Genome Biology

METHOD

Open Access

Descart: a method for detecting spatial chromatin accessibility patterns with inter-cellular correlations

Xiaoyang Chen[1†], Keyi Li[1†], Xiaoqing Wu[1], Zhen Li[1], Qun Jiang[1], Xuejian Cui[1], Zijing Gao[1], Yanhong Wu[1] and Rui Jiang[1*]

[†]Xiaoyang Chen and Keyi Li contributed equally to this work.

*Correspondence: ruijiang@tsinghua.edu.cn

[1] Ministry of Education Key Laboratory of Bioinformatics, Bioinformatics Division at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing 100084, China

Abstract

Spatial epigenomic technologies enable simultaneous capture of spatial location and chromatin accessibility of cells within tissue slices. Identifying peaks that display spatial variation and cellular heterogeneity is the key analytic task for characterizing the spatial chromatin accessibility landscape of complex tissues. Here, we propose an efficient and iterative model, Descart, for spatially variable peaks identification based on the graph of inter-cellular correlations. Through the comprehensive benchmarking, we demonstrate the superiority of Descart in revealing cellular heterogeneity and capturing tissue structure. Utilizing the graph of inter-cellular correlations, Descart shows its potential to denoise data, identify peak modules, and detect gene-peak interactions.

Keywords: Spatially variable peak, Spatial ATAC-seq, Feature selection, Inter-cellular correlations, Data imputation, Peak module, Gene-peak interactions

Background

Spatial molecular profiling enables the measurement of biomolecules within intact tissue sections, facilitating the construction of spatially resolved cell atlas [1, 2], analysis of cellular communication [3, 4], and exploration of the cancer tumor microenvironment [5]. Recent innovations in spatial sequencing technologies have integrated spatial barcoding schemes with assays for transposase-accessible chromatin using sequencing (ATAC-seq), allowing for the capture of spatial epigenetic information at the tissue level [6, 7]. Moreover, spatial multi-omics sequencing enables the detection of connections between chromatin accessibility and gene expression in the spatial context and provides valuable insights into spatiotemporal gene regulatory mechanisms of complex tissues [8, 9].

A key analytic task in spatial sequencing data is to identify spatially variable (SV) features that display spatial patterns of chromatin accessibility or gene expression [10]. Numerous methods [11–19] specifically developed for spatial RNA-seq (spRNA-seq)

data have achieved notable success in identifying SV genes [20, 21], while there is still a lack of methods tailored for modeling spatial ATAC-seq data [21] (spATAC-seq). Given the characteristic of higher sparsity in spATAC-seq data, the accessibility pattern of features (peaks in common scenarios) in a single slice is more discrete than spRNA-seq data, rendering the assumptions underlying spRNA-seq data-based methods inapplicable to spATAC-seq data. Furthermore, since the dimensionality of spATAC-seq data (the number of peaks) is an order of magnitude larger than the number of genes, methods designed for spRNA-seq data, which typically model, evaluate, and rank genes individually, are notably inefficient for spATAC-seq and require even several hours for computation. On the other hand, several methods for identifying peaks with high heterogeneity, as used for single-cell ATAC-seq (scATAC-seq), such as selecting peaks with the highest degree of accessibility (commonly used in scATAC data analysis) [22–26], a correlation-based method named Cofea [27], and specific functions provided in analytic pipelines [28, 29], ignore spatial information and thus cannot capture the spatial variations. Intuitively, the above two types of approaches fail to take full advantage of the intrinsic information from spatial distribution and data matrices, suggesting the pressing demand for methods to identify SV peaks. Besides this, other crucial analytic tasks, such as spatially peak module identification, gene-peak interaction detection, and data imputation, also lack the tailored modeling for spATAC-seq data.

To fill these gaps, we present Descart, a graph-based model, for DEtection of Spatial Chromatin Accessibility patteRns with inTer-cellular correlations. Leveraging the graph of inter-cellular correlations, Descart adeptly evaluates and identifies SV peaks by analyzing the self-correlations of peaks within the graph. To navigate the inherent challenge of highly dispersed accessibility patterns in spATAC-seq data, Descart incorporates chromatin accessibility information with spatial locations during graph construction and iteratively updates the graph to capture the intricate relationships between neighboring cells. Based on comprehensive benchmarking on 16 slices from 4 datasets, we demonstrate the superiority of Descart in identifying SV peaks that reveal cellular heterogeneity and tissue structure. Beyond its analytic advantages, Descart also surpasses other methods with spatial assumptions in computational efficiency. By leveraging neighboring relationships of the graph, Descart can impute data through signals from adjacent cells, thereby enhancing the accuracy of downstream analyses. Utilizing the inter-correlation of features within the graph, Descart enables the capture of inherent relationships between features: when applied to spATAC-seq data, Descart can obtain a peak-peak correlation matrix, facilitating peak module identification; when applied to spatial multi-omics data, Descart can produce a gene-peak correlation matrix, enabling the detection of gene-peak interaction and facilitating the discovery of gene regulatory networks.

Results

The Descart model

The Descart model aims to identify informative peaks that simultaneously characterize cell heterogeneity at cellular level and spatial continuity at tissue level (Fig. 1). Given a peak-by-spot matrix with spatial locations of spots (also can be replaced by cells), Descart evaluates and ranks peaks based on the graph of inter-cellular correlations, which are integrated from both spatial and chromatin accessibility information. More

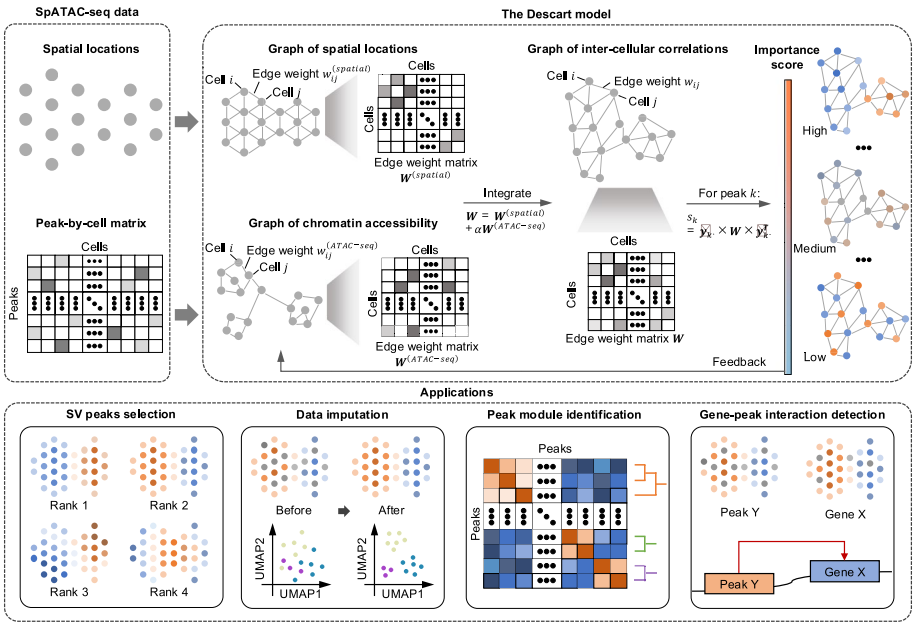

Fig. 1 The overview of Descart. Descart first constructs two distinct graphs based on spatial locations of spots and the peak-by-spot matrix, that is, the graph of spatial locations and the graph of chromatin accessibility. Next, Descart integrates the two graphs to derive the graph of inter-cellular correlations and utilizes the self-correlation of each peak within the graph to calculate the importance score. Based on these importance scores, Descart ranks all peaks and selects SV peaks. The SV peaks identified in each iteration are utilized to feedback and update the graph of chromatin accessibility, thereby refining the accuracy of neighborhood relationships among cells. Besides SV peaks identification, Descart can also be applied to data imputation, peak module identification, and detection of gene-peak interaction

specifically, the procedure of Descart can be divided in to five main steps: (i) constructing a spatial graph based on spatial locations of spots; (ii) performing principal component analysis (PCA) transformation on the peak-by-spot matrix with 50,000 (a default value that can be customized by users) selected peaks to obtain the latent embeddings of spots; (iii) constructing a graph of chromatin accessibility based on the latent embeddings and integrating the edge weight matrix of this graph with the edge weight matrix of the spatial graph to obtain a graph of inter-cellular correlations; (iv) utilizing the self-correlation of each peak within the graph to calculate the importance score; (v) evaluating and ranking all peaks based on importance scores and then feeding the current ranking back into step (ii). Descart iteratively performs steps (ii) through (v) until the obtained ranking of peaks undergoes minimal changes (4 iterations as default), and in step (ii) of the initial iteration, Descart directly selects peaks based on their decreasing order of accessible degree. The concept of modeling employed here is akin to Moran's *I* [30], a statistical measure frequently applied to spRNA-seq data [17–19]. Descart, however, tailors its modeling specifically to the characteristics of spATAC-seq data, integrating spatial information with chromatin accessibility data into a cohesive framework. The technical details of Descart are provided in the "Methods" section. When applying Descart, researchers could designate a specific number or a predetermined proportion of peaks as SV peaks, based on the ranking or importance scores of peaks. Leveraging the graph of inter-cellular correlations, Descart can accomplish data imputation. Besides, Descart can generate peak-peak similarity matrix based on the graph of inter-cellular

correlations and further utilize the similarity matrix for peak module identification. For spatial multi-omics data, such as the simultaneous capture of chromatin accessibility and gene expression information in a single slice, Descart can obtain a gene-peak similarity matrix in a similar manner, enabling the detection of gene-peak interactions.

Benchmarking performance of Descart using labeled spATAC-seq data

At the outset, we used the mouse brain dataset [8], which comprises four tissue slices with well-annotated domain labels (Additional file 2: Table S1), to assess the performance of Descart in SV peaks identification. Due to the absence of methods specifically designed for spATAC-seq data, Descart was benchmarked against two types of published methods: methods tailored for spRNA-seq data, including SOMDE [15], Moran's *I* [17], SpatialDE2 [11], SpatialDE [12], SPARK-X [13], SPARK [31], scGCO [14], Sepal [16], and methods designed for scATAC-seq data, including directly selecting peaks with high degree of accessibility (commonly used for scATAC-seq data analysis) [22–26], epiScanpy [28], Signac [29], and Cofea [27] (the "Methods" section). Drawing from scIB [32] and our previous work [27], we assessed different methods from two perspectives: the ability to facilitate clustering performance and capture domain-specific signals. The evaluation process is detailed in the "Methods" section. For cell clustering performance, we employed normalized mutual information (NMI), adjusted Rand index (ARI), and adjusted mutual information (AMI) scores as metrics. To evaluate the capture of domain-specific signals, we used the overlap proportion (OP) of domain-specific peaks with SV peaks as the metric, where OP1, OP2, and OP3 correspond to overlaps identified by the "tl.rank_features" function in epiScanpy, the "FindAllMarkers" function in Signac, and the "tl.diff_test" function in snapATAC2 [33], respectively. Higher scores in these metrics indicate better method performance. For a fair comparison, we tested each method by selecting 10,000 SV peaks and conducted all tests on a server with 128 GB of memory and equipped with 32 units of 13th Gen Intel(R) Core(TM) i9-13900 K to simulate typical personal computing conditions. The rationale for using a fixed number of 10,000 SV peaks for benchmark is twofold: disparate numbers of SV peaks identified by different methods complicate fair comparisons; 10,000, a common number in scATAC-seq or spATAC-seq analyses, prevent downstream clustering performance from reaching threshold levels that could bias comparisons (Additional file 1: Note S1 and Additional file 1: Fig. S1a). SPARK and SpatialDE2 encountered memory overflow errors during the process, and scGCO did not converge even after 24 h. Given that the mouse brain dataset represents the smallest scale within our collected data, we decided not to include these two methods in further comparisons. The benchmark results for other methods are depicted in Fig. 2a (with pre-processed results in Additional file 1: Fig. S2, S3, and S4). Descart not only gets the highest in overall scores but also excels in all metrics of cell clustering and uncovering domain-specific signals, indicating the superiority in identifying SV peaks rich in the capture of cellular heterogeneity and tissue structure. Furthermore, we provided clustering results of different methods when using SV peak counts of 3000 and 30,000, with Descart still performing exceptionally well as measured by the NMI metric (Additional file 1: Note S2 and Additional file 1: Fig. S1b). Besides, due to lacking the incorporation of spatial information, methods based on scATAC-seq data perform significantly worse in uncovering domain-specific signals, leading to lower

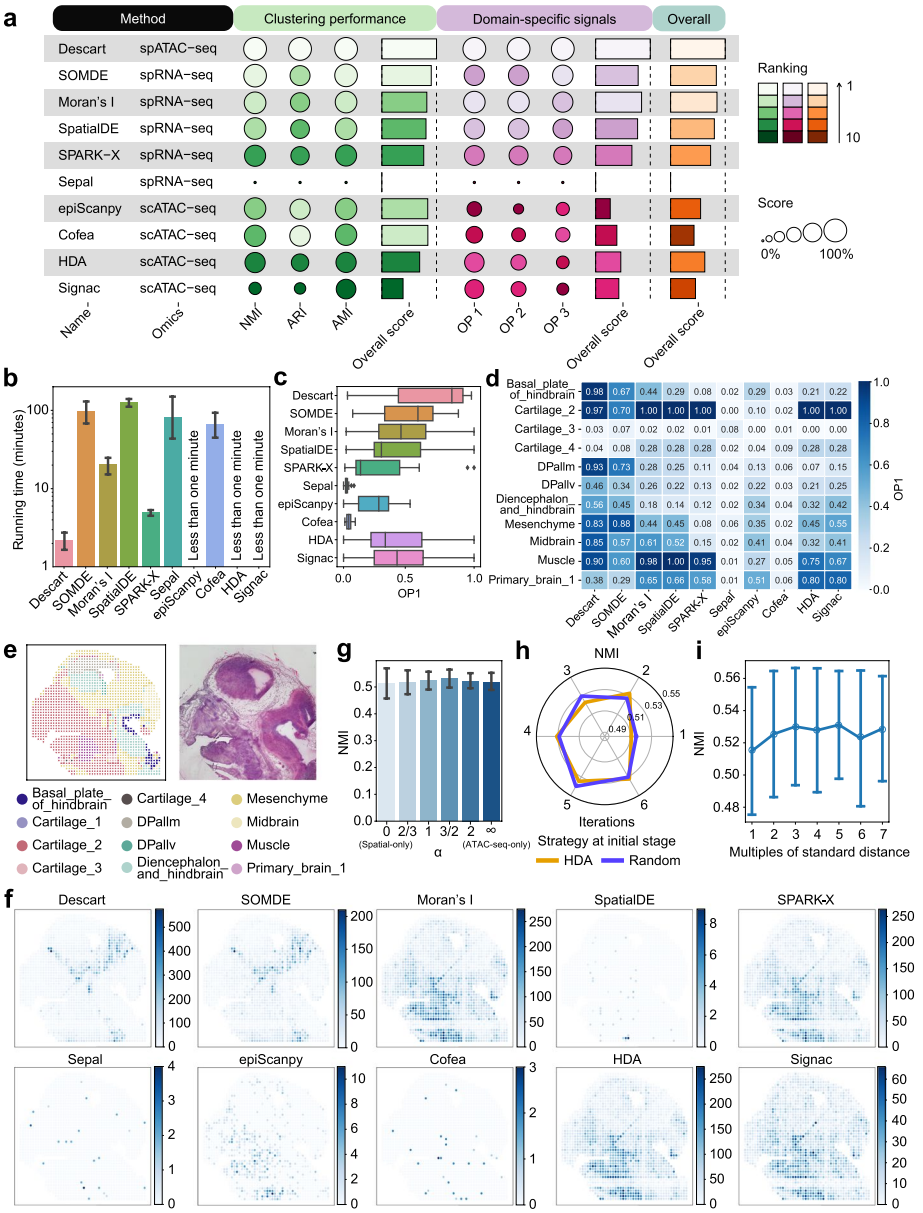

Fig. 2 Benchmarking performance of SV peaks identification on the mouse brain dataset. **a** Overview for benchmarking results of different methods from three perspectives, that is, the ability to facilitate clustering performance and capture domain-specific signals (see the "Methods" section for further visualization details). **b** Running time of different methods. **c**, **d** Overlapped proportion of SV peaks identified by Descart and baseline methods with domain-specific peaks related to overall domains (**c**) or each domain (**d**). Using the "tl.rank_features" function in episcanpy, we defined the top 100 peaks with the lowest *p*-values in each domain as domain-specific peaks. **e** Visualization of domains within the tissue space (left) and the corresponding histological image (right). **f** Top-ranked SV peak identified by each method on the E13_5-S1 slice, with the raw count values visualized in the tissue space. **g** Clustering performance using SV peaks identified by Descart and its variants. ATAC-seq-only and spatial-only represents the variants of Descart that only utilizes the graph of chromatin accessibility and the graph of spatial locations, respectively. **h**, **i** Clustering performance using SV peaks identified by Descart with a different number of iteration (**h**), different strategies for peak selection at the initial stage (**h**), and different multiples of standard distance (**i**). Clustering performance is evaluated by NMI scores. In **b**, **g**, and **i**, the error bars denote the 95% confidence interval, and the centers of the error bars denote the average value

overall scores. In terms of computational efficiency, except for HDA, epiScanpy, and Signac, which evaluate and rank peaks based on simplistic statistical information, Descart is over two times more efficient than the other methods (Fig. 2b). This is primarily because Descart is a transparent and intuitive method that does not rely on complex assumptions and involves fewer than 20 matrix operations in total (Additional file 1: Note S3). Thus, Descart demonstrates a significant advantage in both accuracy and computational efficiency in identifying SV peaks from spATAC-seq data. Furthermore, we have collected a dataset comprising four slices from various organs of humans and mice, serving as the mixed-species A dataset (Additional file 2: Table S1). The benchmarking procedure is generally consistent with that used in the mouse brain dataset. The primary distinction lies in the lack of explicit domain labels for these four slices; instead, we utilized reliable clustering labels provided by the original publication. Detailed results and analyses of this dataset are available in Additional file 1: Note S4, Additional file 1: Fig. S5, S6, S7, and S8. SV peaks identified by Descart significantly enhance downstream analysis performance, and the advantages of spatial methods are further amplified due to the labels being derived from spatial clustering algorithms.

Next, we conducted an in-depth analysis of the benchmark results, to delve into the intrinsic differences between various methods. We performed pairwise comparisons of the Spearman's correlation between the peak ranks evaluated by each method. As shown in Additional file 1: Fig. S9, spatial methods generally exhibit higher similarity with other spatial methods and lower similarity with non-spatial methods, and the opposite is also true for non-spatial methods. When evaluating the overlapped proportion of their top important peaks, we found that many spatial methods, such as SPARK-X, Moran's *I*, and SpatialDE, exhibit a high degree of overlap with HDA and Signac (Additional file 1: Fig. S10). This indicates that these spatial methods tend to select highly accessible peaks based on their underlying assumptions. In contrast, Descart shows a lower overlap with the peaks identified by other methods, but it consistently demonstrates the highest overlap with domain-specific signals across varying numbers of selected peaks (Additional file 1: Fig. S11). The detailed analyses are available in Additional file 1: Note S5. Taking the E13_5-S1 slice as a case study, we observed that the overall distribution of SV peaks identified by Descart shows a higher overlap with domain-specific peaks than baseline methods (Fig. 2c). When focusing on individual domains, we found that Descart also outperforms baseline methods in capturing cellular heterogeneity across various domains, including those with few samples, such as "DPallv" (13 spots) (Fig. 2d). Domains where Descart underperforms, such as "Cartilage_4" (2 spots) and "Primary_brain_1" (6 spots), typically suffered from an extremely low number of spots and the lack of spatial continuity, which may lead to excessive noise in identifying domain-specific peaks and diminish the informative value for evaluation. Similar trends could also be observed in other slices (Additional file 1: Fig. S3 and S4). We then visualized the top-ranked peak selected by each method, and compared it against the spatial coordinates of the domain and the corresponding histological image (Fig. 2e, f and Additional file 1: Fig. S12). The top-ranked peak identified by Descart and SOMDE closely corresponds to specific tissue regions, while SOMDE, which utilizes self-organizing maps, tends to select peaks accessible over larger areas. Moran's *I* and SPARK-X only capture accurate SV peaks in half of slices but still perform reasonably well. SpatialDE, following a

multivariate Gaussian assumption, might mistakenly identify peaks that are only accessible in adjacent spots as having strong spatial clustering, leading to erroneous SV peak identification. Sepal, on the other hand, due to its diffusion model, is not suitable for data with extremely low signal-to-noise ratios, which might lead to noise-rich peaks being identified as SV peaks. Essentially, this is because spATAC-seq data is sparser compared to spRNA-seq data, resulting in distributions of peaks that are unlikely in genes. Methods based on scATAC-seq data did not show clear spatial patterns in the visualization. Given that spATAC-seq data has deeper sequencing depth compared to scATAC-seq data, domain-specific peaks generally exhibit high accessibility, but Cofea tends to treat these as background peaks, leading to poor performance in related experiments. However, the significant peaks it identifies still retain rich heterogeneity, leading to satisfactory performance in downstream clustering.

In our final analysis, to dissect the key to superior performances of Descart, we conducted a series of ablation experiments and employed NMI from clustering performance as the evaluative standard. The essence of Descart lies in the unique integration of spatial and chromatin accessibility information in graph construction. The parameter α determines the ratio of spatial and chromatin accessibility information; by altering α values, we generated different variants of Descart and evaluated Descart against them. The most distinctive variants are as follows: one utilizing only spatial information (spatial-only) and another relying only on chromatin accessibility information (ATAC-seq-only). These variants echo strategies applying in prior spRNA-seq methods [18, 19], yet neither incorporates a fusion of these elements. As shown in Fig. 2g, the results demonstrate that Descart with α between 0 and 1, which integrates both spatial information and chromatin accessibility, enhances the accuracy of downstream applications more effectively than any single-element-focused variant. Notably, an α value of 1.5 (the default) yielded the best performance, as evidenced by the ablation experiments, thereby confirming the superiority of $\alpha = 1.5$ over other values. Intriguingly, the variant ATAC-seq-only outperforms that relying only on spatial locations of spots, highlighting the discrete nature of spATAC-seq data in space and the introduction of noise when only spatially accessible pattern of individual peaks is considered. Nonetheless, the graph constructed from spatial locations is indispensable, encapsulating structural information of tissues. We also compared the correlation of SV peaks identified by Descart with the two single-element-focused variants and found that Descart effectively integrates the results of these two variants, thereby enhancing the accuracy of identified SV peaks (Additional file 1: Note S6 and Additional file 1: Fig. S13). To mitigate noise, Descart constructs the spatial graph using neighbors of spots within five standard deviations, akin to applying a low-pass filter to the signal in space, which proved more effective than considering only a few neighbors around a spot (Fig. 2h). For efficient and precise graph construction based on ATAC-seq matrix information, an iterative strategy was adopted in the initial graph construction and updates peaks required for each iteration. As the results shown in Fig. 2i, we demonstrated that the strategy can enhance the accuracy of SV peak identification, and using the HDA peaks in the initial phase can accelerate convergence. By comparing the overlapped proportion of SV peaks obtained in each iteration by Descart, we found that the results stabilize after 4 iterations; thus, we set the default number of iterations to 4 (Additional file 1: Note S7 and Additional file 1: Fig. S14). Details on

ablation experiments to other parameters are available in Additional file 1: Note S8 and Additional file 1: Fig. S15. Overall, the key to advantages of Descart lies in how to construct the graph of inter-cellular correlations from both spatial and ATAC-seq perspectives, tailored to the specific characteristics of spATAC-seq data.

SV peaks identified by Descart align well with spatial structure of tissues

Besides the mouse brain dataset, we also collected two datasets: (i) a dataset comprising six slices of mouse embryos, served as the mouse embryo dataset, and (ii) a dataset consisting of five slices from a mix of human and mouse tissues, served as the mixed-species B dataset (Additional file 2: Table S1). The absence of well-defined domain labels in these datasets precludes us from utilizing the aforementioned benchmarking procedures for quantitatively assessing different methods. Here, we utilized the SV peaks identified by different methods to cluster spots and then referenced spatialPCA to evaluate the spatial clusters using three metrics: the spatial chaos score (CHAOS), the median low local inverse Simpson index (LISI), and the percentage of abnormal spots (PAS). Lower scores across these metrics signify more continuous spatial distribution of clusters, thereby indicating superior performance of the corresponding method. Compared to the mouse brain dataset, both two datasets are larger-scale and impose higher demands on the robustness of methods. When identifying SV peaks, SOMDE fails to produce results within 24 h on both datasets, while Sepal encountered errors on the mixed-species B dataset. As the results showed in Fig. 3a, b and Additional file 1: Fig. S16, lower metrics indicate the clustered domains using SV peaks of Descart are more spatially continuous and smooth, thereby demonstrating the advantages of Descart in detecting spatial structure of tissues. Taking the CHAOS metric as an example, Descart achieves the best results in 5 slices of the mouse embryo dataset and in 3 slices of the mixed-species B dataset. SpatialDE gets the second-best performance in both two datasets, but the running time requires at least 10 h and is hundreds of times longer than Descart, underscoring a significant efficiency gap (Fig. 3c, d). In these datasets, methods with spatial assumptions outperform those that do not consider spatial information, consistent with observations drawn from the mouse brain dataset. We then visualized the top-ranked peaks identified by different methods and compared them with histological images (Fig. 3e and Additional file 1: Fig. S17 and S18). The top-ranked peaks identified by methods with spatial assumptions are generally associated with specific regions within the tissue, corroborating the insights gained from the metrics. Notably, Descart, additionally incorporating chromatin accessibility information, identified top-ranked peaks that align well with spatial structure of tissues in all slices.

Descart captures cellular heterogeneity of metastatic melanoma

Next, we turned our attention to the performance of various methods on spATAC-seq data with single-cell resolution. We collected a metastatic melanoma dataset, where Russell et al. integrated their developed Slide-tags technique with scATAC-seq, facilitating simultaneous acquisition of chromatin accessibility and spatial information in individual cells [9] (Additional file 2: Table S1). In their study, Russell et al. also provided well-annotated labels of cell types, allowing us to conduct a benchmarking of Descart and baseline methods. The benchmarking is analogous to that applied to the mouse brain

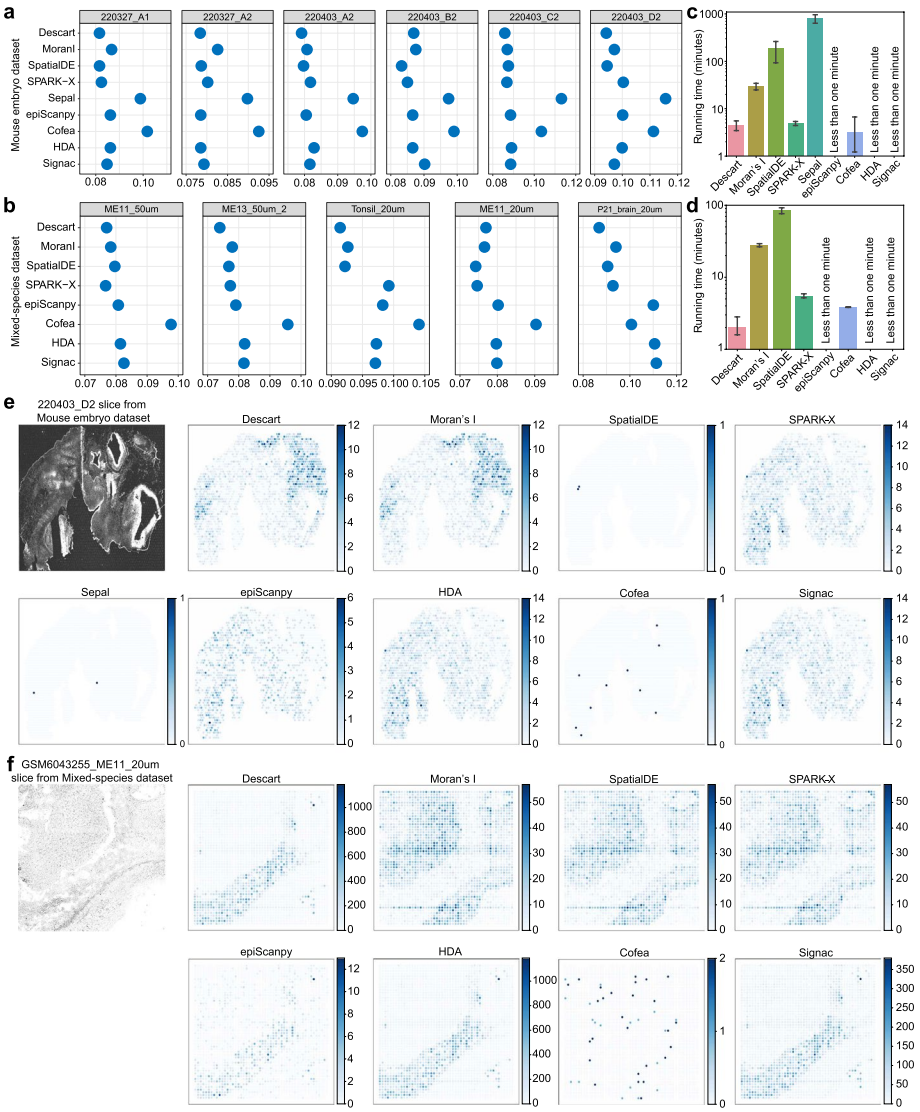

Fig. 3 Evaluation for different methods on the mouse embryo dataset and the mixed-species B dataset. **a**, **b** Clustering performance evaluated by CHAOS scores using SV peaks identified by different methods on the mouse embryo dataset (**a**) and the mixed-species B dataset (**c**), respectively. Due to the lack of well-annotated labels in the two datasets, we are unable to utilize label-dependent metrics for evaluation, such as NMI, ARI, and AMI scores. **c**, **d** Running time of different methods on the mouse embryo dataset (**b**) and the mixed-species B dataset (**d**), respectively. **e**, **f** The histological image (the first plot) and top-ranked SV peak identified by each method on the 220403_D2 slice from the mouse embryo dataset and the GSM6043255_ME11_20um slice from the mixed-species B dataset, with the raw count values visualized in the tissue space

dataset, with the distinction that cell type labels were employed instead of domain labels. Sepal reported errors on this dataset, while SOMDE exceeded a 24-h computation time. Benchmarking results for other methods, as illustrated in Fig. 4a (with pre-processed results in Additional file 1: Fig. S19), demonstrate that Descart excels in facilitating cell clustering and revealing cell type-specific signals, thereby affirming its superiority in identifying SV peaks. Given the dataset encompassed a total of 53,431 peaks, closely aligning with our predetermined number of SV peaks, the cluster performance across

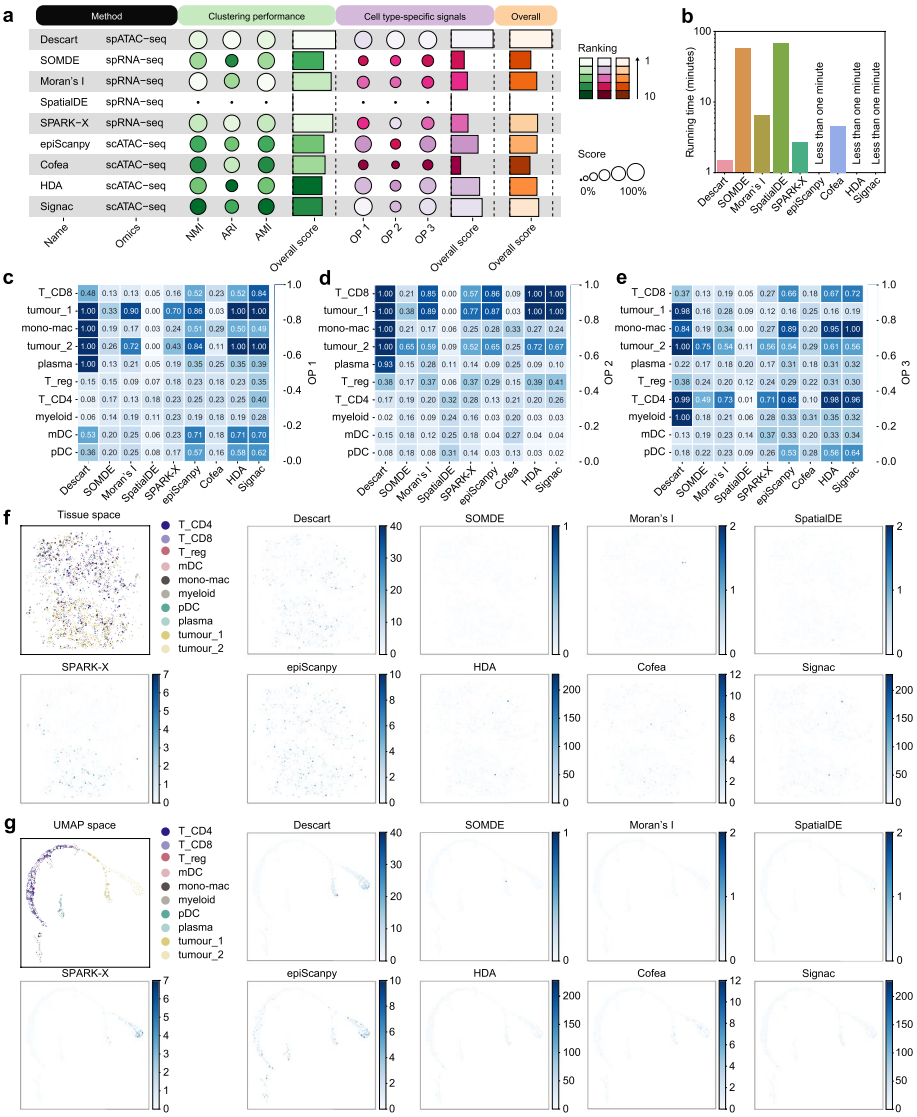

Fig. 4 Benchmarking performance of SV peaks identification on the metastatic melanoma dataset. **a** Overview for benchmarking results of different methods from three perspectives, that is, the ability to facilitate clustering performance and capture cell type-specific signals (see the "Methods" section for further visualization details). **b** Running time of different methods. **c**–**e** Overlapped proportion of SV peaks identified by Descart and baseline methods with domain-specific peaks identified by the "tl.rank_features" function in episcanpy (**c**), "FindAllMarkers" function in Signac (**d**), and the "tl.diff_test" function in snapATAC2 (**e**). **f** The top-ranked SV peak identified by each method in the tissue space, compared to histological image (the first subplot). **g** Top-ranked SV peak identified by each method, compared with cell type labels (the first subplot), in the UMAP space

different methods is relatively consistent. Note that, with the exception of two tumor subtypes ("tumour_1" and "tumour_2"), the spatial distribution patterns of various cell types are not significantly distinct, allowing us to assess the robustness of different methods when sequencing sample spatial distribution contains significant noise. Thus, methods with spatial assumptions do not show the pronounced advantages as in the mouse brain dataset. Within this dataset, Descart also demonstrates its notable efficiency on running time over other methods with spatial assumptions (Fig. 4b). In terms

of cell type-specific signals, methods that show superior performance, such as Descart, Moran's *I*, SPARK-X, epiScanpy, HDA, and Signac, primarily focus on cell types with a higher cell count (cell types are ordered by cell count from top to bottom in Fig. 4c–e). Notably, Descart can focus on a broader range of cell types, contributing to its distinguished clustering performance. We further visualized cells, on the tissue space using actual spatial locations and the latent space using UMAP on the scATAC-seq data matrix, respectively (Fig. 4f, g). The two spaces correspond to the spatial and chromatin accessibility information of cells, serving as the foundational elements for constructing the graph in Descart. The majority of methods do not show discernible patterns among top-ranked peaks in the two spaces. In contrast, top-ranked peak identified by Descart is associated with two tumor subtypes, and SPARK-X with one, marking them as relatively superior methods.

Descart imputes data using the graph of inter-cellular correlations

SpATAC-seq data typically suffer from noise and a large number of missing values, leading to the inaccuracy of downstream analysis. A key feature of Descart is its ability to denoise data and restore missing values, by utilizing the graph-based neighbor relationships between spots. The data imputation procedure can be categorized into four cases (details in the "Methods" section): (i) case 1: based on the graph of spatial locations; (ii) case 2: based on the graph of chromatin accessibility; (iii) case 3: based on the graph of inter-cellular correlations, that is, the integration of case 1 and case 2; (iv) case 4: augmenting case 3 with raw data. Taking the E13_5-S1 slice of the mouse brain dataset as an example, we applied Descart on to select 10,000 SV peaks and then performed cell clustering and uniform manifold approximation and projection (UMAP) visualization on the data before and after imputation. As shown in Fig. 5a, b, except for case 2, all other cases improve the accuracy of clustering, and different domains are better separated in the low-dimensional space after imputation. Using only spatial information, except for marginally enhanced spatial continuity (CHAOS), case 2 does not surpass the results using the original data in other metrics. Using only chromatin accessibility information (case 1) significantly improves clustering results but slightly disrupts spatial continuity (LISI). In contrast, the fusion of two types of information (cases 3 and 4) leads to better performance than using either type of information alone, suggesting that indispensability of both spatial and chromatin accessibility information when constructing the graph of inter-cellular correlations.

Inspired by SCALE, we next focused on Descart's effectiveness in imputing signals across different domains. We aggregated signals within spots of the same domain to serve as a meta-spot (ground truth) and then calculated the Pearson correlation coefficients between each spot's signal and the meta-spot before and after data imputation. Higher correlation coefficient indicates greater accuracy of the imputation results. As shown in Fig. 5c, except for domains "Cartilage_3" (1 spot) and "Cartilage_4" (2 spots) containing few spots, the correlation coefficients between post-imputation signals and meta-spot signals are higher across other domains, suggesting the superior performance of Descart in data imputation. Notably, cases 3 and 4, which involved using an inter-cellular correlation graph that combines spatial and chromatin accessibility information for imputation, show superior performance, aligning with conclusions derived

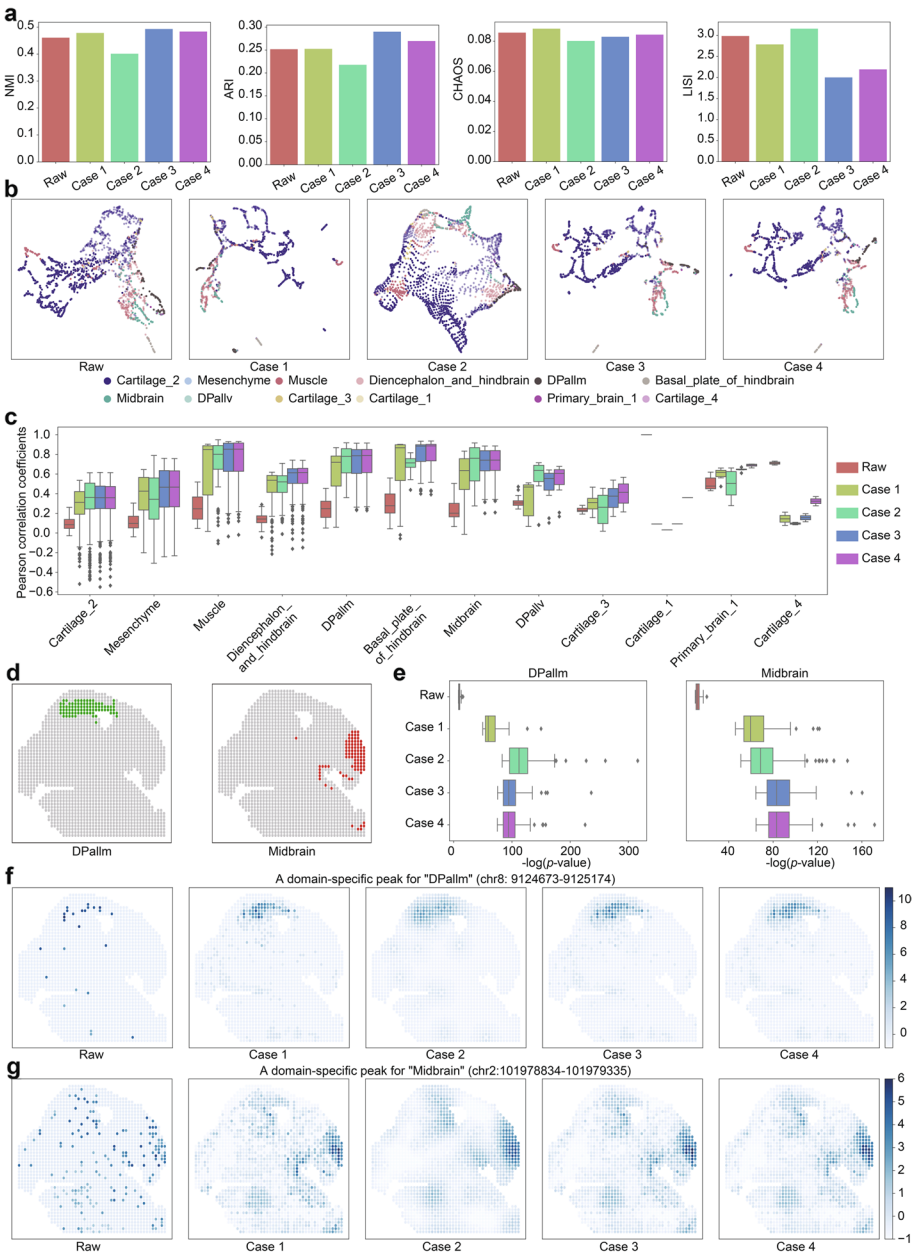

Fig. 5 Descart enables data imputation using the graph of inter-cellular correlations. **a** Evaluation of clustering performance on the E13_5-S1 slice from the mouse brain dataset, assessed using NMI, ARI, CHAOS, and LISI scores. **b** Visualization of spots in the UMAP space. **c** Pearson correlation coefficients between each spot's signal and the corresponding meta-spot, comparing results before and after data imputation. The central line of the boxplot represents median correlation coefficients of each spot, with the whiskers indicating the upper and lower quartiles. **d** Spatial locations of the domains "DPallm" and "Midbrain" in the tissue space. **e** Statistical significance of domain-specific peaks for "DPallm" and "Midbrain," evaluated through *p*-values generated by the "tl.rank_features" function in epiScanpy. **f**, **g** Visualization of domain-specific peaks for "DPallm" (chr8: 9,124,673–9,125,174) (**e**) and "Midbrain" (chr8: 9,124,673–9,125,174) (**f**), in tissue space. In **a**, **b**, **c**, **e**, **f**, and **g**, the comparison between raw data and data imputed by Descart is showcased. Cases 1 to 4 denote as different imputation strategies implemented in Descart (details in the "Methods" section): (i) case 1: based on the graph of spatial locations; (ii) case 2: based on the graph of chromatin accessibility; (iii) case 3: based on the graph of inter-cellular correlations, that is, the integration of case 1 and case 2; (iv) case 4: augmenting case 3 with raw data

from clustering metrics. For domain-specific signals, taking "DPallm" and "Midbrain" domains (the spatial locations are illustrated in Fig. 5d) as examples, we found that p -values (outputted by the "tl.rank_features" function in epiScanpy) of domain-specific peaks significantly decreases after imputation (Fig. 5e), demonstrating the ability of Descart in recovering signal disparities across different domains. To look deep in the differences of data imputation cases, we selected and visualized a domain-specific peak from each domain, respectively (Fig. 5f, g). Using only chromatin accessibility information (case 1) precisely boosts the signal within the domain but disrupts spatial continuity. Using only spatial information (case 2) is akin to applying a low-pass filter to the signal in space, which enhances signals within the domain but also inadvertently amplifies out-of-domain noises. Due to the incorporation of more diverse information, cases 3 and 4 show more precise imputation effects, proving to be more applicable in real-world scenarios.

To further explore the optimal parameters for imputation, we tested the impact of varying the number of nearest neighbors (assumed as k) for the graph construction of chromatin accessibility and different times of d_s (the mean Euclidean distance between each spot and its nearest neighbor) for the graph construction of spatial locations. Details on the settings and results of the ablation experiment are available in Additional file 1: Note S9 and Additional file 1: Fig. S20. Considering the overall effectiveness of imputation, ease of use, and preservation of cellular heterogeneity, we employ the default parameters of Descart ($k = 20$, times of d_s being 5) for data imputation represents an optimal solution.

Descart enables peak module identification

Features with co-variation can be clustered into a module, facilitating the identification of genes or peaks specifically associated with the development of a cell type even a specific tissue structure. Utilizing modules of features for analysis, as opposed to focusing on individual features, can mitigate noise and be more efficient for researchers, making the precise identification of modules with specific patterns crucial. However, compared to gene module identification via spRNA-seq data, the field for identifying peak modules based on spATAC-seq data still remains a significant gap. Moreover, due to the high-dimensional nature of spATAC-seq data and the highly discrete patterns of chromatin accessibility in space, precise peak module identification is a formidable challenge. To fill this gap, Descart leverages constructed graphs to calculate the similarity between peaks in terms of spatial locations and chromatin accessibility, facilitating the identification of peak modules through hierarchical clustering. In testing Descart on the E13.5-S1 slice of the mouse brain dataset, we initially identified 10,000 SV peaks and clustered them into eight peak modules of varying sizes (Fig. 6a). Next, we averaged read counts of peaks within these modules, visualized peak modules on the tissue space, and compared them with the spatial distribution of domains and tissue structures (Fig. 6b, c). The results reveal that nearly every peak module corresponds to one or more specific domains: modules 1 and 3 align with the "Mesenchyme" and "Muscle" domains, respectively; module 2 correlates with peripheral regions across multiple tissues encompassing various domains, including "Basal_plate_of_hindbrain," "Dpallm," "Dpallv," "Midbrain," and "Diencephalon_and_hindbrain"; modules 4, 6, and 7 correspond to two distinct regions

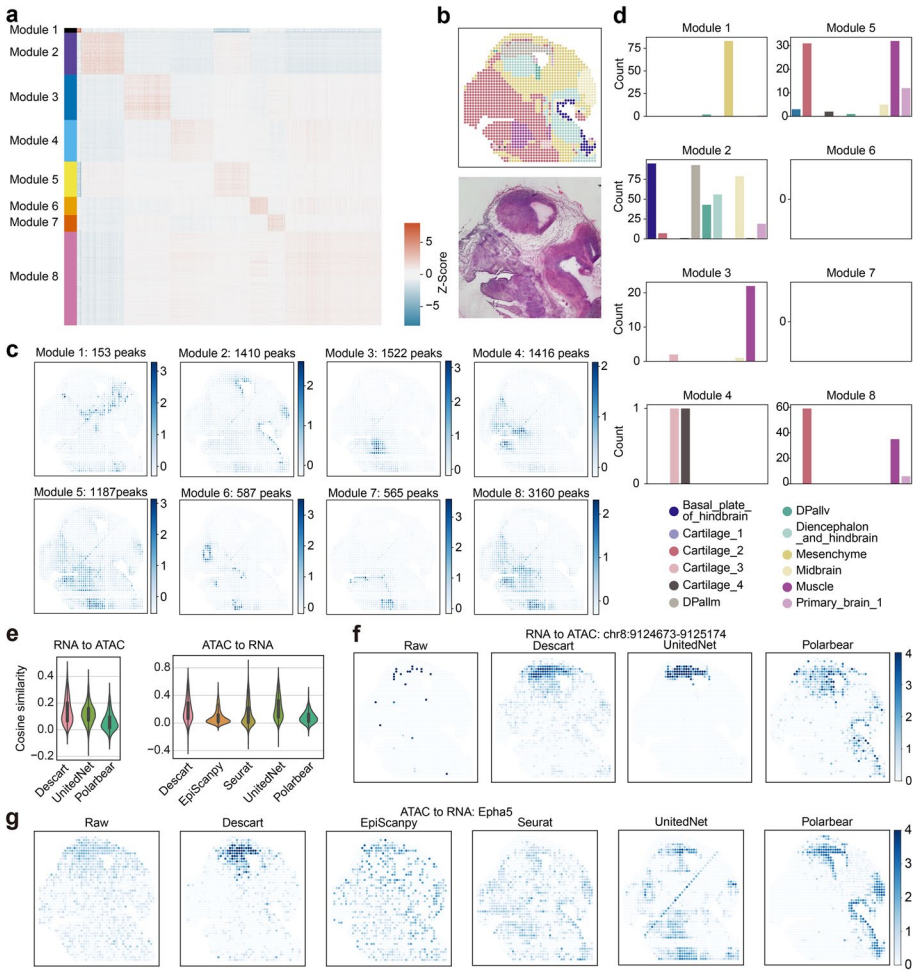

Fig. 6 Descart facilitates peak module identification and detection of gene-peak interaction. **a** Heatmap of peak-peak correlations generated by Descart. 10,000 SV peaks identified by Descart are grouped into 8 modules. **b** Visualization of domains in the tissue space (top) and the corresponding histological image (right). **c** Visualization of signals for each peak module in tissue space. Signals are averaged using raw counts of peaks from each module. **d** Overlapped counts between domain-specific peaks and peaks in each module. **e** Cosine similarities between the raw data and predicted data by different methods in the RNA to ATAC (left) and ATAC to RNA (right) transformation task. **f**, **g** Comparison between the raw data and predicted data by different methods, on the domain-specific peak (chr8: 9,124,673–9,125,174) and gene (Epha5) for "DPallm," in the RNA to ATAC (**f**) and ATAC to RNA (**g**) translation task. The raw data shown in **f** and **g** is performed TF-IDF and z-score transformation, and the predicted data is performed *z*-score transformation. All experiments and subplots corresponding to the whole figure are performed on the E13.5-S1 slice of the mouse brain dataset

within the "Cartilage_2" domain; modules 5 and 8 exhibit highly similar accessible patterns, covering regions within both the "Cartilage_2" and "Muscle" domains. We then performed GREAT analysis [34] to identify significant pathways associated with peaks from each module. As illustrated in Additional file 1: Fig. S21, the top-5 most significant pathways for each module are respectively associated with the development of various tissues in the mouse brain, aligning with the fetal stage of the mice used in the dataset. Furthermore, we assessed the overlap between peaks in each module and domain-specific peaks across different domains and found a strong alignment with the spatial visualization, demonstrating the potential of Descart in peak module identification (Fig. 6d).

Descart links gene expression and chromatin accessibility from spatial multi-omics data

Linking gene expression patterns to chromatin accessibility is a crucial step for constructing gene regulatory networks. Leveraging the graph of inter-cellular correlations, Descart can obtain a gene-peak correlation matrix, which captures the intensity of interactions between genes and peaks (the 'Methods" section). Utilizing the E13_5-S1 slice from the mouse brain dataset as a case study, we applied the Descart framework to identify 2000 SV genes and 20,000 SV peaks. To assess the accuracy of gene-peak correlations identified by Descart, we conducted a novel evaluation that uses the correlation matrix for reciprocal prediction of gene expression and chromatin accessibility and measures the performance by cosine similarity between predicted and original values. We only used the top twenty and bottom five peaks (or genes) most correlated with each gene (or peak) for prediction, with the corresponding correlation values for computation. For benchmarking, we compared two data-driven cross-omics translation methods, UnitedNet and Polarbear, as baselines, and two knowledge-based methods, functions for calculating gene activity scores in epiScanpy and Seurat. UnitedNet and Polarbear were trained and predicted on the same dataset, and such tasks are simpler than their real-world applications. As shown in Fig. 6e, except in the task transforming ATAC-seq to RNA-seq where UnitedNet excels, Descart outperforms methods specifically designed for cross-omics prediction, suggesting the potential for elucidating gene regulatory networks from spatial multi-omics data. Knowledge-based methods (epiScanpy and Seurat), focusing only on distances between peaks and genes without updating information from training data, generally underperform. The low cosine similarity values across methods may be attributed to differences in the sparsity of real and predicted data. Raw spatial multi-omics data, such as the domain-specific gene Epha5 and peak chr8:9,124,673–9,125,174, initially disperses in space, while the predicted data are imputed and exhibits greater spatial continuity (Fig. 6f, g). Visualization results demonstrate that Descart effectively restored and enhanced the raw signals. UnitedNet, a more complex neural network-based method, precisely enhances domain-specific signals in the RNA-seq to ATAC-seq transformation task but did not perform as well in other scenarios, but, like other baseline methods, falls short in another task. Furthermore, to validate gene-peak interactions from a biological perspective, we compared the interaction strength with corresponding Hi-C data [35]. For important genes in mouse brain development, Myog and Tcf7l1, the results show that their transcription start sites (TSS) and exons are located within the same topologically associating domain (TAD) structure as the peaks associated with them, identified by Descart (Additional file 1: Note S10 and Additional file 1: Fig. S22).

Discussion

Recent involutions in spatial sequencing technologies can simultaneously capture spatial location and chromatin accessibility of cells, and also increase the demand for SV peaks identification tailored for modeling spATAC-seq data. In this article, we introduce Descart, a method based on the graph of inter-cellular correlations, for identifying peaks characterized by both spatial variation and cellular heterogeneity. To our best knowledge, Descart is the first method for SV peaks identification tailored for spATAC-seq data. To deal with the challenge posed by the overly discrete spatial patterns of

chromatin accessibility in spATAC-seq data, Descart employs the following strategies to capture precise neighborhood relationships among cells: (i) integrating the graph constructed from chromatin accessibility information into the spatial graph; (ii) considering a broad range of neighboring cells when constructing the spatial graph, and (iii) iteratively updating the graph built from chromatin accessibility information with each iteration of SV peaks. To comprehensively evaluate our method, our benchmarking pipeline spans 16 slices from 4 datasets and incorporates three aspects from enhancing clustering performance, capturing domain-specific signals, and preserving spatial continuity. The benchmarking results demonstrate that Descart surpasses other methods with only spatial assumptions in both the accuracy and efficiency of SV peak identification. Due to the lack of spatial information, methods based on scATAC-seq data failed to maintain spatial continuity in the identified SV peaks, thereby undermining the accuracy of domain identification. Through case studies on the E13.5-S1 slice of the mouse brain dataset, we demonstrate the potential of Descart for data imputation, peak module identification, and gene-peak interaction detection. Overall, Descart offers an effective and valuable tool for spATAC-seq data analysis, contributing to detection of spatial chromatin accessibility patterns with inter-cellular correlations.

Despite the progress achieved so far, Descart still has several potential directions that need to be improved. First, to ensure high-efficiency of Descart on large-scale datasets, we will introduce downsampling or self-organizing maps for keeping the number of nodes in the constructed graph within a manageable range. The key concept of Descart is based on the graph of inter-cellular correlations, and its computational complexity scales quadratically with the number of cells. As spATAC-seq technologies evolve, increasing the spots (or cells) within datasets, Descart may face efficiency challenges. Second, we aim to refine existing simulation methods, such as simCAS, to provide multi-scenario simulated spATAC-seq data for systematic evaluation. Most feature selection methods based on spRNA-seq and scATAC-seq leverage simulated data for evaluation, as such data typically come with precise annotations. Finally, considering the availability of other spatial omics data, such as spatial metabolic data [36] and spatial CITE-seq data [37], we will broaden the application scope of Descart in our future work.

Conclusions

Utilizing the graph of inter-cellular correlations, Descart emerges as a groundbreaking method for SV peaks identification within spATAC-seq data. Through the comprehensive benchmark that incorporates three aspects from enhancing clustering performance, capturing domain-specific signals, and preserving spatial continuity, Descart demonstrates its superior performance compared with existing methods, significantly enhancing the efficacy and efficiency in identifying SV peaks. Besides, Descart also shows its ability for data imputation, peak module identification, and gene-peak interactions detection, making Descart a valuable tool for spATAC-seq data analysis.

Methods

Construction for graph of spatial locations

To capture spatial accessibility pattern of each peak, Descart construct a spatial graph based on spatial locations of spots (can also be replaced by cells). We suppose that

$\boldsymbol{X} = (\boldsymbol{x}_{\bullet 1}, \ldots, \boldsymbol{x}_{\bullet N})$ denotes as spatial locations of all N spots in a single slice, and the coordinates of the spatial locations are typically two-dimensional (i.e., $\boldsymbol{x}_{\bullet i}$ can be represented as $(x_{i1}, x_{i2})^T$). Descart first calculates the mean Euclidean distance between each spot and its nearest neighbor, denoting it as the standard distance $d_s = \frac{\sum_{i=1}^{N} \min_{j \neq i} Dist(\boldsymbol{x}_{\bullet i}, \boldsymbol{x}_{\bullet j})}{N}$, and then connects each spot to its neighbors within five times d_s to obtain a spatial graph. The edge weight of the spatial graph is inversely proportional to the Euclidean distance between spots, i.e., the edge weight between spot i and spot j is given by

$$w_{ij}^{(spatial)} = \begin{cases} \frac{d_s}{Dist(\boldsymbol{x}_{\bullet i}, \boldsymbol{x}_{\bullet j})}, Dist\left(\boldsymbol{x}_{\bullet i}, \boldsymbol{x}_{\bullet j}\right) \leq 5 d_s and i \neq j \\ 0, else \end{cases}$$

where $w_{ij}^{(spatial)}$ is an element in edge weight set $\boldsymbol{W}^{(spatial)}$ of the spatial graph, and $Dist(\boldsymbol{x}_{\bullet i}, \boldsymbol{x}_{\bullet j})$ denotes the Euclidean distance between spot i and spot j in tissue space.

Construction for graph of chromatin accessibility

The majority of graph-based existing methods for selecting primarily rely on spatial positional information for graph construction, while the constructed graph tends to overlook the inherent spot-spot relationships associated with gene expression or chromatin accessibility. In contrast to those methods, Descart incorporates a graph constructed from the ATAC-seq matrix and integrates it with the spatial graph. Due to the characteristics of high dimensionality, sparsity, and noise in the ATAC-seq matrix, constructing a graph directly from the raw count matrix is deemed impractical. Assuming that $\boldsymbol{Y} \in \mathbb{R}^{M \times N}$ denotes as the raw peak-by-spot (or peak-by-cell) count matrix with N spots and M peaks, Descart first applies term frequency-inverse document frequency (TF-IDF) transformation as in Signac [29] to obtain a continuously valued matrix (denoted as $\widetilde{\boldsymbol{Y}}$). Then, Descart selects a subset of peaks (50,000 peaks as the default) and performs PCA transformation to obtain a PC-by-spot matrix $\boldsymbol{P} = (\boldsymbol{p}_{\bullet 1}, \ldots, \boldsymbol{p}_{\bullet N}) \in \mathbb{R}^{10 \times N}$. Note that as an iterative method, in the initial iteration, Descart directly selects peaks based on their decreasing order of accessible degree and, in subsequent iterations, involves the selection according to the ranking from the previous iteration. With the PC-by-spot matrix $\boldsymbol{P}$, Descart connects each spot to its 20 nearest neighbors and defines the edge weight between spot i and spot j as

$$w_{ij}^{(ATAC-seq)} = \begin{cases} Cosine(\boldsymbol{p}_{\bullet i}, \boldsymbol{p}_{\bullet j}), j \in \left\{ Top\ 20\ nearest\ neighbors\ of\ cells\ i \right\} \\ 0\ , else \end{cases}$$

where $Cosine(\boldsymbol{p}_i, \boldsymbol{p}_j)$ denotes as the cosine similarity between vector $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$, $w_{ij}^{(ATAC-seq)}$ is an element in edge weight set $\boldsymbol{W}^{(ATAC-seq)}$ of the chromatin accessibility graph.

SV peaks selection

Given the graph of spatial locations and the graph of chromatin accessibility, Descart directly integrates the edges of them to obtain a graph of inter-cellular correlations. The edge weight of spot i and spot j in the newly-formed graph can be calculated as

$$w_{ij} = w_{ij}^{(spatial)} + \alpha w_{ij}^{(ATAC-seq)}$$

where α is the factor for balancing the two types of edge weights (1.5 in default), and w_{ij} is an element in edge weight set $\boldsymbol{W}$ of the inter-cellular correlations graph. On the other hand, Descart perform z-score transformation on the matrix to $\widetilde{\boldsymbol{Y}}$ obtain the matrix $\widetilde{\boldsymbol{Y}}$. The key concept of Descart is grounded in that peaks with regional continuity in the graph will be identified as informative peaks (i.e., SV peaks), while peaks with disparate accessibility levels between the two spots connected by the majority of graph edges will be considered as non-informative peaks (i.e., peaks that need to be filtered out). Under the concept, Descart evaluated each peak by an importance score using self-correlations, which can be obtained by

$$s_k = \widetilde{\boldsymbol{y}}_{k\bullet} \times \boldsymbol{W} \times \widetilde{\boldsymbol{y}}_{k\bullet}^{\boldsymbol{T}}$$

where $\widetilde{\boldsymbol{y}}_{k\bullet}$ is a row vector in the matrix $\widetilde{\boldsymbol{Y}}$ representing the peak k, and s_k is the importance score of the peak k. In our practical code implementation, this is accomplished using the formula $\boldsymbol{s} = diag(\widetilde{\boldsymbol{Y}} \times \boldsymbol{W} \times \widetilde{\boldsymbol{Y}}^T)$, where a single matrix operation completes the evaluation of all peaks. Based on the importance score, Descart sorts all peaks and feeds the ranking back into the step of constructing the graph of chromatin accessibility. The iterative process continues until the ranking stabilizes. Descart involves four iterations as the default setting, with outputting the final iteration's importance scores as the ultimate results. It is noteworthy that after performing omics-specific preprocessing and transforming the values of each feature into z-score-transformed values, Descart can also be employed to obtain the importance scores of features for any type of spatial sequencing data. Researchers can utilize the importance scores to fit various distributions to obtain p-values for all peaks. Alternatively, they can straightforwardly designate a specific number or a predetermined proportion of peaks as SV peaks.

Peak module identification

Based on the graph of inter-cellular correlations and its edge weights, Descart can obtain the peak-peak similarity matrix $\boldsymbol{S}^{(peak-peak)}$ by.

$$\boldsymbol{S}^{(peak-peak)} = \widetilde{\boldsymbol{Y}} \times \left(\frac{\boldsymbol{W} + \boldsymbol{W}^{\boldsymbol{T}}}{2} \right) \times \widetilde{\boldsymbol{Y}}^T$$

Subsequently, Descart transforms the matrix $\boldsymbol{S}^{(peak-peak)}$ to a peak-peak distance matrix $\boldsymbol{D}^{(peak-peak)}$ through (i) being subtracted by the 99.5th percentile of the elements in $\boldsymbol{S}^{(peak-peak)}$ from each element in $\boldsymbol{S}^{(peak-peak)}$ and (ii) setting the diagonal elements and negative values of the resulting matrix to zero. Finally, Descart utilizes the scipy package for hierarchical clustering (with the method parameter set to "ward") to identify peak modules.

Gene-peak interaction detection

For transcriptomics data within spatial multi-omics data, Descart uses the same preprocessing procedure as in Seurat and Scanpy. Specifically, Descart (i) scales the library size of each spot to 10,000, (ii) performs $\log(x + 1)$ transformation, and (iii) performs z -score transformation, on the raw gene-by-cell (or gene-by-spot) matrix. Then, Descart integrates the processed matrix with the z-score-transformed peak-by-spot matrix to

obtain a feature-by-spot matrix $\widetilde{\boldsymbol{Y}}'$, where features encompass all genes and peaks. Using the same calculation formula as for $\boldsymbol{S}^{(peak-peak)}$, Descart can derive a correlation matrix among all features, from which the corresponding submatrix reveals the strength of gene-peak or peak-gene interactions.

Data imputation

By incorporating the chromatin accessibility information of each spot with that of other cells, Descart enables data imputation on the raw data. Specifically, utilizing the graph of inter-cellular correlations and the weights of its edges, Descart performs a weighted averaging of the data from nearest neighbor spots and adds the result onto the raw data, that is

$$\widetilde{\boldsymbol{Y}}^{(enhanced)} = Scale\left(\widetilde{\boldsymbol{Y}} \times \boldsymbol{W}\right)$$

where $Scale(\bullet)$ denotes the function of z-score transformation, and $\widetilde{\boldsymbol{Y}}^{(enhanced)}$ represents the enhanced matrix that can be directly used for downstream analysis. Descart also provides two alternatives: (i) imputing data only based on neighborhood relationships from the spatial graph or the chromatin accessibility graph, that is, substituting $\boldsymbol{W}$ in the formula for obtaining $\widetilde{\boldsymbol{Y}}^{(enhanced)}$ with $\boldsymbol{W}^{(spatial)}$ or $\boldsymbol{W}^{(CAS)}$, and (ii) incorporating the original data with the imputed data, that is, $\widetilde{\boldsymbol{Y}}^{(enhanced)} = Scale\left(\widetilde{\boldsymbol{Y}} \times \boldsymbol{W} + \widetilde{\boldsymbol{Y}}\right)$.

Data collection

The mouse brain dataset [8], which simultaneously provides capture chromatin accessibility and gene expression of spots, consists of four tissue slices from mouse fetal brain at stages E11.0, E13.5, E15.5, and E18.5, respectively. In addition, the dataset also provides accurate manual anatomical annotations which reveal major tissue organizations of these slices, with reference to Kaufman's Atlas of Mouse Development and Allen Brain Atlas. The mixed-species A dataset [38] encompasses four annotated slices, two from mouse brains, one from the human brain, and another from the mouse embryo. Although this dataset lacks explicit domain labels, reliable clustering labels provided by the original authors are available. If the slices have multi-omics-derived clustering labels available, we use these as ground truths; otherwise, we use the clustering labels derived from the spATAC-seq data. The mouse embryo [6, 7] dataset composes of six tissue sections from three stages of mouse gestational development, including E12.5, E13.5, and E15.5 embryonic days. The mixed-species B dataset [7] comprises a collection of tissues from human and mouse, including five slices—three from mouse embryos, one from the mouse brain, and another from human tonsil. The metastatic melanoma dataset is a single-cell resolution dataset, also serving as a spatial multi-omics dataset [9]. In this dataset, Russell et al. have ingeniously integrated their Slide-tags technology with scRNA-seq and scATAC-seq, allowing for the simultaneous capture of chromatin accessibility and gene expression at the cellular level. Besides, Russell et al. also provided well-annotated cell type labels of the dataset in their study. A summary of the above datasets is shown in Additional file 2: Table S1.

Model evaluation

SV peaks selection

Due to the limited knowledge about the congruent relationship between domains (or cell types) and peaks, systematically benchmarking methods for SV peaks selection remains a tough challenge. For datasets with well-annotated labels, such as the mouse brain and metastatic melanoma datasets, we follow and improve the quantitative evaluation procedure from our prior research [27], that is, evaluating whether SV peaks identified by Descart can (i) perform better in facilitating cell clustering and (ii) capture more domain-specific (or cell type-specific) signals. The design of the evaluation framework and metric visualization is also informed by scIB [32], with specific calculations detailed as follows.

From the perspective of facilitating cell clustering, we utilize different methods to select SV peaks, and then perform unsupervised clustering to obtain annotated spatial domains. Due to the absence of analytic methods tailored for spATAC-seq data, we resort to the Signac [29] to obtain low-dimensional representations and clustering labels of spots (or cells). When clustering spots by the Leiden clustering (the default clustering method in Signac), we use a binary search to ensure that the number of clusters matches the number of domain labels. The clustering accuracy is assessed by NMI, ARI, and AMI scores, and higher values of the three metrics indicate that the method retains more spatial domain information in the identified SV peaks.

From the perspective of capturing domain-specific signals, we assess the performance using the overlap proportion between identified SV peaks and domain-specific accessible peaks. Specifically, given the domain labels in a specific dataset, we first use the "FindAllMarkers" function in Signac [29], the "tl.rank_features" function in epiScanpy [28], and the "tl.diff_test" function in snapATAC2 [33] to extract 100 domain-specific peaks for each domain and then compare the OP between specific peaks for each domain and SV peaks identified by various methods. OP1, OP2, and OP3 correspond to the overlap using the "tl.rank_features" function in epiScanpy, the "FindAllMarkers" function in Signac, and the "tl.diff_test" function in snapATAC2, respectively. A higher OP value indicates that SV peaks encompass more heterogeneity information related to specific domains. We evaluated the three aforementioned methods, along with several other approaches (including scaDA [39], scATAC-pro [40], ArchR [41], and snapATAC [42]) for identifying differentially accessible peaks, on the E13_5-S1 slice of the mouse brain dataset. Considering the reliability and specificity of the available ground truth, we ultimately used these three methods to obtain domain-specific signals (Additional file 1: Note S11 and Additional file 1: Fig. S23). Notably, the feature selection functions of Signac and epiScanpy are also used as baseline methods, but we verified that the operator does not introduce bias when benchmarking methods (Additional file 1: Note S11).

Referencing scIB, to ensure that each metric contributes equally to each perspective and possesses the same discriminative power, we perform min–max scaling on individual metrics among different methods. Specifically, for each metric, we scale the highest value among different methods into 1 and the lowest value into 0. After scaling, all metrics of each method for each perspective are averaged to obtain a score, with a higher value indicating superior performance of the method in that perspective. The overall score of each method is then determined by averaging the scores of the two perspectives,

serving as the final measure of each method's performance. During the benchmarking, all the methods are performed on the raw spATAC-seq datasets after peak calling, without filtering any low-quality peaks. Under the strategy, our benchmarking procedure aligns with current practices in the review [43], ensuring that comparisons directly reflect the intrinsic capabilities of the methods rather than their adaptability to the preprocessing.

For the mouse embryo and mixed species datasets, the absence of well-annotated labels precludes the same quantitative assessment as aforementioned. We here assess SV peaks identified by each method based on the ability to preserve spatial continuity as in spatialPCA [44]. Specifically, we perform cell clustering based on SV peaks and then use CHAOS, median LISI, and PAS scores to measure the spatial continuity of the clusters. Lower values of the three metrics indicate better performance of the corresponding method. The workflow for cell clustering is consistent with the aforementioned setting, and the number of clusters is set to 10 for all slices.

Detailed formulas for all the metrics mentioned above are provided in Additional file 1: Note S12. In addition to assessing the accuracy of SV peak identification, we conducted evaluations of the running time for different methods, i.e., the time they require from initial data processing to obtaining the ranking of each peak. All experiments were conducted on a server with 128 GB of memory and equipped with 32 units of 13th Gen Intel(R) Core(TM) i9-13900 K.

Baseline methods

Given the absence of methods specifically designed for spATAC-seq data, we assessed the performance of Descart against two categories of methods: (i) those based on spRNA-seq data, including SOMDE [15], SpatialDE2 [11], SpatialDE [12], SPARK-X [13], SPARK [31], scGCO [14], Sepal [16], and Moran's I [17], and (ii) those based on scATAC-seq data, including Cofea [27], HDA [22–26], Signac [29], and epiScanpy [28]. HDA, notably the most commonly used feature selection method for scATAC-seq data analysis, selects peaks with at least one read count in a majority of cells. We implemented HDA in Python following its instruction. EpiScanpy and Signac, both important pipelines for scATAC-seq data processing, were utilized solely for the feature selection functions as baseline comparison methods. The implementation of other methods was conducted using source code and default parameters provided in their respective studies.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03458-6.

Additional file 1: Note S1-S12. Note S1. The reason for using 10,000 SV peaks for benchmark. Note S2. Clustering performance with different SV peak numbers. Note S3. The rationale behind the superior efficiency of Descart. Note S4. Benchmarking methods using the mixed-species A dataset. Note S5. Similarity of the SV peaks identified by different methods. Note S6. Differences in the SV peaks identified by Descart and two variants. Note S7. Overlapped proportion of SV peaks obtained in each iteration by Descart. Note S8. Ablation experiments of Descart on the mouse brain dataset. Note S9. Ablation experiments for data imputation using different parameters in Descart. Note S10. Validation for gene-peak interactions from Descart using Hi-C data. Note S11. Detailed clarifications for the choice of domain-specific peak detection methods. Note S12. Detailed formulas for all the metrics. Fig S1-S23. Fig. S1. Clustering performance on the mouse brain dataset using different number of peaks. Fig. S2. Benchmarking results before metric transforming of different methods on the mouse brain dataset. Fig. S3. Overlapped proportion (OP) of SV peaks identified by Descart and baseline methods with domain-specific peaks related to overall domains on the mouse brain dataset. Fig. S4. Overlapped proportion (OP) of SV peaks identified by Descart and baseline methods with domain-specific peaks related to each domain on the mouse brain dataset. Fig. S5. Overview for benchmarking

results and running time on the mixed-species A dataset. Fig. S6. Benchmarking results before metric transforming of different methods on the mixed-species A dataset. Fig. S7. Overlapped proportion (OP) of SV peaks identified by Descart and baseline methods with domain-specific peaks related to overall domains. Fig. S8. Overlapped proportion (OP) of SV peaks identified by Descart and baseline methods with domain-specific peaks related to each domain. Fig. S9. Concordance of SV peak rankings evaluated by each method. Fig. S10. Overlapped proportion of SV peaks identified by each method under different peak numbers. Fig. S11. Overlapped proportion of SV peaks identified by Descart and baseline methods. Fig. S12. Visualization of domains and top-ranked SV peak on the slices from the mouse brain dataset. Fig. S13. Venn diagrams and box plot on the mouse brain dataset showing the overlap of the top 10,000 peaks identified by Descart and its two single-element-focused variants. Fig. S14. The overlap proportion of the top 10,000/50,000 important peaks identified by Descart between successive iterations. Fig. S15. Quantitive assessment of the impact of different parameters with NMI scores as the metric. Fig. S16. Clustering performance on the mouse embryo dataset. Fig. S17. The top-ranked SV peak in the tissue space on the slices from the mouse embryo dataset. Fig. S18. The top-ranked SV peak in the tissue space on the slices from the mixed-species B dataset. Fig. S19. Benchmarking results on the metastatic melanoma dataset. Fig. S20. The impact of varying the number of nearest neighbors for the graph construction of chromatin accessibility and different times of d_s for the graph construction of spatial locations. Fig. S21. The top-5 most significant pathways for each module with p-values calculated through GREAT analysis. Fig. S22. Contact matrix of the mouse brain development Hi-C data. Fig. S23. Overlapped proportion among domain-specific peaks and visualization of marker peaks on the E13_5-S1 slice of the mouse brain dataset.

Additional file 2: Table S1. Summary of the datasets used in our study.

Additional file 3. Review history.

Acknowledgements
Not applicable.

Peer review information
Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 3.

Authors' contributions
R.J. conceived the study and supervised the project. X.C. and K.L. designed, implemented, and validated Descart. X.W., Z.L., Q.J., and Y.W. helped with analyzing the results. X.C., K.L., and R.J. wrote the manuscript, with input from all the authors.

Funding
This research was funded by the National Key Research and Development Program of China, grant number 2021YFF1200902 (R.J.), the National Natural Science Foundation of China, grant numbers 62273194 (R.J.) and 61721003 (R.J.), and Beijing Natural Science Foundation, grant numbers L242026.

Data availability
The Descart software, including detailed documents and tutorial, is freely available on GitHub under a MIT license (https://github.com/likeyi19/Descart) [45], with the version used in the manuscript also deposited in Zenodo (https://zenodo.org/records/14248995) [46].
The mouse brain dataset [8] is available accessed under National Genomics Data Center accession number (OEP003285 [47], www.biosino.org/node/project/detail/OEP003285). The mixed-species A dataset [38], mouse embryo dataset [6], mixed-species B dataset [7], and metastatic melanoma dataset [9] can be accessed from the Gene Expression Omnibus (GEO) under accession number GSE205055 [48], GSE214991 [49], GSE171943 [50], and GSE244355 [51], respectively.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 25 March 2024 Accepted: 9 December 2024
Published online: 30 December 2024

References

1. Zhang M, Eichhorn SW, Zingg B, Yao Z, Cotter K, Zeng H, Dong H, Zhuang X. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. Nature. 2021;598:137–43.

2. Asp M, Giacomello S, Larsson L, Wu C, Furth D, Qian X, Wardell E, Custodio J, Reimegard J, Salmen F, et al. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. Cell. 2019;179(1647–1660): e1619.
3. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, et al: Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 2018, 361.
4. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, Macosko EZ. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019;363:1463–7.
5. Lomakin A, Svedlund J, Strell C, Gataric M, Shmatko A, Rukhovich G, Park JS, Ju YS, Dentro S, Kleshchevnikov V, et al. Spatial genomics maps the structure, nature and evolution of cancer clones. Nature. 2022;611:594–602.
6. Llorens-Bobadilla E, Zamboni M, Marklund M, Bhalla N, Chen X, Hartman J, Frisen J, Stahl PL. Solid-phase capture and profiling of open chromatin by spatial ATAC. Nat Biotechnol. 2023;41:1085–8.
7. Deng Y, Bartosovic M, Ma S, Zhang D, Kukanja P, Xiao Y, Su G, Liu Y, Qin X, Rosoklija GB, et al. Spatial profiling of chromatin accessibility in mouse and human tissues. Nature. 2022;609:375–83.
8. Jiang F, Zhou X, Qian Y, Zhu M, Wang L, Li Z, Shen Q, Wang M, Qu F, Cui G, et al. Simultaneous profiling of spatial gene expression and chromatin accessibility during mouse brain development. Nat Methods. 2023;20:1048–57.
9. Russell AJC, Weir JA, Nadaf NM, Shabet M, Kumar V, Kambhampati S, Raichur R, Marrero GJ, Liu S, Balderrama KS, et al. Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. Nature. 2024;625:101–9.
10. Palla G, Fischer DS, Regev A, Theis FJ. Spatial components of molecular tissue biology. Nat Biotechnol. 2022;40:308–18.
11. Kats I, Vento-Tormo R, Stegle O: SpatialDE2: fast and localized variance component analysis of spatial transcriptomics. *Biorxiv* 2021**:**2021.2010. 2027.466045.
12. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. Nat Methods. 2018;15:343–6.
13. Zhu J, Sun S, Zhou X. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. Genome Biol. 2021;22:184.
14. Zhang K, Feng W, Wang P. Identification of spatially variable genes with graph cuts. Nat Commun. 2022;13:5488.
15. Hao M, Hua K, Zhang X. SOMDE: a scalable method for identifying spatially variable genes with self-organizing map. Bioinformatics. 2021;37:4392–8.
16. Andersson A, Lundeberg J. sepal: identifying transcript profiles with spatial patterns by diffusion-based modeling. Bioinformatics. 2021;37:2644–50.
17. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, Rybakov S, Ibarra IL, Holmberg O, Virshup I, et al. Squidpy: a scalable framework for spatial omics analysis. Nat Methods. 2022;19:171–8.
18. DeTomaso D, Yosef N. Hotspot identifies informative gene modules across modalities of single-cell genomics. Cell Syst. 2021;12(446–456): e449.
19. Wu Y, Hu Q, Wang S, Liu C, Shan Y, Guo W, Jiang R, Wang X, Gu J. Highly Regional Genes: graph-based gene selection for single-cell RNA-seq data. J Genet Genomics. 2022;49:891–9.
20. Chen C, Kim HJ, Yang P. Evaluating spatially variable gene detection methods for spatial transcriptomics data. Genome Biol. 2024;25:18.
21. Charitakis N, Salim A, Piers AT, Watt KI, Porrello ER, Elliott DA, Ramialison M. Disparities in spatially variable gene calling highlight the need for benchmarking spatial transcriptomics methods. Genome Biol. 2023;24:209.
22. Chen X, Chen S, Song S, Gao Z, Hou L, Zhang X, Lv H, Jiang R. Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. Nature Machine Intelligence. 2022;4:116–26.
23. Chen S, Yan G, Zhang W, Li J, Jiang R, Lin Z. RA3 is a reference-guided approach for epigenetic characterization of single cells. Nat Commun. 2021;12:2177.
24. Zamanighomi M, Lin Z, Daley T, Chen X, Duren Z, Schep A, Greenleaf WJ, Wong WH. Unsupervised clustering and epigenetic classification of single cells. Nat Commun. 2018;9:2410.
25. Xiong L, Xu K, Tian K, Shao Y, Tang L, Gao G, Zhang M, Jiang T, Zhang QC. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. Nat Commun. 2019;10:4576.
26. Liu Q, Chen S, Jiang R, Wong WH. Simultaneous deep generative modeling and clustering of single cell genomic data. Nat Mach Intell. 2021;3:536–44.
27. Li K, Chen X, Song S, Hou L, Chen S, Jiang R: Cofea: correlation-based feature selection for single-cell chromatin accessibility data. *Brief Bioinform* 2023, 25.
28. Danese A, Richter ML, Chaichoompu K, Fischer DS, Theis FJ, Colome-Tatche M. EpiScanpy: integrated single-cell epigenomic analysis. Nat Commun. 2021;12:5228.
29. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. Nat Methods. 2021;18:1333–41.
30. Moran PA. Notes on continuous stochastic phenomena. Biometrika. 1950;37:17–23.
31. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat Methods. 2020;17:193–200.
32. Luecken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colome-Tatche M, Theis FJ. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19:41–50.
33. Zhang K, Zemke NR, Armand EJ, Ren B: SnapATAC2: a fast, scalable and versatile tool for analysis of single-cell omics data. *bioRxiv* 2023.
34. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28:495–501.
35. Chakraborty A, Wang JG, Ay F. dcHiC detects differential compartments across multiple Hi-C datasets. Nat Commun. 2022;13:6827.
36. Sun C, Wang A, Zhou Y, Chen P, Wang X, Huang J, Gao J, Wang X, Shu L, Lu J, et al. Spatially resolved multi-omics highlights cell-specific metabolic remodeling and interactions in gastric cancer. Nat Commun. 2023;14:2692.

37. Liu Y, DiStasio M, Su G, Asashima H, Enninful A, Qin X, Deng Y, Nam J, Gao F, Bordignon P, et al. High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq. Nat Biotechnol. 2023;41:1405–9.
38. Zhang D, Deng Y, Kukanja P, Agirre E, Bartosovic M, Dong M, Ma C, Ma S, Su G, Bao S, et al. Spatial epigenome-transcriptome co-profiling of mammalian tissues. Nature. 2023;616:113–22.
39. Zhao F, Ma X, Yao B, Lu Q, Chen L. scaDA: a novel statistical method for differential analysis of single-cell chromatin accessibility sequencing data. PLoS Comput Biol. 2024;20: e1011854.
40. Yu W, Uzun Y, Zhu Q, Chen C, Tan K. scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. Genome Biol. 2020;21:94.
41. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, Greenleaf WJ. Author Correction: ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat Genet. 2021;53:935.
42. Fang R, Preissl S, Li Y, Hou X, Lucero J, Wang X, Motamedi A, Shiau AK, Zhou X, Xie F, et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nat Commun. 2021;12:1337.
43. Li Z, Patel ZM, Song D, Yan G, Li JJ, Pinello L: Benchmarking computational methods to identify spatially variable genes and peaks. *bioRxiv* 2023.
44. Shang L, Zhou X. Spatially aware dimension reduction for spatial transcriptomics. Nat Commun. 2022;13:7203.
45. Chen X, Li K, Wu X, Li Z, Jiang Q, Cui X, Gao Z, Wu Y, Jiang R. Detection of spatial chromatin accessibility patterns with inter-cellular correlations. GitHub; 2024. https://github.com/likeyi19/Descart.
46. Chen X, Li K, Wu X, Li Z, Jiang Q, Cui X, Gao Z, Wu Y, Jiang R. Detection of spatial chromatin accessibility patterns with inter-cellular correlations. 2024. Zenodo. https://zenodo.org/records/14248995.
47. Jiang F, Zhou X, Qian Y, Zhu M, Wang L, Li Z, Shen Q, Wang M, Qu F, Cui G, Chen K, Peng G. Simultaneously spatiotemporal gene expression and chromatin accessibility for mouse brain development. National Genomics Data Center. https://www.biosino.org/node/project/detail/OEP003285.
48. Zhang D, Deng Y, Kukanja P, Agirre E, Bartosovic M, Dong M, et al. Spatial epigenome–transcriptome co-profiling of mammalian tissues. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE205055.
49. Llorens-Bobadilla E, Zamboni M, Marklund M, Bhalla N et al. Solid-phase capture and profiling of open chromatin by spatial ATAC. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE214991.
50. Deng Y, Bartosovic M, Ma S, Zhang D et al. Spatial profiling of chromatin accessibility in mouse and human tissues. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE171943.
51. Russell AJC, Weir JA, Nadaf NM, Shabet M et al. Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE244355.

Publisher's Note