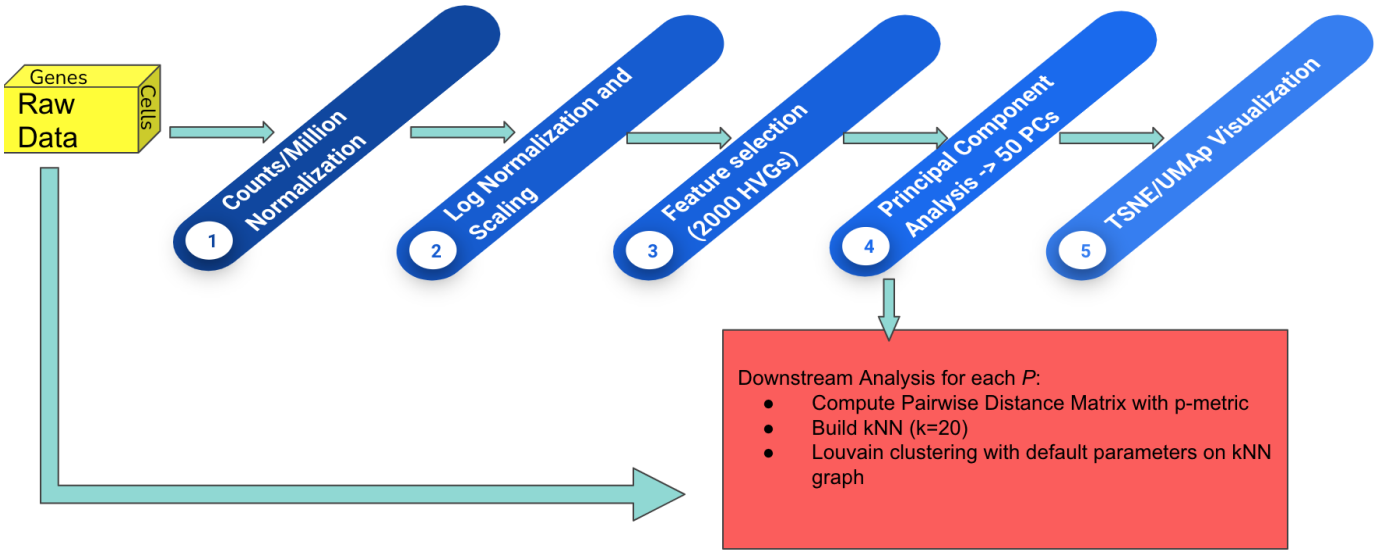
The Effectiveness of Minkowski p-norms in recovering Biological Variance

Timothy Hamilton
Gunalan Natesan
March 2022

Single-Cell RNA sequencing projects have increased drastically in scope and scale in recent years, allowing for numerical measurements of cellular gene expression at the transcript level. To this end, gene expression profiles can thus be built on a cell by cell level for any number and type of cellular tissues. Indeed, many scRNA-seq studies are characterized by counting RNA transcripts for individual cells, and using unsupervised machine learning models to find underlying similarities in the resulting expression profiles.

A current trend is the use of unsupervised clustering to group cells together, particularly in the context of establishing and validating cell types using gene expression profiles. Many possible pipelines have developed to process a vector of RNA counts for a specific cell to compute distances between various cells, before processing and clustering them in high dimensional space. In this report, we benchmark an analysis pipeline introduced in [1], shown in Figure 1 below.

Figure 1: Illustration of clustering pipeline



Pipeline for scRNA-seq tissue composition analysis, from [1], with preprocessing pipeline shown with upper right arrow sequence and raw references shown in lower left arrow sequences. Both are piped into the clustering program shown in the lower right corner.

The pipeline shown in the upper right corner is an established pipeline in that Raw RNA counts are Counts/Million [2] and Log Normalized [3] before Feature selection and PCA

Dimensionality reduction [4] are applied and the data is clustered. As a control, the raw counts of the RNA in each cell were also directly fed into the clustering process as well. The clustering process itself computes a pairwise distance matrix with an established p - metric (traditionally p=2, the euclidean norm) before building a k Nearest Neighbor model [5] and applying louvain clustering [2] on the dataset. In this experiment, however, distance matrices in both preprocessed and raw data will be computed with various p-norms, ranging from p = 1 to 6 with variable increments. The data selected to cluster are pancreatic islet sc-RNA-seq data from [6], with generated annotations for each label.

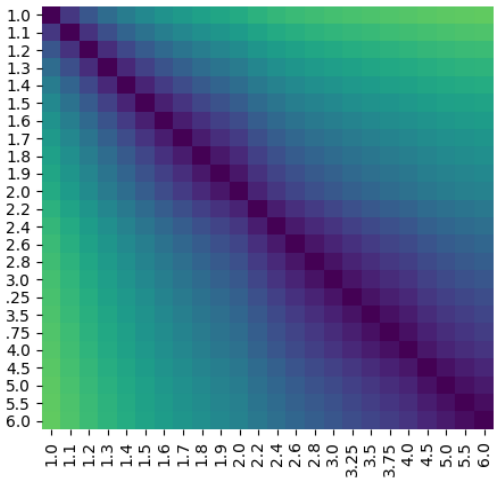
To measure the consistency between the various generated clusters, we introduce several metrics, the first of which is the jaccard distance [7]. The Jaccard Distance measures a relative distance between two sets according to the set of their shared and excluded elements. The results of this clustering is shown in Figure 2

$$J = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

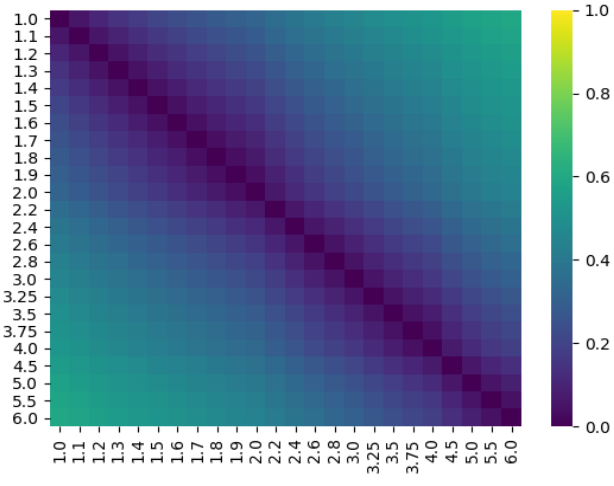
The formula of Jaccard Distance between sets A and B.

Figure 2:

A AJD of P-norm Heatmap of Raw Data



ADJ of P-norm Heatmap of Preprocessed Data **B**

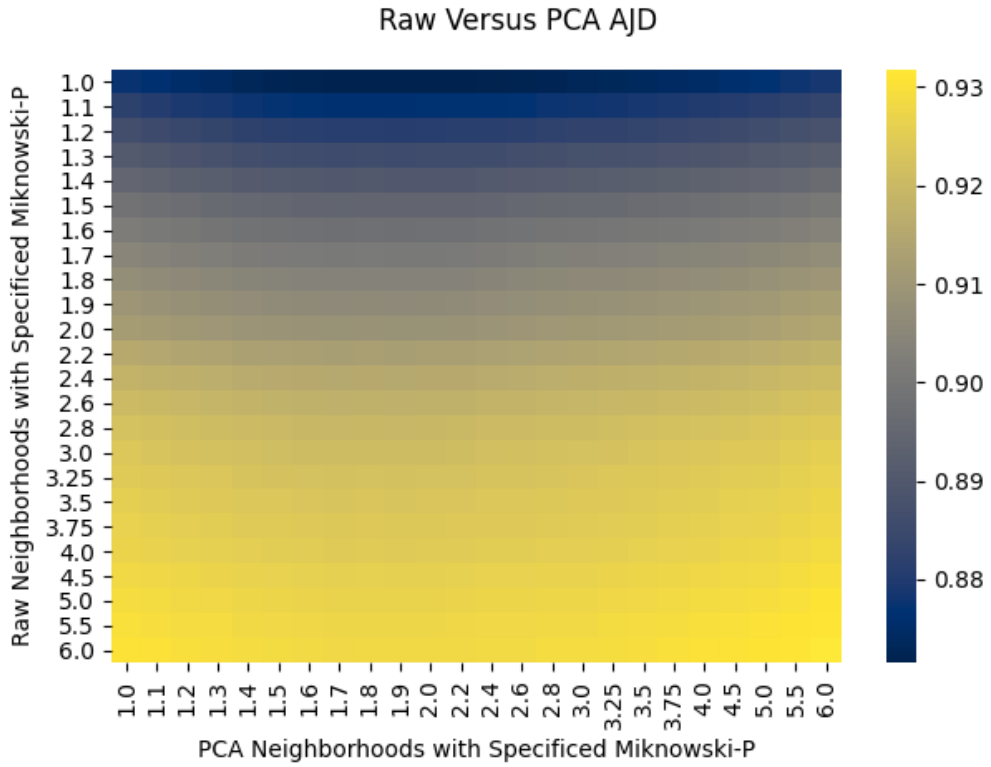


The Average Jaccard Distances shown between the sets of clusters formed for each p-norm. Figure 2A illustrates the AJD values for the p-norm produced raw data, while Figure 2B illustrates the AJD values for the p-norm produced preprocessed data

Figure 2A and 2B both show high levels of self similarity, which tapers to higher levels of dissimilarity as the p-norms for each set differ. As the set of clusters from p-norms are highly close to other similarly valued p-norm clusters (AJD ~= 1), and taper to almost complete

incongruence (AJD ~= 0) in both sets, we note that a high level of discrepancy exists in the set of possible clusterings, based on the p-norm identity alone. Furthermore, we note that levels of dissimilarity are significantly higher in the unprocessed data, as normalization and dimensionality reduction are shown to homogenize variances.

Figure 3



The Average Jaccard Distances shown between the sets of clusters formed for each p-norm for the raw data against the preprocessed data .

This is corroborated in Figure 3, as there is relatively little variance in the AJD of raw datasets against PCA neighborhood, especially in the direction of the preprocessed data, further suggesting that variance is homogenized when PCA/normalization is applied. To validate these clusterings against reference annotations for each cell, we introduce the Adjusted Rand Index [8] below.

Adjusted Rand Index (ARI)

$X \setminus Y$	Y_1	Y_2	$\cdots$	Y_s	sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
sums	b_1	b_2	$\cdots$	b_s	

$$n_{ij} = |X_i \cap Y_j|$$

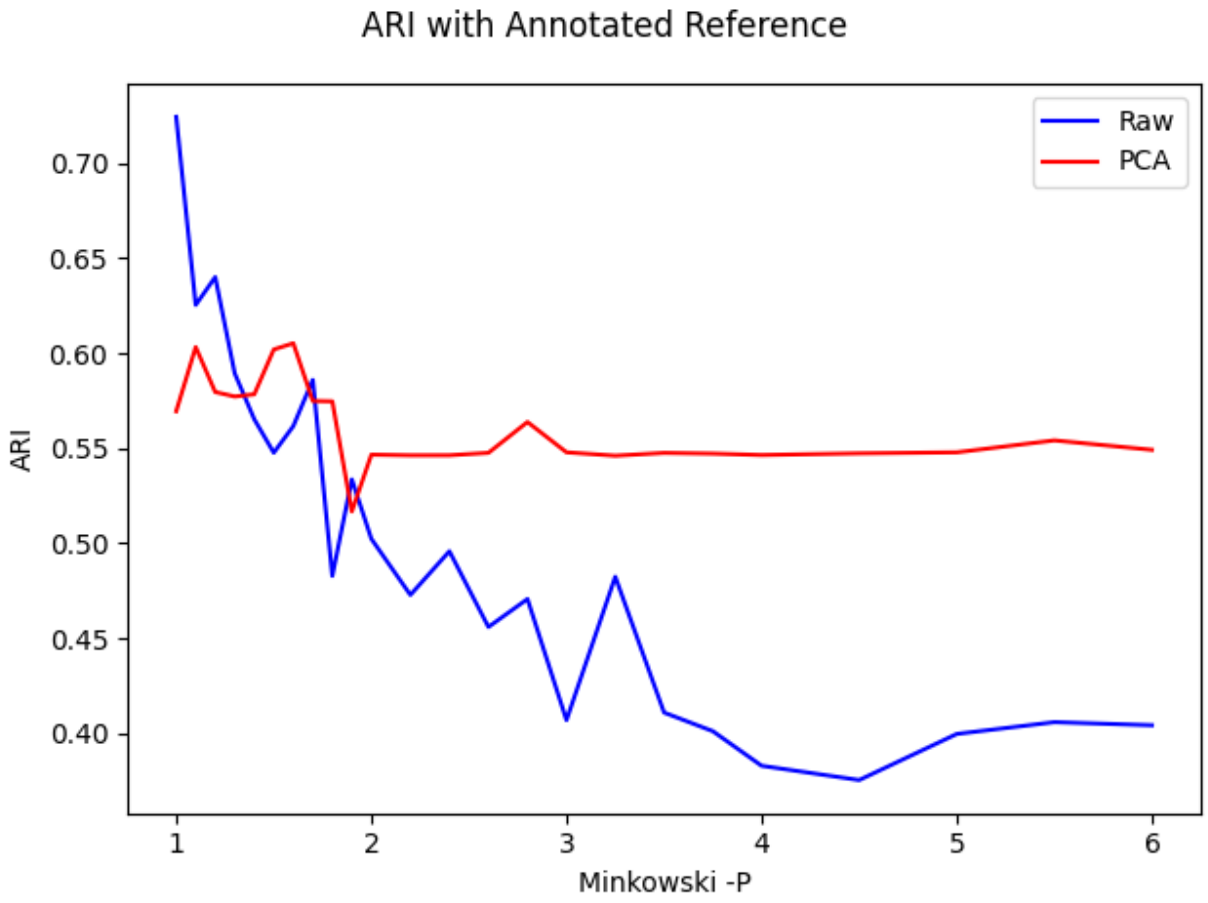
$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

X_i - Set of all cells in cluster *i* in cluster schema X

Y_i - Set of all cells in cluster *i* in cluster schema Y

Determines the difference in clustering from random assignment of cluster labels to cells

Figure 4



Above, we show the ARI of the Raw and Preprocessed RNA datasets with the annotated reference set, each of which is varied by its p-norm.

As shown in Figure 4, the PCA preprocessed data above consistently has an ARI of ~0.55 with the annotated reference set for p-norm values above 2, while p-norm values below result in an ARi that fluctuates ~0.5 above and below 0.55. The large amount of homogeneity once again insinuates that a large amount of smoothing has taken place in the preprocessing stage, especially as the ARI for p-norm values in the raw dataset range from ~0.75 to 0.4. Furthermore, the trend shows that low p-norm values tend to align with the reference dataset by a degree of almost 0.2 more than the higher p-norm values for raw counts, and 0.05 for p-norm values for preprocessed counts. This insinuates that the Manhattan norm is the most useful for comparing distinct RNA counts between cells, rather than the euclidean or maximal metric. This may, in part, be due to the fact that cellular vectors are not rotation invariant by any means. For example, a cell with X amount more of a certain gene and another cell with Y amount more of another gene do not have a distance of $(X^p+Y^p)^{1/p}$, as the notion of non-integer gene counts do not have any biological counterparts. Mechanistically, there are a total number of X + Y RNA molecules that differ in each cell, which is best encapsulated by the Manhattan distance.

tSNE of Pancreatic Cell scRNA-seq dataset colored by Clustering Dataset:

Figure 5

A Annotated Reference

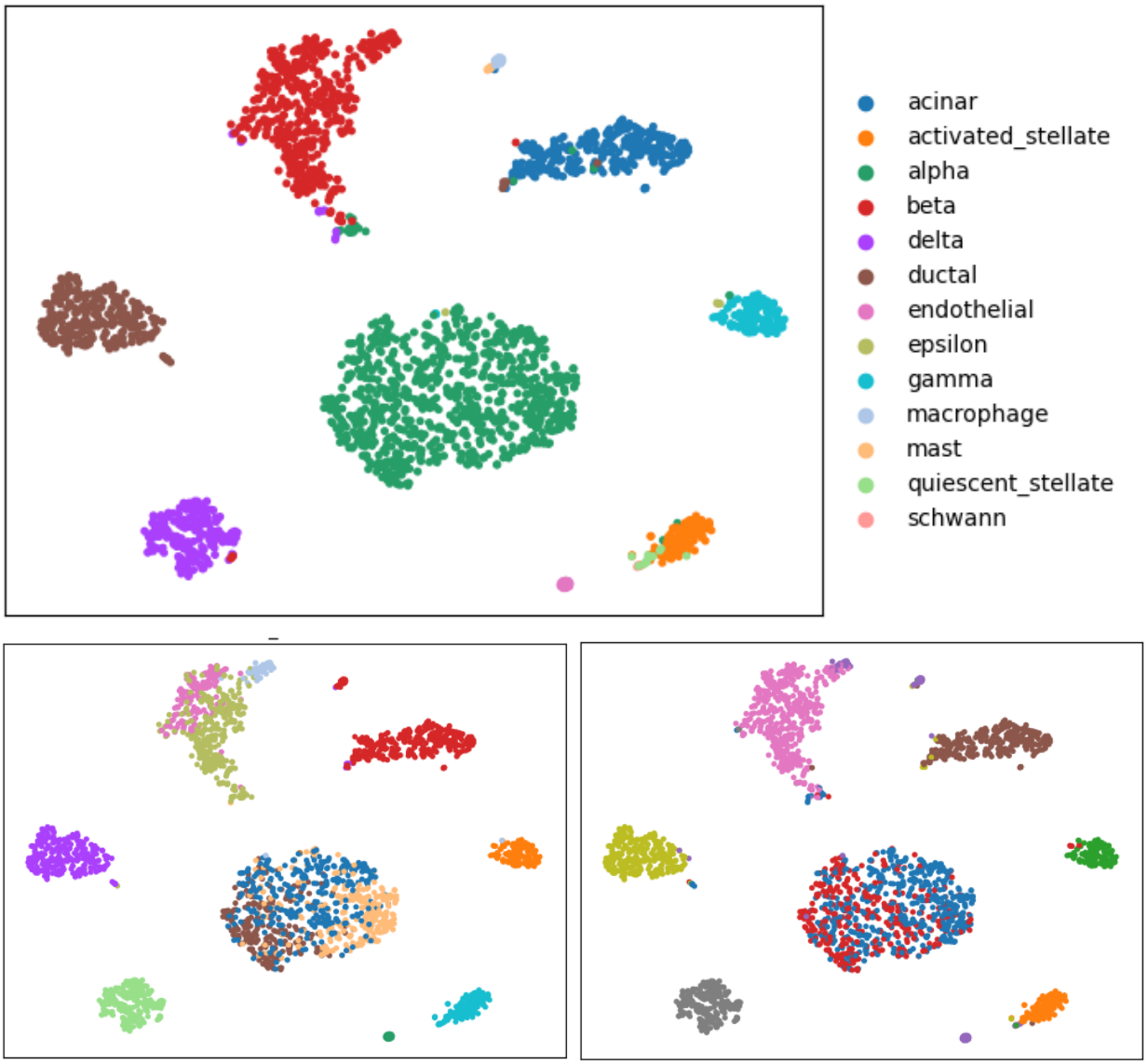


Figure 5A shows the tSNE Pancreatic Cell scRNA-seq dataset colored by the annotated reference assignments into 14 labeled cell types, shown on its left. Figure 5B shows tSNE Pancreatic Cell scRNA-seq dataset colored by the clusters produced by preprocessing the data before clustering with a p-norm of 2. Figure 5B shows tSNE Pancreatic Cell scRNA-seq dataset colored by the clusters produced by processing the raw data with a p-norm of 1.

Figure 5A shows tSNE [9] Pancreatic Cell scRNA-seq dataset, with its original reference annotations. It is worth noting that the annotations were reconstructed from clustering the data in [2], not necessarily through directly tracking the cells. Nonetheless, it shows mostly distinct clusters between the cells though there is some admixing of cell types between clusters. For example, the cluster in 5A shows alpha cells that are proximally near a cluster consisting of

mostly beta cells. This may be due to the inherent heterogeneity present in pancreatic cells [10], or due to the phenomenon in which pancreatic alpha cells turn into beta cells [11]. Nonetheless, despite the fact that tSNE groups the cells similarly, the cells are identifiably different in the annotated dataset, but not the preprocessed dataset with a p-norm of 2. As a matter of fact, this clustering does not pick up on any of the rare cell types that are oriented positioned near other cells in tSNEm while clusters produced by processing the raw data with a p-norm of 1 do. The raw data clustering finds many of the rare cell types in the same position of the original annotated set in tSNE space, though with varying accuracies regarding cell assignment.

Given the more consistent ARI and alignment of the raw reads with a p-norm of 1 do indicate the effects of variable norms and preprocessing regarding alignment with real assignments. More rigorous testing on more possible datasets are required before large conclusions can be made.

Resources:

[1]

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). Scanpy: Large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1). https://doi.org/10.1186/s13059-017-1382-0

[2]

Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-seq based on a Multinomial model. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1861-6

[3]

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single‐cell RNA‐SEQ Analysis: A tutorial. *Molecular Systems Biology*, *15*(6). https://doi.org/10.15252/msb.20188746

[4]

Wagner, F., Barkley, D., & Yanai, I. (2019). Accurate denoising of single-cell RNA-seq data using unbiased principal component analysis. https://doi.org/10.1101/655365

[5]

Friedman, J. H. (1997). *Data Mining and Knowledge Discovery*, *1*(1), 55–77. https://doi.org/10.1023/a:1009778005914

[6]

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420. https://doi.org/10.1038/nbt.4096

[7]

Levandowsy , Michael, & Winter, David (1971). Distance between sets. *Nature*, *234*(5323), 34–35. https://doi.org/10.1038/234034a0

[8]

Santos, J. M., & Embrechts, M. (2009). On the use of the adjusted Rand Index as a metric for evaluating supervised classification. *Artificial Neural Networks – ICANN 2009*, 175–184. https://doi.org/10.1007/978-3-642-04277-5_18

[9]

Cai, T. T., & Ma, R. (2022, March 28). *Theoretical Foundations of T-Sne for visualizing high-dimensional clustered data*. arXiv.org. Retrieved September 23, 2022, from https://arxiv.org/abs/2105.07536v3

[10]

Tyler, S. R., Rotti, P. G., Sun, X., Yi, Y., Xie, W., Winter, M. C., Flamme-Wiese, M. J., Tucker, B. A., Mullins, R. F., Norris, A. W., & Engelhardt, J. F. (2019). Pyminer finds gene and autocrine-paracrine networks from human islet scrna-seq. *Cell Reports*, *26*(7). https://doi.org/10.1016/j.celrep.2019.01.063

[11]

Thorel, F., Népote, V., Avril, I., Kohno, K., Desgraz, R., Chera, S., & Herrera, P. L. (2010). Conversion of adult pancreatic α-cells to β-cells after extreme β-cell loss. *Nature*, *464*(7292), 1149–1154. https://doi.org/10.1038/nature08894