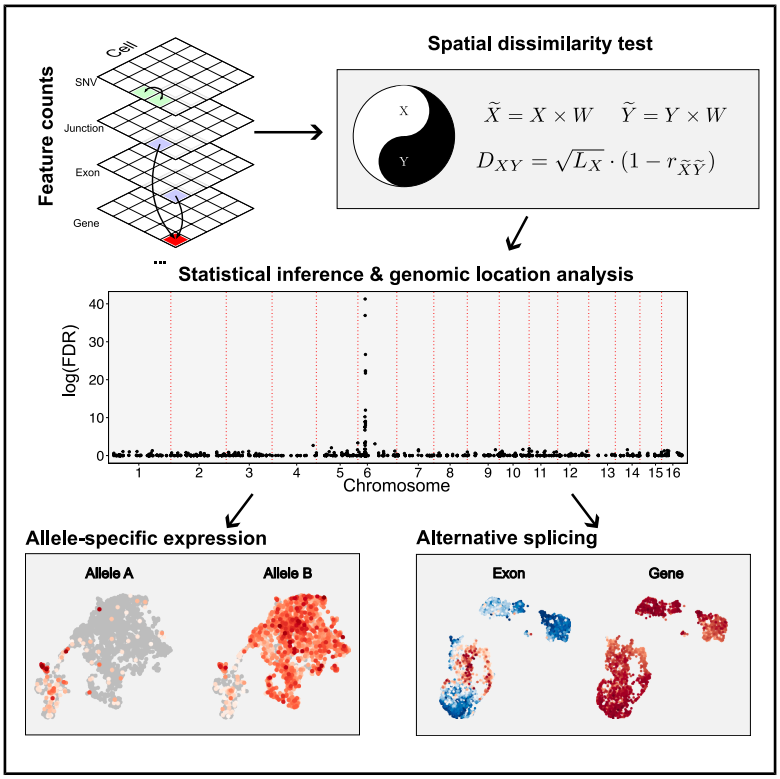
Cell Reports
Methods

Article

Spatial dissimilarity analysis in single-cell transcriptomics

Graphical abstract



Highlights

- We develop a statistical framework to analyze linked feature pairs in single-cell data
- We use this method to detect and visualize alternative splicing events in neurons
- We also demonstrate application to allele-specific gene expression in tumors
- We provide software tools (PISA/Yano) to implement the method

Authors

Quan Shi, Karsten Kristiansen

Correspondence

quan.shi@bio.ku.dk (Q.S.),
kk@bio.ku.dk (K.K.)

In brief

Shi and Kristiansen introduce a statistical framework to analyze differential expression patterns between biologically linked feature pairs. This approach enables detection of alternative splicing and allele-specific gene expression in single-cell and spatial transcriptomics data, offering new insights into complex gene regulation across diverse biological contexts.

Shi & Kristiansen, 2025, Cell Reports Methods *5*, 101141
September 15, 2025
https://doi.org/10.1016/j.crmeth.2025.101141

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Spatial dissimilarity analysis in single-cell transcriptomics

Quan Shi[1,*] and Karsten Kristiansen[1,2,3,4,*]
[1]Laboratory of Integrative Biomedicine, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark
[2]BGI-Shenzhen, Shenzhen 518103, China
[3]Institute of Metagenomics, Qingdao-Europe Advanced Institute for Life Sciences, BGI-Qingdao, Qingdao 166555, China
[4]Lead contact
*Correspondence: quan.shi@bio.ku.dk (Q.S.), kk@bio.ku.dk (K.K.)
https://doi.org/10.1016/j.crmeth.2025.101141

MOTIVATION Many statistical tools developed for single-cell and spatial transcriptomics analysis primarily focus on gene-level features. However, sequence-level signals such as alternative splicing and allele-specific expression may offer deeper insights into cellular dynamics and regulatory mechanisms. Compared with bulk RNA-seq, the application of such analyses to single-cell data has been limited by the biased nature of RNA library preparation and the typically low coverage across gene bodies. We sought to develop a statistical framework to analyze differences in expression patterns between biologically linked feature pairs even when expression is sparse, making it suited for current-state single-cell and spatial transcriptomics data.

SUMMARY

We develop the spatial dissimilarity method to uncover complex bivariate relationships in single-cell and spatial transcriptomics data, addressing challenges such as alternative splicing and allele-specific gene expression. Applying this method to detect alternative splicing in neurons demonstrates improved accuracy and sensitivity compared to existing tools, notably identifying neuron subtypes. In tumor cells, spatial dissimilarity analysis reveals somatic variants that emerge during tumor progression, validated through whole-exome sequencing. These findings highlight how allele-specific genetic variants contribute to the subclone architecture of cancer cells, offering insights into cellular heterogeneity. Applied on a human cell atlas, we uncover numerous cases of allele-specific expression of genes in normal cells. We provide a software package for spatial dissimilarity analysis to enable enhanced understanding of cellular complexity and gene expression dynamics under homeostatic conditions and during states of transitions.

INTRODUCTION

Since the generation of the first single-cell transcriptome dataset using next-generation sequencing,[1] the past 15 years have seen remarkable advancements in single-cell RNA sequencing (scRNA-seq)[2,3] and spatial transcriptomics,[4–6] greatly enhancing our understanding of cell heterogeneity[7,8] and cell-cell interactions.[9] Despite these advancements, most research continues to focus primarily on gene expression, often overlooking other characteristics of sequence data, such as genome coordinate information, alternative splicing (AS),[10,11] alternative polyadenylation,[12] natural antisense transcripts,[13,14] somatic variants,[15] and allele-specific expression (ASE).[16–21] While some of these features have been extensively studied in bulk RNA-seq, their analysis at the single-cell level remains limited. For instance, AS analysis in bulk RNA-seq typically involves assembling short reads into full transcripts,[22] identifying differential exon usage between groups,[23] and quantifying splicing events using percent spliced in (PSI).[24,25] However, applying these approaches to single-cell data introduces additional complexities. The high dropout rates and low gene coverage inherent in scRNA-seq make it challenging to perform reliable AS analysis at the individual cell level.[26]

To address these challenges, researchers often aggregate cells within a group to create pseudo-bulk samples, enabling the use of bulk RNA-seq tools,[27] or resample group labels for cells multiple times to evaluate the distribution of delta PSI scores[28] to identify alternative splicing between groups. Yet, the number of cells within a group can vary significantly, from a few cells to thousands or more, making it difficult to balance sample sizes across pseudo-bulk samples. Furthermore, this approach results in the loss of nuanced biological information, such as continuous phenotypic transitions observed as cells progress from one state to another.[29] To address these

limitations, recent tools designed specifically for single-cell data employ imputation methods or Bayesian models to assign AS events to cells and perform differential splicing analysis by testing the significance of the contributions of cell group factors in linear regression models.[25,30] However, as these methods require known cell labels and/or primarily rely on PSI or junction reads as input, they are often constrained to specific exon exclusion events.

In parallel, spatial statistics are increasingly being applied to spatial transcriptomics analyses to address biological questions.[31] Perhaps the most widely used spatial statistic in spatial transcriptomics is Moran's *I*, which measures spatial autocorrelation or dependence for a given feature.[31,32] Nevertheless, examining the spatial relationship between two genes is also important, as it provides insights into gene regulatory networks and cell-cell interactions.[33] It is well known that Pearson's correlation fails to account for spatial dependence,[34] underscoring the need for methods that can describe differences between two spatial processes. Numerous bivariate spatial correlation methods have been developed, which can be broadly categorized into two groups: modified t test, often adjusted for effective sample size,[35,36] and new statistical measures such as Wartenberg's bivariate extension of Moran's *I*,[37] Lee's L score,[38,39] and spatial cross-correlation index.[40] In spatial transcriptomics analysis, statistical measures are more commonly used than modified t test to assess the relationship between two genes, as the measures can be further utilized for clustering and defining expression modules in spatial contexts.[41,42] This approach can also be applied in single-cell analysis, where cell-cell similarity graphs are used instead of spatial distances to identify informative genes and co-expression modules in cell space.[42,43]

Despite advancements in spatial transcriptomics and the application of spatial statistics, most existing methods focus on assessing the spatial autocorrelation of a feature and its similarity patterns with other features.[41,42,44–46] This leaves the exploration of distinct expression patterns between two features in spatial contexts largely underexplored. For example, considering the spatial autocorrelation of an exon while simultaneously examining differences in its expression pattern relative to its gene can provide critical insights into differential exon usage in spatial transcriptome and single-cell analysis. However, to the best of our knowledge, no existing statistical method fully addresses this requirement.

In this article, we introduce a spatial measure to analyze the spatial autocorrelation of a feature across cell trajectory (1D), spatial location (2D), or multiple dimensions and test its dissimilarity with a biologically connected binding feature. We term this method spatial dissimilarity (SD) test and extend it to single-cell analysis. The spatial autocorrelation of the test feature ensures its relevance within the cell space, while distinct expression patterns between the test and binding features often reveal specific biological phenomena. To demonstrate the utility of our method, we apply it to two common cases: alternative splicing and allele-specific gene expression. For alternative splicing, we compare the distinct spatial expression patterns of an exon or junction with the overall expression of its gene in single cells. Similarly, for allele-specific gene expression, we examine the differences in spatial expression patterns between one allele (test feature) and the other allele (binding feature). Beyond these cases, our method has the potential to serve as a foundational tool for a wide range of single-cell analyses, providing insights into spatial expression dynamics and underlying biological mechanisms. To facilitate these analyses, we present an updated version of the PISA toolkit[47] for annotating and quantifying various transcriptomic features from BAM files, along with an new R package, Yano, that integrates seamlessly with the popular Seurat package.[48] Additionally, Yano includes a visualization function to plot unique molecule identifier (UMI) depth for each cell group and an annotation function to annotate genetic variants with various databases and predict the molecular consequence. The package is open-source and available at https://github.com/shiquan/Yano.

RESULTS

Overview of the spatial dissimilarity analysis method

Our goal was to systematically prioritize spatially dissimilar expression events between molecular features (test features) and their associated features (binding features). Notably, this method requires the test feature to exhibit spatial autocorrelation, while no specific distribution is assumed for the binding feature. The analysis is structured into five distinct steps. In step one, we begin by using our preprocessing tool, PISA,[47] to annotate and quantify a wide array of molecular features through aligned reads (Figure 1A). These features include genome location-based entities, such as genes, exons, junctions, exon-excluded reads, and sequence-based entities like single nucleotide variants (SNVs) (Figure 1B).

Next, we construct a cell-to-cell distance/similarity matrix and set up the test and binding features. The matrix is formulated by the sequential arrangement of cells along a developmental trajectory,[29] spatial coordinates,[31] or shared nearest neighbors (SNNs) within a high-dimensional space.[48,49] The SNN method is preferred for its robustness in defining cell neighbors, surpassing the conventional K-nearest neighbors (KNNs) approach,[48,49] being calculated by principal-component analysis (PCA) or integrated space such as Harmony.[50] The distance matrix is then transformed into a cell-cell weight matrix, which is column normalized and is directly used to compute Moran's *I* for each feature and only features that are spatial autocorrelated (Moran's $I > 0$ and Benjamini-Hochberg adjusted p value < 0.01) are used in downstream analysis.

In the third step, the association between a test feature and its binding feature is established. The association relationship is usually guided by the biological question. For instance, to study alternative splicing at the exon level, the corresponding gene serves as the binding feature. For allele-specific gene expression, one allele expression can be used as test feature while the binding feature may be the genomic locus or the alternative allele. Normalized counts of the test and binding features are used as input for downstream analysis. Depending on the specific question, we employ three operational modes for binding features: (1) the binding feature's normalized counts are used as input (mode 1); (2) the binding feature's counts are subtracted by the test feature's counts before normalization (mode 2); and

CellPress
OPEN ACCESS

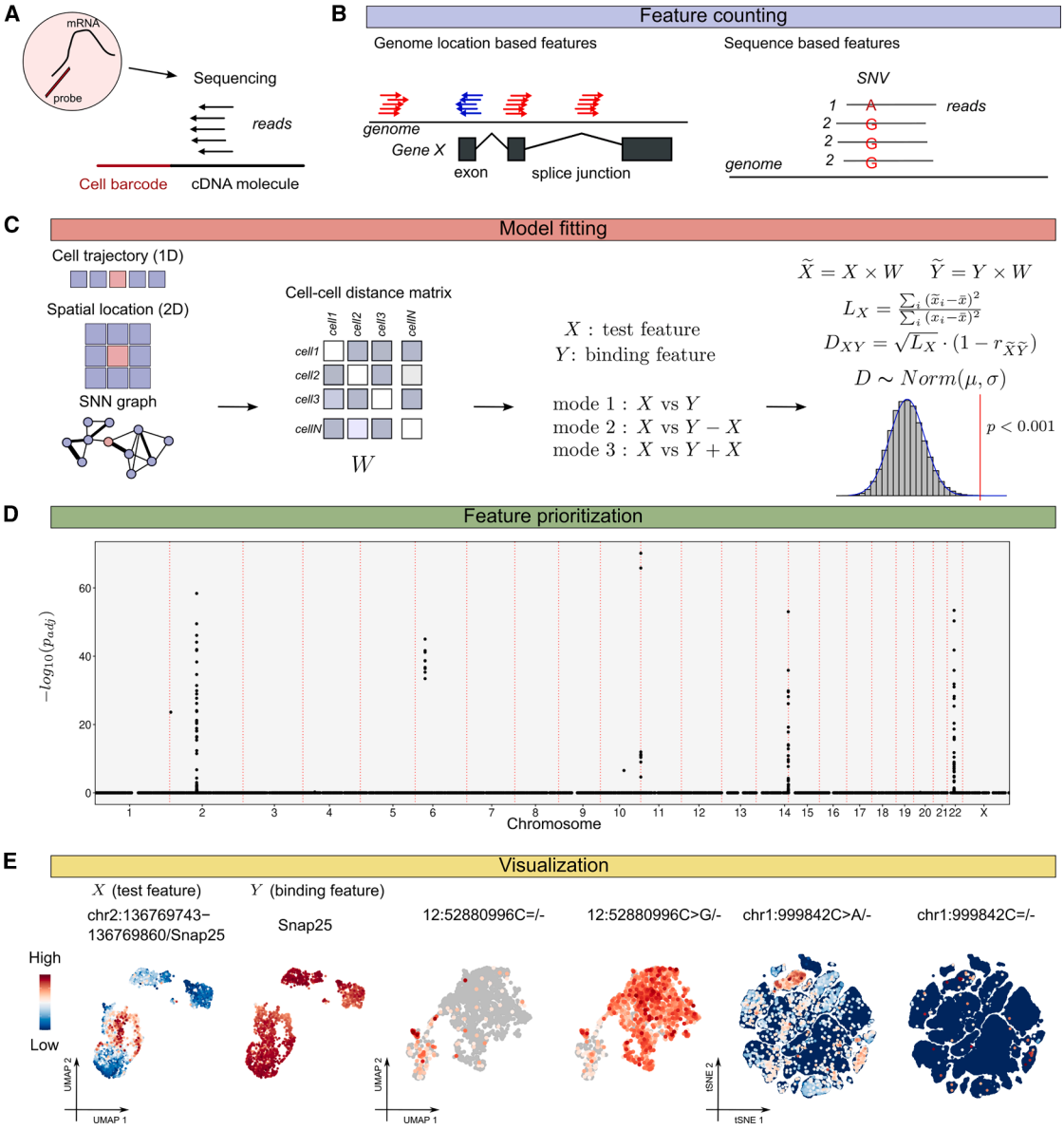


Figure 1. Overview of the spatial dissimilarity analysis method

(A) Simplified pipeline of single-cell RNA sequencing, highlighting key steps such as molecule capture and the sequencing strategy.

(B) Counting features derived from alignment reads. This includes genome location-based features like genes, exons, and junctions, as well as sequence-based features such as SNVs.

(C) Outlines the process of conducting a spatial dissimilarity test. Initially, a cell-cell distance graph is constructed based on either the cells' positions in a single-cell linear trajectory (1D), their spatial locations (2D), or an SNN graph. Subsequently, cell-cell distance matrix is normalized by column. Then the binding relationship is set up, and the analysis mode is selected. The D score is calculated with feature counts and weight matrix, and it follows a normal distribution. The *p* value for D score is calculated based on the permutated mean value and standard deviation to assess the significance of spatial dissimilarity.

(D) Genomic locations of features and *p* values plot on a Manhattan plot. The *p* value is transformed using a negative log10 scale for visualization. The genomic locations are sorted by chromosome and coordinates.

(E) Visualizing feature pairs on dimensionality reduction maps. (Left) An example demonstrating differential exon usage, with the exon expression and corresponding gene expression displayed on a UMAP plot. (Middle) An example illustrating allele-specific gene expression, showing both alleles at chromosome 12 (position 52880996) plotted on UMAP. (Right) Another example of differential allele expression at chromosome 1 (position 999842), visualized using t-SNE.

(3), the counts of the test and binding features are added together before normalization (mode 3).

In the fourth step, we perform the SD test for each feature pair and identify significant events. A spatial measure, the D score, is introduced to link the spatial autocorrelation of the test feature (L) and the differential expression pattern relative to the binding feature ($1 - r$, where r is the Pearson's correlation score of the spatial lag of the test and binding features) (Figure 1C). Although

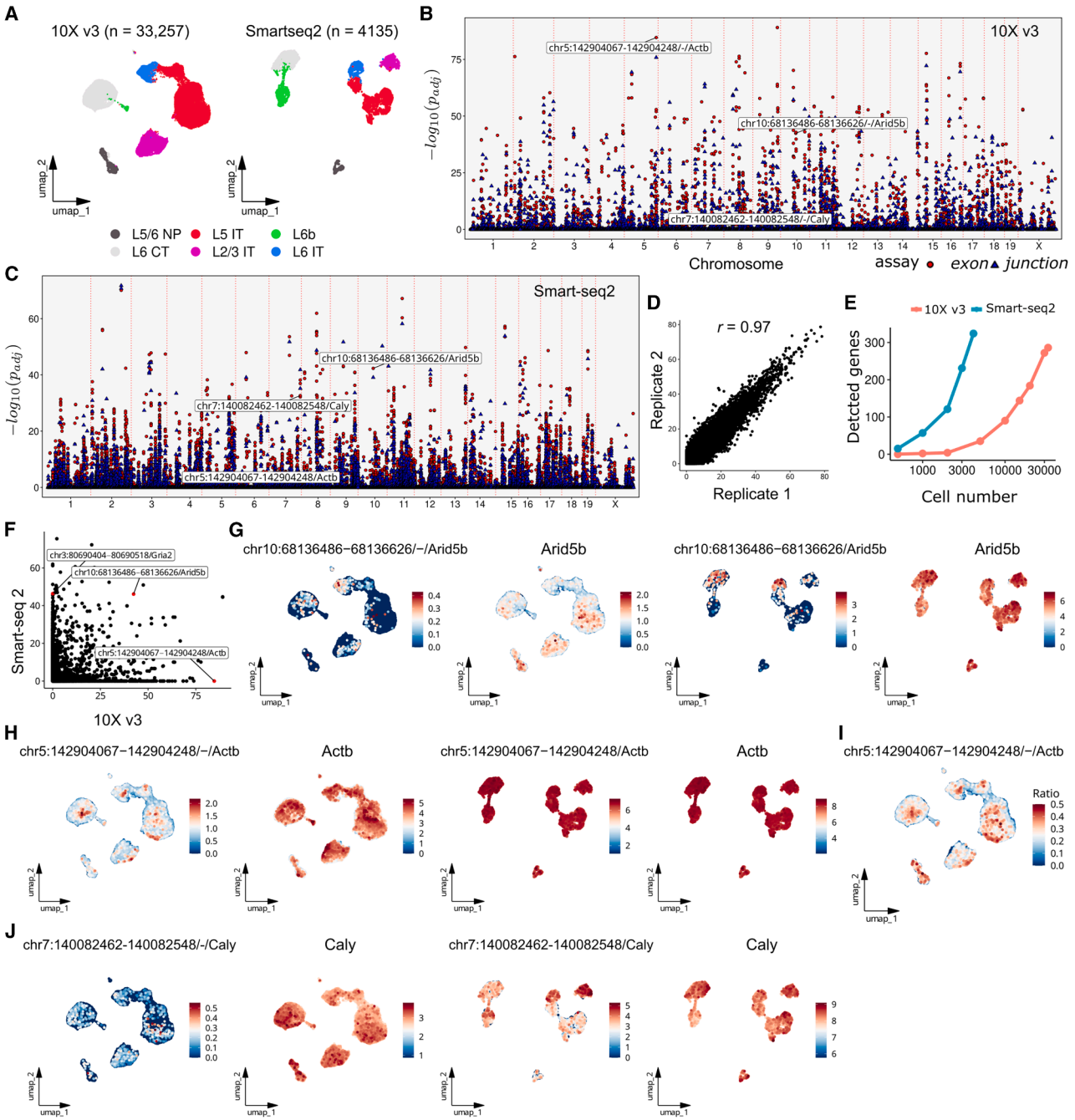


Figure 2. Spatial dissimilarity analysis reveals alternative splicing features in mouse primary motor cortex

(A) UMAP visualization of the MOp dataset from the 10× Genomics v3 library (left) and the Smart-seq2 library (right), with cells color-coded by cell types. CT, corticothalamic; IT, intratelencephalically projecting; NP, near-projecting.

(B) Manhattan plot illustrating the results of the spatial dissimilarity test between exon/junction and gene pairs for 10× Genomics v3 data. Points are color coded and shape coded by the assay types and sorted by genomic coordinates along the *x* axis. *p* value of each event was transformed using a negative log10 scale for visualization.

(C) Like (B), this Manhattan plot focuses on the spatial dissimilarity test between exon/junction and gene pairs for Smart-seq2 data. Points are differentiated by color and shape based on the assay type and sorted by genomic coordinates along the *x* axis. *p* value of each event was transformed using a negative log10 scale for visualization.

(D) Correlation plot of adjusted *p* values from two groups of cells. These *p* values were derived from 10× Genomics v3 data, where cells were randomly divided into two groups, and the SD test was performed separately for each group. The *p* values were transformed using a negative log10 scale for visualization. Each point represents a feature, illustrating the consistency of adjusted *p* values between the two groups.

(legend continued on next page)

the initial distribution of D scores is unknown, a permutation method reveals that D scores follow a normal distribution (Figures S1A and S1B). To ensure comparability, the *p* value for each event is calculated. For multiple tests, the Benjamini-Hochberg procedure is used to calculate the false discovery rate.[51] Our package, Yano, conducts random sampling of the test feature's distribution across cells 100 times, while the binding feature remains fixed, to generate the distribution of D scores. Using the mean and standard deviation of these permuted D scores, *p* values are calculated via a one-tailed t statistic. High D scores correspond to (1) large L scores (indicating spatial autocorrelation of the test feature); (2) low *r* scores (indicating differing expression patterns between the test and binding features).

Finally, we prioritize features based on their *p* values and genomic locations (Figure 1D). While selecting features solely by *p* values is acceptable, clustered *p* values at nearby genomic locations provide stronger evidence of significant events. The prioritized features are then visualized using dimensionality reduction maps (Figure 1E, left) and track plots. For instance, a distinct spatial dissimilarity pattern for an exon on chromosome 2 (positions 136769743 to 136769860) and its associated gene *Snap25* indicates an alternative splicing event. On chromosome 12 (position 52880996), contrasting expression patterns of alleles C and G across cells suggest the loss of allele-specific gene expression in certain cells (Figure 1E, middle). Notably, our SD test evaluates the spatial dependency of the test feature without imposing restrictions on the distribution of the binding feature. This characteristic is particularly advantageous for identifying genes that have an informative expressed allele and another low-expressed allele, which may result from somatic mutations or allele-selective gene repression (Figure 1E, right).

Spatial dissimilarity analysis reveals alternative splicing features in mouse primary motor cortex

We applied our SD analysis to four published datasets generated using Smart-seq2[52] and 10× Genomics 3′ and 5′ kits, covering the most commonly used library structures. Initially, we applied the SD analysis to 33,257 neurons from six 10× Genomics 3′ libraries and 4,135 neurons from a Smart-seq2 library, both derived from the mouse primary motor cortex (MOp) (Figure 2A).[53] These cells were clustered and integrated at the gene level using the Seurat[48] and Harmony[50] pipelines, with cell annotations and selected cells adapted from the original dataset.[53] To ensure robust detection, we excluded features with low coverage, retaining only those detectable in a minimum of 20 cells. This process identified 62,422 junctions (approximately twice the number of detected genes) and 584,173 exons for the 10× libraries, and 166,257 junctions (approximately seven times the number of detected genes) and 304,217 exons for the Smart-seq2 library. Eventually, we identified 1,317 junctions and exons with *p* values below the threshold of 1e−10 for the 10× data (Figure 2B) and 2,100 junctions and exons for the Smart-seq2 data (Figure 2C).

To test the robustness of our SD method, we divided the 10× library cells into two groups of 16,628 cells each and re-ran the SD test on exon data. The *p* values of the two groups showed a high correlation ($r = 0.97$) (Figure 2D). To compare the performance of full-length sequencing and 3′ biased sequencing data using the SD test, we downsampled cells from both the 10× and Smart-seq2 libraries and performed the SD test for exons and genes. As expected, with an equal number of cells, the 10× library detected fewer genes than the Smart-seq2 library due to its lower gene coverage (Figure 2E). Furthermore, the number of detected genes increased with a higher cell count, though saturation was not observed even with over 30,000 cells in the 10× data (Figure 2E). Comparisons of exons between the two kinds of libraries revealed that 142 exons were prioritized by both methods (adjusted $p < 1e{-}10$), while 684 exons were uniquely detected by Smart-seq2 (adjusted $p < 1e{-}10$ for Smart-seq2, adjusted $p < 1e{-}2$ for 10×) and 574 exons were uniquely detected by the 10× data (adjusted $p < 1e{-}10$ for 10×, adjusted $p < 1e{-}2$ for Smart-seq2) (Figure 2F). As expected, we observed that shared alternative splicing events show similar expression pattern between both libraries (Figure 2G). We also found that exons detected only in the 10× data often displayed unique spatial expression patterns; for example, exon chr5:142904067–142904248 was spatially expressed in nearby cells, while its gene, *Actb*, appeared globally expressed across all cells in the 10× data (Figure 2H). In the Smart-seq2 data, however, this exon's expression pattern aligned with that of the gene (Figure 2H). When plotting the ratio of this exon to gene counts on a reduction map, the ratio followed the exon's expression pattern rather than the gene's pattern (Figure 2I). This suggests that the distinct expression pattern is unlikely to be caused by low gene body coverage, as the reads overlapping this exon did not increase proportionally with the gene's total reads. For another exon, chr7:140082462–140082548, prioritized only by Smart-seq2,

(E) Downsampling analysis of 10× Genomics v3 and Smart-seq2 data. Detected genes represent those with at least one event having an adjusted *p* value < 1e−10.

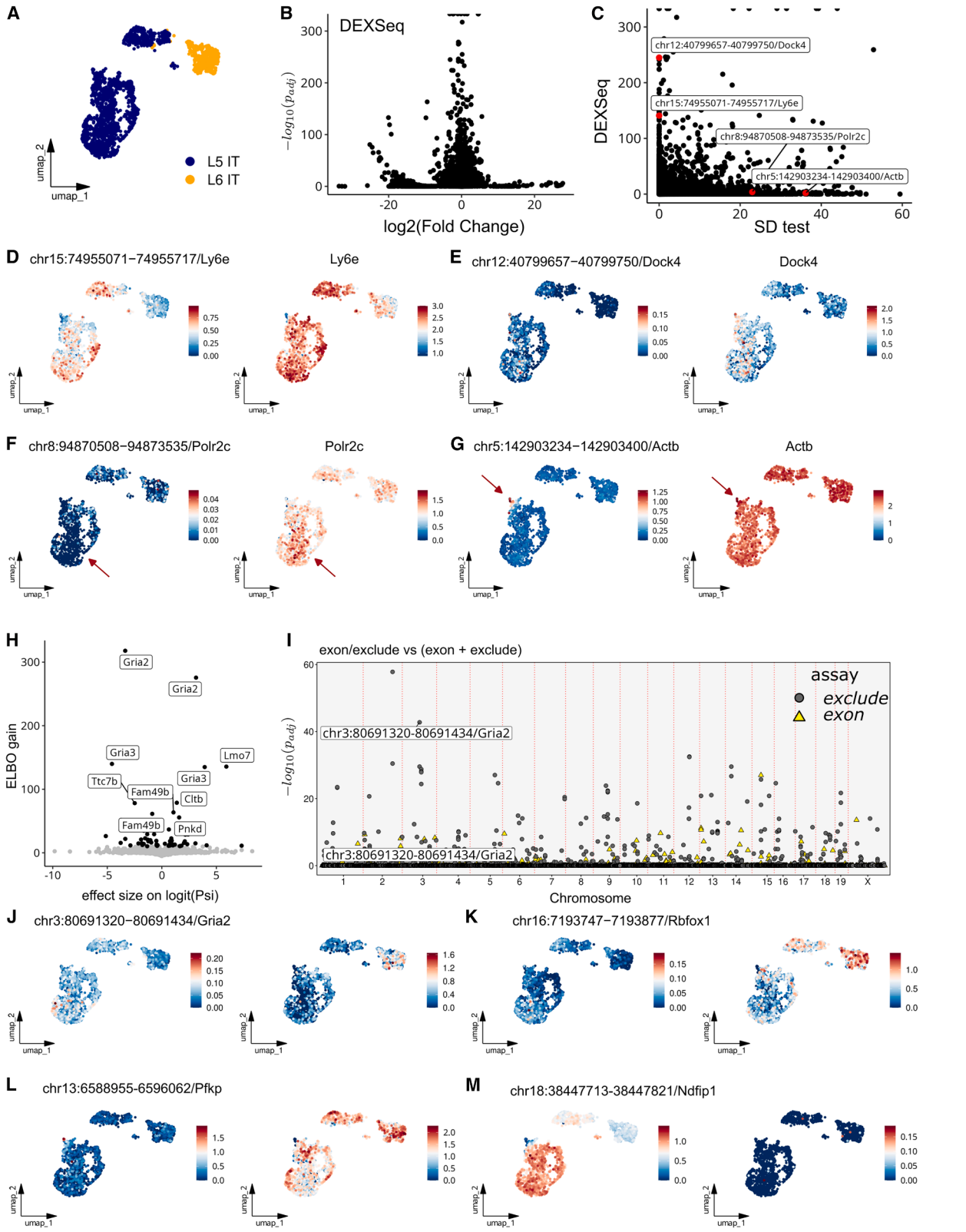
(F) Correlation plot of adjusted *p* values for exon and gene pairs from 10× Genomics v3 and Smart-seq2 data. *p* values along the *x* axis and *y* axis were transformed using a negative log10 scale for visualization. Strand information in exon names from 10× data was removed to enable comparison with Smart-seq2 data.

(G) UMAP plots showing the expression of exon chr10:68136486–68136626 and the gene Arid5b in 10× Genomics v3 data (left) and Smart-seq2 data (right). Expression levels are normalized on a log scale and adjusted for library size. Note that due to differences in library sizes between assays, expression values are not directly comparable across plots.

(H) UMAP plots showing the normalized expression of exon chr5:142904067–142904248 and the gene *Actb* in 10× Genomics v3 data (left) and Smart-seq2 data (right).

(I) UMAP plots showing the ratio of raw counts for exon chr5:142904067–142904248 to the gene *Actb* across individual cells.

(J) UMAP plots showing the normalized expression of exon chr7:1400882462–149982548 and the gene *Caly* in 10× Genomics v3 data (left) and Smart-seq2 data (right).



(legend on next page)

the pattern appeared to be introduced by low gene body coverage (Figures 2J and S2A–S2E).

To further understand alternative expression events across different cell clusters and systematically summarize these features, we co-clustered exon and gene counts and generated a heatmap for all cells using Smart-seq2 data (Figure S3). This analysis revealed complex alternative splicing cases within the dataset. For example, exon chr16:18629808–1862911 was expressed across all cell groups but showed low expression in group L2/3 IT, while the gene *Sept5* was highly expressed in both L5 IT and L2/3 IT (Figure S3, highlighted). Although Smart-seq2 has greater power to detect alternative splicing due to its ability to capture full-length gene sequences, its lack of strand specificity can hinder analysis. For example, the gene *Stk16* overlaps with gene *Tuba4a* on the opposite strand. Since *Tuba4a* is highly expressed and independent of *Stk16*, exons of *Stk16* appear with low *p* values in Smart-seq2 data and introduce false positive events. In contrast, the strand-specific nature of 10× data allows these genes to be distinguished based on read strands (Figures S4A–S4C).

By applying our SD analysis to MOp neurons from both 10× and Smart-seq2 libraries, we identified thousands of alternative splicing events, demonstrating that this method is robust across different library preparations. Although 3′ biased sequencing has lower sensitivity for detecting alternative splicing compared with full-length sequencing at the same sample sizes, analyzing a larger cell population enables the discovery of novel splicing events.

Benchmarking of the spatial dissimilarity test method against cluster-based methods

Despite the availability of established methods for investigating AS in single-cell analyses, most techniques are designed to detect differential AS events between two cell groups. To benchmark the performance of our SD test, we selected two cell clusters (L5 IT and L6 IT) from the Smart-seq2 dataset (Figure 3A). Rerunning SD analysis for these cells revealed new AS events for all cells, suggesting that "local" events may be masked during a "global" test (Figures S5A and S5B). For instance, the exon chr9:78478449–78478858 of gene *Eef1a1* was prioritized in selected cells due to its distinct expression pattern. However, when analyzing all cells together, this different expression pattern was diluted and not sufficient to be highlighted (Figure S5C).We then used two software tools for comparison: DEXSeq,[23] which was originally developed for bulk RNA-seq and uses exon and gene counts as input, and BRIE2,[30] which uses PSI values as input. These tools represent the most commonly used approaches for AS analysis. To begin, we generated three pseudo-samples for each cell group and ran the DEXSeq workflow, detecting 2,372 exons with adjusted *p* values below 1e−10 (Figure 3B). In contrast, our SD test identified 400 exons and 122 junctions, considerably fewer than DEXSeq. Examining the differences between DEXSeq and our method revealed that 2,064 exons were detected solely by DEXSeq (adjusted $p < 1e{-}10$ for DEXSeq, adjusted $p > 0.01$ for SD test) (Figure 3C). DEXSeq's high sensitivity detects differences in exon usage between groups, but it is easy to find many differences that arise from low gene expression (as in *Ly6e*; Figures 3D and S5D). Especially when both gene expression and exon expression are low in one group, the lineage regression model can become biased, leading to artificially low *p* values due to dispersion (Figure 3E).

Conversely, 224 exons were exclusively detected by our SD test (adjusted $p > 0.01$ for DEXSeq, adjusted $p < 1e{-}10$ for SD test). These exons generally showed pronounced differences in expression patterns between exons and genes, such as chr15:74955071–74955717 (Figure 3F) and chr12:40799657–40799750 (Figure 3G). Because our SD test assesses the entire cell population without considering cell cluster labels, subclusters within the cell population may also be detected.

In the case of BRIE2, which prioritizes AS events using the evidence lower bound (ELBO) score[30] (Figure 3H), we conducted a fair comparison by first testing exon to exon-excluded reads and then exon-excluded reads to exon with mode 3 (Figure 3I). In

Figure 3. Benchmarking of the spatial dissimilarity test method against cluster-based methods

(A) UMAP plot of L5 IT and L6 IT cells from Smart-seq2 data, with cells colored by cell type.

(B) Volcano plot of DEXSeq results. The *x* axis represents the log2 fold change between L5 IT and L6 IT, while the *y* axis shows *p* values transformed using a negative log10 scale for visualization.

(C) Comparison of same exon analyzed with DEXSeq and the spatial dissimilarity test results. *p* values are negative log10 scaled.

(D) UMAP plots showing the expression of exon chr15:74955071–74955717 (left) and the gene *Ly6e* (right). Expression levels are normalized on a log scale and adjusted for library size. Note that due to differences in library sizes between gene assay and exon assay, expression values are not directly comparable across plots.

(E) UMAP plots showing the normalized expression of exon chr12:40799657–40799750 (left) and the gene *Dock4* (right).

(F) UMAP plots showing the normalized expression of exon chr8:94870508–94873535 (left) and the gene *Polr2c* (right). The arrow highlights cells exhibiting an inverse expression pattern between the exon and the gene.

(G) UMAP plots showing the normalized expression of exon chr5:142903234–142903400 (left) and the gene *Actb* (right). The arrow highlights cells exhibiting an inverse expression pattern between the exon and the gene.

(H) Volcano plot of BRIE2 results. The *x* axis represents the effect size difference of PSI between L5 IT and L6 IT, while the *y* axis shows the ELBO gain. Exons with low ELBO scores (<10) are colored in gray, while other exons are colored in black.

(I) Manhattan plot illustrating the results of the spatial dissimilarity test between exon assay and exon-excluded assay (and vice versa) in mode 3. Points are color and shape coded based on assay type and sorted by genomic coordinates along the *x* axis. *p* values for each event were transformed using a negative log10 scale for visualization.

(J) UMAP plots showing the normalized expression of exon chr3:80691320–80691434 (left) and exon-excluded reads for the same exon (right).

(K) UMAP plots showing the normalized expression of exon chr16:7193747–7193877 (left) and exon-excluded reads for the same exon (right).

(L) UMAP plots showing the normalized expression of exon chr13:6588955–6596062 (left) and exon-excluded reads for the same exon (right).

(M) UMAP plots showing the normalized expression of exon chr18:38447723–38447821 (left) and exon-excluded reads for the same exon (right).

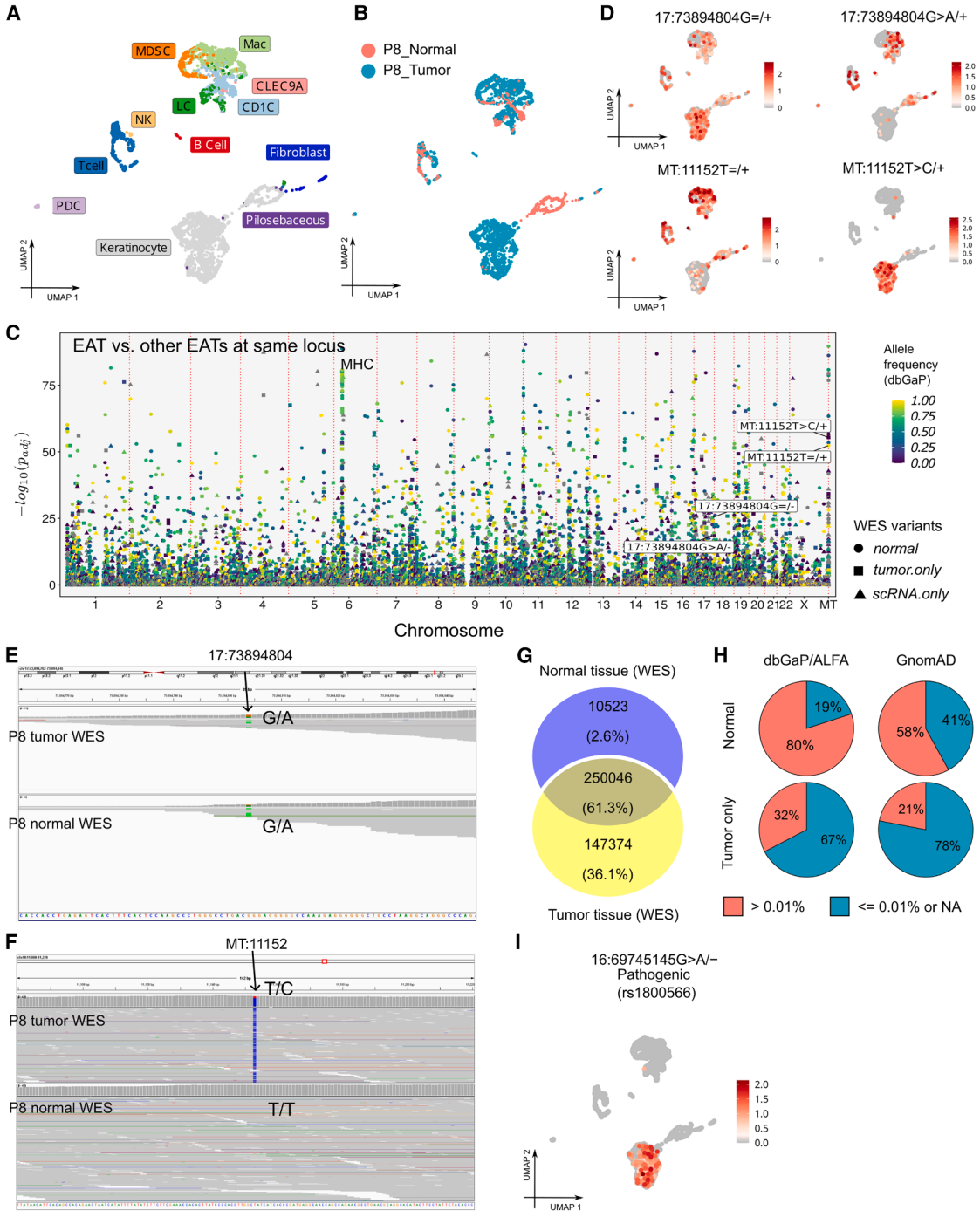


Figure 4. Spatial dissimilarity analysis of EATs reveals allele-specific gene expression in cutaneous squamous cell carcinoma

(A) UMAP visualization of cutaneous squamous cell carcinoma (cSCC) cells, each color-coded according to identified cell clusters. Myeloid-derived suppressor cells, MDSC; natural killer, NK; macrophage, Mac; CD1C^{+} dendritic cells, CD1C; CLEC9A^{+} dendritic cells, CLEC9A; Langerhans cells, LC.

(B) UMAP visualization of cSCC cells colored by tissue type.

(C) Manhattan plot illustrating the results of the spatial dissimilarity test between EATs and other EATs at the same locus. Points are color coded by AF from the dbGaP/ALFA database, shaped by variant origin, and sorted by genomic coordinates along the *x* axis. *p* values for each event were transformed using a negative log10 scale for visualization.

(D) UMAP plots showing the expression of EAT 17:73894804 allele G and allele A (top) and EAT MT:11152 allele T and allele C (bottom). Expression levels are normalized on a log scale and adjusted for library size. As both EATs at each locus are derived from the same assay, their expression values are directly comparable.

(legend continued on next page)

mode 3, exon-included and exon-excluded reads were summed as the binding assay, ensuring that only events with strong differential expression patterns between exon-included and exon-excluded reads were highlighted, resulting in the identification of 46 genes (adjusted p value < 1e−10). We also performed this test using mode 1, which is more sensitive and prioritized 1,985 genes, which is much higher than mode 3 (Figure S6A). Notably, all events detected with mode 3 were also detected with mode 1. Our SD test successfully identified genes also prioritized by BRIE2, such as *Gria2* (Figure 3J) and *Rbfox1* (Figure 3K). However, BRIE2 only provides a score for each gene without identifying specific alternative splicing sites, complicating result interpretation. Five genes detected by BRIE2 (ELBO > 30) were not identified by our SD test with mode 3 but can be detected with mode 1. Upon inspection, we found that minor differences were observed in spliced reads for these genes, though these differences were not substantial (Figures S6B–S6D). Additionally, we noted that the genes exclusively detected with our SD test often were influenced by alternative expression events within the L5 IT group (Figure 3L). Furthermore, high L scores for the test feature combined with sparse expression of the binding feature can also lead to high D scores under mode 3 (Figure 3M), as the D score only considers the spatial dependency of the test feature. Therefore, under mode 3, it is recommended to intersect the called events with both exon-included and exon-excluded data.

This benchmarking demonstrated that methods like DEXSeq, which rely on linear regression models that treat cell clusters as test factors, are highly sensitive but not suitable for biased sequenced data. They often introduce many false positives and make the results hard to interpret. Tools like BRIE2 that use PSI as input can detect only a limited number of events because they require high gene body coverage, which is an obstacle for cell groups with few cells. In contrast, our SD test strikes a balance between sensitivity and specificity by focusing on expression pattern differences in cell space. Furthermore, our method does not rely on predefined cell group information, making it well suited for uncovering hidden cell populations within the data.

Spatial dissimilarity analysis of SNVs reveals allele-specific gene expression in squamous cell carcinoma

We called SNVs in 3,464 cells from both the tumor and paired normal tissues of a human cutaneous squamous cell carcinoma (cSCC) sample[54] (Figure 4A). Genetic variants were collated from whole-exome sequencing (WES), and scRNA-seq data of the tumor and normal tissues (Figure 4B) subsequently merged into a variant call format (VCF) file.[55] By post-annotation of reads with the VCF file, we quantified expression levels for both strands for all alleles at these loci (STAR Methods). Alleles with expression in fewer than 10 cells were filtered out to diminish noise and errors. Because our allele-specific expression analysis not only considers SNVs but also considers the reference allele, for convenience, we use expressed allele specific tag (EAT) to stand for SNV and reference allele throughout this article. Unlike gene expression, which is directly counted from alignment reads, we define the binding assay here as locus expression, calculated by summing up the counts of all EATs at a given locus. Using the SD test in mode 2 (comparing one allele to another at the same locus), we identified 229,507 EATs across 194,018 loci, treating each strand-specific locus independently. For loci to be analyzed in the SD test, at least two EATs per locus were required. Of the 31,876 loci meeting this criterion, 2,240 showed significant alternative expression of one allele compared to the other at the same locus (adjusted p < 1e−10) (Figure 4C).

Certain EATs with cell-type-specific expression were observed; for example, EAT 11:252818 allele G was expressed in myeloid cells and keratinocytes across all samples, whereas allele A was lost in epithelial cells within the tumor (Figure 4D, top). In another case, EAT MT:11152 allele C was primarily expressed in keratinocytes in the tumor, though it was also detected in a single macrophage and one $CD1C^+$ dendritic cell (Figure 4D, bottom). Furthermore, an analysis of WES sequence reads revealed that both allele G and allele A at chromosome 17, position 73894804, were detected in normal and tumor DNA, indicating that the expression of allele A was silenced in keratinocytes within the tumor (Figure 4E). In contrast, the gain of allele C at mitochondrial position 11152 was attributed to somatic variants (Figure 4F), suggesting it may have been acquired during tumor progression. The presence of this allele in non-epithelial cells could be explained by library contamination or mitochondrial transfer between tumor and immune cells.[56]

Our package, Yano, then predicted the consequences of genetic variants and linked genotype information with phenotype data from published databases (STAR Methods). We observed that over 36% of mutations in tumor cells were not present in paired normal cells (Figure 4G). After annotation with human population allele frequency databases,[57–60] these mutations typically exhibited lower allele frequencies (Figure 4H), suggesting that rare genetic variants (allele frequency < 0.0001) or those undetected in public databases are more likely to be somatic mutations compared to variants with higher frequencies. Annotating of variants using the ClinVar database[61] identified a known pathogenic mutation expressed only in tumor cells, indicating that these mutations may be acquired during tumor progression (Figure 4I). In summary, our SD test, in conjunction with genetic and phenotype databases, successfully revealed ASE events and informative somatic mutations in cancer cells. This approach holds a significant potential for enhancing our understanding of the etiology of cancer development.

(E) IGV plot showing reads from WES data of tumor and normal tissues overlapping with chromosome 17, position 73894804.
(F) IGV plot showing reads from WES data of tumor and normal tissues overlapping with mitochondria, position 11152.
(G) Venn diagram showing the genetic variants detected in WES data from normal tissues and tumor tissues.
(H) Pie charts representing the proportion of low AF somatic mutations (detected in tumor tissues only) across different human population AF databases: dbGaP/ALFA database (left) and the Genome Aggregation database (GnomAD) (right).
(I) UMAP plots showing the expression of a known pathogenic mutation across cells.

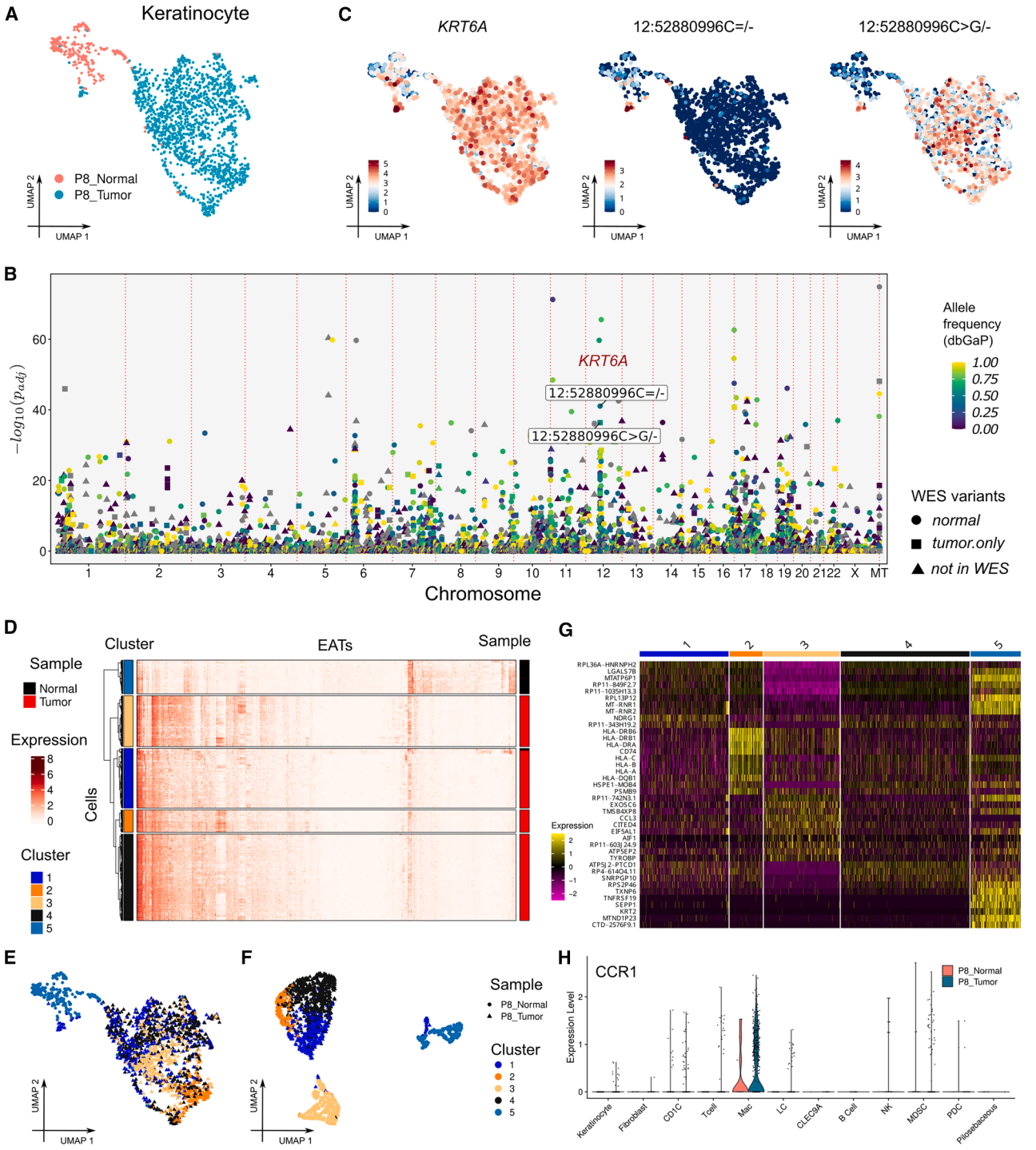


Figure 5. Identification of cutaneous squamous cell carcinoma subclones with alternatively expressed EATs

(A) UMAP visualization of keratinocytes, colored by tissue type.

(B) Manhattan plot illustrating the results of the spatial dissimilarity test between EATs and other EATs at the same locus. Points are color coded by AF from the dbGaP/ALFA database, shaped by variant origin, and sorted by genomic coordinates along the *x* axis. *p* values for each event were transformed using a negative log10 scale for visualization.

(C) UMAP plots showing the expression of gene *KRT6A* (left), EAT 12:52880996 allele C (middle), and allele G (right). Expression levels were retrieved from the "data" layer of Seurat object with the default method.

(legend continued on next page)

Identification of squamous cell carcinoma subclones with alternatively expressed EATs

We investigated the utility of alternative EAT expression patterns in distinguishing tumor subclones within a tumor sample. Traditional analyses based on gene expression typically rely on highly variable genes (HVGs) for PCA and downstream cell clustering, which may obscure ASE genes. Focusing on keratinocytes, we re-clustered these cells using PCA and UMAP analyses based on highly variable EATs in place of HVGs (Figure 5A). We then performed SD tests for all EATs, highlighting 265 EATs across 195 loci (mode 2) (Figure 5B). Notably, a concentration of alternatively expressed EATs was observed at chromosome 12 q13.13, predominantly originating from gene *KRT16A*. Additionally, certain EATs demonstrated the ability to differentiate between tumor and normal cells. For example, EAT 12:52880996 allele G was present across all cell groups, while allele A showed high expression in normal tissue cells but only sporadic expression in tumor cells. This suggests the presence of distinct tumor subclones (Figure 5C).

Subsequently, keratinocytes were hierarchically clustered into 5 groups with the EATs from the 195 loci, with most normal tissue cells predominantly in group 5 (Figure 5D). When mapping these cluster labels onto the UMAP plot, the original reduction method did not distinctly separate cell groups 1, 2, and 3 (Figure 5E). However, re-clustering the cells with alternatively expressed EATs achieved clear segregation of these groups (Figure 5F and STAR Methods). Further gene-level expression analysis among these cell groups revealed variations in cancer-related genes, such as *DDIT4*[62] and *HILPDA*,[63] across the tumor subclones (Figure 5G). Of particular interest was the high expression of the ligand gene *CCL3* in group 2 and the corresponding receptor gene *CCR1* in tumor cells (Figure 5H). The *CCL3-CCR1* axis has been implicated in various cancers,[64,65] suggesting a potential involvement in cSCC subclone progression.

Spatial dissimilarity analysis of EATs in 15 adult human tissues reveals new allele-specific expressed genes

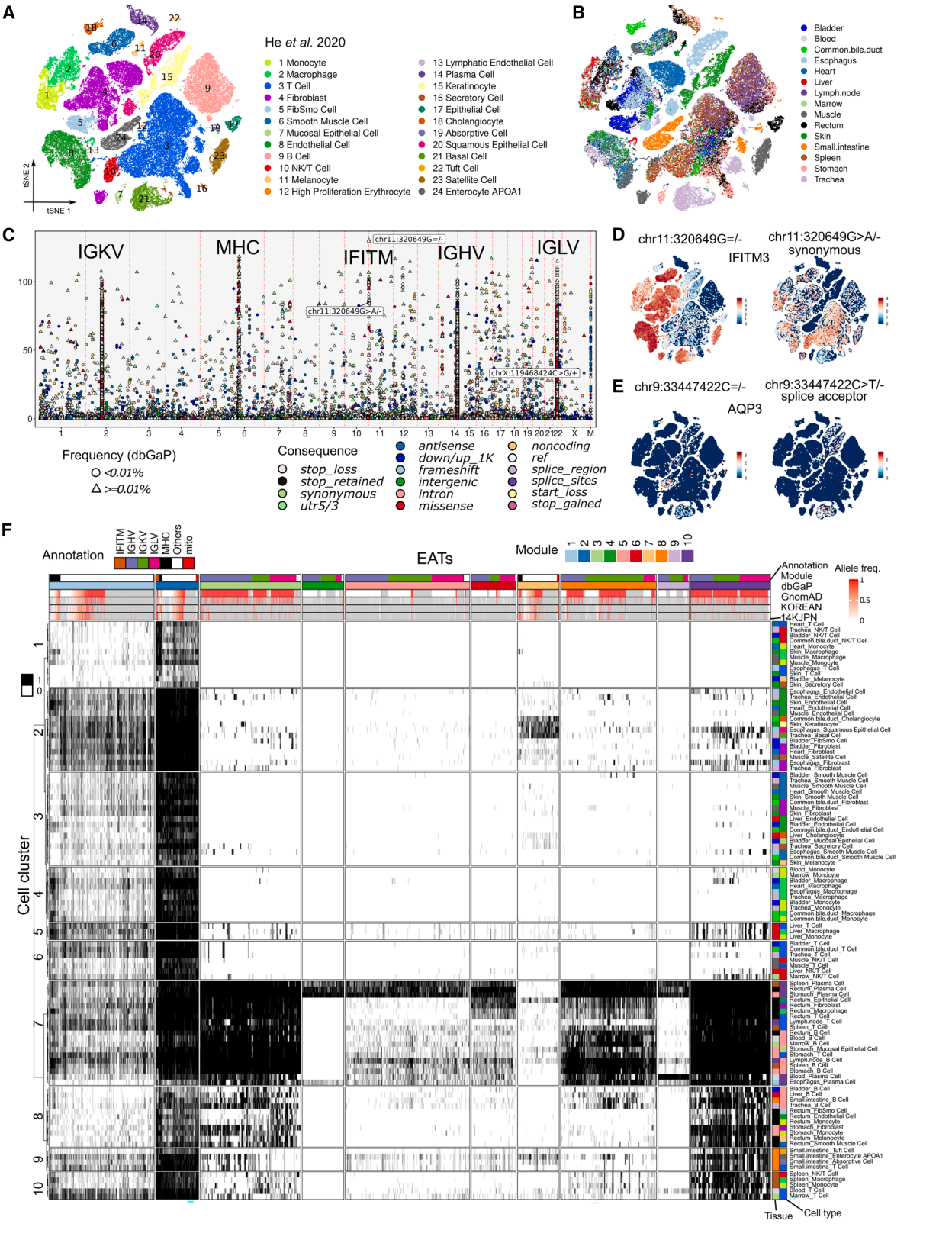
We explored allele-specific gene expression across 15 organs from an adult male donor, using 84,363 cells[66] (Figure 6A). Cell clusters and annotations were transferred from the original research.[66] For simplicity, we consolidated subclusters into major clusters, such as grouping all T cell subtypes under "T cells" (Figure 6B). From 176,883 loci, we identified 219,671 EATs, with 2,894 showing significant alternative expression (adjusted $p < 1e{-}10$) (Figure 6C). The Manhattan plot revealed six prominent peaks, including five in autosomal chromosomes and one in mitochondria, with four peaks associated with human immunoglobulin (Ig) heavy- and light-chain variants (IGHV, IGKV, IGLV) and the major histocompatibility complex (MHC) (Figures 6C and S7A). Notably, a peak at chromosome 11 p15.5 corresponded with interferon-induced transmembrane (IFITM) genes, which are known for their role in antiviral activities.[67] A closer inspection of this genomic region revealed that most EATs overlapped with *IFITM3* and *IFITM1* (Figure S7B). The ASE events were cell type specific, suggesting that these genes are likely to be coordinately expressed in *cis*. For instance, EAT chr11:230649 allele G is predominantly found in fibroblast, FibSmo cells, and some immune cells, whereas its counterpart, allele A, exhibited higher expression in T cells, and differential expression between monocytes and macrophages was observed (Figure 6D).

To systematically evaluate ASE genes across different cell clusters, we categorized EATs into two classes: class I and class II. For class I, both EATs at the same locus exhibited significant SD test results (adjusted $p < 1e{-}10$), indicating that both alleles display informative expression patterns but in different cell populations. This suggests bona fide ASE events. For class II, only one EAT at the locus had a significant p value, while its complementary EAT did not. This often indicates that the complementary allele is either randomly or sparsely expressed. These EATs are more likely to arise from somatic mutations or stochastic expression[19,68–71] in cells that typically involves independent expression from each allele. For example, EAT chrX:119468424G/+ was found across multiple tissues, while its binding EAT allele C showed sparse and lower levels of expression. Given males have only one X chromosome, allele C's presence across many tissues suggests that a somatic mutation likely occurring early in embryogenesis is tolerated by cells (Figure S7C).

We identified 5,726 significant EATs, of which 1,023 EATs from 508 loci were classified as class I, with 7 loci containing more than 2 EATs. The remaining EATs were categorized as class II. Most class I EATs were located within MHC/Ig regions, with 16 EATs derived from IFITM genes. Additionally, 62 EATs from 31 loci across 24 genes displayed cell-type-specific expression patterns (Figures S8 and S9A). None of these genes exhibited allelic exclusion but displayed allele-biased expression patterns. These genes have not previously been associated with known mechanisms, such as somatic rearrangements,[72] somatic mutations, olfactory gene regulation,[73] X chromosome inactivation,[74] and gene imprinting.[75] Among these genes, we discovered that *CD68*, a marker gene for macrophages and monocytes, was also expressed in fibroblasts, governed by different promotors.[76] In this gene, the EAT chr17:7579637 allele T was more likely expressed in fibroblasts than in myeloid cells, whereas allele G at the same locus was expressed in a different pattern (Figure S9A).

For another case, prediction of EAT consequences revealed that EAT chr9:33447422 is in the splice acceptor region of the *AQP3* gene. Allele T was absent in enterocytes, suggesting that cells carrying this allele were eliminated. Given that *AQP3*

(D) Heatmap representing the expression of prioritized EATs across all keratinocytes. Rows correspond to cells, clustered into five groups, while columns represent EATs.
(E) UMAP visualization of keratinocytes, color-coded by the cell clusters defined in the heatmap (D). Points are additionally shaped according to tissue type.
(F) Refined UMAP plot of keratinocytes, further highlighting clustering based on selected EAT expression. Points are color-coded by cell clusters and shaped according to tissue type.
(G) Heatmap of scaled gene marker expression across the defined cell clusters in (D).
(H) Violin plot showing *CCR1* gene expression across various cell types and tissues, separated by tissue type to highlight differences in expression patterns.



(legend on next page)

deficiency is known to impair enterocyte proliferation,[77] this result highlights the potential elimination of mutated cells, leading to ASE of *AQP3*. Additionally, beyond these allele-specific expressed genes, we observed mutations causing changes in the primary structure of proteins, which were expressed in subclones of cells compared to their corresponding genes overall (Figures S9B and S9C).

To assess the distribution of ASE genes and somatic mutations across tissues and cell types, we aggregated cells into pseudo-bulk samples, requiring a minimum of 50 cells per sample. After filtering, we analyzed 5,420 EATs across 103 samples, performing hierarchical clustering to identify 10 clusters of samples and 10 modules of EATs (Figure 6F). Notably, while modules 1, 2, and 7 contained significant EATs from Ig variants, modules 3 and 10 had high allele frequency in the dbGAP/ALFA database (www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/), suggesting reference allele expression. In contrast, somatic mutations in Ig genes were enriched in the other modules and predominantly expressed in plasma cells, indicating these mutations mostly occurred as B cells differentiate into plasma cells. Interestingly, fibroblasts from the rectum (cluster 7) and stomach (cluster 8) expressed EATs from module 2 and module 10, unlike those from the heart, skin (cluster 2), and muscle (cluster 3), suggesting organ-specific expression patterns for these alleles.

Performing SD analysis with EATs across different organs and cell types from the same donor provides valuable insights into the ASE of genes in normal tissues. Notably, 24 allele-specific expressed genes were identified in human cells. Differential promoter usage and phenotype consequence of genetic variants were found to contribute to ASE, expanding our understanding of ASE mechanisms. Furthermore, this work underscores the value of analyzing allele-biased expressed EATs across different cell types, combined with consequence predictions of genetic variants, to gain deeper insight into gene functions within specific cellular contexts. When combined with allele frequency databases, this method can also highlight informative somatic mutations and further facilitate lineage development analysis of cells in healthy tissues.

DISCUSSION

In this study, we extended spatial statistics to single-cell analysis by utilizing a cell-cell similarity matrix instead of a spatial distance matrix and addressed the spatial dissimilarity challenge by introducing a spatial measure to examine distinct spatial expression patterns between two features in single cells. Our approach focuses solely on the spatial autocorrelation of the test feature, leaving the binding feature unexamined. A permutation test is then used to determine the distribution shape and assess statistical significance. By permutating only the test feature's distribution, our method ensures robust evaluation.

The D score harmonizes two components. The first (L) captures the spatial autocorrelation of the test feature, designed to address the sparsity inherent in scRNA-seq by evaluating the spatial distribution of features rather than their absolute expression levels. The second ($1 - r$) evaluates the dissimilarity between the test and binding features, specifically overcoming challenges associated with low gene body coverage. Only features showing inverse expression patterns relative to their corresponding binding features are prioritized. Notably, our approach can highlight features with low expression levels if they exhibit spatial autocorrelation and distinct expression patterns compared to their binding features. This makes our method particularly suitable for biological contexts in biased scRNA-seq datasets, representing the most widely used technology in single-cell sequencing.

Furthermore, our method requires only feature counts and a cell-cell distance matrix as input, eliminating the need for prior cell clustering, which is a step often considered non-robust due to ambiguities in current cell cluster definitions.[78,79] The flexibility in defining the binding relationship between the test feature and the binding feature makes our SD test applicable to a wide range of studies beyond AS and ASE. For example, to investigate gene expression dynamics, the spliced and unspliced expressions of the same gene can be analyzed. Similarly, for studying natural antisense transcripts, the expression of antisense and sense strands of the same gene can be tested to explore the dynamics of antisense activity in single cells. Because our package, Yano, works directly with Seurat objects, different feature expressions can be read as "assays" within a Seurat object. This eliminates the need to regenerate reduction maps or reanalyze cell clustering for ongoing or published projects, avoiding unintended inconsistencies and making the method highly practical for diverse research applications.

A further advancement in our SD test is its integration with the Manhattan plot, which combines genome location data with *p* values. This allows us to identify genomic hotspot regions with heightened spatial expression differences. Additionally, we extend the application of the SD test to ASE analysis and cell lineage tracing using EATs, eliminating the need for prior read assembly. This approach leverages allele-specific EATs and somatic mutations as "natural barcodes" to trace

Figure 6. Spatial dissimilarity analysis of EATs in 15 adult human tissues reveals allele-specific expressed genes

(A) tSNE plot with cells color coded based on their tissue or organ of origin.

(B) tSNE visualization of cells from the adult human cell atlas (AHCA) project, color coded to indicate major cell types.

(C) Manhattan plot illustrating the results of the spatial dissimilarity test between EATs and other EATs at the same locus. Points are color coded by molecular consequences and shaped by AF from the dbGaP/ALFA database and sorted by genomic coordinates along the *x* axis. *p* values for each event were transformed using a negative log10 scale for visualization.

(D) tSNE visualization of the normalized expression level of EAT chr11:320649/– allele G (left) and allele A (right).

(E) tSNE visualization of the normalized expression level of EAT chr9:33447422/– allele C (left) and allele T (right).

(F) Heatmap of alternatively expressed EATs, with rows clustered into 10 clusters based on merged cells and columns clustered into 10 modules based on EATs. Expression values of EATs were merged by cell type and tissue, then binarized before clustering. Within each module, EATs are sorted by AF from dbGaP/ALFA databases, and within each cluster, cell groups are sorted by type. GnomAD, the Genome Aggregation database; KOREAN, Korean Reference Genome Project; 14KJPN, Japanese population reference panel.

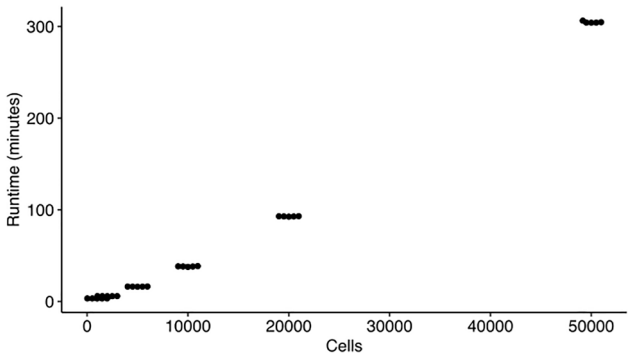


Figure 7. Runtime benchmark across a range of cell counts from the AHCA project, limited to 100,000 exons

Tests were performed on a server with 20 CPU cores and 40 threads.

developmental trajectories within human tissues. Besides the SD test, our tool also offers genetic variant annotation with various databases and functional prediction. By investigating pathogenic mutations in large-scale single-cell atlases, it is possible to understand new gene functions through naturally occurring somatic mutations or by conducting perturbation experiments.

Other directions for this work include integrating single-cell transcriptome analysis with other omics data to enhance our understanding of gene expression dynamics. In summary, while inspired by single-cell and spatial transcriptome analysis, the spatial dissimilarity test represents a versatile new approach with potential for expansion within spatial statistics.

Limitations of the study

While our SD test method offers significant innovations, it also has limitations. Its global application across all cells may overlook specific expression changes within particular cell groups (local events). Addressing this limitation requires reapplying the SD test to subclusters for a more detailed, cell-type-specific analysis. Another limitation is the computational intensity of the permutation test. To accelerate this process, we implemented core functions in pure C and utilized multithreading with the OpenMP library. A runtime benchmark, in which we downsampled cells from the AHCA project, is shown in Figure 7.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for the resources should be directed to and will be fulfilled by the lead contact, Karsten Kristiansen (kk@bio.ku.dk).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- This study analyzes published sequencing datasets and genome reference databases that are listed in the key resources table.
- Codes to generate the figures for this paper can be reached at https://github.com/shiquan/Yano_paper. The Yano software package for implementing spatial dissimilarity analysis is freely available at https://github.com/shiquan/Yano. Archival DOIs are listed in the key resources table.
- Any additional information required to re-analyze the results reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank Albin Sandelin, Rickard Sandberg, and Lin Lin for discussions and comments on the early versions of this work.

AUTHOR CONTRIBUTIONS

Q.S. and K.K. conceived the project. Q.S. developed the concept, implemented the code, performed the analyses, and drafted the manuscript. K.K. supervised the work and reviewed and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Feature annotation and counting with PISA
  - Implementation of the spatial dissimilarity test method
  - Visualization of single cell alignments with track plots
  - EAT calling
  - EAT annotation
  - Analysis of mouse primary motor neuron data
  - Cluster based alternative splicing analysis with DEXSeq and BRIE2
  - Analysis of squamous cell carcinoma data
  - Analysis of adult human cell atlas
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2025.101141.

Received: March 10, 2025
Revised: June 4, 2025
Accepted: July 24, 2025
Published: August 20, 2025

REFERENCES

1. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382. https://doi.org/10.1038/Nmeth.1315.
2. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.
3. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002.
4. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. Science *348*, aaa6090. https://doi.org/10.1126/science.aaa6090.
5. Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J.F., et al. (2019).

High-definition spatial transcriptomics for in situ tissue profiling. Nat. Methods *16*, 987–990. https://doi.org/10.1038/s41592-019-0548-y.

6. Stahl, P.L., Salmen, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82. https://doi.org/10.1126/science.aaf2403.
7. Almendro, V., Marusyk, A., and Polyak, K. (2013). Cellular Heterogeneity and Molecular Evolution in Cancer. Annu. Rev. Pathol. *8*, 277–302. https://doi.org/10.1146/annurev-pathol-020712-163923.
8. Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. *33*, 155–160. https://doi.org/10.1038/nbt.3102.
9. Armingol, E., Officer, A., Harismendy, O., and Lewis, N.E. (2021). Deciphering cell-cell interactions and communication from gene expression. Nat. Rev. Genet. *22*, 71–88. https://doi.org/10.1038/s41576-020-00292-x.
10. Kalsotra, A., and Cooper, T.A. (2011). Functional consequences of developmentally regulated alternative splicing. Nat. Rev. Genet. *12*, 715–729. https://doi.org/10.1038/nrg3052.
11. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476. https://doi.org/10.1038/nature07509.
12. Tian, B., and Manley, J.L. (2017). Alternative polyadenylation of mRNA precursors. Nat. Rev. Mol. Cell Biol. *18*, 18–30. https://doi.org/10.1038/nrm.2016.116.
13. Yelin, R., Dahary, D., Sorek, R., Levanon, E.Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., et al. (2003). Widespread occurrence of antisense transcription in the human genome. Nat. Biotechnol. *21*, 379–386. https://doi.org/10.1038/nbt808.
14. Balbin, O.A., Malik, R., Dhanasekaran, S.M., Prensner, J.R., Cao, X., Wu, Y.M., Robinson, D., Wang, R., Chen, G., Beer, D.G., et al. (2015). The landscape of antisense gene expression in human cancers. Genome Res. *25*, 1068–1079. https://doi.org/10.1101/gr.180596.114.
15. Forsberg, L.A., Gisselsson, D., and Dumanski, J.P. (2017). Mosaicism in health and disease - clones picking up speed. Nat. Rev. Genet. *18*, 128–142. https://doi.org/10.1038/nrg.2016.145.
16. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772. https://doi.org/10.1038/nature08872.
17. Chess, A. (2013). Random and Non-Random Monoallelic Expression. Neuropsychopharmacology *38*, 55–61. https://doi.org/10.1038/npp.2012.85.
18. Eckersley-Maslin, M.A., Thybert, D., Bergmann, J.H., Marioni, J.C., Flicek, P., and Spector, D.L. (2014). Random Monoallelic Gene Expression Increases upon Embryonic Stem Cell Differentiation. Dev. Cell *28*, 351–365. https://doi.org/10.1016/j.devcel.2014.01.017.
19. Reinius, B., and Sandberg, R. (2015). Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. Nat. Rev. Genet. *16*, 653–664. https://doi.org/10.1038/nrg3888.
20. Chess, A. (2016). Monoallelic Gene Expression in Mammals. Annu. Rev. Genet. *50*, 317–327. https://doi.org/10.1146/annurev-genet-120215-035120.
21. Knowles, D.A., Davis, J.R., Edgington, H., Raj, A., Favé, M.-J., Zhu, X., Potash, J.B., Weissman, M.M., Shi, J., Levinson, D.F., et al. (2017). Allele-specific expression reveals interactions between genetic variation and environment. Nat. Methods *14*, 699–702. https://doi.org/10.1038/nmeth.4298.
22. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515. https://doi.org/10.1038/nbt.1621.
23. Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. *22*, 2008–2017. https://doi.org/10.1101/gr.133744.111.
24. Shen, S., Park, J.W., Lu, Z.-X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. USA *111*, E5593–E5601. https://doi.org/10.1073/pnas.1419161111.
25. Song, Y., Botvinnik, O.B., Lovci, M.T., Kakaradov, B., Liu, P., Xu, J.L., and Yeo, G.W. (2017). Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation. Mol. Cell *67*, 148–161.e5. https://doi.org/10.1016/j.molcel.2017.06.003.
26. Lahnemann, D., Koster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biol. *21*, 31. https://doi.org/10.1186/s13059-020-1926-6.
27. Patrick, R., Humphreys, D.T., Janbandhu, V., Oshlack, A., Ho, J.W.K., Harvey, R.P., and Lo, K.K. (2020). Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. Genome Biol. *21*, 167. https://doi.org/10.1186/s13059-020-02071-7.
28. Wen, W.X., Mead, A.J., and Thongjuea, S. (2023). MARVEL: an integrated alternative splicing analysis platform for single-cell RNA sequencing data. Nucleic Acids Res. *51*, e29. https://doi.org/10.1093/nar/gkac1260.
29. Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554. https://doi.org/10.1038/s41587-019-0071-9.
30. Huang, Y., and Sanguinetti, G. (2021). BRIE2: computational identification of splicing phenotypes from single-cell transcriptomic experiments. Genome Biol. *22*, 251. https://doi.org/10.1186/s13059-021-02461-5.
31. Moses, L., and Pachter, L. (2022). Museum of spatial transcriptomics. Nat. Methods *19*, 534–546. https://doi.org/10.1038/s41592-022-01409-2.
32. Moran, P.A.P. (1950). Notes on Continuous Stochastic Phenomena. Biometrika *37*, 17–23. https://doi.org/10.1093/biomet/37.1-2.17.
33. Cang, Z., Zhao, Y., Almet, A.A., Stabell, A., Ramos, R., Plikus, M.V., Atwood, S.X., and Nie, Q. (2023). Screening cell–cell communication in spatial transcriptomics via collective optimal transport. Nat. Methods *20*, 218–228. https://doi.org/10.1038/s41592-022-01728-4.
34. Student (1914). The Elimination of Spurious Correlation due to Position in Time or Space. Biometrika *10*, 179.
35. Clifford, P., Richardson, S., and Hémon, D. (1989). Assessing the Significance of the Correlation between Two Spatial Processes. Biometrics *45*, 123–134. https://doi.org/10.2307/2532039.
36. Vallejos, R., and Osorio, F. (2014). Effective sample size of spatial process models. Spatial Statistics *9*, 66–92. https://doi.org/10.1016/j.spasta.2014.03.003.
37. Wartenberg, D. (1985). Multivariate Spatial Correlation: A Method for Exploratory Geographical Analysis. Geogr. Anal. *17*, 263–283. https://doi.org/10.1111/j.1538-4632.1985.tb00849.x.
38. Lee, S.-I. (2001). Developing a bivariate spatial association measure: An integration of Pearson's r and Moran's I. J. Geogr. Syst. *3*, 369–385. https://doi.org/10.1007/s101090100064.
39. Lee, S.-I. (2004). A Generalized Significance Testing Method for Global Measures of Spatial Association: An Extension of the Mantel Test. Environ. Plan. A. *36*, 1687–1703. https://doi.org/10.1068/a34143.
40. Chen, Y. (2015). A New Methodology of Spatial Cross-Correlation Analysis. PLoS One *10*, e0126158. https://doi.org/10.1371/journal.pone.0126158.
41. Miller, B.F., Bambah-Mukku, D., Dulac, C., Zhuang, X., and Fan, J. (2021). Characterizing spatial gene expression heterogeneity in spatially resolved

single-cell transcriptomic data with nonuniform cellular densities. Genome Res. *31*, 1843–1855. https://doi.org/10.1101/gr.271288.120.

42. DeTomaso, D., and Yosef, N. (2021). Hotspot identifies informative gene modules across modalities of single-cell genomics. Cell Syst. *12*, 446–456.e9. https://doi.org/10.1016/j.cels.2021.04.005.
43. DeTomaso, D., Jones, M.G., Subramaniam, M., Ashuach, T., Ye, C.J., and Yosef, N. (2019). Functional interpretation of single cell similarity maps. Nat. Commun. *10*, 4376. https://doi.org/10.1038/s41467-019-12235-0.
44. Svensson, V., Teichmann, S.A., and Stegle, O. (2018). SpatialDE: identification of spatially variable genes. Nat. Methods *15*, 343–346. https://doi.org/10.1038/nmeth.4636.
45. Edsgärd, D., Johnsson, P., and Sandberg, R. (2018). Identification of spatial expression trends in single-cell gene expression data. Nat. Methods *15*, 339–342. https://doi.org/10.1038/nmeth.4634.
46. Sun, S., Zhu, J., and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat. Methods *17*, 193–200. https://doi.org/10.1038/s41592-019-0701-7.
47. Shi, Q., Liu, S., Kristiansen, K., and Liu, L. (2022). The FASTQ plus format and PISA. Bioinformatics *38*, 4639–4642. https://doi.org/10.1093/bioinformatics/btac562.
48. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502. https://doi.org/10.1038/nbt.3192.
49. Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.A.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., et al. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell *162*, 184–197. https://doi.org/10.1016/j.cell.2015.05.047.
50. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0.
51. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B *57*, 289–300.
52. Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181. https://doi.org/10.1038/nprot.2014.006.
53. Yao, Z., Liu, H., Xie, F., Fischer, S., Adkins, R.S., Aldridge, A.I., Ament, S. A., Bartlett, A., Behrens, M.M., Van den Berge, K., et al. (2021). A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. Nature *598*, 103–110. https://doi.org/10.1038/s41586-021-03500-8.
54. Ji, A.L., Rubin, A.J., Thrane, K., Jiang, S., Reynolds, D.L., Meyers, R.M., Guo, M.G., George, B.M., Mollbrink, A., Bergenstråhle, J., et al. (2020). Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell *182*, 497–514.e22. https://doi.org/10.1016/j.cell.2020.05.039.
55. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M. A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.
56. Borcherding, N., and Brestoff, J.R. (2023). The power and potential of mitochondria transfer. Nature *623*, 283–291. https://doi.org/10.1038/s41586-023-06537-z.
57. L. Phan, Y.J., H. Zhang, W. Qiang, E. Shekhtman, D. Shao, D. Revoe, R. Villamarin, E. Ivanchenko, M. Kimura, et al. (2020). ALFA: Allele Frequency Aggregator. www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/.
58. Gudmundsson, S., Singer-Berk, M., Watts, N.A., Phu, W., Goodrich, J.K., Solomonson, M., Genome Aggregation Database Consortium; Rehm, H. L., MacArthur, D.G., and O'Donnell-Luria, A. (2022). Variant interpretation using population databases: Lessons from gnomAD. Hum. Mutat. *43*, 1012–1030. https://doi.org/10.1002/humu.24309.
59. Jung, K.S., Hong, K.W., Jo, H.Y., Choi, J., Ban, H.J., Cho, S.B., and Chung, M. (2020). KRGDB: the large-scale variant database of 1722 Koreans based on whole genome sequencing. Database *2020*, baz146. https://doi.org/10.1093/database/baaa030.
60. Tadaka, S., Katsuoka, F., Ueki, M., Kojima, K., Makino, S., Saito, S., Otsuki, A., Gocho, C., Sakurai-Yageta, M., Danjoh, I., et al. (2019). 3.5KJPNv2: an allele frequency panel of 3552 Japanese individuals including the X chromosome. Hum. Genome Var. *6*, 28. https://doi.org/10.1038/s41439-019-0059-5.
61. Landrum, M.J., Chitipiralla, S., Brown, G.R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., et al. (2020). ClinVar: improvements to accessing data. Nucleic Acids Res. *48*, D835–D844. https://doi.org/10.1093/nar/gkz972.
62. Zhang, Z., Zhu, H., Zhao, C., Liu, D., Luo, J., Ying, Y., and Zhong, Y. (2023). DDIT4 promotes malignancy of head and neck squamous cell carcinoma. Mol. Carcinog. *62*, 332–347. https://doi.org/10.1002/mc.23489.
63. Liu, C., Zhou, X., Zeng, H., Wu, D., and Liu, L. (2021). HILPDA Is a Prognostic Biomarker and Correlates With Macrophage Infiltration in Pan-Cancer. Front. Oncol. *11*, 597860. https://doi.org/10.3389/fonc.2021.597860.
64. Lu, P., Nakamoto, Y., Nemoto-Sasaki, Y., Fujii, C., Wang, H., Hashii, M., Ohmoto, Y., Kaneko, S., Kobayashi, K., and Mukaida, N. (2003). Potential interaction between CCR1 and its ligand, CCL3, induced by endogenously produced interleukin-1 in human hepatomas. Am. J. Pathol. *162*, 1249–1258. https://doi.org/10.1016/S0002-9440(10)63921-1.
65. Silva, T.A., Ribeiro, F.L.L., Oliveira-Neto, H.H.d., Watanabe, S., Alencar, R. D.C.G., Fukada, S.Y., Cunha, F.Q., Leles, C.R., Mendonça, E.F., and Batista, A.C. (2007). Dual role of CCL3/CCR1 in oral squamous cell carcinoma: Implications in tumor metastasis and local host defense. Oncol. Rep. *18*, 1107–1113.
66. He, S., Wang, L.H., Liu, Y., Li, Y.Q., Chen, H.T., Xu, J.H., Peng, W., Lin, G. W., Wei, P.P., Li, B., et al. (2020). Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. Genome Biol. *21*, 294. https://doi.org/10.1186/s13059-020-02210-0.
67. Bailey, C.C., Zhong, G., Huang, I.-C., and Farzan, M. (2014). IFITM-Family Proteins: The Cell's First Line of Antiviral Defense. Annu. Rev. Virol. *1*, 261–283. https://doi.org/10.1146/annurev-virology-031413-085537.
68. Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science *343*, 193–196. https://doi.org/10.1126/science.1245316.
69. Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic Gene Expression in a Single Cell. Science *297*, 1183–1186. https://doi.org/10.1126/science.1070919.
70. Cook, D.L., Gerber, A.N., and Tapscott, S.J. (1998). Modeling stochastic gene expression: Implications for haploinsufficiency. Proc. Natl. Acad. Sci. USA *95*, 15641–15646. https://doi.org/10.1073/pnas.95.26.15641.
71. McAdams, H.H., and Arkin, A. (1997). Stochastic mechanisms in gene expression. Proc. Natl. Acad. Sci. USA *94*, 814–819. https://doi.org/10.1073/pnas.94.3.814.
72. Mostoslavsky, R., Singh, N., Tenzen, T., Goldmit, M., Gabay, C., Elizur, S., Qi, P., Reubinoff, B.E., Chess, A., Cedar, H., and Bergman, Y. (2001). Asynchronous replication and allelic exclusion in the immune system. Nature *414*, 221–225. https://doi.org/10.1038/35102606.
73. Chess, A., Simon, I., Cedar, H., and Axel, R. (1994). Allelic inactivation regulates olfactory receptor gene expression. Cell *78*, 823–834. https://doi.org/10.1016/S0092-8674(94)90562-2.
74. Heard, E., and Disteche, C.M. (2006). Dosage compensation in mammals: fine-tuning the expression of the X chromosome. Genes Dev. *20*, 1848–1867. https://doi.org/10.1101/gad.1422906.
75. Reik, W., and Walter, J. (2001). Genomic imprinting: parental influence on the genome. Nat. Rev. Genet. *2*, 21–32. https://doi.org/10.1038/35047554.

76. Gottfried, E., Kunz-Schughart, L.A., Weber, A., Rehli, M., Peuker, A., Müller, A., Kastenberger, M., Brockhoff, G., Andreesen, R., and Kreutz, M. (2008). Expression of CD68 in non-myeloid cell types. Scand. J. Immunol. *67*, 453–463. https://doi.org/10.1111/j.1365-3083.2008.02091.x.

77. Thiagarajah, J.R., Zhao, D., and Verkman, A.S. (2007). Impaired enterocyte proliferation in aquaporin-3 deficiency in mouse models of colitis. Gut *56*, 1529–1535. https://doi.org/10.1136/gut.2006.104620.

78. Zeng, H. (2022). What is a cell type and how to define it? Cell *185*, 2739–2755. https://doi.org/10.1016/j.cell.2022.06.031.

79. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The Human Cell Atlas. eLife *6*, e27041. https://doi.org/10.7554/eLife.27041.

80. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

81. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595. https://doi.org/10.1093/bioinformatics/btp698.

82. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics *31*, 2032–2034. https://doi.org/10.1093/bioinformatics/btv098.

83. Cliff, A.D., and Ord, J.K. (1981). Spatial Processes : Models & Applications (Pion).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Mouse primary motor cortex datasets	Yao et al.[53]	RRID: SCR_016152
Cutaneous squamous cell carcinoma datasets	Ji et al.[54]	GSE144240 and GSE159929
The adult human cell atlas	He et al.[66]	GSE159929
Mouse genome reference and annotation (mm10)	10× Genomics resource website	refdata-gex-mm10-2020-A
Human genome reference and annotation (hg19)	10× Genomics resource website	refdata-cellranger-hg19–3.0.0
Human genome reference and annotation (GRCh38)	10× Genomics resource website	refdata-gex-GRCh38-2020-A
NCBI RefSeq annotation file for genetic variants annotation	UCSC goldenPath website	https://hgdownload.soe.ucsc.edu/goldenPath/archive/hg38/ncbiRefSeq/110/hg38.110.ncbiRefSeq.gtf.gz
Population-based allele frequency database (GRCh38)	NCBI/dbsnp website	https://ftp.ncbi.nih.gov/snp/latest_release/VCF/GCF_000001405.40.gz
Population-based allele frequency database (hg19)	NCBI/dbsnp website	https://ftp.ncbi.nih.gov/snp/latest_release/VCF/GCF_000001405.25.gz
Clinvar database (GRCh38)	NCBI/Clinvar website	https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/archive_2.0/2024/clinvar_20240317.vcf.gz
Clinvar database (hg19)	NCBI/Clinvar website	https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/archive_2.0/2024/clinvar_20240312.vcf.gz
Software and algorithms		
PISA	This paper and Shi et al.[47]	https://github.com/shiquan/PISA; https://doi.org/10.5281/zenodo.15857280
Yano	This paper	https://github.com/shiquan/Yano; https://doi.org/10.5281/zenodo.15857280
Seurat	CRAN and Satija et al.[48]	https://satijalab.org/seurat/
bcftools	Danecek et al.[80]	https://www.htslib.org/
samtools	Danecek et al.[80]	https://www.htslib.org/
Bwa	Li and Durbin[81]	https://github.com/lh3/bwa
sambamba	Tarasov et al.[82]	https://github.com/biod/sambamba
BRIE2	Huang and Sanguinetti[30]	https://github.com/huangyh09/brie
DEXSeq	Bioconductor and Anders et al.[23]	https://www.bioconductor.org/packages/release/bioc/html/DEXSeq.html
harmony	CRAN and Korsunsky et al.[50]	https://github.com/immunogenomics/harmony
scCustomize	CRAN	https://github.com/samuel-marsh/scCustomize
ggpubr	CRAN	https://cran.r-project.org/web/packages/ggpubr/index.html
RColorBrewer	CRAN	https://cran.r-project.org/web/packages/RColorBrewer/index.html
data.table	CRAN	https://cran.r-project.org/web/packages/data.table/index.html
ggvenn	CRAN	https://github.com/NicolasH2/ggvenn
ComplexHeatmap	Bioconductor	https://github.com/jokergoo/ComplexHeatmap

METHOD DETAILS

Feature annotation and counting with PISA

The feature annotation process begins with BAM files, using the enhanced functionality of the PISA[47] software. New parameters, '-exon' and '-psi', have been added to the PISA anno (v1.2). By default, the annotation process is restricted to reads that are entirely enclosed within exonic regions. Gene names are recorded under the GN tag, while exon names are documented under the EX tag. For reads spanning junctions or multiple exons, an additional JC tag is generated. Excluded exons are represented with the ER tag. The naming convention for exons, junctions, and excluded exons is consistent and follows the format: *chromosome:start-end/strand/gene.* In cases where strand specificity is not applicable, such as non-strand-sensitive libraries, the strand field is omitted. Notably, exon and exon-excluded reads share the same name but have distinct meanings: (1) Exon counts represent reads entirely enclosed within exons or junction reads overlapping the exon; (2) Exon-excluded counts represent junction reads that exclude the exon.

For SNV annotation, the '-ref-alt' parameter has been introduced in PISA anno (v1.2). By default, the annotation process is strand-sensitive; however, for non-strand-sensitive data (e.g., Smart-seq2), the '-is' parameter (short for "ignore strand") can be used. The name for EAT follows the format: "*chromosome:pos/reference allele > alternative allele/strand*" for SNVs or "*chromosome:pos/reference allele = /strand*" for reference allele, the strand field is omitted when '-is' set.

After feature annotation, the PISA count function quantifies UMI counts for each feature within individual cells. For example, gene expression counts can be obtained using the following command: "*PISA count -tag CB -umi UB -outdir exp -anno-tag GN anno.bam".* In this command, CB specifies the cell barcode; UB specifies the UMI; '-anno-tag' determines the feature type to count, such as genes (GN), exons (EX), junctions (JC), excluded exons (ER), or EATs (VR).

Implementation of the spatial dissimilarity test method

Our approach comprises a multi-step process to perform the spatial dissimilarity test of features within single cell space.

Step 1: Preparing the Seurat object. A Seurat object is prepared using gene expression counts, following standard preprocessing steps, including normalization, selection of HVGs, scaling of HVGs, and PCA analysis. Cell clustering is optional for this analysis but can usually be performed for visualization. Then, the weight matrix graph is built with 'GetWeights' function in our Yano package, by creating an SNN graph and then normalized column-wise, ensuring that the sum of each column is one.

Step 2: Incorporating additional features. Next, we incorporate other features (e.g., exons or junctions) as separate assays within the Seurat object using the 'CreateAssayObject' function. We then switch the default assay to the test assay. By executing 'RunAutoCorr' from Yano, Moran's *I* is calculated for the test features with the weight matrix. The Moran's *I* is defined as

$$I = \frac{N}{W} \frac{\sum_{i}^{n} \sum_{j}^{n} W_{ij} X_i X_j}{\sum_{i}^{n} X_i^2},$$

where *N* is the cell number, and *W* represents the sum of all values in the weight matrix. Consider that the sum of each row is always equal to 1, and $\frac{N}{W}$ always should be 1, then, the formula simplifies to

$$I = \frac{\sum_{i}^{n} \sum_{j}^{n} W_{ij} X_i X_j}{\sum_{i}^{n} X_i^2},$$

with *p* values derived from the Z scores calculated by the standard deviation of *I*.[83] *p* values are then adjusted with the Benjamini-Hochberg procedure.[51] Features with significant spatial autocorrelation are identified using function 'SetAutoCorrFeatures' in Yano. The default cutoff for spatial autocorrelated features is Moran's $I > 0$ and adjusted *p* value <0.01.

Step 3: Feature annotation and spatial dissimilarity testing. Specific functions from Yano, like 'ParseExonName' and 'annoVAR', are designed to parse the genomic coordinates and annotate various feature types from the feature names which generated with PISA. The 'RunBlockCorr' function conducts spatial dissimilarity tests on selected features with binding assay. This step involves setting binding relationships between features and calculating the D score, which combines spatial smoothing and dissimilarity measures:

$$D_{XY} = \sqrt{L_X} \times (1 - r_{\tilde{x}\tilde{y}}),$$

where *X* is the test feature, *Y* is the binding feature. L_X is a spatial smooth scalar of test feature, defined as

$$L_X = \frac{\sum_i (\tilde{x}_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}.$$

L_X is used to describe spatial association degree.[38] $\tilde{X}$ is the spatial lag value of X by weight matrix W. The 1- Pearson correlation coefficient r describe the dissimilarity between spatial lag of X and Y. Therefore, the D score integrates the spatial smoothing degree of test feature and the dissimilarity between test feature and the binding feature. Yano then permutates the order of X and calculate the D score 100 times to evaluate the expected value and variance, and estimate significance, with p values calculated by one-tail t-test via the 'pt' function from the R stats package. p values were automatically adjusted using the Benjamini-Hochberg procedure, implemented with the 'p.adjust' function from the R stats package. Finally, the 'FbtPlot' function visualizes these p values on a Manhattan plot.

Visualization of single cell alignments with track plots

To elucidate the distribution of single-cell alignments across genomic regions, including junction reads, we have implemented a track plot visualization tool, 'TrackPlot', within Yano. 'TrackPlot' provides an intuitive representation of read coverage across different cell groups, leveraging UMI depth as a metric. The process begins with the preparation of a sorted and indexed BAM file. The track plot function then computes the UMI depth at each genomic coordinate. To enhance visualization performance and achieve a smoother appearance of the tracks, the target region is segmented into 1,000 bins. The average UMI depth across these bins is calculated and utilized for the plot.

To account for cell number variations in group sizes, the UMI depth for each cell group is normalized to the number of cells in the group, resulting in a normalized UMI depth per cell. This step ensures that the tracks are directly comparable between groups, allowing for a clear visual comparison of alignment distributions. But if cell groups are not specified, the track plot will generate raw UMI depth for all cells. By default, 'TrackPlot' distinguishes the strand of reads during visualization. UMI depth for the reverse strand is adjusted by applying a negative conversion, and reads are color-coded by strand, with forward strand reads displayed in red and reverse strand reads in blue. When the UMI tag is not set, 'TrackPlot' defaults to using raw read depth instead of UMI depth for visualization. If the strand parameter is set to 'ignore', reads from both strands are summed up, providing a combined view without strand differentiation.

For analyses involving multiple sequencing runs or datasets, Yano supports the input of a list of BAM files, allowing for the concurrent visualization of alignment tracks from several samples or conditions. This feature is particularly useful for comparing and contrasting alignment patterns across different sequence libraries, samples or experimental groups.

EAT calling

For the sorted and indexed BAM files, bcftools[80] is employed for genetic variant calling. Given the intrinsic non-strand-specific nature of DNA variant callers, the standard filtering methods typically used for DNA may not be suitable for RNA datasets due to potential strand biases, particularly in the context of ASE. To address this, the unfiltered, raw variants are used during the EAT annotation phase. The annotation of all detected alleles at a given locus with '-ref-alt' parameter. Additionally, PISA anno enables the strand-sensitive annotation of EATs in default. For paired-end sequencing data, the strand information for EATs from the second read is inverted to accurately reflect the original cDNA fragment's strand.

EAT annotation

The 'annoVAR' function in the Yano package is designed to annotate EATs using various databases and predict their molecular consequences based on gene annotation files. This function accepts essential inputs such as EAT locations, reference and alternative alleles, and strand information. When using databases in VCF format, the tags parameter must be specified to extract desired fields, adhering strictly to VCF specifications (VCFv4.2, https://samtools.github.io/hts-specs/VCFv4.2.pdf). For instance, records with a 'Number' type of A, which stores allele-specific information separated by commas, the function extracts data exclusively for the alternative allele.

For molecular consequence predictions, the function requires a GTF annotation file to define gene and transcript boundaries and an indexed reference genome in FASTA format. Unlike DNA variant annotations, 'annoVAR' accounts for strand-sensitive RNA reads, as different transcripts or genes on opposite strands can produce distinct consequences for the same allele. The function predicts a consequence for each overlapping transcript and ranks them by impact, exporting only the most significant annotation and gene name for simplicity in downstream analysis. The ranking prioritizes coding genes over non-coding and antisense genes for reference alleles, while for non-reference alleles, a detailed ranking system places highly impactful consequences. Consequences are ranked in the following order: *whole_gene* > *exon_loss* > *stop_gained* > *exon_splice_sites* > *splice_donor* > *plice_acceptor* > *frameshift_truncation* > *frameshift_elongation* > *stop_loss* > *start_loss* > *inframe_indel* > *missense*>*synonymous*>*stop_retained*>*start_retained*>*utr5*>*utr3*>*splice_region*>*intron*>*utr5_intron*>*utr3_intron*>*noncoding_exon*>*noncoding_splice_region*>*noncoding_intron*>*upstream_1K*>*downstream_1K*>*antisense_utr3*>*antisense_utr5*>*antisense_exon*>*antisense_intron* > *antisense_upstream_1K* > *antisense_downstream_1K* > *intergenic*. Moreover, if a chromosome is missing in the annotation file, the consequence is marked as *unknown*. For non-strand-sensitive EATs, the antisense records will not be generated.

Analysis of mouse primary motor neuron data

The raw FASTQ files for the Smart-seq2 data were downloaded from NeMO Archive. Reads for each cell were aligned to the mm10 reference genome using STAR (v2.7.1a), with the RG tag added during alignment. The resulting BAM files were sorted and merged using sambamba (v0.8.0). The sorted and indexed BAM files were subsequently used for feature annotation and counting. Feature annotation was performed with PISA (v1.2), using the '-exon', '-psi', and '-is' parameters. A GTF file (gencode vM23) downloaded from the 10X Genomics website was used to annotate genes and exons. Non-strand-sensitive features were counted with PISA count, with the '-cb' parameter set to RG. No UMI tag was used for Smart-seq2 data.

The resulting gene count file was loaded into the R environment using the 'ReadPISA' function from the Yano package. The loaded count matrix was converted into a Seurat object and normalized using a log-scaled approach, followed by adjustment for library size, as implemented in the Seurat (v5.1.0) pipeline. Default parameters and the top 20 PCs were used for cell clustering and dimensional reduction analysis.

Counts for exons, junctions, and exon-excluded reads were loaded into the Seurat object as new assays using the 'CreateAssayObject' function from the Seurat package. Features detected in fewer than 20 cells were excluded. Raw counts were normalized using Seurat's 'NormalizeData' function. The top 20 PCs calculated with the gene assay were used to construct a cell-cell similarity matrix, which was subsequently employed to calculate Moran's *I* and perform spatial dissimilarity analysis follow procedures of *"Implementation of spatial dissimilarity test method"*. For alternative expression analysis with exon and junction assays, Mode 1 is used for comparison with gene assay. While Mode 1 and Mode 3 are used when comparison between exon assay and exon-excluded assay.

As most published scRNA-seq datasets are generated with strand-sensitive kits, both PISA and Yano enable strand-sensitive processing by default. Track plot visualization with the 'TrackPlot' function by setting the UB parameter to *NULL*, CB parameter set to RG and the strand parameter was set to 'ignored' to accommodate Smart-seq2 data.

The aligned BAM files for the 10X Genomics v3 data were directly downloaded from NeMO Archive. Feature annotation was performed with PISA (v1.2), using the '-exon' and '-psi' parameters. During feature counting, the CB tag was used to identify cell barcodes, and the UB tag was used to count UMIs. Post-processing for the 10X data was carried out using the same Seurat and Yano workflow as described for Smart-seq2 data. During visualization of track plots for 10X data, default parameters were used.

Cluster based alternative splicing analysis with DEXSeq and BRIE2

To run DEXSeq on selected cells, we implemented the 'RunDEXSeq' function in the Yano package. The process involves generating pseudo-bulk groups from single-cell raw feature counts for each cell group. For each comparison, three pseudo-bulk samples are generated for each group by aggregating raw feature counts from the single cells. The group 1, specified with 'ident.1', is the primary cell group for comparison. By default, all cells not in group 1 are aggregated into group 2 unless explicitly specified with 'ident.2'. A count matrix of the test assay with six pseudo-samples (three for each group) is prepared as the input for DEXSeq. Simultaneously, for the binding assay, six pseudo-samples from the same cells are generated from the raw counts of the binding assay and used as 'alternativeCountData' in DEXSeq. DEXSeq is run using a single thread, which makes the process slower when analyzing all features. To optimize runtime and avoid potential false positive events, only spatially autocorrelated features are input for benchmarking purposes.

For the BRIE2 pipeline, the PISA pick tool is used to split the merged BAM file into a one-cell-per-file format, as this is the required input format for brie-count. The brie-count and brie-quant pipelines were executed according to the standard protocols provided in the BRIE2 manual (https://brie.readthedocs.io/).

Analysis of squamous cell carcinoma data

Aligned BAM files were downloaded from Gene Expression Omnibus (GEO) datasets, adopting filtered cells and annotations directly from the original studies. Due to discrepancies in the versions of the human reference and mitochondrial genomes used in WES and scRNA-seq datasets, WES reads were realigned to the hg19 human reference genome using BWA.[81] This was followed by the repair of paired-end information and marking of duplicate reads using samtools.[80]

Genetic variants in both tumor and normal libraries were identified using bcftools.[80] This process resulted in four variant files in VCF format, two each from scRNA-seq and WES datasets. These files were merged, and indels were normalized using bcftools. The merged and normalized VCF file was used to annotate the BAM files employing PISA. The quantified gene expression data served as the basis for cell clustering, executed via the Seurat package employing its default parameters. The EAT counts read with 'ReadPISA' and load into Seurat as a new assay. Features detected in fewer than 10 cells were excluded from downstream analysis to maintain data quality and relevance. We then performed spatial dissimilarity analysis follow procedures of *"Implementation of spatial dissimilarity test method".*

Analysis of adult human cell atlas

Raw FASTQ files were downloaded and processed using the CellRanger pipeline (version 6.1.2) with 'SC5P-PE' chemistry. This step produced BAM files, which were then utilized for subsequent annotation and feature counting phases. To maintain consistency with previous findings and leverage established insights, filtered cells, cell annotations, and tSNE maps were directly adopted from the original study.[66] Genetic variants for each tissue sample following procedure of section *"EAT calling".* To ensure the robustness of

downstream analysis, features detected in fewer than 20 cells were filtered out. The gene expression data was employed for PCA analysis using the Seurat package, adhering to its default settings. EAT counts were read and loaded as a new assay for Seurat object. The top 20 PCs built by gene expression were then used to build cell-cell weight matrix. The locus names and positions of EATs were parsed with 'ParseVarName' function. The UCSC RefSeq gene annotation for GRCh38 was used to predict the consequences of EATs with 'annoVAR'.

QUANTIFICATION AND STATISTICAL ANALYSIS

This study reports the development of a new statistical measures with details described in the 'method details' section.