BioRSP: a method for characterizing enrichment patterns in single-cell embeddings

Zeyu Yao[1] and Jake Y. Chen[2*]

[1]Stamford American International School, Singapore 357684

[2]Department of Biomedical Informatics and Data Science, Systems Pharmacology AI Research Center, the University of Alabama at Birmingham, Birmingham, Alabama, US

[*]Corresponding author: Dr. Jake Chen, jakechen@uab.edu

Abstract

Low-dimensional embeddings such as UMAP and t-SNE are routinely used to visually interpret high-dimensional omics data, yet claims based on embedding geometry are often qualitative, embedding-sensitive, and weakly calibrated. We present BioRSP (Biological Radar Scanning Plots), a geometry-first framework that quantifies how a user-defined foreground subset is distributed across the embedding footprint of a fixed analysis set. BioRSP converts 2D coordinates to polar form around a robust vantage point, scans the set by angle, and computes a radar profile that summarizes signed radial enrichment of foreground relative to background within sliding angular windows using a distance-based radial discrepancy. The profile is reduced to interpretable summaries including anisotropy magnitude, peak directionality, and coverage, and is accompanied by explicit adequacy rules and subsampling-based stability diagnostics so the method can abstain when the geometry is underpowered. We demonstrate BioRSP in a community-standard human kidney single-nucleus reference by analyzing thick ascending limb (TAL) nuclei using published UMAP coordinates and a standardized within-set top-decile

foreground rule. Within TAL, BioRSP distinguishes sharply localized rim-enrichment patterns, broadly supported but structured within-type heterogeneity, and near-isotropic profiles, with anisotropy spanning more than an order of magnitude in the pooled TAL analysis. Donor-aware reruns show that per-donor adequacy is frequently limiting, but that directional profiles are stable when donor-level support is sufficient. BioRSP is provided as open-source software producing standardized plots, summary tables, and run manifests to support reproducible embedding-aligned enrichment analysis.

1. Introduction

Low-dimensional neighbor embeddings such as UMAP and t-SNE are now routine for exploring single-cell and other high-dimensional omics data (McInnes et al., 2018; van der Maaten and Hinton, 2008). In these layouts, investigators commonly interpret gradients, asymmetries, and patchy structure as biological organization, including developmental progression, injury responses, tissue zonation, and condition-specific shifts (Becht et al., 2018; Farrell et al., 2018; Yost et al., 2019). Yet this interpretive step remains weakly formalized. Most quantitative approaches either partition cells into discrete groups or assess how faithfully an embedding preserves local neighborhoods. Far fewer methods directly address a common question that motivates visual inspection: whether a biologically defined foreground population is non-uniformly distributed across the footprint of an embedded cell set, and if so, where that enrichment is concentrated within the 2D layout. Polar Gini Curves partially closed this gap by quantifying how uniformly gene-positive cells occupy a cluster footprint, thereby distinguishing broadly expressed markers from patchy signals (Nguyen et al., 2021). However, as embedding-centric analyses become more widely used, geometry-grounded claims are expected to be accompanied by explicit diagnostics for artifact susceptibility, sampling limitations, and

reproducibility rather than relying on qualitative inspection alone (Kobak and Berens, 2019; Gorin et al., 2022; Chari and Pachter, 2023; Johnson et al., 2022; Saquib et al., 2022; Hart and Tavolara, 2024).

Here, we introduce BioRSP (Biological Radar Scanning Plots), a geometry-first framework for quantifying angularly resolved radial enrichment of a foreground population in a two-dimensional representation. BioRSP operates on a user-defined analysis set (e.g., an annotated cell type or subtype) and a user-defined foreground rule (e.g., cells exceeding a feature-specific threshold or cells belonging to a condition of interest). Given fixed 2D coordinates, BioRSP selects a robust vantage point, transforms coordinates into polar form, and scans the analysis set by angle using sliding windows. Within each angular window, BioRSP compares the radial distributions of foreground and background cells using a signed, distance-based radial discrepancy, yielding a radar profile R(\theta) over the full angular range. We then reduce R(\theta) to a compact set of interpretable summaries, including an anisotropy magnitude that captures spatial structure strength, peak angles that localize extremal enrichment, and a coverage measure that reports how much of the profile is supported by adequate data.

BioRSP is neither a clustering nor a trajectory method. Instead, it treats published labels and user-defined subsets as strata for analysis and provides a standardized way to describe how a foreground is arranged within a given embedding footprint. Because embedding-aligned statistics can be underpowered or unstable when foreground support is sparse or uneven across angles, BioRSP includes explicit adequacy rules that mask under-supported angular sectors and can abstain from reporting scalar summaries when geometry is not identifiable. To support reproducible use, BioRSP additionally provides subsampling-based stability diagnostics that quantify the reproducibility of radar profiles and anisotropy summaries under controlled

resampling, and it records analysis parameters and random seeds in machine-readable output manifests.

In this manuscript, we detail the BioRSP estimand, summary statistics, adequacy rules, and robustness diagnostics, and we demonstrate the method in a community-standard human kidney single-nucleus reference by analyzing thick ascending limb (TAL) nuclei using published UMAP coordinates and a standardized within-set top-decile foreground rule. We further perform donor-aware reruns to illustrate how per-donor sample size and foreground support limit inference, and when donor-level profiles are stable once adequacy is met. Together, these components position BioRSP as a practical, embedding-aligned framework for characterizing enrichment patterns within fixed analysis sets while explicitly reporting when the available geometry does not support reliable interpretation.

2. Method

2.1. Inputs and notation

We analyze a dataset of N cells indexed by $i \in \{1, \ldots, N\}$. Each cell i has a two-dimensional coordinate $z_i \in \mathbb{R}^2$, representing a user-supplied embedding or 2D layout (e.g., UMAP, t-SNE, or a custom projection). We do not assume Euclidean distances in this embedding are globally faithful. Instead, the embedding is treated as a coordinate system for defining angular directions and radial distances relative to a reference point.

For each feature g (e.g., a gene, protein abundance, chromatin accessibility score, module score, or pathway activity score), let $x_i^{(g)} \geq 0$ denote the feature value in cell i after user-chosen

preprocessing/normalization. All definitions apply unchanged provided $x_i^{(g)}$ is numeric and comparable within the analyzed cell set.

BioRSP operates on a user-defined cell set $S \subseteq \{1, \ldots, N\}$ provided upstream (e.g., by filtering the input matrix and coordinates to the desired subset). The current implementation does not perform clustering or metadata-based subsetting; all quantities are computed on the provided cells.

When available, per-cell library size (total counts) is denoted by u_i. In the current implementation, u_i is used only for optional stratified permutation inference; it does not affect the radar estimand itself. If u_i is not provided, the CLI may fall back to using per-cell expression row-sums as a proxy, which is recommended only when those row-sums reflect true library sizes.

2.2. Vantage point and polar coordinates

A vantage point $v \in \mathbb{R}^2$ serves as the origin for radar scanning. By default, v is set to the geometric median of $\{z_i : i \in S\}$, which is robust to outliers and irregular embedding boundaries. The geometric median is computed via an iterative procedure with fixed convergence tolerance recorded in the run manifest.

For each cell $i \in S$, we define polar coordinates relative to v:

$$r_i = \|z_i - v\|_2, \qquad \theta_i = \operatorname{atan2}(z_{i,y} - v_y,\ z_{i,x} - v_x) \in [-\pi, \pi).$$

Let $\Theta = \{\theta^{(1)}, \ldots, \theta^{(B)}\}$ be an equally spaced grid on $[-\pi, \pi)$. Unless stated otherwise, we use $B = 360$. For each grid angle $\theta \in \Theta$, we define a sliding angular window (sector) of width Δ radians:

$$\mathcal{W}(\theta) = \left\{ i \in S : \mathrm{dist}_{\mathbb{S}^1}(\theta_i, \theta) \leq \frac{\Delta}{2} \right\},$$

where wrapped angular distance is

$$\mathrm{dist}_{\mathbb{S}^1}(\alpha, \beta) = \min_{k \in \mathbb{Z}} |\alpha - \beta + 2\pi k|.$$

All computations use radians internally; degrees are used only for reporting. The default window width is $\Delta = 180^\circ$.

2.3. Foreground definition

BioRSP compares the spatial distribution of "foreground" cells (high feature values) against "background" cells (the remainder). The current implementation uses a quantile-based binary foreground to stabilize estimation across features. For feature g within cell set S, we define

$$y_i^{(g)} = \mathbf{1}(x_i^{(g)} > t_g), \qquad t_g = Q_q(\{x_i^{(g)} : i \in S\}),$$

where Q_q is the q-quantile (default $q = 0.90$) and $\mathbf{1}(\cdot)$ is the indicator function. We use a strict inequality $x_i^{(g)} > t_g$ to reduce tie-related instabilities common in sparse single-cell data.

We report the realized foreground fraction

$$c_g^{(q)} = \frac{1}{|S|} \sum_{i \in S} y_i^{(g)},$$

which should be interpreted as the fraction of cells assigned to the upper tail under the chosen quantile rule (not the fraction of expressing cells). If $c_g^{(q)} = 0$ or if the total number of foreground cells is too small (adequacy rules below), the feature is treated as underpowered for stable directional estimation.

2.4. Radar Scanning Plot estimand

For a fixed feature g, cells in S are partitioned into a foreground set

$$F_g = \{i \in S : y_i^{(g)} = 1\},$$

and a background set $B_g = S \setminus F_g$.

Within each angular window $\mathcal{W}(\theta)$, we compare radial samples

$$\mathcal{R}_F(\theta) = \{r_i : i \in \mathcal{W}(\theta) \cap F_g\}, \qquad \mathcal{R}_B(\theta) = \{r_i : i \in \mathcal{W}(\theta) \cap B_g\}.$$

Let $W_1(\mathcal{R}_F, \mathcal{R}_B)$ denote the one-dimensional Wasserstein-1 (earth mover's) distance between the two empirical samples of radii (computed directly from the samples without binning).

BioRSP defines the signed, IQR-normalized radial discrepancy

$$R_g(\theta) = s(\theta)\, \frac{W_1(\mathcal{R}_F(\theta),\, \mathcal{R}_B(\theta))}{\max(\mathrm{IQR}(\mathcal{R}_B(\theta)),\, \tau_{\mathrm{IQR}})}.$$

Here IQR is the interquartile range, and τ_{IQR} is a global IQR floor used to stabilize sectors whose background radial dispersion is near-degenerate. In the implementation,

$$\tau_{\mathrm{IQR}} = \max(0.1 \cdot \mathrm{IQR}(\{r_i : i \in B_g\}),\ \varepsilon)\,, \qquad \varepsilon = 10^{-8}.$$

The sign term is

$$s(\theta) = \mathrm{sign}(\mathrm{median}(\mathcal{R}_B(\theta)) - \mathrm{median}(\mathcal{R}_F(\theta)))\,.$$

With this convention, $R_g(\theta) > 0$ indicates that in direction θ, foreground cells are more proximal (smaller radii; closer to the vantage point) than background cells, while $R_g(\theta) < 0$ indicates that foreground cells are more distal (larger radii; rim-enriched) than background cells. The software records whether τ_{IQR} was applied in each sector as a diagnostic.

The function $R_g(\theta)$ is evaluated on the grid Θ. Any optional smoothing is used for visualization only; all summaries and inference are computed from unsmoothed values.

2.5. Power and adequacy rules

Directional statistics are unreliable when computed from too few foreground or background cells. We therefore impose explicit adequacy criteria and treat under-supported sectors as missing. For each direction θ, define:

$$n_{\mathrm{fg}}(\theta) = |\mathcal{W}(\theta) \cap F_g|, \qquad n_{\mathrm{bg}}(\theta) = |\mathcal{W}(\theta) \cap B_g|.$$

Adequacy counts are computed using the same sliding window $\mathcal{W}(\theta)$ used by the radar statistic.

A sector is considered adequate only if

$$n_{\mathrm{fg}}(\theta) \geq n_{\mathrm{fg}}^{\min} \quad \text{and} \quad n_{\mathrm{bg}}(\theta) \geq n_{\mathrm{bg}}^{\min},$$

with defaults $n_{\mathrm{fg}}^{\min} = 10$ and $n_{\mathrm{bg}}^{\min} = 50$. If a sector fails these criteria, $R_g(\theta)$ is set to missing and excluded from scalar summaries.

In addition to per-sector thresholds, we apply a gene-level "coverage of adequate sectors" criterion. Let

$$\Theta_g = \{\theta \in \Theta : R_g(\theta) \text{ is defined}\}, \qquad f_g = \frac{|\Theta_g|}{|\Theta|}.$$

A feature is treated as adequate for downstream summarization and inference only if $f_g \geq f_{\min}$ (default $f_{\min} = 0.9$). We report f_g for every feature.

Finally, gene-level inference requires a minimum total number of foreground cells:

$$|F_g| \geq n_{\mathrm{fg,tot}}^{\min},$$

with default $n_{\mathrm{fg,tot}}^{\min} = 100$. Features failing this criterion are labeled underpowered and excluded from inference (they may still be visualized cautiously).

2.6. Scalar summaries

Each feature's radar profile is summarized using a primary anisotropy magnitude and interpretable peak directions computed over Θ_g.

1. **Quantile foreground fraction (reported).**

$$c_g^{(q)} = \frac{1}{|S|} \sum_{i \in S} y_i^{(g)}.$$

2. **Anisotropy magnitude (primary score).** Using only valid sectors Θ_g, the anisotropy score is

$$A_g = \left(\frac{1}{|\Theta_g|} \sum_{\theta \in \Theta_g} R_g(\theta)^2 \right)^{1/2}.$$

   If $|\Theta_g| = 0$, A_g is reported as missing.

3. **Peak directions and strengths.** Peaks are computed over Θ_g:

   a. Peak distal (rim-enriched):

$$\theta_{g,\mathrm{dist}}^* = \arg\min_{\theta \in \Theta_g} R_g(\theta), \qquad P_{g,\mathrm{dist}} = \min_{\theta \in \Theta_g} R_g(\theta).$$

   b. Peak proximal (core-enriched):

$$\theta_{g,\mathrm{prox}}^* = \arg\max_{\theta \in \Theta_g} R_g(\theta), \qquad P_{g,\mathrm{prox}} = \max_{\theta \in \Theta_g} R_g(\theta).$$

   c. Peak extremal (largest absolute magnitude):

$$\theta^*_{g,\mathrm{ext}} = \arg\max_{\theta\in\Theta_g} |R_g(\theta)|, \qquad P_{g,\mathrm{ext}} = R_g(\theta^*_{g,\mathrm{ext}}).$$

We also report $\max_\theta R_g(\theta)$, $\min_\theta R_g(\theta)$, and the integrated profile $\sum_{\theta\in\Theta_g} R_g(\theta)$ as diagnostics.

2.7. Feature typing into four classes

To support biomarker discovery and prioritization, BioRSP can subtype features into four qualitative categories based on how widespread the feature is (coverage) and how spatially structured it is (anisotropy). Because the radar statistic uses a quantile foreground by default (which compresses $c_g^{(q)}$ across features), we recommend defining two coverage notions:

1. **Radar foreground fraction $c_g^{(q)}$:** the quantile-based fraction used to compute $R_g(\theta)$ (reported by the software).
2. **Prevalence coverage $c_g^{(\mathrm{abs})}$:** a biologically interpretable prevalence metric computed from the raw feature values, e.g.

$$c_g^{(\mathrm{abs})} = \frac{1}{|S|}\sum_{i\in S} \mathbf{1}(x_i^{(g)} > t_{\mathrm{detect}}),$$

   where t_{detect} can be 0, a data-driven detection threshold, or a user-specified cutoff appropriate to the modality. This metric is recommended for "coverage typing" even when the radar uses quantile foreground internally.

Given a coverage metric c_g (typically $c_g^{(\mathrm{abs})}$ for interpretation) and anisotropy A_g, we define thresholds c_{hi} and A_{hi} (by default, median splits across adequately powered features within S, or user-defined cutoffs recorded in the manifest). Features are then assigned to one of four types:

1. **Type I: High coverage, high anisotropy ($c_g \geq c_{\mathrm{hi}}$ and $A_g \geq A_{\mathrm{hi}}$).** Widespread but strongly structured; often marks broad programs with spatial/niche bias.
2. **Type II: Low coverage, high anisotropy ($c_g < c_{\mathrm{hi}}$ and $A_g \geq A_{\mathrm{hi}}$).** Rare but strongly structured; high-priority candidate biomarkers.
3. **Type III: High coverage, low anisotropy ($c_g \geq c_{\mathrm{hi}}$ and $A_g < A_{\mathrm{hi}}$).** Broad and relatively uniform; often housekeeping-like or globally expressed programs.
4. **Type IV: Low coverage, low anisotropy ($c_g < c_{\mathrm{hi}}$ and $A_g < A_{\mathrm{hi}}$).** Rare and weakly structured; typically lower priority unless supported by other evidence.

In all cases, features must satisfy adequacy criteria ($f_g \geq f_{\min}$ and $|F_g| \geq n_{\mathrm{fg,tot}}^{\min}$) to be typed and compared reliably.

2.8. Pairwise feature relationships using radar profiles

To identify complementary or synergistic feature pairs, we compare their directional signatures $R_g(\theta)$ over shared valid directions. For two features g and h, define a common valid set $\Theta_{g,h} = \Theta_g \cap \Theta_h$. We recommend two complementary measures:

1. **Directional similarity (synergy):** correlation between radar profiles $\mathrm{corr}(R_g(\Theta_{g,h}), R_h(\Theta_{g,h}))$. Strong positive correlation suggests co-localization of directional bias.
2. **Directional complementarity:** correlation between R_g and $-R_h$, or angular separation between extremal peaks $|\theta^*_{g,\mathrm{ext}} - \theta^*_{h,\mathrm{ext}}|$ (wrapped), to detect opposing spatial biases (e.g., one distal where the other is proximal).

These pairwise analyses are performed downstream using the computed radar profiles and are recommended only for adequately powered features with stable robustness diagnostics.

2.9. Inference (optional permutation test)

BioRSP can assess whether the observed anisotropy exceeds what is expected if foreground membership is unrelated to spatial position in the embedding. Inference is performed on the anisotropy statistic A_g using permutation tests.

UMI-stratified permutation of foreground labels.

For each feature g, the embedding coordinates $\{z_i\}$ (and thus $\{r_i, \theta_i\}$) are held fixed. Cells are optionally partitioned into $Q = 10$ strata based on library size u_i (deciles within S), and the foreground labels $y_i^{(g)}$ are permuted independently within each stratum. If u_i is not provided, all cells are treated as a single stratum.

Fixed-direction RMS for permutations.

Let $\Theta_g \subseteq \Theta$ denote the set of directions where the observed radar profile is defined. To compute permutation statistics over a consistent set of directions, BioRSP uses the observed valid-direction mask Θ_g for all permutations. For permutation k, we compute a permuted radar profile $R_g^{(k)}(\theta)$ and define

$$A_g^{(k)} = \left(\frac{1}{|\Theta_g|} \sum_{\theta \in \Theta_g} \tilde{R}_g^{(k)}(\theta)^2 \right)^{1/2},$$

where $\tilde{R}_g^{(k)}(\theta) = R_g^{(k)}(\theta)$ if defined and $\tilde{R}_g^{(k)}(\theta) = 0$ if the permuted value is missing at a direction in Θ_g. If $|\Theta_g| = 0$, the p-value is reported as missing.

The empirical one-sided p-value is

$$p_g = \frac{1 + \sum_{k=1}^{K} \mathbf{1}(A_g^{(k)} \geq A_g)}{1 + K}.$$

By default, $K = 200$ permutations are used for exploratory runs (and can be increased for final reporting). False discovery rate control (e.g., Benjamini–Hochberg) can be applied downstream across the set of adequately powered features.

2.10. Robustness diagnostics (subsampling stability)

To assess sensitivity to sampling variation, we perform repeated subsampling of cells. For each feature, we draw $M = 20$ random subsamples (default 80% of cells without replacement), recompute the radar profile and scalar summaries, and quantify two stability metrics.

Profile reproducibility is the mean Pearson correlation between the subsampled and full radar profiles, computed over directions where both profiles are finite. If fewer than two directions are jointly defined, the correlation for that subsample is recorded as missing and excluded from the mean.

Anisotropy stability is the coefficient of variation of anisotropy across subsamples:

$$\mathrm{CV}(A_g) = \frac{\mathrm{SD}(A_g)}{\mathrm{mean}(A_g)}.$$

These diagnostics are reported alongside the primary summaries and are used to flag features whose directional signatures are unstable.

2.11. Default constants

Unless otherwise specified:

- Angular grid: $B = 360$.

- Sector width: $\Delta = 180^\circ$.
- Minimum per-sector support: $n_{\mathrm{fg}}^{\min} = 10, n_{\mathrm{bg}}^{\min} = 50$.
- Minimum adequate-sector fraction: $f_{\min} = 0.9$.
- Minimum total foreground for inference: $n_{\mathrm{fg,tot}}^{\min} = 100$.
- UMI strata: $Q = 10$.
- Numerical stabilization: $\varepsilon = 10^{-8}$.
- Permutations: $K = 200$ (exploratory; user-adjustable).
- Robustness subsampling: $M = 20$, subsample fraction $= 0.8$.

2.12. Runtime complexity and scaling

Let $n = |S|$ and G be the number of analyzed features. Polar coordinate computation is $O(n)$. For each feature, the current implementation evaluates each of the B window centers by computing wrapped angular distances against all cells, yielding $O(Bn)$ time to form sector masks and counts. Sector-wise Wasserstein distances are computed on radii within each window; overall runtime is dominated by the $O(Bn)$ masking step. Permutation inference scales as $O(KBn)$ per feature in the worst case, with underpowered features filtered by adequacy criteria prior to inference.

2.13. Determinism and reproducibility

All stochastic components (permutations and subsampling) are controlled by explicit random seeds recorded in output metadata. Each analysis produces machine-readable results (including adequacy diagnostics and summary statistics) and a run manifest containing parameters and seeds to support exact reproduction.

3. Case Study: Within–cell-type enrichment structure in the human kidney

We first demonstrate how BioRSP can be used to characterize spatial heterogeneity within a single annotated cell type in a large, community-standard single-nucleus reference dataset. We demonstrate that BioRSP complements existing marker-based analyses by systematically characterizing the distribution of high-expression cells across the embedding footprint of a fixed cell population.

We applied BioRSP to thick ascending limb (TAL) nuclei from the KPMP Human Kidney reference using the published 2D UMAP coordinates, normalized expression, TAL annotations, and donor identifiers, without reclustering or re-embedding. Using the default within-set quantile foreground rule (90th percentile; strict ">" threshold), the typical foreground size was ~3,839 nuclei (≈10% of the TAL set; total TAL nuclei $N \approx 38390$ inferred from the realized foreground counts).

Within the TAL cell set, we apply BioRSP gene by gene using a standardized foreground definition based on within-set expression quantiles. We first include a small panel of canonical TAL markers to serve as baseline controls, ensuring that broadly expressed markers exhibit low directional structure. We then apply BioRSP to a larger discovery panel of genes expressed in TAL cells, rank genes by the primary BioRSP anisotropy score, and retain only those genes that meet explicit adequacy and coverage criteria. This ranking step is used solely to identify representative examples of different within–cell-type enrichment phenotypes.

For statistical calibration, we compute permutation-based significance for selected genes using depth-aware shuffling within the TAL cell set, and we report corrected significance values

alongside BioRSP summary metrics. To assess robustness, we repeat the analysis under controlled perturbations, including subsampling of TAL nuclei and donor-aware reruns where donor sample sizes permit.

4. Results

4.1. BioRSP reveals a spectrum of within-TAL enrichment phenotypes

Across the pooled TAL set, BioRSP anisotropy scores spanned more than an order of magnitude (A_g range: 0.21-2.66), demonstrating that even within a single annotated cell type, "high-expression" cells for different genes can exhibit markedly different spatial structure within the embedding footprint.

Highly anisotropic, low-coverage focal pattern (candidate rare/focal marker behavior).

The strongest structure in this panel was observed for VCAM1 ($A_g = 2.66$), which also showed substantially reduced realized coverage under the quantile foreground definition ($n_{\text{fg}} = 689$; $\hat{c}_g \approx 1.8\%$), consistent with a sparse foreground caused by strong zero inflation/ties at low expression. VCAM1 exhibited an extreme directional excursion in the pooled TAL radar curve (most negative sector value $P_g = -5.64$), indicating a sharply localized rim-enrichment (distal) phenotype in a specific angular region of the TAL embedding.

Moderate anisotropy with typical coverage (structured within–cell-type heterogeneity).

Several genes with near-standard foreground sizes ($\hat{c}_g \approx 10\%$) nonetheless showed clear directional structure, including TPT1 ($A_g = 0.93, P_g = -1.91$), ITGB8 ($A_g = 0.76, P_g = -1.79$), PROM1 ($A_g = 0.62, P_g = -1.63$), CDH1 ($A_g = 0.57, P_g = -1.43$

), and SLC12A1 ($A_g = 0.32, P_g = -0.69$). In practical terms, these profiles indicate that the top-expression decile for these genes is not uniformly distributed over the TAL embedding footprint; instead, high-expression nuclei preferentially occupy particular angular sectors and (depending on sign) tend toward either the core or rim in those directions. This phenotype is the regime BioRSP is designed to capture for within–cell-type substructure discovery.

Core-enriched signatures with weak-to-moderate directionality (foreground generally more proximal).

A subset of genes showed radar profiles whose minima were positive, meaning the foreground was radially more proximal than background across essentially all directions in the pooled TAL set under the current sign convention. Notably, DEFB1 ($A_g = 0.89, P_g = 0.43$) and UMOD ($A_g = 0.81, P_g = 0.27$) fell in this category, consistent with a predominantly "core-enriched" phenotype rather than a rim-localized one.

Near-isotropic controls / weak structure.

At the low end of the score distribution, MT-CO1 ($A_g = 0.21, P_g = -0.61$), CD24 ($A_g = 0.26, P_g = -0.58$), and ATP1A1 ($A_g = 0.29, P_g = -0.74$) showed comparatively small anisotropy magnitudes, producing flatter radar curves consistent with weak directional structure under the top-decile foreground definition.

Together, these results illustrate that BioRSP can separate (i) sparse, highly localized patterns (VCAM1-like), (ii) broadly supported but directionally structured within-type patterns (PROM1/CDH1/ITGB8-like), and (iii) comparatively uniform patterns (MT-CO1/CD24/ATP1A1-like), all while operating strictly within a fixed TAL cell set and without relying on reclustering or trajectory inference.

Permutation calibration on selected genes

All reported genes achieved the minimum resolvable one-sided p-value under the stored permutation setting. This is consistent with evaluating a deliberately selected set enriched for pronounced structure and indicates that, at this permutation resolution, the observed anisotropy exceeded all permuted values for each tested gene.

4.2. Donor-aware robustness highlights power limits per donor, but stable profiles when adequate

To assess robustness to donor structure, we reran BioRSP within each donor where TAL nuclei were available (56 donors total) for a representative set of 12 genes. A key practical observation was that donor-level adequacy is often the limiting factor: for most genes, only ~7-10 of 56 donors (≈12-18%) met the gene-level adequacy criteria under the default thresholds. VCAM1 was not adequate in any single-donor run, consistent with its sparse foreground even in the pooled TAL set.

Importantly, when a donor passed adequacy, the directional profiles were typically well-supported across angles (mean adequacy fraction ≈ 0.997 among adequate donor runs; minimum observed ≈ 0.903), indicating that failures were primarily due to insufficient TAL counts/foreground counts rather than localized angular dropout. Directional agreement across adequate donors was gene-dependent: some genes showed moderate-to-strong angular concentration (e.g., SPP1 with concentration $R \approx 0.68$, implying relatively consistent peak direction), whereas others were more weakly aligned (e.g., UMOD, $R \approx 0.26$), reflecting either genuine donor variability or residual sampling noise at donor-specific cell counts. Overall, these donor-aware diagnostics support the pooled TAL analysis as the primary, well-powered

characterization, while also showing that donor-stratified confirmation is feasible but requires sufficient per-donor TAL nuclei (or relaxed thresholds / larger sector widths) to be informative.

5. Conclusion

BioRSP provides a practical, embedding-aligned statistic for quantifying enrichment structure within a fixed analysis set without reclustering, trajectory inference, or assumptions that Euclidean distances are globally meaningful in the embedding. By scanning a cell set in angular windows around a robust vantage point, BioRSP summarizes foreground-versus-background radial differences as a radar profile $R(\theta)$, and reduces that profile to interpretable outputs (anisotropy magnitude, peak directions, and coverage) accompanied by explicit adequacy and instability flags. This design directly targets a common gap in embedding-based workflows: investigators often observe gradients, asymmetries, and patchiness within an annotated population but lack a calibrated, standardized way to characterize these patterns feature-by-feature while transparently reporting when the geometry is not identifiable.

In a human kidney reference application focused on thick ascending limb (TAL) nuclei, BioRSP reveals that “high-expression” cells can occupy markedly different spatial configurations even within a single annotated cell type. Using a standardized within-set top-decile foreground rule on published UMAP coordinates, BioRSP separates sparse, highly localized rim-enrichment phenotypes (VCAM1-like) from broadly supported but directionally structured within-type patterns (PROM1/CDH1/ITGB8/SLC12A1-like), and from comparatively uniform, near-isotropic profiles (MT-CO1/CD24/ATP1A1-like). These results illustrate how BioRSP complements conventional marker-based summaries by making within-type heterogeneity explicit and comparable across features through a common estimand and reporting scheme.

Donor-aware reruns further clarify how BioRSP should be used in multi-donor atlases: donor-level adequacy is often the limiting factor, and many genes do not meet per-donor support thresholds under default settings, especially when the foreground is sparse. However, when donor runs are adequately powered, radar profiles are generally well-supported across angles, enabling donor-stratified confirmation of pooled signals in settings where sufficient per-donor cell counts exist. Practically, this indicates that pooled analyses are typically the primary, well-powered characterization, while donor-aware analyses are most informative as a robustness layer (and may require relaxed thresholds or larger sector widths when per-donor counts are modest).

BioRSP has clear boundaries that should guide interpretation. It does not assert that embedding geometry is a faithful physical space, nor does it infer developmental direction, temporal movement, or causal trajectories; instead, it provides a calibrated description of foreground distribution in the provided coordinate system. The default quantile-based foreground stabilizes comparisons across features, but prevalence-oriented interpretations should incorporate a separate, biologically meaningful detection/coverage metric when appropriate. Going forward, the same framework can be extended with additional foreground definitions (e.g., continuous weighting or covariate-adjusted thresholds), alternative coordinate systems, and richer cross-replicate models, while preserving the core principle that embedding-aligned claims should be accompanied by adequacy reporting and stability diagnostics rather than qualitative inspection alone.

6. Code and data availability

BioRSP is released as open-source software and is publicly available on GitHub at https://github.com/cytronicoder/biorsp. All analyses reported in this manuscript were executed from a fixed repository snapshot to ensure reproducibility.

The TAL case study uses the Azimuth Human Kidney reference, a KPMP/HuBMAP-derived single-nucleus RNA-seq reference distributed by the HuBMAP consortium. The reference is publicly available via Zenodo (DOI: 10.5281/zenodo.10694842). For analysis, the reference was converted from the distributed Seurat object format into an AnnData (.h5ad) representation using utilities provided in the BioRSP repository. The exact reference file used in this study, along with its checksum, is documented in the accompanying run manifest.

All BioRSP analyses generate a machine-readable run manifest that records the complete computational provenance of each execution. This manifest includes:

- Analysis parameters and foreground definitions
- Random seeds and execution timestamps
- Dataset sizes and gene filtering outcomes
- Donor composition and adequacy diagnostics (where applicable)
- Paths to all generated result tables and figures

The manifest enables exact reproduction of reported results and is intended to accompany any redistributed outputs. For the TAL case study, BioRSP produces standardized result artifacts, including ranked gene summary tables, angular enrichment statistics, and visualization outputs. These files, together with the run manifest and repository commit identifier, fully specify the analysis performed in this manuscript.

7. Author contributions

Z.Y. conceived the study, developed the methodological framework, implemented the software, curated and processed the datasets, performed the computational analyses, and drafted the manuscript. J.Y.C. designed the experimental and evaluation framework, supervised the project, and provided scientific guidance throughout. Both authors reviewed, edited, and approved the final manuscript.

8. Competing interests

The authors declare no competing interests.

9. Funding

This work was supported by the University of Alabama at Birmingham (UAB) research startup fund (to J.Y.C.) and by the National Institutes of Health (NIH) Common Fund award OD036472 (to J.Y.C.).

References

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*(1), 38–44. https://doi.org/10.1038/nbt.4314

Bergam, N., Snoeck, S., & Verma, N. (2025). t-SNE exaggerates clusters, provably. *ArXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2510.07746

Binnie, J. A. D., Dłotko, P., Harvey, J., Malinowski, J., & Yim, K. M. (2025). A Survey of Dimension Estimation Methods. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2507.13887

Biondo, M., Cirone, N., Valle, F., Lazzardi, S., Caselle, M., & Osella, M. (2025). The intrinsic dimension of gene expression during cell differentiation. *Nucleic Acids Research*, *53*(16). https://doi.org/10.1093/nar/gkaf805

Chari, T., & Pachter, L. (2023). The specious art of single-cell genomics. *PLOS Computational Biology*, *19*(8), e1011288–e1011288. https://doi.org/10.1371/journal.pcbi.1011288

Facco, E., d'Errico, M., Rodriguez, A., & Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-11873-y

Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A., & Schier, A. F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, *360*(6392). https://doi.org/10.1126/science.aar3131

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-13056-x

Kobak, D., & Linderman, G. C. (2021). Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, *39*(2), 156–157. https://doi.org/10.1038/s41587-020-00809-z

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. de, Cappuccio, A., & Corleone, G. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-1926-6

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/joss.00861

Narayan, A., Berger, B., & Cho, H. (2021). Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature Biotechnology*, *39*(6), 765–774. https://doi.org/10.1038/s41587-020-00801-7

Saelens, W., Cannoodt, R., Todorov, H., & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, *37*(5), 547–554. https://doi.org/10.1038/s41587-019-0071-9

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature Biotechnology*, *32*(4), 381–386. https://doi.org/10.1038/nbt.2859

Tritschler, S., Büttner, M., Fischer, D. S., Lange, M., Bergen, V., Lickert, H., & Theis, F. J. (2019). Concepts and limitations for learning developmental trajectories from single cell genomics. *Development*, *146*(12). https://doi.org/10.1242/dev.170506

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

Xia, L., Lee, C., & Jingyi Jessica Li. (2024). Statistical method scDEED for detecting dubious 2D single-cell embeddings and optimizing t-SNE and UMAP hyperparameters. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-024-45891-y

Yost, K. E., Satpathy, A. T., Wells, D. K., Qi, Y., Wang, C., Kageyama, R., McNamara, K. L., Granja, J. M., Sarin, K. Y., Brown, R. A., Gupta, R. K., Curtis, C., Bucktrout, S. L., Davis, M. M., Chang, A. L. S., & Chang, H. Y. (2019). Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nature Medicine*, *25*(8), 1251–1259. https://doi.org/10.1038/s41591-019-0522-3