RESEARCH ARTICLE

REVISED The need to reassess single-cell RNA sequencing datasets: the importance of biological sample processing [version 2; peer review: 2 approved]

Previously titled: The need to reassess single-cell RNA sequencing datasets: more is not always better

Alex M. Ascensión[1,2], Marcos J. Araúzo-Bravo[1,3-6], Ander Izeta[2,7]

[1]Computational Biology and Systems Biomedicine Group, Biodonostia Health Research Institute, San Sebastian, Gipuzkoa, 20014, Spain
[2]Tissue Engineering Group, Biodonostia Health Research Institute, San Sebastian, Gipuzkoa, 20014, Spain
[3]Computational Biomedicine Data Analysis Platform, Biodonostia Health Research Institute, San Sebastian, Gipuzkoa, 20014, Spain
[4]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[5]CIBER of Frailty and Healthy Aging (CIBERfes), Madrid, Spain
[6]Computational Biology and Bioinformatics Group, Max Planck Institute for Molecular Biomedicine, Münster, Germany
[7]Department of Biomedical Engineering and Science, Tecnun-University of Navarra, School of Engineering, San Sebastian, Gipuzkoa, 20009, Spain

v2 **First published:** 06 Aug 2021, **10**:767
https://doi.org/10.12688/f1000research.54864.1
Latest published: 08 Mar 2022, **10**:767
https://doi.org/10.12688/f1000research.54864.2

Abstract

Background: The advent of single-cell RNA sequencing (scRNAseq) and additional single-cell omics technologies have provided scientists with unprecedented tools to explore biology at cellular resolution. However, reaching an appropriate number of good quality reads per cell and reasonable numbers of cells within each of the populations of interest are key to infer relevant conclusions about the underlying biology of the dataset. For these reasons, scRNAseq studies are constantly increasing the number of cells analysed and the granularity of the resultant transcriptomics analyses.
Methods: We aimed to identify previously described fibroblast subpopulations in healthy adult human skin by using the largest dataset published to date (528,253 sequenced cells) and an unsupervised population-matching algorithm.
Results: Our reanalysis of this landmark resource demonstrates that a substantial proportion of cell transcriptomic signatures may be biased by cellular stress and response to hypoxic conditions.
Conclusions: We postulate that careful design of experimental conditions is needed to avoid long processing times of biological samples. Additionally, computation of large datasets might undermine the extent of the analysis, possibly due to long processing times.

Open Peer Review

Approval Status ✓ ✓

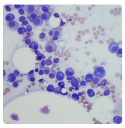
	1	2
version 2 (revision) 08 Mar 2022	✓ view	✓ view
version 1 06 Aug 2021	? view	? view

1. **Xiao Long**, Chinese Academy of Medical Sciences, Beijing, China
Zhujun Li, Chinese Academy of Medical Sciences, Beijing, China

2. **(Jo) Huiqing Zhou**, Radboud University, Nijmegen, The Netherlands
Radboudumc, Nijmegen, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
single-cell RNA-seq, skin, fibroblasts, reproducibility, computational analysis, Python



This article is included in the Cell & Molecular Biology gateway.

This article is included in the Max Planck Society collection.

Corresponding authors: Marcos J. Araúzo-Bravo (mararabra@yahoo.co.uk), Ander Izeta (ander.izeta@biodonostia.org)

Author roles: Ascensión AM: Conceptualization, Investigation, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Araúzo-Bravo MJ**: Funding Acquisition, Project Administration, Writing – Review & Editing; **Izeta A**: Funding Acquisition, Investigation, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Instituto de Salud Carlos III grant AC17/00012, co-funded by the European Union (ERDF/ESF, "Investing in your future") (MJAB) ERA-Net program Era-coSysMed, JTC-2 2017 (MJAB) Instituto de Salud Carlos III grant PI19/01621, co-funded by the European Union (ERDF/ESF, "Investing in your future") (AI) Basque Government PhD fellowship PRE_2019_2_0233 (AMA), Horizon 2020 Eracosysmed, Grant Agreement No 643271.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Ascensión AM, Araúzo-Bravo MJ and Izeta A. **The need to reassess single-cell RNA sequencing datasets: the importance of biological sample processing [version 2; peer review: 2 approved]** F1000Research 2022, **10**:767 https://doi.org/10.12688/f1000research.54864.2

First published: 06 Aug 2021, **10**:767 https://doi.org/10.12688/f1000research.54864.1

REVISED **Amendments from Version 1**

We have incorporated the reviewers' suggestions and recommendations in the new version. Accordingly, the revisions have been made as follows:

1) Title has been updated
2) It has been given more relevance to the biological processing effects of the samples.
3) The relationship between long processing times and biases in the analysis has been updated, adding Figure 4 and Supplementary Table 1.

Any further responses from the reviewers can be found at the end of the article

Introduction

The quest for deciphering the underlying biology of numerous phenomena at the single-cell level has exponentially increased the number of published single-cell RNA sequencing (scRNAseq) studies.[1] Additionally, individual studies are gradually increasing in scale, and in most tissues a correlation between the numbers of cells sequenced and the number of identified cell types is found.[1] Unfortunately, many (if not most) of the studies concentrate their efforts on individual dataset analyses and perform relatively little correlative study to meta-analyse previously published scRNAseq datasets. However, the amount of information that could be retrieved from the already existing corpus of literature is enormous.[2]

Within identified cell clusters (what we normally would define as "cell types"), the existing cell heterogeneity may be indicative of cell subsets that respond to particular conditions (such as cell cycle phase, cell stress, response to local signals, etc.) or reflect underlying functional/positional differences.[3-6] It is thus of utmost importance that the scientific community interested in a specific tissue or cell type agrees on the existing subsets within particular cell types and their defining molecular profiles, so that a common reference atlas may be used to understand homeostasis and response to varying insults.[7]

In a re-analysis of 13,823 human adult dermal fibroblasts obtained from four independent scRNAseq studies,[8-11] we recently proposed that human skin presents a common set of fibroblast subsets, irrespective of donor area.[12] These subsets can be categorised into three main fibroblast types (type A, B, and C), with a total of 10 minor subpopulations (A1–A4, B1–B2, C1–C4). In a recent landmark paper published in Science, Reynolds *et al.* produced a dataset of 528,253 sequenced cells obtained from healthy adult skin (five female patients undergoing mammoplasty surgery) and fetal samples, as well as inflamed skin from atopic dermatitis and psoriasis patients.[13] In healthy dermal fibroblasts, the authors described three populations: a main cluster termed Fb1, and two minor subpopulations, Fb2 and Fb3. Fb2 was additionally described as enriched in fetal and inflamed skin samples.[13] We aimed to analyse whether the Fb1, Fb2 and Fb3 populations were consistent with the A–C fibroblast types and subtypes that we had just described. More specifically, we reasoned that at least the most abundant subpopulations that we had defined, namely A1, A2, B1 and B2, should be clearly detected in a >500k cell dataset, thus further validating our previous scRNAseq study. In contrast, we found that a substantial proportion of the Reynolds *et al.* scRNAseq dataset shows a predominancy of differential expression of stress and hypoxia-related genes. Thus, data extracted from this source should be interpreted in the light of this bias. It is possible that other existing large datasets suffer from similar methodological problems, which might be due to insufficient oversight.

Methods

Preprocessing of fibroblast sample data

Fibroblast sample data originated from five donors, as described by Reynolds *et al.*,[13] and were processed from raw fastq files (E-MATB-8142). The ID numbers are 4820STDY7388991 [S1], 4820STDY7388999 [S2], 4820STDY7389007 [S3], SKN8104899 [S4], SKN8105197 [S5]. Fastq files were processed using the `loompy fromfq` pipeline described in https://linnarssonlab.org/loompy/kallisto/index.html. *Loompy* (RRID:SCR_016666) and *kallisto* (RRID: SCR_016582) versions are 3.0.6 and 0.46.0. Genome fasta index and annotations were based on GRCh38 Gencode v31 (RRID:SCR_014966). Additionally, for other annotations and analysis of other populations, the processed h5ad adata from[13] was downloaded from the Zenodo repository (ID: 4536165).

Each individual sample (S1–C fibroblast types and subtypes that we had just descS5) data was processed equally using the following *scanpy* (RRID:SCR_018139, v1.7.0rc1)[14] procedure. To map the clusters from the original publication, cells from the processed data set were extracted and mapped to the samples. Genes with fewer than 30 counts were discarded. The sample was normalised (`sc.pp.normalize_per_cell`) and log-transformed. Then, Principal Component Analysis (PCA) with 30 components was calculated and feature selection was performed with

triku[15] (RRID:SCR_020977, v1.3.1), and *k*NN with cosine metric were computed. Finally, UMAP (RRID:SCR_018217, v0.4.6)[16] and leiden (v0.8.3)[17] were applied to detect the fibroblast populations.

Most of the cells from the preprocessed adata were mapped to the raw dataset. However, additional unmapped cells appeared, some of them related to other cell types (e.g. keratinocytes, immune cells or perivascular cells). To assign unmapped cells to their corresponding cell types a population matching algorithm was applied (described below). This algorithm requires a dictionary of cell types and markers. The markers used were the following:

- Fibroblast: *LUM*, *PDGFRA*, *COL1A1*, *SFRP2*, *CCL19*.
- Perivascular cell: *RGS5*, *MYL9*, *NDUFA4L2*.
- Erithrocyte: *HBB*, *HBA2*, *HBA1*.
- Immune cell: *TPSB2*, *TPSAB1*, *HLA-DRA*, *FCER1G*, *CD74*.
- Melanocyte: *PMEL*, *MLANA*.
- Endothelial vascular cell: *CLDN5*, *PECAM1*.
- Keratinocyte: *DMKN*, *KRT1*, *KRT5*.
- Mitochondrial content (low quality): *MTND2P8*, *MTND4P12*, *MTCO1P40*, *ADAM33*, *RN7SL2*, *MTRNR2L6*.

Once cell types have been assigned, non-fibroblast cells were discarded, and the PCA, triku, *k*NN, UMAP, leiden cycle was repeated to recalculate the new cell projection.

The sample S5 was discarded from the analysis due to its lack of *SFRP2* expression, a well established fibroblast marker that is expressed in the rest of samples.[12]

Then, we separated the Fb2 population from the Fb1 and Fb3 populations for each dataset and applied the population matching algorithm to annotate them with the labels assigned from.[12] The genes used for the population assignation were the following:

- A1: *PI16*, *QPCT*, *SLPI*, *CCN5*, *CPE*, *CTHRC1*, *MFAP5*, *PCOLCE2*, *SCARA5*, *TSPAN8*
- A2: *APCDD1*, *COL18A1*, *COMP*, *NKD2*, *F13A1*, *HSPB3*, *LEPR*, *TGFBI*
- B1: *CXCL2*, *MYC*, *C7*, *SPSB1*, *ITM2A*
- B2: *SOCS3*, *CCL19*, *CD74*, *RARRES2*, *CCDC146*, *IGFBP3*, *TNFSF13B*
- C: *CRABP1*, *PLXDC1*, *RSPO4*, *ASPN*, *F2R*, *POSTN*, *TNN*

Next, all datasets with Fb1 and Fb3, or Fb2 populations were joined. We applied the previous processing routine and, to correct for batch effects, we used *bbknn* (v1.4.0)[18] with `metric=angular` and `neighbors_within_batch=2` parameters.

To analyse the transcriptomic profile between Fb1 and Fb3, and Fb2 populations, we joined the two datasets and applied the same processing pipeline as before. We first characterised the genes driving the differences by obtaining the DEGs between the two sets of populations, and running GOEA with the first 150 DEGs of each category. The set of ontologies used was *GO Biological Process 2018* with the module *gseapy* (v0.10.4).[19] Then, to assess that the differences were due to cellular stress in the Fb2 population, we downloaded the lists of genes mentioned in the Results section (gene lists are available in the Github repository below), and genes appearing in more than two lists were selected. Then, the population matching algorithm was run against this list, and clusters with scores lower than 0.55 were assigned as ”Non-stress” clusters.

To analyse the differences in transcriptomic profiles within Fb1 and Fb3 populations, we obtained the DEGs between the two sets of A2 populations, which were the easiest to separate in clusters. By using that list of DEGs, we applied the population matching algorithm and divided the Fb1 and Fb3 populations into two halves. We then obtained the DEGs between the two halves and ran GOEA with the first 150 DEGs of each category, which revealed a hypoxia pattern in one of the halves. To assess that the differences were due to hypoxia, we downloaded the lists of hypoxia-related genes, and genes appearing in more than two lists were selected. Since some key genes (some glycolysis genes, or important genes appearing in one list) were missing, they were manually added to obtain a more robust list. Then, the population matching algorithm was run against this list, as well as the list of stress-related genes, and clusters with scores lower than 0.5 were assigned as "Normal" clusters.

To replicate the analysis on the rest of the cell types, we used the processed h5ad file.

Correction of stress and hypoxia cell states

In order to correct for stress and hypoxia cell states we used the `sc.pp.regress_out` implementation from *scanpy* on the stress and hypoxia scores. We first created two sub-datasets, one containing stress and normal cells, and another one with hypoxia and normal cells, and then the scores were regressed out. Finally, the common processing pipeline was applied. Additional correction methods can be seen in the notebooks in the Zenodo repository.[20]

Population matching algorithm

The aim of this algorithm is to assign a set of clusters to a set of labels, where each label contains a list of representative markers. For each label we extracted the matrix of counts of the genes belonging to the label. Then, we created a new matrix, where we assigned to each cell and gene the sum of the counts of the gene within its *k*NN, divided by the number of neighbours. This step reduced the noisiness of the expression, and also exacerbated the local expression of a gene and dampened the expression of sparse genes.

Gene expression values were substituted by the ranked index of their expression; and the values were divided by the largest index to sum 1. Therefore, the cell with the highest expression had a value of 1 for that gene, while the lowest expressed cell had a near 0 value. After this normalisation was applied to the rest of genes within the label, the mean of the normalised values across genes was computed, so that each cell had one value for that label.

After the previous steps were computed for the rest of labels, a new matrix with the number of clusters by the number of labels was computed. For each label and each cluster, the percentile of the normalised values within cells of that cluster was computed (percentile 70 by default). This helped reduce noise on normalised values, and assigned a unique number per cluster.

This algorithm allowed choosing for intermediate states, that is cell labels with a high similarity. By default, the label with the highest score per cluster was chosen. With the intermediate state option, labels that had a similar value as the label with the highest value were included. The difference in values was set as a threshold (0.05 by default), and labels with a difference of a value greater than the threshold were not merged.

Results

Reassessment of the main cell populations in a large skin dataset reveals the presence of clusters with stress- and hypoxia-related gene signatures

By using an unsupervised population-matching algorithm (details in processed notebooks available online[20]) we observed that in each of the healthy donors analysed by Reynolds *et al.*,[13] at least two independent fibroblast clusters expressed signature markers of the A1, A2, B1 and B2 populations. One set of cells corresponded to the Fb2 population, and the second set corresponded to the Fb1 and Fb3 populations. A joint analysis of all donors after batch effect correction showed that the cluster duplication observed in each individual donor could be replicated jointly. We therefore assumed that some global effect should be affecting the cells, i.e. Fb2 might be a copy of Fb1+Fb3 cells, although perhaps affected by some alteration. Differential gene expression (DEG) analysis between Fb2 and Fb1+Fb3 revealed an enrichment in ontology terms associated to cell stress (e.g. unfolded protein response, regulation of apoptotic process, mRNA catabolic process). We then designed a signature gene list composed of 50 DEGs commonly associated to stress in very different scRNAseq settings (e.g. *ATF3*, *BTG2*, *FOS*, *FOSB*, *GADD45B*, *HSPA1A/B*, *IER2/3*, *JUN*, *JUNB*, *NFKBIA*, *NR4A1/2*, *PPP1R15A*, *RHOB*).[21-25] Using this signature, the Fb2 population over-expressed *BTG2*, *EGR1*, *FOSB*, *IER2*, *SOCS3*, and *ZFP36*, among others, indicating that these cells clustered together mainly due to cellular stress.

In a further analysis of the Fb1 and Fb3 cells, we observed that the A1, A2, B1 and B2 populations appeared twice again. A DEG analysis between each pair of duplicated populations disclosed genes in one of the split populations that were

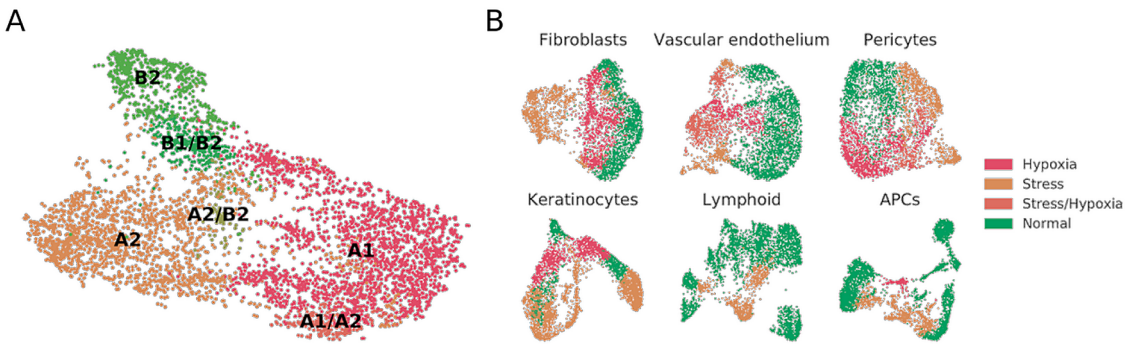


Figure 1. A re-analysis of the Reynolds et al. dataset in search of dermal fibroblast subpopulations reveals the presence of substantial proportions of stressed and hypoxic cells. (A) UMAP plot of normal fibroblasts (after removal of hypoxic and stressed cell subsets) reveals conservation of some, but not all, cell types previously described in independent datasets.[12] (B) UMAP plot of fibroblast, vascular endothelium, pericyte, keratinocyte, lymphoid and APC cell populations from healthy donors, labeled to highlight hypoxic and stressed cell subpopulations as characterized by overexpression of defined gene signatures.

related to glycolysis (*ALDOC*, *ENO2*, *GAPDH*, *PGK1*, *PDK1*, *PFKFB4*, *PYGL*), cell integrity, hypoxia and apoptosis (*BNIP3*, *BNIP3L*, *ANGPTL4*, *LOX*, *HILPDA*); whereas the second split population over-expressed units of the mitochondrial ATPase and complex I, indicating an active oxidative metabolism. It is well known that cells under hypoxic conditions switch from aerobic to anaerobic metabolism to keep energy homeostasis within the cell.[26-28] We therefore generated a curated list of hypoxia-related genes, and managed to separate the non-hypoxic from the hypoxic group with the population-matching algorithm. Once stressed or hypoxic cells were removed on the basis of a set threshold of expression of signature genes, we mapped the main types of fibroblasts in what we termed normal cell subset of Reynolds *et al.* (Figure 1A). Fibroblast A1, A2 and B2 populations were independently mapped, and we also found clusters which seemingly were mixtures of previously defined populations e.g. B1/B2, A1/A2, or A2/B2. No type C fibroblasts were detected.

To understand whether the stress and hypoxic signatures were only present in fibroblast subsets or could also be traced to other populations within the Reynolds *et al.* dataset, we mapped the stress and hypoxia gene signatures to perivascular cell, keratinocyte, vascular endothelial cell, lymphoid cell, and antigen presenting cell (APC) clusters. In our reanalysis of healthy donors, fibroblasts, perivascular cells, keratinocytes, and vascular endothelial cells showed clear hypoxia and stress-related clusters (Figure 1B). For instance, the VE3 population, described by Reynolds *et al.* as increased in patients suffering from inflammatory conditions, presented a clear stress-related transcriptomic profile. On the other hand, most of the VE2 population over-expressed hypoxia-related genes. On lymphoid cells we did observe a sub-cluster of stressed Tc/Th cells but no clear hypoxic profiles. On APCs, an inflammatory macrophage cluster showed hypoxia, and the M2 and DC2 clusters showed stress-related profiles. Some of these results may be expected in physiological conditions for immune cells, but others could be attributed to sample handling.

Finally, we tested if the aforementioned stress and hypoxia related signatures were present in the previously published scRNAseq datasets of human skin.[8-11] The levels of expression of these genes were clearly higher in the Reynolds *et al.* dataset as compared to other available resources (Figure 2).

Correction of stress and hypoxia signatures shows that stressed cells show a non-recoverable gene signature

Since the stress and hypoxia related expression profiles are apparent, we were interested in studying the ”reversibility” of the transcriptomic signatures, and creating a normalised dataset where hypoxic and stressed cells could merge with the normal cells, and classifying the whole dataset into the original cell types described in.[12] To this end, we applied two approaches with similar results. On the one hand, we considered cell states as batches, and applied batch effect correction with *bbknn* and *harmony*. On the other hand, we applied regression on the stress and hypoxia scores shown in Figure 2 based on the *Seurat*'s linear regression function implemented in *scanpy*. Since both approaches showed similar results, we show the results of the latter case in Figure 3.

To further study if stress and hypoxia transcriptomic profiles are ”recoverable”, we generated two types of datasets, one each with the stress or hypoxia cells, and another one containing normal cells. When applied the correction to the stress + normal dataset we observed that there was no integration between the two states (Figure 3A). On the other hand, there was

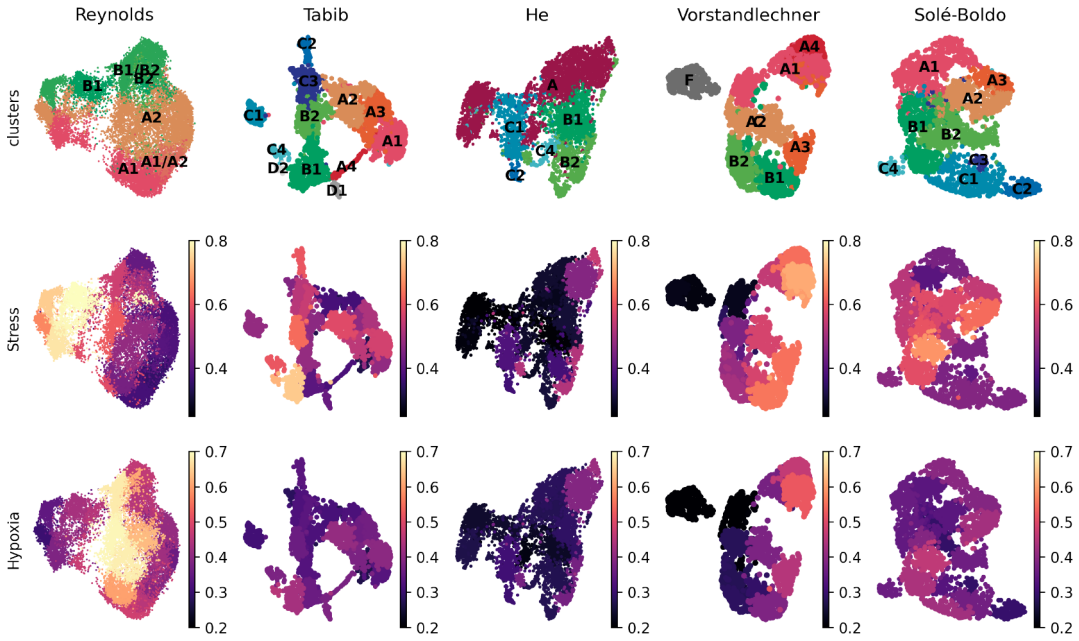


Figure 2. Stress and hypoxia-related signatures in published human dermal fibroblast datasets. (A) UMAP plot of normal fibroblasts (after removal of hypoxic and stressed cell subsets) reveals conservation of some, but not all, cell types previously described in independent datasets (1). (B) UMAP plot of human dermal fibroblast subsets as defined in[12] are shown here for five published datasets,[8-11,13] and depicted by the average levels of expression of stress and hypoxia gene signatures.

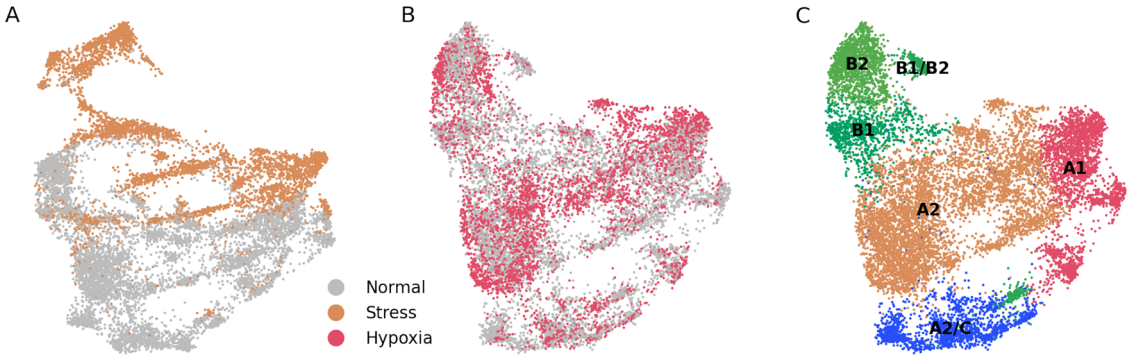


Figure 3. Dataset merging of stress and hypoxia populations show mixed degrees of integration with the "normal" dataset. (A) UMAP plot of merged "stress" and "normal" cells. There is a low degree of integration between both cell types. (B) UMAP plot of merged "hypoxia" and "normal" cells. There is a high degree of integration between both cell types. (C) Unsupervised assignation of fibroblast types from (B) reveals, similar to results from Figure 1A, major fibroblast types.

a good integration between the hypoxia and normal cell states (Figure 3B), and the main fibroblast populations could be correctly mapped (Figure 3C). From these results we infer that the transcriptome from stressed cells is much more altered than the one from hypoxic cells, to the extent that stressed cells are in a computationally non-reversible state.

Analysis on running times of the Reynolds dataset with different numbers of cells

The Reynolds dataset contains, after some basic filterings, approximately 450k cells. We became interested in analysing the runtimes of a standard single-cell pipeline procedure – consisting on quality control, PCA, graph neighbor construction, dimensionality reduction, clustering and DEG calculation – using different cell numbers, to see how this analysis is scaled. The results of the analysis are observed in Figure 4 and detailed in Supplementary Table 1.

The analysis shows that running the pipeline a single time in the whole dataset in a working station takes approximately 1 hour. The parts with the longest running times are the batch, clustering and DEG calculation. Additionally, when analysing trends in the processing times, we observe an inflection point at around 30,000 cells, marking two clear runtime

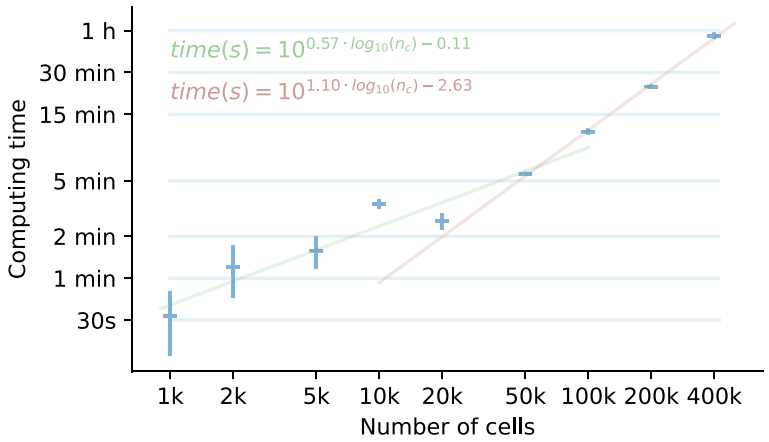


Figure 4. Running times of a basic single-cell pipeline on Reynolds dataset. For each number of cells, three running times are collected, and the mean (horizontal bar) and standard deviation (vertical bar) are shown. For the [1k - 20k] and [50k - 400k] intervals, two linear regressions were trained and extended outside of these intervals to show that there is a change in the processing time rate at ~30k cells. A doubling in the cell number implies 1.48 times more runtime in the [1k - 20k] interval, and 2.14 times in the [50k - 400k] interval.

trends. For higher numbers of cells, the processing times are further increased – 2.1 times per doubling of cells – compared with a lower number of cells – 1.5 times per doubling of cells.

For good measure, a single run of this pipeline analysis on a extended dataset with 1M cells would take 2 hours and 25 minutes.

Discussion

The results from the efforts to compare, correlate, and compile the information present in available scRNAseq datasets could be condemned to short longevity since they can be overpassed by new resources that appear almost on a daily basis. However, it is to be expected that, at some tipping point, robust cell types and subtypes will be fully defined for each tissue and organ. Then, new scRNAseq datasets will only add information on the transcriptomically defined cell states of each of the robustly defined cell subpopulations, in response to specific perturbations such as injury or disease.

Here, we aimed to validate results we had obtained with a few thousand cells with a large scRNAseq dataset including over half a million cells. Instead, we have found that clustering of this large dataset shows an apparent pattern of expression of stress and hypoxia-related genes. In our opinion, the origin of stress- and hypoxia-related signatures in healthy donor cells might be caused by two factors: a tough tissue processing that put the cells under stress and hypoxia, and an underdeveloped analysis caused by the long processing time derived from the high number of cells, which hindered the detection of bad quality cells and propagated this artefact downstream the analysis.

The first factor is related to the very exhaustive and complex protocol for cell isolation chosen by authors. The top 200 μm-thick layer of the skin was cut with a dermatome, digested with dispase (1 h at 37°C) to separate dermal and epidermal layers. Both layers were digested in collagenase for 12 h at 37°C, cells were filtered and subjected to FACS sorting before library generation and sequencing.[13] While this strategy warrants high purity of the obtained cell populations, the long processing times (≥16 h) and the use warm dissociation for a long period might have significantly affected patterns of gene expression of relevant numbers of cells in this setting. In this sense, aiming to process large numbers of cells involves longer processing times. High processing times (even ≥ 60') have previously been reported to generate significant transcriptomic alterations.[21,25] and, in particular, warm dissociation is associated with stress response,[29] which is apparent in the transcriptomic profiles of part of the cells . An alternative to warm dissociation may be the use of cold-active psychrophilic proteases.[30]

The second factor is related to the computational and analytic challenges of such a complex dataset. As it has been observed in Figure 4, runtime of analytic pipelines vastly increase with the numbner of cells. Thus, if the processing time is expected to be the same for a small dataset and a big dataset, due to the low time to perform a beginning-to-end analysis adapted to the current fast-paced times of publication, this leads to a more shallow exploration at the initial stages. Before a pipeline is run on a single-cell dataset, researchers usually have to spend some time doing an exploratory analysis, where they select the cutoff values for quality control, explore different batch effect removal methods, or tune the parameters for clustering, neighbour graph calculation, and other steps in the pipeline.

These decisions are made on the basis of the output of the differently-preprocessed versions on downstream analyses: how the datasets look on UMAP plots, how robust their DEGs are, etc. Usually, this part of the analysis requires several reruns of the same pipeline to find the best parameters and obtain an overall view of the limitations of the dataset and the general information elements that will be obtained from it. This means that single-cell pipelines are not a linear, but rather an iterative process where researchers have to make decisions based on the output of previous steps. As a consequence, if the results from initial stages of the analysis are overlooked and biases go unnoticed, these effects propagate downstream the pipeline – e.g. observing differences in healthy vs diseased samples, search for rare populations, pathway/ontology analysis, or RNA velocity analysis – and artefacts can be presented as genuine results, hindering the dissemination of quality results to the scientific community.

In conclusion, understanding skin fibroblast heterogeneity is of great relevance not only in skin homeostasis, but also in ageing[11,31] and disease.[32-36] We sincerely hope these reanalyses help further advance the field of single-cell transcriptomics of human skin. Further refinement of fibroblasts subsets and their identity-defining features will provide a fruitful framework for the advancement of knowledge as well as for the development of novel therapeutic approaches in dermatological disease and skin cancer.

Data availability

Extended data

Repository: Extended data for "The need to reassess single-cell RNA sequencing datasets: the importance of biological sample processing". https://doi.org/10.5281/zenodo.6324956.[37]

This project contains the following underlying data:

Supplementary Table 1 (Processing times of different elements of the single-cell pipeline, varying the number of cells analysed).

Software availability

Notebooks to replicate this work can be found at: https://github.com/alexmascension/revisit_reynolds_fb.

Processed notebooks and AnnData files can be found at: https://doi.org/10.5281/zenodo.4596374.[20]

References

1. Svensson V, da Veiga Beltrame E, Pachter L: **A curated database reveals trends in single-cell transcriptomics.** *Database.* 2020; **2020**.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Phan QM, Driskell IM, Driskell RR: **The three rs of single-cell rna sequencing: reuse, refine, and resource.** *J Invest Dermatol.* 2021; **141**(7): 1627–1629.
PubMed Abstract | Publisher Full Text

3. Schuster R, Rockel JS, Kapoor M, *et al.*: **The inflammatory speech of fibroblasts.** *Immunol Rev.* 2021.
PubMed Abstract | Publisher Full Text

4. Sawant M, Hinz B, Schönborn K, *et al.*: **A story of fibers and stress: Matrix-embedded signals for fibroblast activation in the skin.** *Wound Repair Regen.* 2021; **29**: 515–530.
PubMed Abstract | Publisher Full Text

5. Correa-Gallegos D, Rinkevich Y: **Cutting into wound repair.** *FEBS J.* 2021.
PubMed Abstract | Publisher Full Text

6. Jiang D, Rinkevich Y: **Distinct fibroblasts in scars and regeneration.** *Curr Opin Genet Dev.* 2021; **70**: 7–14.
PubMed Abstract | Publisher Full Text

7. Puntambekar S, Hesselberth JR, Riemondy KA, *et al.*: **Cell-level metadata are indispensable for documenting single-cell sequencing datasets.** *PLoS Biol.* 2021; **19**(5).
PubMed Abstract | Publisher Full Text | Free Full Text

8. Tabib T, Morse C, Wang T, *et al.*: **Sfrp2/dpp4 and fmo1/lsp1 define major fibroblast populations in human skin.** *J Invest Dermatol.* 2018; **138**(4): 802–810.
PubMed Abstract | Publisher Full Text | Free Full Text

9. He H, Suryawanshi H, Morozov P, *et al.*: **Single-cell transcriptome analysis of human skin identifies novel fibroblast subpopulation and enrichment of immune subsets in atopic dermatitis.** *J Allergy Clin Immunol.* 2020; **145**(6): 1615–1628.
PubMed Abstract | Publisher Full Text

10. Vorstandlechner V, Laggner M, Kalinina P, *et al.*: **Deciphering the functional heterogeneity of skin fibroblasts using single-cell rna sequencing.** *FASEB J.* 2020; **34**(3): 3677–3692.
PubMed Abstract | Publisher Full Text

11. Solé-Boldo Llorenç, Raddatz Günter, Schütz S, *et al.*: **Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming.** *Commun Biol.* 2020; **3**(1).
PubMed Abstract | Publisher Full Text | Free Full Text

12. Ascensión AM, Fuertes-Álvarez S, Ibañez-Solé O, *et al.*: **Human dermal fibroblast subpopulations are conserved across single-cell rna sequencing studies.** *J Invest Dermatol.* 2021; **141**(7): 1735–1744.
PubMed Abstract | Publisher Full Text

13. Reynolds G, Vegh P, Fletcher J, *et al.*: **Developmental cell programs are co-opted in inflammatory skin disease.** *Science.* 2021;

371(6527).
PubMed Abstract | Publisher Full Text

14. Wolf FA, Angerer P, Theis FJ: **Scanpy: large-scale single-cell gene expression data analysis.** *Genome Biol.* 2018; **19** (1).
PubMed Abstract | Publisher Full Text | Free Full Text

15. Ascensión AM, Ibañez-Solé O, Inza I, *et al.*: **Triku: a feature selection method based on nearest neighbors for single-cell data.** *bioRxiv.* 2021.
Publisher Full Text

16. McInnes L, Healy J, Melville J: **Umap: Uniform manifold approximation and projection for dimension reduction.** *arXiv preprint arXiv:1802.03426.* 2018.

17. Traag VA, Waltman L, van Eck NJ: **From louvain to leiden: guaranteeing well-connected communities.** *Sci Rep.* 2019; **9**(5233).
Publisher Full Text

18. Polańs ki K, Young MD, Miao Z, *et al.*: **Bbknn: Fast batch alignment of single cell transcriptomes.** *Bioinformatics.* 2020; **36**(3): 964–965.
PubMed Abstract | Publisher Full Text

19. Fang Z, Wolf A, Liao Y, *et al.*: **zqfang/gseapy: gseapy-v0.10.3.** February 2021.
Publisher Full Text

20. Ascensión AM, Araúzo-Bravo MJ, Ander I: **The need to reassess single-cell rna sequencing datasets: more is not always better.** *Zenodo.* 2021.
Publisher Full Text

21. van den Brink SC, Sage F, Vértesy Ábel, *et al.*: **Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations.** *Nat Methods.* 2017; **14**(10): 935–936.
PubMed Abstract | Publisher Full Text

22. O`Flanagan CH, Campbell KR, Zhang AW, *et al.*: **Dissociation of solid tumor tissues with cold active protease for single-cell rna-seq minimizes conserved collagenase-associated stress responses.** *Genome Biol.* 2019; **20**(1).
PubMed Abstract | Publisher Full Text | Free Full Text

23. Denisenko E, Guo BB, Jones M, *et al.*: **Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows.** *Genome Biol.* 2020; **21**(1).
PubMed Abstract | Publisher Full Text | Free Full Text

24. Adam M, Potter AS, Potter SS: **Psychrophilic proteases dramatically reduce single-cell rna-seq artifacts: a molecular atlas of kidney development.** *Development.* 2017; **144**(19): 3625–3632.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Waise S, Parker R, Rose-Zerilli MJ, *et al.*: **An optimised tissue disaggregation and data processing pipeline for characterising fibroblast phenotypes using single-cell rna sequencing.** *Sci Rep.* 2019; **9**: 9580.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Xiao Z, Dai Z, Locasale JW: **Metabolic landscape of the tumor microenvironment at single cell resolution.** *Nat Commun.* 2019; **10**: 3763.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Mohyeldin A, Garzón-Muvdi Tomás, Quiñones-Hinojosa A: **Oxygen in stem cell biology: A critical component of the stem cell niche.** *Cell Stem Cell.* 2010; **7**(2): 150–161.
PubMed Abstract | Publisher Full Text

28. Simon MC, Keith B: **The role of oxygen availability in embryonic development and stem cell function.** *Nat Rev Mol Cell Biol.* 2008; **9**(4): 285–296.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Denisenko E, Guo BB, Jones M, *et al.*: **Systematic assessment of tissue dissociation and stor-age biases in single-cell and single-nucleus rna-seq work-flows.** *Genome Biol.* 2020; **21**(1): 130.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Potter AS, Potter SS: **Dissociation of tis-sues for single-cell analysis.** *Methods Mol Biol.* 2019; 55–62.
PubMed Abstract | Publisher Full Text

31. Zou Z, Long X, Zhao Q, *et al.*: **A single-cell transcriptomic atlas of human skin aging.** *Dev Cell.* 2021; **56**(3): 1–15.
PubMed Abstract | Publisher Full Text

32. Rojahn TB, Vorstandlechner V, Krausgruber T, *et al.*: **Single-cell transcriptomics combined with interstitial fluid proteomics defines cell-type-specific immune regulation in atopic dermatitis.** *J Allergy Clin Immunol.* 2020; **146**(5): 1056–1069.
PubMed Abstract | Publisher Full Text

33. Gao Y, Yao X, Zhai Y, *et al.*: **Single cell transcriptional zonation of human psoriasis skin identifies an alternative immunoregulatory axis conducted by skin resident cells.** *Cell Death Dis.* 2021; **12**: 450.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Kim J, Lee J, Kim HJ, *et al.*: **Single-cell transcriptomics applied to emigrating cells from psoriasis elucidate pathogenic vs. regulatory immune cell subsets.** *J Allergy Clin Immunol.* 2021.
PubMed Abstract | Publisher Full Text

35. Liu X, Chen W, Zeng Q, *et al.*: **Single-cell RNA-seq reveals lineage-specific regulatory changes of fibroblasts and vascular endothelial cells in keloid.** *J Invest Dermatol.* 2021.
PubMed Abstract | Publisher Full Text

36. Liu J, Chang H-W, Huang Z-M, *et al.*: **Single-cell rna sequencing of psoriatic skin identifies pathogenic TC17 cell subsets and reveals distinctions between CD8+ T cells in autoimmunity and cancer.** *J Allergy Clin Immunol.* 2021; **147**(6): 2370–2380.
PubMed Abstract | Publisher Full Text

37. Ascensión AM, Araúzo-Bravo MJ, Izeta A, *et al.*: **Extended data for "The need to reassess single-cell RNA sequencing datasets: the importance of biological sample processing" [Data set].** *Zenodo.* 2022.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 05 April 2022

https://doi.org/10.5256/f1000research.122038.r126626

✔ **Xiao Long**
Department of Plastic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China
Zhujun Li
Department of Plastic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China

The authors made improvements according to the reviewers' comments. The revised title is better than the previous version.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Adipose derived stem cells

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 22 March 2022

https://doi.org/10.5256/f1000research.122038.r126627

✔ **(Jo) Huiqing Zhou**
[1] Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University, Nijmegen, The Netherlands
[2] Human Genetics, Radboudumc, Nijmegen, The Netherlands

I am fully satisfied with the additions and comments that the authors have put together.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: (Epidermal) stem cells; transcriptomis, epigenomics, developmental disease

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 20 December 2021

https://doi.org/10.5256/f1000research.58386.r95141

?

(Jo) Huiqing Zhou
[1] Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University, Nijmegen, The Netherlands
[2] Human Genetics, Radboudumc, Nijmegen, The Netherlands

In this article, the authors re-analysed a large dataset published in a high-impact journal and compared the outcome of analyzing this dataset with several other previously published works. This effort is highly appreciated and should be encouraged in the field, to critically evaluate and make sense of published work, with the ultimate aim to understand biology.

Specifically, the authors re-analysed the skin single-cell RNA-seq data published by Reynolds *et al*., 2021. They initially focused on the healthy fibroblast population, using the authors' previously published annotations, and identified an enhanced signature of stress- and hypoxia-related genes. They then further analysed several other cell populations of the dataset and reported that this enhanced stress and hypoxia signature is also present in other cell populations. Furthermore, using a normalised dataset where hypoxia and stressed cells could merge with the normal cells, they showed that the stress signature seems to be 'irreversible', in contrast to the hypoxia signature.

The workflow and methods seem to be appropriate, and the conclusion on the enhanced stress and hypoxia gene signature in the analysed dataset is also convincing (for a non-computational biologist who has a good understanding of common bioinformatics analysis and interpretation). This is consistent with the authors' discussions on the cell dissociation/processing methods. However, the title 'more is not always better' can be discussed and reconsidered. The authors should probably give advice on careful dissociation methods when retrieving a large number of cells, rather than proposing that a large number of cells is not necessary or desirable. In addition,

the authors did not give an extensive discussion on data analysis effort in analyzing the large datasets. It is also appropriate for the authors to comment on the data retrieval process of the analyses dataset, and to encourage the authors of the original paper to annotate their data properly and clearly (e.g., which methods were used to generate which batch of data). This is true for all publications and it is the only way for scientists to share and re-use the published data according to the FAIR principle (https://www.go-fair.org/fair-principles/).

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: (Epidermal) stem cells; transcriptomis, epigenomics, developmental disease

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 02 Mar 2022

Ander Izeta, Biodonostia Health Research Institute, San Sebastian, Spain

We thank the reviewer for the helpful and insightful comments. We will now comment on the issues point-by-point.

The title 'more is not always better' can be discussed and reconsidered

This comment is also in line with the other reviewer. Therefore we have changed the title to focus the paper mainly on biological sample processing.

The authors should probably give advice on careful dissociation methods when retrieving a large number of cells.

We agree with this excellent point. The original article had already quoted the reference by Denisenko *et al* (2020) where the importance of cold dissociation was demonstrated. The use of cold active psychrophilic proteases has been proposed to avoid use of enzymes that need incubation at 37ºC (Potter and Potter, 2019). Both references are now included in the text.

[...] the authors did not give an extensive discussion on data analysis effort in analyzing the large dataset

We analyzed the dadaw- Source code can be located as scripts in the GitHub repository https://github.com/haniffalab/HCA_skin as well as in the Zenodo repository https://zenodo.org/record/4249674. Although the scripts from the GitHub repository are conveniently ordered and are legible, the Zenodo repository does not include the output results and intermediate figures from the scripts, so we were unable to check the values and intermediate results. Also, for some parts of the scripts, further commenting would have been appreciated. Additionally, we observed that the file structure of the GitHub and Zenodo repository were not comparable. For instance, the \textit{Pipeline} folder with the structured analysis is lacking in the Zenodo repository, although some of its scripts are scattered through different folders. Another issue from the GitHub repository is that analysis scripts were uploaded in their final form to the repository (commits 1766cd, 9b520f, 9a361e and 123b7d, 8 Dec 2020). Considering that scripts from Zenodo were uploaded on November, some efforts to tidy the GitHub repository were made afterwards. However, a quick look at the scripts from the different sections shows a lack of variable consistency and file I/O, which implies a lack of reproducibility on their scripts. Additionally, despite a succinct explanation of the README file of the GitHub repository, the lack of commentaries on the scripts and the apparition of entire scripts that are not reflected in the methods difficult any replication effort.

Regarding the initial analysis of the dataset, from the python scripts we observed a set of common thresholds for all batches (*n_genes $<$ 6000, n_genes $>$ 400, n_counts $>$ 1000, percent_mito $<$ 0.2*). Although this is common practice, maybe a thresholding per batch would have been more convenient. For instance, we observed that sample SKN8105197 does not provide enough consistency in fibroblast marker expression. For convenience, this sample was removed from the analysis.

After QC and feature selection, bbknn is run with default parameters, although in other notebooks harmony has also been partially used. bbknn seems to be favored as the reference batch correction method. The lack of commit history does not allow us to look for previous attempts with other methods.

In the Pipeline/02 folder we found some scripts that use a logistic regression model for label stability prediction. We do not observe this script in the methods section.

In the Pipeline/03 folder we observe that an enhanced reclustering was made with hand-picked DEGs which, according to the authors, favored the separation of clusters. Interestingly, some of these DEGs are found on the hypoxia/stress lists (ZFP36, HSPA1A,

HSPA1B, DNAJB1, JUNB, ATF3, SOCS3, GADD45B, FOS), and others are ribosomal protein associated genes (RPL22, RPL37, RPL34). In our opinion, the selected DEGs might bias the reclustering for the segregation of hypoxic and stress populations.

It is also appropriate for the authors to comment on the data retrieval process of the analyses dataset, and to encourage the authors of the original paper to annotate their data properly and clearly

Regarding the data retrieval process, neither the GitHub nor the Zenodo repositories host the code for data retrieval and preprocessing described in materials and methods.

Competing Interests: No competing interests

Reviewer Report 01 December 2021

https://doi.org/10.5256/f1000research.58386.r98248

?

Xiao Long
Department of Plastic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China
Zhujun Li
Department of Plastic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China

The study reanalyzed data from several single-cell RNA sequencing resources, proposing an interesting point that a larger number of cells could lead to compromised results. It is well designed and conducted. Bioinformatic and statistic methods were valid.

The authors brought up an interesting point of view. However, large numbers of cells might not have a definite positive correlation with longer processing time. For example, as the authors discussed, the choice of the method during tissue processing could be an important contributing factor to the biased results. Therefore, the conclusion should focus more on the bias caused by processing time and choice of method, and maybe quality control measures, instead of larger cell numbers, because that would require further experiments and analysis to rule out these confounding factors.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Adipose derived stem cells

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 02 Mar 2022

Ander Izeta, Biodonostia Health Research Institute, San Sebastian, Spain

We thank the reviewer for the insight on focusing more on the biological processing of the sample. Although we are limited by the information from the article, we believe that a poor execution of the sample processing is a leading factor inducing the artifacts from the analysis. Indeed, we added some references supporting the claim that a general stress transcriptomic profile is produced by the long times that cells endured warm dissociation.

However, we believe that the high processing time and its probable consequence of insufficient analytical insight at the dataset cannot be ignored. In fact, we added a new figure (Figure 4) showing a relevant increase in running times when increasing number of cells analyzed. Additionally, We have observed a trend in publications at top-ranked journals that consist of the presentation of datasets with a diverse and complex set of cells, but which lack an in-depth analysis and do not produce biological results with enough insights. This trend affects, in the end, the quality of the datasets. In our opinion, this might be related to the limited time data scientists have been able to iterate analyses with that dataset.

We thus postulate that prospective authors should revise sample processing strategies as well as data analysis protocols, so that sampling errors can be pinpointed and corrected upstream in the analysis pipeline. In conclusion, we have reformulated our statements to increment the importance of biological sample processing as suggested by the reviewer.

Competing Interests: No competing interests