Information Geometry Reconciles Discrete and Continuous Variation in Single-Cell and Spatial Transcriptomic Analysis

Jinpu Cai[1,2], Yuxuan Wang[2], Yunhao Qiao[2], Cheng Wang[2], Ziqi Rong[3], Luting Zhou[1], Haoyang Liu[2], Meng Jiang[4], Hongbin Shen[5], Jingyi Jessica Li[6,7,8⋆], and Hongyi Xin[2,9⋆⋆]

[1] Global College, Shanghai Jiao Tong University, Shanghai, China
[2] Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai, China
[3] Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA, USA
[4] Division of Cardiology, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China
[5] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China
[6] Biostatistics Program, Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA
[7] Department of Biostatistics, University of Washington, Seattle, WA, USA
[8] Department of Statistics and Data Science, University of California, Los Angeles, CA, USA
[9] School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China

Abstract. Single-cell and spatial transcriptomics provide high-resolution cellular characterization, yet standard analytical approaches remain theoretically misaligned with the probabilistic nature of the data. After UMI normalization, current pipelines rely on Euclidean or log-transformed Euclidean distance for similarity measurement. Both are fundamentally ill-suited to model the multinomial count data. Euclidean distance in normalized space overemphasizes high-variance genes, while log-transformation inverts this bias but at the cost of distorting subtle, continuous expression modulations. Neither approach naturally captures the dual nature of gene expression: both discrete presence/absence transitions and continuous quantitative variation. To overcome these limitations, we introduce GAIA (Geometric Analysis from an Information Aspect), an information-geometric framework for cell representation learning and inter-cell similarity measurement. By anchoring analysis in the true probabilistic model, treating cells as multinomial distributions over genes and projecting cells to a statistical manifold, GAIA organically reconciles both the presence/absence effect and the more continuous expression modulations. Mathematically, GAIA exploits the equivalence between Fisher-Rao distance in multinomial space and geodesic distance on the unit hypersphere, a property that enables both theoretical guarantees and computational efficiency. Experiments in synthetic and real scRNA-seq and spatial transcriptomic datasets demonstrate that GAIA preserves robust and consistent cell-to-cell relationships, delineates biologically nuanced sub-types, mitigates batch effects arising from sequencing depth variation, and eliminates the dependence on knowledge-restricted gene selection for learning meaningful cell representations. Overall, GAIA offers a knowledge-lean, variance-stabilizing framework for analyzing single-cell and spatial transcriptomic data, enhancing discrimination between nuanced cell sub-type and -states.

Keywords: Single-cell RNA-seq · Information geometry · Cell representation learning · Spatial transcriptomics · Qualitative · Quantitative.

⋆ Corresponding author. lijy03@fredhutch.org
⋆⋆ Corresponding author. hongyi.xin@sjtu.edu.cn

1 Introduction

Recent advances in single-cell and spatial multi-omics technologies have led to increasingly high-resolution characterization of cellular states, enabling the discovery of new cell types, trajectories, and spatial interactions[1]. However, most analytical frameworks still rely on heuristic similarity measures applied directly to normalized gene expression profiles[2, 3]. Typically, cell-to-cell similarities are computed using Euclidean or cosine distances, optionally over log-transformed data, implicitly assuming that the space of gene expression vectors is Euclidean[4]. This assumption is theoretically inconsistent with the underlying probabilistic nature of single-cell sequencing data, which arise from stochastic molecular sampling processes governed by multinomial or related count distributions[5].

Because the Euclidean metric is not derived from the generative model of transcript counts, the relative relationships between cells are highly situational and strongly dependent on the subset of genes used[6]. The geometry implied by these metrics can distort both local and global neighborhood structures in the low-dimensional latent space[7]. As a result, small changes in the selected gene set can lead to substantially different neighborhood structures. Such intricacy requires extensive knowledge and effort in gene selection, in order to produce a meaningful latent landscape. In practice, this often reverts single cell and spatial transcriptomic analysis into a trial-and-error process, increasing the labor effort, reducing the reliability of the analysis, and risking the confinement of the analysis within existing knowledge boundaries[8, 9].

From a probabilistic stand point, Euclidean similarity measurements are dominated by high-expressing genes. Because of the intrinsic stochasticity of the RNA capturing process, high-expressing genes exhibit greater absolute variation and contribute disproportionately to the total distance[10]. A common practical workaround is the log-transformation, which shifts emphasis from absolute expression differences to relative (fold-change) relationships[11]. However, this transformation disproportionately amplifies qualitative differences, such as genes switching from zero to positive expression, introducing strong discontinuities in the geometry of the data space. Consequently, analyses conducted in log-transformed Euclidean space tend to overemphasize presence/absence effects rather than capturing subtler quantitative modulations across genes (Fig 1a).

From a geometric perspective, the Euclidean distance between cells on the normalized-and-log-transformed manifold is not geodesic and is highly sensitive to sequencing depth variations (Fig 1b). After log-transformation, the simplex plane stretches into a convex curvature where any Euclidean straight line cuts underneath the surface. Consequently, the Euclidean straight line between two cells is not a geodesic and contains no valid states, as points along it violate the simplex constraint. Thus, the undersurface straight line between two cells on the log-transformed curvature does not represent a viable transition path and its associated Euclidean distance lacks biological meanings. Log transformation also complicates batch integration. When mRNA counts are randomly down-sampled, simulating a shift in sequencing depth, the inter-cell distance experiences an irregular transformation that co-depends on the sequencing depth and gene expression strength: as the relative mean expression a gene approaches zero after down-sampling and gravitating towards becoming a dropout gene, its influence over the inter-cell distance in the log-transformed space grows stronger, disproportionately expanding the dispersion of the cells in its dimension. This irregular distance transformation exacerbates the batch effects by sequencing depth variation and complicates the data integration procedure.

To overcome these limitations, we propose an alternative formulation grounded in *informational geometry*[12]. In our framework GAIA (Geometric Analysis from an Information Aspect), each cell is modeled as a multinomial distribution over genes, representing its normalized expression proportions. The space of such multinomial distributions forms a statistical manifold equipped with a natural Riemannian metric: the Fisher-Rao metric, which measures infinitesimal differences between probability distributions in terms of information distinguishability as Fig 1c,d shows. A key property of this metric is that, for multinomial distributions, under the square-root transformation of the normalized expression proportions, the Fisher-Rao distance between two cells reduces to the angular/arc distance on the positive (1st) quadrant of a unit hypersphere(Fig 1e)[13, 14]. This elegant metric provides both a principled probabilistic foundation and a geometrically interpretable similarity measure which is also computationally efficient.

The GAIA informational geometric representation is knowledge-lean and reconciles the quantitative and qualitative aspects of gene expression variation within a unified metric structure. Quantitative changes correspond to smooth displacements along the manifold, while qualitative transitions correspond to larger geodesic separations reflecting substantial information divergence. Unlike Euclidean distance on a simplex, which steers the measurement towards the quantitative changes of high-expressing genes, nor Euclidean distance on the log-transformed curvature, which is dominated by the qualitative differences

of the presence/absence effects, the arc distance on the unit hypersphere after square-root transformation strikes a good balance between preserving quantitative differences while moderately amplifying qualitative shifts (Fig 1a). Most importantly, a geodesic arc on the unit hypersphere define a valid transition path that minimize the integrated information difference between cell states. With the geodesic arc distance, GAIA prevents the underlying statistical manifold to be dominated by either ultra-high-expression housekeeping genes, as in the case of Euclidean distance on a simplex, nor ultra-high-dropout genes, as in the case of Euclidean distance on a log-transformed manifold. Thus, GAIA preserves robust and consistent cell-to-cell distance relations without requiring prudent gene selection, liberating cell representation learning from the tedious, potentially biased and domain-knowledge-intensive gene selection tuning. GAIA also helps mitigating batch effects that stem from sequencing depth variations. Unlike log-transformation, in GAIA, the dispersion growth from the sequencing-depth-induced conversion of a decent-expressing gene to a high-dropout gene is mild (Fig 1b). Thus, sequencing depth variations only introduce mild geometric distortions to the cell distribution in the statistical manifold, naturally improving alignment without sacrificing cell type delineation resolution. In addition to scRNA-seq, GAIA is also instrumental to spatial transcriptomics (ST) analysis. In ST, many genes that are qualitatively distinct at the single-cell level are weakened into merely quantitative or semi-quantitative differences, due to molecule diffusion or oversized spots. By harmonizing qualitative and quantitative differences, GAIA amplifies the nuanced transcriptomic shifts between spots, which is otherwise lost in log-transformation, and improves the spatial domain segmentation accuracy.

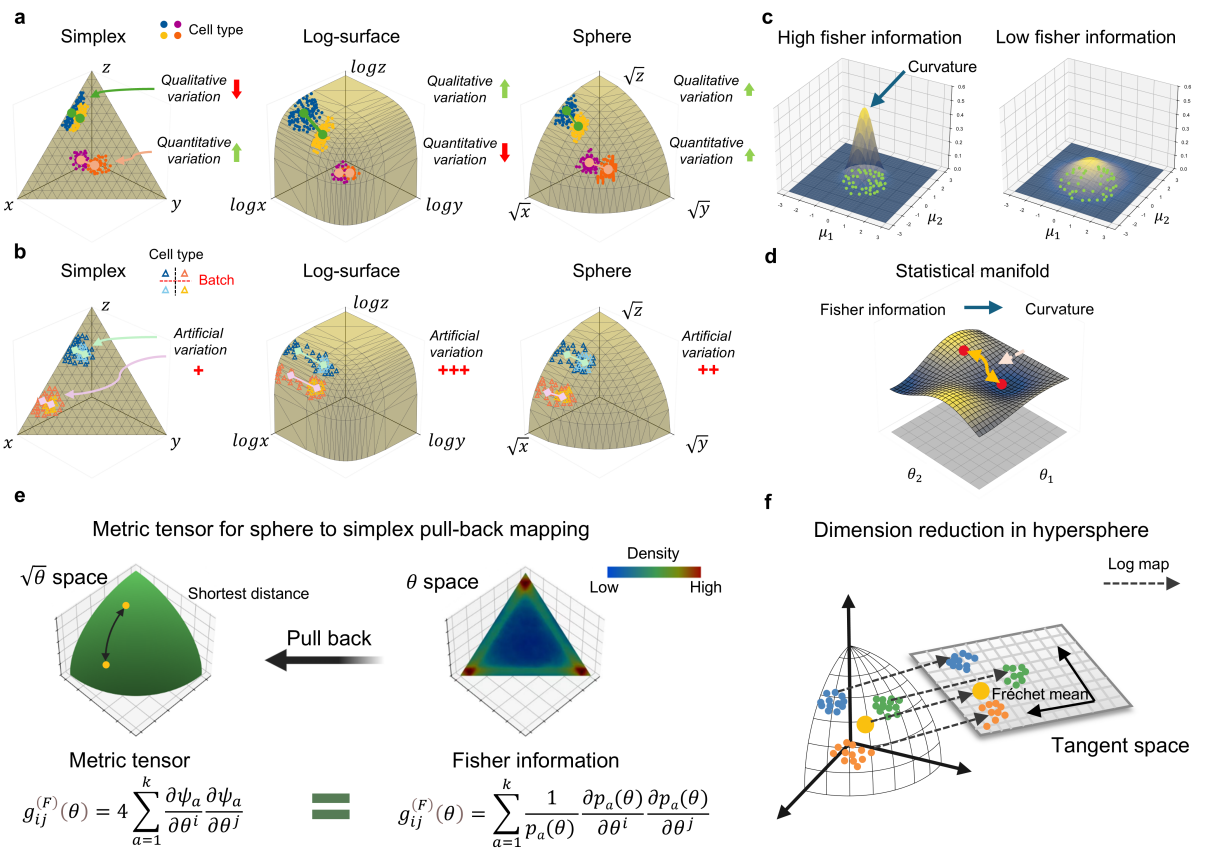

Fig. 1: (a) Geometry comparison: Simplex emphasizes quantitative differences, log-surface highlights qualitative variations, and the sphere balances both. (b) Batch effects: Log-normalization introduces artificial variation due to sequencing depth. (c) Fisher information: Measures likelihood curvature, with regions of high and low curvature. (d) Statistical manifold: Parameters form a manifold; distances correspond to geodesics. (e) Pull-back mapping: Under a multinomial model, Fisher information equals geodesic distance on a unit hypersphere after square-root transformation. (f) PCA reduction: Projects the hyperspherical geometry into lower-dimensional space.

2 Methods

2.1 Probabilistic Framework for ScRNA-seq Data

Single-cell RNA sequencing (scRNA-seq) captures gene expression by sampling a limited number of mRNA molecules from the transcriptome of each cell. Conceptually, each cell can be viewed as a random sample drawn from an underlying gene expression distribution that characterizes its transcriptional state. Thus, the observed count vector for cell i, denoted as $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iG})$, represents a finite sampling realization from this distribution. Previous work in [15] has established that single-cell RNA capture can be modeled as a hypergeometric sampling process, in which transcripts are sampled without replacement from the cellular transcriptome. Formally, for a total of N_i transcripts distributed among G genes, the probability of observing counts $\mathbf{x}_i$ given the true transcript abundances $\mathbf{p}_i = (p_{i1}, \ldots, p_{iG})$ is:

$$P(\mathbf{x}_i \mid \mathbf{p}_i, N) = \frac{\prod_{g=1}^{G} \binom{Np_{ig}}{x_{ig}}}{\binom{N}{n_i}}, \tag{1}$$

where $n_i = \sum_{g=1}^{G} x_{ig}$ denotes the total number of captured molecules in cell i.

When N is sufficiently large and the sampling fraction is small (i.e., $n_i \ll N$), the hypergeometric distribution can be well approximated by a **multinomial distribution** $\mathbf{x}_i \sim \text{Multinomial}(n_i, \mathbf{p}_i)$, with the probability density function as follows:

$$P(\mathbf{x}_i \mid \mathbf{p}_i, n_i) = \frac{n_i!}{\prod_{g=1}^{G} x_g!} \prod_{g=1}^{G} p_g^{x_g} \tag{2}$$

Under this formulation, each cell corresponds to a probability vector $\mathbf{p}_i$ lying on the parameter simplex:

$$\Delta^{G-1} = \left\{ \mathbf{p} \in \mathbb{R}^G \,\middle|\, p_g \geq 0, \sum_{g=1}^{G} p_g = 1 \right\}. \tag{3}$$

Therefore, the problem of measuring the difference between cells can be naturally reframed as defining a *distance metric between probability distributions* on the simplex, rather than directly comparing raw counts or normalized expression values.

2.2 Distance Measurement in the Gene Expression Space

A central task in single-cell analysis is quantifying similarity between cells. Traditionally, distances are computed directly on raw counts or normalized expression vectors. For example, the Euclidean distance between cells i and j is

$$d_{\text{Euclidean}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{g=1}^{G} (x_{ig} - x_{jg})^2}. \tag{4}$$

Such approaches, however, are sensitive to sparsity, high dimensionality, and technical noise.

By modeling each cell as a multinomial distribution with probabilities $\mathbf{p}_i$, the natural representation becomes

$$\mathbf{p}_i = \frac{\mathbf{x}_i}{n_i}, \quad n_i = \sum_{g=1}^{G} x_{ig}. \tag{5}$$

One might consider classical divergence measures between distributions, such as the Kullback-Leibler (KL) divergence

$$D_{\text{KL}}(\mathbf{p}_i \,\|\, \mathbf{p}_j) = \sum_{g=1}^{G} p_{ig} \log \frac{p_{ig}}{p_{jg}}, \tag{6}$$

or the Jensen-Shannon (JS) divergence

$$D_{\text{JS}}(\mathbf{p}_i, \mathbf{p}_j) = \frac{1}{2} D_{\text{KL}}(\mathbf{p}_i \,\|\, \mathbf{m}) + \frac{1}{2} D_{\text{KL}}(\mathbf{p}_j \,\|\, \mathbf{m}), \quad \mathbf{m} = \frac{\mathbf{p}_i + \mathbf{p}_j}{2}. \tag{7}$$

However, KL is asymmetric and unbounded[16], while JS, though symmetric, does not satisfy the triangle inequality, hence not a proper metric distance[17]. From a geometric perspective, neither serves a proper

distance metric for cell similarity measurement in the transcriptomic space. Another Wasserstein distance provides an interpretable notion of the "minimal transport cost" between two distributions, it has several limitations in the single-cell context: i) High computational cost, especially for high-dimensional gene spaces; ii) Sensitivity to the choice of cost matrix $c(g, h)$, which is often arbitrary for genes[18].

While classical distance measures and divergence metrics provide some notion of similarity between cells, they often fail to respect the intrinsic geometry of probability distributions, especially in high-dimensional, sparse single-cell RNA-seq data. This motivates the need for a framework that accounts for the manifold structure of gene expression probabilities, ensuring that distances reflect both local and global variations in the data.

2.3 Riemannian Metric Tensor, Fisher-Rao Metric and the Information Geometry

Quantifying distances between cells in single-cell transcriptomics requires a metric that respects the underlying geometry of gene expression distributions. Riemannian geometry, and in particular the Fisher-Rao metric from information geometry, offers a principled approach by defining geodesic-aware distances directly on the statistical manifold of a family of parameterized distributions[19].

A **Riemannian metric** tensor g_p assigns a positive-definite inner product on the tangent space $T_p\mathcal{M}$ of a smooth manifold $\mathcal{M}$ at a coordinate p in an Euclidean space:

$$g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}. \tag{8}$$

To understand the purpose of the Riemannian metric tensor, one can consider it as a translation between unit vectors from two coordinate systems: the real coordinate in the Euclidean space and a perceived coordinate on the Riemannian manifold. For a coordinate that is defined on the Riemannian manifold, at a given location, the metric tensor provides the means to translate the straight line distance traveled on the manifold, measured in the Riemannian manifold coordinate system, to the aggregated Euclidean distance of the corresponding curve in the real Euclidean space. The real Euclidean length of a tangent vector $v \in T_p\mathcal{M}$ on the Riemannian manifold, defined as the inner product of the vector, can be computed as:

$$\|v\|_p = \sqrt{g_p(v, v)}, \tag{9}$$

and the geodesic distance between two points $p, q \in \mathcal{M}$ on the non-Euclidean smooth manifold is the length of the shortest geodesic curve connecting them:

$$d(p, q) = \inf_{\gamma\,:\,\gamma(0)=p,\gamma(1)=q} \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\, dt. \tag{10}$$

For parametrized distributions, the Fisher information matrix serves as a natural metric tensor that defines a Riemannian geometry in the parameter space. Intuitively, Fisher information matrix, defined as:

$$[I(\theta)]_{i,j} = \mathbb{E}_{x\sim p(x;\theta)} \left[\left(\frac{\partial}{\partial\theta_i} \log p(x;\theta) \right) \left(\frac{\partial}{\partial\theta_j} \log p(x;\theta) \right) \middle| \theta \right] \tag{11}$$

quantifies the variance (if $i = j$) or covariance (if $i \neq j$) of the score function (derivative of the log-likelihood function with respect to a parameter θ_i or a parameter pair $[\theta_i, \theta_j]$). Fisher information is always positive-definite. Thus, it can be perceived as defining a curvature of the support curve representing the local geometry of a parametrized distribution in its parameter space. High Fisher information for a parameter(s) indicates that the curvature along the parameter(s) direction is sharp, therefore minute perturbation of the parameter(s) would significantly alter the distribution. Low Fisher information indicates the curvature is blunt and thus the nature of the distribution is less affected by changes in the respected parameter(s).

Given the natural equivalence of the Fisher information matrix and the Riemannian metric tensor, we can harness the Fisher information matrix as defining a mapping between a Riemannian geometry, in which the distributions locates on a Riemannian surface, and the Euclidean parameter space. The distance, known as the **Fisher-Rao distance**, between two distributions $p(x;\theta_1)$ and $p(x;\theta_2)$ with respective parameters can thus be computed as the geodesic distance of the shortest path on the Riemannian statistical manifold:

$$D_{\mathrm{FR}}(\theta_1, \theta_2) = \inf_{\gamma} \int_0^1 \sqrt{\dot\gamma(t)^\top I(\gamma(t))\dot\gamma(t)}\, dt, \tag{12}$$

where $\gamma(t)$ is a smooth path connecting θ_1 and θ_2. The Fisher-Rao distance is symmetric, invariant under reparametrization, and respects the underlying probabilistic structure, making it well-suited to quantify cell–cell similarity in single-cell RNA-seq data.

The Fisher-Rao distance quantifies how distinguishable two probability distributions are given their coordinates in the parameter space. Intuitively, with a high Fisher Information value defining a sharp curvature, distributions in a local pocket are more different in nature as they are less capable of explaining samples drawn from each other. These distributions are more distant on the Riemannian manifold as they have to overcome a large vertical distance on the statistical manifold despite their proximity in the parameter space. The Fisher-Rao distance, an integral of the shortest curve on the Riemannian manifold, therefore measures the aggregated difficulty of reshaping one distribution into another through gradually updating the parameters.

2.4 Square-root Transformation and the Simplex-to-Sphere Pull-back Mapping

The Fisher-Rao distance does not always produce a well-known, regular geometry (e.g., spherical, hyperbolic). Multinomial distribution, among a few others, is an exception where the information geometry is the first quadrant of a hypersphere. To show this, we need to demonstrate that the Fisher information matrix for multinomial distributions is equivalent (amplified by a constant factor) to the metric tensor for a pullback mapping from a simplex to a unit hypersphere, with the coordinates on the sphere being substituted by its squared variant, which is the coordinates on a hyper-simplex. The coordinate substitution could be comprehended as a pullback mapping of the metric tensor on a hyper-sphere space to the Euclidean hyper-simplex system, with square-root transformation.

Let $\boldsymbol{\theta}$ denote the n-dimensional Cartesian coordinates of points on a simplex, where $|\boldsymbol{\theta}|_1 = 1$. Let $\boldsymbol{\omega}$ denote the square-root transformed Cartesian coordinates, in which the simplex is transformed into a sphere with $|\boldsymbol{\omega}|_2 = 1$. For a point on the sphere, let the bases $\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_n}$ denote the orthogonal bases of the local tangent space. A vector $\boldsymbol{z}$ in the local tangent space, represented as $\boldsymbol{z} = \sum_{i=1}^{n} \beta_i \boldsymbol{e_i}$, can be converted to a representation in the $\boldsymbol{\omega}$ space as $\boldsymbol{z} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\omega_i}$ with chain rule, where $\alpha_i = \sum_{i=1}^{n} \beta_i \frac{\partial \boldsymbol{\omega_i}}{\partial \boldsymbol{e_i}}$. The inner product of a vector $\boldsymbol{z}$ in the $\boldsymbol{\omega}$ coordinate system can thus be translated to the local tangent space as following:

$$|\boldsymbol{z}|_2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \beta_i \beta_j \frac{\partial \boldsymbol{\omega_k}}{\partial \boldsymbol{e_i}} \frac{\partial \boldsymbol{\omega_k}}{\partial \boldsymbol{e_j}} \tag{13}$$

The above equation can be perceived as a translation of L-2 norm between the local tangent space and the $\boldsymbol{\omega}$ coordinate system, as $|\boldsymbol{z}|_2 = \boldsymbol{z}\top \boldsymbol{G}\boldsymbol{z}$, where $\boldsymbol{G}$ is the metric tensor defined as $G_{i,j} = \sum_{k=1}^{n} \frac{\partial \boldsymbol{\omega_k}}{\partial \boldsymbol{e_i}} \frac{\partial \boldsymbol{\omega_k}}{\partial \boldsymbol{e_j}}$. By substituting in $\omega_i = \sqrt{\theta_i}$ and applying chain rule in reverse ($\theta_i = \omega_i^2$), we have:

$$\boldsymbol{G}_{i,j} = \sum_{k=1}^{n} \frac{\partial \boldsymbol{\omega_k}}{\partial \boldsymbol{e_i}} \frac{\partial \boldsymbol{\omega_k}}{\partial \boldsymbol{e_j}} = \sum_{k=1}^{n} (\frac{1}{2\boldsymbol{\omega_k}} \frac{\partial \boldsymbol{\theta_k}}{\partial \boldsymbol{\omega_k}} \frac{\partial \boldsymbol{\omega_k}}{\partial \boldsymbol{e_i}})(\frac{1}{2\boldsymbol{\omega_k}} \frac{\partial \boldsymbol{\theta_k}}{\partial \boldsymbol{\omega_k}} \frac{\partial \boldsymbol{\omega_k}}{\partial \boldsymbol{e_j}}) = \sum_{k=1}^{n} \frac{1}{4\boldsymbol{\theta}_k} \frac{\partial \boldsymbol{\theta_k}}{\partial \boldsymbol{e_i}} \frac{\partial \boldsymbol{\theta_k}}{\partial \boldsymbol{e_j}} \tag{14}$$

given that $\frac{\partial \boldsymbol{\theta_k}}{\partial \boldsymbol{e_i}} = \frac{\partial \boldsymbol{\theta_k}}{\partial \boldsymbol{\omega_k}} \frac{\partial \boldsymbol{\omega_k}}{\partial \boldsymbol{e_i}} = 2\boldsymbol{\omega_k} \frac{\partial \boldsymbol{\omega_k}}{\partial \boldsymbol{e_i}}$. Equation 14 establishes the metric tensor between a simplex and its square-root-transformed unit hypersphere, where the geodesic distance on the equivalent hypersphere manifold between two points $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ on the hyper-simplex can be computed as:

$$D_{\mathrm{FR}}(\theta_1, \theta_2) = arccos(\omega_1 \cdot \omega_2) \tag{15}$$

with $\theta_1 = (\omega_1)^2$ and $\theta_2 = (\omega_2)^2$.

For multinomial distributions, the Fisher information matrix takes the same form as the metric tensor for the simplex-to-sphere pullback mapping. To see this, we use an alternative form of the Fisher information matrix

$$[I(\theta)]_{i,j} = -\mathbb{E}_{x\sim p(x;\theta)} \left[\left(\frac{\partial}{\partial \theta_i \partial \theta_j} \log p(x;\theta) \right) \middle| \theta \right] \tag{16}$$

By plugging Equation 2 into Equation 16 and apply chain rule, we have:

$$\begin{aligned} \frac{\partial}{\partial p_i \partial p_j} \log p(x;\boldsymbol{p}) &= \sum_{k=1}^{G} \frac{\partial}{\partial p_i} \left(\frac{\partial p_k}{\partial p_j} \cdot \left(\frac{\partial \log p(x;\boldsymbol{p})}{\partial p_k} \right) \right) = \sum_{k=1}^{G} \frac{\partial}{\partial p_i} \left(\frac{\partial p_k}{\partial p_j} \cdot \left(\frac{x_k}{p_k} \right) \right) \\ &= \sum_{k=1}^{G} \frac{\partial}{\partial p_i \partial p_j} \left(p_k \frac{x_k}{p_k} \right) - \sum_{k=1}^{G} \frac{\partial p_k}{\partial p_i} \cdot \frac{\partial p_k}{\partial p_j} \cdot \left(\frac{x_k}{p_k^2} \right) \end{aligned} \tag{17}$$

Take 17 into 16, we have:

$$[I(\theta)]_{i,j} = -\sum_{k=1}^{G} \frac{\partial \mathbb{E}[x_k]}{\partial p_i \partial p_j} + \sum_{k=1}^{G} \frac{\partial p_k}{\partial p_i} \cdot \frac{\partial p_k}{\partial p_j} \cdot \left(\frac{\mathbb{E}[x_k]}{p_k^2} \right). \tag{18}$$

Notice that $\mathbb{E}[x_k] = N \cdot p_k$ with N being the total RNA molecule count of a cell. Thus Equation 18 becomes

$$[I(\theta)]_{i,j} = \frac{\partial}{\partial p_i \partial p_j} \sum_{k=1}^{G} (x_k) + \sum_{k=1}^{G} \frac{\partial p_k}{\partial p_i} \cdot \frac{\partial p_k}{\partial p_j} \cdot \left(\frac{N}{p_k} \right) = N \sum_{k=1}^{G} \frac{1}{p_k} \frac{\partial p_k}{\partial p_i} \cdot \frac{\partial p_k}{\partial p_j}. \tag{19}$$

Notice that Equation 19 has the same form as Equation 14 but amplified by a factor of $4N$. This establishes the geometric equivalence between information distance and arc distance: The Fisher-Rao distance between two multinomial distributions is a constant multiple of the arc distance between $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_j$ on the hypersphere, after taking square-root transformation: $\boldsymbol{\omega}_i^2 = \boldsymbol{p}_i$ and $\boldsymbol{\omega}_j^2 = \boldsymbol{p}_j$.

2.5 Spherical Embedding Preserves Transcriptomic Heterogeneity in High Dimensions

An important property of the spherical embedding for multinomial distributions is that it makes feature engineering such as PCA less sensitive to gene selection. Specifically, the spherical embedding preserves the geometry that appears in a low-dimensional informative gene space, after expanding the gene space into higher dimensions by adding uninformative noisy genes. This property of being robust to noisy housekeeping genes is an direct corollary of an even stronger statement, which states that Fisher information, and the spherical embedding for the multinomial distributions, is isometry for every Markov embedding. This is known as the Chentsov theorem[20].

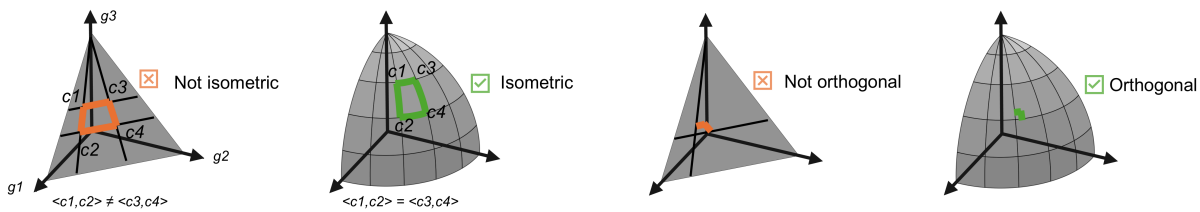

Fig. 2: Spherical embedding ensures isometry and orthogonalization of distances.

The significance of Chentsov theorem is that it establishes the theoretical basis for Fisher information metric being the most natural geometry to quantify the distance between distributions. It shows that Fisher information is the only isometric Riemannian geometry for all Markov embeddings. Retaining isometry is a crucial property for an effective dimension reduction regime. Assume mRNA molecules from G genes are partitioned into M biological functions (mRNA but not genes are partitioned). Let $\bar{p}_1, \ldots, \bar{p}_M$ denote the probability of drawing mRNA molecules from the respective biological functions. Let $q_{m,1}, \ldots, q_{m,G}$ with $\sum_i q_{m,i} = 1$ denote the probability of a mRNA molecule of a biological function m belonging to specific genes. Then the probability $p_1, \ldots p_G$ of drawing mRNA from specific genes, could be seen as a compound two-step multinomial sampling: first from $\bar{\boldsymbol{p}}$ and then from $\boldsymbol{q}$. We call $\boldsymbol{q}$ being the Markov embedding which maps one multinomial distribution function space of M dimensions ($\bar{\boldsymbol{p}}$) to another multinomial distribution function space of G dimensions ($\boldsymbol{p}$). Chentsov shows that Fisher-Rao distance is the only metric where the distance relations between distributions remain the same after any Markov embedding: for two pairs of distributions, $< \boldsymbol{p}_1, \boldsymbol{p}_2 >$ and $< \boldsymbol{p}_3, \boldsymbol{p}_4 >$ where $D_{\mathrm{FR}}(\boldsymbol{p}_1, \boldsymbol{p}_2) = D_{\mathrm{FR}}(\boldsymbol{p}_3, \boldsymbol{p}_4)$, then $D_{\mathrm{FR}}(\bar{\boldsymbol{p}}_1, \bar{\boldsymbol{p}}_2) = D_{\mathrm{FR}}(\bar{\boldsymbol{p}}_3, \bar{\boldsymbol{p}}_4)$ after transformation by any Markov embedding regime $\boldsymbol{q}$.

The Euclidean distance between the proportions of genes $\boldsymbol{p}$, for instance, is not isometry. Consider two pair of cells $< c1, c2 >$ and $< c3, c4 >$ in a 3-gene space. Let the ratio between $g1 : g2 = r1$ in both $c1$ and $c3$, and $g1 : g2 = r2$ in both $c2$ and $c4$. From the perspective of the dual-gene space $g1, g2$, $< c1, c2 >$ and $< c3, c4 >$ has the same Euclidean distance: $d_{[g1,g2]}(c1, c2) = d_{[g1,g2]}(c3, c4)$. However, once we consider the third gene $g3$, the distance equivalence is no longer maintained. Let $c1$, $c2$ has the same, high (but less than 50%) $g3$ proportion; while $c3$, $c4$ has equal, low $g3$ proportion. It is easy to see that $d_{[g1,g2,g3]}(c1, c2) < d_{[g1,g2,g3]}(c3, c4)$. This means that a constant ratio between $g1, g2$ is interpreted differently with respect to different expression strengths of an irrelative gene $g3$, creating an unnecessary

attachment between genes. The spherical embedding on the other hand, always maintain the distance equivalence between $< c1, c2 >$ and $< c3, c4 >$ in both 2-gene space and 3-gene space. Therefore the difference between $g1 : g2 = r1$ and $g1 : g2 = r2$ remains a constant irrespective of including other irrelevant genes (Fig 2).

The disentanglement between genes could also be explained from a geometric perspective. Assume there are two groups of cells in a 3-gene space. The first group $S1$ has the same $g3$ proportion but varies in $g1 : g2$ ratios. The second group $S2$ has a constant $g1 : g2$ ratio but varies in $g3$ proportions. In the Euclidean simplex, $S1$ and $S2$ are distributed along two straight lines but these two lines are only perpendicular if either cells in $S2$ have $g3 = 50\%$ or cells in $S1$ have $g1 : g2 = 1$. Otherwise, $S1$ and $S2$ are not truly orthogonal and $g3$ becomes entangled with $g1 : g2$ ratio. On the Riemannian sphere, in contrast, $S1$ and $S2$ are always on a pair of strictly perpendicular longitude and latitude. Thus, $g3$ is always locally orthogonal to $g1 : g2$ ratio shifts at their intersection (a latitude is not a straight line on the spherical surface but is locally orthogonal to any longitude). This geometric orthogonality is key to the establishment of a stable, gene-selection-invariant dimension reduction regime in the spherical manifold. Variance in irrelative uninformative genes will always be orthogonal to meaningful ratio shifts between biological functions and are expelled from the top low-dimensional components.

2.6 Harmonization of Qualitative and Quantitative Differences with Square-Root Transformation

Both the log transformation and the square-root transformation are special cases of the Box-Cox transformation family[21]. Formally, the Box-Cox transformation is a parametric family of power transformations:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y) & \text{if } \lambda = 0 \end{cases} \tag{20}$$

The log and the square-root transformations are special cases where $\lambda = 0$ and $\lambda = 0.5$ respectively. The Box-Cox transformation is designed to stabilize variance and induce approximate normality in non-normal data. After Box-Cox transformation, the variance is expected to be detached from the mean of the random variable, such that distance measurements would not be biased towards the highest-expressing genes. In Box-Cox transformation, λ characterizes the anticipated relationship between the mean and the variance of a random variable. For distributions where standard deviation scales proportionally to mean, such as members in the Gamma distributions family, log transformation is particularly effective where it converts multiplicative relationships into additive ones. For distributions where variance scales proportionally to mean, such as Poisson, square-root transformation is the most effective. Multinomial distributions, with its mean ($\mathbb{E}[X_i] = np_i$) and variance ($Var(X_i) = np_i(1 - p_i) = (1 - p_i)\mathbb{E}[X_i]$) being approximately linearly dependent, square-root transformation is therefore the more appropriate transformation.

Imprudent application of log transformation to multinomial observations could lead to a number of adverse effects: 1) amplification of low-count/boundary effects: When p_i is small, log transformation pulls values near zero close to $-\infty$, amplifying the variance of low-expressing genes and introduces distortions to distance measurements. 2) Loss of quantitative information: With an over-emphasis on the multiplicative structure, log transformation assigns equal distances between $1 \to 5$ and $100 \to 500$. A difference between $100 \to 200$, generally considered a significant shift in expression strength, pales in comparison to the minute perturbation of a gene growing from $1 \to 5$ due to sequencing randomness. 3) Asymmetric variance stabilization across the data range: For a large p_i, log transformation significantly compresses the scale, leading to a collapse of the variance. But for a small p_i, log transformation could drastically expand the scale, leading to an inverse inflation of the variances for small means. The co-dependence between mean and variance in log transformation is best illustrated by computing the asymptotic variance. Using the delta method by leveraging the first-order Taylor expansion, we have $Var[log(p_i)] \approx \frac{1-\pi_i}{n\pi_i^2}$, where $\pi_i = \mathbb{E}[X_i] = np_i$. As $\pi_i \to 0$, the variance explodes to ∞. In contrast, for square-root transformation, we have $Var[\sqrt{p_i} \approx \frac{1-\pi_i}{n}]$, where the variance is significantly more stable across the gene expression strength spectrum.

To systematically evaluate how data transformation affects the balance between qualitative and quantitative variation in single-cell data, we adopted the Box–Cox transformation as a unifying analytical framework. We simulated datasets containing both discrete and continuous expression differences among cell types and quantified their separation using class-center distances. To measure the overall preservation of biologically meaningful variation, we defined a harmonization score: the harmonic mean of qualitative

and quantitative distances, which penalizes transformations that overemphasize one type of variation at the expense of the other. This framework allowed us to identify the transformation that most effectively maintains a balanced representation of discrete and continuous transcriptional differences across the expression range. We found that as λ increases, the captured variation gradually shifts from qualitative to quantitative, and the harmonization score reaches its maximum at $\lambda = 0.5$, corresponding to the square-root transformation.

While square-root transformation is effective at stabilizing variance for multinomial distributions and does not completely diminish the intrinsic variances for high-expressing genes, it is worth noting that simply regarding the square-root transformed gene proportion space as a Euclidean space is still inappropriate. Euclidean distance, a natural metric for Normal distributions, is fundamentally incompatible with multinomial distributions, despite that square-root transformation reshapes multinomial distributions to be more Normal-like. Fundamentally, the Riemannian arc distance on the square-root-transformed hypersphere is the most natural metric for multinomial distributions, which doubles as the Fisher-Rao metric on the probability simplex.

2.7 Dimensionality Reduction on the Square-Root Transformed Sphere

After applying the square-root transformation to map multinomially-distributed single-cell data onto the unit sphere, standard linear dimensionality reduction methods such as PCA are no longer directly appropriate, since the data now reside on a curved manifold rather than in Euclidean space. To address this, we employ *Tangent PCA*[22], which respects the spherical geometry while providing an approximate low-dimensional embedding.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{S}^{G-1}$ denote the square-root transformed expression profiles of N cells on the unit sphere in G dimensions. The first step is to estimate the *Fréchet mean* $\bar{\mathbf{x}}$ of the points on the sphere:

$$\bar{\mathbf{x}} = \arg\min_{\mathbf{y} \in \mathbb{S}^{G-1}} \sum_{i=1}^{N} d_g^2(\mathbf{x}_i, \mathbf{y}), \tag{21}$$

where $d_g(\mathbf{x}_i, \mathbf{y}) = \arccos(\mathbf{x}_i^\top \mathbf{y})$ is the geodesic distance on the sphere. Next, each point is mapped to the tangent space at $\bar{\mathbf{x}}$ using the logarithm map:

$$\mathbf{z}_i = \log_{\bar{\mathbf{x}}}(\mathbf{x}_i) = \frac{\theta_i}{\sin\theta_i}(\mathbf{x}_i - \cos\theta_i\, \bar{\mathbf{x}}), \quad \theta_i = \arccos(\mathbf{x}_i^\top \bar{\mathbf{x}}), \tag{22}$$

yielding vectors $\mathbf{z}_i$ in the $(G-1)$-dimensional tangent space. Standard PCA can now be applied in this tangent space:

$$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N], \quad \mathbf{U}, \mathbf{\Lambda} = \mathrm{PCA}(\mathbf{Z}), \tag{23}$$

where $\mathbf{U}$ contains the principal directions and $\mathbf{\Lambda}$ the corresponding variances.

The tangent plane $T_{\bar{\mathbf{x}}}\mathbb{S}^{G-1}$ provides a local linearization of the sphere around the Fréchet mean $\bar{\mathbf{x}}$. For points $\mathbf{x}_i$ close to $\bar{\mathbf{x}}$, the Euclidean distance in the tangent plane approximates the geodesic distance on the sphere:

$$d_g(\mathbf{x}_i, \mathbf{x}_j) \approx \| \log_{\bar{\mathbf{x}}}(\mathbf{x}_i) - \log_{\bar{\mathbf{x}}}(\mathbf{x}_j) \|_2.$$

Consequently, performing PCA in the tangent plane yields a local linear approximation to spherical PCA, preserving the principal directions of variation while respecting the sphere's geometry.

3 Results

3.1 GAIA Robustly Identifies Stable B Cell Subtypes Across Feature Selections

We applied GAIA to the bone marrow mononuclear cell (BMMC) dataset from [23], and performed clustering using the Leiden algorithm on the GAIA embedding. Using the GAIA framework, we delineated four transcriptionally distinct B cell subtypes (Fig 3a), each characterized by coherent gene expression programs. In contrast, the four B cell sub-types become considerably inter-mixed in the normalization-only and the normalization-and-log-transformation frameworks. Marker genes for each cluster were identified via Wilcoxon rank-sum test performed on normalized count matrices. In GAIA-derived clusters, we observed distinct immunoglobulin expression programs. Specifically, two clusters were enriched for CD27, consistent with a memory B cell phenotype, while others showed differential enrichment of IGKC

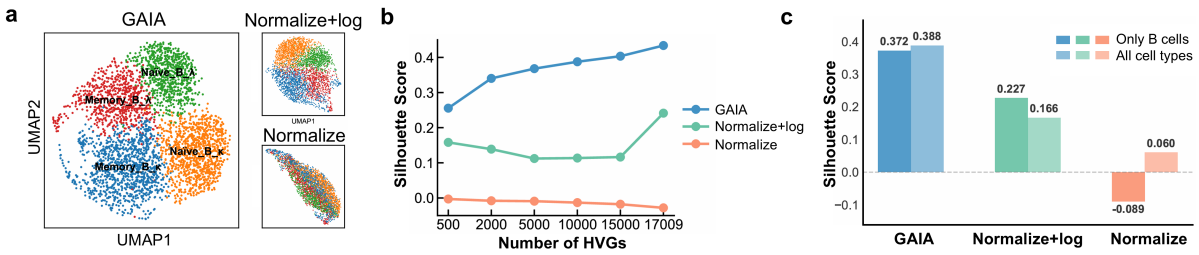

Fig. 3: (a) GAIA identifies four B-cell subtypes. (b) SI scores of GAIA, Normalize, and Normalize+log when selecting different numbers of top HVGs. (c) Comparison of SI scores for the four B-cell subtypes under GAIA, Normalize, and Normalize+log using the full dataset and the B-cell–only subset.

(κ light chain) and IGLC2 (λ light chain), reflecting mutually exclusive light chain usage [24]. These marker genes, together with IGHM and IGHD, are well-documented canonical markers distinguishing B cell maturation stages and functional states, as curated in the CellMarker database [25]. Notably, we observed that these marker genes exhibited both quantitative differences in expression levels and qualitative differences in presence/absence patterns across the four clusters (Appendix A).

To quantitatively assess the robustness and feature-independence of the GAIA classification, we computed the Silhouette score (SI) on the PCA embeddings obtained under each of the three frameworks, across multiple sets of highly variable genes (HVGs). Across all tested HVG configurations, the four GAIA-derived subtypes remained consistently well-separated, showing higher SI scores compared to normalization or normalization with log-transformation, as Fig 3b shows. This consistency indicates that GAIA captures intrinsic transcriptional geometry within the B cell population rather than being sensitive to arbitrary choices of input features. In other words, GAIA effectively extracts stable, geometry-preserving manifolds that are resilient to feature selection variations. Furthermore, when embedding B cells together with other major immune cell types, the same four GAIA-identified subtypes remained distinctly resolved within the global transcriptional landscape (Appendix A). The preservation of high SI scores (Fig 3c) in this broader context further demonstrates that GAIA achieves feature-independent, biologically reproducible delineation of B cell subtypes that generalizes beyond isolated clustering analyses. Collectively, these results establish GAIA as a robust and biologically grounded framework for subtyping immune cells, capable of maintaining consistent delineation across varying feature spaces and heterogeneous cellular environments.

As an additional validation, we applied GAIA to a human liver single-cell RNA-seq dataset (Appendix E) [26]. Consistent with the B cell analysis presented in the main text, GAIA achieved higher silhouette scores than normalization and normalization+log across different highly variable gene (HVG) selections. This supplementary analysis further supports the robustness of GAIA across datasets and feature choices. We also evaluated the computational efficiency of GAIA compared with standard normalization and normalization+log pipelines (Appendix F). As expected, GAIA incurs additional computational cost due to square-root mapping and tangent-space projection. In practice, the overall runtime scales linearly with the number of cells and genes and remains comparable to standard PCA-based workflows. Empirical benchmarks on datasets of comparable size indicate that GAIA incurs a moderate increase in preprocessing time, while remaining feasible for typical single-cell datasets. Detailed runtime and memory usage statistics are provided in Appendix F.

3.2 GAIA Enhances Spatial Domain Identification in Human DLPFC

Organic integration of qualitative and quantitative information is particularly important in spatial transcriptomics, where the diffusion of transcripts and the fusion of neighboring cells into spatial spots attenuate qualitative differences. We demonstrate the advantages of GAIA by benchmarking the domain segmentation performance of *scNiche* [27] on a human dorsolateral prefrontal cortex (DLPFC) dataset [28]. We compare annotations derived from GAIA-PCA, Euclidean PCA on normalized expression, and Euclidean PCA with normalization followed by log transformation. Fig 4 shows the domain segmentation results relative to manual annotation. Consistent with expectations, GAIA embeddings yield the most anatomically accurate domain delineations, which is reflected in a strong alignment with manually curated cortical layers and a high adjusted Rand index (ARI).

Notably, normalization followed by log transformation performs worst in domain segmentation. A primary contributor to its suboptimal performance is the attenuation of qualitative differences between

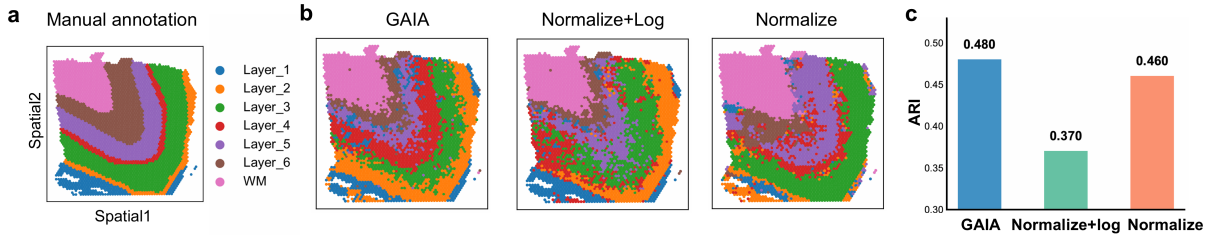

Fig. 4: Domain detection results of the mouse dorsolateral prefrontal cortex layers with manual annotations using GAIA, normalized, and normalized+log embeddings via scNiche, along with ARI scores.

cortical layers. Due to the limited spatial resolution of 10x Visium, each measured spot captures transcripts from multiple cells in a close vicinity. Transcriptomic differences between cortex layers manifest mostly as subtle quantitative shifts in average gene expression rather than clear presence/absence signals. Appendix B provides a detailed account of these effects. Together, these findings illustrate the benefit of harmonizing qualitative and quantitative variation with GAIA for spatial transcriptomics analyses.

3.3 GAIA Preserves Biological Structure Despite Sequencing Depth Variations

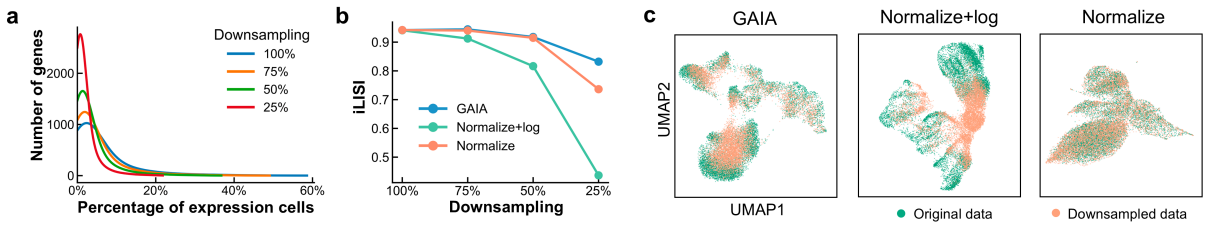

Fig. 5: (a) Density plot of gene-expression percentages across downsampling rates. (b) iLISI scores for GAIA, normalize + log, and normalize at varying downsampling levels. (c) UMAP of GAIA, normalize + log, and normalize across downsampling rates.

Sequencing-depth variation is a major source of batch effects in data integration. As sequencing depth decreases, moderately expressed genes progressively shift toward becoming high-dropout genes. When integrating datasets with drastically different depths, log-transformation amplifies inter-batch differences between low-dropout and high-dropout genes, creating spurious biological distinctions between batches.

To systematically evaluate the robustness of distance metrics against sequencing-depth variations, we benchmarked alignment of a public BMMC CITE-seq dataset against itself following random mRNA down-sampling [29]. Cell types were manually annotated based on surface marker expression [30]. We selected one BMMC dataset and randomly down-sampled mRNA molecules to 75%, 50%, and 25% of the original depth. Fig 5a displays the distribution of non-zero expression percentages across genes following down-sampling. As sequencing depth decreases, a substantial fraction of genes transition into high-dropout genes, accompanied by a corresponding decrease in non-zero expression cell percentages.

When a functionally important gene transitions from a non-dropout to a high-dropout gene across datasets, log-transformation dramatically amplifies inter-batch differences. iLISI[31] scores, which quantify cross-batch cell intermixing, reveal that alignment degrades more rapidly under log-transformation than under GAIA or normalization-only approaches (Fig 5b). This degradation is visually apparent in UMAP plots(Fig 5c), where batch effects are most pronounced in the log-transformation framework.

GAIA uniquely maintains biological distinctions between cell types while minimizing batch misalignment. Compared to normalization-only approaches, GAIA achieves superior cell-type separation, reflected in higher cell-type silhouette indices (Appendix D). This advantage is evident in the four-batch integration results, where four BMMC samples are integrated into a common transcriptomic space without explicit batch correction. As Appendix D demonstrates, GAIA outperforms alternative methods by simultaneously delineating cell types and reducing batch effects. Collectively, these results demonstrate the superior performance of the GAIA dimension reduction and distance metric in multi-batch analysis.

4 Conclusion

In this study, we present GAIA, a framework that overcomes the limitations of conventional Euclidean or log-transformed embeddings by modeling cells as multinomial distributions and computing distances within an informational-geometric framework. This approach effectively balances quantitative and qualitative variations in gene expression, preserves robust cell-to-cell relationships, mitigates the batch effects from sequencing depth variations, and reduces reliance on feature selection. Notably, GAIA is applicable to both scRNA-seq and spatial transcriptomics, enhancing subtle transcriptomic differences, improving domain-level annotations, and providing a principled, interpretable, and depth-robust framework for single-cell and spatial analyses. **GAIA** is available at https://github.com/Carroll105/GAIA.

5 Acknowledgements

This work is supported by STI2030-Major Projects 2022ZD0212400, Lingang Laboratory grant LGL-8888, STCSM grant 24510714300 and 20DZ2254400, GuangDong Basic and Applied Basic Research Foundation grant 2023B1515120006 and SJTU Science-Medicine interdisciplinary grant YG2026ZD09 and 24X010301456. We also thank Prof. Hongyu Zhao for helpful discussions and insightful suggestions.

References

1. Alev Baysoy, Zhiliang Bai, Rahul Satija, and Rong Fan. The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology*, 24(10):695–713, 2023.
2. Yuge Ji, Tessa D Green, Stefan Peidli, Mojtaba Bahrami, Meiqi Liu, Luke Zappia, Karin Hrovatin, Chris Sander, and Fabian J Theis. Optimal distance metrics for single-cell rna-seq populations. *BioRxiv*, pages 2023–12, 2023.
3. Ebony Rose Watson, Ariane Mora, Atefeh Taherian Fard, and Jessica Cara Mar. How does the structure of data impact cell–cell similarity? evaluating how structural properties influence the performance of proximity metrics in single cell rna-seq data. *Briefings in bioinformatics*, 23(6):bbac387, 2022.
4. Taiyun Kim, Irene Rui Chen, Yingxin Lin, Andy Yi-Yang Wang, Jean Yee Hwa Yang, and Pengyi Yang. Impact of similarity metrics on single-cell rna-seq data clustering. *Briefings in bioinformatics*, 20(6):2316–2326, 2019.
5. F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, 20(1):295, 2019.
6. Pengyi Yang, Hao Huang, and Chunlei Liu. Feature selection revisited in the single-cell era. *Genome Biology*, 22(1):321, 2021.
7. Shaoheng Liang, Vakul Mohanty, Jinzhuang Dou, Qi Miao, Yuefan Huang, Muharrem Müftüoğlu, Li Ding, Weiyi Peng, and Ken Chen. Single-cell manifold-preserving feature selection for detecting rare cell populations. *Nature computational science*, 1(5):374–384, 2021.
8. Lucy Xia, Christy Lee, and Jingyi Jessica Li. Statistical method scdeed for detecting dubious 2d single-cell embeddings and optimizing t-sne and umap hyperparameters. *Nature Communications*, 15(1):1753, 2024.
9. Ian Covert, Rohan Gala, Tim Wang, Karel Svoboda, Uygar Sümbül, and Su-In Lee. Predictive and robust gene selection for spatial transcriptomics. *Nature Communications*, 14(1):2091, 2023.
10. Ziqi Rong, Jinpu Cai, Jiahao Qiu, Pengcheng Xu, Lana X Garmire, Qiuyu Lian, and Hongyi Xin. L2 normalization and geodesic distance for enhanced information preservation in visualizing high-dimensional single-cell sequencing data. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–11, 2024.
11. Constantin Ahlmann-Eltze and Wolfgang Huber. Comparison of transformations for single-cell rna-seq data. *Nature Methods*, 20(5):665–672, 2023.
12. Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.
13. Henrique K Miyamoto, Fábio CC Meneghetti, Julianna Pinele, and Sueli IR Costa. On closed-form expressions for the fisher–rao distance. *Information Geometry*, 7(2):311–354, 2024.
14. Frank Nielsen. Fisher-rao and pullback hilbert cone distances on the multivariate gaussian manifold with applications to simplification and quantization of mixtures. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 488–504. PMLR, 2023.
15. Min Cheol Kim, Rachel Gate, David S Lee, Andrew Tolopko, Andrew Lu, Erin Gordon, Eric Shifrut, Pablo E Garcia-Nieto, Alexander Marson, Vasilis Ntranos, et al. Method of moments framework for differential expression analysis of single-cell rna sequencing data. *Cell*, 187(22):6393–6410, 2024.
16. Yufeng Zhang, Jialu Pan, Li Ken Li, Wanwei Liu, Zhenbang Chen, Xinwang Liu, and Ji Wang. On the properties of kullback-leibler divergence between multivariate gaussian distributions. *Advances in neural information processing systems*, 36:58152–58165, 2023.
17. Sreangsu Acharyya, Arindam Banerjee, and Daniel Boley. Bregman divergences and triangle inequality. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 476–484. SIAM, 2013.
18. Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431, 2019.
19. Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information geometry*, volume 64. Springer, 2017.
20. James G Dowty. Chentsov's theorem for exponential families. *Information geometry*, 1(1):117–135, 2018.
21. Jason Osborne. Improving your data transformations: Applying the box-cox transformation. *Practical Assessment, Research, and Evaluation*, 15(1), 2010.
22. P Thomas Fletcher, Conglin Lu, Stephen M Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005, 2004.
23. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.
24. Lisa M Rimsza, William A Day, Sarah McGinn, Anne Pedata, Yasodha Natkunam, Roger Warnke, James R Cook, Teresa Marafioti, and Thomas M Grogan. Kappa and lambda light chain mrna in situ hybridization compared to flow cytometry and immunohistochemistry in b cell lymphomas. *Diagnostic pathology*, 9(1):144, 2014.
25. Congxue Hu, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, et al. Cellmarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scrna-seq data. *Nucleic acids research*, 51(D1):D870–D876, 2023.

26. Benjamin J Stewart, John R Ferdinand, Matthew D Young, Thomas J Mitchell, Kevin W Loudon, Alexandra M Riding, Nathan Richoz, Gordon L Frazer, Joy UL Staniforth, Felipe A Vieira Braga, et al. Spatiotemporal immune zonation of the human kidney. *Science*, 365(6460):1461–1466, 2019.
27. Jingyang Qian, Xin Shao, Hudong Bao, Yin Fang, Wenbo Guo, Chengyu Li, Anyao Li, Hua Hua, and Xiaohui Fan. Identification and characterization of cell niches in tissue from spatial omics data at single-cell resolution. *Nature Communications*, 16(1):1693, 2025.
28. Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436, 2021.
29. Christopher Lance, Malte D Luecken, Daniel B Burkhardt, Robrecht Cannoodt, Pia Rautenstrauch, Anna Laddach, Aidyn Ubingazhibov, Zhi-Jie Cao, Kaiwen Deng, Sumeer Khan, et al. Multimodal single cell data integration challenge: results and lessons learned. *BioRxiv*, pages 2022–04, 2022.
30. Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.
31. Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.