# A single-cell clusters similarity measure for different batches, datasets, and samples

*Óscar González-Velasco[1*], Elena Sanchez-Luis[1], Enrique De La Rosa[1], José Manuel Sánchez-Santos[1,2], Javier De Las Rivas[1]*

[1] *Bioinformatics and Functional Genomics Group, Cancer Research Center (CIC-IBMCC, CSIC/USAL), Consejo Superior de Investigaciones Cientificas (CSIC), University of Salamanca (USAL), Salamanca, Spain.*

[2] *Department of Statistics, University of Salamanca (USAL), Salamanca, Spain.*

## Abstract

**Summary:** Since the inception of single-cell level measuring techniques, identification of distinct cell stages, phenotypes and populations has been a challenge. Cell clustering and dimensionality reduction methods are the most popular approaches to identify heterogeneity of single-cell data. But, as public repositories continue to grow in number, integrative analyses and merging of large pools of samples from different and heterogeneous datasets becomes a difficult challenge, which showcases the impossibility of scalability of some of the existing methods. Here we present *ClusterFoldSimilarity*, an R package that calculates a measure of similarity between clusters from different datasets/batches, without the need of correcting for batch effect or normalizing and merging the data, thus avoiding artifacts and the loss of information derived from these kinds of techniques. The similarity metric is based on the average vector module and sign of the product of logarithmic fold-changes. *ClusterFoldSimilarity* compares every single pair of clusters from any number of different samples/datasets, including different number of clusters for each sample. Additionally, the algorithm is able to select the top genes which contribute the most to the similarity of two specific clusters, serving also as a feature selection tool.

**Availability and implementation:** The algorithm is freely available as an R package at: https://github.com/OscarGVelasco/ClusterFoldSimilarity

**Contact:** oscargvelasco@gmail.com

## Introduction

Single-cell technologies (such as scRNA-seq, scATAC-seq or scCyToF) constitute a new collection of advanced methodologies capable of capturing the biomolecular state of the cell and measuring multiple cellular markers for hundreds to millions of cells. Using these

technologies, researchers are able to elucidate genomic, epigenomic, and transcriptomic heterogeneity in cellular populations and subpopulations, and the variability at these levels (Kashima et al., 2020). The number of large-scale single-cell omic datasets generated using different technologies is continually growing with the advent of accessible datasets in public repositories. Despite this, integration of single-cell data remains difficult due to inherent dataset-specific systematic variability that present a challenge for batch-effect removal. A number of tools for batch-effect removal exist, but each has its advantages and limitations, with no clearly superior method (Tran et al., 2020). Here we present a new application tool that is able to analyze the similarity between clusters from different and heterogeneous datasets, without the need of batch-effect removal or merging. The approach is based on the fold change of the measured signal between clusters, standing on the idea that the overall signal change across different cell populations, subpopulations or phenotypes in a dataset should be conserved in different datasets measuring similar populations.

# Materials and methods

## *Pairwise fold changes*

In single-cell data, signal fold-changes (FC) can be defined as a Euclidean vector with origin at 0 and direction equal to the signal change, the scalar product of two vectors in the same space gives as a result a perpendicular vector to the plane. In the case of the dot product of two fold-changes the direction of the perpendicular vector will act as a concordance value of the direction of the FCs (see Figure 1A,B,C). The FCs are computed using the normalized count matrix. They are calculated pairwise between each cluster $i$ and the other clusters belonging to the same sample. FC for each gene is calculated as the log2 of the ratio of the mean expression from cells in cluster $i$ and the mean expression from cells in cluster $j$: $\forall j \neq i$. In such manner, we obtain $k$-1 FCs ($k$: number of clusters in a given sample) for each of the clusters in that sample.

## *Similarity measurement*

Once all the possible combinations of FCs between all clusters from each sample/batch are computed, we start by comparing the cluster 1 from sample/batch 1 with the cluster 1 in sample/batch 2 and continue iteratively. We achieve this by computing the scalar product of every possible combination of FCs from the two clusters: $combinations = (k_1 - 1) \times (k_2 - 1)$ being $k_1$: number of clusters in sample 1; $k_2$: number of clusters in sample 2. Then we calculate for each gene the mean of the scalar products. This value will serve as a measure of gene importance, being the gene with the highest positive value the one that contributes the most to the similarity between the two clusters (negative values indicate opposite FC values).

Finally, the similarity value for a pair of clusters is calculated as the square root of the sum of the mean of genes (calculated in the previous step):

$$Similarity = \sqrt{\left|\sum \overline{FC}\right|} * SIGN\left(\sum \overline{FC}\right) * \omega$$

The weight $\omega$ is based on the concordance of positive versus negative FC's and is calculated as:

$$\omega = \sqrt{\frac{|NEG_{FC} - POS_{FC}|}{NEG_{FC} + POS_{FC}}}$$

The *maximum similarity positive value* (calculated as described) will correspond with the most similar cluster: it is expected that, for a specific population of cells, overall changes in gene expression across different clusters (i.e., in different single cell datasets with similar cell populations) are conserved on different samples, and thus the sign of the product of FCs should be positive and its module higher than less concordant expression changes.

## *Compatibility*

*ClusterFoldSimilarity* is compatible with the most popular single-cell pipelines for an easy downstream integration: Seurat (Hao et al., 2021) and SingleCellExperiment (Amezquita et al., 2020).

# Results

*ClusterFoldSimilarity* has been tested in several single-cell datasets in which cell-type information was available. The first is comprised of four human pancreas datasets from the scRNAseq Bioconductor package (Risso D, 2021): GSE84133 (Baron et al., 2016); GSE86469 (Lawlor et al., 2017); GSE81608, (Xin et al., 2016); ArrayExpress: E-MTAB-5061 (Segerstolpe et al., 2016). Additionally, it has been tested in four peripheral blood mononuclear cell (PBMC) single-cell RNA-Seq datasets: GSE149938 (Xie et al., 2021); 10x Genomics 10k Peripheral blood mononuclear cells (PBMCs) from a healthy donor, Single Indexed; GSE120221 (Oetjen et al., 2018); PRJEB40448 (Lawlor et al., 2021).

Our results show that ClusterFoldSimilarity metric is able to pair clusters by cell type and cell subpopulation with high accuracy: all clusters corresponding with shared cell types across datasets are paired together, including subpopulations in the case of PBMCs. Moreover, the top gene markers selected by similarity correspond to markers reported on scientific literature (see Supp. Material).

# References


Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M. L., Huber, W., Morgan, M., Gottardo, R., & Hicks, S. C. (2020). Orchestrating single-cell analysis with Bioconductor. *Nature Methods*, *17*(2), 137–145. https://doi.org/10.1038/s41592-019-0654-x

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., & Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, *3*(4). https://doi.org/10.1016/j.cels.2016.08.011

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573-3587.e29. https://doi.org/https://doi.org/10.1016/j.cell.2021.04.048

Kashima, Y., Sakamoto, Y., Kaneko, K., Seki, M., Suzuki, Y., & Suzuki, A. (2020). Single-cell sequencing techniques from individual to multiomics analyses. *Experimental & Molecular Medicine*, *52*(9), 1419–1427. https://doi.org/10.1038/s12276-020-00499-2

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., Kycia, I., Robson, P., & Stitzel, M. L. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Research*, *27*(2), 208–222. https://doi.org/10.1101/gr.212720.116

Lawlor, N., Nehar-Belaid, D., Grassmann, J. D. S., Stoeckius, M., Smibert, P., Stitzel, M. L., Pascual, V., Banchereau, J., Williams, A., & Ucar, D. (2021). Single Cell Analysis of Blood Mononuclear Cells Stimulated Through Either LPS or Anti-CD3 and Anti-CD28. *Frontiers in Immunology*, *12*, 691. https://doi.org/10.3389/fimmu.2021.636720

Oetjen, K. A., Lindblad, K. E., Goswami, M., Gui, G., Dagur, P. K., Lai, C., Dillon, L. W., McCoy, J. P., & Hourigan, C. S. (2018). Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. *JCI Insight*, *3*(23). https://doi.org/10.1172/jci.insight.124928

Risso D, C. M. (2021). *scRNAseq: Collection of Public Single-Cell RNA-Seq Datasets. R package version 2.8.0.*

Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E. M., Andréasson, A. C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M. K., Smith, D. M., Kasper, M., Ämmälä, C., & Sandberg, R.

(2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism*, *24*(4). https://doi.org/10.1016/j.cmet.2016.08.020

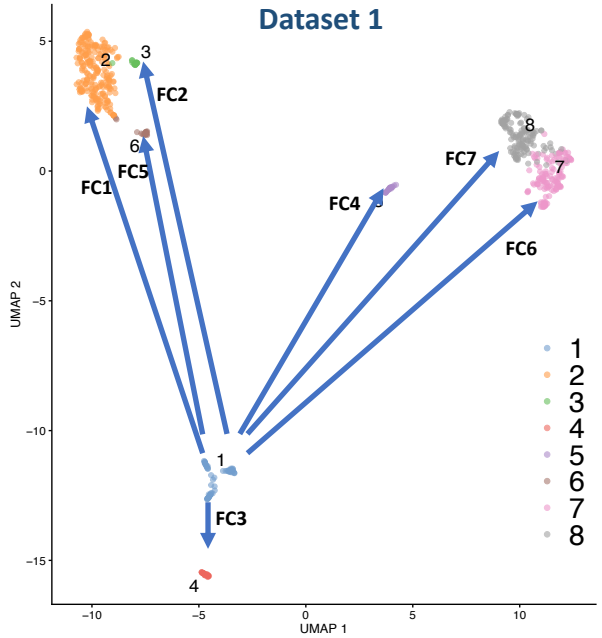
Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, *21*(1), 12. https://doi.org/10.1186/s13059-019-1850-9

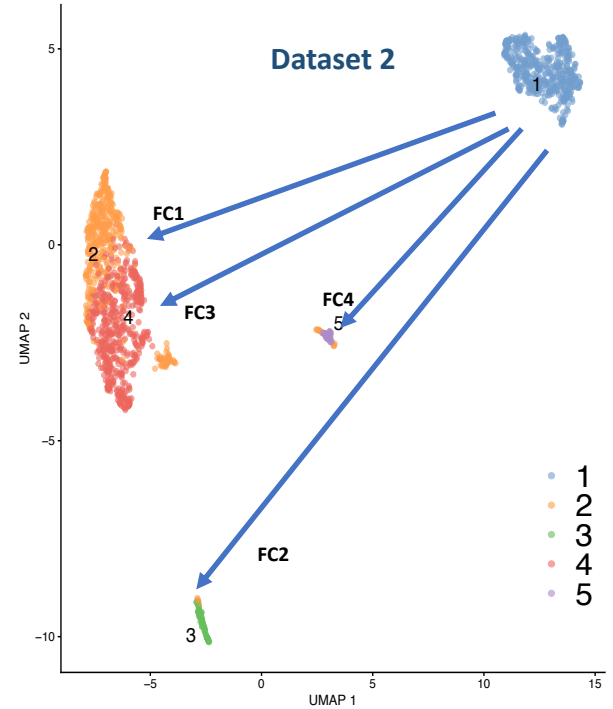
Xie, X., Liu, M., Zhang, Y., Wang, B., Zhu, C., Wang, C., Li, Q., Huo, Y., Guo, J., Xu, C., Hu, L., Pang, A., Ma, S., Wang, L., Cao, W., Chen, S., Li, Q., Zhang, S., Zhao, X., … Cheng, T. (2021). Single-cell transcriptomic landscape of human blood cells. *National Science Review*, *8*(3). https://doi.org/10.1093/nsr/nwaa180

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., Murphy, A. J., Yancopoulos, G. D., Lin, C., & Gromada, J. (2016). RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metabolism*, *24*(4). https://doi.org/10.1016/j.cmet.2016.08.018

*Figure 1: Pipeline representation of the similarity measurement done by ClusterFoldSimilarity. A: Foldchanges (FC) are calculated between all possible combination of clusters for each dataset, figure shows the FC calculation for cluster 1 in both datasets 1 and 2 in a UMAP corresponding with pancreas cells. B: The scalar product of the computed FCs between all pairs is calculated. C: Representation of the vectorial space of FCs and their product (in the graphic, a and b represent FCs of a same gene between clusters), FC that points in the same direction will have positive product sign, meanwhile opposite FC will have negative product sign. D: Table with the similarity values and output information given by ClusterFoldSimilarity. E: UMAP representation of two pancreatic datasets with annotated cell-types. Note that exact cell-type clusters are matched together by the table given at D. F: Graph plot made by ClusterFoldSimilarity using the similarity values calculated (given at table D), nodes correspond with clusters from a specific dataset, edges correspond with the similarity value, arrows point in the direction of the similarity relationship.*

A

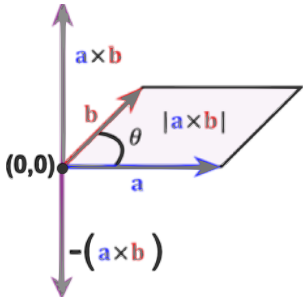



B

| Dataset1 | \| | Dataset2 |
|---|---|---|
| FC1 | * | FC1 |
| FC1 | * | FC2 |
| FC1 | * | FC3 |
| FC1 | * | FC4 |
| FC2 | * | FC1 |
| FC2 | * | FC2 |
| FC2 | * | FC3 |
| ... | | ... |
| FC7 | * | FC1 |
| FC7 | * | FC2 |
| FC7 | * | FC3 |
| FC7 | * | FC4 |

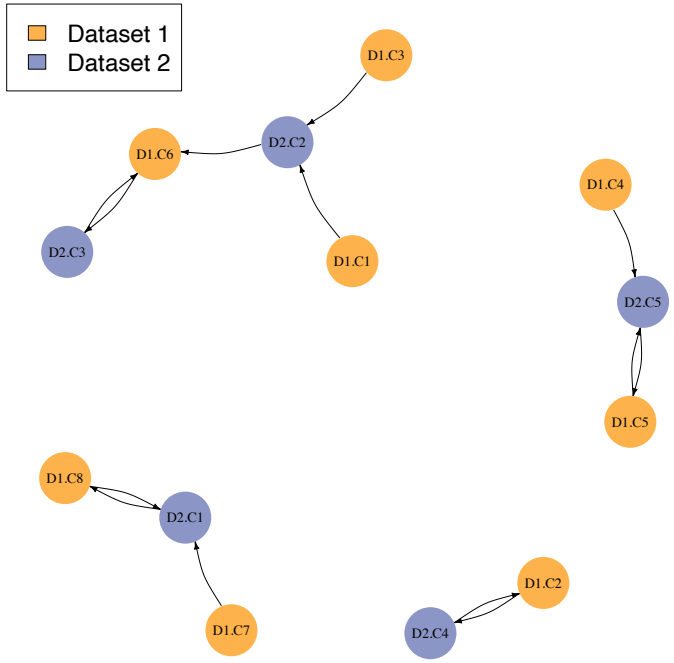
C

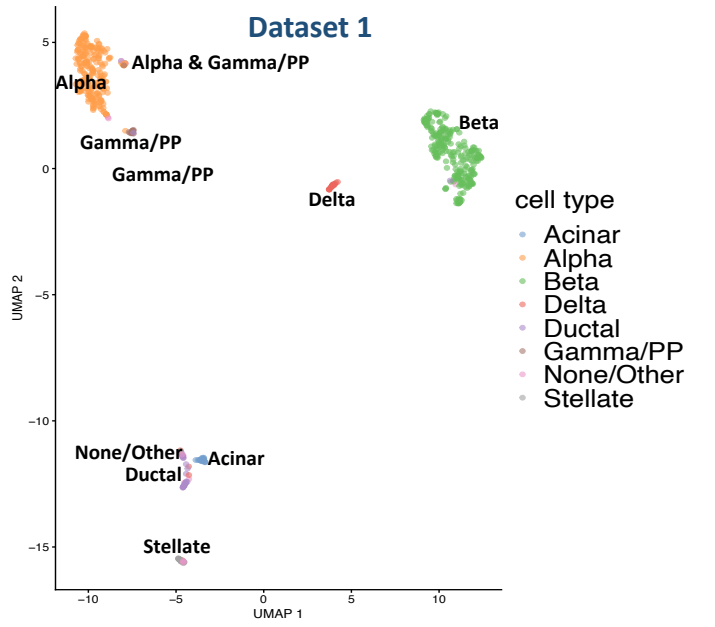
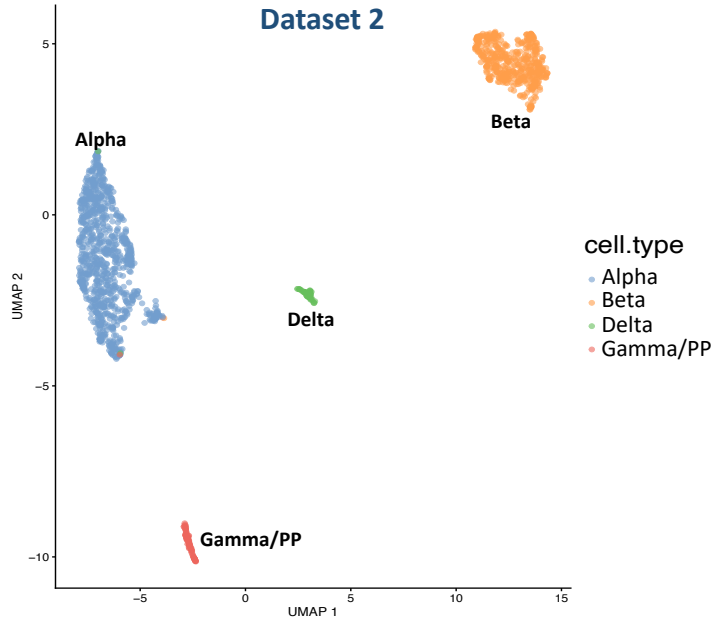



D

| Similarity value | sem | Dataset_l | cluster_l | Dataset_r | cluster_r | top_gene_conserved |
|---|---|---|---|---|---|---|
| 2.3804 | 0.0089 | 1 | 1 | 2 | 2 | ENSG00000175426 |
| 8.6026 | 0.015 | 1 | 2 | 2 | 4 | ENSG00000138131 |
| 2.4667 | 0.0096 | 1 | 3 | 2 | 2 | ENSG00000145321 |
| 1.4384 | 0.0125 | 1 | 4 | 2 | 5 | ENSG00000152254 |
| 3.2800 | 0.014 | 1 | 5 | 2 | 5 | ENSG00000157005 |
| 6.3570 | 0.0144 | 1 | 6 | 2 | 3 | ENSG00000108849 |
| 7.1224 | 0.0173 | 1 | 7 | 2 | 1 | ENSG00000185559 |
| 8.5061 | 0.0178 | 1 | 8 | 2 | 1 | ENSG00000141433 |
| 8.5061 | 0.0178 | 2 | 1 | 1 | 8 | ENSG00000141433 |
| 3.1865 | 0.0084 | 2 | 2 | 1 | 6 | ENSG00000145321 |
| 6.3570 | 0.0144 | 2 | 3 | 1 | 6 | ENSG00000108849 |
| 8.6026 | 0.015 | 2 | 4 | 1 | 2 | ENSG00000138131 |
| 3.2800 | 0.014 | 2 | 5 | 1 | 5 | ENSG00000157005 |

F




E