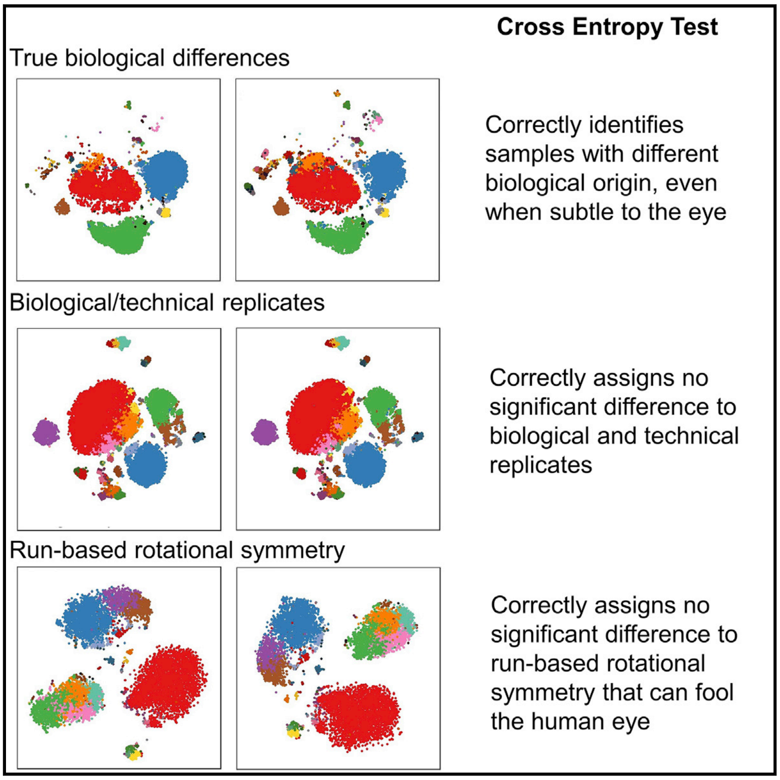
Cell Reports
Methods

Article

A cross entropy test allows quantitative statistical comparison of t-SNE and UMAP representations

Graphical abstract



Cross Entropy Test

True biological differences

Correctly identifies samples with different biological origin, even when subtle to the eye

Biological/technical replicates

Correctly assigns no significant difference to biological and technical replicates

Run-based rotational symmetry

Correctly assigns no significant difference to run-based rotational symmetry that can fool the human eye

Highlights

- A cross entropy test enables evaluation of differences between t-SNE and UMAP projections
- The cross entropy test can distinguish biological variation from technical variation
- The cross entropy test can quantify differences between multiple samples
- Full code and instructions are given for applying the test to single cell datasets

Authors

Carlos P. Roca, Oliver T. Burton, Julika Neumann, ..., Rafael V. Veiga, Stéphanie Humblet-Baron, Adrian Liston

Correspondence

al989@cam.ac.uk

In brief

Dimensionality-reduction tools such as t-SNE and UMAP allow visualizations of single-cell datasets. Roca et al. develop and validate the cross entropy test for robust comparison of dimensionality-reduced datasets in flow cytometry, mass cytometry, and single-cell sequencing. The test allows statistical significance assessment and quantification of differences.

Roca et al., 2023, Cell Reports Methods *3*, 100390
January 23, 2023 © 2022 The Author(s).
https://doi.org/10.1016/j.crmeth.2022.100390

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

A cross entropy test allows quantitative statistical comparison of t-SNE and UMAP representations

Carlos P. Roca,[1,4] Oliver T. Burton,[1,4] Julika Neumann,[2,3] Samar Tareen,[1] Carly E. Whyte,[1] Vaclav Gergelits,[1] Rafael V. Veiga,[1] Stéphanie Humblet-Baron,[2] and Adrian Liston[1,5,*]

[1] Immunology Programme, The Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, UK
[2] VIB Center for Brain and Disease Research, 3000 Leuven, Belgium
[3] KU Leuven - University of Leuven, Department of Microbiology and Immunology, 3000 Leuven, Belgium
[4] These authors contributed equally
[5] Lead contact
*Correspondence: al989@cam.ac.uk

https://doi.org/10.1016/j.crmeth.2022.100390

MOTIVATION Dimensionality-reduction tools, such as t-SNE and UMAP, are frequently used to visualize highly complex single-cell datasets in single-cell sequencing, flow cytometry, and mass cytometry. Despite the ubiquity of these approaches and the clear need for quantitative comparison of single-cell datasets, t-SNE and UMAP have largely remained data visualization tools, with a lack of robust statistical approaches available. We sought to fulfill the need for a statistical test to evaluate the difference between dimensionality-reduced datasets and provide a quantification of differences between multiple datasets.

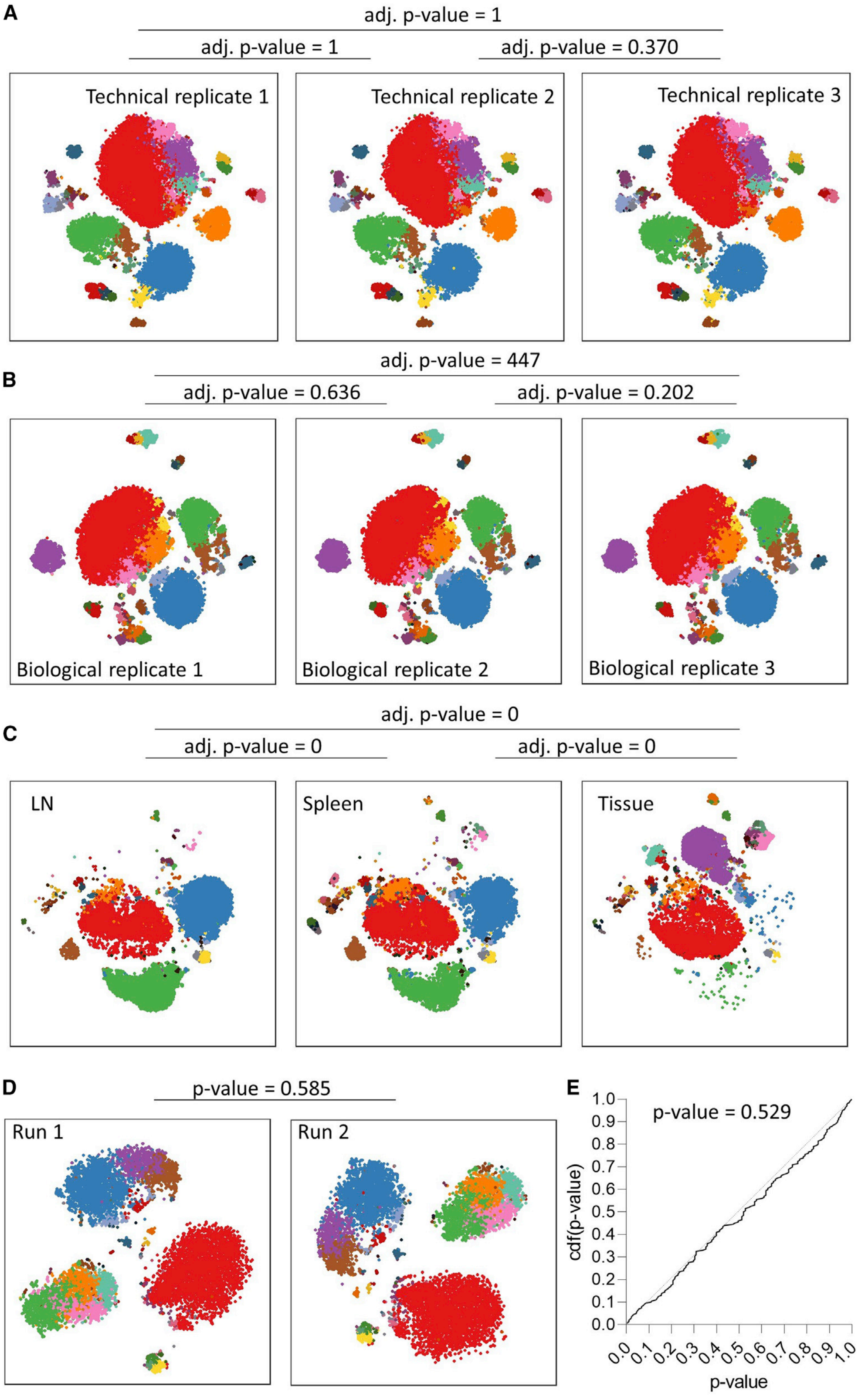
SUMMARY

The advent of high-dimensional single-cell data has necessitated the development of dimensionality-reduction tools. t-Distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) are the two most frequently used approaches, allowing clear visualization of complex single-cell datasets. Despite the need for quantitative comparison, t-SNE and UMAP have largely remained visualization tools due to the lack of robust statistical approaches. Here, we have derived a statistical test for evaluating the difference between dimensionality-reduced datasets using the Kolmogorov-Smirnov test on the distributions of cross entropy of single cells within each dataset. As the approach uses the inter-relationship of single cells for comparison, the resulting statistic is robust and capable of identifying true biological variation. Further, the test provides a valid distance between single-cell datasets, allowing the organization of multiple samples into a dendrogram for quantitative comparison of complex datasets. These results demonstrate the largely untapped potential of dimensionality-reduction tools for biomedical data analysis beyond visualization.

INTRODUCTION

Single-cell technologies are capable of generating vast datasets where multiple parameters (tens to thousands) are assessed on large numbers of cells (hundreds to tens of millions). With the rapid advance of these technologies, increasing the ease of data generation while decreasing the cost,[1] data analysis approaches are rapidly becoming the limiting factor.[2] Low-dimensional visualizations are an attractive entry point to single-cell analysis, with dimensionality-reduction tools providing an overview of the data. Beyond the science communication advantages of low-dimensional representation, such exploratory analyses often identify pattern distortions worthy of detailed analysis. Unfortunately, there is a shortage of statistical tools with which to detect such pattern distortions, forcing the use of subjective assessment by ‘‘eye balling’’ the data.

Traditional approaches, such as principal-components analysis (PCA) and multidimensional scaling (MDS) produce linear-scaled, low-dimensional representations, which are of high utility for comparing datapoints that are most different from each other. By contrast, the structure of most single-cell experiments places the emphasis on identifying cells that are highly similar to each other; for these processes, non-linear approaches provide superior resolution. t-Distributed stochastic neighbor embedding (t-SNE) is the most commonly used non-linear dimensionality-reduction algorithm for single-cell biology. In its common usage for visualizing high-dimensionality single-cell data, the algorithm starts with the single cells distributed at random points along a



(legend on next page)

Gaussian distribution in transformed space. In an iterative process, the cells move along a cost gradient, which provides a penalty for mismatch between the distances between two cells in the original high-dimensional space versus the representational low-dimensional space.[3] When sufficient iterations have occurred to reach stability, the outcome produces clusters of similar cells based on the input data. Membership of a cluster indicates shared properties; however, the non-linear nature of the penalty cost does not allow relationships to be inferred by the relative positioning of the clusters.

The wide-scale use of t-SNE to visualize single-cell datasets has spurred the development of alternative non-linear dimensionality-reduction tools. Uniform manifold approximation and projection (UMAP) takes a similar iterative approach in positioning cells based on mismatch between high- and low-dimensional data; however, it uses a distinct mathematical basis for the calculations. Cell positions are calculated in topological space, with mismatch between manifolds driving the iterative process.[4] While the results are broadly consistent with t-SNE visualization, UMAP has superior run times and is argued to better preserve the global distances between cell clusters,[4] although this is due to the choice of initialization defaults rather than the UMAP algorithm.[5] Density-preserving SNE (den-SNE) and densMAP are modifications of t-SNE and UMAP, respectively, that attempt to solve the problem inherent to these approaches whereby cluster size is driven by the number of cells in the cluster. den-SNE and densMAP use an auxiliary function to optimize cell density between high- and low-dimensional space, with the size of clusters in the resulting low-dimensional plots reflecting the degree of heterogeneity present in the cluster.[6] Additional non-linear tools have been built, such as TMAP,[7] to deal with limitations of t-SNE when applied beyond the single-cell space.

The near ubiquity of t-SNE or analogous tools in the visualization of single-cell sequencing data demonstrates the utility of low-dimensional representations of complex data. Despite this, t-SNE is generally not used as an analytical tool, with downstream analysis most commonly based on subsetting data and treating as pseudo-bulk populations. t-SNE and UMAP representations are used primarily as a way of displaying high-parameter data. Both techniques, however, use K-nearest neighbor (KNN) graph-based metrics to maintain the local relationships of points from the original high-parameter space, thus generating good representations of the data structure while avoiding distortions caused by asymmetry and outliers. t-SNE and UMAP therefore carry important information about the high-parameter data structure that is lost in downstream cluster analysis. Here, we sought to develop a statistical replacement for the “eye-balling” approach to identifying different patterns present in the output of t-SNE and UMAP projects. Taking advantage of cross entropy as a measure of the relationship between multiple datapoints (e.g., the coordinates of individual cells following a dimensionality reduction), we developed a cross entropy test allowing robust statistical comparisons of t-SNE representations. The test has appropriate sensitivity, with the identification of biological differences and the dismissal of difference in technical replicates, biological replicates, or repeat t-SNE runs. The test responds to differences in either inter-cluster frequency or intra-cluster phenotype shifts, and differences between multiple t-SNE plots can be quantified for the production of dendrograms. The cross entropy test is broadly applicable to single-cell technologies, including flow cytometry, mass cytometry, and single-cell sequencing, and can be performed on either t-SNE and UMAP transformations, providing a highly versatile statistical tool to the single-cell toolkit.

RESULTS

A cross entropy test provides a robust statistical test for t-SNE comparison

t-SNE is an iterative algorithm for dimensionality reduction. In its common usage for visualizing high-dimensionality single-cell data, the cost gradient of t-SNE places greater weight on pairs of cells close to each other, with medium- and long-range pairs ignored. The low-dimensional representation of high-dimensional data makes t-SNE an attractive visualization tool, yet its value as an analytical tool has been hampered by the paucity of statistical tests. We sought to derive a statistical test capable of distinguishing biological differences in single-cell t-SNE representations while being robust against false detection of differences in technical replicates or the seed-dependent variation in t-SNE generation. As the t-SNE algorithm is driven by the cross entropy of the individual cells in the dataset, and t-SNE fixes the average point entropy, each t-SNE can be considered a distribution of cross entropy divergences. Deriving a distribution of cross entropy divergences per t-SNE plot therefore allows the use of the Kolmogorov-Smirnov (KS) test to evaluate the degree of difference between two or more t-SNE plots (see STAR Methods).

A robust t-SNE statistical test should (1) reliably fail to detect differences when comparing technical replicates; (2) reliably fail to detect differences when comparing biological replicates; (3) reliably identify differences when comparing different biological samples; and (4) not get fooled by different t-SNE runs on the same sample. The cross entropy test was beta tested by collaborating immunologists on >130 datasets over the course of 24 months of validation. We initially focused on flow cytometry datasets as

Figure 1. A cross entropy test provides a robust statistical test for t-SNE comparison

The MUS flow cytometry dataset was used to test the sensitivity of the cross entropy test.

(A) Three technical replicates of a single splenocyte sample were plotted on a t-SNE with FlowSOM clustering in overlay. p values for cross entropy testing are shown.

(B) t-SNE plots of splenocytes from three individual age-/sex-matched mice (biological replicates). p values for cross entropy testing are shown.

(C) t-SNE plots from lymph nodes, spleen, and tissue (small intestinal lamina propria). p values for cross entropy testing are shown.

(D) Two independent t-SNE runs of lymph node cells. p values for cross entropy testing are shown.

(E) Cumulative distribution function of p values obtained from 200 comparisons of independent t-SNE runs of the same lymph node sample. Statistical comparison with uniform distribution using KS test.

See also Figure S1.

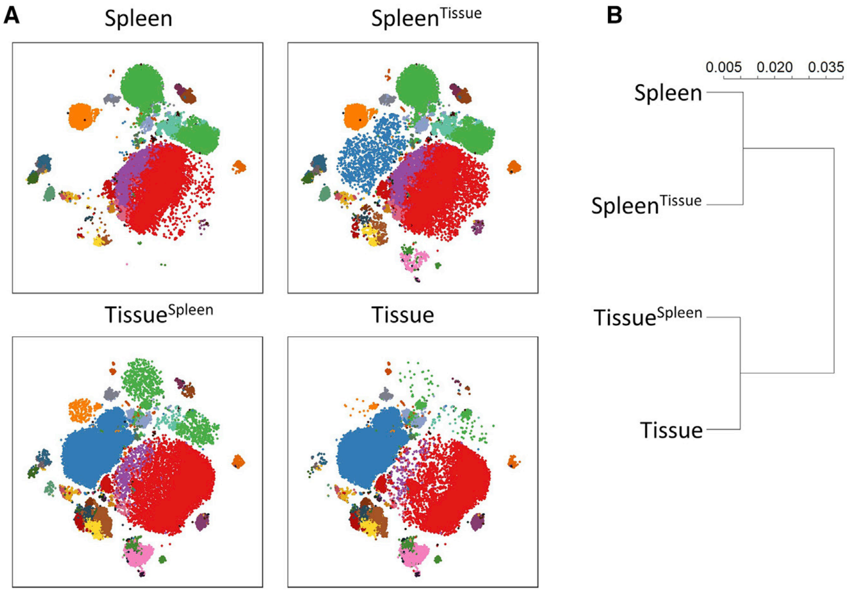


Figure 2. L^{∞} provides a quantitative comparison of different t-SNE visualizations

In order to create a series of samples with defined degrees of biological difference, we started with the spleen and tissue (small intestinal lamina propria) samples from the MUS sample set and generated two artificial samples, one composed of 90% spleen and 10% tissue cells (spleen^{tissue}) and one composed of 10% spleen and 90% tissue cells (tissue^{spleen}).

(A) t-SNE plots showing spleen, spleen^{tissue}, tissue^{spleen}, and tissue datasets.

(B) Dendrogram based on the L^{∞} cross entropy distance.

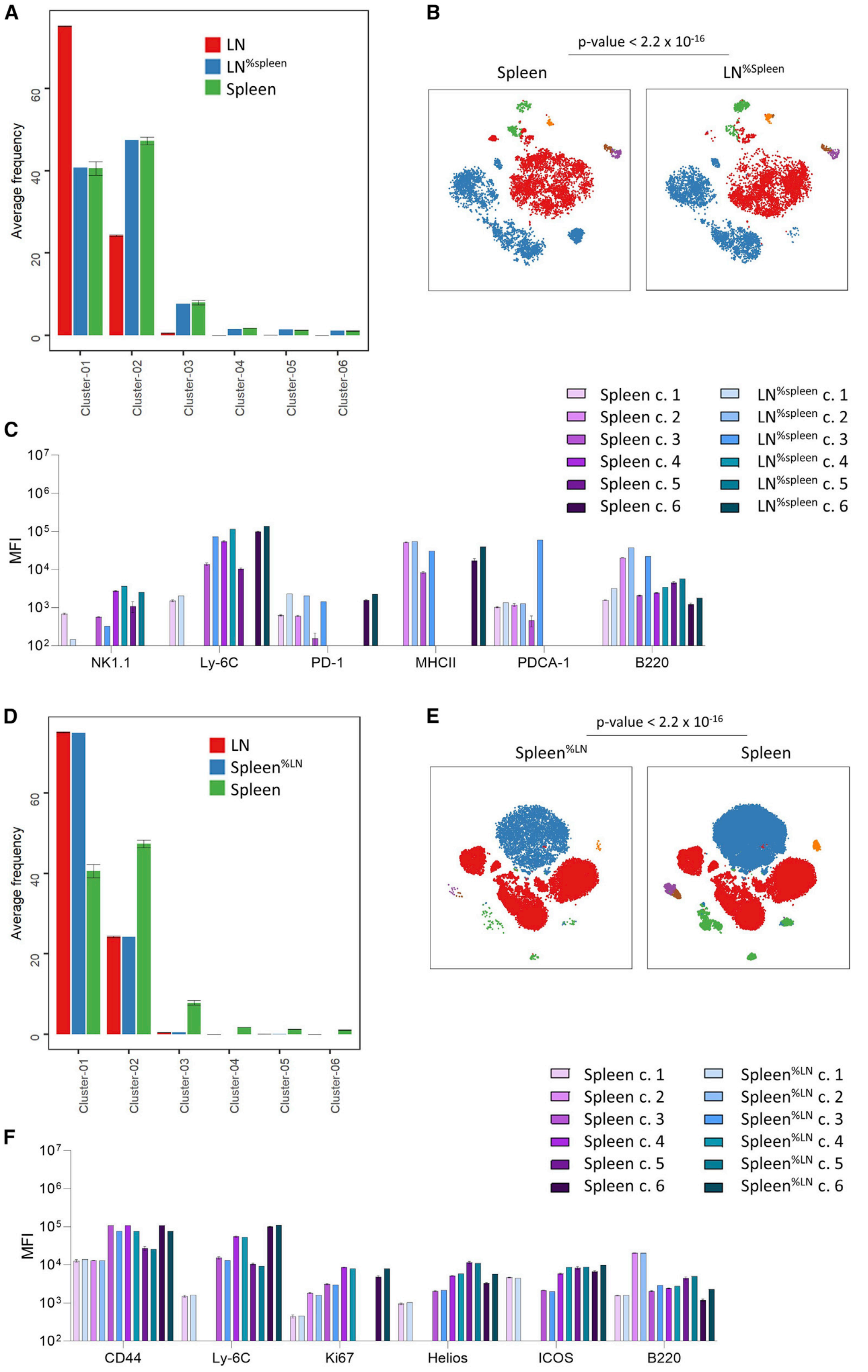
the very high cell number achievable by flow cytometry provides superior ability to challenge and validate a statistical test. As an example of the robustness observed, we present here a high-dimensional flow cytometry dataset based on immunological profiling of lymphocytes from the lymph nodes, spleen, and tissues from the C57BL/6 inbred mouse strain (MUS dataset). This design allows us to compare technical replicates (splitting a single sample), biological replicates (comparing analogous samples from different mice), and biological samples (comparing lymphocytes from different tissues of the same mice). First, to test for robustness in technical replicates, we compared the t-SNE plots generated by splitting a single sample of splenocytes with FlowSOM clustering, allowing comparative visualization of cell identity (Figure 1A). Visually similar plots were generated, with cross entropy test values showing p values of 0.370–1, supporting the null hypothesis of no difference (Figure 1A). Second, to test for robustness in biological replicates, we compared the t-SNE plots generated by multiple samples of splenocytes from age-/sex-matched mice. Again, visually similar plots were generated, with the cross entropy test p values of 0.202–0.636 supporting the null hypothesis (Figure 1B). Third, to test for the capacity to identify true biological differences, we compared the t-SNE plots generated on lymphocytes profiled from the spleen, lymph node, and tissues of mice. Lymphocytes from the lymph nodes and spleen are phenotypically similar, but the relative proportions of each population vary between the sites, while non-lymphoid tissue lymphocytes are phenotypically distinct. t-SNE visualizations mirrored this biological background knowledge, with highly significant cross entropy test values across each comparison (Figure 1C). Fourth, we produced independent runs of t-SNE generation on the lymph node sample, generating similar t-SNE plots with the characteristic rotational symmetry (Figure 1D). Despite the visual disparity being greater than the highly significant spleen-lymph node comparison, the cross entropy test gave a test p value of 0.585, supporting the null hypothesis (Figure 1D). Multiple runs of this false comparison gave an appropriate sensitivity of the cross entropy test, with neither over- nor under-reporting of false positives (Figure 1E). Furthermore, we find that the cross entropy test is robust to changes in t-SNE settings, such as perplexity, iteration values, or analysis of independent runs (Figure S1). Together, this example dataset demonstrates the robust ability of the cross entropy test to distinguish biological signal from noise.

Calculation of relative distance between t-SNE comparisons

A key advantage of the utilization of the K-S test in the cross entropy test is the ability to calculate the L^{∞} distance, which can be used as a measure of the difference between t-SNE plots. In order to test the ability of the L^{∞} distance to correctly assign distances between t-SNE plots, we created two artificial datasets of known distance from the splenic and tissue flow cytometry data. First, we created a spleen^{tissue} dataset where 90% of the events were extracted from the spleen sample and 10% of events were extracted from the tissue sample prior to concatenation to create a single merged sample. Second, we created the reciprocal tissue^{spleen} dataset with 90% tissue events and a 10% splenic event spike in. Generation of t-SNE plots using the parental samples and constructed spike-in samples provides visual confirmation of the intermediate status of the artificial datasets (Figure 2A). Calculation of the L^{∞} distance allowed the construction of a dendrogram that correctly assigned spleen^{tissue} as biologically closer to spleen and tissue^{spleen} as biologically closer to tissue (Figure 2B). This demonstrates the L^{∞} distance as a tool for gauging the relative closeness of t-SNE plots when comparing multiple samples.

A cross entropy test responds to both quantitative and qualitative changes in single-cell phenotype

Next, we sought to determine the sensitivity of the cross entropy test to biological differences based on inter-cluster frequency versus intra-cluster phenotype. Currently, most single-cell analysis ultimately measures either the frequency of individual clusters (typically annotated based on FlowSOM[8] or similar clustering tools[9]) or treats individual clusters as pseudo-bulk populations for comparison of gene/protein expression changes. In principle, as t-SNE incorporates the full phenotype of individual cells, a useful



(legend on next page)

t-SNE-based statistical test should be sensitive to both changes in inter-cluster frequency and in intra-cluster phenotype. To test this, we generated artificial datasets derived from the MUS dataset of lymphocytes from the spleen and lymph node, which exhibit changes in both inter-cluster frequency and phenotype. First, we used selective downscaling to create a lymph$^{\%spleen}$ dataset where the number of cells in each major FlowSOM cluster were normalized to match those in the spleen dataset while only using cells from the lymph node data. Comparison of the spleen sample with the rescaled lymph$^{\%spleen}$ sample generated t-SNE plots with cluster frequencies normalized between spleen and lymph node (LN) (Figure 3A). Despite this, the cross entropy test identified the samples as significantly different (Figure 3B) based on the phenotype difference within clusters (Figure 3C). Next, we created a reciprocal artificial, spleen$^{\%lymph}$, taking biological replicates of the spleen sample and selectively downscaling populations to create cluster frequencies similar to that of the LN (Figure 3D). To determine whether the cross entropy test could detect differences solely based on cluster frequency, we compared spleen with spleen$^{\%lymph}$. The test found significant differences between the two samples (Figure 3E) despite identical phenotypes (Figure 3F). These results demonstrate that the cross entropy test is sensitive to both qualitative and quantitative changes between clusters.

The cross entropy test has broad utility in comparison of dimensionally reduced single-cell datasets

Having demonstrated the validity of the cross entropy test on a mouse-based, high-dimensional flow cytometry panel, we sought to test it on the independent single-cell technologies of mass cytometry and single-cell sequencing from human samples. We used available datasets generated from healthy humans and patients with COVID-19 based on mass cytometric analysis of the peripheral blood[10] and 10x single-cell sequencing analysis of bronchoalveolar lavage.[11] Using the mass cytometry dataset, we compared peripheral blood lymphocyte subsets from 12 patients at different time points, namely admission to the intensive care unit (ICU), during their stay at the ICU (intermediate), and upon discharge from the ICU. We identified major subsets based on characteristic marker expression, with the dataset reproducing key features of severe COVID-19, such as increased neutrophils and decreased CD4^{+} as well as CD8^{+} T cells[12,13] (Figure 4A). The L^{∞} distance between the three time points suggests that the overall immune landscape during the ICU stay more closely resembles that of the admission time point (Figure 4A). However, when performing analysis of the monocyte subset alone, we observed a closer resemblance of the intermediate time point to that of discharge (Figure 4B). These results are in accordance with findings from the original publication, generated through traditional gating and marker analysis, as well as other studies that report monocytes as the first immune population to recover following severe COVID-19.[10,14] Turning to the single-cell sequencing dataset, we used a 10x single-cell sequencing comparison of bronchoalveolar lavage from patients with COVID and patients with non-COVID pneumonia.[11] Using the annotated cell clusters, we compared the transcriptional profile of COVID with non-COVID pneumonia cells for each of the epithelial, neutrophil, monocyte/macrophage, CD4 T cell, CD8 T cell, dendritic cell, B cell, and natural killer (NK) cell clusters (Figure 4C). Highly significant changes were observed in the t-SNE cross entropy of COVID versus non-COVID epithelial, neutrophil, monocyte/macrophage, CD4 T cell, and CD8 T cell, in accordance with the detailed classical analysis used.[11] Use of the L^{∞} distance (Figure 4D) identified the largest change being in the neutrophil compartment, consistent with the results from multiple studies,[11,15,16] and found the degree of change in CD8 T cells to be greater than that observed in CD4 T cells, again consistent with traditional analysis.[11] Together, these results demonstrate that the cross entropy test is compatible with t-SNE analysis generated by multiple independent technologies, with a simple test recapitulating many of the key features identified through high-depth traditional methods.

The cross entropy test accurately detects differences in comparisons of UMAP dimensionally reduced single-cell datasets

Recently, alternative non-linear data visualization tools have been generated for single-cell analysis. UMAP performs the same basic functions as t-SNE with regards to using two dimensions to separate cells based on complex multidimensional phenotype data.[4] Despite the different mathematical basis for generating the dimensionality reduction, cross entropy dictates the separation of UMAP as well as t-SNE. We therefore performed validation tests for the cross entropy test on UMAP representations. Using the MUS dataset, we produced UMAP plots comparing technical replicates (Figure 5A), biological replicates (Figure 5B), biological differences (Figure 5C), and repeat UMAP runs (Figure 5D). In each case, the cross entropy test produced the appropriate conclusions: failing to detect significant differences between technical and biological replicates while appropriately detecting differences in biologically distinct samples. The cross entropy test was also appropriately powered, with neither over- nor under-sensitivity detected (Figure 5E). As with the utility in the t-SNE test, the UMAP cross entropy test

Figure 3. Cross entropy test detects both qualitative and quantitative changes in single-cell datasets

In order to create a series of samples with biological difference restricted only to qualitative or quantitative changes, we started with the spleen and lymph node samples from the MUS sample set and generated artificial samples. Artificial samples were generated by first defining and exporting the clusters in each parental sample, then downscaling to match the new proportions and concatenating to reconstruct the artificial sample. This resulted in a sample where lymph node clusters were rescaled such that each cluster was the same frequency as the analogous cluster in spleen (LN$^{\%spleen}$) and the reciprocal sample where spleen clusters were rescaled to the same frequency as the analogous cluster in LNs (spleen$^{\%LN}$).

(A) FlowSOM cluster frequencies in spleen and LN$^{\%spleen}$ samples.

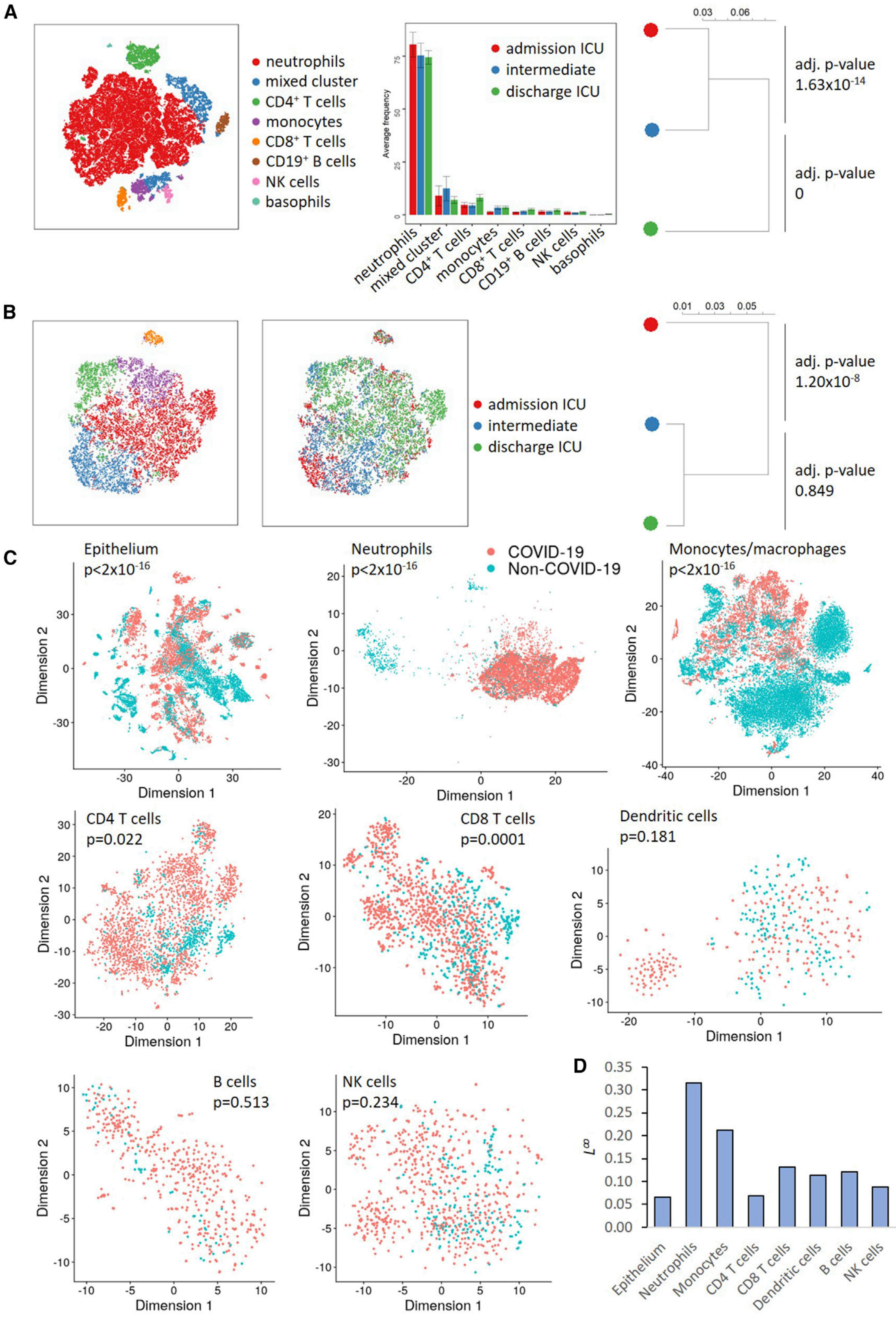
(B) t-SNE plots of in spleen and LN$^{\%spleen}$ samples. p values for cross entropy testing are shown.

(C) Bar chart showing selected mean fluorescence intensity (MFI) values per cluster for spleen and LN$^{\%spleen}$ samples.

(D) FlowSOM cluster frequencies in spleen, spleen$^{\%LN}$, and LN samples.

(E) t-SNE plots of spleen$^{\%LN}$ and spleen samples. p values for cross entropy testing are shown.

(F) Bar chart showing MFI values for markers in clusters from spleen$^{\%LN}$ and spleen samples.



(legend on next page)

responds to both shifts in subcluster cellular phenotype and cluster frequency, using the artificial samples spleen$^{\%LN}$ and LN$^{\%spleen}$, compared with the parental spleen sample (Figure 5F). Cross entropy testing of UMAP plots was not sensitive to changes in UMAP iterations or the number of neighbors, nor to the number of dimensions used in the dimensionality reduction (Figure S2). Furthermore, application of the cross entropy test to additional dimensionality-reduction approaches, giving highly similar p values regardless of the reduction technique applied (Figure S3), suggests it may have broad utility in this area. The recent proliferation and rapid turnover of these embedding algorithms, however, puts validation outside the scope of this work. These results demonstrate the broad utility of the cross entropy test for multidimensionally scaled single-cell data across both t-SNE and UMAP approaches.

DISCUSSION

The advent of single-cell sequencing has led to an urgent need for analyzing and visualizing the complex high-dimensional datasets generated. While similar utility can be found in flow and mass cytometry datasets, the lower number of parameters and increased reliability of individual parameters allowed the continuation of the pairwise gating approach, derived from the historical progress of biological knowledge and technical capacity.[9] Both t-SNE and UMAP approaches fulfill the visualization role created by single-cell technologies, especially when paired with clustering tools such as FlowSOM,[8] to provide biological meaning to the two-dimensional space that represents complex shifts in phenotype or transcriptome. At an analytical level, however, t-SNE and UMAP have been under-utilized. Most analytical approaches treat single-cell data as pseudo-bulk data. Single-cell data are typically only used to generate clusters as an intermediate step.[9] Downstream analysis typically compares cluster size change between samples or compares cluster-aggregated data for cluster-level statistics (such as mean fluorescence intensity in flow cytometry or average transcript expression in single-cell sequencing). This large information loss provides an opportunity for the generation of new statistical tests that include single-cell information in the sample comparisons.

Unlike the commonly used pseudo-bulk approach, t-SNE and UMAP encapsulate the nuance of data at the single-cell level. These methods, being topological and non-metric, will never fully preserve the distances in the original data but concentrate on the closeness of closest points. Cross entropy captures, as a single statistic, both changes in the relative frequency of cells of different phenotype classes and also the shift of phenotype of cells within phenotype classes. We have used this tool to quantify the complex shifts in phenotypic markers present in T cells of different regions of brain or transcriptional shifts occurring in microglia in the presence or absence of CD4 T cells.[17] Valid targets for the cross entropy test include any high-dimensional single-cell dataset routinely generated by single-cell sequencing, flow cytometry, or mass cytometry in fields ranging from immunology[18] to neuroscience[19] to cancer.[20] The growth of single-cell technologies expands the potential range to techniques such as CODEX or other high-dimensional imaging technology,[21] single-cell proteomics,[22] or mutation analysis in single-cell sequencing.[23] More broadly, t-SNE and UMAP use can be implemented in cases where the cell is not the individual unit. For example, using people as the individual unit, high-dimensional data such as immune response,[24] genomic variation,[25] or microbiome composition[26] can drive the algorithm. The cross entropy test would then allow statistical comparison of disparate groups (e.g., based on disease state) to identify differences in the underlying data. In each case, full utility of the data generated requires a statistical tool that can compare the high-dimensional data distributions between two groups.

The use of the K-S tests, and the ability to quantify differences between samples, in the cross entropy test provides a potential route to incorporate t-SNE or UMAP into routine diagnostics. Flow cytometric analysis of blood leukocytes is routinely used as a single-cell technology in the clinic;[27] however, data analysis is largely limited to calculate population frequencies compared with reference populations. Consortia such as EuroFlow are standardizing protocols to enable a broader uptake and cross center comparison.[28] The use of the cross entropy L^{∞} distance would enable a simple analysis for deviations in leukocyte number or activation status by determining whether an individual sample mapped to the healthy or disease nodes of a t-SNE-based dendrogram. For example, we used cross entropy to draw a dendrogram of T cell phenotypes in individuals with mild or severe COVID.[29] Broader use of a cross entropy test approach within the diagnostics field may provide higher sensitivity to the detection of atypical immunological disorders or hematological malignancies, as the integrated statistical would incorporate the detection of both aberrant marker expression and the development of rare/unusual populations not routinely gated for.[30] While flow cytometry is the currently the dominant single-cell technology in diagnostics, both mass cytometry[31] and single-cell sequencing[32] are actively being developed for diagnostics and may soon become routine; parallel development of the cross entropy statistic may enhance the utility and sensitivity of these new technologies.

Figure 4. Cross entropy test provides utility on mass cytometry and single-cell sequencing datasets

(A) Whole blood from patients with COVID-19 sampled upon admission, during their stay (intermediate), and upon discharge from the ICU was analyzed by CyTOF. Total cells were clustered using the t-SNE algorithm with FlowSOM clusters overlaid and annotated (left). Bar chart represents the mean frequency of each cluster per condition (middle). Dendrogram showing the comparative similarity between conditions based on the KS statistic (right).

(B) t-SNE run on monocytes from patients with COVID-19 as in (A) overlaid with FlowSOM clusters (left) and events per condition (middle). Dendrogram showing the comparative similarity between conditions based on the KS statistic (left).

(C) Bronchoalveolar lavage from patients with COVID and patients with non-COVID pneumonia was assessed by 10x single-cell sequencing. Cells were annotated as epithelium, neutrophils, monocytes/macrophages, CD4 T cells, CD8 T cells, dendritic cells, B cells, and NK cells based on transcriptional profiles. For each cell type, cells were clustered using the t-SNE algorithm with patient source overlaid.

Cross entropy p value and (D) L^{∞} distance between patients with and without COVID and for each cell type.

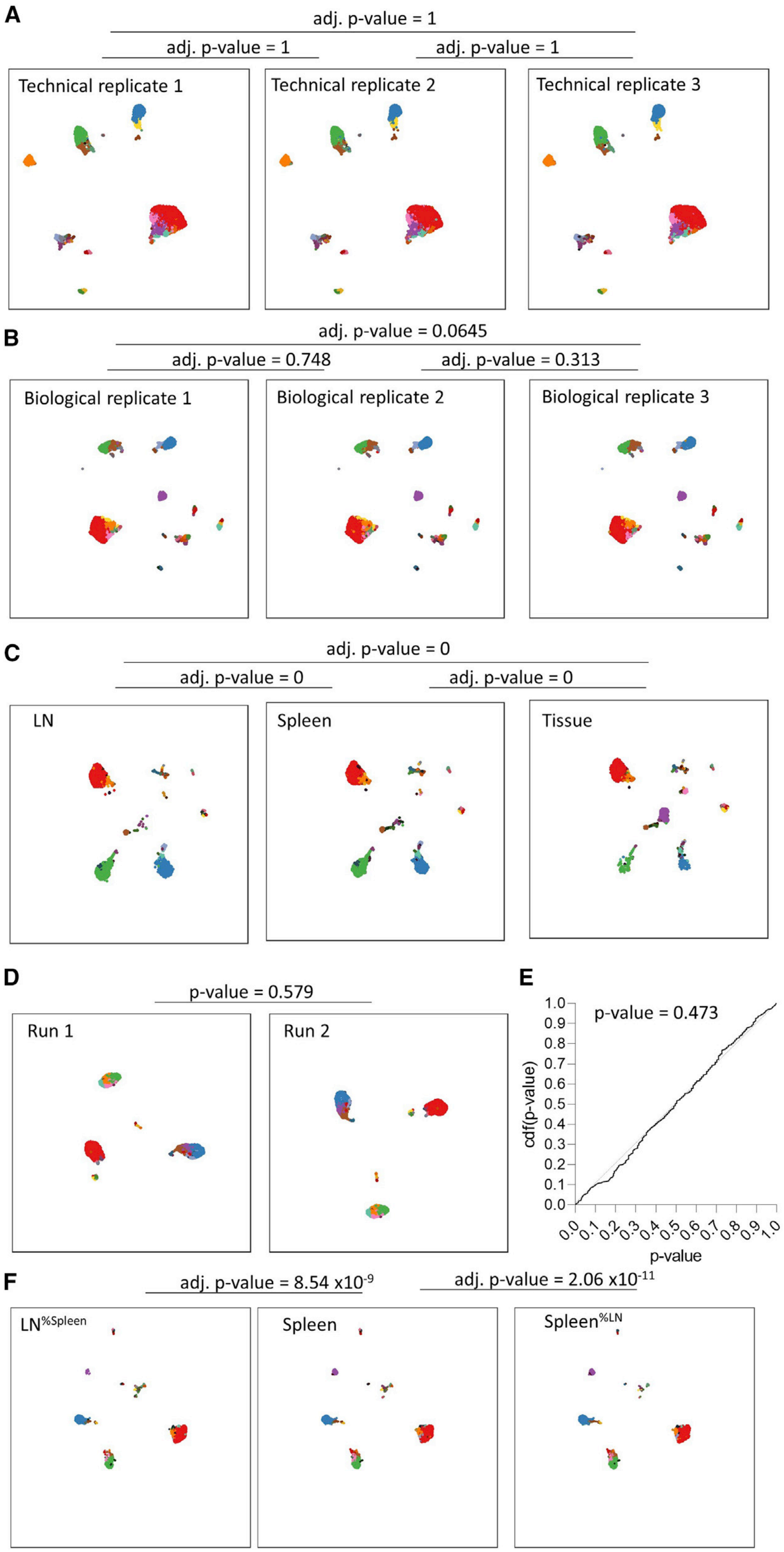


Figure 5. A cross entropy test provides a robust statistical test for UMAP comparison

The MUS flow cytometry dataset was used to test the sensitivity of the cross entropy test.

(A) Three technical replicates of a single splenocyte sample were plotted on a UMAP with FlowSOM clustering in overlay. p values for cross Entropy testing are shown.

(B) UMAP plots of splenocytes from three individual age-/sex-matched mice (biological replicates). p values for cross entropy testing are shown.

(C) UMAP plots from LNs, spleen, and tissue (small intestinal lamina propria). p values for cross entropy testing are shown.

(D) Two independent UMAP runs of LN cells. p values for cross entropy testing are shown.

(E) Cumulative distribution function of p values obtained from 400 comparisons of independent UMAP runs of the same LN sample. Statistical comparison with uniform distribution using KS test.

(F) In order to compare samples with biological difference restricted only to qualitative or quantitative changes, we used the previously generated artificial samples where LN clusters were rescaled to splenic frequencies (LN$^{\%spleen}$) or splenic clusters were rescaled to the LN frequencies (spleen$^{\%LN}$) and compared both against the parental spleen sample in UMAP. p values for cross entropy testing are shown.

See also Figures S2 and S3.

Limitations of the study

Despite the high utility of the cross entropy test for calculating the significance of differences between two or more dimensionality-reduced plots, we caution against the misuse of a p value as a measure of biological meaningfulness. The power of cross entropy tests is dependent on cell number, and thus even biologically distinct samples can return non-significant p values if the number of cells is highly limiting. In practice, sample sizes in the low thousands range can return non-significant p values if the biological difference is slight, while datasets above 10,000 cells are highly powered for even sensitive change. Conversely, very high cell numbers can provide sufficient power to capture the differences between biological replicates, which, while real, are rarely considered to be biologically meaningful. In the datasets tested here, for high cell number analysis, a threshold of 0.001 excludes biological replicates while successfully identifying biological differences. We suggest that analyses include biologically distinct positive controls and the use of the L^∞ distance, in addition to the p value calculation. The L^∞ distance, as a measure of difference magnitude rather than statistical difference, is less sensitive to cell number changes, and the use of a biologically distinct positive control provides a reference point for the magnitude of the differences observed. Negative controls, such as technical replicates, can also provide utility in that a significant difference between technical replicates should be interpreted as unacceptable technical variability, reducing confidence in any sample-to-sample variation identified by the analysis. We also caution that the cross entropy test does not aim to evaluate questions regarding preservation of distance or metric but rather those of the similarity between t-SNE or UMAP representations. Neither does cross entropy, used with any dimensionality-reduction tool, replace the value of full analysis of the original high-dimensional dataset; the choice to use a dimensionality-reduction tool such as t-SNE and UMAP should be made with awareness of the limitation of these methods. Finally, as with all statistical tests, the utility of the analysis is fully dependent on the quality of the data and experimental design and the appropriateness of the test implementation.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Mice
- METHOD DETAILS
  - ○ t-SNE representation of multidimensional data
  - ○ Distributions of entropy and cross entropy
  - ○ Test on distributions of cross entropy
  - ○ Additional dimensionality reduction analysis
  - ○ General implementation details
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2022.100390.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 779295. This work was also supported by the ERC Consolidator Grant TissueTreg (to A.L.), the Biotechnology and Biological Sciences Research Council through BB/S019189/1, Institute Strategic Program grant funding BBS/E/B/000C0427 and BBS/E/B/000C0428, and the Biotechnology and Biological Sciences Research Council Core Capability Grant to the Babraham Institute. The authors thank the Babraham Institute bioinformatics and flow cytometry core facilities.

AUTHOR CONTRIBUTIONS

The study was conceived by C.P.R. and A.L. C.P.R. developed and tested the cross entropy test. O.T.B., C.E.W., J.N., S.T., and S.H.-B. provided the datasets and evaluated results. V.G. and R.V.V. provided additional mathematical expertise. The manuscript was written by C.P.R. and A.L. and revised and approved by all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 3, 2022
Revised: October 29, 2022
Accepted: December 20, 2022
Published: January 13, 2023

REFERENCES

1. Aldridge, S., and Teichmann, S.A. (2020). Single cell transcriptomics comes of age. Nat. Commun. *11*, 4307. https://doi.org/10.1038/s41467-020-18158-5.

2. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biol. *21*, 31. https://doi.org/10.1186/s13059-020-1926-6.

3. van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

4. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.

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

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

7. Probst, D., and Reymond, J.L. (2020). Visualization of very large high-dimensional data sets as minimum spanning trees. J. Cheminform. *12*, 12. https://doi.org/10.1186/s13321-020-0416-x.

8. Quintelier, K., Couckuyt, A., Emmaneel, A., Aerts, J., Saeys, Y., and Van Gassen, S. (2021). Analyzing high-dimensional cytometry data using FlowSOM. Nat. Protoc. *16*, 3775–3801. https://doi.org/10.1038/s41596-021-00550-0.

9. Cheung, M., Campbell, J.J., Whitby, L., Thomas, R.J., Braybrook, J., and Petzing, J. (2021). Current trends in flow cytometry automated data analysis software. Cytometry A. *99*, 1007–1021. https://doi.org/10.1002/cyto.a.24320.

10. Penttilä, P.A., Van Gassen, S., Panovska, D., Vanderbeke, L., Van Herck, Y., Quintelier, K., Emmaneel, A., Filtjens, J., Malengier-Devlies, B., Ahmadzadeh, K., et al. (2021). High dimensional profiling identifies specific immune types along the recovery trajectories of critically ill COVID19 patients. Cell. Mol. Life Sci. *78*, 3987–4002. https://doi.org/10.1007/s00018-021-03808-8.

11. Wauters, E., Van Mol, P., Garg, A.D., Jansen, S., Van Herck, Y., Vanderbeke, L., Bassez, A., Boeckx, B., Malengier-Devlies, B., Timmerman, A., et al. (2021). Discriminating mild from critical COVID-19 by innate and adaptive immune single-cell profiling of bronchoalveolar lavages. Cell Res. *31*, 272–290. https://doi.org/10.1038/s41422-020-00455-9.

12. Giamarellos-Bourboulis, E.J., Netea, M.G., Rovina, N., Akinosoglou, K., Antoniadou, A., Antonakos, N., Damoraki, G., Gkavogianni, T., Adami, M.E., Katsaounou, P., et al. (2020). Complex immune dysregulation in COVID-19 patients with severe respiratory failure. Cell Host Microbe *27*, 992–1000.e3. https://doi.org/10.1016/j.chom.2020.04.009.

13. Qin, C., Zhou, L., Hu, Z., Zhang, S., Yang, S., Tao, Y., Xie, C., Ma, K., Shang, K., Wang, W., and Tian, D.S. (2020). Dysregulation of immune response in patients with coronavirus 2019 (COVID-19) in Wuhan, China. Clin. Infect. Dis. *71*, 762–768. https://doi.org/10.1093/cid/ciaa248.

14. Trombetta, A.C., Farias, G.B., Gomes, A.M.C., Godinho-Santos, A., Rosmaninho, P., Conceição, C.M., Laia, J., Santos, D.F., Almeida, A.R.M., Mota, C., et al. (2021). Severe COVID-19 recovery is associated with timely acquisition of a myeloid cell immune-regulatory phenotype. Front. Immunol. *12*, 691725. https://doi.org/10.3389/fimmu.2021.691725.

15. Shaath, H., Vishnubalaji, R., Elkord, E., and Alajez, N.M. (2020). Single-cell transcriptome analysis highlights a role for neutrophils and inflammatory macrophages in the pathogenesis of severe COVID-19. Cells *9*, 2374. https://doi.org/10.3390/cells9112374.

16. Park, J.H., and Lee, H.K. (2020). Re-Analysis of single cell transcriptome reveals that the NR3C1-CXCL8-neutrophil Axis determines the severity of COVID-19. Front. Immunol. *11*, 2145. https://doi.org/10.3389/fimmu.2020.02145.

17. Pasciuto, E., Burton, O.T., Roca, C.P., Lagou, V., Rajan, W.D., Theys, T., Mancuso, R., Tito, R.Y., Kouser, L., Callaerts-Vegh, Z., et al. (2020). Microglia require CD4 T cells to complete the fetal-to-adult transition. Cell *182*, 625–640.e24. https://doi.org/10.1016/j.cell.2020.06.026.

18. Guilliams, M., Dutertre, C.A., Scott, C.L., McGovern, N., Sichien, D., Chakarov, S., Van Gassen, S., Chen, J., Poidinger, M., De Prijck, S., et al. (2016). Unsupervised high-dimensional analysis aligns dendritic cells across tissues and species. Immunity *45*, 669–684. https://doi.org/10.1016/j.immuni.2016.08.015.

19. Armand, E.J., Li, J., Xie, F., Luo, C., and Mukamel, E.A. (2021). Single-cell sequencing of brain cell transcriptomes and epigenomes. Neuron *109*, 11–26. https://doi.org/10.1016/j.neuron.2020.12.010.

20. Nath, A., and Bild, A.H. (2021). Leveraging single-cell approaches in cancer precision medicine. Trends Cancer *7*, 359–372. https://doi.org/10.1016/j.trecan.2021.01.007.

21. Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G.P. (2018). Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. Cell *174*, 968–981.e15. https://doi.org/10.1016/j.cell.2018.07.010.

22. Vistain, L.F., and Tay, S. (2021). Single-cell proteomics. Trends Biochem. Sci. *46*, 661–672. https://doi.org/10.1016/j.tibs.2021.01.013.

23. Luquette, L.J., Bohrson, C.L., Sherman, M.A., and Park, P.J. (2019). Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. Nat. Commun. *10*, 3908. https://doi.org/10.1038/s41467-019-11857-8.

24. Phillips, B.G., Wang, Y., Ambati, S., Ma, P., and Meagher, R.B. (2020). Airways therapy of obstructive sleep apnea dramatically improves aberrant levels of soluble cytokines involved in autoimmune disease. Clin. Immunol. *221*, 108601. https://doi.org/10.1016/j.clim.2020.108601.

25. Platzer, A. (2013). Visualization of SNPs with t-SNE. PLoS One *8*, e56883. https://doi.org/10.1371/journal.pone.0056883.

26. Xu, X., Xie, Z., Yang, Z., Li, D., and Xu, X. (2020). A t-SNE based classification approach to compositional microbiome data. Front. Genet. *11*, 620143. https://doi.org/10.3389/fgene.2020.620143.

27. Betters, D.M. (2015). Use of flow cytometry in clinical practice. J. Adv. Pract. Oncol. *6*, 435–440. https://doi.org/10.6004/jadpro.2015.6.5.4.

28. Pedreira, C.E., Costa, E.S.D., Lecrevise, Q., Grigore, G., Fluxa, R., Verde, J., Hernandez, J., van Dongen, J.J.M., and Orfao, A.; EuroFlow (2019). From big flow cytometry datasets to smart diagnostic strategies: the EuroFlow approach. J. Immunol. Methods *475*, 112631. https://doi.org/10.1016/j.jim.2019.07.003.

29. Neumann, J., Prezzemolo, T., Vanderbeke, L., Roca, C.P., Gerbaux, M., Janssens, S., Willemsen, M., Burton, O., Van Mol, P., Van Herck, Y., et al. (2020). Increased IL-10-producing regulatory T cells are characteristic of severe cases of COVID-19. Clin. Transl. Immunology *9*, e1204. https://doi.org/10.1002/cti2.1204.

30. van Rhenen, A., Moshaver, B., Kelder, A., Feller, N., Nieuwint, A.W.M., Zweegman, S., Ossenkoppele, G.J., and Schuurhuis, G.J. (2007). Aberrant marker expression patterns on the CD34+CD38- stem cell compartment in acute myeloid leukemia allows to distinguish the malignant from the normal stem cell compartment both at diagnosis and in remission. Leukemia *21*, 1700–1707. https://doi.org/10.1038/sj.leu.2404754.

31. Behbehani, G.K. (2017). Applications of mass cytometry in clinical medicine: the promise and perils of clinical CyTOF. Clin. Lab. Med. *37*, 945–964. https://doi.org/10.1016/j.cll.2017.07.010.

32. Zhang, J., Wang, W., Huang, J., Wang, X., and Zeng, Y. (2020). How far is single-cell sequencing from clinical application? Clin. Transl. Med. *10*, e117. https://doi.org/10.1002/ctm2.117.

33. Durbin, J. (1973). Distribution Theory for Tests Based on the Sample Distribution Function (Society for Industrial and Applied Mathematics).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
MUS dataset	This paper	https://flowrepository.org/id/FR-FCM-Z48W
MC dataset	Penttila et al.[10]	https://flowrepository.org/id/FR-FCM-Z34U
SCS dataset	Wauters et al.[11]	https://ega-archive.org/studies/EGAS00001004717
Software and algorithms		
Cross Entropy test	https://github.com/AdrianListon/Cross-Entropy-test	https://doi.org/10.5281/zenodo.7420921
Guide to running the test	https://www.liston.babraham.ac.uk/flowcytoscript/	

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to the lead contact, Adrian Liston (al989@cam.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Collaborating immunologists beta-tested the Cross Entropy test, testing for robustness. Among these datasets, four are used as examples here, covering high dimensional flow cytometry in mice (MUS), and flow cytometry (FC), mass cytometry (MC) and single cell sequencing (SCS) datasets in humans. The MUS dataset may be accessed via FlowRepository (https://flowrepository.org/id/FR-FCM-Z48W). For the mass cytometry (MC) dataset, we used data published by Pentilla et al.[10] and available in the Flow repository (https://flowrepository.org/id/FR-FCM-Z34U). For the SCS dataset, we used data published by Wauters et al.[11] and available in the EGA European Genome-Phenome Archive database (EGAS00001004717).
- The script used is available on GitHub at https://github.com/AdrianListon/Cross-Entropy-test. Parallel scripts are available for use on cytometry and single cell sequencing datasets.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mice

C57BL/6 mice were housed under SPF conditions. Both male and females (8–15 weeks old) mice were fed a standard chow diet, *ad libitum*. All experiments were performed in accordance with the Babraham Institute Animal Welfare and Ethics Review Body. Animal husbandry and experimentation complied with existing European Union and national legislation and local standards under license PP3981824.

METHOD DETAILS

For the MUS dataset, spleen, lymph nodes and small intestinal lamina propria from 10-week old female C57Bl/6 mice were stained for flow cytometry. Spleen and lymph nodes were disrupted with glass slides, filtered through 100 μm mesh, and, in the case of the spleen, red blood cells were lysed. Intestinal tissue was incubated in staining buffer (HBSS with 10mM HEPES, 2% FCS and 2mM EDTA) (Gibco) for 15 min at 37°C in a shaking platform. Dislodged cells were discarded and the remaining tissue was digested with 1 mg/ml Collagenase D (Roche), 100 μg/ml hyaluronidase and 40 μg/ml DNAse I (Sigma) in IMDM with 10mM HEPES and 10% FCS) for a further 30 min at 37°C with shaking. The tissue was then filtered through 100 μm mesh to obtain a single cell suspension. The intestinal single cell suspension was passed through a 40% Percoll (GE Healthcare) gradient by centrifugation at 600*g* for 10 min at 4°C. Cells ($2x10^6$) were blocked for 30min with 2.4G2 hybridoma supernatant supplemented with 1% Monocyte Blocker (BioLegend) prior to staining. Cells were stained for 60 min at 37°C with CD45-NovaBlue 530, CD4-NovaBlue 585, CD8-NovaBlue

610, CD19-NovaYellow 690, CD3-NovaRed 685, CD69-biotin (ThermoFisher), CCR2-BV750 and CCR9-BB515 (BD Biosciences) in staining buffer containing presence of CellBlox (NovaBlock). Cells were washed and stained with ViaKrome 808 (Beckman Coulter) and Qdot705-streptavidin (ThermoFisher) in HBSS at room temperature for 20min. After washing, cells were fixed and permeabilized with Foxp3 Transcription Factor Staining Buffer Set (eBioscience) according to the manufacturer's instructions. Cells were then washed twice with permeabilization buffer (eBioscience) and stained overnight (16hrs) at 4°C in permeabilization buffer with 2.4G2 anti-CD16/32 for the following antibodies: CD103-BUV395, IgD-BUV496, NK1.1-BUV563, CTLA-4-BUV615, c-Kit-BUV661, CD62L-BUV737, GITR-BUV805, CXCR3-BV480, Siglec F-BV510, TCRγδ-BB660-P2, PDCA-1-BB700, CD11c-BB755-P, Ly-6C-BB790-P (BD Biosciences), CD150-BV421, Helios-Pacific Blue, TCRβ-BV570, PD-1-BV605, XCR1-BV650, CD127-BV711, I-A/I-E-PerCP, CD64-PE-Dazzle 594, CD11b-PE-Fire 640, CD172a-PE-Cy7, CD38-PE-Fire 810, CD44-APC-Fire 750, Gr-1-APC-Fire 810 (BioLegend), ICOS-Super Bright 436, KLRG1-Super Bright 780, CD86-PE-Cy5, NKp46-PerCP-eFluor710, F4/80-AF561, Foxp3-PE-Cy5.5, RORγT-APC, GATA-3-PE, T-bet-eFluor660, Ki67-AF700 (Thermo Fisher), B220 (CD45R)-StarBright UV 445, CD25-StarBright Violet 515 (Bio-Rad), CD24-Pacific Orange, IgM-AF532 and CD90.2-AF790 (conjugated in house). Cells were washed twice with permeabilization buffer and once with staining buffer prior to acquisition. Samples were acquired on a 5-laser Aurora Spectral Analyser (Cytek), with unmixing based on single stained cell controls. An autofluorescence channel was created using unstained intestinal cells, gating on high SSC cells.

For the artificial MUS dataset, new FCS files were created using events extracted and concatenated from the MUS dataset. For lymph-spleen normalization, clusters mapped by FlowSOM were exported as independent FCS files, downscaled in FlowJo to match the new proportions, and concatenated. For tissue-spleen spike-in data, events were extracted and concatenated in R using flowCore and premessa.

For the mass cytometry dataset, we used data published by Pentilla et al.[10] and available in the Flow repository (https://flowrepository.org/id/FR-FCM-Z34U). We selected 12 individuals that were serially sampled at the time of admission to the ICU, 6–8 days after (intermediate time point) and upon discharge from the ICU. Briefly, whole blood was collected and stained with antibodies outlined in the original publication within 2–4 h of collection and acquired on a Helios Mass Cytometer (Fluidigm). For the analysis performed in this manuscript, living single cells were gated and beads removed. Monocytes were identified as being $CD45^{+}CD3^{-}CD19^{-}$ $CD14^{+}$ or $CD16^{+}$ and $CD56^{-}$.

For the SCS dataset, we used data published by Wauters et al.[11] and available in the EGA European Genome-Phenome Archive database (EGAS00001004717). Briefly, bronchoalveolar lavage was collected from 22 COVID-19 patients and 13 non-COVID-19 pneumonia patients during standard-of-care treatment. The cellular component of the lavage was used for single cell sequencing on the 10x platform. We used the data from the bronchoalveolar lavage of both COVID and non-COVID pneumonia patients, with the published cell identity. For the presented analysis, t-SNE representations were calculated based on the count matrices using the Seurat pipeline and its coordinates used to calculate the cross entropy distributions and comparison.

t-SNE representation of multidimensional data

The t-SNE algorithm aims to provide a simplified, low-dimensional representation of high-dimensional data, while preserving as much as possible the local distances between data points, that is, the differences between similar data points.

Let us represent a dataset in original space as a collection of n data points $\{x_i\}$, with $i = 1 \ldots n$. All data points x_i are vectors of $\mathbf{R}^d$, with d, the number of dimensions, being usually a large number.

The low-dimensional representation that t-SNE produces is another collection of data points $\{y_i\}$, again with $i = 1 \ldots n$. In this case, data points are vectors of a real space with low number of dimensions, typically $\mathbf{R}^2$ or $\mathbf{R}^3$.

What defines the t-SNE algorithm, as well as any other technique of dimensionality reduction, is the way in which the new collection of points $\{y_i\}$ is obtained from the original collection $\{x_i\}$. In the case of t-SNE, Gaussian probabilities are defined for each pair of points, as follows. Given a pair in the original space (x_i, x_j), with $i \neq j$, the probability p_{ij} is defined by assigning first to each tuple (i,j):

$$p_{j|i} = \frac{exp\left(-\|x_j - x_i\|2\sigma_i^2\right)}{\sum\limits_{k \neq i} exp\left(-\|x_k - x_i\|2\sigma_i^2\right)} \qquad \text{(Equation 1)}$$

and then symmetrized by

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n} \qquad \text{(Equation 2)}$$

which implies the global normalization property

$$\sum_{i=1}^{n} \sum_{j=1\, j \neq 1}^{n} p_{ij} = 1 \qquad \text{(Equation 3)}$$

The parameters σ_i are adjusted for each point x_i, so that the perplexity ρ (or equivalently the entropy log ρ) of the distribution $\{p_{j|i}\}$ has a predefined value.

Similarly, probabilities in the representation space are defined for each pair of points (y_i, y_j), with $i \neq j$, but following in this case a Cauchy distribution:

$$q_{ji} = \frac{\left(1 + \|y_j - y_i\|^2\right)^{-1}}{\sum_{k \neq i}\left(1 + \|y_l - y_k\|^2\right)^{-1}} \qquad \text{(Equation 4)}$$

which implies global normalization without the need of symmetrisation

$$\sum_{i=1}^{n} \sum_{j=1\, j \neq 1}^{n} q_{ij} = 1 \qquad \text{(Equation 5)}$$

With the definition of these probability distributions, which give more weight to pairs of points close to each other, the problem of finding the transformed points $\{y_i\}$ from the original points $\{x_i\}$ is converted into the problem of making the probabilities $\{q_{ij}\}$ as similar as possible to $\{p_{ij}\}$. This is achieved by minimizing numerically the Kullback-Leibler divergence between the distributions $P = P = \{p_{ij}\}$ and $= \{q_{ij}\}$,

$$D(P,Q) = \sum_{i=1}^{n} \sum_{j=1\, j \neq 1}^{n} p_{ij} \, log \frac{p_{ij}}{q_{ij}} \qquad \text{(Equation 6)}$$

This is equivalent to making the cross entropy between P and Q

$$H(P,Q) = -\sum_{i=1}^{n} \sum_{j=1\, j \neq 1}^{n} p_{ij} \, log \, q_{ij} \qquad \text{(Equation 7)}$$

as close as possible to the entropy of P

$$H(P) = -\sum_{i=1}^{n} \sum_{j=1\, j \neq 1}^{n} p_{ij} \, log \, p_{ij} \qquad \text{(Equation 8)}$$

given that

$$D(P,Q) = H(P,Q) - H(P) \qquad \text{(Equation 9)}$$

Because the entropy of the distribution P is fixed by construction to a value determined by the perplexity ρ, the optimization carried out by t-SNE only operates over the cross entropy between P and Q.

Distributions of entropy and cross entropy

The key insight for the test on t-SNE representations proposed here resides in considering the distributions of entropy and cross entropy per point, instead of their global counterparts. For a data point with index i, the entropy and cross entropy, $\{H_i^{(P)}\}$ and $\{H_i^{(P,Q)}\}$, respectively, are defined as:

$$H_i^{(P)} = -\sum_{j=1\, j \neq 1}^{n} p_{ij}^{*} \, log \; p_{ij}^{*} \qquad \text{(Equation 10)}$$

$$H_i^{(P,Q)} = -\sum_{j=1\, j \neq 1}^{n} p_{ij}^{*} \, log \; q_{ij}^{*} \qquad \text{(Equation 11)}$$

with

$$p_{ij}^{*} = np_{ij} \qquad \text{(Equation 12)}$$

$$q_{ij}^{*} = nq_{ij} \qquad \text{(Equation 13)}$$

The local pair probabilities p_{ij}^{*} and q_{ij}^{*} are introduced to have quantities with a sum per point closer to 1:

$$\sum_{j=1\, j \neq i}^{n} p_{ij}^{*} \approx 1 \qquad \text{(Equation 14)}$$

$$\sum_{j=1\, j \neq i}^{n} q_{ij}^{*} \approx 1 \qquad \text{(Equation 15)}$$

The distribution of point Kullback-Leibler divergences can also be obtained as

$$D_i^{(P,Q)} = H_i^{(P,Q)} - H_i^{(P)} \qquad \text{(Equation 16)}$$

This way, the global Kullback-Leibler divergence can be written as the average of the distribution of divergences per point

$$D(P,Q) = \frac{1}{n}\sum_{i=1}^{n} D_i^{(P,Q)} = \frac{1}{n}\sum_{i=1}^{n} H_i^{(P,Q)} - \frac{1}{n}\sum_{i=1}^{n} H_i^{(P)} \qquad \text{(Equation 17)}$$

By design, t-SNE makes all the point entropies $\{H_i^{(P)}\}$ close to a fixed value determined by the perplexity ρ, with $H_i^{(P)} = loglog\ \rho$ for all *i* before symmetrization. Therefore, their average is fixed *a priori* and almost the same for any input dataset $\{x_i\}$. As a consequence, differences between results obtained from different input datasets will necessarily appear as differences between the respective distributions of cross entropy per point $\{H_i^{(P,Q)}\}$.

Test on distributions of cross entropy

Given two input datasets $\{x_i\}$ and $\{x_i'\}$, the t-SNE algorithm will produce two representations $\{y_i\}$ and $\{y_i'\}$, with associated distributions of cross entropy $\{h_i\}$ and $\{h_i'\}$.

Under the null hypothesis, $\{x_i\}$ and $\{x_i'\}$ will follow the same probability distribution, which implies that local distances between pair of points will also be equal in distribution. As a consequence, after applying the t-SNE algorithm, obtained datasets in representation space $\{y_i\}$ and $\{y_i'\}$, local distances, and point cross entropies will also be equal in distribution.

Under the alternative hypothesis, $\{x_i\}$ and $\{x_i'\}$ will differ in probability distribution.

This will likely imply differences in the distributions of local distances, which will in turn produce differences in distribution for the local distances in representation space. As a result, point cross entropies will likely differ in distribution.

Therefore, the test consists in evaluating the difference between the distributions of cross entropy, which allows us to use already available tests for comparing one-dimensional empirical distributions. The main test chosen for this purpose is the Kolmogorov-Smirnov test, which offers the additional advantage of providing a statistic that is a valid distance between functions, namely the L^∞ distance. This distance can be used as a measure of the difference between t-SNE plots, for example to organize them into a dendrogram.

L^∞ was calculated as:

$$L^{\infty(f,g)} = \max(|f(x) - g(x)|)$$

for any x in the domin of f and g

The Kolmogorov-Smirnov test was used as in Durbin.[33]

Additional dimensionality reduction analysis

Principal Component Analysis (PCA) was performed in R using the function prcomp and the package factoextra. Multidimensional Scaling (MDS) was implemented using the landmarked MDS R package lmds in order to reduce the computational time required. Independent Component Analysis (ICA) was run using the package ica in R. Isomap was implemented using the RDRToolbox library. PacMAP and triMAP were plotted in Python 3.9 via the R package reticulate.

General implementation details

The Cross Entropy test was implemented in R v.4.0.5, using the packages ConsensusClusterPlus, digest, dunn.test, flowCore, FlowSOM, ggplot2, ggridges, RANN, RColorBrewer, reshape2, Rtsne and umap.

QUANTIFICATION AND STATISTICAL ANALYSIS

This paper describes the development of a statistical test, which is described in the above 'method details' section.

ADDITIONAL RESOURCES

The script used is available on GitHub at https://github.com/AdrianListon/Cross-Entropy-test. Parallel scripts are available for use on cytometry and single cell sequencing datasets, and are available as part of the same GitHub repository. As a practical guide for biologists with low experience at working in R, we have provided a walk-through on the use of this code at: https://www.liston.babraham.ac.uk/flowcytoscript/