nature communications

Article

https://doi.org/10.1038/s41467-026-70652-4

Systematic background selection with BasCoD enhances contrastive dimension reduction in single cell genomics

Received: 30 July 2025

Accepted: 1 March 2026

Published online: 17 March 2026

Check for updates

Kwangmoon Park[1], Zhongxuan Sun[2], Ruiqi Liao[3], Emery H. Bresnick[3] & Sündüz Keleş[2,4]

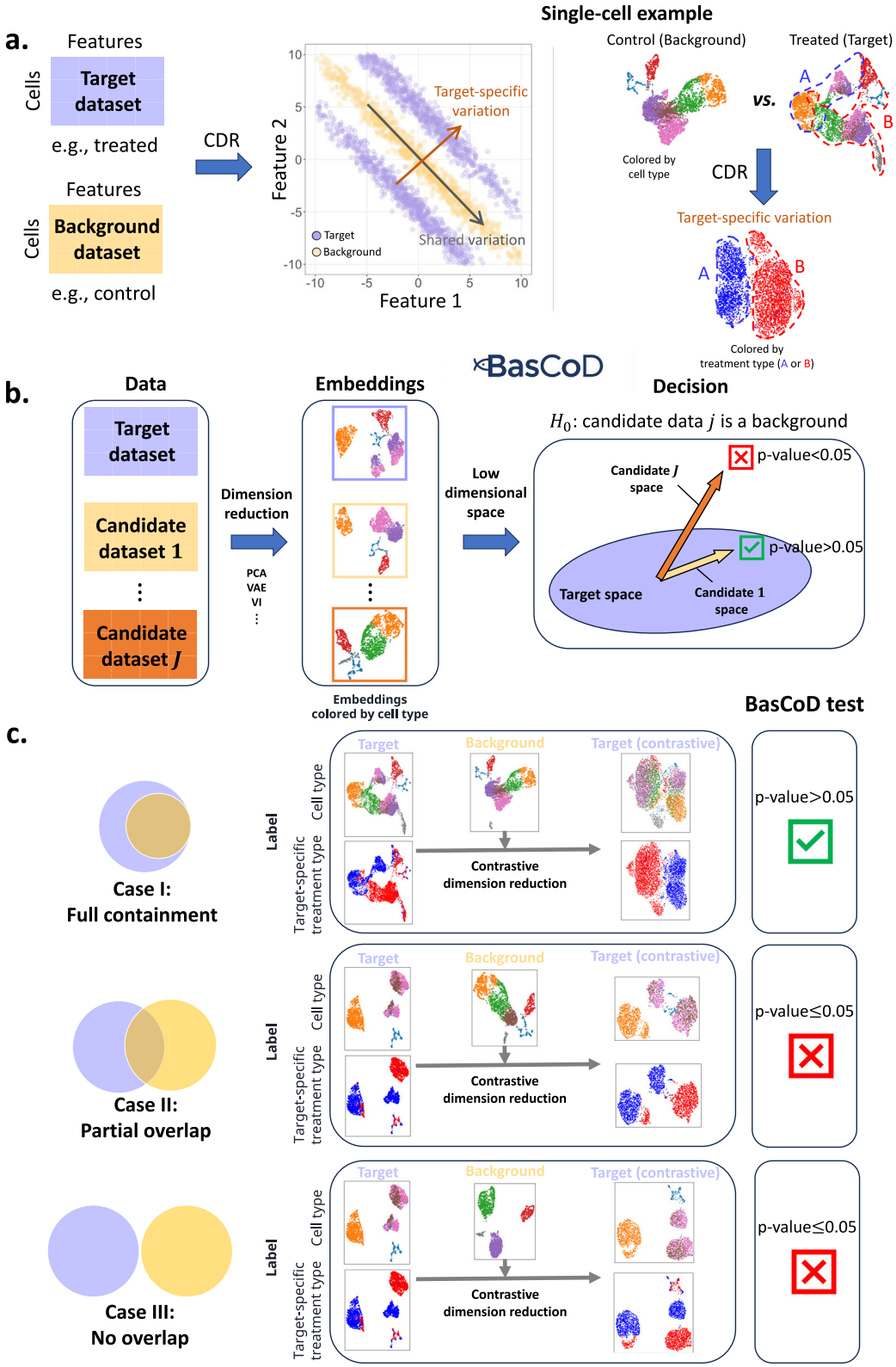
In single-cell experiments spanning diverse conditions, distinguishing variation specific to one condition (e.g., treatment) from shared or background variation (e.g., control) is critical for uncovering treatment-specific molecular responses. However, these studies typically yield ultra-high-dimensional data, necessitating effective dimension reduction for reliable biological interpretation. Contrastive dimension reduction methods address this challenge by identifying low-dimensional features enriched in a target dataset relative to a background dataset that captures shared variation. Despite their growing utility, the success of such methods critically depends on the choice of background, yet no formal criterion exists for evaluating or selecting backgrounds. To address this gap, we introduce BasCoD, a statistical testing framework based on spectral subspace inclusion theory, that enables rigorous evaluation and systematic selection of background datasets. Applying BasCoD across a range of single-cell datasets, we show that it effectively identifies suitable backgrounds, substantially improving the contrast and interpretability of the resulting target representations. We further demonstrate how BasCoD can guide the design of contrastive analyses in large-scale single-cell experiments conducted under heterogeneous conditions and elucidate potential interaction effects in perturbation studies.

Advances in high-throughput sequencing yield massive genomic datasets across diverse conditions (e.g., treatment *vs.* control). For example, population-scale scRNA-seq[1,2] or Perturb-seq[3] experiments provide high dimensional features that can be compared across different treatments or perturbations. However, their ultra-high dimensionality requires effective dimension reduction to uncover trustworthy biological insights. In addition, to better contrast high-dimensional datasets from different conditions, a systematic approach is needed to extract low-dimensional features that are specific to the target dataset in comparison to background datasets.

Contrastive dimension reduction (CDR) methods[4–9] identify variation specific to a target dataset (e.g., treatment) by contrasting its low-dimensional representation with that of a background dataset (e.g., control) (Fig. 1a). These approaches enhance contrast by effectively filtering out variation shared between the target and background datasets. We note that, while related, CDR is conceptually distinct from contrastive learning in the machine learning literature (e.g., sCIN[10], scContrast[11]), where the goal is to generate synthetic pairs through data augmentation[12,13]. CDR methods have been shown to perform well across a variety of high-dimensional domains including genomics,

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Biostatistics and Medical Informatics, University of Wisconsin - Madison, Madison, WI, USA. [3]Wisconsin Blood Cancer Research Institute, Department of Cell and Regenerative Biology, Carbone Cancer Center, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA. [4]Department of Statistics, University of Wisconsin - Madison, Madison, WI, USA. e-mail: keles@stat.wisc.edu

https://doi.org/10.1038/s41467-026-70652-4



proteomics, and pattern recognition by extracting target-specific low-dimensional features. In the context of single-cell genomics, particularly in treatment-versus-control comparisons, CDR has proven effective in isolating biologically meaningful variation specific to the treated cell population[4,5,7,8]. For example, when dimension reduction applied to a target single-cell dataset, such as one involving a specific treatment, fails to clearly reveal treatment-related variation, contrastive dimension reduction using a background dataset that captures shared cell-type variation can amplify treatment-specific variation in the resulting contrastive low-dimensional embeddings (Fig. 1a).

Despite the increasing utility of CDR methods, the critical issue of background dataset selection remains largely unaddressed. Current applications typically analyze only a small number of fixed background datasets without assessing their suitability. As a result, the influence of background choice on the quality and interpretability of contrastive

Fig. 1 | A schematic overview of BasCoD for background dataset selection in contrastive dimension reduction. a A general overview of contrastive dimension reduction. A target dataset of interest is contrasted against a background dataset (left panel) in a low-dimensional space. Contrastive dimension reduction isolates the target data-specific variation (brown arrow in the middle panel). When applied to single cell datasets from an experiment comparing treatment and control groups, contrastive dimension reduction captures variations of cells unique to target, such as treatment type or infection status, that is present only in the target dataset (right panel). **b** Overview of BasCoD. A target dataset of interest and a collection of J candidate background datasets are considered as inputs (Left). BasCoD utilizes these datasets along with their low-dimensional embeddings derived from a user-selected dimension reduction method (e.g., PCA, VAE, or VI; Middle). It then systematically compares the embedding subspaces of the target dataset to those of each candidate background (Right). Specifically, BasCoD evaluates the null hypothesis $H_{0,j}$: the embedding subspace of candidate dataset j is fully contained within that of the target dataset, for each $j = 1, \cdots, J$. **c** An illustration of typical outcomes when applying BasCoD. Case I (top row) represents an ideal scenario where the candidate background's embedding subspace (yellow region) is fully included within that of the target (blue region). For this setting, contrastive dimension reduction effectively highlights target-specific variation (e.g., experimental condition or treatment type) rather than shared variation (e.g., cell type), and BasCoD appropriately accepts the null hypothesis (green checkmark). In contrast, Case II and Case III depict scenarios where the candidate backgrounds possess distinct, salient low-dimensional structures differing from the target. In these cases, contrastive dimension reduction is unlikely to enhance target-specific signals, and consequently, BasCoD rejects the null hypothesis (red X mark).

analysis remains poorly understood and poses a key limitation to the broader and more principled application of CDR.

To address this critical gap, we introduce BasCoD (Background Selection for Contrastive Dimension Reduction), a principled framework for evaluating the suitability of background datasets in contrastive analysis. BasCoD is grounded in the central premise that a valid background should not contain salient structures that are unique or dominant relative to those in the target dataset. This assumption underlies a broad class of CDR methods[6–9]. BasCoD formalizes this principle through a statistical testing framework that compares low-dimensional representations of the target and background datasets, derived from a user-specified dimension reduction method (Fig. 1). This yields a goodness-of-fit test for CDR, providing an interpretable decision rule that empirically aligns with the quality of contrastive representations observed in practice. Beyond serving as a diagnostic tool, BasCoD offers a systematic approach for identifying or constructing appropriate background datasets in large-scale, high-dimensional single-cell experiments under heterogeneous conditions. By ensuring that background-driven artifacts do not confound contrastive signals, BasCoD improves the reliability of downstream analyses, such as gene ontology and gene set enrichment analysis, for the target dataset. We demonstrate the utility of BasCoD across previously published CDR analyses and new applications, including single-cell trajectory analysis, large-scale single-cell experiments under heterogeneous conditions, and Perturb-seq studies.

Results

The BasCoD testing framework

BasCoD leverages spectral subspace inclusion theory to evaluate whether a candidate dataset is appropriate as a background for contrastive dimension reduction of a target dataset. In typical applications involving high-dimensional single-cell data, a dimension reduction method such as principal component analysis (PCA) or a variational autoencoder (VAE) is first applied to both the target and candidate background datasets to obtain low-dimensional embeddings (Fig. 1b, first and second columns). From these embeddings, BasCoD constructs corresponding subspaces (Methods, Fig. 1b, third column) and compares them to assess the relationship between the target and each candidate background. Effective contrastive analysis requires the target to exhibit a richer or more complex low-dimensional structure than the background. BasCoD quantifies this complexity by determining the dimensionality of the target representation and testing whether the candidate background's representation lies within the span of the target representation (Fig. 1b, decision component). A test statistic based on spectral projections is then used to formally assess this inclusion relationship (Methods, Algorithm 1). When a candidate background contains salient structures distinct from those in the target (Candidate background J in Fig. 1b), BasCoD returns a small p-value, indicating it is unsuitable for contrastive analysis (Cases II and III in Fig. 1c). Conversely, backgrounds lacking such additional structure yield large p-values (Candidate background 1 in Fig. 1b), supporting their use as valid contrasts (Case I in Fig. 1c). Simulations based on data generative model of BasCoD confirm that p-values from this framework are well calibrated under the null hypothesis (Fig. S1a–c, Supplementary Notes) and that the method maintains high power and controlled false discovery rate across a range of subspace perturbation scenarios that reflect varying difficulty in background selection (Fig. S1d, Supplementary Notes). Furthermore, a biologically informed simulated single-cell dataset generated using a Negative Binomial model[14] illustrates that BasCoD consistently rejects candidate background datasets exhibiting strong background-specific signals (Fig. S2, Supplementary Notes).

BasCoD as a goodness-of-fit test for contrastive dimension reduction analyses

We first demonstrated BasCoD's capability to provide a rigorous goodness-of-fit test for contrastive dimension reduction analyses. We examined a mouse protein expression dataset[15,16] previously analyzed in the cPCA study[4], where mice given shock therapy (target group) were contrasted against control mice (background group). Within the target group, Ts65Dn mice with Down Syndrome relevant neurological phenotypic features (DS) and control (non-DS) mice were nearly indistinguishable using standard PCA (Adjusted Rand Index (ARI) score = 0.01, Normalized Mutual Information (NMI) score = 0.01, average Silhouette score = 0.06, Fig. 2a). However, when the control mice were used as background in a contrastive PCA, a clear separation emerged between the DS and non-DS mice (ARI score = 0.78, NMI score = 0.43, average Silhouette score = 0.74, Fig. 2a). As emphasized by Abid et al.[4], the control mice capture natural variation in the absence of treatment-induced effects, i.e., a favorable situation essential for a valid background in contrastive dimension reduction. BasCoD directly quantifies the rarity of this favorable situation, and its application to this dataset yields a p-value of 0.16, thereby supporting the validity of the contrastive dimension reduction for the target dataset with the chosen background dataset.

Next, to illustrate the behavior of BasCoD under a relatively inappropriate background choice, we analyzed population-scale scRNA-seq data of human induced pluripotent stem cells (iPSCs) differentiating toward a midbrain neural fate, measured at Day 11 (D11), Day 30 (D30), and Day 52 (D52)[1]. At D52, a subset of cells was exposed to rotenone-induced oxidative stress (ROT), while the remaining cells served as untreated controls. Previously, Jerber et al. demonstrated that cell type compositions at D11 differ substantially from those at D52. Based on this, we defined the ROT-treated cells at D52 as the target dataset and considered D11 cells as a candidate background. Notably, Jerber et al. also introduced a *Differentiation Efficiency (DE)* metric for each of the 213 donor iPSC lines, quantifying the proportion of dopaminergic and serotonergic-like neurons at D52, to capture donor-specific variation in differentiation. When PCA was applied to the target dataset alone (Fig. 2b), cells from donors with high and low DE were intermingled ($R^2 = 0.26$ in a regression of DE on the first two PCs). Applying cPCA with D11 cells as the background yielded nearly

https://doi.org/10.1038/s41467-026-70652-4

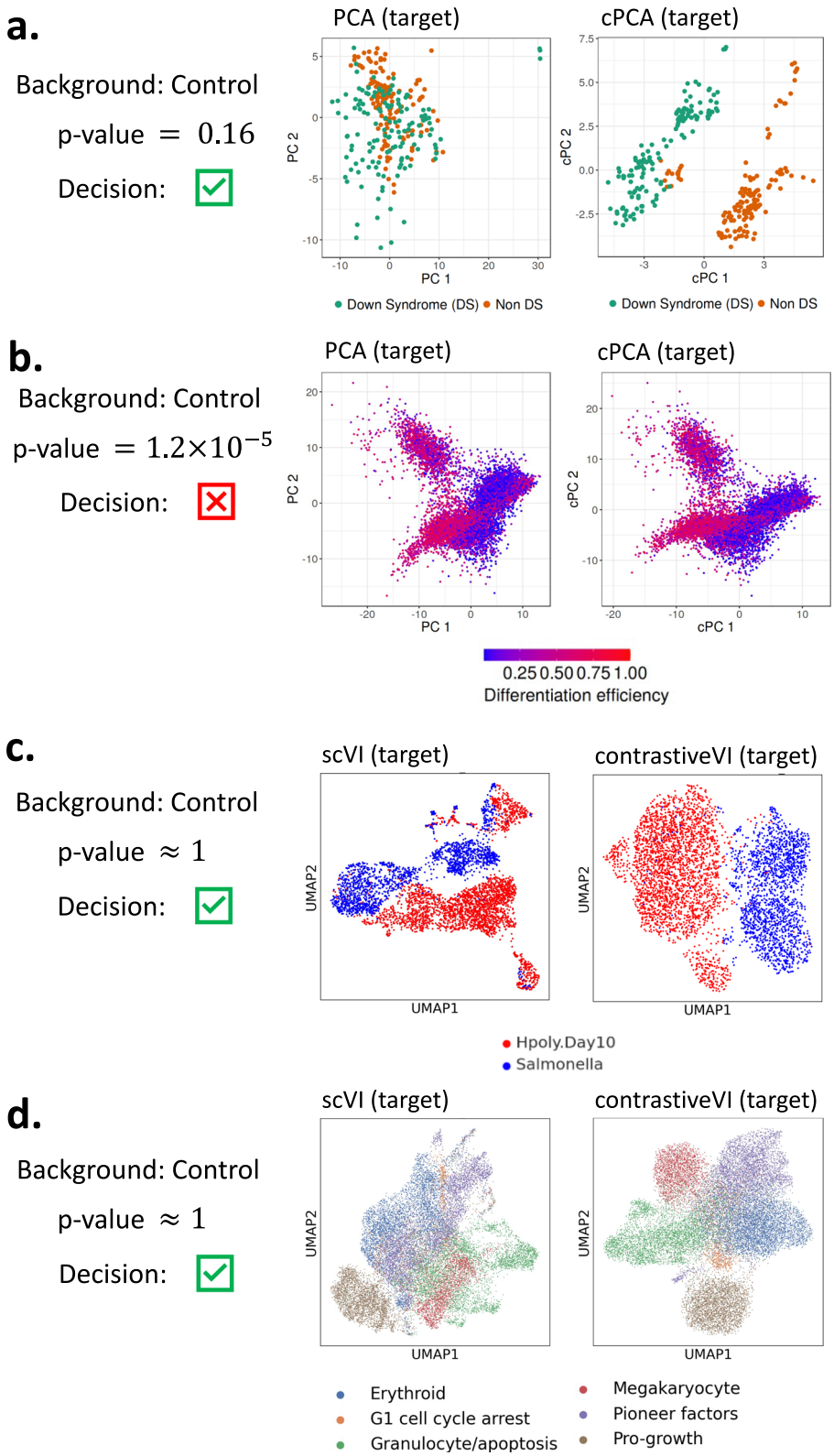


Fig. 2 | BasCoD as a goodness-of-fit test for previously reported results from the contrastive dimension reduction literature. a Left: PCA coordinates of the protein expression data of mice that received shock therapy (target dataset) colored by Down Syndrome label. Right: Same figure with cPCA coordinates[4] of the target dataset against a group of mice that did not receive shock therapy (background dataset). BasCoD test for the control group mice provides p-value = 0.16, which indicates that the background is valid (green check mark). **b** Left: UMAP coordinates of the scVI[18] cell embeddings obtained from the perturbed cells (target dataset) of Norman et al. (2019)[17] colored by gene program label. Right: Same figure with contrastiveVI[7] salient embeddings, where the target dataset was contrasted against the non-perturbed control cells (background dataset). BasCoD test for the control K562 cells provides p-value ≈ 1. **c** Left: UMAP coordinates of the scVI embeddings of the entire mouse intestinal cells[32] that received pathogen treatments (target dataset). Right: Same figure with contrastiveVI[7] salient embeddings, where the target dataset was contrasted against the control cells without pathogen treatment (background dataset). The right panels (contrastiveVI) of **c**, **d** were generated using the analysis pipeline implemented in the GitHub repository associated with Weinberger et al. (2023)[7]. Source data are provided as a Source Data file.

identical embeddings, with no meaningful improvement in separation by DE (R^2 = 0.28). This suggests that D11 cells do not serve as an effective background for contrastive analysis. Consistent with this observation, BasCoD returned a p-value of 1.2×10^{-5}, formally rejecting the suitability of D11 as a valid background for contrastive dimension reduction.

Having confirmed that BasCoD performs reliably in the linear contrastive dimension reduction setting, we extended its application to a state-of-the-art nonlinear approach, contrastiveVI[7], to demonstrate its broader applicability. Utilizing a subset of the large-scale Perturb-seq dataset from Norman et al.[17] (filtered as in Weinberger et al.), we designated perturbed cells as the target and control cells as the candidate background. Notably, embeddings derived from the standard, non-contrastive scVI[18] exhibited clear associations with cell cycle phase labels in both datasets (Fig. S3a), whereas the target dataset additionally revealed subtle separation according to gene-program labels (Fig. 2b). We subsequently employed BasCoD to rigorously assess whether the low-dimensional loadings of the candidate background, calculated via eqn.(4) (Methods), were encompassed within the subspace spanned by those of the target dataset. BasCoD yielded a p-value of ≈ 1, confirming the suitability of control cells as an appropriate background. As previously reported by Weinberger et al.[7], contrastiveVI generated salient target-specific embeddings clearly separated according to gene-program labels (Fig. 2b). We observed similar results in an analysis of a mouse intestinal single-cell dataset originally generated by Haber et al.[19]. Here, cells exposed to two pathogens (*H. polygyrus* or *Salmonella*) served as the target, and cells from healthy mice as the candidate background. As already shown in Weinberger et al.[7], ContrastiveVI successfully captured pathogen-specific variation in the target dataset, and BasCoD yielded a p-value of ≈ 1, validating the use of healthy cells as an appropriate background (Fig. 2c, Fig. S4). Taken together, these findings demonstrate that BasCoD corroborates the results of the previous contrastive dimension reduction studies.

BasCoD offers insights for post-trajectory analysis

Next, we illustrate how BasCoD can yield insights for post-trajectory analysis. Trajectory analysis traditionally involves constructing a minimum spanning tree on a low-dimensional representation of high-dimensional data[20–22]. Following this, one might wish to contrast the low-dimensional space at a specific time point with that at a branch point of the tree. However, transitions along pseudo-time do not guarantee that data at one time point are optimally contrasted against those at another. In this setting, BasCoD can be leveraged to compare datasets at different time points and identify which pairs are best suited for contrastive analysis. As an illustration, we applied BasCoD to a subset of the Human Cell Atlas bone marrow (HCA-BM) dataset from Hou et al., 2023[23]. Using TSCAN[21], we inferred three hematopoietic stem cell (HSC) lineages - Lymphoid, Myeloid, and Erythroid (Fig. 3a) - consistent with Hou et al. Firstly, an exploratory analysis based on PCA revealed that HSC cells exhibit higher rank than their differentiated progeny (Fig. S7b), suggesting that the lineages have reduced gene expression variation relative to HSC. We thus designated HSC as the target dataset and treated each lineage cluster as a candidate background. Applying BasCoD to each cluster, we found that only the two clusters corresponding to the Erythroid lineage (green and yellow in Fig. 3a) were unsuitable as backgrounds. Specifically, BasCoD yielded p-values of 0.29 for the Lymphoid lineage, 0.6 for the Myeloid lineage, and 0.005 and 0.0004 for the two Erythroid clusters, indicating that the Erythroid lineage exhibits the most drastic deviation in low-dimensional gene expression structure from HSC. To support these findings biologically, we examined the expression patterns of the genes SRSF7 and TPM4, which have the largest weights for the first two principal components of the HSC dataset. In the Erythroid lineage (green and yellow violin plots in Fig. S5a), these genes showed significantly different expression ($p < 2 \times 10^{-16}$) compared with the other lineages, indicating lineage-specific differential behavior of these key driver genes.

To further validate the BasCoD test result, we applied contrastive PCA to the HCA-BM dataset by regarding the HSC cells as the target dataset and the rest of the cell clusters as background datasets individually. The application of contrastive PCA followed by Hotspot[24] analyses on the learned cPCA components picked PBX1 as a key gene distinguishing HSCs from Myeloid and Lymphoid lineages, which were accepted by BasCoD as valid backgrounds. The result is also evidenced by strong correlations between PBX1 expression and cPCA components for those backgrounds compared to that of the PCA components obtained from the target dataset (Fig. 3b). In addition, while PBX1 was not chosen as one of the top genes based on a Hotspot analysis applied to the entire set of cells (Fig. 3a), it shows a differential behavior in the HSC cells as illustrated in Fig. S6. This finding is consistent with PBX1's function in sustaining stem cell identity and selectively biasing lineage outcomes: suppressing myeloid maturation while supporting lymphoid potential[25]. On the other hand, the use of Erythroid lineage, which was rejected by BasCoD, as a background dataset for the cPCA did not highlight PBX1 as strongly as the other lineages did when used as a background (Fig. 3b). We further note that the ordering of the BasCoD p-values across backgrounds aligns with that of the association strengths between cPC1 and the PBX1 expression levels and also with that of the Hotspot Z-scores (Fig. 3b). In summary, these results reinforce that contrastive analysis with valid backgrounds enhances the downstream inference from the trajectory analysis and that BasCoD provides a criterion that aligns with the validity of the backgrounds.

To illustrate the importance of appropriate background selection in contrastive dimension reduction along experimentally defined time trajectories, we revisited the population-scale scRNA-seq dataset of human iPSCs from Jerber et al.[1], previously introduced. This dataset includes four cell groups corresponding to Day 11 (D11), Day 30 (D30), Day 52 (D52), and D52 with rotenone-induced oxidative stress (ROT-treated) (Fig. 3c). Initial PCA on the full dataset revealed a substantial mismatch between D11 and D52 (ROT-treated) cells, whereas D30 and D52 cells showed considerable overlap in the low-dimensional embedding (Fig. 3c, top panel). We also examined how cells in the PCA embedding were separated by differentiation efficiency (DE). As shown in the bottom panel of Fig. 3c, high and low DE cells were not clearly distinguishable (R^2 = 0.24 from a regression of DE on the first two PCs). Next, we defined D52 ROT-treated cells as the target dataset and considered each of the other three groups (D11, D30, D52 control) as candidate backgrounds. Consistent with earlier observations, PCA on the target dataset alone yielded poor separation by DE (R^2 = 0.26, top right panel of Fig. 3d). Similarly, using D11 as the background for contrastive PCA (cPCA) led to minimal improvement (R^2 = 0.28; leftmost panel). In contrast, cPCA using D30 or D52 control cells as the background yielded improved separation by DE, with R^2 = 0.35 and 0.36, respectively (middle and right panels). To formally evaluate background suitability, we applied BasCoD. The resulting p-values further supported our findings: D30 (p = 0.06) and D52 control ($p \approx 1$) were valid backgrounds, while D11 ($p = 1.2 \times 10^{-5}$) was not. These results align with Jerber et al.'s observation that D11 cells differ markedly in cell type composition compared to later time points, whereas D30 and D52 are more similar. In summary, these analyses highlight how BasCoD provides a principled and quantitative tool for verifying the reliability of contrastive dimension reduction across temporal conditions, whether defined by experimental design or inferred via trajectory inference.

BasCoD enables background data calibration in single-cell experiments

After demonstrating BasCoD's ability to evaluate whether a given dataset can serve as a suitable background for contrastive dimension

https://doi.org/10.1038/s41467-026-70652-4

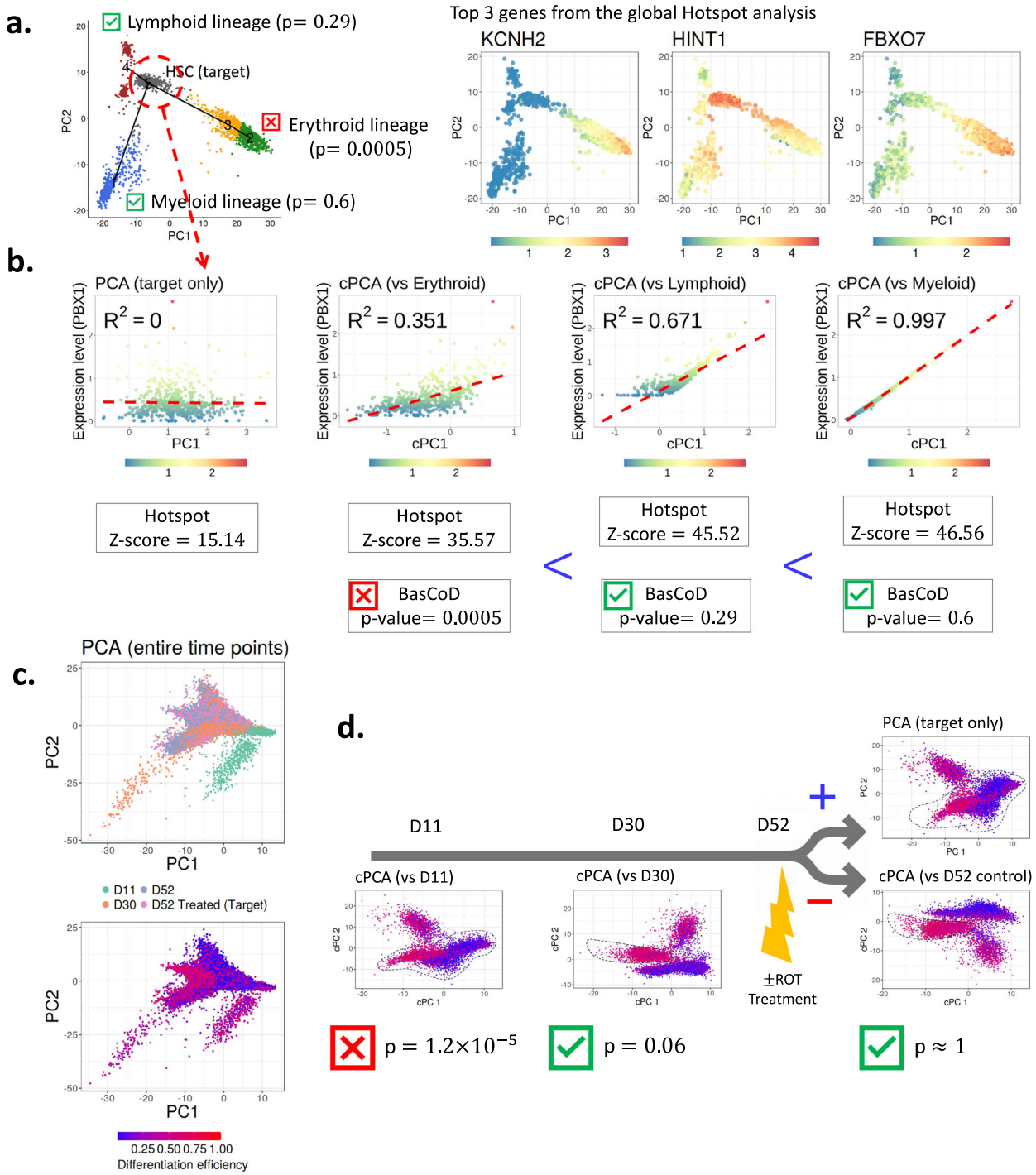


Fig. 3 | Background selection for post-trajectory analysis contrastive dimension reduction with BasCoD. a PCA coordinates of HCA-BM cells[23] with the TSCAN[21] minimum spanning tree. Each lineage (candidate background dataset) learned from TSCAN is evaluated by BasCoD for a contrastive analysis against the HSC cells (target dataset) and BasCoD p-values are reported for each lineage. The coordinates colored by the expression levels of the top 3 genes identified by Hotspot[24] were also displayed. **b** PCA or cPCA coordinates of the target dataset (HSC cells) colored by the expression level of PBX1 gene. A direct comparison of the PBX1 gene expression level and the first PC (or cPC) is also displayed with the R^2 coefficient of determination. **c** Top: PCA coordinates of the human iPSCs across all the conditions, colored by time points and treatment: D11, D30, D52 and D52 (Treated). Bottom: Same as the top panel, but colored by the differential efficiency level of each donor. **d** Top right: PCA coordinates of the human iPSCs at D52 with ROT treatment (target dataset), colored by the differential efficiency level of each donor. Bottom right: cPCA coordinates of the same D52 (ROT) cells after cPCA against control cells at D52 (candidate background dataset). Cells marked with dashed-line circle represent the ones from donors with high differentiation efficiency. BasCoD *p*-values are reported at the bottom of the cPCA plots. Middle and left panels depict the same plots with D30 and D11 cells as candidate background datasets, respectively. The left most figure of panel a was generated using the analysis pipeline implemented in the GitHub repository associated with Hou et al. (2023)[23]. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-026-70652-4

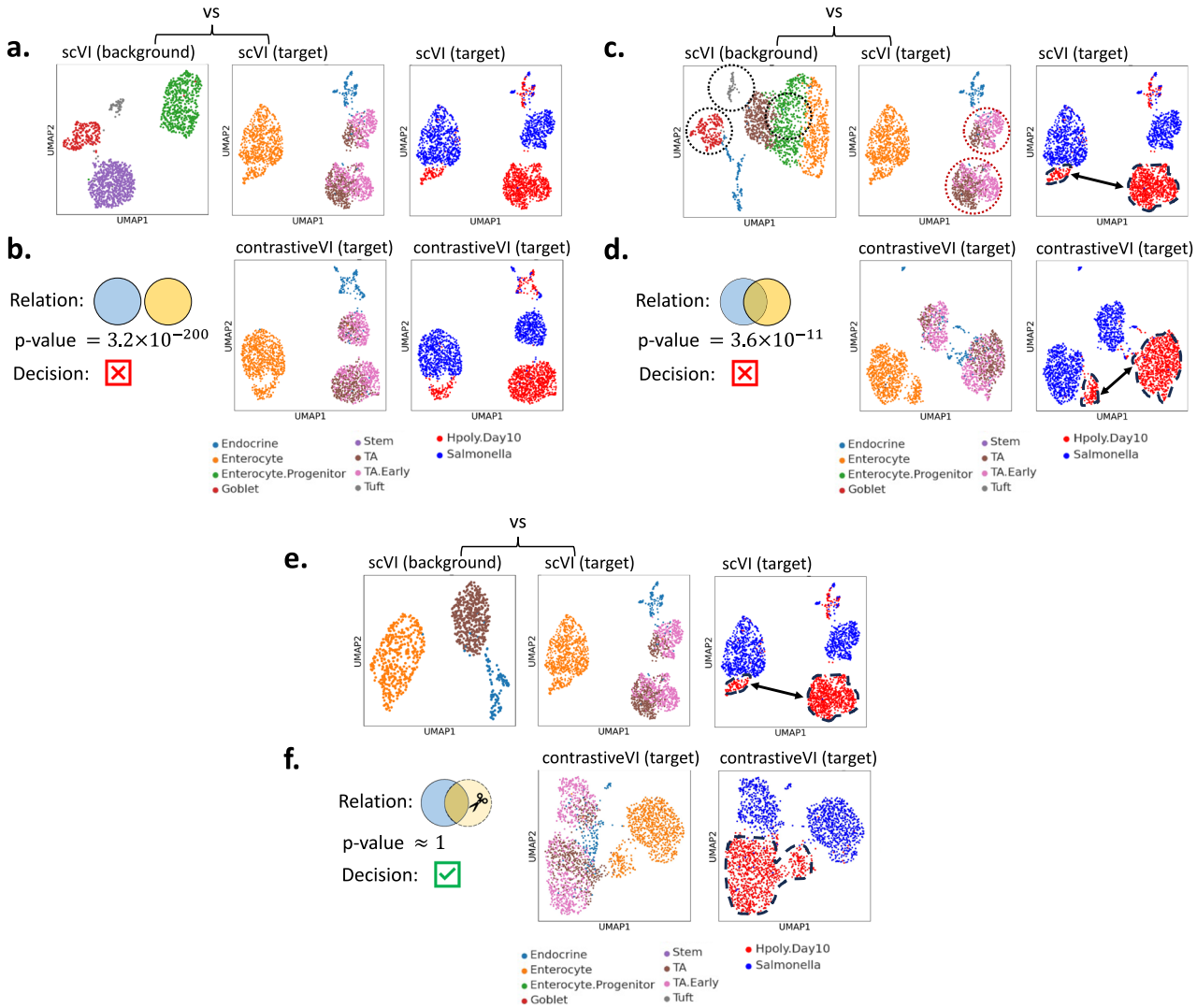


Fig. 4 | Background data calibration for improved contrastive dimension reduction with BasCoD. a Left: UMAP coordinates of the scVI embeddings of a subset of mouse intestinal cells[32] without pathogen treatment (candidate background designed to have no overlapping cell types with the target dataset) colored by cell type. Middle: same figure with the cells that received pathogen treatments (target dataset). Right: Same figure as the middle panel, but colored by the pathogen type. **b** Left: Venn-diagram depicts the designed relation between target and candidate background. A BasCoD p-value with a decision on the candidate background is reported. Middle: UMAP coordinates of contrastiveVI salient embedding of the target dataset contrasted against the candidate background dataset, colored by cell type. Right: same as the middle panel, but colored by pathogen type. **c** Left: UMAP coordinates of the scVI embeddings of a subset of mouse intestinal cells[32] without pathogen treatment (candidate background) colored by cell type. The cells within black dotted-line circles are the ones that are only present in the candidate background dataset. Middle: same figure with the cells that received pathogen treatments (target dataset). Right: Same figure as the middle panel, but colored by the pathogen type. **d** Left: Venn-diagram depicts the designed relation between target and candidate background. A BasCoD p-value with a decision on the candidate background is reported. Middle: UMAP coordinates of contrastiveVI salient embedding of the target dataset contrasted against the candidate background dataset, colored by cell type. Right: same as the middle panel, but colored by the pathogen type. **e**, **f** Same figures as **c**, **d**, but with a calibrated candidate background dataset obtained by removing the cells (circled cells in **e**) that are specific only to the candidate background dataset. Source data are provided as a Source Data file.

reduction of a target dataset, we now illustrate how it can further guide the design and refinement of candidate backgrounds. As detailed in the Methods section, a valid background according to BasCoD should not harbor any distinct low-dimensional features beyond those present in the target dataset, and its low-dimensional embedding space must lie entirely within that of the target (Fig. 1b). To exemplify a pathological scenario, we revisited the mouse intestinal single-cell dataset, constructing target and candidate datasets with completely non-overlapping cell types (Fig. 4a). Under these conditions, contrastiveVI fails to reveal any additional insights on pathogen types, as its salient embeddings show (Fig. 4b). Correspondingly, BasCoD returned an extremely low p-value (3.2×10^{-200}), clearly rejecting the candidate background. As an intermediate scenario, we constructed a candidate background dataset containing both cell types common to the target (Endocrine, Enterocyte, and TA) and candidate-specific cell types (Goblet, Enterocyte. Progenitor, and Tuft), highlighted by black dotted circles in Fig. 4c. Here, the target dataset also exhibited distinct variation within TA cells (red dotted circles in Fig. 4c). Although contrastiveVI improved pathogen-type clustering compared to the target-only dimension reduction (ARI score = 0.2, NMI score = 0.18, average Silhouette score = 0.47), the clustering was still suboptimal. Specifically, the cells with Hploy.Day 10 pathogen type (red colored cells marked by black dashed lines in Fig. 4d) still remain separated even within the contrastive VI salient embeddings. Consistently, BasCoD yielded a p-value of 3.6×10^{-11}, indicating that the candidate background was not valid.

Following this rejection, we applied a sequential refinement procedure by removing candidate-specific cell types and reassessing background validity using BasCoD. Specifically, we recalibrated the candidate dataset by excluding the Goblet, Enterocyte. Progenitor, and Tuft cell types (black dashed circles in Fig. 4c), separately applied scVI to the target and calibrated candidate datasets, and carried out testing with BasCoD (Fig. 4e). This recalibrated candidate yielded a p-value of ≈ 1, affirming its suitability for contrastive analysis. Correspondingly, the contrastiveVI salient embeddings derived from this refined background exhibited markedly improved separation by pathogen type (ARI score = 0.30, NMI score = 0.34, average Silhouette score = 0.45) compared to embeddings obtained from the target-only scVI (Fig. 4f). In particular, the cells with Hploy.Day 10 pathogen type (red colored cells marked by black dashed lines in Fig. 4f) now form a more homogeneous cluster against those in Fig. 4d. In summary, these results demonstrate that BasCoD not only rigorously assesses candidate backgrounds but also provides a systematic framework for their effective design and calibration in contrastive dimension reduction.

Practical utility of BasCoD for designed experiments with high-dimensional data

We demonstrate the practical utility of BasCoD for designing valid contrasts in single-cell experiments. In an erythroid differentiation analysis with primary human hematopoietic stem/progenitor cells (HSPCs), cells were subjected to culture under conditions or upon inflammation with a mixture of the cytokines TNFα, IFNγ, and IL-6 on day 4 of the culture and harvested for single-cell RNA sequencing on days 9 (D9) and 13 (D13). An initial exploratory analysis revealed that, while control cells differentiate as a function of time, inflammation largely blocks differentiation (Fig. S9). We focused on cells that still differentiated by D13 (orange cells within the dashed circle in Fig. S9), defining them as "escaped cells" post-inflammation (Supplementary Note for details). We set these D13 inflammation-treated cells as the target and considered three candidate backgrounds: inflammation D9, control D9, and control D13. PCA on the target dataset alone failed to separate the escaped cells (ARI = 0.18, Fig. S10a), a misalignment also evident with the Leiden clustering of the principal components (Fig. S10b, Fig. S11). We applied cPCA[4] to contrast the target against each candidate background. When contrasting inflammation-treated D13 and D9 cells (Contrast I in Fig. 5a), the escaped cells were clearly separated (ARI = 0.69), and BasCoD returned a p-value of ≈ 1, confirming that inflammation-treated D9 cells represents a valid background. This supports our expectation that fixing one factor (treatment) while comparing different times enhances the capacity to identify temporal effects. In contrast, comparing inflammation-treated D13 cells with control D13 cells (Contrast III in Fig. 5a) yielded poor separation (ARI = 0.16) and a BasCoD p-value of $< 2 \times 10^{-146}$, due to the candidate's strong, independent low-dimensional features (as evident by the sparse distribution of earlier stage cells in Fig. S9). Similarly, using inflammation-treated D9 cells as the candidate (Contrast II in Fig. 5a) resulted in a p-value < 0.001 and only moderate clustering (ARI = 0.49). Notably, the UMAP embedding of inflammation-treated D9 cells is similar to that of control D9 cells, but with fewer late-stage cells, reinforcing its validity as a background. Collectively, this analysis highlights the capacity of BasCoD to identify and calibrate appropriate background dataset, thereby avoiding misleading inferences in contrastive dimension reduction. Critically, inappropriate background selection can lead to misleading downstream findings. For instance, Gene Set Enrichment Analysis (GSEA)[26] on cell clusters derived from cPCA with a valid background (D9 with inflammation) showed that escaped cells exhibit significantly lower expression of immune function genes (e.g., leukocyte and lymphocyte activation) in comparison to other cells (Fig. 5b). This result was consistent with a gold standard GSEA based on inferred cell labels, with seven out of ten gene sets matching (Fig. 5c). In contrast, GSEA on other contrasts matched only one out of ten gene sets (Fig. S12a,b), emphasizing that invalid background selection can lead to incomplete/misleading inference.

Practical usage of BasCoD in single cell perturbation studies

Finally, we demonstrate a broader application of BasCoD's subspace inclusion test for detecting interaction effects of treatments in low-dimensional spaces, using a subset of the Perturb-seq dataset from Norman et al.[17], as analyzed in Weinberger et al.[7].

The Norman et al. (2019) Perturb-seq dataset includes CRISPRi interventions targeting both single genes and selected gene pairs; that is, simultaneous CRISPRi perturbations on two genes. Leveraging this design, we asked whether the BaSCoD subspace inclusion testing framework could offer insights into potential interaction effects of such joint perturbations. For each intervened gene pair, we assessed whether the *double-perturbed* cells (target) span a low-dimensional subspace that contains the subspace spanned by the *singly-perturbed* cells (candidate background). We applied the BaSCoD test to each of the nine available gene (i.e., perturbation) pairs. Gene pairs from the same family, such as CEBPA/CEBPB, CDKN1C/CDKN1B, CDKN1B/CDKN1A, and CEBPA/CEBPE, produced significant BaSCoD p-values (< 0.05) rejecting the appropriateness of background datasets. In contrast, more functionally distinct pairs, such as KLF1/CEBPA, KLF1/CEBPB, KLF1/CEBPE, ETS2/MAPK1, and CEBPE/ETS2, yielded non-significant p-values (> 0.05) (Fig. 6a). These findings suggest that joint perturbations involving dissimilar genes tend to generate a low-dimensional subspace that includes the variability observed in the corresponding single perturbations.

To further investigate the testing results, we focused on two gene pairs with the most contrasting BaSCoD p-values: CEBPA/CEBPB ($< 2 \times 10^{-16}$) and KLF1/CEBPA (≈ 1). We examined the low-dimensional embeddings of the singly-perturbed cells (blue, Fig. 6b, c) and double-perturbed cells (red, Fig. 6b, c) for each pair. CEBPA and CEBPB are members of the same transcription factor family (C/EBP), with partially redundant regulatory roles[27], whereas KLF1 and CEBPA belong to distinct transcription factor families with divergent gene regulatory functions[28]. In the case of KLF1/CEBPA, visual inspection revealed that the double-perturbed cells are broadly distributed across clusters occupied by the singly-perturbed cells (Fig. 6b), but also form distinct clusters of their own (black dashed circles), suggesting that the subspace spanned by the double perturbation encompasses that of the single perturbations (see Venn diagram in Fig. 6a). In contrast, for the CEBPA/CEBPB pair, the singly-perturbed cells span distinct subspaces (golden dashed circles in Fig. 6c), and most of the double-perturbed cells fall within this combined subspace (Fig. 6c). We then applied contrastive PCA (cPCA) to both gene pairs. For KLF1/CEBPA, the contrastive components (cPCs) revealed that the double-perturbed cells (red, Fig. 6d) vary primarily along a single direction (solid arrow), while the singly-perturbed cells (blue) are dispersed across the embedding. This indicates that the learned contrastive components capture the unique structure of the double-perturbed cells relative to the singly-perturbed ones, consistent with the high BaSCoD p-value. By contrast, for CEBPA/CEBPB, the double-perturbed cells showed no distinct directional variation (Fig. 6e), aligning with the strong subspace inclusion result.

Following visual exploration of cPCA, we hypothesized that the simultaneous perturbation of KLF1 and CEBPA induces interaction effects on the cells' transcriptomic profiles, which may be reflected in novel variation captured by the subspace of double-perturbed cells. From the resulting contrastive principal components (cPCs), we identified the top 10 genes most correlated with each of the first five cPCs, collectively referred to as the cPC genes. We repeated this procedure for the CEBPA-CEBPB pair to obtain a comparable set of cPC

https://doi.org/10.1038/s41467-026-70652-4

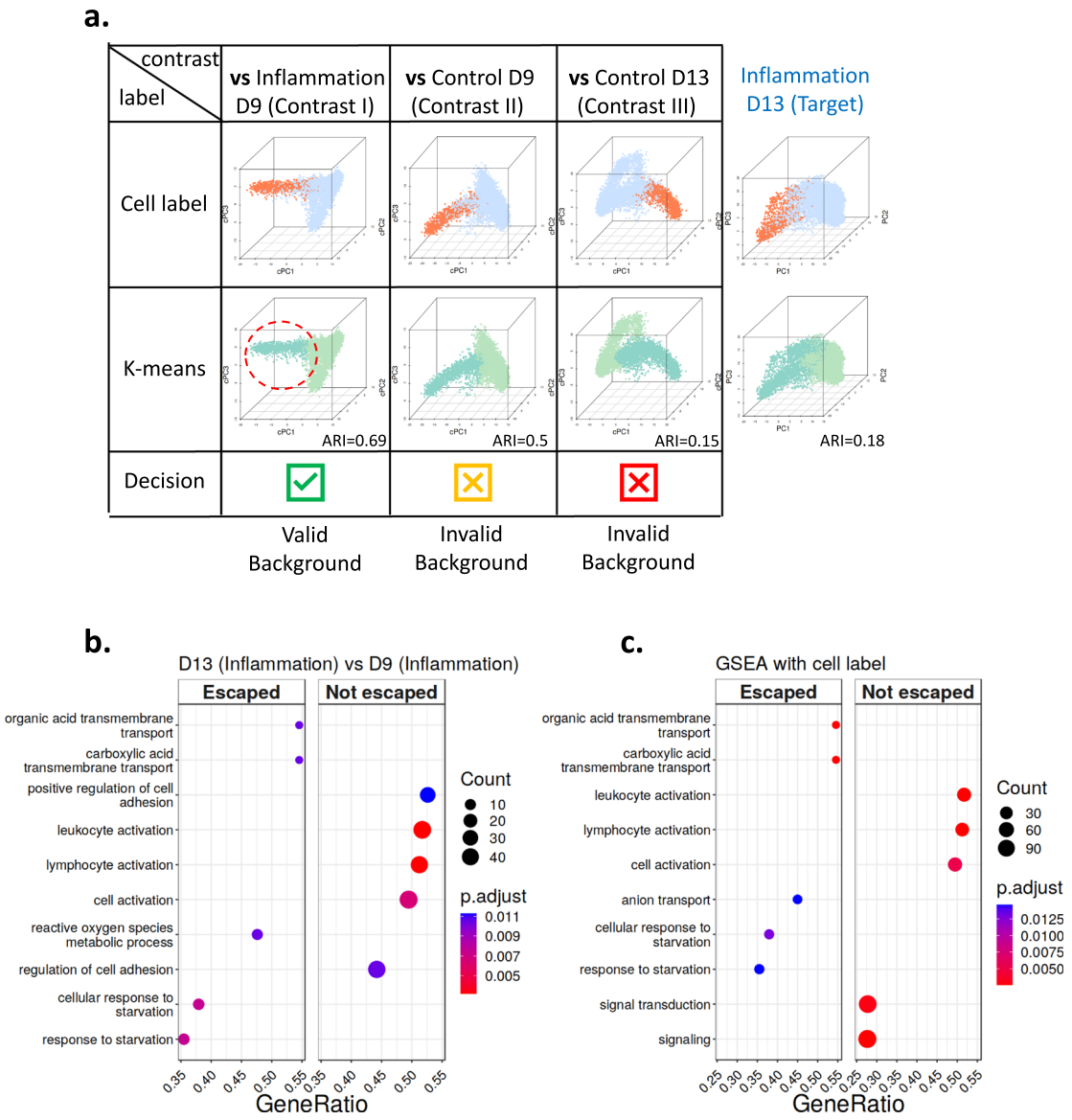


Fig. 5 | Application of BasCoD for background selection in a designed scRNA-seq experiment. a The right-most column depicts PCA embeddings (first three components) of HSPC cells at day 13 (D13) under inflammation treatment (target dataset), colored by the cell labels (top) and k-means derived cell clusters (bottom). The ARI at the bottom quantifies the concordance between the cell labels and clustering. Each column in the table depicts cPCA embeddings contrasting the target dataset against candidate backgrounds: D9 cells with inflammation, D9 cells without inflammation, and D13 cells with inflammation. The BasCoD decisions are reported in the bottom row. **b** Gene set enrichment analysis (GSEA) comparing the two cell clusters derived from contrastive PCA embeddings when the target dataset (D13 with inflammation) is contrasted against inflammation-treated D9 cells. The "Escaped" panel contains the gene functions enriched in escaped cells and the "Not escaped" panel contains those enriched in not escaping cells. **c** GSEA result comparing escaped and not escaping cells of the target dataset (D13 with inflammation). For both b and c, one-sided hypergeometric test *p*-values with BH correction are reported. Source data are provided as a Source Data file.

genes. Next, we conducted a formal analysis to quantify the interaction effects of each gene pair perturbation on their respective cPC genes using a linear model that included an interaction term between the two perturbation indicators. This analysis revealed stronger interaction effects on the cPC genes for the KLF1-CEBPA pair (red circles in Fig. 6f) compared to the CEBPA-CEBPB pair (gray circles in Fig. 6f). For example, the gene RPL14, identified as a cPC gene for the KLF1-CEBPA pair, exhibited a synergistic response: its expression decreased when only KLF1 was perturbed (1,0) relative to the control (0,0), but increased significantly under double perturbation (1,1) relative to CEBPA-only perturbation (0,1), indicating a clear interaction effect (Response Fig. 6g). In contrast, the CEBPA-CEBPB pair did not show such reversals in expression (Fig. 6h). Collectively, the statistical testing revealed that 60% (30/50) of the cPC genes for the KLF1-CEBPA pair exhibited significant interaction effects (i.e., effects beyond what is explainable by an additive perturbation effect of two genes) after Benjamini-Hochberg multiplicity correction, compared to only 32% (16/50) for the CEBPA-CEBPB pair, yielding a significant difference in proportions (p-value < 0.01).

Collectively, these results demonstrate the utility of BasCoD for detecting low-dimensional structure that captures strong interaction-effect signals in single-cell Perturb-seq datasets involving multiple interventions per cell.

Discussion

We introduced BasCoD, a framework that leverages low-dimensional representations of target and candidate background datasets to provide a statistical goodness-of-fit test for contrastive dimension reduction in large-scale, heterogeneous genomic datasets. The p-values produced by BasCoD are well calibrated under the null hypothesis that

https://doi.org/10.1038/s41467-026-70652-4

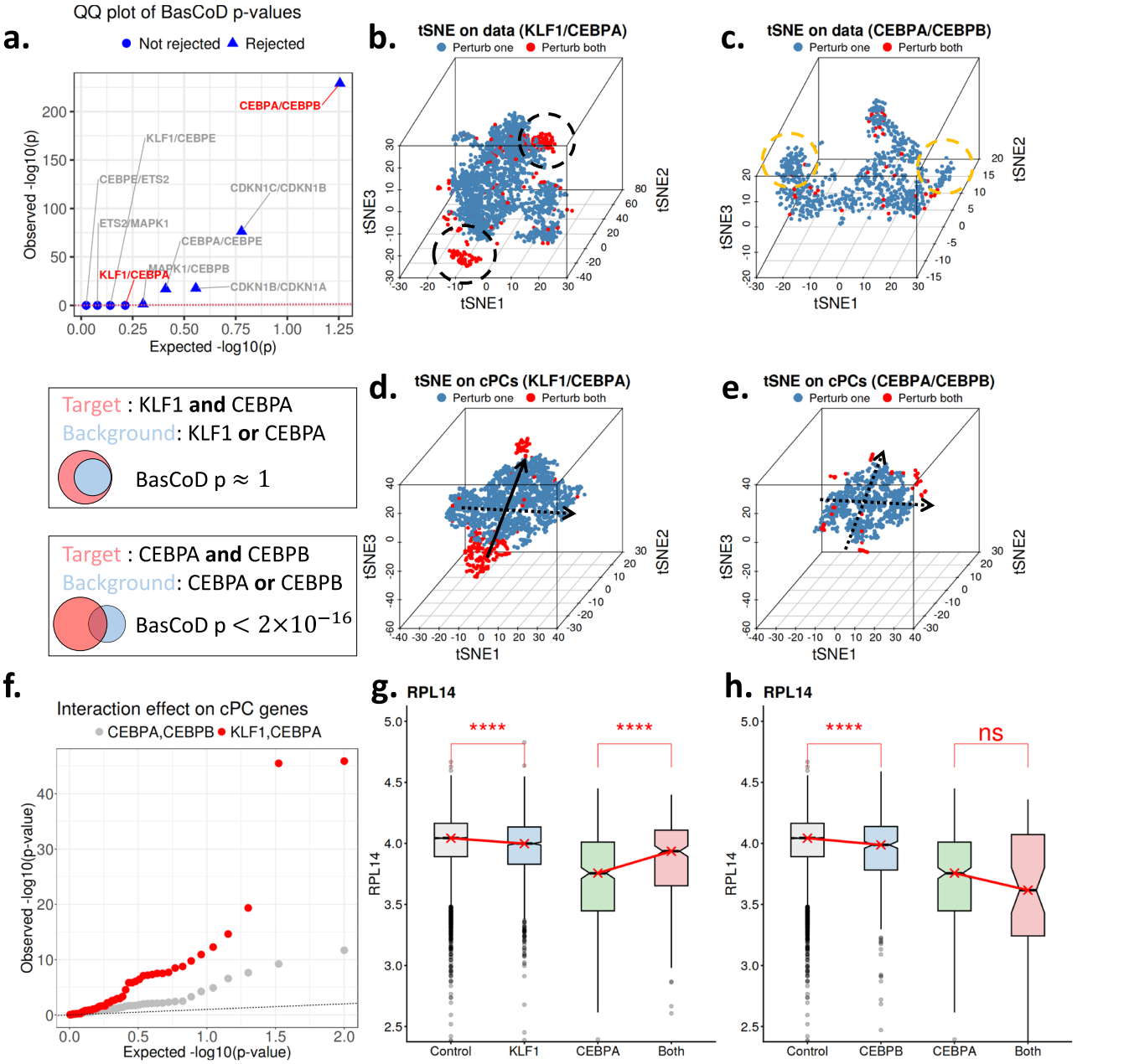


Fig. 6 | Practical utility of BasCoD for identifying interaction effect in Perturb-seq experiments. a Quantile-Quantile plot for the BasCoD p-values of each of the gene pairs that were both simultaneously and individually perturbed in a subset of Norman et al. (2019) dataset processed by Weinberger et al., (2023). For each gene pair, the target data corresponds to the cells with double-perturbation and the background corresponds to the cells with single-perturbation. **b** A three-dimensional tSNE embeddings of the cells with single perturbation (blue) and double perturbation (red) for the KLF1 and CEBPA pair. **c** As in panel **b**, but for the CEBPA and CEBPB pair. **d** As in **b**, but the tSNE embeddings were obtained from the cPCs of cPCA of the pair KLF1/CEBPA. **e** As in panel **d**, but the cPCs were obtained from the cPCA of the pair CEBPA/CEBPB. **f** Quantile-Quantile plot of two-sided t-test p-values for interaction effects of cPC genes under the KLF1-CEBPA perturbations (red circles) and the CEBPA-CEBPB perturbations (gray circles). **g** Normalized expression level of RPL14 (y-axis) across control cells (#of cells = 7, 275), KLF1 intervened cells (#of cells = 1, 629), CEBPA intervened cells (#of cells = 387), and cells with intervention on both of the genes (#of cells = 267). The summary statistics---minimum, maximum, center, lower and upper bounds of the box, lower and upper whiskers, and the 10th and 90th percentiles---for each boxplot, from left to right, are (1.21, 5.36, 4.04, 3.89, 4.16, 3.48, 4.55, 3.69, 4.26), (2.38, 4.82, 3.99, 3.82, 4.13, 3.37, 4.54, 3.62, 4.24), (1.44, 4.44, 3.75, 3.44, 4.01, 2.61, 4.44, 3.14, 4.18), and (2.60, 4.39, 3.93, 3.65, 4.10, 2.97, 4.39, 3.40, 4.20). **h** As in panel **g**, but with the pairs CEBPA and CEBPB. There are 402 cells with the CEBPB intervention and 50 cells with intervention on both of the genes. The summary statistics---minimum, maximum, center, lower and upper bounds of the box, lower and upper whiskers, and the 10th and 90th percentiles---for the boxplots labeled `CEBPB and `Both are (3.29, 4.58, 3.98, 3.78, 4.13, 3.29, 4.58, 3.49, 4.23) and (2.23, 4.36, 3.61, 3.24, 4.07, 2.23, 4.36, 2.98, 4.15), respectively. The stars in panels **g**, **h** indicate the significance (with $p < 0.01$) of two-sided Wilcoxon rank sum test. In **g**, the comparison between control and KLF1-perturbed cells yields $p = 3.4 \times 10^{-12}$, and the other comparison yields $p = 9.9 \times 10^{-7}$. In **h**, the comparison between control and CEBPB-perturbed cells yields $p = 5 \times 10^{-6}$, and the other comparison yields $p = 0.27$. Source data are provided as a Source Data file.

a candidate dataset is a valid background. Moreover, simulation studies demonstrate that BasCoD achieves well-controlled false discovery rates and high power in identifying unsuitable backgrounds, underscoring its reliability as a diagnostic tool for contrastive analysis. In this work, we showcase the utility of BasCoD as a general-purpose goodness-of-fit test for contrastive dimension reduction, as well as a post-trajectory analysis tool to elucidate biological signals specific to a given time point. We further demonstrated that BasCoD can be used to calibrate background datasets in single-cell experiments involving contextual variation, such as differences in cell types. Finally, we showed that BasCoD can yield insights from designed experiments with high-dimensional measurements by identifying appropriate conditions for contrast and formally testing subspace inclusion. In particular, we applied BasCoD to Perturb-seq data to distinguish gene pairs with synergistic transcriptional effects, highlighting its value for interpreting complex perturbational designs.

A recently proposed method[29] tests whether the target dataset contains additional low-dimensional structure relative to a candidate background. While effective at detecting target-specific variation, it does not address the complementary and critical task of validating the background dataset itself. In particular, if a completely unrelated dataset is used as the background, their procedure still concludes that the target dataset contains distinct components. However, this does not prevent users from proceeding with contrastive dimension reduction using an invalid background. For instance, in the pathological case shown in Fig. 4a, their method returns a significant p-value, suggesting target-specific signal (Supplementary Notes). Yet, applying contrastive VI with this irrelevant background fails to effectively reveal that signal.

In determining the dimensionality of embeddings for target and background datasets, we followed common empirical practices informed by exploratory analysis (Supplementary Notes). However, developing a more systematic approach to dimensionality selection remains an important direction for future work. Additionally, while we approximated subspaces for nonlinear CDR methods using a standard linear embedding technique, non-parametric alternatives could provide broader generalization. For instance, comparing reconstruction errors from encoder-decoder models trained on the target versus background datasets via non-parametric testing could offer another avenue for evaluating subspace inclusion.

BasCoD operates at the embedding level and can, in principle, be applied to a variety of single-cell modalities, including scATAC-seq and scMethyl-seq. Moreover, it is compatible with any contrastive dimension reduction method that has a natural non-contrastive counterpart. These features position BasCoD as a broadly applicable and versatile tool for advancing contrastive analysis across diverse single-cell genomic settings.

Methods

Latent variable models of target and background datasets

Throughout the paper, we denote the *target dataset* as $X_0 = [\mathbf{x}_{1,0}^T, \cdots, \mathbf{x}_{n_0,0}^T] \in \mathbb{R}^{n_0 \times p}$, where $\{\mathbf{x}_{1,0}, \cdots, \mathbf{x}_{n_0,0}\}$ are n_0 i.i.d. copies of $\mathbf{x}_0 \in \mathbb{R}^p$, and candidate background dataset j, $j \in \mathcal{C}$, as $X_j = [\mathbf{x}_{1,j}^T, \cdots, \mathbf{x}_{n_j,j}^T] \in \mathbb{R}^{n_j \times p}$, where $\{\mathbf{x}_{1,j}, \cdots, \mathbf{x}_{n_j,j}\}$ are n_j i.i.d. copies of $\mathbf{x}_j \in \mathbb{R}^p$. Here, $\mathcal{C}$ is the index set for the candidate datasets. We model $\mathbf{x}_j$ as a function of two sets of latent variables named as (i) data-common embeddings $\mathbf{c}_j \in \mathbb{R}^{R_c}$ and (ii) data-specific embeddings $\mathbf{s}_j \in \mathbb{R}^{R_{\Delta j}}$, and a random noise $\boldsymbol{\epsilon}_j \in \mathbb{R}^p$, for each $j \in \{0\} \cup \mathcal{C}$ as:

$$\mathbf{x}_j = f_{\boldsymbol{\epsilon}_j}^j(\mathbf{c}_j, \mathbf{s}_j). \tag{1}$$

In the contrastive dimension reduction framework, we expect the target dataset to be generated from both $\mathbf{c}_0$ and $\mathbf{s}_0$, and an appropriate background dataset to be generated only through $\mathbf{c}_j$. With this notion, we define an appropriate *background dataset* as a candidate dataset satisfying $\mathbf{s}_j = 0$, i.e.,

$$\mathbf{x}_j = f_{\boldsymbol{\epsilon}_j}^j(\mathbf{c}_j, \mathbf{0}), \tag{2}$$

and denote the data index set $\mathcal{B} := \{j \in \mathcal{C} | \mathbf{s}_j = 0\}$. Then, the main task is to identify the set $\mathcal{B} \subset \mathcal{C}$, utilizing the low-dimensional embeddings learned from X_j for $j \in \{0\} \cup \mathcal{C}$.

It is worth noting that model (1) with a single background dataset, i.e., $\mathcal{B} = \{1\}$, encompasses a large class of contrastive dimension reduction method models. For example, when $f_{\boldsymbol{\epsilon}}^j(\mathbf{x}, \mathbf{y}) = \Gamma_c \mathbf{x} + \Gamma_{s,j} \mathbf{y} + \boldsymbol{\epsilon}$, where $[\Gamma_c, \Gamma_{s,j}] \in \mathbb{R}^{p \times R_j}$ are orthogonal, with $\Gamma_c \in \mathbb{R}^{p \times R_c}$ and $\Gamma_{s,j} \in \mathbb{R}^{p \times R_{\Delta_j}}$, we have

$$\mathbf{x}_0 = \Gamma_c \mathbf{c}_0 + \Gamma_{s,0} \mathbf{s}_0 + \boldsymbol{\epsilon}_0,$$
$$\mathbf{x}_1 = \Gamma_c \mathbf{c}_1 + \boldsymbol{\epsilon}_1$$

as target and background datasets, respectively. This corresponds to Contrastive Latent Variable Model (CLVM)[6] and a specific case of Contrastive PCA (cPCA)[4] model. Here, the loading matrix Γ_s encodes the low-dimensional space specific to the target dataset and Γ_c encodes the low-dimensional space common to both of the datasets. When $f_{\boldsymbol{\epsilon}}^j(\mathbf{x}, \mathbf{y}) = f_{\boldsymbol{\epsilon}, \boldsymbol{\theta}}^j(\mathbf{x}, \mathbf{y})$ is constructed through a Variational Auto Encoder or Variational Inference with unknown parameters $\boldsymbol{\theta}$, the model corresponds to that of ContrastiveVAE[5] or ContrastiveVI[7].

Statistical test for selecting background datasets

With the set up outlined above, we now introduce the procedure of selecting $\mathcal{B}$ out of $\mathcal{C}$. We first illustrate the testing procedure for linear $f_{\boldsymbol{\epsilon}}^j(\mathbf{x}, \mathbf{y})$ and describe how a similar procedure can be implemented in general. Consider a linear latent variable model $f_{\boldsymbol{\epsilon}}^j(\mathbf{x}, \mathbf{y}) = \Gamma_c \mathbf{x} + \Gamma_{s,j} \mathbf{y} + \boldsymbol{\epsilon}$ with $\Gamma_j := [\Gamma_c, \Gamma_{s,j}] \in \mathbb{R}^{p \times R_j}$. Then, we have the datasets

$$\mathbf{x}_0 = \Gamma_c \mathbf{c}_0 + \Gamma_{s,0} \mathbf{s}_0 + \boldsymbol{\epsilon}_0,$$
$$\mathbf{x}_j = \Gamma_c \mathbf{c}_j + \Gamma_{s,j} \mathbf{s}_j + \boldsymbol{\epsilon}_j, \ \forall j \in \mathcal{C}.$$

When the jth candidate dataset is a valid background, i.e., $j \in \mathcal{B}$, we have

$$\mathbf{x}_j = \Gamma_c \mathbf{c}_j + \boldsymbol{\epsilon}_j.$$

This implies that the candidate's subspace Γ_j, $j \in \mathcal{B}$, must match or be contained in the same subspace that the target dataset occupies (Γ_c): $C(\Gamma_j) = C(\Gamma_c) \subset C(\Gamma_0)$. This is because $\Gamma_j = \Gamma_c$ for valid background $j \in \mathcal{B}$, which equivalently implies $\Gamma_j = P_0 \Gamma_j$, where P_0 is the projection matrix onto the space $C(\Gamma_0)$. If this condition fails, the candidate background dataset j introduces additional, unwanted variation and thus is not a suitable background. As a result, we formulate the null hypothesis $H_{0,j}^* : j \in \mathcal{B}$ as $H_{0,j}$: $\Gamma_j = P_0 \Gamma_j$, which holds if and only if

$$H_{0,j} : \rho_{ij} = 1, \forall i \in [R_j],$$

where $R_j = R_c + R_{\Delta j}$, $\rho_{ij} = \mathrm{corr}(\boldsymbol{\gamma}_{ij}, P_0 \boldsymbol{\gamma}_{ij})$, $\mathrm{corr}(\cdot, \cdot)$ denotes the sample correlation, and $\boldsymbol{\gamma}_{ij}$ denotes the ith column vector of Γ_j.

Testing whether the jth candidate is an appropriate background requires dimension reduction for each $j \in \{0\} \cup \mathcal{C}$ to obtain $\widehat{\Gamma}_j$ and evaluation of how close $\widehat{\rho}_{ij}$ is to 1 for $\forall\ i \in R_j$, where $\widehat{\rho}_{ij} = \mathrm{corr}(\widehat{\boldsymbol{\gamma}}_{ij}, \widehat{P}_0 \widehat{\boldsymbol{\gamma}}_{ij})$. For such evaluation, we consider the Fisher transformation on each correlation estimate $\widehat{\rho}_{ij}$ to generate a z-score $Z_{ij} = \sqrt{p-3}(\mathrm{arctanh}(\widehat{\rho}_{ij}) - \mu)$, where $\mu = \mathrm{arctanh}(1 - \epsilon)$ for a small $\epsilon > 0$ (to avoid $\mu = \infty$). Then, under the null hypothesis $H_{0,j}$, we have each $Z_{ij} \sim N(0, 1)$ for all $i \in [R_j]$. Leveraging Fisher's method for combining p-

values, we have

$$\chi_j^2 = -2\sum_{i=1}^{R_j} \ln\left\{1-\Phi\left(Z_{ij}\right)\right\} \sim \chi^2(df = 2R_j), \tag{3}$$

where df denotes degrees of freedom and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. While the Fisher's method to combine p-values rely on independence of the individual p-values across $i \in [R_j]$, variants of this procedure for accommodating dependencies among p-values are available[19,30]. With this null distribution for the chi-squared statistic of the combined p-values, $H^*_{0,j} : j \in \mathcal{B}$ (candidate dataset j is a background) is rejected as an appropriate background for small p-values.

When $f^j_{\boldsymbol{\epsilon}}(\mathbf{c}_j, \mathbf{s}_j)$ is in a general form (i.e., as in contrastiveVI), a direct analogue of Γ_j may not be available unless f^j contains a partially linear format depending on the latent embeddings $\mathbf{c}_j$ and $\mathbf{s}_j$. As a result, to adopt the BasCoD testing procedure for these settings, we estimate Γ_j through the following linear approximation:

$$\widehat{\Gamma}_j := \underset{B\in\mathbb{R}^{p\times R_j}}{\operatorname{argmin}} \left\| X_j - L_j B^T \right\|_2^2, \text{ for } \forall j \in \{0\}\cup\mathcal{C}. \tag{4}$$

where $L_j \in \mathbb{R}^{n_j \times R_j}$ is any low-dimensional embedding obtained from data X_j. For contrastive PCA, L_j corresponds to the R_j principal components of X_j and for contrastiveVI or VAE, L_j corresponds to R_j latent embedding variables using only jth dataset. Then, using $\widehat{\Gamma}_j$ across $j \in \{0\}\cup\mathcal{C}$, we can follow the same procedure introduced above. Algorithm 1 summarizes the entire BasCoD testing procedure. After constructing/selecting a valid background dataset $\{X_j | j \in \mathcal{B}\}$, one can proceed with contrastive dimension reduction against the target dataset X_0. The testing procedure requires two key parameters: the size R_j of the low-dimensional space of each $j \in \{0\}\cup\mathcal{C}$ and the slack parameter $\epsilon > 0$. We refer to Supplementary Notes for details on how to set these parameters.

Algorithm 1. BasCoD

Require: Target dataset X_0; Candidate background datasets $\{X_j | j \in \mathcal{C}\}$; Ranks from dimension reduction of choice R_j for $j \in \{0\}\cup\mathcal{C}$; $\epsilon > 0$.

1: Set $\mathcal{B} = \mathcal{C}$.

2: **for** $j \in \mathcal{C}$ **do**

3: Obtain $\widehat{\Gamma}_0 \in \mathbb{R}^{p\times R_0}$ and $\widehat{\Gamma}_j \in \mathbb{R}^{p\times R_j}$ from a dimension reduction on X_0 and X_j, respectively from (4).

4: Calculate $\widehat{\rho}_{ij} = \operatorname{corr}(\widehat{\gamma}_{ij}, \widehat{P}_0\widehat{\gamma}_{ij})$ for each $i \in [R_j]$.

5: Transform $Z_{ij} = \sqrt{p-3}(\operatorname{arctanh}(\widehat{\rho}_{ij}) - \mu)$ for $i \in [R_j]$, where $\mu = \operatorname{arctanh}(1-\epsilon)$.

6: Construct

$$\chi_j^2 = -2\sum_{i=1}^{R_j} ln\left\{1-\Phi\left(Z_{ij}\right)\right\}.$$

7: If $p_j = \mathbb{P}(\chi^2(2R_j) \le \chi_j^2) \le \alpha$ (default: $\alpha = 0.05/(\#$ of candidate background datasets)), reject the null (exclude j from $\mathcal{B}$).

8: **return** $\mathcal{B}$

9: **end for**

It is also instructive to consider the setting where the target subspace is entirely contained within the candidate background subspace, i.e., $C(\Gamma_0) \subset C(\Gamma_1)$. In this setting, as the contrastive analysis of the target dataset relative to the candidate background is unlikely to yield additional insights, BasCoD is expected to deem the contrastive dimension reduction invalid. As an illustrative example, when $\Gamma_0 = \Gamma_1[\,,1]$, we expect $\operatorname{corr}(\widehat{\Gamma}_1[\,,1], \widehat{P}_0\widehat{\Gamma}_1[\,,1]) \approx 1$, whereas $\operatorname{corr}(\widehat{\Gamma}_1[\,,i], \widehat{P}_0\widehat{\Gamma}_1[\,,i]) \approx 0$ for $i \neq 1$, leading to the rejection of BasCoD's null hypothesis. Combined with the settings in Fig. 4, this underscores BasCoD's capacity to assess the validity of contrastive dimension reduction across varied contexts.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

• Mouse protein expression data. The processed mouse protein expression data used in this study are available at the cPCA GitHub repository. • Perturb-seq data. The processed Perturb-seq data used in this study are available at Figshare through the ContrastiveVI tutorial. • Mouse intestinal single-cell RNA-seq data. The processed mouse intestinal single-cell RNA-seq data used in this study are available at the ContrastiveVI tutorial. • Human Cell Atlas bone marrow (HCA-BM) data. The processed HCA-BM data used in this study are available at the Lamian GitHub repository. • Population-scale single-cell RNA-seq data with ROT treatment. The processed population-scale single-cell RNA-seq data used in this study are available in the Zenodo database under accession code 4333872. • Single-cell RNA-seq data with inflammation treatment. The processed single-cell RNA-seq data used in this study are available at Zenodo: https://zenodo.org/records/18776758. Source data are provided with this paper.

Code availability

The code used to develop the model, perform the analyses and generate results in this study is publicly available and has been deposited in the GitHub repository at https://github.com/keleslab/BasCoD, under the MIT license. The specific version of the code associated with this publication is archived in Zenodo and is accessible via https://doi.org/10.5281/zenodo.18291183[31].

References

1. Jerber, J. et al. Population-scale single-cell rna-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* **53**, 304–312 (2021).
2. Soskic, B. et al. Immune disease risk variants regulate gene expression dynamics during cd4+ t cell activation. *Nat. Genet.* **54**, 817–826 (2022).
3. Dixit, A. et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell* **167**, 1853–1866 (2016).
4. Abid, A., Zhang, M. J., Bagaria, V. K. & Zou, J. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nat. Commun.* **9**, 2134 (2018).
5. Abid, A., Zou, J. Contrastive variational autoencoder enhances salient features. arXiv preprint arXiv:1902.04601 (2019).
6. Severson, K.A., Ghosh, S., Ng, K. Unsupervised learning with contrastive latent variable models. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. 4862–4869 (2019).
7. Weinberger, E., Lin, C. & Lee, S. I. Isolating salient variations of interest in single-cell data with contrastivevi. *Nat. Methods* **20**, 1336–1345 (2023).
8. Weinberger, E., Covert, I., & Lee, S. I. Feature selection in the contrastive analysis setting. *Adv. Neural. Inf. Process. Syst.* **36**, 66102–66126 (2023).
9. Zhang, B., Nyquist, S., Jones, A., Engelhardt, B. E. & Li, D. Contrastive linear regression. *Ann. Appl. Stat.* **19**, 1868 (2025).
10. Ebrahimi, A., Siahpirani, A. F. & Montazeri, H. scin: a contrastive learning framework for single-cell multi-omics data integration. *Brief. Bioinforma.* **26**, bbaf411 (2025).
11. Li, W., Murtaza, G. & Singh, R. sccontrast: A contrastive learning based approach for encoding single-cell gene expression data. *bioRxiv* 2025–04 (2025).

12. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In: *Proc. of the IEEE/CVF conference on computer vision and pattern recognition* 9729–9738 (2020).
13. Hawke, S., Zhang, E., Chen, J. & Li, D. Contrastive dimension reduction: A systematic review. arXiv preprint arXiv:2510.11847 (2025).
14. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome Biol.* **20**, 295 (2019).
15. Ahmed, M. M. et al. Protein dynamics associated with failed and rescued learning in the ts65dn mouse model of down syndrome. *PloS one* **10**, e0119491 (2015).
16. Higuera, C., Gardiner, K. J. & Cios, K. J. Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PloS one* **10**, e0129126 (2015).
17. Norman, T. M. et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**, 786–793 (2019).
18. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. methods* **15**, 1053–1058 (2018).
19. Wilson, D. J. The harmonic mean p-value for combining dependent tests. *Proc. Natl. Acad. Sci.* **116**, 1195–1200 (2019).
20. Trapnell, C. et al. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nat. Biotechnol.* **32**, 381 (2014).
21. Ji, Z. & Ji, H. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids Res.* **44**, e117–e117 (2016).
22. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics* **19**, 1–16 (2018).
23. Hou, W. et al. A statistical framework for differential pseudotime analysis with multiple single-cell rna-seq samples. *Nat. Commun.* **14**, 7286 (2023).
24. DeTomaso, D. & Yosef, N. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell Syst.* **12**, 446–456 (2021).
25. Ficara, F. et al. Pbx1 restrains myeloid maturation while preserving lymphoid potential in hematopoietic progenitors. *J. cell Sci.* **126**, 3181–3191 (2013).
26. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a J. Integr. Biol.* **16**, 284–287 (2012).
27. Murray, C. W. et al. LKB1 drives stasis and C/EBP-mediated reprogramming to an alveolar type II fate in lung cancer. *Nat. Commun.* **13**, 1090 (2022).
28. Lara-Astiaso, D. et al. In vivo screening characterizes chromatin factor functions during normal and malignant hematopoiesis. *Nat. Genet.* **55**, 1542–1554 (2023).
29. Hawke, S., Ma, Y. & Li, D. Contrastive dimension reduction: when and how?. *Adv. Neural Inf. Process. Syst.* **37**, 74034–74057 (2024).
30. Liu, Y. & Xie, J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* **115**, 393–402 (2020).
31. Park, K., Sun, Z., Liao, R., Bresnick, E.H. & Keleş, S. Systematic background selection with bascod enhances contrastive dimension reduction in single cell genomics. *GitHub Repository: BasCoD*https://doi.org/10.5281/zenodo.18291183 (2026).
32. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).

Acknowledgements

We thank Shuchen Yan from the University of Wisconsin–Madison for sharing the processed scRNA-seq dataset from Jerber et al.[1]. We thank Dr. Siqi Shen (Fred Hutchinson Cancer Center) and Coleman Breen (University of Wisconsin–Madison) for insightful discussions. This work was supported by NIH grants R01HG003747 (S.K.) and R21HG012881 (S.K.), and a Chan Zuckerberg Initiative Data Insights Award (S.K.).

Author contributions

K.P. and S.K. conceived the project. K.P. and S.K. designed the research and developed the method. K.P. performed the experiments and simulation studies. K.P. and S.K. contributed to the preparation of the manuscript. R.L. and E.B. generated the single-cell RNA-seq dataset with inflammation treatment. Z.S. processed the dataset and designed experimental ideas involving double-perturbed cells.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-70652-4.

Correspondence and requests for materials should be addressed to Sündüz Keleş.

Peer review information *Nature Communications* thanks Chaojie Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.