requires selective absence

In marker gene discovery, scRNA-seq data is typically represented in a chosen representation of cells, such as tissue coordinates, a low-dimensional latent space (PCA, tSNE or UMAP), or a cell–cell neighborhood graph. In any of these representations, a gene can be locally enriched in this space without being a reliable marker. The key point of failure for existing tools is background expression: a gene may peak in one region yet remain detectably expressed across all other regions. Such a gene correlates with structure but does not identify or demarcate a particular cell population.

This motivates a stricter operational definition of localization. A localized marker should exhibit two complementary properties: (i) concentrated expression in compact regions of high density (concentration), and (ii) selective absence from substantial portions of the remaining cell population, particularly where many cells reside (depletion). The schematic in Fig. 1 illustrates why these properties are not equivalent: "enriched" describes a local maximum, whereas "enriched and depleted" describes a maximum paired with meaningful contrast against the global background density. Only the latter provides the necessary specificity required for population demarcation.

We illustrate this distinction using murine embryonic dermis data from [7] and four representative genes in Fig. 1. These examples are chosen solely to demonstrate the conceptual cases below, whereas a more exhaustive evaluation of the dataset is given in Sections 2.3, 2.5, and 2.6.

- Concentrated only: *H19*, *Nfib*. Expression is clearly concentrated into a compact region, but appreciable expression remains visible across other populated

regions of the embedding. These genes look localized under enrichment-first scoring because they have a strong peak, yet they do not cleanly define a population boundary.

- Concentrated and depleted: *Sox2*, *Hist1h2bb*. Expression is concentrated into a compact domain and is also markedly absent elsewhere, including in regions where many cells reside. Compared to *H19* and *Nfib*, these genes are better markers in the practical sense: presence is informative and absence is informative.

Depletion is thus not an incremental modification of enrichment scoring, but a necessary criterion for marker specificity. Having established the conceptual distinction between concentration and depletion, we formalize both as hypothesis tests for each individual gene in a fixed low-dimensional cell embedding, where proximity reflects similarity in transcriptional state. The concentration test evaluates whether a gene's expression mass is more aggregated within this embedding than expected under a null model in which expressing cells are randomly distributed according to the background transcriptional activity.

Concentration

We quantify concentration by contrasting two density functions over the embedding. The first, f_g, represents the density of gene g's expression, and the second, f_0, represents the background cell density. We define the concentration score as:

$$\tilde{z}_g = \sum_{c \in C} w_{g,c} \frac{f_g(\mathbf{x}_c) - f_0(\mathbf{x}_c)}{f_g(\mathbf{x}_c) + f_0(\mathbf{x}_c)} \tag{1}$$

where C denotes the set of all cells and $w_{g,c} = e_{g,c} / \sum_{c'} e_{g,c'}$ denotes the normalized expression weight of gene g in cell c (Section 4). The symmetric difference form $(f_g - f_0)/(f_g + f_0)$ is bounded in $[-1, 1]$ and measures the relative deviation of gene density from background at each cell location. The score aggregates these deviations across all expressing cells: positive values indicate that expression concentrates in regions of elevated density relative to the null expectation.

We estimate $f_g(\mathbf{x})$ from cells expressing gene g, weighted by that gene's expression, and we estimate $f_0(\mathbf{x})$ from all cells in the embedding (see Section 4.1.2). The full operational definition of the concentration statistic is given in Section 4.1.3, including prevalence weighting and normalization. Significance is then obtained by calibrating this statistic against prevalence-matched randomized expression patterns, yielding a concentration p-value that controls for expression prevalence.

Depletion

We quantify depletion by testing whether gene expression is under-represented across a substantial portion of the embedding density. For each gene g and contrast threshold $\lambda > 1$, we define a region of depletion

$$\Gamma_\lambda = \{\mathbf{x} : f_0(\mathbf{x}) > \lambda f_g(\mathbf{x})\}. \tag{2}$$

Within this region of depletion, we compute the observed expressing-cell fraction and the background density fraction

$$\widehat{\beta}_g(\lambda) = \frac{\sum_{c\in\Gamma_\lambda} \mathbf{1}\{w_{g,c} > 0\}}{\sum_{c\in C} \mathbf{1}\{w_{g,c} > 0\}} \quad \text{and} \quad \widehat{\beta}_0(\lambda) = \frac{\sum_{c\in\Gamma_\lambda} f_0(\mathbf{x}_c)}{\sum_{c\in C} f_0(\mathbf{x}_c)}. \tag{3}$$

Intuitively, $\widehat{\beta}_g(\lambda)$ measures the fraction of gene g's expressing cells that lie in regions where background density substantially exceeds gene density, while $\widehat{\beta}_0(\lambda)$ measures the fraction of overall cellular density in those same regions.

Depletion is assessed by testing whether

$$\widehat{\beta}_g(\lambda) \ll \widehat{\beta}_0(\lambda), \tag{4}$$

i.e., whether the gene avoids regions where many cells reside. The resulting depletion statistic corresponds to the tail probability under a null model in which gene expression follows the background density (see Section 4.1.4).

In practice, we evaluate depletion for different values of $\lambda \in \Lambda$ (see Section 4.1.4) to capture depletion at different stringency levels, apply effective sample size corrections to account for autocorrelation of neighboring cells, and adjust for multiple testing across thresholds. The final depletion p-value represents the minimum across thresholds after correction.

Joint Testing: Localization

Locat evaluates gene localization by jointly testing for concentration and depletion (see Section 4.1). A gene is considered localized only if it shows significant evidence for both aggregation/compactness and exclusion from large portions of transcriptional space (see Section 4.1.5), aligning the statistical criterion with the biological goal of acting as a compact population indicator.

To motivate clear examples of localized genes, we again consider four genes from murine embryonic dermal condensates dataset [8]. Under our criterion, *Sox2* and *Hist1h2bb* are prioritized because they satisfy both requirements, whereas *H19* and *Nfib* fail the selective-absence criterion despite exhibiting strong local enrichment. All four genes show highly significant concentration ($p < 10^{-6}$).

For *Sox2* and *Hist1h2bb*, depletion is strongly supported (depletion $p < 1 \times 10^{-12}$), resulting in significant localization p-values ($p = 9.89 \times 10^{-3}$ and 2.48×10^{-2}, respectively). In contrast, *H19* and *Nfib* show weak or non-significant depletion (depletion $p = 0.140$ and 0.103), and consequently do not achieve significant localization (localization $p = 0.089$ and 0.082; Fig. 1).

We provide a secondary set of examples using a widely used 10X PBMC dataset [9] to reinforce these concepts in Supplementary Fig. S1.

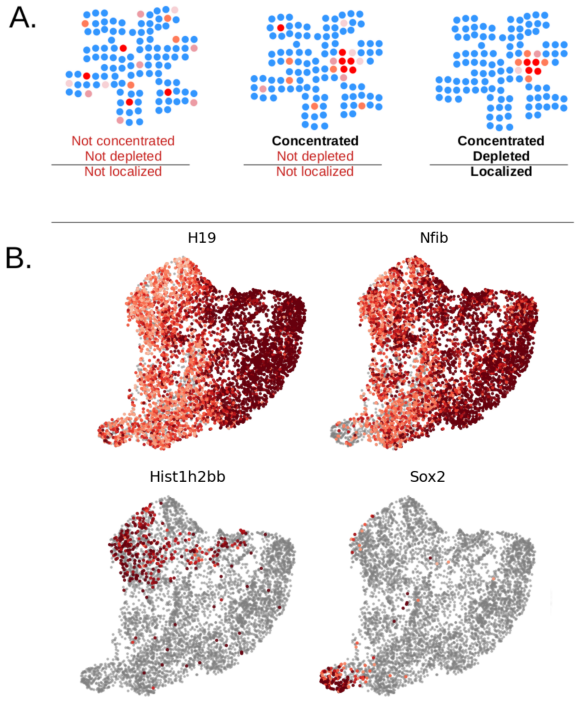

Fig. 1 Illustrative examples of localized and not localized gene expression profiles. **A.** Illustrative representations of the expression of three genes on the same manifold. Each dot represents a cell. Blue cells do not express the gene, red cells express the gene, and the intensity of expression is proportional to the intensity of the color. On the left the expression is not localized in a specific region and it is thus not depleted outside such region. In the middle, the expression is concentrated in a region, but is not depleted outside of it. On the right the expression is concentrated in a region and depleted outside of it, thus the gene expression is said to be localized. **B.** Expression profiles on two-dimensional UMAP of four illustrative genes selected to demonstrate the occurrence of concentration and localization in real data. UMAP embeddings with expressed cells labeled illustrate two nonspecific genes with concentrated enrichment, *H19* and *Nfib*, as well as two concentrated and highly specific local genes, *Sox2* and *Hist1h2bb*.

2.2 Locat shows high sensitivity and robustness in synthetic data experiments

To characterize how Locat behaves under controlled conditions, we generated synthetic datasets with known ground truth and evaluated its response to changes in sample size, sparsity, and localization radius. Each simulation consisted of a unit-variance multivariate Gaussian background distribution of 1,500 cells, with expression for a simulated localized gene drawn from a compact neighborhood (red points, Fig. 2A, C, and E). We show three statistical approaches for testing simulated genes: Locat's localization p-value (see Section 4.1.5), a likelihood ratio test comparing gene-weighted versus background density models (see Section 4.6), and a permutation-based variant of Locat (see Section 4.7). The likelihood ratio test serves as a natural parametric baseline that directly compares distributional fits but lacks explicit constraints

for compactness or specificity. The permutation-based approach represents a gold-standard resampling method that preserves gene-specific expression patterns using randomized shuffling of expression weights and simulated bootstrapping, but it incurs prohibitive computational cost. For each experimental setup, 50 replicate trials were performed.

Sensitivity to sample size.

We varied the number of expressing cells between 5 and 250 while maintaining localization around a fixed center (Fig. 2A). Locat's localization p-value remained consistently below the significance threshold $\alpha = 0.05$ for much of the data, indicating high sensitivity even for genes with small sample sizes (Fig. 2B, left panel). In contrast, the likelihood ratio test performed poorly, with sensitivity decreasing as the number of expressing cells increased (Fig. 2B, middle panel). This reflects increasing overlap between the gene and background densities as expression becomes more widespread. The p-values from the permutation-based approach indicated that the approach yielded similar sensitivity to Locat's localization test but exhibited greater variability in p-values at larger sample sizes due to the finite number of resampling iterations (Fig. 2B, right panel). These results show that Locat maintains high sensitivity across a large range of gene-expressing population sizes, spanning simulated genes expressed in 1.6% to 16% of the total number of cells in our simulation.

Robustness to gene sparsity.

In our second simulation scenario we introduced a multimodal localized population with increasing positional jitter with respect to each mode, to represent different levels of compactness in a local support region. Locat is intended to identify extremely compact genes as well as genes that are moderately compact but very specific to a particular population of cells. This experiment evaluated simulated genes with progressively more spread out expression patterns, representing increasing gene sparsity in the expressed population. The simulated genes are also multimodal, which is an additional source of deviance from the background distribution (Fig. 2C). The likelihood ratio test lost power rapidly, with p-values exceeding the significance threshold already at a jitter of 5% (Fig. 2D, middle panel). In contrast, Locat's localization p-value remained significant across most of the tested jitter range and only lost significance once the jitter was sufficiently large that the expression pattern became indistinguishable from the background distribution (Fig. 2D, left panel). The permutation-based test maintained nominal significance across all jitter levels, including at extreme dispersions, suggesting overly permissive behavior when gene expression structure is nearly indistinguishable from the background signature. (Fig. 2D, right panel).

Sensitivity to localization radius.

In our third simulation scenario we varied the dispersion radius of the localized expression region while holding the background variance constant (Fig. 2E). As the dispersion radius increases, the signal gradually merges with the background distribution, thus losing its localization property. Modeling concentration as a naïve single-sided z-test we expected no identifiable localization for a dispersion radius greater than 1.6 at a

significance threshold of $\alpha = 0.05$, i.e. the observed distribution of the gene expression becomes undistinguishable from the distribution of the background. Locat's localization p-value crossed the significance threshold $\alpha = 0.05$ at dispersion radius around 1.6. The likelihood-ratio test lost significance by radius ≈ 0.5 (Fig. 2F, left and middle panels). The permutation-based approach remained significant across nearly all radii, confirming its anti-conservative tendency and indicating a greater tendency toward false positives compared to Locat (Fig. 2F, right panel). These findings indicate that Locat yields higher sensitivity than the likelihood-ratio test approach and higher specificity than the permutation-based approach.

These simulations demonstrate that Locat maintains power across rare populations ($\geq 1.6\%$ prevalence), handles multimodal distributions and positional noise, and correctly fails to detect localization when the signal concentration equals the background. The latter property is critical for non-uniform embeddings e.g., single-cell RNA data, where likelihood ratio and permutation tests show systematic biases.

2.3 Comparison of marker gene selection approaches

To evaluate Locat's performance relative to existing approaches, we used a murine embryonic dermis dataset of approximately 5,000 cells spanning the transition from upper dermal fibroblasts to *Sox2*-positive dermal condensates, early precursors of hair follicles [7]. This dataset presents a challenging test case: it contains two intersecting sources of variation. There is continual progression of fibroblasts toward dermal condensate identity, and these fibroblasts likewise move through phases of the cell cycle. Together, these form a continuous transcriptional manifold without discrete cluster boundaries (Supplementary Fig. S2A). Traditional cluster-based differential expression methods cannot cleanly separate markers for these contiguous processes, motivating the need for clustering-free gene prioritization approaches.

We compared Locat gene rankings to six marker gene selection methods representing diverse methodological frameworks: LMD (diffusion statistic over a cell-cell graph, [5]), Scanpy rank_genes_groups (Wilcoxon rank-sum test on user-defined groups, [10]), Haystack (2D kernel density comparison, [2]), Hotspot (graph autocorrelation based on Geary's C, [3]), GSPA (gene-specific pattern analysis, [6]), and GiniClust (Gini index-based rare cell detection, [4]). We also included a spectral baseline (SpectralRH) that scores genes by graph smoothness and concentration of spectral energy in low-frequency Laplacian eigenmodes on a k-nearest-neighbor graph and serves as a comparator with a different objective from the other methods (see Supplementary Methods S1.7.2). Several of these methods are designed explicitly to capture local clustering patterns, whether through autocorrelation (Hotspot) or diffusion-based statistics (LMD). Whereas the other approaches prioritize compactness, local autocorrelation, or slow diffusion across neighboring cells, SpectralRH is intended to highlight genes whose expression aligns with well-supported transitions and boundaries in the embedding (see Supplementary Methods S1.7.1, S1.7.2).

This collection of methods spans concentration-based approaches (LMD, Haystack, Hotspot), enrichment-only differential expression (Scanpy), rare cell detection (GiniClust), pattern-based scoring (GSPA), geometric structure alignment (SpectralRH) and Locat's novel dual concentration-depletion framework. For each method, we

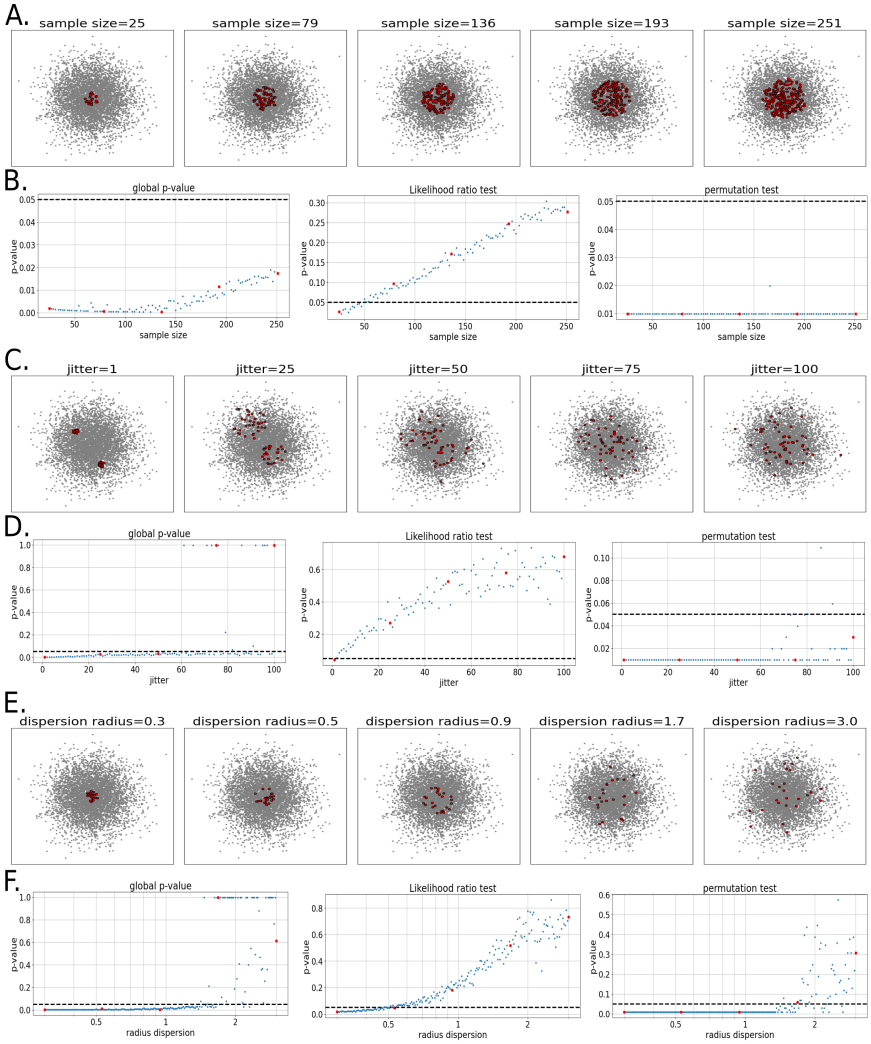

Fig. 2 Validation of Locat on simulated data with controlled ground truth. (A, C, E) UMAP embeddings of 1,500 cells sampled from a Gaussian background distribution. Colored points show five replicates of a simulated localized gene under varying conditions: (A) Simulated genes with sample size ranging from 25 to 251 expressing cells; (C) Simulated genes with positional jitter from 1 to 100 embedding units; (E) Simulated genes with dispersion radius from 0.3 to 3.0 embedding units. (B, D, F) scatter plots of p-values from three statistical tests across 50 replicates per condition: global Locat (left), likelihood ratio test (middle), and permutation-based Locat (right). Dashed lines indicate the significance level $\alpha = 0.05$. (B) Global Locat maintains $p \approx 0.025$ across all sample sizes (blue points), while likelihood ratio test p-values degrade from ~ 0.05 to > 0.20 as sample size increases (red trend). (D) Locat remains significant ($p < 0.05$) through moderate jitter levels (≤ 67), whereas likelihood ratio test loses power when jitter is introduced. (F) Locat detects localization up to dispersion radius ~ 2.0; likelihood ratio test sensitivity drops by radius ~ 1.0. Permutation-based approach shows significant p-values at extreme parameter values (panels D, F, right).

ranked all genes and identified high-discrepancy genes, i.e. those ranked highly by one method but poorly by Locat or vice versa.

Genes prioritized by Locat, including *Histh2ab* (Locat: 98th percentile; LMD: 94th percentile, scHaystack 73rd percentile), *Casq1* (Locat: 99th percentile; GSPA: 87th percentile; LMD: 95th percentile, Hotspot: 64th percentile), and *Qprt* (Locat: 99th percentile; GSPA: 91st percentile; LMD: 96th percentile; Hotspot: 74th percentile), exhibited tightly confined expression with minimal background (Supplementary Fig. S3). In contrast, genes ranked more highly by enrichment-focused methods, including *Col8a1* (LMD: 93rd percentile; GSPA: 99th percentile, Locat: 76th percentile) and *Ptn* (LMD: 99th percentile, SpectralRH: 98th percentile; Haystack: 99th percentile; Locat: 5th percentile) showed moderate local enrichment accompanied by substantial background expression (Supplementary Fig. S3). Such diffuse patterns achieve high scores under concentration-only criteria and lack well-defined localized domains, otherwise enforced by the depletion score included in Locat. We provide additional visualizations of genes with high rank discrepancies between Locat and existing models in Supplementary Fig. S4.

Rank correlation analysis across all genes revealed low to moderate concordance between Locat and alternative approaches (Fig. 3A). Locat shows highest concordance with GSPA and LMD (Spearman $\rho = 0.66$ and $\rho = 0.63$, respectively) and moderate agreement with SpectralRH ($\rho = 0.38$), reflecting a shared emphasis on detecting genes with highly concentrated expression, though Locat's dual concentration-depletion framework further penalized genes with broad and nonspecific background expression, as evidenced by Supplementary Fig. S4. We additionally computed Jaccard indices between the top 500 genes identified by each method, and evaluated rank correlations between methods restricted to the top 500 genes identified by Locat (Supplementary Fig. S5). In both analyses, concordance remained modest (Supplementary Fig. S5). This indicates that the depletion test introduces a distinct selectivity criterion that separates truly localized markers from broadly expressed genes with local peaks. The rank- and overlap-based comparisons in this section establish that Locat prioritizes a distinct class of genes relative to existing approaches. The following sections provide broader context for these differences through applications in dermal condensate differentiation (Section 2.5), interferon-β stimulation (Section 2.7), ESC differentiation time series (Section 2.8), and controlled perturbations of a real-data template gene (Section 2.4).

2.4 Comparing methods using simulated perturbations of a real dataset

To determine whether Locat's distinct gene selections reflect a meaningful advantage rather than merely a different ranking criterion, we performed controlled gene-level simulations to evaluate how different gene prioritization methods respond to simulated genes with variable sample size and background expression. Simulations were based on the murine dermis dataset, using *Sox2* as a template gene due to its compact and well-defined expression pattern in dermal condensates. Synthetic variants were generated by perturbing *Sox2* expression while preserving the underlying cell embedding.

Each simulation retained the original cell embedding and expression values for approximately 1,000 randomly sampled background genes, together with multiple simulated copies of *Sox2*. This experiment was conducted independently of the full-dataset correlation analysis and isolates method behavior under controlled perturbations of a single gene. For methods that do not report p-values, including GSPA and LMD, significance thresholds are defined empirically from the distribution of gene scores (see Section 4.8.1). All methods introduced above were applied identically in this setting.

Robustness to subsampling.

We assessed robustness to decreasing prevalence by progressively subsampling the number of cells expressing the simulated *Sox2* gene while maintaining its expression pattern (Supplementary Fig. S6A). Locat maintained detectability across nearly the full range of subsampling levels (Supplementary Fig. S6B). SpectralRH, Hotspot, and Haystack similarly showed limited sensitivity to subsampling in this regime. By contrast, cluster-based differential expression tests showed a rapid sensitivity drop as expressing cells became rare. GSPA and LMD were significant at some moderate subsampling levels, but once more than half of expressing cells were removed, both showed a monotonic loss of significance with further subsampling.

Sensitivity to background expression.

We next evaluated sensitivity to loss of localization by progressively redistributing *Sox2* expression into the background population while holding the number of expressing cells fixed (Supplementary Fig. S7A). In this setting, Locat p-values became less significant and crossed the detection threshold as expression became broadly distributed and indistinguishable from background (Supplementary Fig. S7B). SpectralRH, GSPA, and LMD similarly exhibited monotonic score increases with increasing background expression, indicating these methods were also sensitive to this attribute. In contrast, Hotspot, Haystack, and cluster-based differential testing remained largely insensitive until background expression signal became extreme, at which point they no longer detected the simulated genes.

These simulations demonstrate that Locat uniquely balances two competing but essential criteria for marker discovery: robustness to decreasing gene prevalence and sensitivity to loss of localization due to background expression. While methods such as LMD, GSPA, and SpectralRH also penalize increasing background expression, those same methods exhibited unstable or rapidly degrading behavior under subsampling, indicating that their apparent sensitivity to background signal is not preserved when expression becomes sparse. Conversely, methods that remained robust to subsampling largely failed to penalize background expression, continuing to assign high confidence to genes whose expression was broadly distributed rather than localized.

This trade-off has direct implications for real data analysis. Some genes are expressed in many cells or show strong local peaks, yet still remain broadly distributed across the embedding. Methods that do not adequately distinguish this pattern may rank such genes highly even when they are not specific markers of a localized cell state. Locat is better able to separate these genes from genes whose expression is concentrated within a restricted region of the embedding, even when those localized genes are

present in relatively few cells. As a result, gene prioritization depends more strongly on spatial specificity than on prevalence or expression magnitude alone.

2.5 Localized markers recover major axes of biological variation

Localized markers along developmental and cell-cycle axes.

In the murine embryonic dermis dataset (see 2.3) Locat identified 430 out of 11,700 genes at $p < 0.05$ (calibration of this threshold under simulation-based null analyses is shown in Supplementary Fig. S17; see Supplementary Methods S1.3) with compact expression patterns aligned with both cell-cycle progression and differentiation-related variation . Several cell-cycle-associated genes including *Hist1h2bb* ($p = 2.34 \times 10^{-2}$), *Ccnb1* ($p = 4.27 \times 10^{-2}$), and *Kif20a* ($p = 4.90 \times 10^{-2}$) formed sharply confined domains corresponding to distinct cell-cycle phases, consistent with established phase-specific programs [11]. Genes associated with dermal condensate identity, including *Sox2* ($p = 1.05 \times 10^{-2}$), *Foxd1* ($p = 1.97 \times 10^{-2}$), and *Gal* ($p = 1.96 \times 10^{-2}$) localized to terminal regions of the fibroblast-to-dermal-condensate trajectory, in agreement with their known expression patterns in developing skin [12, 13]. Several genes linked to ephrin signaling and cytoskeletal dynamics localized to a cycling population adjacent to the terminal dermal condensate (Supplementary Fig. S8). These included *Epha8* ($p = 3.49 \times 10^{-3}$) and *Stmn2* ($p = 4.48 \times 10^{-2}$). This pattern is consistent with a transitional population undergoing cytoskeletal and signaling remodeling prior to terminal condensate differentiation [14, 15].

Manifold reconstruction from localized genes.

To test whether Locat-prioritized genes retain biologically meaningful structure in the dataset, we constructed a localized-gene embedding (LG-embedding): a low-dimensional representation computed using only the 403 significant localized genes discovered by Locat (Supplementary Fig. S9B). We compared this to a baseline UMAP embedding derived from 4000 highly variable genes (HVG-embedding, Supplementary Fig. S2B, left panel). Despite using substantially fewer genes, the LG-embedding recapitulated both the fibroblast-to-dermal-condensate differentiation trajectory and the cell-cycle–associated variation observed in the HVG embedding (Supplementary Fig. S10B, right panel). These results indicate that genes selected for compact localization are sufficient to represent at least two well-characterized axes of biological variation in this system, demonstrating that localized markers, though individually restricted to confined expression domains, can collectively preserve continuous developmental and cell-cycle structure using a substantially smaller and more interpretable feature set than the HVG baseline.

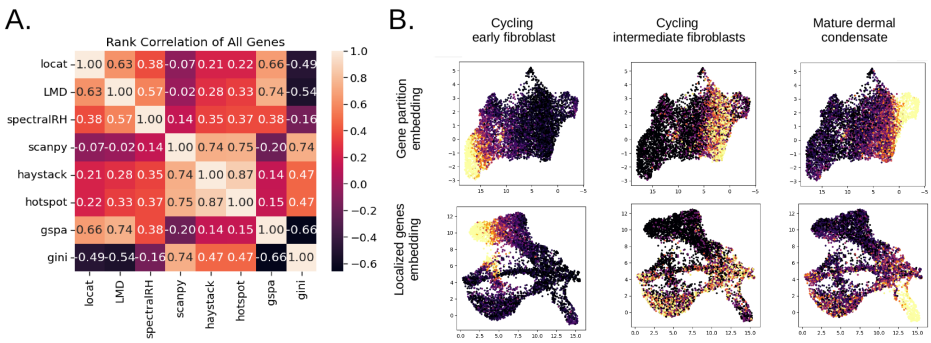

Fig. 3 Locat identifies distinct gene programs and shows moderate concordance with alternative prioritization methods. A. Rank correlation heatmap (Spearman ρ) comparing genome-wide gene rankings across eight local gene selection methods applied to the murine embryonic dermis dataset. Moderate correlations indicate partial agreement in prioritized genes while reflecting method-specific differences. **B.** Average expression of genes in three representative Locat-derived modules. Modules were identified by applying Locat to principal components computed from a subset of genes selected to emphasize differentiation-related variation [8]. The modules correspond to cycling early fibroblasts (left), cycling intermediate fibroblasts (middle), and mature post-mitotic dermal condensates (right). Top row: module-average expression projected onto a UMAP constructed from the differentiation-focused principal components used for discovering localized genes and modules, illustrating alignment with progression along the fibroblast-to-dermal-condensate trajectory. Bottom row: average expression of the same modules projected onto the localized-genes embedding constructed from all significant Locat genes. In this representation, each module occupies a coherent and spatially restricted region, demonstrating that differentiation-associated gene programs remain well-defined when visualized using only localized genes. Average expression of all modules from this analysis is shown in Supplementary Fig. S11.

2.6 Localized markers reveal distinct stages of development in murine embryonic dermis

Module discovery reveals cell-cycle exit and lineage commitment.

A key question in dermal development is identifying the molecular signatures that mark fibroblast commitment to the dermal condensate lineage and the timing of cell-cycle exit during this transition. To address this, we constructed a differentiation-focused embedding from the murine embryonic dermis dataset (see 2.3) by removing cell-cycle variation using gene partitions from [8] (Supplementary Fig. S9). We then applied Locat to identify which genes from the entire dataset were localized on this new embedding. By construction, such localized genes along the developmental trajectory correspond to specific developmental stages.

Locat identified 1302 distinct localized genes along the developmental trajectory, which we grouped in 8 gene modules (see 4.2). These modules exhibited particular expression patterns corresponding to early fibroblast, intermediate, and mature dermal condensate states (Fig. 3B, upper row). When we projected these module expression signatures back onto the LG-embedding (which retains cell-cycle structure, Fig. 3B, lower row), the modules traced a continuous developmental path revealing discrete

transitions from proliferative to post-mitotic states, highlighting a region of the embedding associated with cell-cycle exit and dermal condensate identity (Fig. 3B, right column).

This analysis highlights a key feature of Locat: because it is embedding-agnostic (see Section 4.1.1), it can be applied to any user-defined low-dimensional representation and will identify genes localized with respect to that representation. The LG-embedding derived from all localized genes captured both developmental and cell-cycle variation (Supplementary Fig. S10). When an embedding is constructed to emphasize a single biological axis, such as differentiation, applying Locat to that representation prioritizes gene programs whose expression patterns align with the chosen trajectory. In this setting, the resulting modules reflect genes whose localization is most consistent with the differentiation-focused embedding, reducing contributions from other sources of variation (e.g., cell-cycle state) to the extent that they are not encoded in that representation.

2.7 Single-sample localization analysis avoids integration artifacts in multi-condition experiments

Multi-sample single-cell studies commonly integrate datasets using batch-correction methods that align cells from different conditions into a shared representation. These approaches reduce cross-sample variation to facilitate direct comparison of similar cell states. However, not all cross-sample differences reflect technical noise. In many settings, they represent true biological programs induced by perturbation, genotype, or developmental state. Because batch correction explicitly minimizes between-sample variation, it can attenuate condition-specific expression structure that is often the primary focus of analysis.

We therefore take a different approach. Rather than applying batch-correction–based integration to align cells across conditions into a shared embedding prior to analysis, we apply Locat independently within each sample and evaluate gene localization with respect to that sample's native representation. Localization scores can then be compared directly across samples, or cells from multiple samples can be embedded together using the union of condition-specific localized genes as the feature set. In this setting, a combined embedding is constructed without additional batch correction, using only genes identified as localized. This strategy preserves condition-specific structure during localization analysis while still enabling interpretable cross-condition comparison at the level of gene programs. Such combined embeddings are most appropriate when samples share a common biological architecture but differ due to defined perturbations (e.g., treatment versus control or genotype), so that a shared representation remains meaningful. When samples differ more fundamentally (e.g., across developmental time points, tissues, or technologies), cross-sample embedding may be less informative. In these settings, localization patterns are more appropriately interpreted within each sample.

Integration artifacts obscure stimulus-specific localized programs.

We demonstrate this framework using a single-cell RNA interferon-β stimulation dataset [16], comprising approximately 12,000 peripheral blood mononuclear cells per

condition from a single donor. Control and IFN-β-stimulated samples were processed independently: each was normalized separately, embedded in its own PCA space, and analyzed with Locat to identify localized genes without cross-sample alignment. The two samples were subsequently merged into a combined dataset of approximately 24,000 cells for visualization and comparison. This design avoids imposing batch-correction-based alignment during localization analysis.

Given that the two samples originate from the same donor, an embedding constructed from both control and stimulated cells should align shared immune populations while preserving stimulus-induced changes within responsive cell types. We therefore evaluated whether merging the datasets could achieve both goals, either without batch correction or after applying batch correction. A UMAP embedding computed from highly variable genes on the merged data without batch correction showed near-complete separation of control and stimulated cells (Fig. 4A). This separation likely reflects technical differences between samples more than the biological effects of IFN stimulation and prevents meaningful comparison of shared immune populations. Applying Harmony-based batch correction substantially improved cross-condition alignment (Fig. 4B), but visual inspection and differential abundance analysis (Supplementary Fig. S12) indicate that stimulated monocytes no longer form a distinct region in the embedding. Monocytes are expected to mount a robust interferon response characterized by induction of canonical interferon-stimulated genes and inflammatory mediators, including chemokines and antigen-presentation programs [17], which should generate a transcriptionally distinct stimulated state relative to control. In the Harmony-corrected embedding, this separation is no longer apparent. These results highlight a central challenge: merging samples without batch correction exaggerates technical differences, whereas strong batch correction can obscure biologically meaningful stimulus-specific structure.

As an alternative approach to representing both samples jointly, we generated a UMAP embedding using only genes that were significantly localized in at least one condition (1283 genes from control, 1398 genes from stimulated, 944 genes shared). Under our dual concentration–depletion criterion, these genes serve as discriminative features because their expression is confined to well-defined cell populations rather than broadly distributed across the manifold. The resulting embedding aligned shared immune lineages while preserving a distinct region corresponding to IFN-β-responsive monocytes, without requiring batch correction (Fig. 4).

We then used this co-embedding to visualize genes exhibiting differential localization between conditions. For validation, we also plotted expression on embeddings constructed separately for the control and stimulated samples to confirm that specific genes were localized uniquely within IFN-β-responsive monocytes or within control monocyte populations.

Condition-specific differential localization.

We define differential localization as a gene that is significantly localized in one condition but not significantly localized in the other. We identified such genes by comparing Locat's localization p-values across the two samples using a fixed significance threshold. Using this criterion, 454 genes were significantly localized in the stimulated sample

but not in control, and 339 genes were significantly localized in control but not in stimulated.

Genes with much stronger localization in the stimulated sample included *CCNA1*, *CXCL9*, *NUPR1*, *CH25H*, and *JUP*, which were confined to interferon-responsive monocytes (Supplementary Fig. S13A and Supplementary Fig. S14A). Among these, *CXCL9* and *CH25H* are well-established interferon-stimulated genes involved in chemokine signaling and antiviral lipid metabolism, respectively [18, 19]. *NUPR1* reflects stress- and interferon-associated transcriptional responses in activated myeloid cells [16, 20].

In contrast, several genes were significantly localized in the control condition but not in the stimulated condition, including *TPST1*, *MMP9*, *NAB2*, *OSM*, and *HIPK2* (Supplementary Fig. S13B and Supplementary Fig. S14B). *MMP9*, which has been associated with extracellular matrix remodeling and basal monocyte function [21], marked a transcriptionally restricted subset of unstimulated monocytes. *OSM*, a cytokine with context-dependent roles in myeloid cells [22], was likewise confined to a subset of control monocytes. Regulatory genes such as *NAB2*, a transcriptional regulator [23], and *HIPK2*, a kinase involved in transcriptional control and stress signaling [24], were similarly localized specifically in control samples. The loss of localized expression for these genes following interferon stimulation suggests that these restricted baseline programs are disrupted under type I interferon exposure, consistent with the well-established reprogramming of monocyte transcriptional states in response to interferon signaling [25, 26].

These results indicate that single-sample localization analysis, followed by joint visualization using condition-specific localized genes, preserves biologically interpretable stimulus-associated programs that may be obscured by integration-based approaches.

Application to subsequent analyses.

This principle of independent per–sample Locat analysis followed by comparison of localization patterns across samples, or localized-gene-based joint visualization when needed, is applied throughout subsequent analyses (see Section 2.8). By avoiding premature integration, we preserve condition-specific localized structure in expression space that would otherwise be erased by alignment procedures. This strategy is particularly critical for detecting rare cell states, transient activation programs, and responses specific to the one condition or sample group in multi-sample experimental designs.

2.8 Temporal localization dynamics reveal stage-specific regulatory programs during ESC differentiation

Having established the principle of independent per-sample localization analysis (see Section 2.7), we applied this framework to characterize temporal dynamics of gene localization during cellular differentiation. We analyzed a retinoic acid (RA) differentiation time course profiling mouse embryonic stem cells (ESC) at days 0, 2, 4, and 10 following induction, with a total of 8 samples, each representing a biological replicate of one time point [27]. This dataset has been previously used to define lineage-specific

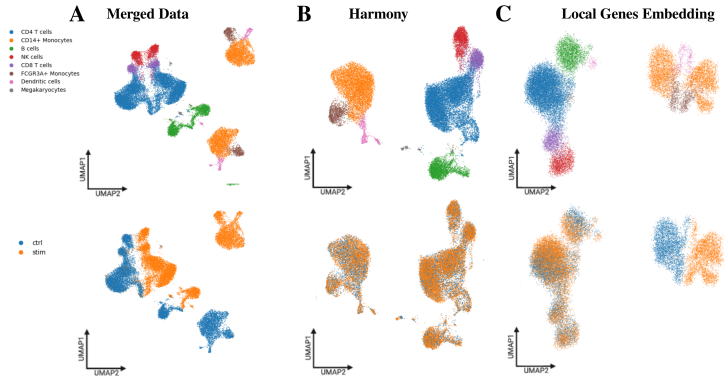

Fig. 4 Locat identifies condition-specific localized gene programs in an informative embedding of IFN-β–stimulated and control PBMCs (a–g) Locat distinguishes shared and stimulus-specific transcriptional programs without batch correction. (a) UMAP of merged control and IFN-β samples showing complete separation before correction. (b) Harmony-corrected embedding with over-alignment and loss of biological signal. (c) Embedding based on top localized Locat genes preserves shared cell types while maintaining stimulus-specific separation of monocytes.

regulatory programs across developmental time [27]. We asked: (i) How do localized gene sets change across developmental stages? (ii) Are localization patterns reproducible across independently analyzed replicates? (iii) Can temporal trajectories of gene localization identify stage-specific regulatory programs? (iv) Can localized genes be used to anchor specific cell populations between samples and time-points?

Each sample was analyzed independently with Locat using the sample's own PCA embedding without batch correction, thus preserving sample-specific embedding structure while enabling comparison of localization scores across developmental stages (Fig. 5). Unless otherwise specified, localization at a given timepoint was determined by comparison of the localization p-values of the given timepoint, combined using geometric mean, to the significance threshold.

Progressive increase in localization marks neural differentiation.

Locat identified 2,431 genes as significantly localized ($p < 0.05$) at at least one timepoint. The number of localized genes increased across the time course: 432 genes at day 0, 515 at day 2, 1,323 at day 4, and 1429 at day 10. Full gene-level results are provided in the Supplementary Data CSV. This increase reflects progressive concentration of gene expression into restricted neighborhoods in latent expression space as differentiation proceeds, consistent with the emergence of increasingly distinct cellular states from initially more homogeneous progenitor populations.

To identify coordinated temporal patterns, we assigned each gene from the localized gene set to one of the 15 possible temporal localization clusters representing the time points in which the gene showed significant localization (by construction we excluded the case in which the gene was not localized in any timepoint). Chi-square analysis revealed highly significant non-independence of temporal localization patterns

($p \sim 2.51 \cdot 10^{-150}$), indicating that localization did not occur randomly but exhibit structured temporal dynamics. The observed temporal structure is likely reflecting the dynamics of biological processes, as supported by our analyses below.

Five dominant temporal patterns capture developmental regulation.

We grouped genes according to their temporal localization patterns across time points and quantified over-representation for each pattern using δ, defined as the difference between the observed frequency of genes assigned to that pattern and the frequency expected if temporal pattern membership were independent across time points. To assess whether observed deviations exceeded what would arise by chance under this independence assumption, we estimated the null distribution of δ using Monte Carlo simulations ($n = 10{,}000$). Under this null model, the mean deviation was $\bar{\delta}_{H_0} = 0 \pm 0.005$, and the largest deviation observed was $\max[\delta_{H_0}] = 0.009 \pm 0.004$. We identified five temporal patterns whose observed δ values substantially exceeded this null range (Supplementary Fig. S15; cluster-level δ values are summarized in Supplementary Table S2 and the full gene-level results are provided in the Supplementary Data CSV).

(1.) Specific emergence on day 10 (cluster 0001, $n = 818$, $\delta = 0.090$). The most over-represented pattern contained genes whose localization emerged at day 10, whose localization became detectable only at day 10, reflecting late-emerging, highly specific expression during later stages of differentiation. Localized genes in this group included regulators associated with ventral neural patterning and lineage specification, including *Nkx2-2* and *Gli2* [28–30], as well as membrane and signaling genes associated with differentiated glial states such as *Gpr17* [31] and *Cldn11* [32]. These 818 genes showed a progressive increase in mean localization scores across the time-course (negative log-10 p-values: 0.93 on day 0, 1.07 on day 2, 1.14 on day 4, 1.42 on day 10), indicating moderate concentration and moderate depletion that strengthened into significant localization only at terminal differentiation. This pattern suggests that lineage-specific programs begin subtly and strengthen through clonal expansion, migration, or selective survival rather than arising de novo at late stages. As a result, while differentiation progresses, cell types become more segregated on the embedding due to increasingly stronger diversification of transcriptional programs.

(2.) Specific transient localization on day 4 (cluster 0010, $n = 624$, $\delta = 0.049$). This cluster captured genes localized at day 4 following retinoic acid exposure, reflecting a short-lived reorganization of transcriptional programs during early post-induction responses. Localized genes in this cluster included factors associated with pluripotency exit and early differentiation, including *Tfcp2l1*, *Zfp42*, *Tbx3*, and *Klf5* [12], together with epithelial-associated genes such as *Krt8*, *Epcam*, and *Esrp2* [13, 15]. The coexistence of pluripotency-linked and epithelial organizational genes suggests that this pattern reflects a transient intermediate state during the early response to retinoic acid, before stabilization of more restricted lineage-specific transcriptional programs [12, 33].

(3.) Specific transient localization on day 2 (cluster 0100, $n = 223$, $\delta = 0.039$). Genes in this cluster localized specifically at day 2, defining an early and transient pattern that coincides with pluripotency exit and the beginning of lineage specification. Localized genes in this cluster included regulators associated with

pluripotency and its early dissolution, such as *Nanog*, *Zscan4c*, and *Esrrb* [34–36], as well as early neural and patterning factors including *Sox21* and *Pax3* [28]. This cluster also included epithelial-associated genes such as *Cldn2* and *Aqp3* [13]. The restriction of localization to day 2 indicates that these programs are transiently organized during an intermediate differentiation stage, after pluripotency exit but before stable lineage commitment [12].

(4.) Specific localization at day 0 (cluster 1000, $n = 199$, $\delta = 0.038$). Genes in this cluster were localized at day 0 prior to induction, reflecting a baseline cell state organization characteristic of the pre-differentiation condition that is rapidly altered upon retinoic acid exposure. Localized genes in this cluster included transcriptional regulators associated with pluripotency maintenance and early developmental programs, such as *Prdm14* [37], *Zbtb16* [38], and *Meis3* [39], alongside genes involved in cell–cell adhesion, extracellular matrix organization, and baseline cellular structure, including *Cdh6*, *Cdh9*, *Dcn*, and *Bgn*, consistent with stable regulatory and structural programs in pluripotent or pre-induction cell states [39, 40]. This cluster also contained genes linked to homeostatic signaling and surface or receptor-associated functions, such as *Thy1*, *Il15ra*, and *Trem2*, consistent with transcriptionally stable but heterogeneous baseline populations [41]. The restriction of localization to day 0 suggests that these programs are specific to the pre-induction state and are rapidly dissolved or reorganized upon retinoic acid exposure, reflecting loss of the initial baseline configuration rather than persistence of a differentiated identity [42].

(5.) Sustained localization: days 0 → 2 → 4 (cluster 1110, $n = 47$, $\delta = 0.009$). Genes in this cluster exhibited sustained localization from day 0 through day 4, but not at day 10, indicating persistent cell populations during early differentiation prior to later state diversification. This set includes transcriptional regulators that define early endodermal and epithelial programs, such as *Sox17* [43], *Foxa2*, *Gata4*, *Hnf1b*, and *Hnf4a*, which are central components of early endodermal patterning and epithelial specification networks [44–46]. The cluster also contains genes associated with epithelial function and metabolic organization, including *Fabp1*, *Muc13*, *Aqp8*, and *Enpep*, reflecting functional programs that accompany early epithelial differentiation [47].

The persistence of localization across early time points suggests stable organization of these transcriptional and functional programs during initial patterning, whereas the loss of localization by day 10 indicates redistribution of these programs as later, more specialized cell states emerge [48].

Together, these temporal localization trajectories provide a dynamic view of cellular differentiation as a complex reorganization where early transcriptional programs dissolve while new lineage-specific organization emerges, marking distinct developmental transitions across time, which we detect as localization events on embedding of intermediate time-points.

Gene localization analysis highlights a developmental watershed between day 4 and day 10.

The most pronounced reorganization of gene localization occurs between day 4 and day 10. Although comparable numbers of genes were localized at each stage (1,323 at day 4, and 1429 at day 10), only 529 genes were shared between these timepoints, indicating

a substantial restructuring of transcriptional organization. This shift reflects a major transition in cellular state organization during retinoic acid–driven differentiation. As described above, genes enriched in day 4 over-represented clusters (clusters 0010 and 0110) are associated with early patterning, epithelial, and progenitor-associated programs, whereas genes localized specifically at day 10 (cluster 0001) reflect increasingly organized neural lineage programs. The near-complete turnover of localized gene sets suggests that the interval between day 4 and day 10 marks a developmental watershed characterized by the resolution of early, multi-lineage transcriptional organization and the emergence of more lineage-restricted neural programs. This transition is consistent with established retinoic acid differentiation dynamics, in which early patterning and endoderm-associated programs diminish as neural identities become increasingly dominant [49].

Differential localization enables tracking of stable cell populations across developmental transitions.

Beyond identifying stage-specific programs, temporal localization analysis reveals a particularly informative form of differential localization: genes that are expressed across conditions but transition from non-localized to localized as cellular heterogeneity increases. This pattern is exemplified by *Id3* and *Cldn3*, which exhibit distinct but complementary localization trajectories over the differentiation time course (Fig. 5). *Id3*, a regulator associated with proliferative progenitor states [50], is broadly expressed and weakly or non-localized at early time points but becomes localized as progenitor populations begin to segregate during differentiation. In contrast, *Cldn3* [51], which encodes a tight-junction component linked to epithelial organization, is unlocalized at day 0, becomes weakly structured at intermediate stages, and is strongly localized by day 10.

These changes do not imply altered gene function or cell-intrinsic regulation. Rather, they reflect shifts in *population context*. At early time points, the culture is dominated by relatively homogeneous, cycling pluripotent cells, causing broadly required regulators and structural components to be widely expressed [52]. As differentiation proceeds, increasing cellular diversity and the emergence of post-mitotic lineages restrict both proliferative and epithelial programs to specific subpopulations, allowing genes such as *Id3* and *Cldn3* to exhibit localized expression patterns [53–55].

The ESC differentiation dataset provides a particularly clear example of this regime of differential localization. As differentiation proceeds and population composition diversifies, genes that are broadly associated with early progenitor populations become increasingly confined to specific subpopulations. Localization often emerges gradually: genes may exhibit modest confinement at intermediate time points before becoming more strongly localized as additional cell states arise. With increasing heterogeneity, these genes are not only enriched within particular lineages but also depleted from surrounding states. Detecting such transitions therefore requires evaluating both concentration within a target population and absence elsewhere.

This behavior is illustrated by genes such as *Id3* and *Cldn3*, which show progressive confinement to progenitor or epithelial subpopulations across the time course. Although their degree of localization varies across stages, their enrichment within

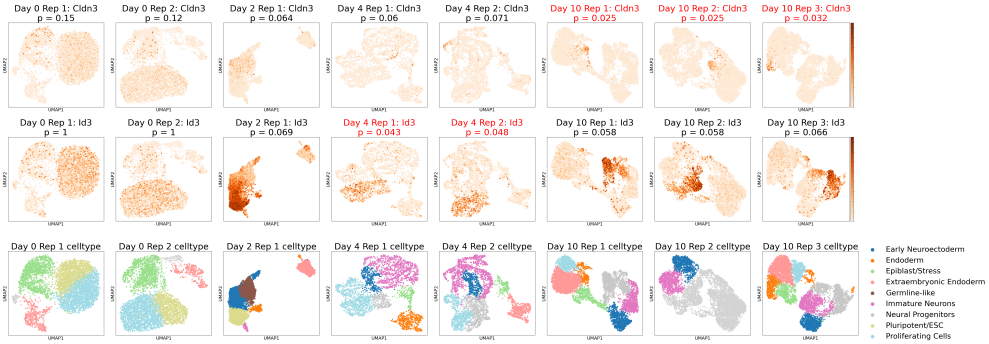

Fig. 5 Locat reveals dynamic localization patterns in RA-induced murine ESC developmental timecourse UMAP embeddings for eight samples from day 0 to day 10 of RA treatment. Top and middle rows plot average expression of two genes that are conditionally localized in phases of differentiation after RA exposure. Titles in the first two rows track magnitude and significance of localization p-values. The third row displays annotated cell types corresponding to each of the eight samples.

specific lineages is accompanied by reduced expression in emerging states. By jointly capturing enrichment and depletion, localization analysis enables identification of cell populations that persist across developmental transitions despite substantial shifts in overall population composition.

3 Discussion

In the present work we refine the definition of a "marker gene" in single-cell analysis as a gene whose expression is both concentrated within a specific subset of cells and significantly depleted in complementary subsets. This refinement addresses a limitation of enrichment-based approaches, which can identify genes that vary across cells yet remain broadly expressed across much of the population. Such genes may exhibit gradual trends or diffuse upregulation without clearly delineating distinct cell populations. By explicitly requiring statistical evidence of depletion outside enriched neighborhoods, we prioritize genes whose expression is restricted to well-defined subsets rather than broadly distributed. In this framework, depletion provides the statistical criterion that distinguishes specific population markers from genes that merely reflect structured variation.

Existing marker gene selection methods use a variety of approaches: gene autocorrelation in embedding space (Hotspot [3]), density-based scoring (Haystack [2]), differential expression tests (Scanpy [10]), and diffusion-based metrics (LMD [5], GSPA [6]). Yet, these methods are designed around enrichment-oriented signals rather than an explicit enrichment-plus-depletion criterion. Our approach, Locat, tests both concentration and depletion simultaneously through density contrasts, distinguishing genes that satisfy this dual criterion from genes that concentrate locally but retain substantial background expression.

Locat identifies localized genes whose expression exhibits sharp contrast against the global background density. These genes therefore form a highly specific, compact,

and interpretable feature set within each sample that is well suited for visualization and downstream analyses. In multi-sample studies, we find it most informative to perform marker gene discovery separately within each sample's embedding, then compare localization scores across samples and, when appropriate, construct joint visualizations using the union of condition-specific localized genes. This strategy avoids applying batch correction during marker gene discovery and directly addresses a central challenge in current analytical pipelines: joint embeddings without batch correction are often dominated by sample-specific and technical variation, while batch-correction–based integration can suppress or obscure the condition-specific gene programs that experimental designs aim to study.

Where controlled perturbations are applied to shared lineages, comparing localization scores across independently embedded samples offers a conservative alternative to aggressive integration. For samples with higher variation, or time courses with varying cell population composition, we demonstrate that localized genes identified in each sample separately enable anchoring populations across samples—for instance, linking minority subpopulations from later developmental stages to the ancestral proliferating pool, or revealing that an apparent global shift in the sample's transcriptional program had been subtly building up in earlier stages, only to appear dominant in the final population.

Locat inherits limitations from both density modeling and the embedding geometry on which it operates. Inference is performed in the user's chosen coordinate system, and distortions of distance scales or neighborhood relationships can therefore influence both concentration and depletion evidence. In this work, we primarily fit models in higher-dimensional PCA coordinates and used UMAP mainly for visualization, as low-dimensional nonlinear embeddings may warp global distances and relative scales in ways that are not faithful to the data. We therefore recommend fitting Locat on embeddings that are well supported for the biological signals of interest, prioritizing representations with stable geometry, and interpreting conclusions that depend on fine-grained distance relationships in two-dimensional nonlinear embeddings with caution. At the same time, cross-fold reproducibility analyses in the IFN-β dataset indicate that localization rankings are reasonably stable across the tested embedding choices, including 4, 8, and 12 PCA dimensions and a BiPCA→UMAP setting, with cross-fold AUROC values of 0.923–0.979 and top-500 Spearman correlations of 0.681–0.809 (Supplementary Tables S3 and S5), while within-fold seed stability remained similarly high across embedding choices (Supplementary Table S4). Another practical consideration is expression prevalence: genes expressed in very few cells provide limited effective sample sizes, which can reduce statistical power and increase sensitivity to modeling choices. While our simulation-based calibration explicitly accounts for the dependence of localization statistics on expression prevalence and includes effective sample size diagnostics, signals from extremely rare genes remain challenging for any localization-based approach.

Several directions for future work arise from this framework. The concentration–depletion logic may be adapted to spatial transcriptomics data, where physical tissue coordinates provide a natural representation for assessing localized expression. In such settings, testing for enrichment within spatial neighborhoods and depletion

outside them could help distinguish genes marking anatomically confined regions from those that are more broadly distributed across tissue. Related opportunities exist in multi-modal assays such as single-cell multiome or CITE-seq, where evaluating whether transcriptional localization is accompanied by coordinated restriction in chromatin accessibility or protein abundance may provide stronger evidence for coherent regulatory programs. Finally, extending the framework to evaluate sets of genes jointly, rather than testing each gene independently, may enable identification of module-level markers defined by coordinated expression patterns. Together, these possibilities suggest that the concentration–depletion principle may provide a unifying perspective for identifying specific biological programs across diverse single-cell modalities.

4 Methods

4.1 Locat

4.1.1 Input Data

Locat takes two inputs: a gene by cell normalized count matrix (the expression matrix) $W \in \mathbb{R}^{m \times n}$, where G is the set of all genes ($|G| = m$) and C is the set of all cells ($|C| = n$), and a d-dimensional cell embedding $X \in \mathbb{R}^{n \times d}$, such that each cell $c \in C$ has normalized count vector $w_{*,c}$ and embedding coordinate $\mathbf{x}_c \in \mathbb{R}^d$. Locat evaluates localization with respect to the embedding X using normalized counts W as informative weights.

4.1.2 Density estimation

First, Locat models both the background cell density and gene-specific expression density using weighted Gaussian mixture models (WGMMs, Supplementary Methods S1.1). By default, the background WGMM uses uniform per-cell weights; optionally, sample-specific weights (for example, total expression count per cell) can be used. Gene-specific WGMMs use the expression level of that gene across cells. In all cases, weights are normalized to sum to one. We use BIC (Supplementary Methods S1.1.3) as a model-selection criterion that balances goodness of fit and model complexity for selecting background model complexity and guiding component selection.

The background WGMM $f_0(\mathbf{x})$ is fit once per dataset using all cells in the embedding. We average over multiple random initializations to stabilize the estimated background density values $f_0(\mathbf{x}_c)$ at each cell c. For each gene g, we fit a signal WGMM $f_g(\mathbf{x})$ using per-cell weights $w_{g,*}$ normalized to sum to one. The number of components in this gene-specific model is adapted to the expression pattern of each gene: genes spread across larger regions of the embedding are assigned more components, while compact patterns use fewer components (Supplementary Methods S1.1.4).

4.1.3 Concentration Score

Next, we quantify the concentration score, which evaluates whether gene expression density concentrates into localized peaks within the embedding. We first compute a

normalized difference between background and signal density. We then convert this raw score to a z-score by comparing it to a null distribution built from randomized pseudo-genes with similar prevalence (fraction of expressing cells), so genes are compared fairly across different detection rates. Finally, we compute statistical significance from that z-score. Before scoring, genes with prevalence above a user-defined upper bound are excluded (parameter `max_freq`, default 0.9).

Let π_g be the fraction of cells with $w_{gc} > 0$, which is the expression prevalence of gene g. Recall that $f_0(\mathbf{x}_c)$ is the background density at cell c, and $f_g(\mathbf{x}_c)$ is the gene-specific density for gene g at cell c.

Concentration is based on a density contrast score comparing gene-specific and background densities, defined as:

$$\gamma_{g,c} \;=\; \sqrt{\pi_g}\,\left(1-\sqrt{\pi_g}\right)\,\frac{f_g(\mathbf{x}_c)-f_0(\mathbf{x}_c)}{f_g(\mathbf{x}_c)+f_0(\mathbf{x}_c)}. \tag{5}$$

This quantity is positive where the gene density exceeds the background and negative otherwise. The prevalence factor $\sqrt{\pi_g}\left(1-\sqrt{\pi_g}\right)$ stabilizes the statistic for sparse genes while downweighting broadly prevalent genes.

The per gene concentration statistic is the expression weighted average of $\gamma_{g,c}$ over expressing cells,

$$z_g = \frac{\sum\limits_{c\in C:, w_{g,c}>0} w_{g,c}\,\gamma_{g,c}}{\sum\limits_{c\in C:, w_{g,c}>0} w_{g,c}}. \tag{6}$$

We then adjust z_g to account for prevalence as described in Supplementary Methods S1.2. Specifically, we estimate $\mu(\pi)$ and $\sigma(\pi)$ from randomized pseudo-genes generated across prevalence levels and interpolate these functions over π. For gene g with prevalence π_g, the prevalence-adjusted concentration z-score is:

$$Z_g = \frac{z_g - \mu(\pi_g)}{\sigma(\pi_g)}, \tag{7}$$

and compute the one-sided concentration p-value

$$p_{c,g} = 1 - \Phi(Z_g), \tag{8}$$

where Φ is the standard normal cumulative distribution function. Genes whose expressing cells lie in regions where f_g is consistently larger than f_0 result in a large positive Z_g and therefore small $p_{c,g}$.

4.1.4 Depletion Score

The second term in our dual criterion is the depletion score. It tests whether a gene is under-represented in parts of the embedding where many cells are expected under the background model. We evaluate this at several contrast levels $\lambda \geq 1$ to examine different region definitions: lower λ values include larger regions of depletion that may extend toward boundaries between depleted and enriched areas, whereas higher λ values define smaller, stricter regions that are more clearly separated from regions

of enrichment. For each λ, we define a region of depletion and compute two fractions in that region: (i) the expected fraction under the background model and (ii) the observed fraction among cells expressing the gene. Using these quantities, we perform a one-sided depletion test for that contrast, where the statistic is more significant when the observed fraction is further below the expected fraction under the null model. The corresponding p-value uses an effective sample size adjustment to account for unequal weights and dependence among nearby cells in the embedding, which would otherwise underestimate variance and inflate significance. Because multiple λ values are evaluated for each gene, we apply a multiple-testing correction across that scan.

Regions of relative depletion and expected density mass.

Let f_0 and f_g be the background and gene g WGMM density functions respectively. Given a specific threshold $\lambda \geq 1$, termed the contrast threshold, we determine the *region of depletion* as the set of cells Γ_λ for which the background density exceeds the gene-specific density by a factor at least equal to λ:

$$\Gamma_\lambda = \{c : f_0(\mathbf{x}_c) > \lambda f_g(\mathbf{x}_c)\}\,, \tag{9}$$

The expected background mass represents the fraction of total background density falling within the region of depletion:

$$\beta_0(\lambda) = \frac{\sum\limits_{c \in \Gamma_\lambda} f_0(\mathbf{x}_c)}{\sum\limits_{c \in C} f_0(\mathbf{x}_c)}. \tag{10}$$

The observed gene-specific mass for gene g is the fraction of expressing cells falling within the region of depletion:

$$\beta_g(\lambda) = \frac{\sum\limits_{c \in \Gamma_\lambda} \mathbf{1}\{w_{g,c} > 0\}}{\sum\limits_{c \in C} \mathbf{1}\{w_{g,c} > 0\}}. \tag{11}$$

The two definitions above are intentionally different. $\beta_0(\lambda)$ is model-based: it is computed from fitted background density values $f_0(\mathbf{x}_c)$ and represents the expected fraction of background density contained in Γ_λ under the null. $\beta_g(\lambda)$ is empirical: it is computed from observed expressing cells via the indicator $\mathbf{1}\{w_{g,c} > 0\}$ and represents the observed fraction of expressing cells in the same region of depletion. Under the null, these two quantities should be similar. Evidence for depletion is obtained when $\beta_g(\lambda)$ is substantially smaller than $\beta_0(\lambda)$. In the default analysis, each expressing cell contributes equally to $\beta_g(\lambda)$, so depletion is tested on expression frequency rather than expression magnitude. Optionally, cells can instead be weighted by nonnegative expression values, yielding a depletion test on total expression amount.

Both expected and observed quantities are computed by summing over the observed cells in the dataset, rather than integrating over the full continuous embedding space. This keeps the test tied directly to the cells and embedding positions that

were actually measured. Integrating over the full embedding would be computationally expensive and would assign weight to regions where no cells were observed.

Tail probability.

Under the null hypothesis that cells expressing gene g are distributed in proportion to the background density, we compute a one-sided lower-tail probability using a beta-binomial model with effective trial count $n_{\text{trials}} = \text{round}(n_{\text{eff}})$. For each contrast λ, let $k_{\text{obs}}(\lambda) = \text{round}(\beta_g(\lambda)\, n_{\text{trials}})$ and $p_0(\lambda) = \beta_0(\lambda)$. Here ρ_{bb} is a fixed overdispersion parameter that controls extra variability relative to a binomial model; the default setting is $\rho_{\text{bb}} = 0.02$. We define

$$\alpha_\lambda = p_0(\lambda)\left(\frac{1}{\rho_{\text{bb}}} - 1\right), \qquad \beta_\lambda = (1 - p_0(\lambda))\left(\frac{1}{\rho_{\text{bb}}} - 1\right), \tag{12}$$

and evaluate

$$p_{d,g}(\lambda) = \sum_{j=0}^{k_{\text{obs}}(\lambda)} \binom{n_{\text{trials}}}{j} \frac{B(j + \alpha_\lambda,\ n_{\text{trials}} - j + \beta_\lambda)}{B(\alpha_\lambda, \beta_\lambda)}, \tag{13}$$

where $B(\cdot,\cdot)$ is the beta function. If $\rho_{\text{bb}} = 0$, the test uses a binomial lower-tail calculation. Here n_{eff} is Kish's effective sample size for the given gene g,

$$n_{\text{eff}} = \frac{\left(\sum_{c \in C} w_{g,c}\right)^2}{\sum_{c \in C} (w_{g,c})^2}, \tag{14}$$

with $w_{g,c}$ being the normalized count vector for gene g in cell c. Equation 14 reflects the reduction in effective degrees of freedom when observations carry unequal weights. Using the raw cell count n would treat all cells as equally informative and yield overly permissive p-values when a small number of highly weighted cells dominate the test statistic. To further ensure stability and conservativeness, we apply a multiplicative scaling factor to n_{eff} and then impose an upper bound on the scaled n_{eff} (see Supplementary Methods for parameterization details). In the primary analyses reported here, we used a fixed effective-sample-size scaling factor of 0.6, an upper cap of $\sqrt{n_{\text{cells}}}$ on effective trials, and a beta-binomial overdispersion parameter $\rho_{\text{bb}} = 0.02$ (Supplementary Table S1). These settings were held constant across datasets as fixed analysis settings rather than tuned per analysis, and their combined effect is assessed empirically through the null-calibration analyses in Supplementary Fig. S17. These adjustments account for the fact that f_0 and f_g are not independent: f_g is fit using a subset of the same cells that define f_0, and both densities are fit to cellular positions in the embedding, where nearby cells share expression patterns.

Selecting a contrast

Genes may exhibit significant depletion at different contrast thresholds, depending on the shape of their density profiles and how sharply expression declines outside concentrated regions. We therefore evaluate the depletion tail probability over a geometric grid of λ values with $\lambda \geq 1$, using a moderate finite range of contrasts, and account for this scan in the final p-value calculation. To ensure that the test is well-conditioned and supported by sufficient evidence, we restrict attention to the subset Λ of λ values satisfying the following criteria:

- the background density mass $\beta_0(\lambda)$ in the region is sufficiently large to support a meaningful comparison;
- the expected effective count $n_{\text{eff}}\beta_0(\lambda)$ exceeds a minimum sample-size threshold;
- the absolute deficit $\beta_0(\lambda) - \beta_g(\lambda)$ exceeds a minimum threshold;
- the observed gene-specific mass does not exceed the level implied by the contrast ratio $1/\lambda$, i.e., $\beta_g(\lambda) < (\beta_0(\lambda)/\lambda)\,(1 - \varepsilon_{\text{rel}})$.

These conditions exclude contrasts that produce extremely small regions, unstable effective sample sizes, or negligible deviations from the null.

Among admissible contrasts, we retain the minimum $p_{d,g}(\lambda)$ over Λ. If no contrast satisfies the above conditions, the depletion p-value is set to 1.

To adjust for multiplicity across scanned λ values, we apply a Šidák-form adjustment based on the effective number of tested contrast thresholds n_Λ, which results in the depletion score:

$$p_{d,g} = 1 - \left(1 - \min_{\lambda \in \Lambda} \left[p_{d,g}(\lambda)\right]\right)^{n_\Lambda} \tag{15}$$

Because tests across λ are correlated, this adjustment should be interpreted as an approximate multiplicity correction across tested λ values. Small $p_{d,g}$ indicates that gene g is under-represented across a substantial portion of the background density (high f_0), consistent with exclusion from areas outside its concentrated regions.

4.1.5 Localization Score

Finally, we combine the concentration and depletion scores to obtain one localization score. The localization score is adjusted with penalization terms aimed at improving robustness for small sample-sizes.

In detail, a raw localization score is computed using a Cauchy combination test [56], which is meant to produce a valid omnibus p-value even under complex dependence:

$$p_{\text{loc},g} = \frac{1}{2} - \frac{1}{\pi}\arctan(T_g), \tag{16}$$

where

$$T_g = \tan\left[\left(\frac{1}{2} - p_{c,g}\right)\pi\right] + \tan\left[\left(\frac{1}{2} - p_{d,g}\right)\pi\right] \tag{17}$$

To compute the adjusted localization score we incorporate two small penalty terms that stabilize the combined statistic. The first penalizes genes supported by very few expressing cells, for which density estimates can be unstable. The second penalizes genes whose fitted signal density does not dominate the background at most of their

expressing cells, thereby discouraging diffuse or poorly separated expression patterns even within their putative regions of concentration.

Let $n_g = \sum_{c \in C} \mathbf{1}\{w_{g,c} > 0\}$ be the number of cells expressing gene g i.e., the support of gene g. We define the sparsity adjustment factor

$$h_g^{(\text{size})} = 1 - e^{\left(-\frac{1}{n_g+1}\right)}, \tag{18}$$

which is close to 0 for well sampled genes and larger for rare genes. We also define the sensitivity adjustment factor

$$h_g^{(\text{sens})} = 1 - \left(\frac{\sum_{c \in C:\, w_{g,c} > 0} \mathbf{1}\{f_g(\mathbf{x}_c) > f_0(\mathbf{x}_c)\}}{n_g} \right), \tag{19}$$

as the fraction of cells expressing gene g for which the gene density exceeds the background density.

The adjusted localization score i.e., the output from Locat, is

$$p_g = 1 - (1 - p_{\text{loc},g})\,(1 - \alpha_{size} h_g^{(\text{size})})\,(1 - \alpha_{sens} h_g^{(\text{sens})}), \tag{20}$$

clipped in $(0, 1)$. The coefficients α_{size} and α_{sens} control the strength of sparsity and sensitivity penalization, respectively. Default parameter values and robustness analyses are described in Supplementary Methods S1.4 and Supplementary Fig. S16. Finally, we apply a monotone smoothing map near the upper tail:

$$\tilde{p}_g = \begin{cases} p_g, & p_g \le 0.99, \\ 1 - \dfrac{0.01}{1 + 100\,(p_g - 0.99) + 10^4\,(p_g - 0.99)^2}, & p_g > 0.99. \end{cases} \tag{21}$$

This transformation only affects values very close to one and preserves rank ordering, while preventing genes from collapsing to nearly identical upper-tail values. As a result, it primarily improves numerical separation among insignificant genes and has negligible effect on which genes are called significantly localized. The reported final localization p-value is $\tilde{p}_g$.

By construction, p_g is a conservative adjusted localization score: Eq. 20 is monotone and satisfies $p_g \ge p_{\text{loc},g}$. Thus, the penalty terms can only increase the score relative to the Cauchy-combined value, stabilizing rankings for sparse or weakly separated genes rather than introducing additional significance. Robustness of gene rankings to the penalty coefficients α_{size} and α_{sens} is shown in Supplementary Methods S1.4 and Supplementary Fig. S16.

4.1.6 Locat Output

For each gene g, Locat is designed to explicitly report:

- p_g: adjusted localization score (eq. 20);
- $p_{c,g}$: concentration significance (eq. 8);

- $p_{d,g}$: depletion based localization significance (eq. 15);
- Z_g: prevalence-adjusted concentration z-score (eq. 7);
- a BIC-like score that quantifies the parsimony of the fitted WGMM (Supplementary Methods S1.1.3);
- K_g: estimated number of mixture components in the gene's signal WGMM (Supplementary Methods S1.1.4);

Genes with p_g below a chosen significance threshold are considered localized. In this setting, we evaluate calibration empirically under simulation-based null analyses (Supplementary Methods S1.3) rather than applying a single genome-wide correction across genes. This is because the gene-level tests are not an exchangeable family: all genes are tested in the same embedding with a shared background/gene-specific GMM density framework, while prevalence and noise vary substantially across genes.

4.2 Modules identification

To identify clusters of co-expressed genes, we extract the gene-by-cell expression matrix for a specified set of input genes. We binarize the expression matrix using a threshold on expression levels. For each gene pair, we quantify overlap between the resulting binary expression vectors across cells and evaluate whether the observed co-expression across cells exceeds baseline expectation using a hypergeometric test. We retain gene pairs with significant co-occurrence across cells as positive edges to form a binary gene–gene adjacency matrix. Louvain community detection then partitions this graph to recover gene modules i.e., clusters of genes with substantial co-expression.

In detail, the entries $b_{g,c}$ of the binarized expression matrix $B = (b_{g,c})$, derived from the normalized expression matrix W as follows:

$$b_{g,c} = \begin{cases} 1 & \text{if } (w_{g,c} > t) \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

where G is a pre-defined list of genes, $w_{g,c}$ is the expression level for gene g in cell c, and t is a threshold on gene expression. In our analyses $t = 0$.

The entries a_{g_1,g_2} of the gene-gene adjacency matrix, where g_1 and g_2 are two genes, are:

$$a_{g_1,g_2} = \begin{cases} 1 & \text{if } p_{h,g_1,g_2} \leq \alpha \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

where p_{h,g_1,g_2} is the hypergeometric p-value for the overlap between binary expression vectors $b_{g_1,*}$ and $b_{g_2,*}$, and α is the chosen significance threshold, in our analyses $\alpha = 0.05$.

To ensure interpretability and stability, communities containing very few genes are discarded. The remaining communities represent gene modules.

4.3 Data Preprocessing

Datasets were accessed as annotated AnnData or Seurat objects when available. When preprocessed matrices or embeddings were not provided, we applied standard normalization and dimensionality reduction to obtain inputs for Locat. In all analyses, we used the log-normalized count matrix as the source of cellular weights in

the gene-specific WGMM models. Prior to analysis, we removed genes expressed in fewer than 1% of cells and excluded genes with symbols indicative of predicted, uncharacterized, or non–protein-coding loci (e.g., Gm-, Rik-, AC-/AA-prefixed genes and related annotations), retaining a predominantly protein-coding gene set. For embedding construction, we used the first several principal components derived from the log-normalized expression matrix, typically between 4 and 8 components across datasets.

4.4 Neighborhood Differential Abundance

To identify neighborhoods associated with condition- or cohort-specific enrichment, we performed neighborhood differential abundance (nDA) analysis, a neighborhood overrepresentation procedure adapted from Landa *et al.* [57]. Given cell coordinates $X \in \mathbb{R}^{n \times d}$ and a binary (or probabilistic) group assignment, nDA produces a final per-cell score $s_c^{\mathrm{nDA}} \in [-1, 1]$, where $s_c^{\mathrm{nDA}} > 0$ indicates local enrichment and $s_c^{\mathrm{nDA}} < 0$ indicates local depletion of the group of interest.

Let $y_c \in \{0, 1\}$ denote the hard group label for cell c (or $q_c \in [0, 1]$ a soft membership probability). We first build a single `NearestNeighbors` index on X and query k-nearest-neighbor sets $\mathcal{N}_k(c)$ for $k \in \{15, 40, 75\}$. For each cell c and each scale k, we compute the local group-1 fraction

$$\widehat{q}_{c,k} = \frac{1}{k} \sum_{c' \in \mathcal{N}_k(c)} z_{c'}, \tag{24}$$

where $z_{c'} = y_{c'}$ for hard labels and $z_{c'} = q_{c'}$ for soft labels. We also compute the global group-1 prevalence

$$q_1 = \frac{1}{|C|} \sum_{c \in C} z_c. \tag{25}$$

We convert each (c, k) into a signed enrichment score using a Bayes-rule-derived transform,

$$s_{c,k} = 2 \left[\frac{\widehat{q}_{c,k}\, q_1}{\widehat{q}_{c,k}\, q_1 + (1 - \widehat{q}_{c,k})(1 - q_1)} \right] - 1, \tag{26}$$

yielding a multiscale score matrix $\mathbf{S} = [s_{c,k}]$.

Scores are computed for each k in parallel chunks to reduce memory overhead. To aggregate information across scales, we fit a logistic regression classifier to predict group membership from $\mathbf{S}$, and linearly rescale the fitted posterior probabilities to $[-1, 1]$ to obtain the final nDA score s_c^{nDA} per cell. This multiscale logistic aggregation increases sensitivity to neighborhood-size heterogeneity while preserving the direction of enrichment.

Implementation details: nDA was implemented in Python using `scikit-learn` for nearest-neighbor search and logistic regression. Analyses were performed on PCA coordinates with Euclidean distance, using balanced class weights. In all experiments, we evaluated neighborhood sizes $k \in \{15, 40, 75\}$ and used default logistic regression settings.

4.5 Embeddings Derived from Gene Subsets

In selected analyses, we constructed additional low-dimensional embeddings restricted to specific gene subsets, such as localized genes identified by Locat or predefined gene partitions (e.g., differentiation trajectory genes in embryonic dermis; see Section 2.6). To do so, we subset the log-normalized expression matrix to the genes of interest, recomputed principal components from this restricted matrix, and generated two- or three-dimensional UMAP embeddings for visualization. When multiple experimental conditions were analyzed jointly (e.g., control and stimulated samples in interferon-β stimulation data; see Section 2.7), embeddings were constructed from the combined dataset using the shared gene subset. These subset-derived embeddings were used for visualization and comparative analysis.

4.6 Likelihood-ratio Test for localization

As a complementary diagnostic for gene localization, we implemented a likelihood-ratio test (LRT) comparing the gene-specific weighted Gaussian mixture model (WGMM) density to the background WGMM density on the full embedding (see Supplementary Methods S1.1). Both models are fit as described above, using the same embedding coordinates and covariance regularization strategy described in Section 4.1.2.

For a given gene g, let $\mathcal{I}_g = \{c : w_{g,c} > 0\}$ denote the set of cells expressing the gene. We evaluated the log-likelihood of these cells under the background model, $f_0(\mathbf{x}_c)$, and under the gene-specific signal model, $f_g(\mathbf{x}_c)$. The likelihood-ratio statistic was computed as

$$\Lambda_g = \frac{-2}{|\mathcal{I}_g|} \sum_{c \in \mathcal{I}_g} \left[\log f_0(x_c) - \log f_g(x_c)\right], \tag{27}$$

where normalization by $|\mathcal{I}_g|$ ensures comparability across genes with different expression prevalences.

Under the null hypothesis that the distribution of expressing cells in the embedding is consistent with the global background structure, Λ_g was compared to a chi-square distribution with degrees of freedom

$$\mathrm{df}_g = (k_0 - k_g) + 1, \tag{28}$$

where k_0 and k_g denote the number of mixture components in the background and gene-specific models, respectively. The corresponding likelihood-ratio p-value is

$$p_{l,g} = 1 - F_{\chi^2(\mathrm{df}_g)}(\Lambda_g), \tag{29}$$

where $F_{\chi^2(\mathrm{df}_g)}$ is the chi-square cumulative distribution function with df_g degrees of freedom.

Genes expressed in fewer than a minimum number of cells or in more than a specified fraction of cells were excluded from testing. In practice, the likelihood-ratio test was used to quantify the extent to which a gene's expressing cells deviate from the background embedding distribution, but was not used as the primary criterion for localization in Locat.

4.7 Permutation Test for localization

Although local and global gene scores are generally conservative in the standard Locat model, we also provide permutation-based empirical p-values as an optional diagnostic filter. These methods estimate localization p-values for each tested gene and can be computationally expensive, so in practice we suggest applying them to a limited set of candidate genes that already pass the combined localization threshold in the standard Locat model. For each null replicate, we sort cells by the gene's normalized weights, divide them into rank-based bins of similar weight magnitude, and permute weights among cells within each bin. We then multiply these shuffled weights by an independent positive random factor per cell, introducing bootstrap-like variability in each cell's effective contribution, and refit the WGMM for that replicate. Empirical p-values are computed by comparing the observed localization statistic to the replicate-based null distribution across $n_{initializations}$ runs (100 by default).

4.8 Benchmarking workflow

We consider multiple clustering-free tools as well as Locat in several benchmarking experiments. Unless otherwise specified, each of these tools uses the same AnnData object [10], input gene list, and log-normalized reads as input, and uses nearest neighbors, PCs, and/or UMAP coordinates derived from these data. We therefore keep relatively consistent preprocessing steps to ensure each method is fairly evaluated. For ranking comparisons we always sort by the most significant corrected p-value/FDR for each method.

Scanpy-DE: Louvain is applied to the neighborhood graph using default parameters in Scanpy [10], yielding a dictionary of clusters assignments. For each cluster, we use default parameters to calculate one-vs-all differential expression tests the Wilcoxon rank-sum Differential Expression tests over all genes. We then rank genes by significance, using the most significant corrected p-value for each gene (thus, if a gene is significant in two clusters and not significant in any others, the smallest of the two significant p-values will be used to rank the gene against other evaluated genes).

scHaystack: We used default parameters in the Python implementation of singleCellHaystack, using PCs as the input embedding parameter in the *train_haystack_logpv* function. We then rank the genes by sorting the *log_pvalue* output.

Hotspot: We used the Python implementation of Hotspot, specifying the "normal" model be used to measure gene autocorrelation. We then recompute the KNN graph and compute gene autocorrelations. In order to rank the genes, we used the computed false discovery rate (FDR) as a proxy for significance of a given gene.

GSPA: We used the Python implementation of GSPA, following the GitHub repository and examples as a guideline for parameter selection. We compute the PHATE operator [58] as the input embedding to GSPA, similarly to the examples on the GitHub repository. We then use the $gspa.GSPA.calculate_localization()$ function to return localization probabilities per each gene. We found that GSPA was sometimes unstable when trained with large data objects. We created a pseudo-GSPA method

that enabled sparse input to the $gspa.GSPA.construct_graph()$ method and allowed us to calculate $gspa.GSPA.get_gene_embeddings()$ in a sparse, batched format.

LMD: We used the published R implementation of LocalizedMarkerDetector (LMD), following the GitHub repository and tutorial workflow as a guideline for inputs and parameter selection[5]. We passed LMD a genes-by-cells expression matrix and a cells-by-dimensions PCA feature space, and used the default kNN-based graph construction to compute per-gene cumulative localization scores returned by $LocalizedMarkerDetector :: LMD$. To ensure compatibility when calling LMD from Python, we implemented an rpy2 wrapper that constructs R matrices with unique gene and cell identifiers, filters genes expressed to a chosen minimum number of cells, and removes ultra low-variance principal components that can otherwise cause failures. For large datasets we additionally used LMD's coarse-grained large-data workflow ($CoarseGrain$ and $fast_get_lmds$) to compute the same cumulative scores efficiently.

SpectralRH: We implemented SpectralRH as a baseline geometric metric defined in this work (see Supplementary Methods S1.7). For each gene, we calculated Rayleigh smoothness and spectral entropy on the cell–cell graph built from PCA coordinates using Scanpy's `pp.neighbors()` function to construct the normalized Laplacian. We used the first $m = 96$ Laplacian eigenvectors to compute spectral entropy and combined the two measures after sample-size–matched null calibration to obtain a single SpectralRH score per gene. Genes were then ranked by increasing SpectralRH value, such that lower scores correspond to smoother, more locally coherent, and potentially localized expression patterns.

4.8.1 Empirical significance calculations

Neither GSPA nor LMD directly return p-values for gene-level significance. To enable direct comparison across methods and datasets, we converted raw localization scores into empirical significance values using within-sample gene rankings. For each method, genes were ranked by their reported score, with more extreme values indicating stronger evidence for non-uniform expression. Empirical p-values were computed as

$$p_g = \frac{\text{rank}_g}{N+1}, \tag{30}$$

where N is the total number of genes evaluated and ranks were computed using average ranking to handle ties. For LMD, lower cumulative scores correspond to stronger signal, and ranks were therefore computed in ascending order. This rank-based calibration preserves each method's internal gene ordering while providing a standardized scale for cross-method comparison.

4.9 Data availability

All datasets analyzed in this study are publicly available from their original publications. The murine embryonic dermis dataset was obtained from [7, 8]. The IFN-β stimulation PBMC dataset was obtained from [16]. The retinoic acid differentiation time course was obtained from [27]. The 10X PBMC 3k dataset was obtained from [9].

4.10 Code availability

Code for running Locat is available via PyPI at https://pypi.org/project/locat and notebooks for generating results in this article are available at https://github.com/locat-paper.

4.11 Acknowledgements

The authors thank Erez Peterfreund, Peggy Myung, and Boaz Nadler for valuable discussions on related data analysis and methodology, and for insightful feedback on the development of the localization framework.

References

[1] J. M. Pullin and D. J. McCarthy, "A comparison of marker gene selection methods for single-cell rna sequencing data," *Genome Biology 2024 25:1*, vol. 25, pp. 56–, 2 2024.

[2] A. Vandenbon and D. Diez, "A universal tool for predicting differentially active features in single-cell and spatial genomics data," *Scientific Reports 2023 13:1*, vol. 13, pp. 1–14, 7 2023.

[3] D. DeTomaso and N. Yosef, "Hotspot identifies informative gene modules across modalities of single-cell genomics," *Cell Systems*, vol. 12, pp. 446–456.e9, 5 2021.

[4] L. Jiang, H. Chen, L. Pinello, and G. C. Yuan, "Giniclust: Detecting rare cell types from single-cell gene expression data with gini index," *Genome Biology*, vol. 17, pp. 1–13, 7 2016.

[5] R. Li, R. Qu, F. Parisi, F. Strino, H. Lam, J. S. Stanley, X. Cheng, P. Myung, and Y. Kluger, "Cluster-independent multiscale marker identification in single-cell rna-seq data using localized marker detector (lmd)," *Communications Biology 2025 8:1*, vol. 8, pp. 1058–, 7 2025.

[6] A. Venkat, S. Leone, S. E. Youlten, E. Fagerberg, J. Attanasio, N. S. Joshi, M. Perlmutter, and S. Krishnaswamy, "Mapping the gene space at single-cell resolution with gene signal pattern analysis," *Nature computational science*, vol. 4, p. 955, 12 2024.

[7] R. Qu, X. Cheng, E. Sefik, J. S. Stanley, B. Landa, F. Strino, S. Platt, J. Garritano, I. D. Odell, R. Coifman, R. A. Flavell, P. Myung, and Y. Kluger, "Gene trajectory inference for single-cell data by optimal transport metrics," *Nature Biotechnology 2024 43:2*, vol. 43, pp. 258–268, 4 2024.

[8] E. Peterfreund, O. Lindenbaum, Y. Kluger, and B. Landa, "Partition first, embed later: Laplacian-based feature partitioning for refined embedding and visualization of high-dimensional data," *OpenReview.net*, 2022.

[9] "3k pbmcs from a healthy donor, single cell rna sequencing dataset by cell ranger v1.1.0, 10x genomics," *arXiv*, 2016.

[10] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: Large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, pp. 1–5, 2 2018.

[11] A. Riba, A. Oravecz, M. Durik, S. Jiménez, V. Alunni, M. Cerciat, M. Jung, C. Keime, W. M. Keyes, and N. Molina, "Cell cycle gene regulation dynamics revealed by rna velocity and deep-learning," *Nature Communications 2022 13:1*, vol. 13, pp. 2865–, 5 2022.

[12] T. Kalkan, N. Olova, M. Roode, C. Mulas, H. J. Lee, I. Nett, H. Marks, R. Walker, H. G. Stunnenberg, K. S. Lilley, J. Nichols, W. Reik, P. Bertone, and A. Smith, "Tracking the embryonic stem cell transition from ground state pluripotency," *Development (Cambridge, England)*, vol. 144, pp. 1221–1234, 4 2017.

[13] J. P. Thiery, H. Acloque, R. Y. Huang, and M. A. Nieto, "Epithelial-mesenchymal transitions in development and disease," *Cell*, vol. 139, pp. 871–890, 11 2009.

[14] G. Martello and A. Smith, "The nature of embryonic stem cells," *Annual review of cell and developmental biology*, vol. 30, pp. 647–675, 10 2014.

[15] C. C. Warzecha, P. Jiang, K. Amirikian, K. A. Dittmar, H. Lu, S. Shen, W. Guo, Y. Xing, and R. P. Carstens, "An esrp-regulated splicing programme is abrogated during the epithelial-mesenchymal transition," *The EMBO journal*, vol. 29, pp. 3286–3300, 10 2010.

[16] H. M. Kang, M. Subramaniam, S. Targ, M. Nguyen, L. Maliskova, E. McCarthy, E. Wan, S. Wong, L. Byrnes, C. M. Lanata, R. E. Gate, S. Mostafavi, A. Marson, N. Zaitlen, L. A. Criswell, and C. J. Ye, "Multiplexed droplet single-cell rna-sequencing using natural genetic variation," *Nature Biotechnology*, vol. 36, pp. 89–94, 1 2018.

[17] C. A. Dinarello, "Immunological and inflammatory functions of the interleukin-1 family," *Annual Review of Immunology*, vol. 27, pp. 519–550, 4 2009.

[18] J. W. Schoggins and C. M. Rice, "Interferon-stimulated genes and their antiviral effector functions," *Current Opinion in Virology*, vol. 1, p. 519, 2011.

[19] I. Rusinova, S. Forster, S. Yu, A. Kannan, M. Masse, H. Cumming, R. Chapman, and P. J. Hertzog, "Interferome v2.0: an updated database of annotated interferon-regulated genes," *Nucleic acids research*, vol. 41, 1 2013.

[20] P. R. Andrade, M. Mehta, J. Lu, R. M. B. Teles, D. Montoya, P. O. Scumpia, E. N. Sarno, M. T. Ochoa, F. Ma, M. Pellegrini, and R. L. Modlin, "The cell fate regulator nupr1 is induced by mycobacterium leprae via type i interferon in human leprosy," *PLOS Neglected Tropical Diseases*, vol. 13, p. e0007589, 7 2019.

[21] A. Yabluchanskiy, Y. Ma, R. P. Iyer, M. E. Hall, and M. L. Lindsey, "Matrix metalloproteinase-9: Many shades of function in cardiovascular disease," *Physiology*, vol. 28, p. 391, 11 2013.

[22] J. P. Lévesque, K. Bisht, K. A. Alexander, and I. G. Winkler, "Role and prognostic value of oncostatin m and its receptor osmr in acute myeloid leukemia, myeloproliferative neoplasms and non-hematological malignancies," *Frontiers in Oncology*, vol. 15, p. 1636570, 2025.

[23] J. Svaren, B. R. Sevetson, E. D. Apel, D. B. Zimonjic, N. C. Popescu, and J. Milbrandt, "Nab2, a corepressor of ngfi-a (egr-1) and krox20, is induced by proliferative and differentiative stimuli," *Molecular and cellular biology*, vol. 16, pp. 3545–3553, 7 1996.

[24] C. Rinaldo, A. Prodosmo, F. Siepi, and S. Soddu, "Hipk2: a multitalented partner for transcription factors in dna damage response and development," *Biochemistry and cell biology = Biochimie et biologie cellulaire*, vol. 85, pp. 411–418, 8 2007.

[25] L. B. Ivashkiv and L. T. Donlin, "Regulation of type i interferon responses," *Nature Reviews Immunology*, vol. 14, pp. 36–49, 1 2014.

[26] G. M. Tannahill, A. M. Curtis, J. Adamik, E. M. Palsson-Mcdermott, A. F. McGettrick, G. Goel, C. Frezza, N. J. Bernard, B. Kelly, N. H. Foley, L. Zheng, A. Gardet, Z. Tong, S. S. Jany, S. C. Corr, M. Haneklaus, B. E. Caffrey, K. Pierce, S. Walmsley, F. C. Beasley, E. Cummins, V. Nizet, M. Whyte, C. T. Taylor, H. Lin, S. L. Masters, E. Gottlieb, V. P. Kelly, C. Clish, P. E. Auron, R. J. Xavier, and L. A. O'Neill, "Succinate is an inflammatory signal that induces il-1β through hif-1α," *Nature*, vol. 496, pp. 238–242, 4 2013.

[27] Z. Feng, X. Chen, Z. Duren, J. Xin, H. Miao, Q. Yuan, Y. Wang, and W. H. Wong, "Modeling combinatorial regulation from single-cell multi-omics provides regulatory units underpinning cell type landscape using cregulon," *Genome Biology 2025 26:1*, vol. 26, pp. 220–, 7 2025.

[28] T. M. Jessell, "Neuronal specification in the spinal cord: inductive signals and transcriptional codes," *Nature reviews. Genetics*, vol. 1, pp. 20–29, 2000.

[29] E. Dessaud, A. P. McMahon, and J. Briscoe, "Pattern formation in the vertebrate neural tube: a sonic hedgehog morphogen-regulated transcriptional network," *Development (Cambridge, England)*, vol. 135, pp. 2489–2503, 8 2008.

[30] M. Persson, D. Stamataki, P. T. Welscher, E. Andersson, J. Böse, U. Rüther, J. Ericson, and J. Briscoe, "Dorsal-ventral patterning of the spinal cord requires gli3 transcriptional repressor activity," *Genes & Development*, vol. 16, p. 2865, 11 2002.

[31] K. Simon, S. Hennen, N. Merten, S. Blättermann, M. Gillard, E. Kostenis, and J. Gomeza, "The orphan g protein-coupled receptor gpr17 negatively regulates oligodendrocyte differentiation via g
*alpha*i/o and its downstream effector molecules," *Journal of Biological Chemistry*, vol. 291, pp. 705–718, 1 2016.

[32] J. Bronstein, K. Chen, S. Tiwari-Woodruff, and H. Kornblum, "Developmental expression of osp/claudin-11," *Journal of Neuroscience Research*, vol. 60, no. 3, pp. 284–290, 2000.

[33] G. Bain, W. J. Ray, M. Yao, and D. I. Gottlieb, "Retinoic acid promotes neural and represses mesodermal gene expression in mouse embryonic stem cells in culture," *Biochemical and Biophysical Research Communications*, vol. 223, pp. 691–694, 6 1996.

[34] T. Kalmar, C. Lim, P. Hayward, S. Muñoz-Descalzo, J. Nichols, J. Garcia-Ojalvo, and A. M. Arias, "Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells," *PLoS Biology*, vol. 7, p. e1000149, 7 2009.

[35] J. Jiang, W. Lv, X. Ye, L. Wang, M. Zhang, H. Yang, M. Okuka, C. Zhou, X. Zhang, L. Liu, and J. Li, "Zscan4 promotes genomic stability during reprogramming and dramatically improves the quality of ips cells as demonstrated by tetraploid complementation," *Cell Research 2013 23:1*, vol. 23, pp. 92–106, 11 2012.

[36] N. Festuccia, R. Osorno, F. Halbritter, V. Karwacki-Neisius, P. Navarro, D. Colby, F. Wong, A. Yates, S. R. Tomlinson, and I. Chambers, "Esrrb is a direct nanog target gene that can substitute for nanog function in pluripotent cells," *Cell Stem Cell*, vol. 11, pp. 477–490, 10 2012.

[37] L. Cao, A. Kuratnik, W. Xu, J. D. Gibson, F. Kolling, E. R. Falcone, M. Ammar, M. D. V. Heyst, D. L. Wright, C. E. Nelson, and C. Giardina, "Development of intestinal organoids as tissue surrogates: cell composition and the epigenetic control of differentiation," *Molecular carcinogenesis*, vol. 54, pp. 189–202, 3 2015.

[38] T. M. Liu, E. H. Lee, B. Lim, and N. Shyh-Chang, "Concise review: Balancing stem cell self-renewal and differentiation with plzf," *Stem cells (Dayton, Ohio)*, vol. 34, pp. 277–287, 2 2016.

[39] J. C. Yeo and H. H. Ng, "The transcriptional regulation of pluripotency," *Cell Research*, vol. 23, p. 20, 1 2012.

[40] D. A. C. Walma and K. M. Yamada, "The extracellular matrix in development," *Development (Cambridge, England)*, vol. 147, p. dev175596, 5 2020.

[41] V. E. Angarica and A. del Sol, "Modeling heterogeneity in the pluripotent state: A promising strategy for improving the efficiency and fidelity of stem cell

differentiation," *Bioessays*, vol. 38, p. 758, 8 2016.

[42] L. J. Gudas and J. A. Wagner, "Retinoids regulate stem cell differentiation," *Journal of Cellular Physiology*, vol. 226, p. 322, 2 2011.

[43] X.-B. Qu, J. Pan, C. Zhang, and S.-Y. Huang, "Sox17 facilitates the differentiation of mouse embryonic stem cells into primitive and definitive endoderm in vitro," *Development, Growth & Differentiation*, vol. 50, no. 7, pp. 585–593, 2008.

[44] A. M. Zorn and J. M. Wells, "Vertebrate endoderm development and organ formation," *Annual review of cell and developmental biology*, vol. 25, pp. 221–251, 11 2009.

[45] S. J. Arnold and E. J. Robertson, "Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo," *Nature reviews. Molecular cell biology*, vol. 10, pp. 91–103, 2 2009.

[46] C. A. Séguin, J. S. Draper, A. Nagy, and J. Rossant, "Establishment of endoderm progenitors by sox transcription factor expression in human embryonic stem cells," *Cell Stem Cell*, vol. 3, pp. 182–195, 8 2008.

[47] D. Pei, X. Shu, A. Gassama-Diagne, and J. P. Thiery, "Mesenchymal-epithelial transition in development and reprogramming," *NATure Cell BioloGy —*, vol. 21, pp. 44–53, 2019.

[48] S. H. Orkin, "Diversification of haematopoietic stem cells to specific lineages," *Nature Reviews Genetics 2000 1:1*, vol. 1, pp. 57–64, 2000.

[49] M. Kim, A. Habiba, J. M. Doherty, J. C. Mills, R. W. Mercer, and J. E. Huettner, "Regulation of mouse embryonic stem cell neural differentiation by retinoic acid," *Developmental biology*, vol. 328, p. 456, 4 2009.

[50] Y. Kee and M. Bronner-Fraser, "To proliferate or to die: role of id3 in cell cycle progression and survival of neural crest progenitors," *Genes & Development*, vol. 19, p. 744, 3 2005.

[51] A. I. Baumholtz, A. Simard, E. Nikolopoulou, M. Oosenbrug, M. M. Collins, A. Piontek, G. Krause, J. Piontek, N. D. Greene, and A. K. Ryan, "Claudins are essential for cell shape changes and convergent extension movements during neural tube closure," *Developmental Biology*, vol. 428, p. 25, 8 2017.

[52] S. Pauklin and L. Vallier, "The cell-cycle state of stem cells determines cell fate propensity," *Cell*, vol. 155, p. 135, 9 2013.

[53] L. M. Farkas and W. B. Huttner, "The cell biology of neural stem and progenitor cells and its significance for their proliferation versus differentiation during mammalian brain development," *Current Opinion in Cell Biology*, vol. 20, pp. 707–715, 12 2008.

[54] J. J. Eckert and T. P. Fleming, "Tight junction biogenesis during early development," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1778, pp. 717–728, 3 2008.

[55] A. Trounson, "The production and directed differentiation of human embryonic stem cells," *Endocrine Reviews*, vol. 27, pp. 208–219, 4 2006.

[56] Y. Liu and J. Xie, "Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures," *Journal of the American Statistical Association*, vol. 115, no. 529, pp. 393–402, 2020.

[57] B. Landa, R. Qu, J. Chang, and Y. Kluger, "Local two-sample testing over graphs and point-clouds by random-walk distributions," *arXiv*, 11 2020.

[58] K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. van den Elzen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, and S. Krishnaswamy, "Visualizing structure and transitions in high-dimensional biological data," *Nature Biotechnology*, vol. 37, pp. 1482–1492, 12 2019.