UNIVERSITAT DE
BARCELONA

Towards High Quality Single-cell Experiments: Approaches, Applications and Performance

Atefeh Lafzi

FACULTAT DE BIOLOGIA

DEPARTAMENT DE GENÈTICA

Programa de Doctorat en Biomedicina

Towards High Quality Single-cell Experiments: Approaches, Applications and Performance

Memòria presentada per

Atefeh Lafzi

per optar al grau de Doctora per la Universitat de Barcelona

Treball realitzat al Centre Nacional d'Anàlisi Genómica (CNAG) i Centre de Regulació Genòmica (CRG)

Doctorand

Atefeh Lafzi

Director	Co-director	Tutor
Holger Heyn	**Ivo Gut**	**Modesto Orozco**
Centre Nacional d'Anàlisi Genómica (CNAG)	Centre Nacional d'Anàlisi Genómica (CNAG)	Universitat de Barcelona Institute for Reseach in Biomedicine (IRB)

Acknowledgments

My PhD passed as one of the most amazing, special, joyful and pivotal periods of my life. I learned, tried, failed, succeeded, competed, fought, suffered, enjoyed, experienced, traveled, broke barriers and eventually grew. I grew to an authentic, independent, curious and brave version of me. A version of me, that today, I'm confident of…

But in this amazing journey, I wasn't alone. There were people who helped me to start and put me at the beginning of the road by believing in me without knowing me. There were people that committed to support me along this way and people that joined me in between, each bringing something valuable into the whole trip. Now that I'm at the end of the way, I would like to thank all those people one by one…

I would like to start with Ivo Gut, the one who believed in me and chose me for one of his best scholarship, providing all I need and more to do a good PhD. He was always there for me to make sure I had everything I needed, regardless of how good or bad I was doing things…

I would like to thank Holger Heyn. The one who sincerely was with me, guiding me, supporting me, hearing me and supervising me all along the way. Thank you very much for caring about what I wanted to do and where I wanted to be, and doing your best to facilitate the way. Thank you for being a good friend beside being a good supervisor. Thank you for all the positive attitude and motivation. For not stressing us even in the most stressful times. Thank you for letting me to enjoy my PhD.

Beside my advisors, I would like to thank Oliver Stegle and Fabian Theis who gave me the opportunity to join their lab for some periods during my PhD and learn more and more. My sincere gratitude goes to Davis McCarthy for all the time he spent to discuss and talk about single cells during my stay at EBI.

My sincere thanks also goes to the whole SINGEK ITN. Thank you Ramon Massana for creating this amazing Marie Curie network and thank you Elena Torrecilla for managing it perfectly for full three years. SINGEK was one of the greatest experiences of my PhD. Thanks to all the organizers, all instructors, thanks to all the workshops and meeting. They all added a lot to my professional and personal knowledge and experiences.
I want to thank the SINGEK ESRs. The other 14 talented PhD-mates. It was great sharing and doing science with you. Ina, your friendship is amazing.

Beside a great advisor, I had amazing labmates who cheered up almost 1000 days of my life. I want to thank Marcos Fernandez for showing me and trying to teach me how a real genuine, good mankind, can be. Thank you for listening to my daily updates, from family news and political situations in Iran to stupid love stories. For being a very good friend and help and for all the "Sobh Bekheir / Asr Bekheir"s of every day. I want to thank Elisabetta Mereu for three years of help, ideas, support, discussion, laughter and arguments. Without you my PhD would have been boring. Ramon Massoni, the encyclopedia, the cool, the PhD-mate thank you for all your help. Thank you Giovanni Iacono for all the scientific insights and the muffins we shared. Juanlu Trincado for being a part of us and not being a part of us. Paula Nieto for showing me how vegan life can be colorful and delicious, Marc Elosua for proving me there shouldn't be gender bias in color of socks, Gustavo Rodriguez to show how someone can be so serious and at the same time a great bachatero and Inma Hernandez for always encouraging me! I want to thank the chicas del wet-lab, Catia Moutinho, Domenica Marchese, Sara Ruiz, Patricia Lorden, Guilia Lunazzi and Amy Guillaumet Adkins for their amazing work to produce the great datasets for us.

Next comes my special thanks to my friends of the last three and half years. The amazing, brilliant people that I got to know and each added a lot of values and happiness to my life. I want to start with the very first, the very whole-hearted group of friends, my Italian 90's girls. You all melted my heart by opening your arms sincerely and embracing the Iranian from the first day. Maria Carla Antonelli, the angle on the earth, my flatmate and guardian in Barcelona, thank you very much for all your kindness and care. Having you in my life for the last years, was like having a second family far away from home. Jacqueline Severino, the Jackie, thank you for being such an active, spontaneous, open-hearted girl. Thank you for calling me the very first day that I arrived to Bcn and thank you for keeping being a good friend. Beatrice Borsari, your kindness has no limits. I appreciate deeply your friendship and genuinity and thank you for all the moments we shared and all the support you have provided. And Elena Tantardini, our blue-eyed beauty, I don't need words to express my gratitude, the fact that I couldn't resist without you so that I decided to move to Switzerland proves everything.

I would like to specially thank Sara Hajian, my one and only Iranian friend in Bcn. Your friendship meant a lot to me. More than a friend, you are like the sister that I've never had. Being truly a friend with an open mind, cheerful sprit, positive attitude and valuable experiences, you are one of the most important things that helped me grow.

Here again I should thank Sara to not only being a great friend but also bringing an amazing group of people to my life. Starting from Ilias, Amalia, Nikolas, Ioannis and my new-old best friend Minos, I want to thank the whole "Barcelona Crew" people. You guys truly cheered up my life and added lots of values to my challenging PhD journey. Knowing you each was a great pleasure. But among all these amazing people, it comes a special one, Antigoni Maria Founta, the Azizam. Thank you very much to become an important part and the most azizam of my life.

I want to also specially thank Nicola Barbieri and Yamile Marquez, true friends. I was so lucky to get to know you, and so fortunate to become your friend. Your friendship and care is so valuable for me.

The last but not the least, I want to thank the purest friends, the ultimate supports, loves and motivations for me to keep going and going, My family. Baba, you are the reason I am who I am today. You tried hard all your life to provide me the strong wings to fly where I want, no matter how far, no matter how hard. And here I am, flying freely over all the limits… Thank you for giving me the best gift of my life. Maman, thank you for being so patient of me, being far away, for my success. And finally, thank you Ali, you are a part of me. Thank you for being my brother, my closest. For being an amazing person with a big heart. Thank you for bringing a new member, lovely Parisa, to our family. Thank you all, you are my most precious belonging in life.

Abstract

Single-cell RNA sequencing has revolutionized the way molecular mechanisms were being studied by allowing the dissection of gene expression at single-cell resolution. The data acquired from scRNA-seq provides great opportunities for scientist to push the limits and go beyond technological boundaries to address biological questions. However, a thoroughly thought experimental design, protocol selection and data analysis strategies are necessary to get the best out of this high potential technology. In this thesis we start with summarizing current methodological and analytical options, and discuss their suitability for a range of research scenarios. We provide information about best practices in every step from separating cells and RNA library preparation to data generation, normalization and analysis. Next, we try to address a biological phenomenon using scRNA-seq. We demonstrate how a correctly designed scRNA-seq experiment and analysis is able to capture in details the process of dermal fibroblast aging. Observing the data produced by different scRNA-seq protocols, their important differences and the challenge to analyse them together, raised the question of their suitability specially in cell atlas projects. Hence, in a big multi-center systematic study we compared 13 commonly used single-cell and single-nucleus RNA-seq protocols using a highly heterogeneous reference sample resource. We pointed at their accuracy, application across distinct cell properties, potential to disclose tissue heterogeneity, reproducibility and integratability with other methods; features in which should be considered when defining guidelines and standards for international consortia, such as the Human Cell Atlas project. Finally, we propose an approach to elevate the data from poor-performing protocols to the quality of the best data coming from best-performing ones using variational autoencoders and vector arithmetic.

Table of Contents

General introduction

Cells and cell type heterogeneity

Cells are the fundamental units of life. They are divided into two main groups of prokaryotic cells that are usually found in single-cell organism, and eukaryotic cells that are found in multicellular organism. But no matter prokaryotic or eukaryotic, for a cell to survive, it should carry out the same basic functions such as acquisition of nutrients and energy sources, disposal of unusable and toxic materials, reproduction and interaction with the environment. Even though cells share fundamental functions, there are different types of cells for different organisms that are specialized to carry out specific sets of functions. These differences are sometimes easy to observe if they lead to phenotypical differences as well, while usually this is not the case. The reason for these differences among cells can range from genetic diversity to variability driven by stochastic molecular interactions, and noise induced cell differentiation [1], which, at a higher level will guarantee the survival and adaptation of the cells to different conditions.

If we focus on mammalian cells, we reach to the theory of stem cells and cell differentiation which causes cellular heterogeneity. Stem cells are the unspecialized, proliferating cells that give rise to specialized cell types in a process called differentiation. Upon differentiation, they go through several stages following some internal and external signals that trigger different steps of this differentiation process. But at the end, all these signals orchestrate the transcriptional activity of each differentiating cells, leading them towards a specific cell type which can be characterized by a snapshot of its transcriptional profile at each moment. Eventually cell type heterogeneity is mostly caused by different levels of genes being expressed in each cell.

Emergence of single cell RNA-seq and advantages compared to bulk

Through the development of RNA sequencing almost a decade ago, scientists were able to examine the quantity and sequence of RNA in a sample using next generation sequencing (NGS). This technique provided the researchers with a tool to get snapshots of transcriptional profiles of a sample. This technique is being used since then for transcriptional profiling, differential gene expression analysis, SNP identification, RNA editing and many other applications.

Early RNA-seq experiments were being performed on bulk [2], homogenized tissue or large population of cells where the extracted RNA represents an average or mixture of thousands to millions of individual cell transcriptomes. Although this technique is still being used and is very useful when overall differences between two or more samples are of interest, it hides the differences at a higher resolution.

Recognizing the degree of cell type heterogeneity in known tissues, tumours and many morphologically identical cell types, scientist started to think about studying individual cells one by one.

Another motivation to increase the resolution of RNA sequencing was the limitations that researchers were facing regarding the large amount of total RNA that is needed for performing bulk RNA-seq, specially in identifying new transcript variants and isoforms. Such experiments need microgram amounts of total RNA, which corresponds to hundreds of thousands of cells. Such high amounts of RNA were impossible to get in some conditions, like early embryonic samples [3], and more sensitive RNA-seq assay were needed, ideally capable of working at single-cell resolution.

Single cell RNA-seq the first techniques and approaches

Like many other technologies, the field of single-cell genomics started in a targeted way. In 1992, Eberwine *et al.* measured the expression of a handful of individual genes based on *in vivo* reverse transcription (RT) followed by amplification through IVT [4]. Two years later, Sheng *et al.* used an oligo(dT) primer with an attached phage promoter, to transcriptionally amplify the whole mRNA pool from a single cell, for the first time [5]. Eventually, untargeted single-cell amplification techniques were developed and used in detangling heterogeneity in cell population. One of the early studies using this approach was to uncover mechanism involved in neuronal differentiation and diversification by Tietjen *et al.* The authors developed an experimental protocol in which single-cell cDNA synthesis was combined with Gene chip analysis and laser-capture mediated cell isolation [6]. Kurimoto *et al.* later improved the coverage, accuracy and reproducibility of the method by directionally amplifying cDNAs from single cells using PCR , a method that was universally applicable to oligonucleotide microarrays [7]. Other studies tried to improve the amplification procedure to perform microarray analysis of single cells [8,9].

So far, all single-cell approaches were based on microarrays till the point that Tang *et al.* in 2009 adapted the technologies to make them compatible with high throughput RNA sequencing, thus allowing completely unbiased whole transcriptome investigation of single cells for the first time [3].

Evolution of single cell RNA-seq in the last decade

In 2009, Tang *et al.* modified the widely used single-cell whole transcriptome amplification method to generate 3 kilobases long cDNAs efficiently and without bias [3]. They showed that it is feasible to get digital gene expression profiles at single-cell resolution using the Applied Biosystems' NGS SOLiD system. Using this technique, they detected around 5K more genes from a single mouse blastomere compared to the microarray using hundreds of blastomeres. Following up on the new proposed technique, the same group utilized single cell RNA-Seq to trace derivation of embryonic stem cells from the inner cell mass [10].

Three years after the first whole transcriptome single cell RNA-Seq, Ramsköld *et al.* proposed Smart-Seq which was a robust mRNA-Seq protocol that was applicable at single-cell level. One of the most important advantage this protocol introduced was the improved read coverage across full length transcripts which enhanced SNP detection and alternative transcript isoform identification [11].

Early single-cell experiments were focused on in depth analysis of gene expression in a few number of precious cells. Guo *et al* came with a shift in the field by carrying out RT-qPCR of 48 genes in parallel on more than 500 cells, skipping pre-sorting step [12]. Even though this method significantly improved on number of profiled cells, it was still limited to few number of genes due to its dependency on q-PCR. Islam *et al*. addressed this limitation by introducing single-cell tagged reverse transcription (STRT), a highly multiplexed method for single cell RNA-seq [13]. In this technique, 96 cells were added to individual wells, and cell specific barcodes were added using the template-switching mechanism of reverse transcriptase during cDNA generation independently for each sample. The material from each well is then pooled before it is amplified by PCR. They demonstrated the feasibility of this strategy by performing whole transcriptome analysis of 85 single cells of two distinct types. This study eventually led to the first steps towards mouse brain atlas by the same group using unbiased RNA sequencing of 3005 single cells [14].

Single cell RNA-seq protocols

With STRT-seq the field of single-cell sequencing had changed greatly in terms of its potentials. This protocol served as a pilot for future studies where cells would be randomly sampled, without having the need to have a priori annotated cell types. Parallel to STRT-seq, Hashimshony *et al*. introduced CEL-seq which was similar to STRT-seq in barcoding cells in individual wells during cDNA synthesis and pooling, but instead of using PCR, the material was amplified by in vitro transcription (IVT) [15]. They showed that IVT gives more reproducible, linear and sensitive results than PCR-based amplification methods and they demonstrated the performance of their method by studying early C. elegans embryonic development.

The SMART-seq protocol was introduced in the same year, giving full length coverage of transcripts using SMART template switching technology. By taking advantage of Illumina’s Nextra XT kits (Illumina, Inc, 2012) which was allowing up to 192 samples per Illumina sequencing lane, for multiplexing and library generation, SMART-seq made single-cell RNA-Seq experiments widely accessible and bioinformatics processing easier.

Even though protocols such as CEL-seq and SMART-seq brought a lot of advances, none of these microwell based techniques processed more than dozens of cells while for a statistically powerful cell type and state identification we would need more samples. With introduction of Fluidigm C1 (IFC) system in 2013, which automatized cell capture from a cell suspension into 96 chambers of the IFC, and reducing reagent cost, number of profiled cells were tackled to reach hundreds for the first time. Fluidigm released a new version of IFC in 2016 which was allowing capture of up to 800 cells. Later Linnarsson group adopted their STRT-Seq protocol to C1 IFC and profiled 3005 cells which leaded to previously mentioned mouse neuronal subtype study by Zeisel *et al.* [14]

Since high price per cell was one of the most important limitations of SMART-seq protocol to profile large samples, Picelli *et al*. tried to address this limitation to some extent by reducing volumes in individual wells and using less expensive off the shelf reagents in the refined version of this protocol, SMARTseq2 [16]. Later the same group decreased the price per cell even more by in-house production of a variant of the Illumina Nextra kit's active enzyme, transposase Tn5 [17].

Amit group on the other hand introduced MARS-Seq which was a modified version of CEL-seq protocol to be compatible with robotic automation in massively parallel single cell RNA sequencing [18]. In this protocol Jaitin *et al*. scaled up the number of profiled cells to 4000 by decreasing the labour of processing plates being filled with isolated cells. After this method many protocols got updated being inspired by this automation like SORT-seq [19] from CEL-seq2 [20].

All these protocols demonstrated the fact that in case we improve the isolation of cells and the ability to generate enough multiplexing barcodes, the rest of the steps can be done in a single unit and we can scale up the throughput. In 2013, Mazutis *et al*. presented a detailed protocol for using droplet-based microfluidics for high-throughput isolation of individual, antibody- secreting cells from a large excess of nonsecreting cells [21]. Two years later, two studies presented protocols for using droplet-based microfluidics for high-throughput isolation of individual cells named inDrops [22] and Drop-seq [23]. After these microfluidics based techniques got stabilised in the fields, companies tried to commercialise the required materials. The company 10X Genomics (10X Genomics, Inc, 2016) and Illumina Bio-Rad (Illumina, Inc., 2017) was among the industrial leaders of this field. 10X Genomics commercialized inDrop and spread it all over the world. It provided the users the possibility to parallelise several experiments in single run by simultaneously processing up to 8 independent cell pools. The technology demonstrated its potentials in a massive study of 250,000 cells by Zheng *et al*. in 2017 [24].

Alternatively, beads can be deposited into picoliter wells and randomly be loaded with cells at limiting dilution [25]. The advantage of this approach is the decreased reaction volume and portability which makes it a good choice for situations where rapid collection of fresh cells is the case. Talking of portability, Shalek group also proposed a portable, low-cost single-cell RNA sequencing technique called Seq-Well, that seals the barcoded beads and single cells in an array of subnanoliter wells using a semipermeable membrane, enabling efficient cell lysis and transcript capture [26]. Also, to increase the sequencing read to UMI (Unique Molecular Identifier) count conversion ratio, Sasagawa *et al*. improved previously publish Quartz-Seq [27] protocol by changing the reaction steps which make the Quartz-Seq2 possible to effectively convert initial reads to UMI counts, at a rate of 30–50%, and detect more genes [28].

So far, all the discussed methods require cell sorters, custom microfluidics, or microwells which are considered as limitations for throughput. Two studies in 2017 proposed to use sequential in situ barcoding to profile single cells. In SPLiT-seq [29], individual transcriptomes are uniquely labeled by passing a suspension of formaldehyde-fixed cells or nuclei through four rounds of combinatorial barcoding, while sci-RNA-seq [30] follows a very similar approach with some small modifications and demonstrates the power of the technique by profiling 50,000 cells from C. elegans.

Even though these approaches had addressed many limitations in profiling the RNA content of individual cells, dissociation of individual cells from some cell types and tissues was still a big challenge. One of these challenging cell types has always been neurons. To address this difficulty in challenging tissues, Habib *et al*. develop Div-Seq, which combines scalable single-nucleus RNA-Seq (sNuc-Seq) with pulse labelling of proliferating cells by 5-ethynyl-2'-deoxyuridine (EdU) to profile individual dividing cells [31]. Meantime, Blue *et al*. proposed a scalable pipeline to sequence and quantify RNA molecules in isolated neuronal nuclei from a postmortem brain. With this technique they identified both known and previously unknown neuronal subtypes across the cerebral cortex in humans [32].

On the other hand, researchers had been conserving cells in cases where obtaining fresh samples was not possible while whether this preservation changes the transcriptional profile of the cells was an ongoing discussion. Guillaumet-Adkins *et al.* in 2017 proposed a preservation method and demonstrated that cryopreservation maintains cellular structures and integrity of RNA molecules for single cells months after archiving by analysing 1486 single-cell transcriptomes from fresh or cryopreserved cells from cell lines or primary tissues [33].

Comparison

Single-cell RNA sequencing protocols differ in many aspects; from the type and amount of information they provide to the costs per cell. Choosing the best protocol highly depends on the goal of the experiment as well as the available resources. We have published a comprehensive guide to choose the best experimental design depending on the research question in 2018 [34] that forms the first chapter of this thesis.

Comparing the amount of information profiled at read level, we can choose between full length transcripts profiles or digital counting of 3´or 5´transcript ends [35]. As mentioned before the decision should be made by prioritizing cost-effectiveness over retention of sequence information or the other way around. Even though full-length transcript profiling is relatively costly, it provides more features to study, such as splice variants detection and alternative transcripts as well as single-nucleotide variants [36] and fusion transcripts [37]. Moreover, genotypes of T and B cell receptors can be obtained from full-length transcriptomes [38]. On the other hand, 3´- and 5´-end methods allow introducing UMIs -which reduces the PCR amplification bias, and higher number of cells can be profiled, due to the lower costs.

The choice between the microtiter-plate-based vs microfluidic-based approaches depends as well on the expected throughput. Microfluidics allows higher throughput and eliminates the technical constrains on scalability associated with microliter plates. Moreover, it reduces the reaction volume to nanoliters which results in lower cost and technical variability. On the other hand, although plate-based techniques are limited in terms of throughput, they usually result in better quality libraries and higher resolution per cell.

Apart from throughput and library resolution, there are other parameters that differ between scRNA-seq protocols which one should consider when deciding experimental designss. Cell doublets, cell capture efficiency and cost are examples of those parameters. Cell doublets are known to be a common problem of microfluid-based techniques that is controllable to some extent by the cell suspension concentration. Some protocols show high cell capture efficiency microfluidics, mainly as a result of efficient cell and bead loading mechanics [34].

The total cost of scRNA-seq experiments is determined by three main components: equipment, reagents and sequencing. For most methods, the cost of scRNA-seq library preparation scales linearly with cell numbers; an exception are custom droplet methods. The actual costs per cell vary widely across methods and institutes, with microfluidic systems being generally cheaper (<$0.30 per cell) than early-indexing plate-based 3′ digital counting methods (~$1–2 per cell). Late-indexing full-length transcriptome profiling is costlier, even with small volumes (~$8–12 per cell). However, costs can be reduced through the use of non-commercial tagmentase [17] or minimum reaction volumes and automated workflows for plate-based formats [34].

As mentioned before, in the first chapter of this thesis we compare different aspects of single-cell RNA sequencing in a Tutorial published in Nature Protocols.

scRNA-seq at data level

Single-cell RNA sequencing opens new doors toward cellular exploration. Information about transcript content of individual cells dramatically increases the resolution of cellular data to be investigated. Though for this type of data to demonstrate its full potentials, a wise and careful chain of analytical approaches should be taken into consideration. We are going to discuss in details these challenges and methods to address them.

Data characteristics and problems

Single-cell RNA expression data is a matrix of 2 dimensions, rows as genes, columns as cells and each entry in the matrix defines the number of mapped UMIs/reads to corresponding gene in corresponding cell. Expanding this matrix to approximately 10,000 profiled cells to detect about 24,000 genes in human, creates a high dimensional matrix which brings in challenges to handle and extract information.

The first problem comes with the large number of genes assayed in scRNA-seq, that is, the high dimensionality, which causes distances between data points (cells) become similar. This problem is known as the curse of dimensionality [39] and forces differences in distances to be small and thus not reliable for identifying cell subpopulations.

The next and one of the most important challenges in dealing with scRNA-seq data that yet does not have a profound solution is the dropout rate. Owing to a low amount of RNA that is obtained initially from single cells (10-40% of total RNA within the cells) this data generally exhibits more zero values and higher levels of noise compared to bulk RNA-seq. The problem with this elevated amount of zero entries is to distinguish between the true zeros and the zeros resulting from technical pitfalls. A true zero means the transcript was not present in the cells and the zero is, thus, an accurate representation of the state of the cell. A technical zero would be result of not reporting the transcript due to not sequencing deep enough, even though it was available in the cell and in the sequencing library. Technical zero can also arise from problems in capturing and amplifying the transcripts in a library preparation step even though it was present in the original cell [40].

Batch effect is a known problem in genomic studies which refers to differences in data that are due to experimental factors like time and laboratory of the experiments, person performing the experiments or even the microliter plate or sequencing run. These challenges the analysis of the data when we are dealing with scRNA-seq data to explore differences within conditions. In this case, differentiating between biological differences or technical ones becomes a big challenge. The best strategy to avoid this problem comes at the initial steps of the experiment which is the design. By having a balanced experimental design which splits samples across technical batches, we can reduce batch effect in scRNA-seq to a good extent. Further, there are computational tools available to correct for this effect [41].

Other technical effects are doublets. Doublets or multiplets arise in scRNA-seq when two or more cells are mistakenly considered as a single cell and they also significantly complicates the analysis of the sample. Doublets arising from cells of two distinct cell types can be easily mistaken for rare transitional cells, as they will exhibit a phenotype that is intermediate between the two originating cell types [39]. This is more controllable in plate-based techniques that allow imaging of captured cells and identification of cell doublets before lysis.

High dimensionality of the data

Due to the high number of transcripts per cell and depending on the number of profiled cells, which with new technologies can scale up to hundreds of thousand, the dimensionality of scRNA-seq data can be very high. This high dimensional data, as any other big data in the field of data science, brings problems to handle, normalise and analyse. The first obvious challenge is usually the computational processing cost. The processing power and memory required to handle such dataset can pass the range of average personal computers.

Another problem that comes at the statistical level is the well known problem of "The Curse of Dimensionality". Even though this problem itself encompasses different domain of data science like sampling, optimization, machine learning, the common theme is by increase in dimensionality, the volume of the space increases so fast that the available data become sparse. This sparsity becomes a big problem for any method that requires statistical significance, since all objects appear to be sparse and dissimilar in many ways, which prevents common data organization strategies from being efficient. Eventually finding patterns and hidden structures within the data becomes difficult.

High Dimensional data is also challenging to visualise. The physical limitations of the displaying devices (2D/3D), and the relatively low capacity of our mind to process complex information at a time makes the visualization of high dimensional data difficult. Findings way to be able to summarize the high dimensional data into representable way is another open area of research.

Dimensionality Reduction

Visualization of high dimensional data in low dimensional space is an essential tool for exploratory data analysis. Not only for visualization, but also to be able to analyse properly the high dimensional data with statistical techniques, we need a low dimension representation of it. That is where Dimensionality Reduction (DR) techniques comes into the game.

Methods for DR provide to understand the hidden structure of high dimensional data. They are meant to reduce variance and redundancy, increase accuracy and recover intrinsic dimensions [42]. However, it is impossible to avoid information loss during this reduction process and the challenge is to develop techniques that preserves the maximum information and relationship from the original data [43,44]

In the following section, we will discuss the most common DR algorithms used in scRNA-seq data exploration.

Principle Component analysis (PCA)

PCA is one of the most commonly used DR algorithms. It is a statistical method to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principle components. For a given high dimensional dataset, PCA find vectors along which the data has maximum variance [45]. Generally, PCA transforms the data into a new coordinate system in such a way that the largest variance by any projection of the data is summarized in the first coordinate PC1, the second largest variance on the second coordinate (PC2) and so on [42]. So, the order of PC gives the information about the vector that explains highest variability to the vector that explains the lowest variance in the data. PCA is useful when the data lies on or close to linear subspace which is not the case in single cell RNA-seq data. Single-cell data has a highly non-linear structure which stems from the large fraction of stochastic zeros in the expression matrix due to the dropout effect.

In summary, even though linear dimensionality reduction techniques like PCA are valid at preserving the global structure of the data, for single-cell data it is important to keep the local structure, thus, linear DR techniques cannot fully resolve the heterogeneity in single-cell data.

T-distributed Stochastic Neighbor Embedding (t-SNE)

In 2008, Van de Maaten presented a DR technique called t-SNE that is able to more precisely reduce the dimensionality of data having nonlinear structure [46]. t-SNE starts by calculating a similarity matrix of the high dimensional data using euclidean distances. Later, it constructs a probability distribution over pairs of high dimensional objects in a way that similar objects would have a higher probability of being selected. In other words, dissimilar cells in the original high-dimensional space are modeled by large distances, and similar cells are modeled by small distances. Then, the technique defines a similar probability distribution over the points in the low dimensional space and minimizes the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the map (Kullback, 1987).

t-SNE captures better the local structure and works on raw or normalised expression matrices, however a pre-dimension reduction with PCA provides more distinct and condensed cluster and reduces the computational cost. This focus on capturing local similarity at the expense of global structure may exaggerate differences between cell populations and overlook potential connections between these populations [47]. An important consideration that should be taken into account while using t-SNE is the influence of its parameters on visual clusters. One of these most influential parameters of t-SNA is "perplexity", which is an estimate about the number of close neighbors each point has and is used to balance attention between local and global aspects of the data. Even though t-SNE is one of the most popular RD algorithms for scRNA-seq, favouring the preservation of local distances over global distances, as well as challenges in setting the correct parameters, opens the way for more robust and comprehensive method.

Uniform Manifold Approximation and Projection (UMAP)

So far, the DR algorithms that we discussed tend to fall into two categories; those that seek to preserve the distance structure within the data (PCA), and those that favour the preservation of local distances over global distance (t-SNE). McInnes *et al*. introduced UMAP in 2018 which is competitive with t-SNE for visualization quality and arguably preserves more of the global structure with superior run time performance [48].

UMAP is based on local manifold approximations and patches together their local fuzzy simplicial set representations to construct a topological representation of the high dimensional data. A similar process can be used to construct an equivalent topological representation given some low dimensional representation of the data. UMAP then optimizes the layout of the data representation in the low dimensional space, to minimize the cross-entropy between the two topological representations [48].

As a summary, UMAP is capable of scaling up to significantly larger datasets due to its topological foundations and tends to preserve better both local and global structure of the data.

Dissecting cellular heterogeneity

It has been proven in both experimental and theoretical level that all cellular systems are heterogeneous [49]. Single-cell RNA sequencing allows the quantitative and unbiased characterization of cellular heterogeneity by providing genome-wide molecular profiles from tens of thousands to millions of individual cells. We also discussed that scRNA-seq data is a high dimensional, challenging data. Hence, finding methods to address and unbiasedly discover these cell-to-cell variabilities and heterogeneity is an active field of research.

One of the challenges to study the heterogeneity within the data is to have a prior knowledge whether the data contains discrete population of cells so that clustering algorithm would be the right approach to address this heterogeneity. Alternatively, the data can describe a continues change in cell states along a differentiation or developmental process that needs to be addressed by trajectory-type analysis.

Clustering algorithms

In this section, we will talk about algorithms that can help to find subpopulation of cells in cases that there are defined clusters present in the data. Techniques in unsupervised learning is our main toolkit is this section. In this approach the first step is to define the similarity of expression matrix profiles using a distance metrics. One of the most common metrics used is Euclidean distance. Later, clusters are obtained by grouping cells based on the gene expression profile similarities.

K-mean

K-means clustering is one of the simplest unsupervised learning algorithms. This algorithm is based on defining initial K number of centroids (which later would be our final number of clusters), assigning them randomly in the data space and iteratively optimizing the position of these centroids. A centroid is the imaginary or real location representing the centre of the cluster. After the first randomly initialization of K random centroids, every data point is assigned to the nearest cluster by the ultimate goal of minimizing the distance between the points within a cluster. Once all points have been assigned to a cluster, the centroids are recalculated by averaging the point in each of the newly created clusters. The algorithm iterates till the point that centroids are stabilized and are not changing anymore.

The main challenge in using k-means is to predict the K value. Also, different initial partitions can result in different final clusters. But is known that this algorithm is very time efficient.

Hierarchical clustering

Like k-means, hierarchical clustering is another simple and popular clustering algorithm. It is also based on distance or similarity matrix. It has two ways of creating clusters: Agglomerative and Divisive.

In agglomerative technique, we ascent from smallest clusters which are every individual point to the biggest cluster which is the cluster encompassing all data point. In every step, two closest clusters are merged and the distance matrix is updated.

On the other hand, divisive hierarchical clustering is opposite of agglomerative in terms of direction. Divisive algorithm starts from the biggest cluster of all points as a single cluster, and in each iteration, it partitions each cluster to two least similar clusters. It iterates until there is one cluster for each observation.

Advantages of hierarchical clustering is the dendogram representation which is very useful in understanding the data. Also, this algorithm does not need a predefined number of cluster to create the dendogram tree even though at some point we need to cut the three at a specific level, which corresponds to defining the number of clusters. Being very time consuming on big dataset is one of its disadvantages.

Graph-based clustering

Graph-based methods or community detection algorithms are based on graph representation of distance matrices. To create the initial graph algorithms like K-Nearest Neighbor (KNN) or Shared Nearest Neighbor (SNN) is used. In KNN graphs two points p and q are connected by an edge, if the distance between p and q is among the k-th smallest distances from p to other points. In other words, each point is connected to its K most similar cells, while SNN defines proximity between two points in terms of the number of neighbors they have in common. In both cases, the result is a graph in which similar points (cells in single-cell data) will be connected to each other. Dense regions of the expression space are represented as densely connected regions of the graph [47] .Next, clustering can be achieved through partitioning the graph into homogeneous and well-separated subgraphs. The partitioning algorithm can be a challenge since some methods require a prior knowledge of the number of subsets to be produced; other methods can produce singletons for sparse graphs [50]. In 2015, Xu *et al*. proposed quasi-clique-based clustering, called SNN-Cliq, to identify tight groups of highly similar nodes that are likely to belong to the same genuine clusters. This technique is based on SNN graphs and is able to automatically determine the number of clusters, as well as identifying clusters with different shapes and densities [51]. Levine *et al*. also developed PhenoGraph, which is also based on nearest neighbor graphs and finding sets of highly interconnected nodes (community detection), borrowed algorithms from field of social networks. One of the main features of PhenoGraph is the ability to construct a graph that faithfully represents the phenotypic relationships between cells. It does so by implementing the graph in two iterations, using the Jaccard similarity coefficient in the second iteration. In other words, it refines the edge weights between any two cells based on the shared overlap in their local neighborhoods (Jaccard similarity). This trick exploits the local density at each data point by removing spurious edges and strengthening well-supported ones [52]. To reach to the final clusters from these refined graphs, modularity optimization techniques are applied. Modularity measures the density of edges inside clusters to the edges outside of the clusters and optimizing this metric would result in the best possible grouping of the nodes of a given network. One of the well-known algorithms for modularity optimization are Louvain [53]. In this technique first, small communities are found by optimizing modularity locally on all nodes, then each small community is grouped into one node and the first step is repeated.

Currently, graph-based clustering techniques are widely used in the field since some of the mostly used scRNA-seq analysis packages like Seurat [54] or Scanpy [55] are based on this algorithm.

There have been also publications regarding performance and comparison of algorithms and tools available for scRNA-seq clustering that can provide insights and guidelines to choose the best algorithm [56,57].

Trajectory Analysis

In specific biological scenarios researchers are interested to follow gene expression of cells through a differentiation process where clustering algorithms will not be the best analytical approach, since we are not expecting cells in defined and discrete clusters. Trajectory or pseudotime inference approaches computationally infer the order of cells along developmental processes using the gene expression data from single cells.

Trajectroeis based on Minimum spanning Trees

Inspired by computational geometry approaches to order bulk cell populations from time-series microarray data along a biological process [58], Trapnell *et al.* proposed an unsupervised algorithm called "Monocle", to order single cells in pseudotime, a quantitative measure of progress through a biological process. Monocle first represents the expression profile in a high-dimensional Euclidean space, with one dimension for each gene and a point for each cell in these dimensions. Then, using independent component analysis (ICA) [59], it reduces the dimensionality of this space. By constructing a minimum spanning tree (MST) on the cells, the algorithm tries to find the longest path, which corresponds to the longest sequence of transcriptionally similar cells. Later, in order to not only model the main differentiation path, but also capture the diverged cells along two or more separate paths, Monocle examines the divergent cells to find alternative trajectories through the MST. It orders these subtrajectories, connects them to the main trajectory and annotates each cell with both a trajectory and a pseudotime value [60].

One of the limitations of the early Monocle [60] is that the tree space is highly complex due to the large number of cells and this can lead to high variability and low stability. Hence, researchers tried to improve this limitation by reducing the complexity of the tree space via approaches like clustering the similar cells. Ji *et al.* published a tool called TSCAN that first clusters similar cells together and then constructs a tree to connect the cluster centers, recovering the true pseudotime of the differentiation [61].

In 2017, the Trapnell group releases the second version of Monocle (Monocle 2) that uses Reverse Graph Embedding (RGE) to learn a parsimonious graph by finding a mapping between the high dimensional gene expression space and a much lower dimensional space, while simultaneously learning the structure of the graph in this reduced space. DDRTree, a scalable RGE algorithm plays the major role in this new version of Monocle to learn the principle tree and enables Monocle to identify branch points that describe significant divergences in cellular states [62].

Slingshot is another well performing trajectory construction tools that can be categorized under tree-based techniques. The first step in Slingshot is similar to TSCAN which is using cluster-based MST to identify the key elements of the global lineage structure. It then uses a method called simultaneous principal curves to fit smooth branching curves to these lineages, thereby translating the knowledge of global lineage structure into stable estimates of the underlying cell level pseudotime variable for each lineage [63].

Trajectories based on graphs

A trajectory construction algorithm based on graphs was proposed in 2014 the Pe'er's group. They developed Wanderlust, a graph-based trajectory approach that receives single-cell data as input and maps it into a one-dimensional developmental path. Wanderlust transforms the data into an ensemble of graphs and selects random waypoints. Each graph is independently analysed where a user-defined starting cell is used to calculate an orientation trajectory. The orientation trajectory is iteratively refined using the waypoint cells and finally the trajectory is an average over all those graphs [64].
A main disadvantage of this algorithms is that it creates pseudo-temporal ordering of cells only if the data comprise a single branch.

Trajectories based on Diffusion maps

Haghverdi *et al.* criticised the use of linear dimensionality reduction techniques like ICA [60], or graph-based techniques [64] to create the hidden temporal order of developmental stages and proposed to use diffusion maps, a previously published tool for harmonic analysis and structure definition of data [65]. They claimed that the distance metric used in diffusion maps is conceptually relevant to the real biological differentiation data, as cells follow noisy diffusion-like dynamics while taking several differentiation lineage paths [66]. Also, diffusion maps preserve the non-linear structure of the data while being robust to noise. However, the tools are mostly for visualisation without ordering cells in pseudotime.

The same group later, extended on the idea and developed Diffusion pseudotime (DPT), which was able to derive a measure from diffusion maps to recover developmental trajectories from single-cell data [67]. DPT uses a weighted k nearest neighbors (kNN) graph on cells and calculates distances using transition probabilities over random walks of arbitrary length.

Trajectories based on mathematical models

Gaussian processes (GP) are Bayesian models that are well suited to model expression profiles and capture the uncertainty inherent in noisy data. Bayesian inference in GPs can be performed analytically and provides posterior mean estimates with a full covariance structure. A GP is parameterized by a mean and a covariance function [68]. These model-based approaches have been also used to construct trajectories.

Lonnberg *et al.*, used a Gaussian process latent variable model (GPLVM) and overlapping mixtures of Gaussian processes (OMGP) to infer trajectories and pseudotimes [69]. Reid *et al.* also propose a principled probabilistic model with a Bayesian inference scheme to impose a priori structure on the latent space. In this model, the latent space is one-dimensional and the imposed structure on the space relates it to the temporal information of the cell capture times. to represents the pseudotime [68]. Later, Campbell *et al.* added factor analyzers into the equation and proposed the first generative, fully probabilistic model based on a Bayesian hierarchical mixture of factor analyzers to infer pseudotime [70].

Optimal transfer

Recently, Schiebinger *et al.*, have proposed a new approach to use the mathematical model of Optimal Transfer (OT) to infer cell trajectories. In this idea, graph-based algorithms impose strong constraints on the model, such as one dimensional trajectories (“edges”) and zero-dimensional branch points (“nodes”). In their proposed model, termed Waddington-OT, they modified the classical algorithm of OT, which was originally developed to redistribute earth for the purpose of building fortifications with minimal work [71], to accommodate cell growth and death. Using this algorithm, they calculated couplings between consecutive time points and then infer couplings over long time intervals by composing the transport maps between every pair of consecutive intermediate time point [72]. In summary, this algorithm can perform OT on scRNA-seq data from a time course and find ancestors, descendants and trajectories, infer regulatory models that drive the temporal dynamics, visualize the cells in 2D using force-directed layout embedding (FLE) [73] and annotate cells by type, expression, trajectories.

RNA velocity

The Linnarsson group in collaboration with the Kharchenko group, developed an algorithm that could reveal the rate and direction of changes of the entire transcriptome in scRNA-seq studies during dynamic processes based on the relative abundance of unspliced and spliced mRNA. To do so, they assumed a simple model for transcriptional dynamics in which the first time derivative of the spliced mRNA abundance (RNA velocity) is determined by the balance between production of spliced mRNA from unspliced mRNA, and the mRNA degradation. They demonstrated the ability of RNA velocity to predict transcriptional dynamics by analysing mouse chromaffin cells [74] and recapitulating the transcriptional dynamics within this dataset. They showed general movement of the differentiating cells towards a chromaffin fate, as well as the movement towards and away from the intermediate differentiation state [75].

Machine Learning

In this final section of introduction, I will introduce the machine learning algorithms that has been used in different chapters of this thesis.

Linear Regression

Linear Regression (or linear models) forms the basis of machine learning. It seeks to find the relationship between the variables that we measure (independent variables) with the variable we are interested to predict (dependent variable). The algorithm defines this

relationship by obtaining a line that best fits the data. Using gradient descent algorithm, the best fit line is calculated in which total prediction error are as small as possible.
The two important pieces of information that can be extracted from linear models are i) the effect size of the covariates in the model and the residuals with an effect on the dependent variable that cannot be explained by covariates included in our linear regression.

This concept is very much used in single-cell data analysis to correct for unwanted sources of variation [76] or for finding differentially expressed genes while controlling for technical covariates [77].

Artificial Neural Network

Artificial neural networks, inspired by biological nervous system, are an information processing network. It is composed of large number of highly interconnected nodes to understand especially non-linear relationships between independent variables. Nodes are computational units they get activated in presence of enough stimuli. Nodes combine data input with weights and coefficients to assign importance by either amplifying or dampening inputs. The input-weighted products are summed up and the output is passed through the node's activation function, to determine whether and to what extent signals should progress further through the network to affect the outcome.

Deep learning networks are distinguished from the normal neural networks by their depth; that is the number of node layers through which data passes in a multistep process. A node layer is a row of nodes that activates or inactivates as the input is fed through the network. In another words, deep neural networks work as a sequence of multiple linear regressions. For each node of a single layer, inputs of the previous layer is recombined with inputs from every other node. Hence, the inputs are mixed in different proportions according to their coefficients, so that the network tests for significant combinations of input to reduce errors.

There is a great flexibility in the topology and layer structure of neural networks that enables development of varieties, each with unique strengths and potentials. In this thesis, I have used a specific kind of neural networks called Autoencoders, which have the purpose to reconstruct its own input.

Autoencoders

Autoencoders are categorised as unsupervised learning networks to reconstruct inputs. They apply backpropagation, setting the target value to be equal to the input. The network consists of 3 layers: An input layer, a hidden layer and an output layer, as well as 2 parts: an encoder function h= f(x) that encodes the data into a latent space with lower dimensionality, and a decoder that produces a reconstruction r= g(h) from the encoded data. The middle layer in this setting is called bottleneck layer, which contains a smaller number of nodes and represents a compression of input data. In this setting the idea is not to set g(f(x)) = x, but to design it in a way that it cannot learn to copy perfectly. Auoencoders are restricted to allow copying only approximately and only use input that resembles the training data. At each iteration that feeds the autoencoder with data, we compare the decoded output with the initial data and backpropagate the error through the network to update the weights of the network.

So in summary, autoencoders are neural network architectures that impose a bottleneck in the network to force a compressed knowledge representation of the original input. If the input data contains uncorrelated and independent features, the encoding and reconstruction would be a difficult task while if the data is structured and correlated, this structure can be learned and leveraged when forcing the input through the network's bottleneck.

Variational Autoencoders (VAE)

As we discussed in previous section autoencoders accept the input, compress it and try to recreate the input from the compressed representation with minimum loss. In this setting, we cannot generate data since the regularity of the latent space strongly depends on the distribution of the data in the initial space, the dimension of the latent space and the architecture of the encoder. In other words, the high degree of freedom of the autoencoder that enables to encode and decode with minimum loss, results in possible severe overfitting in the case new data is generated as some points of the latent space may give meaningless content once decoded. This problem can be solve with correctly organizing the latent space and making it more regular. This expected regularity of the latent space that makes generative processes possible can be defined in two main characteristics: i) Continuity; two close points in the latent space should give consistent outputs once decoded and ii) Completeness; a point sampled from the latent space should give meaningful content once decoded. In order to imply this regularization of the latent space, there is a slight modification of the encoding and decoding processes. In this modification, we encode an input as a distribution over the latent space rather than encoding it as a single point. This encoded distributions are usually chosen to be normal, but they can be selected based on specific parameters and requirements of the model. We apply this regularization in the loss function of the autoencoder as regularization term and it is expressed as the Kulback-Leibler divergence between the returned distribution and the prior distribution that exists over the latent space.

In summary, in a VAE: *First*, the input is encoded as distribution over the latent space. *Second*, a point from the latent space is sampled from that distribution. *Third,* the sampled point is decoded and the reconstruction error can be computed. *Finally*, the reconstruction error is backpropagated through the network.

Through this process, we are able to generate new data by sampling point from the regularized latent space and decoding it through our VAE.

Objectives

Single-cell RNA sequencing (scRNA-seq) is the leading technique for charting the molecular properties of individual cells. The latest protocols are scalable to thousands of cells, enabling in-depth characterization of sample composition without prior knowledge. However, there are important differences between scRNA-seq techniques, and it remains unclear which are the most suitable protocols to address specific biological questions. Within my PhD work, I gained a comprehensive knowledge of single-cell studies from experiments to data generation and analysis as well as best practices to get high quality data. These experiences are summarized and published in Nature Protocols (Lafzi *et al*., 2018) [34] and build the first chapter of my thesis. After my initial training, I worked on collaborative single-cell projects to apply the acquired knowledge on biomedical problems. These collaborative studies gave me a great opportunity to improve my skills in single-cell data analysis and resulted in interesting scientific discoveries. In chapter II of this thesis, I am presenting one of these studies dealing with single-cell analysis of dermal fibroblasts in aging, a work published in Cell (Salzer *et al*., 2018) [78]. Chapter III of this thesis focuses on my main PhD project entitled “Benchmarking Single-Cell RNA Sequencing Protocols for Cell Atlas Projects”. In this project, which is a part of Human Cell Atlas consortium, we proposed to perform a multi-center benchmarking exercise for the systematic evaluation of scRNA-seq methods. A variety of scRNA-seq methods have been developed and proven their utility in single-cell transcriptome analysis of complex and dynamic tissues. Besides technical differences, the methods vary in their efficiency in molecule capture. The resulting difference in library complexity is directly associated with the sensitivity of identifying transcripts and genes. However, how this impacts on the resolution of cellular phenotyping had not been systematically evaluated. The Human Cell Atlas project seeks to comprehensively chart cellular compositions of complex human tissues. Herein, a critical evaluation of protocols was a crucial prerequisite to inform the methodology selection process. In this project, we benchmarked current scRNA-seq protocols to inform the methodology selection process of cell atlas projects by pointing at their accuracy, application across distinct cell properties, potential to disclose tissue heterogeneity, reproducibility and integratability with other methods. This project is *accepted* in Nature Biotechnology and a preprint is available at bioRxiv (Mereu & Lafzi *et al*., 2019). Finally, I developed a method (chapter IV) that allows to transfer data from low quality protocols to the quality of top performing methods. I used autoencoders, a type of artificial neural network, to encode the data into latent spaces and calculate the transformation vector between high and low-quality data points using vector arithmetic. This manuscript is *under preparation* and will be submitted to Bioinformatics Journal.

In summary, the main objectives of this thesis are:

- Guidelines for experimental design of scRNA-seq studies
- Single-cell study of dermal fibroblasts in aging, as a case study
- Benchmarking single-cell RNA-seq protocols
- Improving the data quality of the low performing protocols

Impact and authorship report of the publications

This thesis dissertation is a collection of four scientific work for which Atefeh Lafzi has been contributed significantly during her doctoral studies. The first two manuscripts have been already published in peer-reviewed high impact journals (*Nature Protocols* and *Cell*). The third paper, which was the main project of her PhD studies, has been accepted in *Nature Biotechnology*, and is waiting the journal editorial processes for publication. A pre-print version is already available at *BioRxiv*. With these three papers, Atefeh fulfils the requirements of submitting a paper-based thesis dissertation, as she is a co-first author in the *Nature Biotechnology* and *Nature Protocols* works and second author in the *Cell* paper. In addition, she has prepared a forth paper, which is about to be submitted to high impact, peer-reviewed journal. The specific contributions of Atefeh Lafzi to each publication are indicated in the following sections, together with the 5-year impact factor of the journal, as reported by the Nature[1] and Cell[2] Publishing Groups' journal metrics. Individual author contributions are also available within each published article.

Holger Heyn

[1] https://www.nature.com/nature-research/about/journal-metrics

[2] https://www.cell.com/impact

Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies

Atefeh Lafzi*, Catia Moutinho*, Simone Picelli, **Holger Heyn**

*co-first authors

- Published in Nature Protocols, November 2018
- 5-year impact factor: 15.086
- **URL**: doi.org/10.1038/s41596-018-0073-y
- **Author's contribution**: This manuscript is a review article summarizing key steps in the experimental design for single-cell experiments. Atefeh was involved in the design of the review and contributed to the writing of most sections. Especially, she wrote the part related to the data processing, analysis and interpretation.

Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging

Marion Claudia Salzer, **Atefeh Lafzi**, Antoni Berenguer-Llergo, Catrin Youssif, Andres Castellanos, Guiomar Solanas, Francisca Oliveira Peixoto, Camille Stephan-Otto Attolini, Neus Prats, Monica Aguilera, Juan Martin-Caballero, **Holger Heyn** and Salvador Aznar Benitah

- Published in Cell, November 2018
- 5-year impact factor: 36.430
- **URL**: https://doi.org/10.1016/j.cell.2018.10.012
- **Author's contribution**: This original article included single-cell RNA sequencing experiments to determine the phenotypic changes of fibroblast in aging. Atefeh designed the experiments, analyzed all single-cell data and led the interpretation of the results. This work was performed in collaboration and the experimental mouse work is part of Marion Salzer's PhD thesis. The contributions of Atefeh and Marion were non-overlapping.

Benchmarking Single-Cell RNA Sequencing Protocols for Cell Atlas Projects

Elisabetta Mereu*, **Atefeh Lafzi***, Catia Moutinho, Christoph Ziegenhain, Davis J. MacCarthy, Adrian Alvarez, Eduard Batlle, Sagar, Dominic Grün, Julia K. Lau, Stéphane Boutet, Chad Sanada, Aik Ooi, Robert C. Jones, Kelly Kaihara, Chris Brampton, Yasha Talaga, Yohei Sasagawa, Kaori Tanaka, Tetsutaro Hayashi, Itoshi Nikaido, Cornelius Fischer, Sascha Sauer, Timo Trefzer, Christian Conrad, Xian Adiconis, L an T. Nguyen, Aviv Regev, Joshua Z. Levin, Aleksandar Janjic, Lucas E. Wange, Johannes W. Bagnoli, Swati Parekh, Wolfgang Enard, Marta Gut, Rickard Sandberg, Ivo Gut, Oliver Stegle, **Holger Heyn**

*co-first authors

- Accepted in Nature Biotechnology, January 2020
- 5-year impact factor: 45.117
- **URL**: https://doi.org/10.1101/630087, (BioRxiv)
- **Author's contribution**: This original article is a multi-center benchmarking effort to evaluate scRNA-seq protocols. Atefeh co-led the entire analysis from primary data processing to interpretation. Specifically, she merged the vastly different data types to a common format that presented the input for a systematic comparative work. Atefeh tested the 13 protocols using numerous quality control metrics that were summarized into a scoring system to rank methods for their suitability to process complex samples.

scAutoTransfer: Improving the scRNA-seq data quality by learning from high-quality datasets

Atefeh Lafzi, Fabian J. Theis, **Holger Heyn**

- In preparation for submission to Bioinformatics, January 2020
- 5-year impact factor: 8.136
- **Author's contribution**: This original article establishes a unique application for autoencoders to improve scRNA-seq data quality. Atefeh conceived this work during her internship in the Theis lab. She designed all experiments and conducted the analysis and interpretation.

Publications

Chapter I

Guidelines for experimental design of scRNA-seq studies

Single cell RNA-seq is at the forefront of high-resolution phenotyping experiments for complex samples. Although this methodology requires specialized equipment and expertise, it is now widely applied in research. However, it is challenging to create broadly applicable experimental designs, because each experiment requires the user to make informed decisions about sample preparation, RNA sequencing and data analysis. To facilitate this decision-making process, we summarize current methodological and analytical options, and discuss their suitability for a range of research scenarios. Specifically, we provide information about best practices for the separation of individual cells and provide an overview of current single-cell capture methods at different cellular resolutions and scales. Methods for the preparation of RNA sequencing libraries vary profoundly across applications, and we discuss features important for an informed selection process. An erroneous or biased analysis can lead to misinterpretations or obscure biologically important information. We provide a guide to the major data processing steps and options for meaningful data interpretation. These guidelines will serve as a reference to support users in building a single-cell experimental framework—from sample preparation to data interpretation—that is tailored to the underlying research context. This project is published in Nature Protocols [34]

nature protocols

REVIEW ARTICLE

https://doi.org/10.1038/s41596-018-0073-y

Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies

Atefeh Lafzi[1,5], Catia Moutinho[1,5], Simone Picelli[2,4], Holger Heyn[1,3*]

Single-cell RNA sequencing is at the forefront of high-resolution phenotyping experiments for complex samples. Although this methodology requires specialized equipment and expertise, it is now widely applied in research. However, it is challenging to create broadly applicable experimental designs because each experiment requires the user to make informed decisions about sample preparation, RNA sequencing and data analysis. To facilitate this decision-making process, in this tutorial we summarize current methodological and analytical options, and discuss their suitability for a range of research scenarios. Specifically, we provide information about best practices for the separation of individual cells and provide an overview of current single-cell capture methods at different cellular resolutions and scales. Methods for the preparation of RNA sequencing libraries vary profoundly across applications, and we discuss features important for an informed selection process. An erroneous or biased analysis can lead to misinterpretations or obscure biologically important information. We provide a guide to the major data processing steps and options for meaningful data interpretation. These guidelines will serve as a reference to support users in building a single-cell experimental framework —from sample preparation to data interpretation—that is tailored to the underlying research context.

Single-cell transcriptomics studies have markedly improved our understanding of the complexity of tissues, organs and organisms[1]. Gene-expression profiling in individual cells has revealed an unprecedented variety of cell types and subpopulations that were invisible with traditional experimental techniques. As well as providing profound insights into cell composition, single-cell studies have changed established paradigms regarding cell plasticity in dynamic processes such as development[2] and differentiation[3]. Cell states are now known to be more flexible than previously thought, and present multipotent characteristics before reaching fate-decision endpoints. Although various approaches are available for phenotyping of individual cells (e.g., transcriptomics[4], proteomics[5] and epigenomics[6]), single-cell RNA sequencing (scRNA-seq) is currently at the forefront, facilitating ever-larger-scale experiments. The scalability of scRNA-seq experiments has advanced rapidly through the use of automation and sophisticated microfluidics systems, producing datasets from more than 1 million cells[7]. As a result, experimental designs have shifted from a focus on specific cell types to unbiased analysis of entire organs[8–10] and organisms[11,12], thereby enabling a hypothesis-free approach to exploration of the cellular composition of a sample.

Most scRNA-seq methods are now broadly applied in both basic research and clinically translational contexts, even though they require specialized equipment and expertise in sample handling, sequencing-library preparation and data analysis. As a result, single-cell research has become one of the fastest-growing fields in life science, producing fascinating new insights into tissue composition and dynamic biological processes. Large-scale scRNA-seq experiments have yielded cellular maps of *Caenorhabditis elegans*[12], the planarian *Schmidtea mediterranea*[13], *Drosophila*[11,14] and different mouse organs[8,15] to be defined. In humans, single-cell analysis has improved understanding of developmental processes[16], aging[17] and different diseases such as cancer[18–21]. However, it is challenging to create generalizable designs for single-cell transcriptomic experiments because each one requires the user to make informed decisions in order to obtain interpretable results. These include the selection of sample types, cell numbers and preparation methods; the choice of scRNA-seq techniques and sequencing parameters; and the design of computational analysis strategies to generate insights from single-cell datasets. Ultimately, successful single-cell transcriptomic studies with interpretable datasets and meaningful scientific output can be achieved only through the use of tailored experimental designs. To inform this decision-making process, in this tutorial we provide a comprehensive description of the phases of single-cell transcriptomic studies, including (1) sample preparation, (2) scRNA-seq, (3) data processing and (4) data analysis (as discussed further below; see Fig. 1). We summarize the methodological and analytical options and highlight their suitability for distinct research scenarios to support users in designing an end-to-end experimental framework tailored to the underlying

[1]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [2]Research Institute for Neurodegenerative Diseases (DZNE), Bonn, Germany. [3]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [4]Present address: Institute of Molecular and Clinical Ophthalmology Basel (IOB) Basel, Switzerland . [5]These authors contributed equally: Atefeh Lafzi, Catia Moutinho.
*e-mail: holger.heyn@cnag.crg.eu

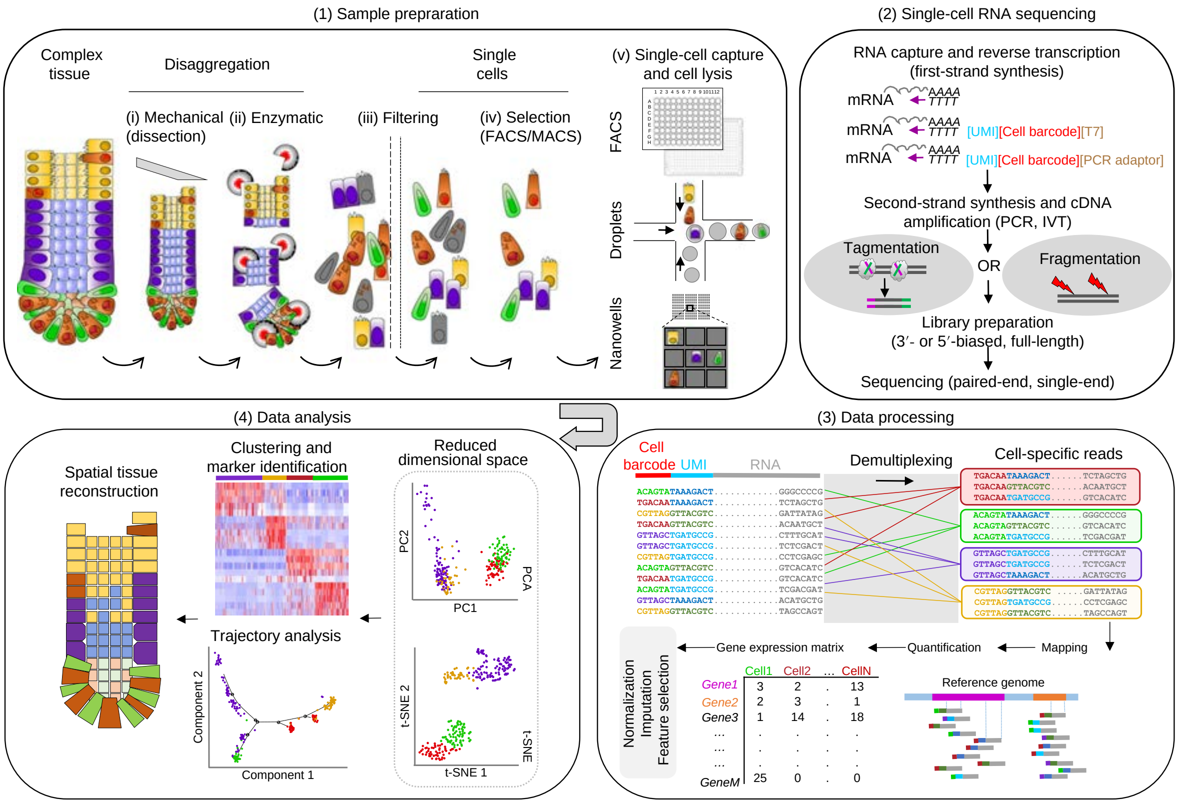

(1) Sample prepraration
Complex tissue
Disaggregation
Single cells
(i) Mechanical (dissection)
(ii) Enzymatic
(iii) Filtering
(iv) Selection (FACS/MACS)
(v) Single-cell capture and cell lysis
FACS
Droplets
Nanowells
(2) Single-cell RNA sequencing
RNA capture and reverse transcription (first-strand synthesis)
mRNA AAAA TTTT
mRNA AAAA TTTT [UMI][Cell barcode][T7]
mRNA AAAA TTTT [UMI][Cell barcode][PCR adaptor]
Second-strand synthesis and cDNA amplification (PCR, IVT)
Tagmentation
OR
Fragmentation
Library preparation (3′- or 5′-biased, full-length)
Sequencing (paired-end, single-end)
(3) Data processing
Cell barcode UMI RNA
ACAGTATAAAGACT...........GGGCCCCG
TGACAATAAAGACT...........TCTAGCTG
CGTTAGGTTACGTC...........GATTATAG
TGACAAGTTACGTC...........ACAATGCT
GTTAGCTGATGCCG...........CTTTGCAT
GTTAGCTGATGCCG...........TCTCGACT
CGTTAGTGATGCCG...........CCTCGAGC
ACAGTAGTTACGTC...........GTCACATC
TGACAATGATGCCG...........GTCACATC
ACAGTATGATGCCG...........TCGACGAT
GTTAGCTAAAGACT...........ACATGCTG
CGTTAGGTTACGTC...........TAGCCAGT
Demultiplexing
Cell-specific reads
TGACAATAAAGACT......TCTAGCTG
TGACAAGTTACGTC......ACAATGCT
TGACAATGATGCCG......GTCACATC
ACAGTATAAAGACT......GGGCCCCG
ACAGTAGTTACGTC......GTCACATC
ACAGTATGATGCCG......TCGACGAT
GTTAGCTGATGCCG......CTTTGCAT
GTTAGCTGATGCCG......TCTCGACT
GTTAGCTAAAGACT......ACATGCTG
CGTTAGGTTACGTC .....GATTATAG
CGTTAGTGATGCCG......CCTCGAGC
CGTTAGGTTACGTC......TAGCCAGT
Gene expression matrix
Quantification
Mapping
Normalization
Imputation
Feature selection
Cell1 Cell2 ... CellN
Gene1 3 2 . 13
Gene2 2 3 . 1
Gene3 1 14 . 18
...
GeneM 25 0 . 0
Reference genome
(4) Data analysis
Spatial tissue reconstruction
Clustering and marker identification
Reduced dimensional space
PC2
PC1
PCA
Trajectory analysis
Component 2
Component 1
t-SNE 2
t-SNE 1
t-SNE

Fig. 1 | The single-cell RNA sequencing process. The successful design of single-cell transcriptomics experiments includes four major phases: (1) During sample preparation, cells are physically separated into a single-cell solution from which specific cell types can be enriched or excluded (optional). After they have been captured in wells or droplets, single cells are lysed, and the RNA is released for subsequent processing. (2) To convert RNA into sequencing-ready libraries, poly(A)-tailed RNA molecules are captured on poly(T) oligonucleotides that can contain unique molecular identifier (UMI) sequences and single-cell-specific barcodes (5′- and 3′-biased methods). To allow for subsequent amplification of the RNA by PCR or IVT, adaptors or T7 polymerase promoter sequences, respectively, are included in the oligonucleotides. After RT into cDNA and second-strand synthesis (optional), the transcriptome is amplified (PCR or IVT). For conversion into sequencing libraries, the amplicons are fragmented by enzymatic (e.g., tagmentation) or mechanical (e.g., ultrasound) forces. Sequencing adaptors are attached during a final amplification step. Full-length sequencing can be carried out, or 5′ or 3′ transcript ends can be selected for sequencing using specific amplification primers (optional). For most applications, paired-end sequencing is required. (3) The sequencing reads are demultiplexed on the basis of cell-specific barcodes and mapped to the respective reference genome. UMI sequences are used for the digital counting of RNA molecules and for correction of amplification biases. The resulting gene-expression quantification matrix can subsequently be normalized, and missing values imputed, before informative genes are extracted for the analysis. (4) Dimensional-reduction representations guide the estimation of sample heterogeneity and the data interpretation. Data analysis can then be tailored to the underlying dataset, which allows cells to be clustered into potential cell types and states, or ordered along a predicted trajectory in pseudotime. Eventually, the spatial cellular organization can be reconstructed through the interrogation of marker genes (experimentally) or through marker-guided computational reconstruction (inference). PC, principal component.

research context (a glossary of relevant terms is provided in Table 1).

Sample preparation

Preparation of high-quality single-cell suspensions is key to successful single-cell studies. Irrespective of the starting material, the condition of the cells is critical for efficient cell capture and optimal performance of the scRNA-seq protocols. Although most methods use fresh viable single cells, alternatives include preserved samples[22–24] and nuclear RNA from frozen tissue[25–29]. Here we provide common general guidelines applicable to all tissues, and optimized parameters tailored to the major tissues of interest. In principle, scRNA-seq applications are not restricted to specific species as long as poly(A)-tailed RNA is present. However, some organisms might require additional processing steps to efficiently release molecules into the reactions (e.g., cell wall removal for plant material).

Good practices for sterile sample handling are recommended, including the use of nuclease-free reagents and consumables. To minimize cell damage, pipetting and centrifugation should be kept to a minimum. Cell concentration and size both influence pelleting efficiency at a given centrifugation speed, time and temperature, and a tightly packed cell pellet may require extra pipetting, which can damage cells

Table 1 | Glossary

Term	Definition
Algorithm	A process or set of rules to be followed in computational calculations or other problem-solving operations.
Barcode	A stretch of sequence used to uniquely label DNA/RNA molecules, cells or sequencing libraries (to allow multiplexing).
Batch effect	A technical source of variation added during sample handling.
Benchmark	Systematic comparison of different techniques (experimental or computational) for their performance in a given scenario.
Binary classifier	A classification function that predicts the assignment of an element to a set of groups.
Bulk RNA sequencing	The sequencing of RNA isolated from pools of cells.
Cell barcode	A cell-specific unique sequence tag that is added to RNA transcripts during library preparation.
Cell capture	Positioning of single cells in reaction volumes (e.g., droplets or wells) for downstream processing.
Cluster annotation	Assigning a function or identity to a group of cells on the basis of the expression of marker genes.
Clustering	The task of grouping cells in such a way that cells in the same group (cluster) are more similar to each other than to cells of another group.
Combinatorial barcoding	The use of combinations of cell barcodes with repeated assignment of barcodes to cells during multiple indexing rounds.
Deconvolution	A process of resolving a complex mixture (e.g., tissue) into its constituent elements (e.g., underlying cell types).
Demultiplexing	The process of separating the elements of interest in a mixed or multiplexed sample.
Digital counting	The counting of RNA molecules using UMI sequences.
Doublets	Two cells that are processed together in a reaction volume (e.g., a well or droplet) and receive the same single-cell barcode.
Dropout events	Transcripts that are not detected in the final dataset even though the gene is expressed in the cell, leading to false zero values in the expression matrix.
FASTQ reads	A sequence composed of the four nucleotides (ACGT) obtained after sequencing in a specific format that represents the chain of nucleotides.
Gene expression matrix	A data matrix containing information about the level of gene expression per cell.
Imputation	The process of replacing missing data with inferred values.
Index sorting	The isolation of single cells by FACS and the retrospective assignment of fluorescence signals during scRNA-seq data analysis.
Library	DNA molecules that contain specific sequences (primers) that enable the initiation of high-throughput sequencing reactions.
Locked nucleic acids	Modified RNA nucleotides with a bridge connecting the 2′ oxygen and 4′ carbon to increase the hybridization properties of oligonucleotides.
Microtiter plates	Also known as microplates or microwell plates; flat plates with multiple wells used as individual reaction sites.
Pipeline	An analysis procedure in which inputs go through a number of processing steps chained together to produce an output.
Poisson distribution	A discrete probability distribution that expresses the probability of the number of events in specified intervals such as distance, area or volume.
Pooling	Combining molecules or cells for their joint processing.
Promoter	A DNA sequence that initiates transcription of the downstream sequence.
Pseudotime	An inferred time line of the progress cells make through a dynamic process such as cell differentiation.
Spike-in RNA	A pool of RNA transcripts of known sequence composition and quantity used to calibrate experiments.
Tagmentation	Reaction that involves the transposase-based cleaving of DNA and the tagging of the double-stranded DNA with universal overhangs.
Template-switching oligonucleotide (TSO)	A DNA oligonucleotide sequence that carries three riboguanosines (rGrGrG) at its 3′ end and binds to the cytosine extension of the cDNA molecules after RT.
Trajectory inference	Computational reconstruction of an underlying cellular developmental or differentiation path.
Unique molecular identifiers (UMIs)	Random sequences attached to transcripts and used as molecular tags to detect and quantify unique RNA molecules.
Zero-inflated data	Data with an excess of zero counts. To model zero-inflated data, a Poisson distribution is used.

through shearing effects; thus, centrifugation conditions should be optimized. Sufficient volumes should be used for cell washing and resuspension, as high concentrations can cause aggregation and clumping. Suspensions should be filtered with appropriately sized cell strainers (pore size larger than cell diameter) to remove clumps and debris. The recommended

Table 2 | Tissue-specific enzymatic treatments to prepare single-cell suspensions (from human and mouse samples)

Tissue	Digestion enzyme	Time (min)	Temperature (°C)	Final concentration	Ref.
Liver	Collagenase IV	10	37	0.16 mg/ml	126
	Liberase Blendzyme 3	5–8	37	40 µg/ml	9
	Collagenase, collagenase D and Pronase, trypsin	20, 20, 10	37	2.5 mg/ml, 10 mg/ml and 10 mg/ml, 0.05%	127
	Collagenase IV	30	37	0.05%	128
Lung	Dispase and elastase	45	37	0.33 U/ml and 3 U/ml	129
	Collagenase and dispase	45	37	0.2% solution	130
	Dispase, elastase and trypsin	60, 30, 15	4, 37 and 37	2 mg/ml, 5 U/ml plus 0.125%,	131
Skin	Trypsin	120	32	1×	132
	Liberase TL	15	37	2 mg/ml	133
Spleen	Collagenase D	45	37	2 mg/ml	134
GI tract	Dispase	20	37	0.4 mg/ml	36
	Trypsin	30	37	2 mg/ml	135
	TrypLE Express	1	37	1×	10
	Collagenase	40	37	1 mg/ml	136
	Collagenase I	60	37	2.5 mg/ml	137
	Collagenase IV	30	37	2 mg/ml	138
Pancreas	Collagenase type CLS IV	30	37	1 mg/ml	139
	Collagenase P	30	37	0.8 mM	140
	TrypLE Express	1	37	1×	141
	Accutase and TrypLE Express	10 and 5–20	37	1×	142
	Accutase	8–10	37	1×	143
	Trypsin	30	37	1×	144
Kidney	Liberase TL	15	37	2 mg/ml	133
Retina	Papain	45	37	4 U/ml	61,145
	Accutase	5	37	1×	146

cell-washing and resuspension solution is phosphate-buffered saline (calcium and magnesium free) containing bovine serum albumin to minimize cell losses and aggregation. Primary cells, stem cells and other sensitive cell types may require washing and suspension in alternative buffers to ensure viability, which also may decrease when cells are kept in suspension for a prolonged period. Cell clumps cause automated cell counters to underestimate the effective concentration of single cells, so suspensions should be processed as soon as possible after preparation, ideally within 30 min. It is important to minimize cellular aggregates, dead cells, noncellular nucleic acids and reverse-transcription (RT) inhibitors in single-cell preparations. To minimize these contaminants while maximizing the purity and unbiased recovery of different cell types, one may need to apply optimization (e.g., adjust the number of wash steps, the composition of the wash solution, centrifugation conditions and/or strainer type).

Preparation of cell suspensions. For isolation of single cells from suspensions (e.g., blood samples), samples are density centrifuged (e.g., using Ficoll-Paque or Histopaque-1077 techniques)[30], after which they can be used directly for single-cell capture. Solid tissues must first be dissociated via mechanical and enzymatic treatment. Initially, tissues are disaggregated by mechanical cutting or mincing with blades. Then enzymatic digestion is used to separate cells, with specific enzymes and digestion times used for different tissues (Table 2). Enzyme types include Accutase, elastase and collagenases, as well as commercial enzymatic mixtures such as TrypLE Express and Liberase Blendzyme 3. Elevated cell lysis can lead to cell clumping, which is reduced through treatment with DNase I during cell separation. Finally, suspensions are cleaned by filtering through a mesh or strainer before capture of single cells.

It is important to note that sample processing might introduce variation in the gene expression profile, as has been shown for the activation of stress-related genes[31]. Also, some more sensitive cell types might be damaged during sample preparation, so processing time should be kept to the minimum required. In contrast, too short digestion times could result in incomplete cell separation and the exclusion of tightly interconnected cells from subsequent single-cell analysis.

To avoid biases in cell type composition, one can use an alternative strategy that involves disruption of cellular membranes and isolation of the nuclei[25–29]. The sequencing of nuclear RNA was shown to be sufficient to deconvolute cell types[29], although this decreases the overall resolution per cell. Single-nuclei sequencing has been applied extensively for differentiated neurons, for example, as it is largely impracticable to isolate intact cells from highly interconnected adult neuronal tissue.

Single-cell capture. For transcriptome profiling in single cells, most methods require the physical isolation of cells in individual reaction volumes. Cells can be isolated by microdissection or pipetting[32], although high-throughput experiments use fluorescence-activated cell sorting (FACS)[33] or microfluidics[34] to guide cells into micro- or nanoliter reaction volumes, respectively. Microfluidic systems capture cells in integrated fluidics circuits (IFCs), droplets or nanowells, thus allowing thousands of cells to be processed simultaneously while minimizing reaction volumes and reagent use. FACS sorts cells into microtiter plates ready for library preparation by manual or automated processing, and facilitates the exclusion of dead or damaged cells, as well as the enrichment of target cell populations (e.g., through surface marker labeling). To reduce background and maximize assay performance, we also recommend FACS or magnetic-activated cell sorting (MACS) processing of single-cell solutions for microfluidic systems to remove debris, damaged/dead cells and cell aggregates.

Sample size and composition. To obtain an unbiased view of the cellular composition of a sample, one must capture all cells during the isolation process. Here attention must be paid to very small or large cells that may be excluded during FACS isolation or captured in microfluidic systems, respectively. However, for many experiments, it may be necessary to enrich for or exclude some cell types to increase the total number of cells of interest in the final scRNA-seq libraries. For example, profiling of specific immune responses requires enrichment of blood cell subtypes, whereas cancer studies might need to exclude blood cells (e.g., CD45$^+$ cells) to increase the overall number of tumor cells. Target populations can be selected by FACS and MACS with appropriate labeling (e.g., antibodies or transgenic systems). Microtiter plates and some nanowell capture systems allow index sorting, in which fluorescence intensity or cell size (FACS information) is associated with capture coordinates and subsequently with single-cell indices. The FACS device records the sorting position and intensity values of a given cell, thereby enabling the subsequent integration of transcriptome profiles with the recorded cell properties. For microfluidic systems, CITE-seq[35] provides a viable alternative that conserves information about surface markers. Here epitopes of interest are targeted with oligonucleotide-labeled antibodies. The antibody-specific sequences are poly(A)-tailed and contain barcodes that allow epitope tracking after scRNA-seq library preparation and sequencing.

To define adequate cell numbers per experiment, one must consider sample heterogeneity and subpopulation frequency (the estimated abundance of the cell type of interest). In particular, larger cell numbers are required to resolve the structure of heterogeneous samples with many expected subpopulations. Also, the total number of cells required increases when rare cell types need to be identified. One can calculate the required cell numbers by estimating both subpopulation structure and low-frequency cell-type abundance and defining the desired cell number per group (computational tool accessible at https://satijalab.org/howmanycells). Because most experiments target poorly described systems, heterogeneity can only be estimated, so pilot experiments are recommended before large-scale data production. For comparative studies across experimental conditions, patient samples or larger population cohorts, control experiments can be used to provide information about optimal cell numbers and the need for subpopulation enrichment steps. Specifically, selected samples can be profiled with high cell numbers to comprehensively identify tissue heterogeneity. Cell numbers in subsequent data production phases can then be adapted according to the required resolution. Similarly, seemingly homogeneous samples can be initially profiled using higher cell numbers and sequencing depth to reveal yet uncharted sample complexity. Note that higher cell numbers can also be beneficial for homogeneous samples, as this increases statistical power during analysis[36].

Sample preservation. All common scRNA-seq methods were initially designed to use freshly isolated cells. However, in research and clinical practice, immediate sample processing can be challenging because of a lack of the required infrastructure or specialized equipment, such as FACS devices. Moreover, although samples may be collected at multiple time points, simultaneous sample processing may be preferred to avoid technical batch effects. Sample preservation is a viable solution because it disconnects the location and time of sampling from the downstream processing steps. In this context, cryopreservation has been established for single-cell transcriptome analysis[22]. After sample storage for up to a year at −80 °C or in liquid nitrogen and subsequent thawing, cryopreserved cells from cell lines and primary samples show complete integrity of the RNA molecules and unchanged expression profiles as compared with those of freshly prepared cells. Note that multiple freeze–thaw cycles should be avoided through the preparation of aliquots or by scraping out still-frozen cells from storage vials. Similarly, methanol fixation has been established as an alternative for droplet-based single-cell methods, and could also be used to avoid technically induced variations in gene expression triggered by prolonged sample processing time[23]. Importantly, both methods allow the archiving and transport of samples and broaden the range of applications of scRNA-seq methods, for example, to the clinical context. However, both approaches have shown a potential bias in cell-type composition, and it is strongly recommended to thoroughly evaluate preservation methods for new cell types that have not been tested. For previously archived samples, such as snap-frozen specimens, nuclei sequencing provides the only solution for scRNA-seq[25–29]. Unlike in cryopreservation, the formation of ice crystals during snap-freezing disrupts the outer cellular membrane, although the nuclei remain intact. Nevertheless, it is preferable to make an initial estimation of the RNA integrity to avoid biases related to sample quality.

Table 3 | Key features of microtiter-plate- and microfluidics-based single-cell RNA sequencing methods

Method	Capture format	Cell loading	Single-cell indexing	Molecule identifier	Additives in RT	cDNA amplification	Fragmentation	Transcript coverage	Sequencing	Ref.
Smart-seq	Plate	FACS	Tagmentation	NA	NA	PCR	Tagmentation	Full length	Paired end	47
Smart-seq2	Plate	FACS	Tagmentation	NA	Betaine	PCR	Tagmentation	Full length	Paired end	147
STRT-seq	Plate	FACS	TSO	UMI	NA	PCR	DNase I	5′ end	Single end	48
STRT-seq-2i	Nanowell	FACS/ Poisson	TSO	UMI	Betaine	PCR	Tagmentation	5′ end	Single end	58
SCRB-seq	Plate	FACS	Oligo(T) primer	UMI	NA	PCR	Tagmentation	3′ end	Paired end	49
mcSCRB-seq	Plate	FACS	Oligo(T) primer	UMI	PEG	PCR	Tagmentation	3′ end	Paired end	50
Quartz-seq	Plate	FACS	Oligo(T) primer	NA	NA	PCR	Ultrasound	Full length	Paired end	51
Quartz-seq2	Plate	FACS	Oligo(T) primer	UMI	NA	PCR	Ultrasound	3′ end	Paired end	52
CEL-seq	Plate	FACS	Oligo(T) primer	NA	NA	IVT	KOAc, MgOAc	3′ end	Paired end	32
CEL-seq2	Plate	FACS	Oligo(T) primer	UMI	NA	IVT	Random priming	3′ end	Paired end	54
MARS-seq	Plate	FACS	Oligo(T) primer	UMI	NA	IVT	Zinc	3′ end	Paired end	53
Seq-Well	Nanowell	Poisson	Oligo(T) beads	UMI	Ficoll	PCR	Tagmentation	3′ end	Paired end	59
inDrops	Droplets	Poisson	Oligo(T) beads	UMI	IGEPAL	IVT	KOAc, MgOAc	3′ end	Paired end	60
Drop-seq	Droplets	Double Poisson	Oligo(T) beads	UMI	Ficoll	PCR	Tagmentation	3′ end	Paired end	61

NA, not applicable.

Single-cell RNA sequencing

Transcriptome profiling of individual cells can be split into four major components: RNA molecule capture, RT and transcriptome amplification, sequencing library preparation, and sequencing. Various scRNA-seq methods exist, but they all apply the same underlying principles. Below we discuss these basic experimental design considerations, and highlight common and emerging microtiter-plate-based and microfluidic scRNA-seq techniques and their applications. Key features of the different scRNA-seq approaches discussed below are also summarized in Table 3. Many of these methods have undergone systematic evaluation, which confirmed their generally high accuracy, although efficiency, scalability and costs vary considerably[37,38]. This should be taken into account during the selection of methods for a given experiment.

RNA molecule capture, reverse transcription and transcriptome amplification for sequencing library preparation. Most scRNA-seq methods, including those described below, capture poly(A)-tailed RNA, although specific protocols are available for profiling total RNA[39,40] or miRNAs[41]. After cell lysis, poly (A)-tailed RNA is captured by poly(T) oligonucleotides, which exclude abundant RNA types such as rRNA and tRNA. After capture, the RNA is reverse-transcribed into stable cDNA, at which point most methods add single-cell-specific barcodes within the poly(T) oligonucleotides that allow cost-effective multiplexed processing of pooled samples. Moreover, random-nucleotide-sequence stretches in the poly(T) oligonucleotide serve as unique molecule identifiers (UMIs) that allow the user to correct for amplification biases and reduce technical noise[42]. RT is a crucial step, and different protocols have been optimized in various ways with efficient enzymes and specific additives that maximize efficiency (Box 1). cDNA can then be amplified by PCR or through in vitro transcription (IVT). For this, adaptor sequences or RNA polymerase promoter sequences are introduced during RT or second-strand synthesis. Although IVT is less prone to biases through linear amplification of molecules, it requires additional downstream steps to convert the amplified RNA into cDNA and sequencing-ready libraries. PCR-based protocols require less hands-on time, but the exponential amplification phase leads to biases in RNA composition in the final libraries. Both approaches were shown to provide interpretable results and were successfully implemented in several scRNA-seq methods (Table 3).

Full-length versus 3′- or 5′-end transcript sequencing. Single-cell transcriptome profiling can be done through full-length transcript analysis or by digital counting of 3′ or 5′ transcript ends[42]. The choice of sequencing method should be dictated by the goal of the experiment—for example, to prioritize cost-effectiveness over retention of sequence information. Digital RNA counting is a cost-effective quantification strategy, although sequence information of the transcripts is lost to a large extent. Full-length transcriptome sequencing allows the detection of splice variants and alternative transcripts, as well as genetic alterations in the transcribed fraction, such as single-nucleotide variants[19,20] and fusion transcripts[43]. Moreover, genotypes of T and B cell receptors can be obtained from full-length transcriptomes[44]. Unlike 3′- and 5′-end methods, full-length protocols do not allow the introduction of UMIs and impede early cellular barcoding and pooling, which results in higher costs for library preparation. This limitation can be overcome through the use of long-read sequencing technologies that do not need library fragmentation[45]. However, such technologies generate smaller quantities of sequencing reads, and transcriptome quantification is not yet possible.

Box 1 | Optimization of reverse transcription for single-cell transcriptome sequencing

Enzymes

Reverse transcription (RT) is one of the most critical steps in the library-preparation workflow. Despite its importance, however, relatively little has been done to improve the efficiency of the underlying enzymes. Reverse transcriptases are based on Moloney murine leukemia virus (MMLV)-derived enzymes, which originally had low processivity and high error rates due to their retroviral origins. Different point mutations have been introduced to improve processivity, resulting in enzymes that can reverse-transcribe even very long RNAs (up to 12–14 kb). SuperScript II is a commonly used enzyme that became popular in the single-cell field because of its template-switching properties, and is used in methods such as Smart-seq2[147] and STRT-seq[48,58]. Most important, SuperScript II carries point mutations that inactivate its RNase H domain, thus impairing competitive RNA degradation during cDNA synthesis. Alternative RT enzymes have been reported to have similar or superior performance, such as Maxima H (used in SCRB-seq[49,50]) and SMARTscribe in the SMARter v4 kit (Takara Bio). Protocols that do not require template switching and that generate second strands by other means, such as poly(A)-tailing or random priming[52,54], can use SuperScript III, which carries different point mutations in the RNA polymerase and has increased thermal stability.

Additives

In an attempt to overcome the limitations of MMLV-based RT enzymes, several additives have been tested over the years. The challenge of generating full-length cDNA libraries has been a constant issue in molecular biology, predating the advent of single-cell RNA sequencing. Carninci et al.[148] showed that the sugar trehalose has a thermo-stabilizing and thermo-protective effect on RT enzymes. Conducting the RT reaction at a higher temperature enhances the unfolding of secondary RNA structures that could hinder enzyme processivity. This finding was confirmed and later extended to the addition of betaine, alone or in combination with trehalose, to improve thermo-protection and related cDNA yield[149,150]. Smart-seq2[147] and STRT-seq-2i[58] use betaine in combination with magnesium chloride; use of the latter at concentrations higher than 1 mM has been suggested to have a synergic destabilizing effect in the presence of betaine[151]. However, the extra magnesium chloride could also reduce the chelating function of 1,4-dithiothreitol (DTT), which is commonly used in RT reactions to guarantee higher cDNA yields and longer transcripts. In the very first published single-cell sequencing method, Tang et al.[152] used the T4 gene 32 protein (T4g32p), a single-stranded binding protein that increases yield and processivity during RT.

Template-switching oligonucleotides

The template-switching reaction relies on 2–5 untemplated cytosine nucleotides, which are added to newly synthesized cDNA (but not to fragmented or uncapped RNAs) when the enzyme reaches the 5′ end of the RNA. The presence of a TSO carrying three complementary guanosines at its 3′ end enables the enzyme to switch templates and to add the complementary sequence of the TSO to the cDNA (including a PCR adaptor for subsequent amplification)). It has been suggested that the reduced RNA capture efficiency of single-cell RNA-seq protocols might be due to the unstable binding of TSO to the untemplated nucleotides. The Smart-seq2 protocol addresses this issue by modifying the last nucleotide of the TSO with a locked nucleic acid. Furthermore, the importance of each nucleotide in the TSO has been extensively evaluated to define its optimal composition[153].

scRNA-seq methodologies: microtiter-plate-based approaches. After isolation of single cells into microtiter plates by FACS, a full-length transcript or 3′/5′-end protocol can be applied. Smart-seq2[46] is a widely used method to reverse-transcribe and amplify full-length transcripts. After RT, the enzyme adds cytosines to the cDNA, providing the basis for a template-switching reaction. Here a template-switching oligonucleotide (TSO) binds to the extra cytosine and provides the template for the addition of PCR adaptor sequences for subsequent cDNA amplification. Compared with the original version[47], the updated protocol improves molecule-capture efficiency and yield by using locked nucleic acids in the TSO and adding betaine to the RT reaction. Sequencing libraries are prepared by tagmentation, which simultaneously fragments and indexes the cells. The Smart-seq2 protocol is highly efficient in capturing RNA molecules[37], although the late indexing step makes it more expensive than other methods. Furthermore, the absence of UMIs makes downstream data analysis more challenging. Nevertheless, the protocol provides an adequate solution if deep single-cell phenotyping is required (e.g., for homogeneous samples or for analysis of weakly expressed genes).

STRT-seq[48] uses a similar strategy for RT and template switching, but it incorporates single-cell barcodes into the TSO. This allows early pooling of cells and cost-effective multiplex processing. STRT-seq enriches 5′ transcript ends through the use of biotinylated purification and 5′-specific PCR primers. Analysis of the 5′ transcript has the advantage of providing information about transcription start sites. Moreover, cell barcodes and transcripts are obtained in a single read, which allows for cost-effective single-end sequencing. Although the original STRT-seq protocol could not correct for amplification biases, later updates for the first time included UMIs in an scRNA-seq method[42]. The SCRB-seq[49] protocol incorporates single-cell barcodes and UMIs in the poly(T) primer, thereby enabling 3′ amplification of transcripts, and, as with STRT-seq, early indexing allows cell pooling to reduce costs. The RNA capture efficiency of the original protocol was improved by an increase in the RT mix density: molecular crowding SCRB-seq (mcSCRB-seq[50]) includes polyethylene glycol to increase binding-event probabilities. In addition, the PCR enzyme was switched from KAPA to the Terra polymerase to further improve library complexity. In Quartz-seq[51], the template-switching reaction is replaced by a poly(A)-tailing step. The additional adenosines provide a template for a poly(T)-primed second-strand synthesis followed by PCR amplification. The amplified transcriptome then undergoes ultrasound fragmentation and sequencing-adaptor ligation. A later version, Quartz-seq2[52], improved the molecule-detection efficiency by using shorter RT primers and improving poly(A)-tagging efficiency.

Amplification biases during exponential PCR are addressed in CEL-seq[32], in which transcripts are copied through IVT. The linear amplification of molecules, made possible by inclusion of a T7 promoter in the poly(T) primer, results in more evenly duplicated transcriptomes. Also, transcriptome amplification by IVT does not require template switching, which improves molecule-capture efficiency. This workflow was further optimized in MARS-seq[53] by inclusion of UMIs in the poly(T) primers and upscaling of cell numbers through automation. In addition, the original CEL-seq protocol was updated in CEL-seq2[54] for more efficient RNA capture and a simplified workflow. Briefly, the CEL-seq2 protocol uses UMIs, a shorter RT primer, and more efficient RT and second-strand synthesis enzymes. Furthermore, cDNA synthesis after IVT is initiated by random priming instead of adaptor ligation.

scRNA-seq methodologies: microfluidic systems-based approaches. Microfluidics allows higher-throughput scRNA-seq workflows, thus eliminating the technical constraints on scalability associated with microtiter plates. Moreover, reducing reaction volumes from microliters to nanoliters reduces costs and technical variability[55] while improving cDNA yield[56]. There are three strategies for capturing cells: IFCs, droplets and nanowells, all of which increase the number of capture sites relative to that achieved with microtiter plates. The first microfluidics system used for scRNA-seq was designed as an automated array solution (Fluidigm C1) in which single cells enter a fluidics circuit and then are immobilized in hydrodynamic traps, lysed, and processed in consecutive nanoliter reaction chambers via a modified Smart-seq2 protocol. Although early versions could use only commercial scRNA-seq assays, a more recent open format accommodates custom scRNA-seq protocols[42] and additional applications for genetics and epigenetics single-cell experiments[57]. Costs were further reduced by an increase in throughput and cell capture from 96 to 800 sites (C1 HT-IFC), and inclusion of an early-indexing strategy that allows cell pooling. Notably, this high-throughput version switched from full-length to 3′ RNA sequencing. Also, the array formats, which are restricted to specific cell sizes (small, medium and large arrays), affect unbiased sampling from complex sample types. To further increase cell numbers, microfluidics progressed to open nanowell systems that allow better scalability. In STRT-seq-2i[58], the original protocol was applied in a nanowell platform with 9,600 sites, with cells loaded by limiting dilution or direct addressable FACS sorting. Positioning cells by FACS allows for index sorting that assigns cell properties (e.g., fluorescence signal or size) to array coordinates and barcodes. Nanowells containing cells can be specifically utilized by targeted dispensing, which substantially reduces reagent costs and contamination by ambient RNA. Moreover, the array format allows imaging to exclude doublets. To guarantee high cell viability during the time-consuming loading into nanowells, FCS can be added to the buffer and sample aliquots can be kept on ice. Alternatively, Seq-Well[59] provides a nanowell-based method that captures cells in 86,000 sub-nanoliter reactions. The underlying principle is the pre-loading of nanowells with barcoded beads before cells enter the capture sites through limited dilution. Subsequently, the arrays are sealed for cell lysis and RNA molecule capture on beads before the immobilized molecules are pooled for 3′-end library production. The Seq-Well system is portable, and so allows sample processing at the sampling sites, as large equipment is not required. The fact that no major investments are required makes the Seq-Well system a flexible and cost-effective alternative. However, although cells can be monitored by microscopy, the random distribution of barcoded beads does not allow the user to integrate imaging data. Also, the method requires experienced users to obtain reproducible, high-quality results.

Although they are scalable to higher throughputs, the IFC and nanowell approaches are intrinsically constrained by the number of reaction sites. Droplet-based systems overcome this by encapsulating cells in nanoliter microreactor droplets. Here, cell numbers scale linearly with the emulsion volume, and large numbers of droplets are produced at high speed, which facilitates large-scale scRNA-seq experiments. Furthermore, droplet size can be adjusted to reduce potential biases during cell capture. Because barcodes are introduced into droplets randomly, this approach does not allow the assignment of cell barcodes to images and so precludes the visual detection of doublets and the integrative analysis of cell properties (e.g., fluorescent signals) with transcriptome profiles. Two droplet-based methods, inDrops[60] and Drop-seq[61], were developed in parallel, with related commercial systems allowing straightforward implementation. inDrops[60,62] encapsulates cells by using hydrogel beads bearing poly(T) primers with defined barcodes, after which the photo-releasable primers are detached from the beads to improve molecule-capture efficiency and initiate in-drop RT reactions. The barcoded cDNAs are then pooled for linear amplification (IVT) and 3′-end sequencing-library preparation. The technique has extremely high cell-capture efficiency (>75%) owing to the synchronized delivery of deformable beads, allowing near-perfect loading of droplets. Therefore, the system is most suitable for experiments with limited total numbers of cells. The inDrops system is licensed to 1CellBio, and a variant protocol has been commercialized as the Chromium Single Cell 3′ Solution (10x Genomics)[63]. The Chromium system is straightforward to implement and standardize, although library-preparation costs are considerably higher than those of the original system. Unlike inDrops protocols, Drop-seq[61] uses beads with random barcodes. After cell lysis and RNA capture, the drops are broken and pooled, covalent binding is carried out through cDNA synthesis, the cDNA is amplified by PCR, and 3′-end sequencing libraries are produced by tagmentation. Drop-seq has lower cell-capture efficiency than inDrops methods because beads and cells are delivered by double limiting dilution (double Poisson distribution), which results in 2–4% barcoded cells. The Drop-seq system is commercially available through Dolomite Bio, and a similar system is provided by Illumina (ddSEQ).

scRNA-seq methodologies: split-pool barcoding-based approaches. Conceptually different from the above techniques are methods based on combinatorial barcoding. Here, cells are not processed as individual units but isolated in pools. These pools are split and mixed, with each round integrating pool-specific barcodes. The combination of such pool indices results in unique barcode combinations for each cell through their random assignment during consecutive pooling processes. Both split-pooling methods, SPLiT-seq (split-pool ligation-based transcriptome sequencing)[64] and sci-RNA-seq (single-cell combinatorial-indexing RNA-seq)[12], were shown to reliably produce single-cell transcriptomes and to be scalable to hundreds of thousands of cells per experiment. SPLiT-seq includes four rounds of indexing, resulting in >20 million possible barcode combinations. After initial indexing during RT, two rounds of index ligation and a final PCR indexing step create cell-specific barcoded 3′-transcript libraries. During the second ligation round, UMIs are incorporated for the subsequent correction of amplification biases. Additional rounds of barcoding or a switch from 96-well to 384-well microtiter formats could further scale up cell numbers. The original sci-RNA-seq protocol includes a two-step indexing workflow with the first index and UMI introduced during RT and a second index during PCR amplification (after tagmentation). The use of indexed tagmentation sequences could further scale up possible barcode combinations and increase cell numbers per experiment. Formaldehyde- and methanol-based fixation of cells, used in SPLiT-seq and sci-RNA-seq, respectively, allows sample storage, thereby providing additional flexibility to the experimental designs. Both methods allow the processing of nuclei and consequently the analysis of more challenging cell types, such as neurons. The split-pool strategy used in sci-RNA-seq was further shown to be applicable in different single-cell epigenomic analysis approaches, including open chromatin (sci-ATAC-seq[65]), chromatin conformation (sci-Hi-C[66]) and DNA methylation (sci-MET[67]) approaches.

Library preparation and sequencing. In library preparation for short-read sequencing applications, the amplified cDNA (PCR) or RNA (IVT) is fragmented before sequencing adaptors are added. Fragmentation can be achieved enzymatically (with tagmentase or DNase), chemically (with zinc, KOAc or MgOAc) or through mechanic forces (e.g., ultrasound) (Table 3). 3′- or 5′-based libraries are subsequently amplified with primers specific for the transcript end or start, respectively. During this step of the protocol, a pool-specific index can be introduced that allows the multiplexed sequencing of multiple experiments. Full-length methods introduce the cell-specific barcodes only after fragmentation, thus impeding pooled processing of cells at earlier stages of the protocol. Apart from STRT-seq, scRNA-seq libraries require paired-end sequencing, in which one read provides information about the transcripts while the other reads the single-cell barcodes and UMI sequences. STRT-seq incorporates the cell barcode and UMI at the 5′-transcript end, which allows cell, molecule and transcript information to be captured in a single read, as no poly(T) stretch separates the respective sequences. High-throughput microfluidics-based experiments generally involve sequencing to lower depths (<100,000 reads per cell), whereas higher read numbers (~500,000 reads per cell) are optimal for many microtiter-plate formats[38]. Nevertheless, single-cell libraries are usually not sequenced to saturation, and the phenotyping resolution (detection of more genes and of those expressed at lower levels) can benefit from further increases in the sequencing depth. Annotation of splice variants from full-length transcriptomes requires deeper sequencing to better resolve the expression levels of transcript variants.

Further technical considerations

Cell doublets. An intrinsic problem for most microfluidics-based methods is that two cells can be captured per reaction site (nanowell or droplet), both receiving identical barcodes. Doublet rates can be experimentally determined in species-mixture experiments, but otherwise can only be estimated. They occur when cells are positioned randomly in reaction sites by limiting dilution and can be controlled by the cell suspension concentration. The relationship between cell loading and doublet rate was systematically quantified for the Chromium system[63]. Up to the maximal recommended loading of 10,000 cells per droplet lane, the doublet rates showed a linear relationship (in line with the Poisson loading of cells into droplets), with inferred rates ranging from 2% (2,500 cells) to 8% (10,000 cells). Other microfluidics approaches yield similar numbers: Drop-seq, 0.36–11.3% (12.5–100 cells/μl; ref. [61]); InDrops, 4% (ref. [60]); and Seq-Well, 1.6% (ref. [59]). The doublet rate decreases at higher dilutions, with a resulting increase in reagent costs per cell, as fewer total cells are captured per experiment. Researchers can partially overcome this handicap by jointly capturing samples from different individuals, such that genotype differences allow the user to distinguish between donors and thereby reliably identify doublets[68]. Specifically, single-nucleotide polymorphisms identified from the RNA sequencing reads are used to determine the donor origin of the cells and to discriminate samples that were processed in a single batch. However, such a workflow is practicable only when the experimental design includes different human individuals or model organisms with distinct genetic backgrounds. Currently, there is no computational method for credibly identifying doublets, so doublet rates must be minimized by experimental design. Doublets can have dramatic consequences for data interpretation, as artifactual mixed transcriptomes can easily be mistaken for intermediate cell states in dynamic systems.

Cell-capture efficiency. Cell-capture efficiency is an important consideration, especially in experiments involving primary or rare samples. The number of cells that receive barcodes is directly related to the proportion of sample that enters downstream analysis. The capture efficiency of FACS-based methods is constrained by the time the device requires to move between wells. To maximize capture rates of FACS-based methods, one can dilute and sort cell suspensions at low speed (e.g., 100 cells/s). Microfluidics technologies differ markedly in capture efficiency, mainly as a result of cell and bead loading mechanics. The HT-IFC system captures a maximum of 800 out of 6,000 injected cells. In nanowell systems that use limiting dilution for cell loading (no sorting), cells enter reaction sites by gravity, with generally high efficiency. For example, 10,000 cells are added to the surface of a Seq-Well array, and around 3,000 cells are captured. For droplet-based systems, the rate at which cells enter the analysis is directly related to the loading efficiency of the beads. When most droplets contain barcoded beads, cell capture is optimal (inDrops). In contrast, if beads and cells are encapsulated by limiting dilution, most cells do not enter a bead-containing droplet, which results in lower capture efficiency (Drop-seq; discussed above).

Costs. The total cost of scRNA-seq experiments is determined by three main components: equipment, reagents and sequencing. For most methods, the cost of scRNA-seq library preparation scales linearly with cell numbers; an exception is custom droplet methods. The actual costs per cell vary widely across methods and institutes, with microfluidic systems being generally cheaper (<$0.30 per cell) than early-indexing plate-based 3′ digital counting methods (~$1–2 per cell). Late-indexing full-length transcriptome profiling is costlier, even with small volumes (~$8–12 per cell). However, costs can be reduced through the use of non-commercial tagmentase[69] or minimum reaction volumes and automated workflows for plate-based formats[70]. Importantly, microtiter plates can be shipped and stored, which disconnects sampling sites from scRNA-seq processes such that expensive devices can be centralized in core units, thus optimizing resource management. Custom microfluidics methods further decrease costs per cell. Commercialized microfluidics methods are more expensive ($0.50–2.00 per cell) than custom systems (<$0.30 per cell), although their automated design reduces hands-on time and personnel costs.

Although the cost of library preparation is decreasing rapidly, sequencing costs are becoming a major factor. Methods with higher molecule-capture efficiency produce more complex sequencing libraries, which makes them informative at low sequencing depths. Consequently, more efficient scRNA-seq methods can compensate for higher library preparation costs by decreasing overall sequencing costs.

Data processing

Data processing includes all the steps necessary to convert raw sequencing reads into gene expression matrices, using workflows similar to those used for bulk RNA-seq. After FASTQ reads have been generated and their quality has been checked (with tools such as FastQC; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), the next important step is de-multiplexing of reads using cell barcodes. Whereas Smart-seq libraries can be directly de-multiplexed using the index reads, the 3′-end-based methods require a dedicated processing step to identify the single-cell indexes in the sequencing reads. De-multiplexed reads are then mapped to reference genomes with alignment tools such as TopHat[71] and STAR[72], the latter of which offers proven accuracy and splice-variant sensitivity. Recent alignment tools were optimized for fast handling of large-scale datasets without loss of accuracy. For example, Kallisto[73] reduces the alignment time by two orders of magnitude through pseudo-alignment, as opposed to alignment of individual bases. In a final processing step, mapped reads are quantified to create a transcript expression matrix. RSEM[74], Cufflinks[75] and HTSeq[76] can be used for full-length transcript datasets, whereas special tools, such as UMI-tools[77], which accounts for sequencing errors in UMI sequences, are available for counting UMI-tagged data types.

In addition to the specific tools available for individual processing steps, single-cell data processing pipelines have been developed that combine mapping and quantification steps and include quality control measures for reads and cells. A pipeline developed by Ilicic et al.[78] supports various mapping and quantification tools, and includes modules for filtering low-quality cells. Scater provides an organized workflow for converting raw sequencing reads into a ‘single-cell expression set’ (SCESet) class, a data structure that facilitates data handling and analysis[79]. Other available pipelines are either protocol specific (e.g., zUMI[80], scPIPE[81] and SEQC[82] for UMI data) or technology specific (e.g., Cell Ranger for Chromium systems). The scRNA-tools database (http://www.scRNA-tools.org) provides a comprehensive list of available computational tools for data processing and analysis[83]. Methods are categorized by analysis task, and researchers can select tools according to the required analysis type.

Normalization. Single-cell RNA-seq datasets show high levels of noise and variability related to nonbiological technical effects, including dropout events due to stochastic RNA loss during sample preparation, biased amplification and incomplete library sequencing. Technical variation also results from batch effects on processing units (e.g., plates or arrays), time points, facilities and other sources. Moreover, natural variability complicates analysis because of, for example, variable cell size and RNA content, different cell cycle stages and gender differences. Therefore, dataset normalization becomes an important step for meaningful data analysis. This can be guided by the addition of artificial spike-in RNA, which is used to model technical noise, as implemented in BASiCS[84]. However, it is not clear whether artificial RNA sufficiently reflects the behavior of endogenous RNA, or whether cellular RNA influences spike-in detection. Recent high-throughput methods distribute cells by limiting dilution, which makes the use of spike-in RNA impracticable because of the high number of otherwise empty reaction volumes. Alternative normalization methods originally developed for bulk RNA sequencing, such as log-expression[85], trimmed mean M-values[86] and upper-quartiles[87], can also be used in scRNA-seq, although more-specialized normalization methods are being developed that can better handle many aspects of this specific type of data. Recent single-cell approaches apply between-sample normalization (SCnorm[88]) or normalize on cell-based factors after pool-based size factor deconvolution (SCRAN[89]). However, for correction for large-scale sources of variation, a recommended and standard procedure is data modeling with the correct distribution. Here, confounding factors can be incorporated as covariates into the model and regressed out. Whereas batch effects are usually detected by visual inspection of reduced-space representations (e.g., principal components), kBET[90] is a batch-effect test based on k nearest neighbors. It quantitatively measures batch effects within and between datasets without directly correcting the data. This approach concludes that a combination of log normalization or SCRAN pooling with ComBat[91] or limma[92] regression provides the best batch-corrected dataset while preserving the biological structure. The batch effect problem is magnified when datasets from different time points, individuals or scRNA-seq methods are integrated. In this case, Haghverdi et al.[93] propose an approach based on mutual nearest neighbors in which a shared subset of populations is sufficient to correct for batch effects across experiments, although predefined or equal population compositions are required. Alternatively, by inferring cell clusters from gene expression similarities and coexpression patterns, Biscuit (Bayesian inference for single-cell clustering and imputing)[82] identifies and corrects for technical variation per cell. Also, the commonly used scRNA-seq package Seurat provides a solution for integrating datasets based on common sources of variation[94], with a new feature that allows the identification of shared populations and facilitates comparative analysis across datasets.

Imputation and gene selection. In addition to having a high noise level, scRNA-seq datasets are also very sparse, which poses further challenges to cellular phenotyping and data interpretation. Non-expressed genes and technical shortcomings, such as dropout events (unsequenced transcripts), result in many zeros in the expression matrix, and thus an incomplete description of a single cell's transcriptome. To reduce sparsity, missing transcript values can be computationally inferred by imputation, for example, with MAGIC[95], which uses diffusion maps to find data structures and restore missing information. Alternatively, scImpute[96] learns a gene's dropout probability by fitting a mixture model and then imputes probable dropout events by borrowing information from similar cells (selected on the basis of genes that are not severely affected).

A common strategy for determining heterogeneity in a sample is to analyze highly variable genes across datasets. A thorough feature-selection step to remove uninformative or noisy genes increases the signal-to-noise ratio but also reduces the computational complexity. Commonly used strategies for extracting variable genes in scRNA-seq tools exploit the relationship between the mean transcript abundance and a measure of dispersion such as the coefficient of variation[97], the dispersion parameter of the negative binomial distribution[98] or the proportion of total variability[84].

Data analysis

Some of the major applications of scRNA-seq experiments include assessment of sample heterogeneity and identification of novel cell types and states. This is achieved through determination of coexpression patterns and clustering of cells by similarity. Cell clusters can subsequently be interpreted through annotation of gene sets that drive clusters (marker genes). A common way to visually inspect cellular subpopulation structures is to carry out dimensionality reduction (DR) and project cells into a two- or three-dimensional space. Principal component analysis (PCA) and *t*-distributed stochastic neighbor embedding (t-SNE) are commonly used approaches for data representation[99,100]. Diffusion components[101] and uniform manifold approximation and projection (UMAP)[102] are viable alternatives that overcome some limitations of PCA and t-SNE by preserving the global structures and pseudo-temporal ordering of cells, as well as being faster[103]. Even though DR techniques can guide the initial data inspection, more-robust clustering algorithms are needed to define subpopulations among cells.

Although prior assumptions and canonical population markers allow supervised clustering (e.g., with Monocle2[104]), hypothesis-free unsupervised clustering is preferred in most cases. A commonly used unsupervised algorithm is hierarchical clustering, which provides consistent results without a predefined number of clusters. Hierarchical clustering can be conducted in an agglomerative (bottom-up) or divisive (top-down) manner, with consecutive merging or splitting of clusters, respectively. Tools such as PAGODA[105], SINCERA[106] and bigSCale[7] implement hierarchical clustering. Another suitable unsupervised clustering algorithm is *k*-means, which estimates *k* centroids (centers of the clusters), assigns cells to the nearest centroid, recomputes centroids on the basis of the mean of cells in the centroid clusters, and then reiterates these steps. SC3, for example, integrates both *k*-means and hierarchical clustering to provide accurate and robust clustering of cells[107]. Other unsupervised approaches, such as SNN-Cliq[108] and Seurat[94], use graph-based clustering, which builds graphs with nodes representing cells and edges indicating similar expression, and then partitions the graphs into interconnected 'quasi-cliques' or 'communities'. Clustering can be done directly on the basis of expression values or more processed data types, such as principal components or similarity matrices, the latter of which shows improved yield in cluster separation. Cluster stability is measured via resampling methods (e.g., bootstrapping) or on the basis of cell similarities within assigned clusters (e.g., silhouette index). To support cluster reproducibility, different algorithms can be compared using adjusted Rand indexes[107]. Clusters can be represented by color-coding in a low-dimensional space produced by the DR algorithms discussed above (e.g., PCA, t-SNE).

Marker genes that discriminate subpopulations can be identified by differential gene expression analysis of clusters using, for example, model-based approaches such as SCDE[109], MAST[110] and scDD[111], which account for data bimodality by using a mixture model. Individual genes can be evaluated to serve as binary classifiers for cell identity with, for example, ROC or LRT tests based on the zero-inflated data[94,107]. A recent publication comprehensively compared differential expression analysis methods for scRNA-seq and can be referred to as a guide for the selection of appropriate differential expression tools[112].

Another important application of scRNA-seq is trajectory inference, which estimates dynamic processes by ordering cells along a predicted differentiation path (pseudotime) using algorithms such as reversed graph embedding (Monocle2[113]) and minimum spanning tree (TSCAN[114]). Also, trajectory inference methods have been comprehensively benchmarked through tests of their accuracy and overall performance[115]. To further facilitate the interpretation of results, tools such as SCENIC[116] provide the opportunity to investigate active regulatory networks in subpopulations of cells. The analysis guides the identification of active transcription factors, eventually providing insights into the cellular mechanisms that drive heterogeneity. For cluster annotation, scmap facilitates comparison of data across experiments by projecting cells from one dataset onto cell types or individual cells from another scRNA-seq experiment[117]. With cell convolution tools such as bigSCale[7], scRNA-seq analysis can be expanded to millions of cells. Eventually, single cells can be mapped back to the spatial tissue context via experimental approaches[118,119] or pseudo-spatial ordering of cells[2,9,94].

To make scRNA-seq data publicly available, one can use data storage and sharing repositories. The Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) is commonly used to provide access to raw data and more-processed formats, such as gene expression quantification matrices. Large-scale projects, such as the Human Cell Atlas, set up specific data coordination platforms to further ease data query and accessibility. For data analysis, many researchers provide free open access to their computational pipelines through public databases such as GitHub (https://github.com/) or offer ready-to-use packages through, for example, Bioconductor (https://www.bioconductor.org/).

Summary

Although it is challenging to define broadly applicable designs for scRNA-seq experiments, we here provide general guidelines to support the production of high-quality datasets and their meaningful interpretation. Thoroughly planned and conducted sample preparation is critical to preserve cellular and RNA integrity and allow unbiased representation of the sample composition. The selection of downstream scRNA-seq techniques is driven by the complexity of the underlying sample and the desired resolution per cell. Although large numbers of cells, processed in microfluidic systems, might better represent the composition of heterogeneous samples, an in-depth analysis of smaller samples could be more appropriate for resolving subtle differences in homogeneous mixtures. Budget restraints and reduced library complexity generally lead to the shallow sequencing of high numbers of cells, whereas cell-type-focused experiments with sensitive methods can benefit from deeper sequencing. Eventually, the analysis and interpretation of single-cell transcriptomes is enabled by a wealth of computational methods specifically tailored to answer biological questions in a hypothesis-free manner or guided by previous knowledge. Despite technical challenges, scRNA-seq experiments are a powerful tool that can be used to fully resolve sample heterogeneity and dynamic cellular systems or to identify perturbation effects at high resolution.

Future directions of the single-cell field

Single-cell transcriptomics technologies are advancing rapidly. Cell numbers that can be analyzed are increasing to hundreds of thousands of cells per experiment, markedly improving statistical power and resolution for detecting rare and transient cell types. However, high-throughput techniques come with the expense of decreased molecule capture rates, and future methods need to better balance cell numbers with cell resolution. This will be accompanied by decreased sequencing costs, eventually allowing comprehensive, high-resolution snapshots of complex tissues to be achieved. Today, tissue- and organism-level projects use 'sky-dive' experimental strategies, initially creating a low-resolution atlas with thousands of cells to estimate sample heterogeneity, and then zooming in on target cell types by means of efficient scRNA-seq methods to achieve higher per-cell resolution. In the future, high-resolution maps will allow users to zoom in on the existing data, circumventing costly and time-consuming sample reprocessing. Microfluidics methods have already driven a paradigm shift in experimental designs, and conceptually different alternative methods such as combinatorial barcoding[12,64] might push the barrier back even farther. Because they do not require physical separation of individual cells, these approaches allow for cost-effective parallel processing of cells, which will make it possible for cell numbers to be scaled up even further.

An additional future avenue of intense investigation will be based on advances in monitoring of transcriptional profiles in spatial contexts. scRNA-seq relies on disconnection of cells from their natural environment, but spatial methods, including in situ sequencing[120] and single-molecule (smFISH[118]) and multiplexed error-robust (MERFISH[119]) fluorescence in situ hybridization, profile gene expression in the tissue context. Although current methods have low transcriptome resolution or require prior marker selection, they are extremely powerful in resolving tissue complexity[9,121]. Future spatial methods should allow the field to advance from the current combinatory experimental designs[122], or pseudo-space analysis[2,94], to a full tissue expression profile in three dimensions. Eventually, phenotype heterogeneity and dynamics in living multicellular systems will be resolved by the fusion of unbiased transcriptome profiling in spatial and temporal dimensions with the combined profiling of additional layers of molecular information, such as genetic variation[123] and gene regulatory marks (e.g., DNA methylation[124] and open chromatin[125]), from the very same cell.

References

1. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
2. Ibarra-Soria, X. et al. Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nat. Cell Biol.* **20**, 127–134 (2018).
3. Grün, D. et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19**, 266–277 (2016).
4. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
5. Bendall, S. C. et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
6. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: recording the past and predicting the future. *Science* **358**, 69–75 (2017).
7. Iacono, G. et al. bigSCale: an analytical framework for big-scale single-cell data. Preprint at *bioRxiv* https://doi.org/10.1101/197244 (2017).
8. *Tabula Muris* Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* **562**, 367–372 (2018).
9. Halpern, K. B. et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).

10. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
11. Karaiskos, N. et al. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
12. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
13. Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M. & Reddien, P. W. Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science* **360**, eaaq1736 (2018).
14. Davie, K. et al. A single-cell transcriptome atlas of the aging *Drosophila* brain. *Cell* **174**, 982–998 (2018).
15. Han, X. et al. Mapping the mouse cell atlas by Microwell-seq. *Cell* **172**, 1091–1107 (2018).
16. Shahbazi, M. N. et al. Pluripotent state transitions coordinate morphogenesis in mouse and human embryos. *Nature* **552**, 239–243 (2017).
17. Enge, M. et al. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell* **171**, 321–330 (2017).
18. Calon, A. et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat. Genet.* **47**, 320–329 (2015).
19. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
20. Tirosh, I. et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
21. Puram, S. V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624 (2017).
22. Guillaumet-Adkins, A. et al. Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biol.* **18**, 45 (2017).
23. Alles, J. et al. Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.* **15**, 44 (2017).
24. Wang, W., Penland, L., Gokce, O., Croote, D. & Quake, S. R. High fidelity hypothermic preservation of primary tissues in organ transplant preservative for single cell transcriptome analysis. *BMC Genomics* **19**, 140 (2018).
25. Lacar, B. et al. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.* **7**, 11022 (2016).
26. Krishnaswami, S. R. et al. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat. Protoc.* **11**, 499–524 (2016).
27. Habib, N. et al. Div-Seq: single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925–928 (2016).
28. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
29. Bakken, T. E. et al. Equivalent high-resolution identification of neuronal cell types with single-nucleus and single-cell RNA sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/239749 (2017).
30. Villani, A.-C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
31. van den Brink, S. C. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
32. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
33. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
34. Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat. Rev. Genet.* **18**, 345–361 (2017).
35. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
36. Barriga, F. M. et al. Mex3a marks a slowly dividing subpopulation of Lgr5$^+$ intestinal stem cells. *Cell Stem Cell* **20**, 801–816 (2017).
37. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643 (2017).
38. Svensson, V. et al. Power analysis of single-cell RNA sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
39. Avital, G. et al. scDual-Seq: mapping the gene regulatory program of *Salmonella* infection by host and pathogen single-cell RNA sequencing. *Genome Biol.* **18**, 200 (2017).
40. Hayashi, T. et al. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.* **9**, 619 (2018).
41. Faridani, O. R. et al. Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.* **34**, 1264–1266 (2016).
42. Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
43. Giustacchini, A. et al. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat. Med.* **23**, 692–702 (2017).
44. Stubbington, M. J. T. et al. T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016).
45. Karlsson, K. & Linnarsson, S. Single-cell mRNA isoform diversity in the mouse brain. *BMC Genomics* **18**, 126 (2017).
46. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
47. Ramsköld, D. et al. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
48. Islam, S. et al. Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nat. Protoc.* **7**, 813–828 (2012).
49. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. Preprint at *bioRxiv* https://doi.org/10.1101/003236 (2014).
50. Bagnoli, J. W. et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. Preprint at *bioRxiv* https://doi.org/10.1101/188367 (2017).
51. Sasagawa, Y. et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* **14**, R31 (2013).
52. Sasagawa, Y. et al. Quartz-Seq2: a high-throughput single-cell RNA sequencing method that effectively uses limited sequence reads. *Genome Biol.* **19**, 29 (2018).
53. Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
54. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
55. Wu, A. R. et al. Quantitative assessment of single-cell RNA sequencing methods. *Nat. Methods* **11**, 41–46 (2014).
56. Streets, A. M. et al. Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl Acad. Sci. USA* **111**, 7048–7053 (2014).
57. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
58. Hochgerner, H. et al. STRT-seq-2i: dual-index 5′ single cell and nucleus RNA-seq on an addressable microwell array. *Sci. Rep.* **7**, 16327 (2017).
59. Gierahn, T. M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
60. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

61. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
62. Zilionis, R. et al. Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* **12**, 44–73 (2017).
63. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
64. Rosenberg, A. B. et al. Scaling single cell transcriptomics through split pool barcoding. Preprint at *bioRxiv* https://doi.org/10.1101/105163 (2017).
65. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
66. Ramani, V. et al. Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017).
67. Mulqueen, R. M. et al. Scalable and efficient single-cell DNA methylation sequencing by combinatorial indexing. Preprint at *bioRxiv* https://doi.org/10.1101/157230 (2017).
68. Kang, H. M. et al. Multiplexed droplet single-cell RNA sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
69. Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
70. Mora-Castilla, S. et al. Miniaturization technologies for efficient single-cell library preparation for next-generation sequencing. *J. Lab Autom.* **21**, 557–567 (2016).
71. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
72. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
73. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
74. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
75. Trapnell, C. et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
76. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
77. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
78. Ilicic, T. et al. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).
79. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
80. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs—a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**, giy059 (2018).
81. Tian, L. et al. scPipe: a flexible R/Bioconductor preprocessing pipeline for single-cell RNA sequencing data. *PLoS Comput. Biol.* **10**, e1006361 (2018).
82. Azizi, E. et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Preprint at *bioRxiv* https://doi.org/10.1101/221994 (2018).
83. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* **14**, e1006245 (2018).
84. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* **11**, e1004333 (2015).
85. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
86. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
87. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
88. Bacher, R. et al. SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* **14**, 584–586 (2017).
89. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
90. Buttner, M., Miao, Z., Wolf, A., Teichmann, S. A. & Theis, F. J. Assessment of batch-correction methods for scRNA-seq data with a new test metric. Preprint at *bioRxiv* https://doi.org/10.1101/200345 (2017).
91. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
92. Ritchie, M. E. et al. limma powers differential expression analyses for RNA sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
93. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
94. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
95. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).
96. Li, W. V. & Li, J. J. scImpute: an accurate and robust imputation method for single-cell RNA-seq data. Preprint at *bioRxiv* https://doi.org/10.1101/141598 (2017).
97. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
98. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
99. Abdi, H. & Williams, L. J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2**, 433–459 (2010).
100. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
101. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
102. McInnes, L. & Healy, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
103. Becht, E. et al. Evaluation of UMAP as an alternative to t-SNE for single-cell data. Preprint at *bioRxiv* https://doi.org/10.1101/298430 (2018).
104. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).
105. Fan, J. et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13**, 241–244 (2016).
106. Guo, M., Wang, H., Potter, S. S., Whitsett, J. A. & Xu, Y. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput. Biol.* **11**, e1004575 (2015).

107. Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
108. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974–1980 (2015).
109. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
110. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
111. Korthauer, K. D. et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* **17**, 222 (2016).
112. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
113. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
114. Ji, Z. & Ji, H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e117 (2016).
115. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. Preprint at *bioRxiv* https://doi.org/10.1101/276907 (2018).
116. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
117. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across datasets. *Nat. Methods* **15**, 359–362 (2018).
118. Ji, N. & van Oudenaarden, A. Single molecule fluorescent in situ hybridization (smFISH) of *C. elegans* worms and embryos. *WormBook* http://www.wormbook.org/chapters/www_smFISH/smFISH.html (2012).
119. Moffitt, J. R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl Acad. Sci. USA* **113**, 11046–11051 (2016).
120. Ke, R., Mignardi, M., Hauling, T. & Nilsson, M. Fourth generation of next-generation sequencing technologies: promise and consequences. *Hum. Mutat.* **37**, 1363–1367 (2016).
121. Ke, R. et al. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).
122. Lein, E., Borm, L. E. & Linnarsson, S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**, 64–69 (2017).
123. Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
124. Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232 (2016).
125. Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
126. Liu, W. et al. Sample preparation method for isolation of single-cell types from mouse liver for proteomic studies. *Proteomics* **11**, 3556–3564 (2011).
127. Dorrell, C. et al. Surface markers for the murine oval cell response. *Hepatology* **48**, 1282–1291 (2008).
128. Su, X. et al. Single-cell RNA-seq analysis reveals dynamic trajectories during mouse liver development. *BMC Genomics* **18**, 946 (2017).
129. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
130. Chapman, H. A. et al. Integrin $\alpha_6\beta_4$ identifies an adult distal lung epithelial population with regenerative potential in mice. *J. Clin. Invest.* **121**, 2855–2862 (2011).
131. Xu, Y. et al. Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. *JCI Insight* **1**, e90558 (2016).
132. Joost, S. et al. Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. *Cell Syst.* **3**, 221–237 (2016).
133. Der, E. et al. Single cell RNA sequencing to dissect the molecular heterogeneity in lupus nephritis. *JCI Insight* **2**, e93009 (2017).
134. Autengruber, A., Gereke, M., Hansen, G., Hennig, C. & Bruder, D. Impact of enzymatic tissue disintegration on the level of surface molecule expression and immune cell function. *Eur. J. Microbiol. Immunol. (Bp.)* **2**, 112–120 (2012).
135. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
136. Glass, L. L. et al. Single-cell RNA sequencing reveals a distinct population of proglucagon-expressing cells specific to the mouse upper small intestine. *Mol. Metab.* **6**, 1296–1303 (2017).
137. Herring, C. A. et al. Unsupervised trajectory analysis of single-cell RNA-seq and imaging data reveals alternative tuft cell origins in the gut. *Cell Syst.* **6**, 37–51 (2018).
138. Merlos-Suárez, A. et al. The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* **8**, 511–524 (2011).
139. Wollny, D. et al. Single-cell analysis uncovers clonal acinar cell heterogeneity in the adult pancreas. *Dev. Cell* **39**, 289–301 (2016).
140. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360 (2016).
141. Segerstolpe, Å. et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).
142. Petersen, M. B. K. et al. Single-cell gene expression analysis of a human ESC model of pancreatic endocrine development reveals different paths to β-cell differentiation. *Stem. Cell Rep.* **9**, 1246–1261 (2017).
143. Muraro, M. J. et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* **3**, 385–394 (2016).
144. Li, D. et al. Complete disassociation of adult pancreas into viable single cells through cold trypsin-EDTA digestion. *J. Zhejiang Univ. Sci. B* **14**, 596–603 (2013).
145. Shekhar, K. et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323 (2016).
146. Daniszewski, M. et al. Single cell RNA sequencing of stem cell-derived retinal ganglion cells. *Sci. Data* **5**, 180013 (2018).
147. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
148. Carninci, P. et al. Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. *Proc. Natl Acad. Sci. USA* **95**, 520–524 (1998).
149. Spiess, A.-N. & Ivell, R. A highly efficient method for long-chain cDNA synthesis using trehalose and betaine. *Anal. Biochem.* **301**, 168–174 (2002).
150. Pinto, F. L. & Lindblad, P. A guide for in-house design of template-switch-based 5′ rapid amplification of cDNA ends systems. *Anal. Biochem.* **397**, 227–232 (2010).
151. Lambert, D. & Draper, D. E. Effects of osmolytes on RNA secondary and tertiary structure stabilities and RNA–Mg^{2+} interactions. *J. Mol. Biol.* **370**, 993–1005 (2007).

152. Tang, F. et al. mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).

153. Zajac, P., Islam, S., Hochgerner, H., Lönnerberg, P. & Linnarsson, S. Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. *PLoS One* **8**, e85270 (2013).

Acknowledgements

H.H. is a Miguel Servet (CP14/00229) researcher funded by the Spanish Institute of Health Carlos III (ISCIII). This work received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement no. H2020-MSCA-ITN-2015-675752 (SINGEK; A.L.) and the Ministerio de Ciencia, Innovación y Universidades (SAF2017-89109-P; AEI/FEDER, UE; H.H.). This project has been made possible in part by grant no. 2018-182827 (H.H.) from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. We thank ThePaperMill for critical reading and scientific editing services. Core funding was provided by the ISCIII and the Generalitat de Catalunya.

Author contributions

The authors contributed to the various sections of this tutorial as follows: A.L., Data processing and Data analysis; C.M., Sample preparation; S.P., Optimization (Box 1); H.H., Design, Sample preparation, Single-cell RNA sequencing, Further technical considerations and Future directions. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to H.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 16 November 2018

Chapter II

Single-cell study of dermal fibroblasts in aging

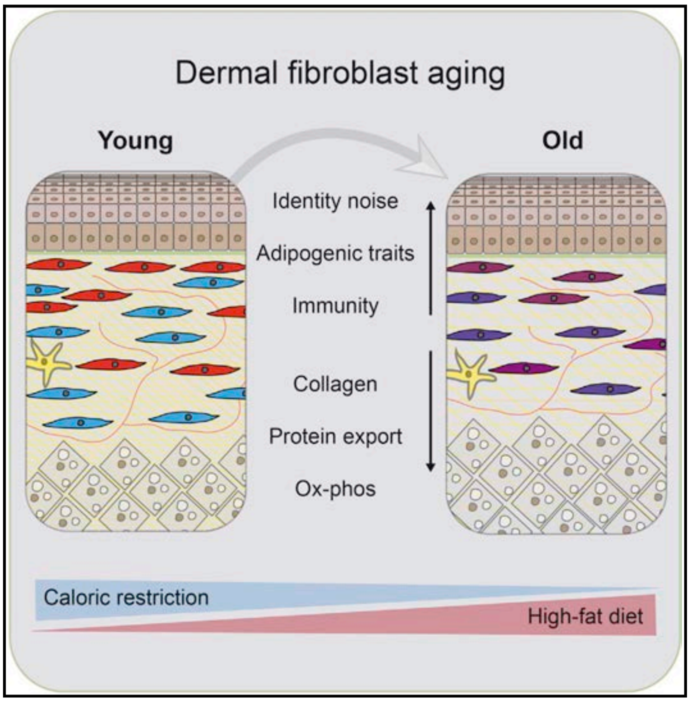
In collaboration with Salvador Aznar's research group at Institute for Research in Biomedicine (IRB), we studied dermal fibroblasts in aging. This project consists of many different experiments, for which I was responsible for the single-cell experimental design, data analysis and interpretation. Little was known whether the changes in stromal function during aging stems from changes in fibroblasts. Using population- and single-cell transcriptomics, as well as long-term lineage tracing, we studied whether murine dermal fibroblasts are altered during physiological aging under different dietary regimes that affect longevity. We tracked single-cell populations from newborn to young and further to old fibroblasts and compared the lost/gained subpopulation and cellular states. We identified specific gene markers for each of the subpopulation at each time point and assessed the functions associated to these genes. We showed that the identity of old fibroblasts becomes undefined, with the fibroblast states present in young skin no longer clearly demarcated. In addition, old fibroblasts not only reduce the expression of genes involved in the formation of the extracellular matrix, but also gain adipogenic traits, paradoxically becoming more similar to neonatal pro-adipogenic fibroblasts. This project is published in Cell [78].

Article

Cell

Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging

Graphical Abstract



Highlights

- The identity of old dermal fibroblasts becomes undefined and noisy
- Old dermal fibroblasts acquire adipogenic traits
- CR and HFD prevent and potentiate fibroblast aging, respectively
- Loss of cell identity is a possible mechanism underlying aging

Authors

Marion Claudia Salzer, Atefeh Lafzi, Antoni Berenguer-Llergo, ..., Juan Martín-Caballero, Holger Heyn, Salvador Aznar Benitah

Correspondence

holger.heyn@cnag.crg.eu (H.H.), salvador.aznar-benitah@irbbarcelona.org (S.A.B.)

In Brief

Single-cell transcriptomics and long-term lineage tracing outline aging-induced molecular identity changes of skin fibroblasts that lead them to acquire an adipogenic profile sensitive to whole-body metabolic changes.

Salzer et al., 2018, Cell *175*, 1–16
November 29, 2018 © 2018 Elsevier Inc.
https://doi.org/10.1016/j.cell.2018.10.012

CellPress

Please cite this article in press as: Salzer et al., Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging, Cell (2018), https://doi.org/10.1016/j.cell.2018.10.012

Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging

Marion Claudia Salzer,[1] Atefeh Lafzi,[3,6] Antoni Berenguer-Llergo,[1,6] Catrin Youssif,[1] Andrés Castellanos,[1] Guiomar Solanas,[1] Francisca Oliveira Peixoto,[1] Camille Stephan-Otto Attolini,[1] Neus Prats,[1] Mònica Aguilera,[1] Juan Martín-Caballero,[5] Holger Heyn,[3,4,*] and Salvador Aznar Benitah[1,2,7,*]
[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
[2]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
[3]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
[4]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[5]Director of PCB-PRBB Animal Facility Alliance
[6]These authors contributed equally
[7]Lead Contact
*Correspondence: holger.heyn@cnag.crg.eu (H.H.), salvador.aznar-benitah@irbbarcelona.org (S.A.B.)
https://doi.org/10.1016/j.cell.2018.10.012

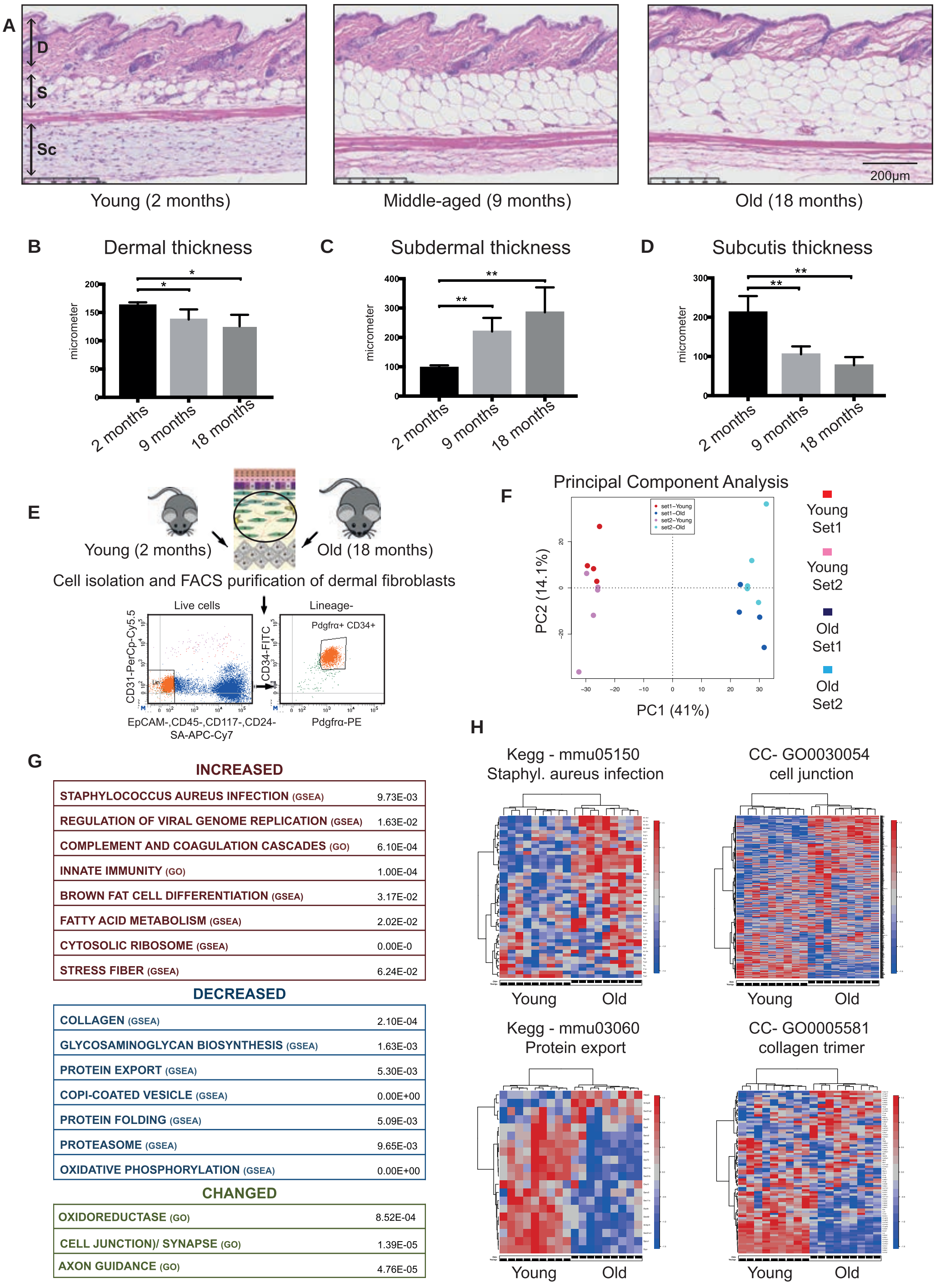
SUMMARY

During aging, stromal functions are thought to be impaired, but little is known whether this stems from changes of fibroblasts. Using population- and single-cell transcriptomics, as well as long-term lineage tracing, we studied whether murine dermal fibroblasts are altered during physiological aging under different dietary regimes that affect longevity. We show that the identity of old fibroblasts becomes undefined, with the fibroblast states present in young skin no longer clearly demarcated. In addition, old fibroblasts not only reduce the expression of genes involved in the formation of the extracellular matrix, but also gain adipogenic traits, paradoxically becoming more similar to neonatal pro-adipogenic fibroblasts. These alterations are sensitive to systemic metabolic changes: long-term caloric restriction reversibly prevents them, whereas a high-fat diet potentiates them. Our results therefore highlight loss of cell identity and the acquisition of adipogenic traits as a mechanism underlying cellular aging, which is influenced by systemic metabolism.

INTRODUCTION

Tissue function declines with age, impairing the ability of tissues to sustain daily homeostasis and repair damage. A major source of physiological tissue aging is the functional decay of adult stem cells through the cell-intrinsic accumulation of damage (such as DNA damage, loss of proteostasis, and oxidative damage). This is exemplified by their ineffectiveness at repopulating young tissues after transplantation, as well as their inability to grow in culture even in the presence of growth factors and nutrients (López-Otín et al., 2013; Price et al., 2014; Sun et al., 2014; Goodell and Rando, 2015). Interestingly, much of this cell-intrinsic damage is a consequence of the predominant function each specific type of stem cell performs during homeostasis; for instance, aged interfollicular epidermal stem cells predominantly accumulate DNA damage due to their continual proliferation, whereas muscle stem cells (which almost never divide) have problems in biomass recycling through autophagy as well as in sensing damage (García-Prat et al., 2016; Price et al., 2014; Sato et al., 2017; Solanas et al., 2017; Sousa-Victor et al., 2014). Other examples of stem cell–intrinsic malfunctions are accumulation of DNA damage and myeloid bias in aged hematopoietic stem cells, and lipid and NAD metabolic alterations in aged hepatocytes (Florian et al., 2013; Signer and Morrison, 2013; Flach et al., 2014; Sun et al., 2014; Xie et al., 2014; Sato et al., 2017).

In addition to the cell-intrinsic accumulation of stress in aged progenitors, other traits associated to aging occur in all tissues, such as increased inflammation and fibrosis, circadian rhythm reprogramming, and imbalances in oxidative phosphorylation and fatty acid metabolism. This suggests that these traits arise from local and systemic signals established during organismal aging (Doles et al., 2012; Florian et al., 2013; Keyes et al., 2013; Keyes and Fuchs, 2018; Loffredo et al., 2013; Matsumura et al., 2016; Neves et al., 2017; Oh, Lee, and Wagers, 2014; Sato et al., 2017; Solanas et al., 2017). Regarding local signals, there is increasing evidence that stem cells continuously engage in a reciprocal communication with surrounding stromal cells (Driskell and Watt, 2015; Gao, Xu, Asada, and Frenette, 2018; Sennett and Rendl, 2012). Examples can be found in the interaction of hematopoietic stem cells with their nearby vasculature, adrenergic nerves, and osteoblastic cells (Acar et al., 2015; Asada et al., 2017; Casanova-Acebes et al., 2013; Gao et al., 2018; Inra et al., 2015; Katayama et al., 2006; Lefrançais et al., 2017; Méndez-Ferrer et al., 2010; Morrison and Scadden, 2014; Zhou et al., 2017); in the local crosstalk between muscle stem cells and macrophages or mesenchymal cells (Gopinath and Rando, 2008; Mashinchian et al., 2018); and in the communication of lung stem cells with fibroblasts (Lee et al., 2017) and communication of hair follicle stem cells with adipogenic, nerve, arrector pili muscle cells, or dermal papilla fibroblasts (Brownell et al., 2011; Donati et al., 2014; Driskell et al., 2013; Festa et al., 2011; Fujiwara et al., 2011; Greco et al., 2009; Hill et al., 2013; Mastrogiannaki et al., 2016; Plikus et al., 2017; Rahmani et al.,



INCREASED

STAPHYLOCOCCUS AUREUS INFECTION (GSEA)	9.73E-03
REGULATION OF VIRAL GENOME REPLICATION (GSEA)	1.63E-02
COMPLEMENT AND COAGULATION CASCADES (GO)	6.10E-04
INNATE IMMUNITY (GO)	1.00E-04
BROWN FAT CELL DIFFERENTIATION (GSEA)	3.17E-02
FATTY ACID METABOLISM (GSEA)	2.02E-02
CYTOSOLIC RIBOSOME (GSEA)	0.00E-0
STRESS FIBER (GSEA)	6.24E-02

DECREASED

COLLAGEN (GSEA)	2.10E-04
GLYCOSAMINOGLYCAN BIOSYNTHESIS (GSEA)	1.63E-03
PROTEIN EXPORT (GSEA)	5.30E-03
COPI-COATED VESICLE (GSEA)	0.00E+00
PROTEIN FOLDING (GSEA)	5.09E-03
PROTEASOME (GSEA)	9.65E-03
OXIDATIVE PHOSPHORYLATION (GSEA)	0.00E+00

CHANGED

OXIDOREDUCTASE (GO)	8.52E-04
CELL JUNCTION)/ SYNAPSE (GO)	1.39E-05
AXON GUIDANCE (GO)	4.76E-05

(legend on next page)

Please cite this article in press as: Salzer et al., Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging, Cell (2018), https://doi.org/10.1016/j.cell.2018.10.012

2014; Rendl et al., 2005; Rendl et al., 2008; Rivera-Gonzalez et al., 2016; Rognoni et al., 2016; Sennett et al., 2015; Telerman et al., 2017; Wojciechowicz et al., 2013; Yang et al., 2017; Zhang et al., 2016). The progressive functional decay of stromal cells is believed to contribute significantly to the inability of tissues to sustain homeostasis and to properly respond to injury during aging. Nonetheless, whether and how stromal cells are altered during aging is unknown in most tissues (Kusumbe et al., 2016; Neves et al., 2017; Stearns-Reider et al., 2017).

Stromal changes are particularly apparent in aged human and mouse dermis. These include decreased dermal thickness and ECM density, and reduced numbers of fibroblasts (Demaria et al., 2015; Harbor and King, 2014; Figures 1A, 1B, S5A, and S5B). These dermal changes contribute to the cosmetic consequences of having reduced skin turgor and increased wrinkling and perhaps also contribute to the increased propensity of aged skin for infections, tumorigenesis, and inefficient wound healing (Driskell and Watt, 2015; Kaur et al., 2016). In the three-dimensional tissue stroma scaffold, fibroblasts play a major role in maintaining the extracellular matrix (ECM) and in heterotypic signaling with epithelial cells during steady-state and injury response. However, a detailed characterization of the cellular and molecular traits of old dermal fibroblasts is lacking. Thus, whether dermal aging is related to a loss of fibroblasts, to acquired alterations in the remaining old fibroblasts, or to both is not known.

RESULTS

In order to study fibroblast aging, we purified dermal fibroblasts by fluorescence-activated cell sorting (FACS) of cells double-positive for the PDGF receptor alpha (Pdgfrα+) and CD34+ from the dermis of young (2 months) and old (18 months) female mice. We determined by immunohistochemistry that the vast majority of fibroblastic dermal cells, but not the α-smooth muscle actin+ arrector pili muscle, were positive for both CD34 and Pdgfrα (Figures S1A and S1B). In addition, Pdgfrα+/CD34+ fibroblasts comprised about 90%–95% of the cells within the dermis, as was evident once pre-adipocytes (CD24+), epithelial cells (EpCAM+), melanocytes (CD117+), hematopoietic cells (CD45+), and endothelial cells (CD31+) were gated out of the single-cell preparations (Figures 1E and S2A). Thus, our purified cells account for the majority of fibroblasts that are embedded in murine dermis. After FACS purification, we compared old and young dermal fibroblasts by whole-transcriptome analysis (Figures 1E and 1F). The old mice had clear histological signs of dermal alterations, such as overall decreased thickness and density of the dermis but with a significantly increased thickness of the adipocyte-containing lower dermal layer (Figures 1A–1D). Interestingly, most of these changes were already visible in middle-aged mice (9-month-old; Figures 1A–1D).

Principal component analysis (PCA) of the transcriptome data indicated that young and old fibroblasts primarily clustered by age and that old fibroblasts expressed approximately 1,000 transcripts that differed from their young counterparts with a fold change of > 1.5 (FDR < 0.05) (Figure 1F, Table S1). Gene ontology (GO) and gene set enrichment analysis (GSEA) of the data indicated that old fibroblasts had a strong reduction in the expression of the main extracellular matrix genes, including collagens and glycosaminogycans, and of genes involved in their secretion (i.e., Golgi, endoplasmic reticulum, and vesicle-mediated transport) (Figure 1G, Table S1). Concomitantly, old fibroblasts showed upregulation of genes involved in inflammation, innate immunity, the formation of stress fibers, and differentially expressed genes related to the establishment of cell contacts (Figures 1G, Table S1). Interestingly, they also had upregulation of a significant number of genes related to adipogenesis, lipid metabolism, and fat cell differentiation (Figure 1G, Table S1). These genes related to signaling downstream of master regulators of adipogenesis, such as PPARγ, VLDLR, and LDLR (Figure 2A; Table S1). Importantly, the genes defining each GO category clustered young and old samples in an unsupervised manner, strongly supporting their fundamental role in determining the traits associated with dermal fibroblast aging (Figures 1H, 2B–2D, and S3).

We confirmed the differential expression of some selected genes by RT-qPCR and by immunohistochemistry in samples obtained from independent biological replicates (Figures S2B and S4). Furthermore, we obtained very similar results when performing transcriptome analysis of old and young fibroblasts isolated from male mice (Figures S5A–S5E, Table S2).

Our comparative transcriptome results indicated that old fibroblasts lose one of their main defining characteristics—that of the production and secretion of extracellular matrix (ECM)

Figure 1. Hallmarks of Dermal Fibroblast Aging

(A) Hematoxylin and eosin (H&E)-stained sections of young (2-month-old), middle-aged (9-month-old), and old (18-month-old) murine skin. D, dermis; S, Subdermis; Sc, Subcutis.

(B–D) Quantification of dermal thickness (B), subdermal thickness (C), and subcutaneous thickness (D) at different ages using H&E-stained murine skin sections. Data are represented as mean ± SD.

(E) Experimental set-up for the isolation of dermal fibroblasts. After tissue digestion, FACS was used to select cells that were negative for the lineage markers CD31 (endothelial marker), CD45 (immune cell marker), CD24 (pre-adipocyte marker), EpCAM (epithelial marker), and CD117 (melanocyte marker), and positive for the fibroblast markers Pdgfrα and CD34.

(F) Principal component analysis (PCA) of 9 young and 9 old dermal fibroblast samples, according to their transcriptome. Samples were combined from two independent experiments (see STAR Methods section Quantification and Statistical Analysis: Microarrays).

(G) Gene set enrichment analysis (GSEA) and gene ontology (GO) analysis of genes differentially expressed between young and old dermal fibroblasts. Categories derived from GO analysis using David 6.8 are marked with "(GO)," categories derived from GSEA are marked with "(GSEA)." For the GSEA categories the FDR is indicated; for the GO categories, the p value is indicated.

(H) Unsupervised clustering of young and old fibroblast samples based on gene ontology signatures. The intensity of the colors represents the expression intensity from blue (low expression) to red (high expression).

See also Figure S3.

A

Signal Transduction Pathway Associations (up in old)

LOW DENSITY LIPOPROTEIN RECEPTOR RELATED PROTEIN	1.06E-05
VERY LOW DENSITY LIPOPROTEIN RECEPTOR RELATED PROTEIN	1.56E-02
PEROXISOME PROLIFERATOR ACTIVATED RECEPTOR GAMMA	9.35E-03
LEPTIN	3.03E-02

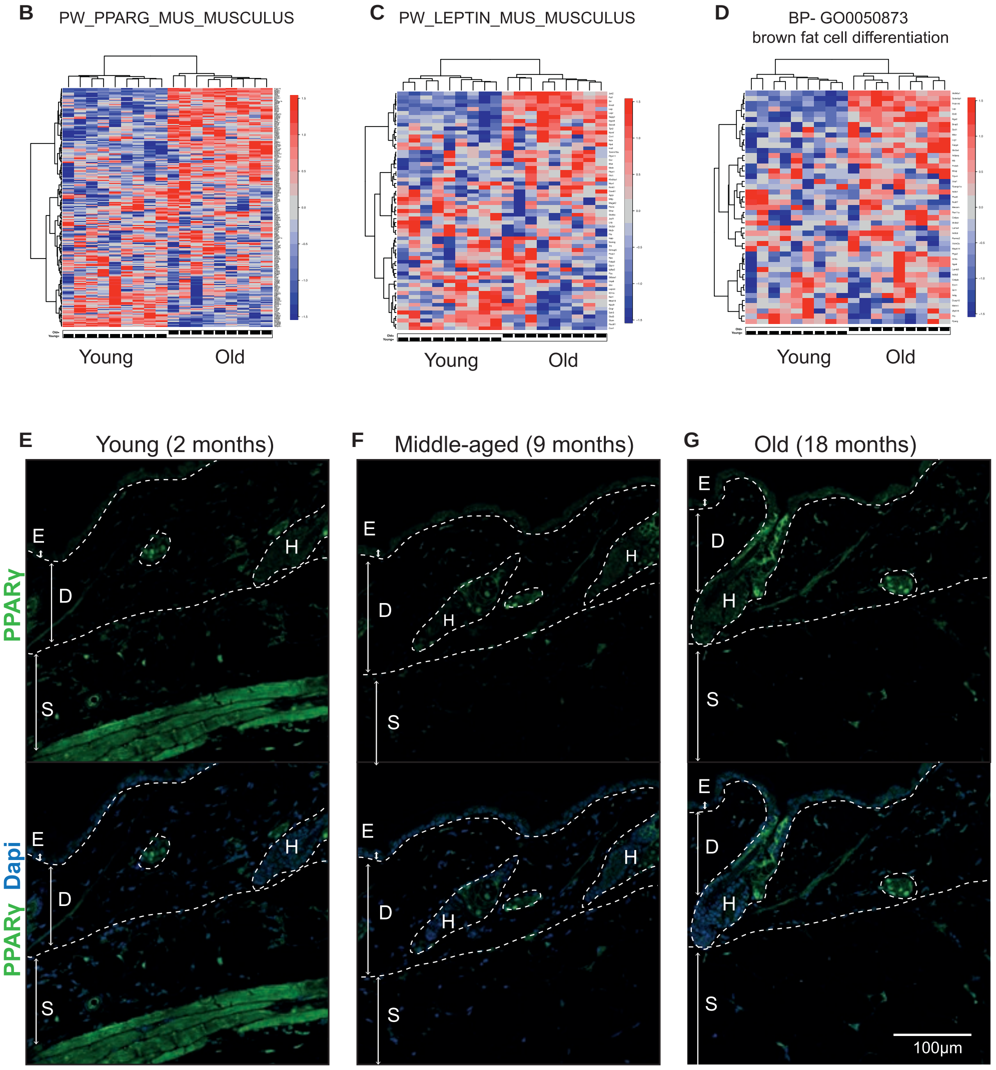


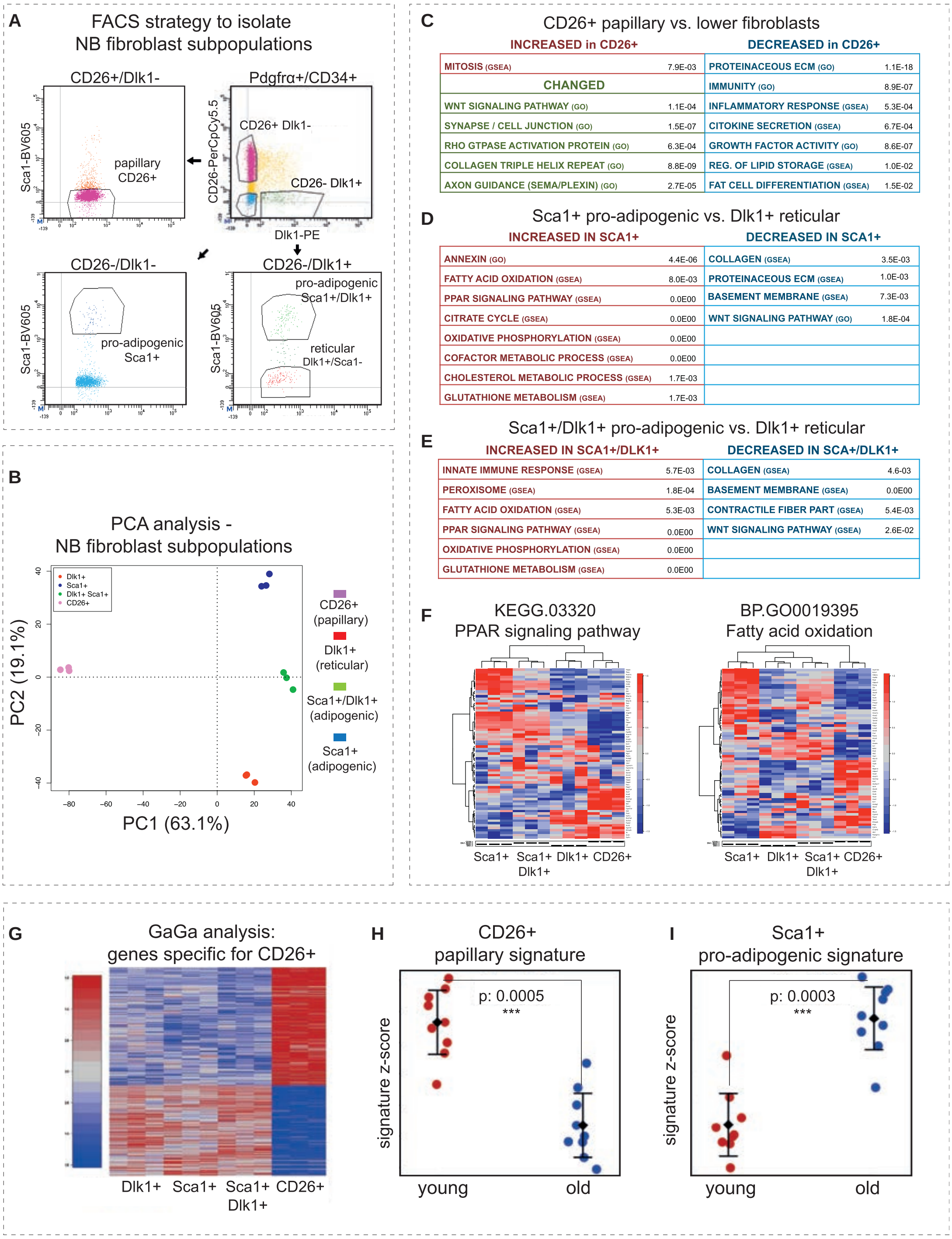
Figure 2. Old Dermal Fibroblasts Are Positive for the Transcription Factor PPARγ Irrespective of their Location

(A) Signal transduction pathway association analysis of young compared to old dermal fibroblasts using Genomatix software.

(B–D) Unsupervised clustering of young and old fibroblast samples based on genomatix and gene ontology signatures. The intensity of the colors represents the expression intensity from blue (low expression) to red (high expression).

(E–G) Immunofluorescent images showing PPARγ in green and DAPI-stained nuclei in blue in skin sections from young (A), middle-aged (B) and old (C) mice. E, epidermis; D, dermis; S, subdermis; H, hair follicle.

Please cite this article in press as: Salzer et al., Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging, Cell (2018), https://doi.org/10.1016/j.cell.2018.10.012



CD26+ papillary vs. lower fibroblasts

INCREASED in CD26+		DECREASED in CD26+	
MITOSIS (GSEA)	7.9E-03	PROTEINACEOUS ECM (GO)	1.1E-18
CHANGED		IMMUNITY (GO)	8.9E-07
WNT SIGNALING PATHWAY (GO)	1.1E-04	INFLAMMATORY RESPONSE (GSEA)	5.3E-04
SYNAPSE / CELL JUNCTION (GO)	1.5E-07	CITOKINE SECRETION (GSEA)	6.7E-04
RHO GTPASE ACTIVATION PROTEIN (GO)	6.3E-04	GROWTH FACTOR ACTIVITY (GO)	8.6E-07
COLLAGEN TRIPLE HELIX REPEAT (GO)	8.8E-09	REG. OF LIPID STORAGE (GSEA)	1.0E-02
AXON GUIDANCE (SEMA/PLEXIN) (GO)	2.7E-05	FAT CELL DIFFERENTIATION (GSEA)	1.5E-02

Sca1+ pro-adipogenic vs. Dlk1+ reticular

INCREASED IN SCA1+		DECREASED IN SCA1+	
ANNEXIN (GO)	4.4E-06	COLLAGEN (GSEA)	3.5E-03
FATTY ACID OXIDATION (GSEA)	8.0E-03	PROTEINACEOUS ECM (GSEA)	1.0E-03
PPAR SIGNALING PATHWAY (GSEA)	0.0E00	BASEMENT MEMBRANE (GSEA)	7.3E-03
CITRATE CYCLE (GSEA)	0.0E00	WNT SIGNALING PATHWAY (GO)	1.8E-04
OXIDATIVE PHOSPHORYLATION (GSEA)	0.0E00		
COFACTOR METABOLIC PROCESS (GSEA)	0.0E00		
CHOLESTEROL METABOLIC PROCESS (GSEA)	1.7E-03		
GLUTATHIONE METABOLISM (GSEA)	1.7E-03		

Sca1+/Dlk1+ pro-adipogenic vs. Dlk1+ reticular

INCREASED IN SCA1+/DLK1+		DECREASED IN SCA+/DLK1+	
INNATE IMMUNE RESPONSE (GSEA)	5.7E-03	COLLAGEN (GSEA)	4.6-03
PEROXISOME (GSEA)	1.8E-04	BASEMENT MEMBRANE (GSEA)	0.0E00
FATTY ACID OXIDATION (GSEA)	5.3E-03	CONTRACTILE FIBER PART (GSEA)	5.4E-03
PPAR SIGNALING PATHWAY (GSEA)	0.0E00	WNT SIGNALING PATHWAY (GSEA)	2.6E-02
OXIDATIVE PHOSPHORYLATION (GSEA)	0.0E00		
GLUTATHIONE METABOLISM (GSEA)	0.0E00		

(legend on next page)

components—with a concomitant upregulated expression of genes involved in inflammation as well as, intriguingly, in lipid metabolism and adipogenesis. Interestingly, decreased production of ECM components and increased lipid metabolism and inflammation are features that naturally distinguish the adipogenic mesenchymal fate from the fibroblast state. Thus, our results led us to hypothesize that dermal fibroblast aging is perhaps associated with the acquisition of pro-adipogenic traits at the expense of fibroblast characteristics. To test this hypothesis, however, we first needed to take several considerations into account. Previous work has shown that the dermis of newborn mice contains four types of fibroblasts located at specific dermal sites: CD26+ papillary fibroblasts that are in close proximity with basal epidermal progenitors in the upper dermis; Dlk1+ reticular fibroblasts that extend throughout the ECM-dense lower region of the dermis; and two additional pro-adipogenic types of fibroblasts, identified as Sca1+/Dlk1– and Sca1+/Dlk1+, which are located in the lower reticular dermis (Driskell et al., 2013). However, as mice reach adulthood, the differential cell surface features that allow these four types of newborn dermal fibroblasts to be distinguished and isolated are lost, which thus prevents analysis of whether the adult dermis still contains these lineages (Driskell et al., 2013; Rinkevich et al., 2015). In addition, already at one month after birth, more than 80% of adult dermal fibroblasts are CD26+ and no longer show any apparent regional distribution, occupying the entire papillary (upper) and reticular (lower) dermal areas (Rinkevich et al., 2015). Hence, it is not clear whether this population of CD26+ adult dermal fibroblasts is homogeneous or still contains papillary, reticular, and pro-adipogenic fibroblasts. Since our comparative transcriptome analysis was based on isolating the majority of fibroblasts spanning the entire upper and lower dermal areas of adult mice (Driskell et al., 2013, Figure S1A), we could not rule out that the over-representation of pro-adipogenic transcripts in old fibroblasts is the consequence of a persistence or expansion of putative pro-adipogenic fibroblasts at the expense of upper papillary fibroblasts or because adult dermal fibroblasts in general (irrespective of location or lineage) acquire pro-adipogenic characteristics.

We took several experimental approaches to discriminate between these possibilities. First, we tested whether old fibroblasts are more predisposed to differentiate into adipocytes (or adipocyte-like cells) in culture. For this, we first isolated dermal fibroblasts from four young (2-month-old) and four old (18-month-old) mice and placed them in culture using standard fibroblast culture conditions. However, after only one passage, fibroblasts in culture showed strong alterations in the expression of many genes involved in cell cycle, metabolism, ECM production, and immunity, irrespective of their age (Figure S5G; Table S3). Strikingly, young and old fibroblasts became indistinguishable from each other (i.e., they clustered together), and the differential expression of most genes that we identified in our transcriptome analysis to be associated with aging *in vivo* were lost (Figure S5F; Table S3). Moreover, the difference in clonogenic potential of young and old fibroblasts observed immediately after placing them in culture was equalized by one passage (Figure S5H). Thus, dermal fibroblasts dramatically alter their transcriptome in culture and do not retain their corresponding *in vivo* age, making it difficult to reach any conclusion about their propensity to differentiate along the adipogenic lineage using *in vitro* experiments and raising caution about using 2D cell culture systems to study aging.

We therefore undertook several *in vivo* approaches to better distinguish between the possibilities described above. First, our comparative gene expression analysis indicated that signaling downstream of the master regulator of adipogenesis, PPARγ, is enhanced in old fibroblasts (Figures 2A–2D). Accordingly, we observed that the majority of old fibroblasts express higher levels of PPARγ than young fibroblasts, as determined by immunohistochemistry of skin sections from young and old mice (Figures 2E–2G). Importantly, old fibroblasts become PPARγ+ irrespective of their location within the dermis and, thus, span the entire papillary and reticular regions. On the other hand, in young skin, nuclear expression of PPARγ is seen in sebocytes and subcutaneous adipocytes, as expected, but is barely detected in dermal fibroblasts (Figure 2E). We verified that the majority of PPARγ+ dermal cells correspond to Pdgfrα+ and CD34+ fibroblasts, but not hematopoietic (CD45) or smooth muscle cells (α-Sma+) (Figures S1A–S1C). Thus, these results further suggest that old fibroblasts acquire adipogenic features irrespective of their location within the dermis.

Figure 3. Old Dermal Fibroblasts Gain Similarities to Newborn Pro-adipogenic Fibroblasts

(A) FACS strategy to isolate newborn (P2) fibroblast subpopulations (Driskell et al., 2013). Lineage-negative (CD45–, CD31–, EpCAM–, CD117–) and Pdgfrα+/CD34+ fibroblasts were further subdivided based on their expression of CD26, Dlk1, and Sca1: papillary fibroblasts were CD26+ and Sca1–; reticular fibroblasts CD26–, Dlk1+, and Sca1–; and pro-adipogenic fibroblasts either Sca1+/Dlk1+ or Sca1+/Dlk1–.

(B) PCA of CD26+ papillary, Dlk1+ reticular, Sca1+/Dlk1+ pro-adipogenic, or Sca1+/Dlk1– pro-adipogenic fibroblast samples according to their transcriptome. All four fibroblast populations were isolated from three newborn female littermates.

(C–E) GSEA and GO analysis of genes differentially expressed in newborn fibroblast subpopulations. Comparisons were made between CD26+ papillary fibroblasts and the other three lower fibroblast populations (C), Sca1+ pro-adipogenic and Dlk1+ reticular fibroblasts (D), and Sca1+/Dlk1+ pro-adipogenic and Dlk1+ reticular fibroblasts (E).

(F) Unsupervised clustering of newborn fibroblast samples based on gene ontology signatures. The intensity of the colors represents the expression intensity from blue (low expression) to red (high expression).

(G) Genes specific for a particular fibroblast subpopulation were selected using the GaGa algorithm (see STAR Methods and Figure S6B). The heatmap shows genes specific for CD26+ papillary fibroblasts.

(H and I) The gene expression profile of old dermal fibroblasts negatively associates with the signature of CD26+ papillary fibroblasts (H) and positively associates with the signature of Sca1+ pro-adipogenic fibroblasts (I). Expression values of genes specific for a fibroblast subpopulation were summarized (signature *Z* score) and compared across old and young samples. Each dot represents one old or young sample. Error bars represent 95% confidence interval.

See also Figure S6.

Please cite this article in press as: Salzer et al., Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging, Cell (2018), https://doi.org/10.1016/j.cell.2018.10.012

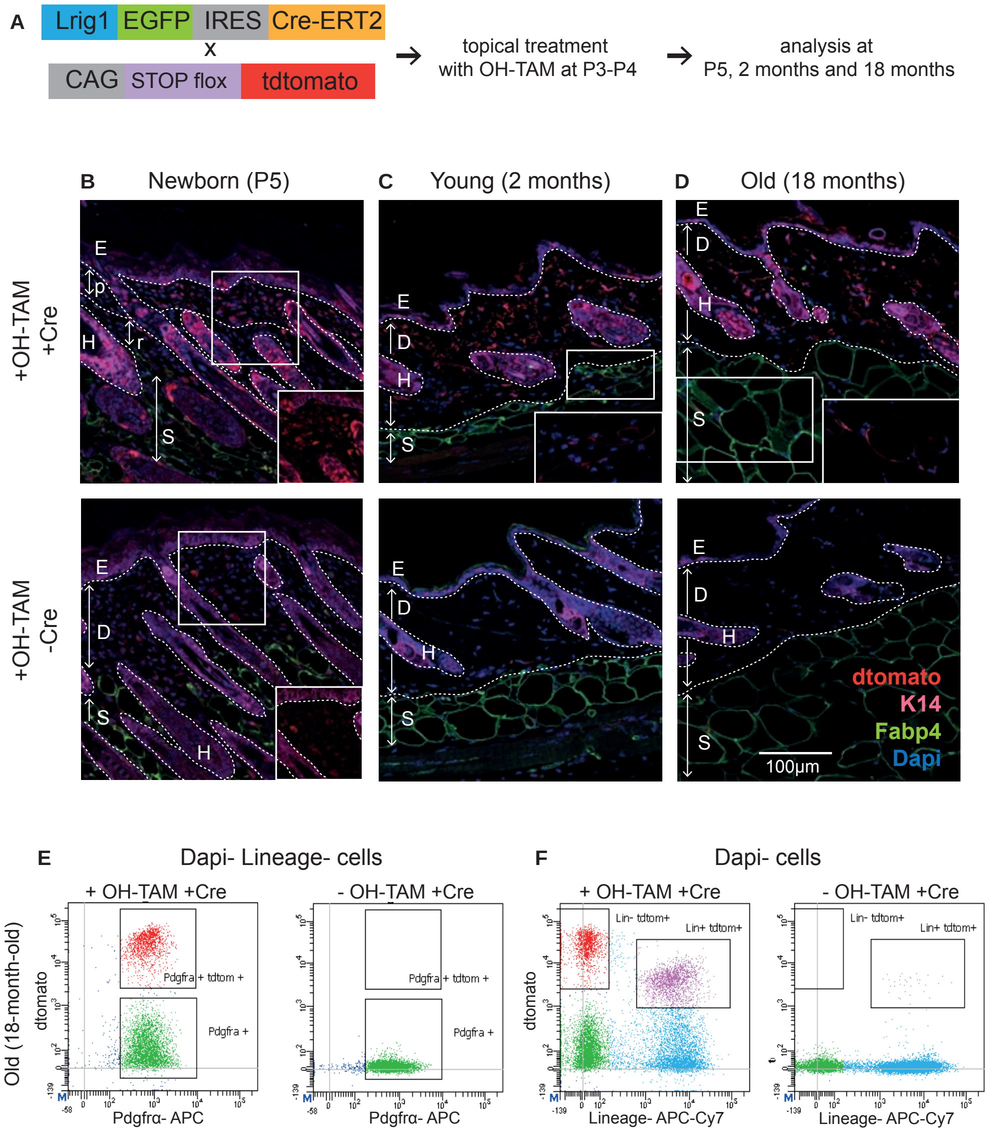


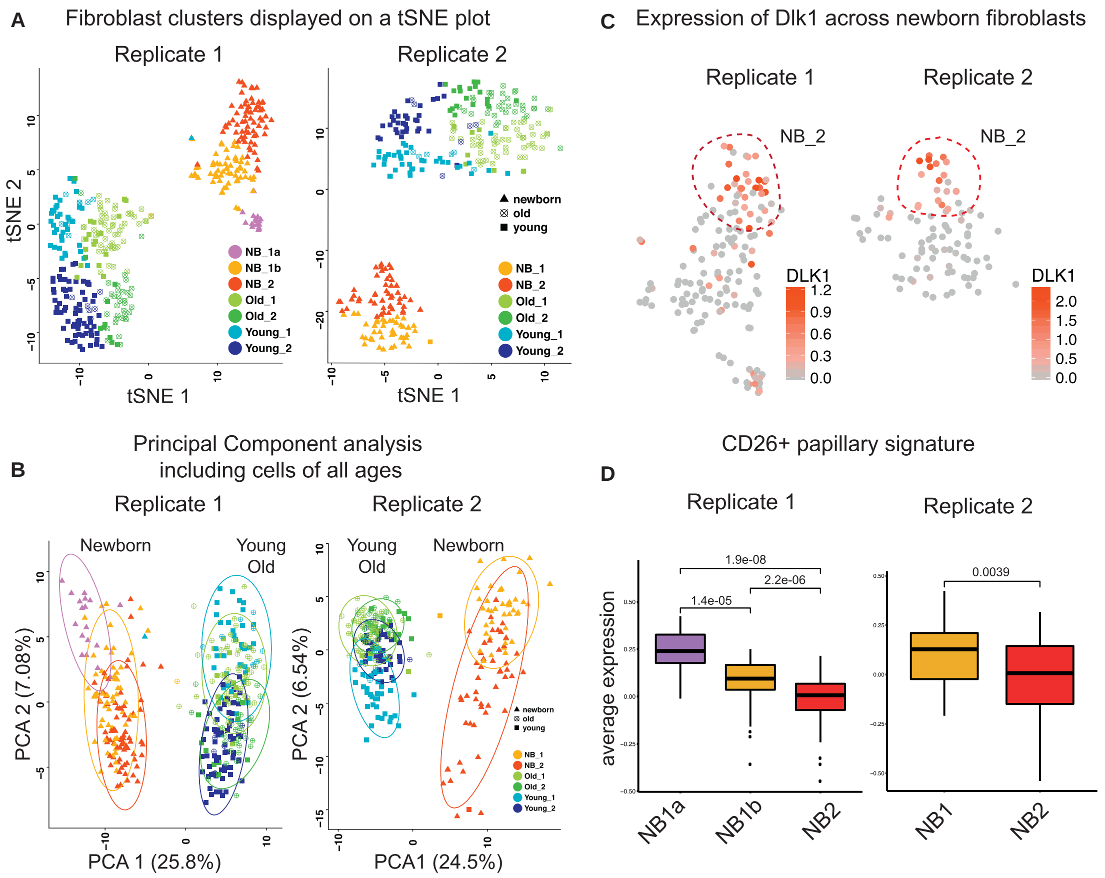
Figure 4. Progeny of Newborn Lrig1+ Papillary Fibroblasts Are Preserved during Aging

(A) 3- to 4-day-old Lrig1-EGFP-IRES-CreERT2/Rosa26-CAG-STOPflox-tdTomato mice were treated topically with 4-hydroxytamoxifen and analyzed at P5, 2 months, and 18 months of age.

(B–D) Immunofluorescent skin images showing DAPI-stained nuclei (blue), keratin 14 (for epithelial stem cells; pink), dTomato (red), and Fabp4 (adipocytes; green) from Lrig1-EGFP-IRES-CreERT2+/Rosa26-CAG-STOPflox-tdTomato+ mice (upper panel) or Lrig1-EGFP-IRES-CreERT2-/Rosa26-CAG-STOPflox-tdTomato+ mice (lower panel) treated with 4-hydroxytamoxifen at P5 (B), 2 months (C), and 18 months of age (D) pap, papillary layer; ret, reticular layer; D, dermis; S, subdermis; H, hair follicle.

(E) FACS plots showing Dapi-, Lin- (CD31, EpCam, CD24, CD117, CD45-), Pdgfrα+ cells of 18-month-old Lrig1-CreERT2+/dtomato+ mice treated with 4-hydroxytamoxifen (left) or not treated with 4-hydroxytamoxifen (right) at P3-P4. No recombination is observed in Pdgfrα + Lin- cells in untreated mice.

(F) FACS plots showing Dapi- cells of 18-month-old Lrig1-CreERT2+/dtomato+ mice treated with 4-hydroxytamoxifen (left) or not treated with 4-hydroxytamoxifen (right) at P3-P4. No spontaneous recombination is observed in Lin- cells, whereas some is observed in Lin+ cells, in untreated mice.



E

Gene Ontology analysis of genes differentially expressed between young clusters

Young 1		Young 2	
Oxidation-reduction process LDHA, UGDH, AKR1C14, SOD3, SOD2, GPX1, PIR, HMOX1, GPX3, HSD11B1, IDH1, AOX3, BDH2, PDHA1,	5.32E-04	**Extracellular matrix** LTBP2, LUM, COL3A1, ELN, SPARC, MMP3, COL16A1, CPZ, BGN, KAZALD1, COL1A2, MFAP2, COL1A1, MFAP4	3.88E-13
Glutathione metabolic process HAGH, GPX1, GPX3, IDH1, GSTT1, MGST1, SOD2	3.69E-04	**Collagen fibril organization** LUM, COL3A1, COL1A2, COL1A1	3.69E-04
Innate immunity IFIH1, MYD88, ZC3HAV1, C4B, C3, IFITM3, ANXA1, SAMHD1, SERPING1, LBP, C1S1	9.57E-05	**Insulin-like growth factor binding protein** RSPO1, LTBP2, KAZALD1, IGFBP7, IGFBP2	9.54E-04
Intracellular ribonucleoprotein complex RALY, MRPL2, ACTB, HNRNPA2B1, EEF2, RPS2, DDX5, HNRNPU, RPL18A, GSK3B, RPLP1, RPS15, NOP58, RPL4, PABPC1	6.76E-06		
Spliceosomal complex RALY, SF3B1, HNRNPA2B1, ALYREF, PABPC1, DDX5, HNRNPU, API5	5.51E-04		

(legend on next page)

Please cite this article in press as: Salzer et al., Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging, Cell (2018), https://doi.org/10.1016/j.cell.2018.10.012

We next hypothesized that old fibroblasts that have acquired adipogenic traits might share lineage similarities to the pro-adipogenic fibroblasts that were previously identified in the dermis of newborn mice (Driskell et al., 2013). We first determined the transcriptomes of CD26+ upper papillary fibroblasts, Dlk1+ reticular fibroblasts, and Sca1+/ Dlk1+ and Sca1+/ Dlk1– pro-adipogenic fibroblasts, which were isolated from the dermis of 2-day-old newborn mice (Figures 3A and 3B; Table S4). We then performed the GaGa analysis, which ascribes a transcriptome signature to each population by only including transcripts that are uniquely up- or downregulated in each population with respect to the three others (Figures 3G and S6B; Table S4). This approach indicated that newborn reticular and pro-adipogenic fibroblasts are more similar to each other than to papillary fibroblasts (Figure 3G). An unbiased PCA analysis of the data reached the same conclusion (Figure 3B). Gene ontology analysis unveiled interesting putative biological differences among the different types of newborn fibroblasts (Figures 3C–3F and S6A; Table S4). For instance, one distinguishing transcriptome feature of papillary fibroblasts is related to the formation of cell junctions as well as to cell division and Wnt signaling; this suggests that these upper dermal fibroblasts are likely prone to establishing cell contacts and engaging in signal transduction with neighboring cells (Figure 3C; Table S4). On the other hand, the three types of lower dermal fibroblasts (i.e., reticular and the two pro-adipogenic ones) expressed more transcripts related to the extracellular matrix, suggesting that one of their main functions is to establish the dense network of ECM proteins that is characteristic of the reticular dermis (Figure 3C; Table S4). Interestingly, lower dermal fibroblasts also expressed more genes involved in innate immunity than did papillary fibroblasts (Figure 3C; Table S4).

To dissect which type of lower dermal fibroblast predominantly contributed to each of these distinguishing features, we next compared their transcriptomes among each other and independently of the papillary signature. This analysis revealed that Dlk1+ reticular cells predominantly expressed ECM genes, whereas the transcriptomes of both types of pro-adipogenic fibroblasts had an overrepresentation of genes involved in fatty acid oxidation, glutathione metabolism, and oxidative phosphorylation (Figures 3D and 3E). Importantly, the genes defining each of these GO categories clustered each type of fibroblast in an unsupervised manner, indicating that they describe important features defining their cellular state (Figures 3F and S6A). Comparing the transcriptome data of young and old fibroblasts described in Figure 1 to these four signatures of newborn fibroblast revealed that old fibroblasts share less features with the newborn papillary signature than young (Figures 3H and S6C). Hence, old fibroblasts are overall more similar to newborn lower dermal fibroblasts (Figures 3H and S6C). Importantly, however, this similarity stemmed primarily from their significant relationship to Sca1+/ Dlk1– pro-adipogenic fibroblasts (Figure 3I).

Altogether, the comparative transcriptomic results between newborn and old fibroblasts (Figure 3) and the localization of PPARγ+ fibroblasts throughout the old dermis (Figure 2) strongly support our hypothesis that old fibroblasts acquire pro-adipogenic traits. These results, however, do not preclude the possibility that the over-representation of pro-adipogenic traits in old fibroblasts is a consequence of losing the upper dermal papillary lineage as the skin ages. To verify this, we undertook two additional approaches: *in vivo* lineage-tracing and single-cell RNA sequencing (RNA-seq) studies. First, we followed the fate of the progeny of papillary fibroblasts from newborn to old mice by crossing Lrig1-CreERT2 mice with ROSA26-STOPflox-dTomato mice (Page et al., 2013). Previous work has shown that Lrig1 is expressed exclusively by papillary fibroblasts in newborn dermis (Driskell et al., 2013). We therefore topically treated 3- to 4-day-old mice with 4-hydroxytamoxifen to permanently tag newborn papillary fibroblasts with the fluorescent protein dTomato (Figures 4A and 4B). We confirmed that the progeny of these papillary fibroblasts remained mainly in the upper dermal region in P5 mice (Figure 4B). By 2 months of age, however, dTomato+ cells were not only located in the upper dermis but were now also visible in the lower reticular region and, interestingly, even within some mature adipocytes below the dermis (Figure 4C). Importantly, in old mice, the progeny of dTomato+ papillary fibroblasts (tagged at the newborn stage) was distributed in the same manner as in young dermis throughout the upper and lower dermal regions (Figure 4D). We did not observe any spontaneous CreERT2-mediated recombination in dermal fibroblasts during the entire experiment (Figures 4E and 4F). These results therefore indicate that the progeny of the upper papillary fibroblasts present in newborn dermis contributes to all dermal layers, even including mature adipocytes, in adulthood. In addition, these observations make it highly unlikely that cell loss in the upper dermal region during aging accounts for the over-representation of adipogenic traits identified in our comparative analysis of the transcriptome of old fibroblasts.

To further characterize the molecular and cellular changes that occur in dermal fibroblasts during aging in an unbiased manner, we performed single-cell RNA-seq of Pdgfrα+/CD34+/Lin– fibroblasts isolated from the dermis of newborn (P1.5–2.5), young (2-month-old), and old (18-month-old) mice. RNA was sequenced using the Smart-seq2 protocol from

Figure 5. Single-Cell RNA-seq Reveals Two Distinct Fibroblast Subpopulations in Newborn, Young, and Old Dermis

(A) Unbiased clustering of transcriptomes of individual dermal fibroblasts isolated from newborn (P1.5–2.5), young (2-month-old) and old (18-month-old) mice from two independent biological replicates termed Replicate 1 (left) and Replicate 2 (right). Each cell is represented as a dot, assigned to a cluster by a clustering algorithm, and plotted on the t-SNE graph.

(B) PCA of newborn, young, and old cells from replicate 1 (left) and replicate 2 (right).

(C and D) Newborn cluster 2 (NB2) from replicate 1 (left) and replicate 2 (right) is more similar to reticular fibroblasts described by Driskell et al. (2013). (C) The reticular fibroblast marker Dlk1 is highly expressed in the majority of NB2 cells. Red circle indicates cluster NB2. (D) Boxplot showing that the CD26+ papillary signature is underrepresented in cluster NB2 as compared to NB1a and NB1b (replicate 1) and NB1 (replicate 2). Each dot represents the average expression of all genes forming the CD26+ papillary signature of one cell.

(E) GO analysis of marker genes identified for the two young clusters in replicate 1 and/or 2.

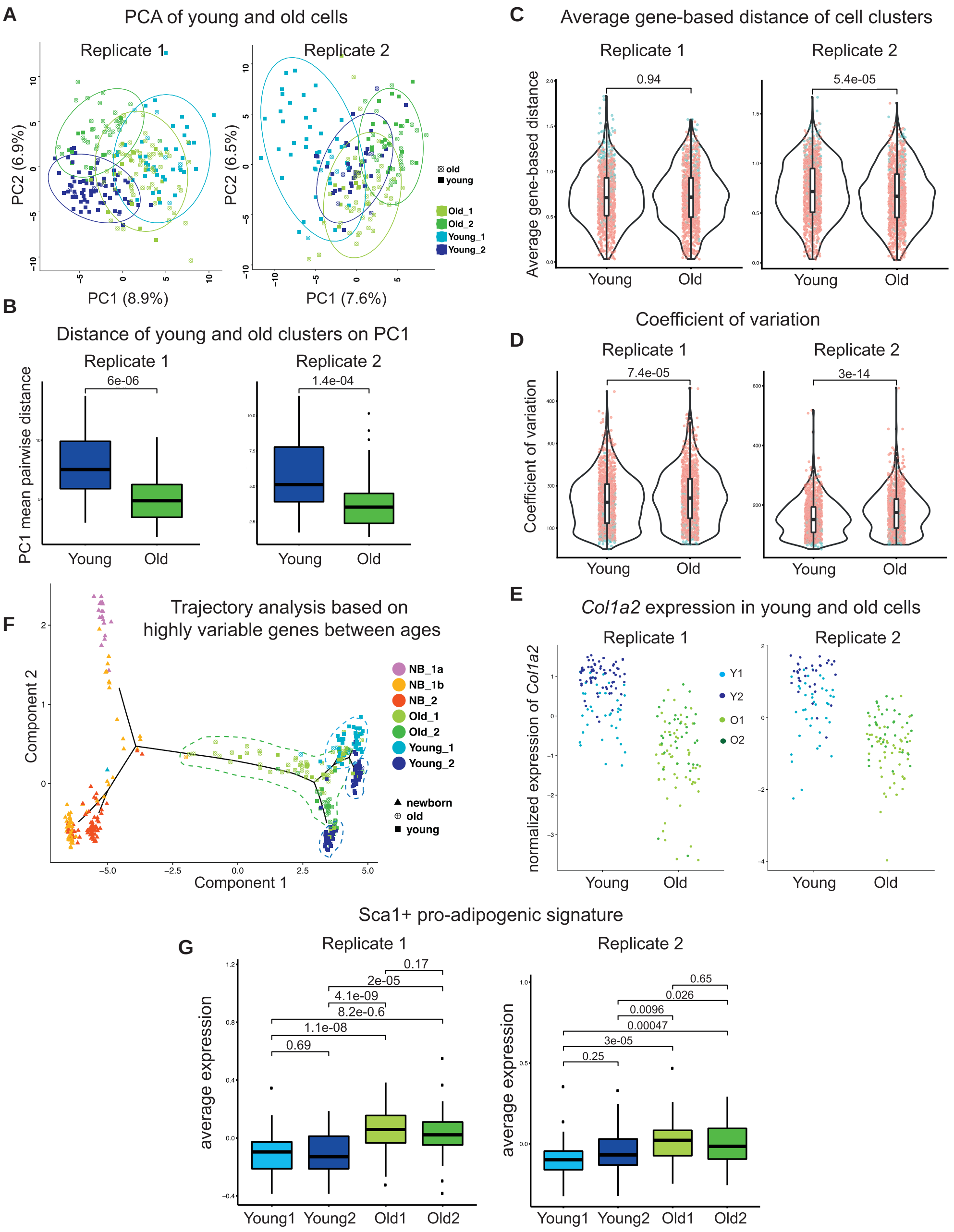


Figure 6. Single-Cell RNA-Seq Reveals Two Distinct Fibroblast Subpopulations in Adult Dermis that Are Less Clearly Demarcated in Old Mice

(A) Principal component analysis of young and old cells excluding newborn cells.

(B) Average of all pairwise distances (average linkage) of fibroblast clusters on PC1 from replicate 1 (left) and replicate 2 (right). Young clusters are more separated than old clusters.

(C) Average of all pairwise distances between the two clusters in young and old for each highly variable gene (HVG). Each dot represents a HVG. Marker genes with a p value < 0.005 are highlighted in blue.

(legend continued on next page)

approximately 300 single cells per age and from two independent biological replicates termed replicate 1 and replicate 2 (Table S5). Of note, the newborn mouse used in replicate 1 was 12–18 hr older than the one used in replicate 2. An independent analysis of the two biological replicates revealed that the results of both experiments strongly overlapped (Figures S7A–S7C). Several interesting conclusions about the lineage relationships between fibroblasts at different ages were drawn from this analysis. First, PCA and tSNE analyses revealed that fibroblasts cluster primarily according to age (Figures 5A and 5B) and suggest that newborn fibroblasts undergo significant molecular changes as mice transition into adulthood. Second, we identified two main types of newborn fibroblasts, termed NB1 and NB2 (Figures 5A and 5B). In replicate 1, cluster NB1 could be further subdivided into NB1a and NB1b, both of which clustered with NB1 from replicate 2 (Figure S7C). Of note, NB1a separated from NB1b fibroblasts due to their higher expression of proliferation genes (Table S5 sheet 2), suggesting that proliferative and quiescent fibroblasts coexist in newborn skin. Comparing the transcriptome signatures of these newborn clusters with those obtained from the GaGa analysis of bulk-sorted newborn fibroblasts (see Figure 3) revealed that NB2 fibroblasts show decreased expression of papillary lineage marker than NB1 fibroblasts (Figure 5D)—in other words, NB2 cells are more similar to CD26− lower dermal fibroblasts, while NB1 cells are more similar to upper CD26+ fibroblasts. Moreover, NB2 fibroblasts were enriched in cells expressing the lower dermal fibroblast marker Dlk1 (Figure 5C). Considering that our single-cell RNA-seq data covered in average less than 50% of the transcriptome per cell, our data did not offer enough resolution to identify the two populations of pro-adipogenic fibroblasts within the cluster of reticular NB2 fibroblasts.

Two clear populations of fibroblasts were also identified in the young adult dermis, termed young1 and young2 (Figures 5A and 5B). These differed mainly in the expression of genes related to the ECM (predominant in young2) and in processes of oxidation-reduction and innate immunity (predominant in young1) (Figure 5E). Interestingly, although the clustering analysis (represented on tSNE) also identified two types of fibroblasts in old dermis (termed old1 and old2), a PCA analysis on young and old cells indicated that the first and strongest principal component better separates young cluster than old cluster (Figures 6A and 6B). In other words, the transcriptome features that define young clusters become blurry with age (Figures 6A and 6B). Additionally, while the average gene-based distance between the two clusters is similar (Figure 6C), the coefficient of variation of these gene-based distances is higher in old, indicating that genes characterizing the young clusters are expressed more consistently throughout their cluster than those genes characterizing the old clusters (Figure 6D). This phenotype is exemplified by *Col1a2*, which in old is expressed less robustly (or more variably) among cells of the same cluster than in young (Figure 6E). Hence, in addition to losing cluster-determining features present in young, the gene expression of old dermal fibroblasts is noisy and very variable among cells from the same cluster. A trajectory analysis of the data based on the genes with the highest significant changes in expression between all ages positioned young fibroblasts at two clearly separate endpoints, whereas old fibroblasts were scattered along the branches with a less well-defined separation (Figure 6F). Strikingly, the trajectory analysis further indicated that old fibroblasts are paradoxically closer to newborn fibroblasts (Figure 6F). Importantly, a comparison of the transcriptomes of old and young fibroblasts obtained by single-cell RNA-seq with the signature of Sca1+/ Dlk1− pro-adipogenic fibroblasts obtained by bulk cell transcriptomics (see Figure 3) indicated that old fibroblasts were more similar to pro-adipogenic newborn fibroblasts than young fibroblasts irrespective of cluster (Figure 6G). This further confirms our hypothesis that the acquisition of adipogenic traits is a general mechanism underlying dermal fibroblast aging which affects the vast majority of old dermal cells. In sum, results from both single-cell RNA-seq and gene expression data of bulk cell populations indicated that, as the dermis ages, its resident fibroblasts have a less well-defined identity (i.e., they become noisy), and acquire adipogenic characteristics reminiscent of newborn pro-adipogenic fibroblasts.

Aging is susceptible to systemic changes in metabolism. For instance, long-term caloric restriction (CR) extends the lifespan of many organisms, including rodents (Froy and Miskin, 2010). On the other hand, a prolonged consumption of a high-fat diet (HFD) accelerates the onset and progression of many age-related pathologies, thereby shortening lifespan (López-Otín et al., 2013). Although the complex interplay of mechanisms underlying the impact of diet on tissue function is still under intense investigation, there is strong evidence indicating that CR prevents the decay of stem cell functions associated to aging, whereas HFD accelerates it (Cerletti et al., 2012; Mihaylova et al., 2014; Sato et al., 2017; Solanas et al., 2017). However, whether and how dietary interventions affect the stroma of tissues during aging is mostly unknown. To study this, we performed two independent experiments aimed at determining whether either CR or HFD affects dermal fibroblast aging by analyzing gene expression from Pdgfrα+/CD34+/Lin− fibroblasts isolated from dermis of the distinct mice groups. Specifically, the CR experiment used: (i) young mice (2-month-old) fed a normal diet (young ND); (ii) old mice (18-month-old) fed a normal diet (old ND); (iii) old mice fed a 30% CR diet for seven months, given in three separate feedings (every 3 hr) during the night to ensure that they ate throughout their most active phase (old CR); and (iv)

(D) Coefficient of variation calculated as the standard deviation of the distances of each gene between cells of opposite clusters divided by the average of these distances for replicate 1 (left) and replicate 2 (right). Each dot represents a HVG. Marker genes with a p value < 0.005 are highlighted in blue. Overall, the variability of gene distances is higher in old than in young.

(E) *Col1a2* expression in young and old cells. *Col1a2* is expressed more robustly between young clusters than old clusters.

(F) Trajectory analysis of newborn, young, and old cells based on highly variable genes between ages. Dashed lines indicate young and old clusters.

(G) Boxplot showing that the Sca1+ pro-adipogenic signature is overrepresented in both old clusters as compared to both young clusters in replicate 1 and replicate 2. Each dot represents the average expression of all genes forming the Sca1+ pro-adipogenic signature of one cell.

Please cite this article in press as: Salzer et al., Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging, Cell (2018), https://doi.org/10.1016/j.cell.2018.10.012

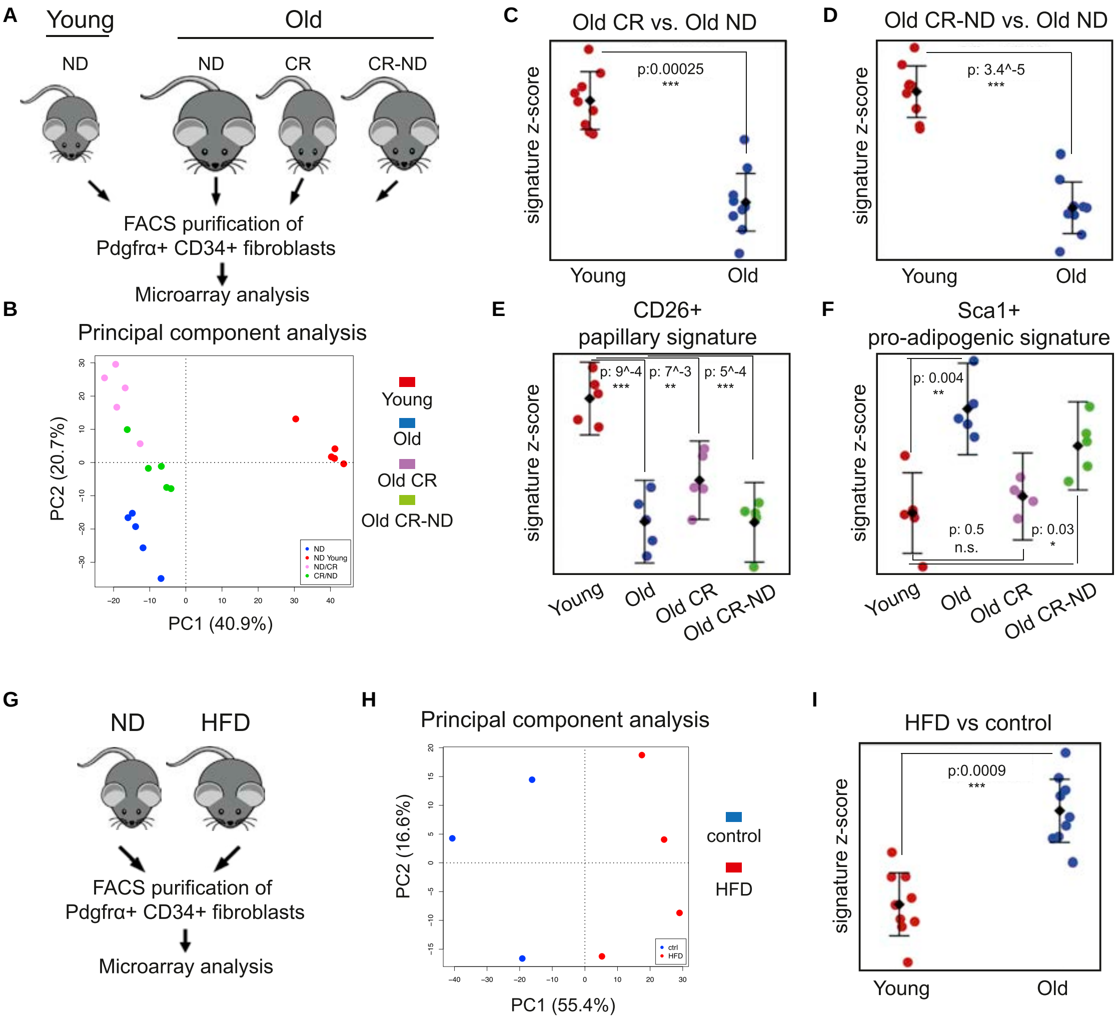


Figure 7. Caloric Restriction and High-Fat Diet Affect Dermal Fibroblast Aging.

(A) Experimental set-up of the caloric restriction experiment. Middle-aged (9-month-old) mice were subjected to diets consisting of two months of normal diet (ND) followed by seven months of caloric restriction (old CR), seven months of CR followed by 2 months of ND (old CR-ND) or 9 months of ND (old ND); mice were sacrificed at 18 months of age. As a control, 1-month-old mice were fed ND for 1 month and then sacrificed (young ND). Pdgfrα+/CD34+ fibroblasts were isolated by FACS and analyzed by whole-genome expression profiling.

(B) PCA of fibroblast samples from the CR experiment (n = 5) according to the microarray expression data.

(C and D) Gene expression changes induced by CR (C) and CR-ND (D) negatively correlate with age. Expression values of genes differentially expressed in old CR (C) or old CR-ND (D) samples compared to Old ND samples were summarized (signature *Z* score) and compared across old and young samples. Each dot represents one old or young sample.

(E and F) A continuous CR diet, but not a CR diet followed by a ND diet (CR-ND), ameliorates the negative correlation of old fibroblasts with CD26+ papillary fibroblasts (E), and the positive correlation of old fibroblasts with Sca1+ pro-adipogenic fibroblasts (F). Expression values of genes forming the CD26+ papillary (E) or the Sca1+ pro-adipogenic (F) signature were summarized (signature *Z* score) and compared across the different samples. Each dot represents one sample. Error bars represent 95% confidence interval.

(G) Experimental set-up of the high-fat diet experiment. Young mice (2-month-old) were subjected to six months of either high-fat diet or control diet and then sacrificed. Pdgfrα+/CD34+ fibroblasts were isolated by FACS and analyzed by whole-genome expression profiling.

(H) PCA of fibroblast samples from the HFD experiment according to the microarray expression data.

(I) Gene expression changes induced by HFD positively correlate with age. Expression values of genes differentially expressed in HFD samples were summarized (signature *Z* score) and compared across old and young samples. Each dot represents one old or young sample. Error bars represent 95% confidence interval.

old mice fed a CR diet for 7 months followed by a normal diet for 2 additional months (old CR-ND) to determine whether the effects of CR, if any, were reversible (Figure 7A; Table S6). The HFD experiment used young mice (2-month-old) fed either a normal diet or HFD for 6 months (Figure 7G; Table S7).

PCA analysis of the CR gene expression data revealed that young and old fibroblasts still primarily separated by age, irrespective of diet (Figure 7B). However, diet exerted a significant change on the transcriptome of old fibroblasts, as shown by the distancing in clustering of fibroblasts isolated from CR mice as compared to ND mice (Figure 7B). Interestingly, old fibroblasts purified from CR-ND mice clustered between those from caloric restricted and normal diet fed old mice, indicating that some of the effects of caloric restriction are maintained even after two months of re-feeding with a normal diet (Figure 7B). Importantly, comparing the transcriptomes of all conditions indicated that old CR and old CR-ND fibroblasts negatively correlated with age (Figures 7C and 7D). In contrast, the signature of adult fibroblasts isolated from HFD mice positively correlated with that of old fibroblasts (Figure 7I). Thus, our results show that dermal fibroblast aging can be significantly delayed by CR or, alternatively, enhanced by HFD. As our previous analyses indicate that dermal fibroblast aging was associated with the acquisition of pro-adipogenic traits, among other changes, we next determined whether CR affects this aspect of the dermal aging process. Strikingly, CR significantly prevented old fibroblasts from losing papillary traits and acquiring adipogenic characteristics (Figures 7E and 7F). However, dermal fibroblasts from mice re-fed a normal diet after caloric restriction (CR-ND mice) started to lose papillary features while re-gaining pro-adipogenic characteristics (Figures 7E and 7F). This suggests the interesting hypothesis that some aspects of the rejuvenating potential of caloric restriction are stable, while others require a constant CR diet.

In conclusion, our results indicate that the lineage heterogeneity of adult dermal fibroblasts is progressively blurred during aging. Of note, our transcriptome results show that old fibroblasts differentially express genes involved in promoting cytoskeletal extensions and cell contacts and downregulate repulsion signals between cells (such as semaphorins). This suggests the intriguing hypothesis that an additional hallmark of dermal fibroblast aging might involve their preference for replenishing the empty dermal space created by any surrounding dying fibroblast by contacting distant fibroblasts through membrane protrusions, rather than by increasing cell proliferation. The dynamics of cell proliferation and cytoskeletal changes of dermal fibroblasts in response to the injury of nearby fibroblasts and during aging, shown in the accompanying article (Marsh et al., 2018, this issue of *Cell*), strongly support this hypothesis. In the future, it will be interesting to study the breaks in signaling that prevent dermal fibroblasts from responding to small-scale neighboring cell death by proliferation in homeostasis, and how and why these are surmounted during wound-healing and tumorigenesis.

Besides the acquisition of identity noise, we show that old fibroblasts lose production and secretion of ECM components yet concomitantly upregulate the expression of genes involved in inflammation, lipid metabolism, and adipogenesis. Moreover, we found LRP signaling to be the most elevated pathway in old dermal fibroblasts, which supports previous findings implicating augmented Wnt signaling in aging (Liu et al., 2007). The acquisition of these traits makes old fibroblasts paradoxically more similar to pro-adipogenic fibroblasts present in newborn dermis. Critically, this transition can be partially—albeit significantly—prevented by caloric restriction. Our results therefore indicate that loss of cell identity is a previously overlooked mechanism underlying cellular aging and offer therapeutic possibilities to delay skin aging through dietary and metabolic interventions.

STAR★METHODS

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

- KEY RESOURCE TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODELS AND SUBJECT DETAILS
  - ○ Mice
- METHOD DETAILS
  - ○ FACS sorting
  - ○ Fibroblast culture
  - ○ Microarrays
  - ○ Single-cell RNA-sequencing
  - ○ qRT-PCR
  - ○ Immunofluorescence
  - ○ Microscopy and image analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Quantification of skin morphology features
  - ○ qRT-PCR analysis
  - ○ Microarrays
  - ○ Single-cell RNA-sequencing analysis
  - ○ Coefficient of variation
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and seven tables and can be found with this article online at https://doi.org/10.1016/j.cell.2018.10.012.

ACKNOWLEDGMENTS

The European Research Council (ERC), the Spanish Ministry of Economy and Development, and the Institute for Research in Biomedicine (IRB-Barcelona) supported this project in the lab of S.A.B. M.C.S. was supported by a *Boehringer Ingelheim Fonds* International PhD fellowship. We would like to thank the IRB Functional Genomics Core Facility for technical assistance in microarray hybridization experiments, the IRB Biostatistics/Bioinformatics Core Facility for assistance in microarray analysis, and the Universitat de Barcelona FACS unit, the PCB (Scientific Park of Barcelona) Animal Unit, IRB Advanced Digital Microscopy Unit, and the IRB Histopathology Unit for technical assistance and support. IRB Barcelona is the recipient of a Severo Ochoa Award of Excellence from MINECO (Government of Spain). H.H. is a Miguel Servet (CP14/00229) researcher funded by the Spanish Institute of Health Carlos III (ISCIII). This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. H2020-MSCA-ITN-2015-675752 (SINGEK). We would like to thank Kim Bak Jensen (BRIC Institute, Copenhagen) for kindly providing the Lrig1-CreERT2 mice and Veronica Raker for manuscript editing.

AUTHOR CONTRIBUTIONS

Design of *in vivo* experiments, M.C.S. and S.A.B.; design of single-cell RNA-seq experiments, H.H., A.L., M.C.S. and S.A.B.; bioinformatics analysis of single-cell RNA-seq data, A.L. and H.H.; bioinformatics analysis of microarray data (including GSEA and GaGa analyses), A.B.-L. and C.S.-O.A.; FACS sorting, M.C.S.; monitoring of high-fat diet mice, G.S. and F.O.P.; technical assistance, C.Y. and A.C.; histological analysis of the skin samples, M.C.S., M.A., and N.P.; establishing correct feeding regimes and protocol in the CR and HFD experiments, J.M.-C.; coordination and manuscript writing, S.A.B.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

Received: February 26, 2018
Revised: August 1, 2018
Accepted: October 2, 2018
Published: November 8, 2018

REFERENCES

Acar, M., Kocherlakota, K.S., Murphy, M.M., Peyer, J.G., Oguro, H., Inra, C.N., Jaiyeola, C., Zhao, Z., Luby-Phelps, K., and Morrison, S.J. (2015). Deep imaging of bone marrow shows non-dividing stem cells are mainly perisinusoidal. Nature *526*, 126–130.

Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Asada, N., Kunisaki, Y., Pierce, H., Wang, Z., Fernandez, N.F., Birbrair, A., Ma'ayan, A., and Frenette, P.S. (2017). Differential cytokine contributions of perivascular haematopoietic stem cell niches. Nat. Cell Biol. *19*, 214–223.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. B *57*, 289–300.

Berry, R., and Rodeheffer, M.S. (2013). Characterization of the adipocyte cellular lineage in vivo. Nat. Cell Biol. *15*, 302–308.

Brownell, I., Guevara, E., Bai, C.B., Loomis, C.A., and Joyner, A.L. (2011). Nerve-derived sonic hedgehog defines a niche for hair follicle stem cells capable of becoming epidermal stem cells. Cell Stem Cell *8*, 552–565.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol *36*, 411–420.

Carlson, M. (2016) org.Mm.eg.db: Genome wide annotation for Mouse, R package version 3.3.0. https://bioconductor.org/packages/release/data/annotation/html/org.Mm.eg.db.html

Casanova-Acebes, M., Pitaval, C., Weiss, L.A., Nombela-Arrieta, C., Chevre, R., A-Gonzalez, N., Kunisaki, Y., Zhang, D., van Rooijen, N., Silberstein, L.E., Weber, C., et al. (2013). Rhythmic modulation of the hematopoietic niche through neutrophil clearance. Cell *153*, 1025–1035.

Cerletti, M., Jang, Y.C., Finley, L.W., Haigis, M.C., and Wagers, A.J. (2012). Short-term calorie restriction enhances skeletal muscle stem cell function. Cell Stem Cell *10*, 515–519.

Chung, N.C., and Storey, J.D. (2015). Statistical significance of variables driving systematic variation in high-dimensional data. Bioinformatics *31*, 545–554.

Demaria, M., Desprez, P.Y., Campisi, J., and Velarde, M.C. (2015). Cell autonomous and non-autonomous effects of senescent cells in the skin. J. Invest. Dermatol *135*, 1722–1726.

Doles, J., Storer, M., Cozzuto, L., Roma, G., and Keyes, W.M. (2012). Age-associated inflammation inhibits epidermal stem cell function. Genes Dev. *26*, 2144–2153.

Donati, G., Proserpio, V., Lichtenberger, B.M., Natsuga, K., Sinclair, R., Fujiwara, H., and Watt, F.M. (2014). Epidermal Wnt/β-catenin signaling regulates adipocyte differentiation via secretion of adipogenic factors. Proc. Natl. Acad. Sci. USA *111*, E1501–E1509.

Driskell, R.R., and Watt, F.M. (2015). Understanding fibroblast heterogeneity in the skin. Trends Cell Biol. *25*, 92–99.

Driskell, R.R., Lichtenberger, B.M., Hoste, E., Kretzschmar, K., Simons, B.D., Charalambous, M., Ferron, S.R., Herault, Y., Pavlovic, G., Ferguson-Smith, A.C., and Watt, F.M. (2013). Distinct fibroblast lineages determine dermal architecture in skin development and repair. Nature *504*, 277–281.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc *4*, 1184–1191.

Eklund, A.C., and Szallasi, Z. (2008). Correction of technical bias in clinical microarray data improves concordance with known biological information. Genome Biol. *9*, R26.

Festa, E., Fretz, J., Berry, R., Schmidt, B., Rodeheffer, M., Horowitz, M., and Horsley, V. (2011). Adipocyte lineage cells contribute to the skin stem cell niche to drive hair cycling. Cell *146*, 761–771.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. *16*, 278.

Flach, J., Bakker, S.T., Mohrin, M., Conroy, P.C., Pietras, E.M., Reynaud, D., Alvarez, S., Diolaiti, M.E., Ugarte, F., Forsberg, E.C., et al. (2014). Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells. Nature *512*, 198–202.

Florian, M.C., Nattamai, K.J., Dörr, K., Marka, G., Uberle, B., Vas, V., Eckl, C., Andrä, I., Schiemann, M., Oostendorp, R.A., et al. (2013). A canonical to non-canonical Wnt signalling switch in haematopoietic stem-cell ageing. Nature *503*, 392–396.

Froy, O., and Miskin, R. (2010). Effect of feeding regimens on circadian rhythms: implications for aging and longevity. Aging (Albany NY) *2*, 7–27.

Fujiwara, H., Ferreira, M., Donati, G., Marciano, D.K., Linton, J.M., Sato, Y., Hartner, A., Sekuguchi, K., Reichardt, L.F., and Watt, F.M. (2011). The basement membrane of hair follicle stem cells is a muscle cell niche. Cell *144*, 577–589.

Gao, X., Xu, C., Asada, N., and Frenette, P.S. (2018). The hematopoietic stem cell niche: from embryo to adult. Development *145*.. https://doi.org/10.1242/dev.139691.

García-Prat, L., Martinez-Vicente, M., Perdiguero, E., Ortet, L., Rodriguez-Ubreva, J., Rebollo, E., Ruiz-Bonilla, V., Gutarra, S., Ballestar, E., Serrano, A.L., et al. (2016). Autophagy maintains stemness by preventing senescence. Nature *529*, 37–42.

Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics *20*, 307–315.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2005). In Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Springer).

Goodell, M.A., and Rando, T.A. (2015). Stem cells and healthy aging. Science *350*, 1199–1204.

Gopinath, S.D., and Rando, T.A. (2008). Stem Cell Review Series: Aging of the skeletal muscle stem cell niche. Aging Cell *7*, 590–598.

Greco, V., Chen, T., Rendl, M., Schober, M., Pasolli, H.A., Stokes, N., Dela Cruz-Racelis, J., and Fuchs, E. (2009). A two-step mechanism for stem cell activation during hair regeneration. Cell Stem Cell *4*, 155–169.

Harbor, C.S., and King, W. (2014). Cell-extracellular matrix interactions in normal and diseased skin. Cold Spring Harb. Perspect. Biol. *3*. https://doi.org/10.1101/cshperspect.a005124.

Hill, R.P., Gardner, A., Crawford, H.C., Richer, R., Dodds, A., Owens, W.A., Lawrence, C., Rao, S., Kara, B., James, S.E., and Jahoda, C.A. (2013). Human hair follicle dermal sheath and papilla cells support keratinocyte growth in monolayer coculture. Exp. Dermatol *22*, 236–238.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Inra, C.N., Zhou, B.O., Acar, M., Murphy, M.M., Richardson, J., Zhao, Z., and Morrison, S.J. (2015). A perisinusoidal niche for extramedullary haematopoiesis in the spleen. Nature *527*, 466–471.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2018). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249–264.

Kaur, A., Webster, M.R., Marchbank, K., Behera, R., Ndoye, A., Kugel, C.H., Dang, V.M., Appleton, J., O'Connell, M.P., Cheng, P., et al. (2016). SFRP2 in the aged microenvironment drives melanoma metastasis and therapy resistance. Nature *532*, 250–254.

Katayama, Y., Battista, M., Kao, W.M., Hidalgo, A., Peired, A.J., Thomas, S.A., and Frenette, P.S. (2006). Signals from the sympathetic nervous system regulate hematopoietic stem cell egress from bone marrow. Cell *124*, 407–421.

Keyes, B.E., and Fuchs, E. (2018). Stem cells: Aging and transcriptional fingerprints. J. Cell Biol. *217*, 79–92.

Keyes, B.E., Segal, J.P., Heller, E., Lien, W.H., Chang, C.Y., Guo, X., Oristian, D.S., Zheng, D., and Fuchs, E. (2013). Nfatc1 orchestrates aging in hair follicle stem cells. Proc. Natl. Acad. Sci. USA *110*, E4950–E4959.

Kusumbe, A.P., Ramasamy, S.K., Itkin, T., Mae, M.A., Langen, U.H., Betsholtz, C., Lapidot, T., and Adams, R.H. (2016). Age-dependent modulation of vascular niches for haematopoietic stem cells. Nature *532*, 380–384.

Lee, J.H., Tammela, T., Hofree, M., Choi, J., Marjanovic, N.D., Han, S., Canner, D., Wu, K., Paschini, M., Bhang, D.H., et al. (2017). Anatomically and Functionally Distinct Lung Mesenchymal Populations Marked by Lgr5 and Lgr6. Cell *170*, 1149–1163.

Lefrançais, E., Ortiz-Munoz, G., Caudrillier, A., Mallavia, B., Liu, F., Sayah, D.M., Thornton, E.E., Headley, M.B., David, T., Coughlin, S.R., et al. (2017). The lung is a site of platelet biogenesis and a reservoir for haematopoietic progenitors. Nature *544*, 105–109.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.

Liu, H., Fergusson, M.M., Castilho, R.M., Liu, J., Cao, L., Chen, J., Malide, D., Rovira, I.I., Schimel, D., Kuo, C.J., et al. (2007). Augmented Wnt signaling in a mammalian model of accelerated aging. Science *317*, 803–806.

Loffredo, F.S., Steinhauser, M.L., Jay, S.M., Gannon, J., Pancoast, J.R., Yalamanchi, P., Sinha, M., Dall'Osso, C., Khong, D., Shadrach, J.L., et al. (2013). Growth differentiation factor 11 is a circulating factor that reverses age-related cardiac hypertrophy. Cell *153*, 828–839.

López-Otín, C., Blasco, M.A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. Cell *153*, 1194–1217.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. Nat. Methods *9*, 1185–1188.

Marsh, E., Gonzalez, D.G., Lathrop, E.A., Boucher, J.D., and Greco, V. (2018). Positional stability and membrane occupancy define skin fibroblast homeostasis in vivo. Cell *175*. Published online November 8, 2018. https://doi.org/10.1016/j.cell.2018.10.013.

Mashinchian, O., Pisconti, A., Le Moal, E., and Bentzinger, C.F. (2018). The muscle stem cell niche in health and disease. Curr. Top Dev. Bio *126*, 23–65.

Mastrogiannaki, M., Lichtenberger, B.M., Reimer, A., Collins, C.A., Driskell, R.R., and Watt, F.M. (2016). β-Catenin Stabilization in Skin Fibroblasts Causes Fibrotic Lesions by Preventing Adipocyte Differentiation of the Reticular Dermis. J. Invest. Dermatol *136*, 1130–1142.

Matsumura, H., Mohri, Y., Binh, N.T., Morinaga, H., Fukuda, M., Ito, M., Kurata, S., Hoeijmakers, J., and Nishimura, E.K. (2016). Hair follicle aging is driven by transepidermal elimination of stem cells via COL17A1 proteolysis. Science *351*, aad4395.

Méndez-Ferrer, S., Michurina, T.V., Ferraro, F., Mazloom, A.R., Macarthur, B.D., Lira, S.A., Scadden, D.T., Ma'ayan, A., Enikolopov, G.N., and Frenette, P.S. (2010). Mesenchymal and haematopoietic stem cells form a unique bone marrow niche. Nature *466*, 829–834.

Mihaylova, M.M., Sabatini, D.M., and Yilmaz, Ö.H. (2014). Dietary and metabolic control of stem cell function in physiology and cancer. Cell Stem Cell *14*, 292–305.

Morrison, S.J., and Scadden, D.T. (2014). The bone marrow niche for haematopoietic stem cells. Nature *505*, 327–334.

Mudge, J.M., and Harrow, J. (2015). Creating reference gene annotation for the mouse C57BL6/J genome assembly. Mamm. Genome *26*, 366–378.

Neves, J., Sousa-Victor, P., and Jasper, H. (2017). Rejuvenating Strategies for Stem Cell-Based Therapies in Aging. Cell Stem Cell *20*, 161–175.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *27*, 29–34.

Oh, J., Lee, Y.D., and Wagers, A.J. (2014). Stem cell aging: Mechanisms, regulators and therapeutic opportunities. Nat. Med. *20*, 870–880.

Page, M.E., Lombard, P., Ng, F., Göttgens, B., and Jensen, K.B. (2013). The epidermis comprises autonomous compartments maintained by distinct stem cell populations. Cell Stem Cell *13*, 471–482.

Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods *10*, 1096–1098.

Plikus, M.V., Guerrero-Juarez, C.F., Ito, M., Li, Y.R., Dedhia, P.H., Zheng, Y., Shao, M., Gay, D.L., Ramos, R., His, T.C., et al. (2017). Regeneration of fat cells from myofibroblasts during wound healing. Science *355*, 748–752.

Price, F.D., von Maltzahn, J., Bentzinger, C.F., Dumont, N.A., Yin, H., Chang, N.C., Wilson, D.H., Frenette, J., and Rudnicki, M.A. (2014). Inhibition of JAK-STAT signaling stimulates adult satellite cell function. Nat. Med. *20*, 1174–1181.

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.A., and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. Nat. Methods *14*, 309–315.

R Development Core Team (2011). R: A language and environment for statistical computing. https://www.R-project.org/.web

Rendl, M., Lewis, L., and Fuchs, E. (2005). Molecular dissection of mesenchymal-epithelial interactions in the hair follicle. PLoS Biol. *3*, 1910–1924.

Rendl, M., Polak, L., and Fuchs, E. (2008). BMP signaling in dermal papilla cells is required for their hair follicle-inductive properties. Genes Dev. *22*, 543–557.

Rinkevich, Y., Walmsley, G.G., Hu, M.S., Maan, Z.N., Newman, A.M., Drukker, M., Januszyk, M., Krampitz, G.W., Gurtner, G.C., Lorenz, H.P., et al. (2015). Identification and isolation of a dermal lineage with intrinsic fibrogenic potential. Science *348*, 1–15.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Rivera-Gonzalez, G.C., Shook, B.A., Andrae, J., Holtrup, B., Bollag, K., Betsholtz, C., Rodeheffer, M.S., and Horsley, V. (2016). Skin adipocyte stem cell self-renewal is regulated by a PDGFA/AKT-signaling axis. Cell Stem Cell *19*, 738–751.

Rognoni, E., Gomez, C., Pisco, A.O., Rawlins, E.L., Simons, B.D., Watt, F.M., and Driskell, R.R. (2016). Inhibition of β-catenin signalling in dermal fibroblasts enhances hair follicle regeneration during wound healing. Development *143*, 2522–2535.

Rossell, D. (2009). GaGa: A parsimonious and flexible model for differential expression analysis. Ann. Appl. Stat. *3*, 1035–1051.

Sato, S., Solanas, G., Peixoto, F.O., Bee, L., Symeonidi, A., Schmidt, M.S., Brenner, C., Masri, S., Benitah, S.A., and Sassone-Corsi, P. (2017). Circadian Reprogramming in the Liver Identifies Metabolic Pathways of Aging. Cell *170*, 664–677.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: An open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.

Sennett, R., and Rendl, M. (2012). Mesenchymal-epithelial interactions during hair follicle morphogenesis and cycling. Semin. Cell Dev. Biol. *23*, 917–927.

Sennett, R., Wang, Z., Rezza, A., Grisanti, L., Roitershtein, N., Sicchio, C., Mok, K.W., Heitman, N.J., Clavel, C., Ma'ayan, A., and Rendl, M. (2015). An integrated transcriptome atlas of embryonic hair follicle progenitors, their niche, and the developing skin. Dev. Cell *34*, 577–591.

Signer, R.A.J., and Morrison, S.J. (2013). Mechanisms that regulate stem cell aging and life span. Cell Stem Cell *12*, 152–165.

Smyth, G. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. *3*.

Solanas, G., Peixoto, F.O., Perdiguero, E., Jardí, M., Ruiz-Bonilla, V., Datta, D., Symeonidi, A., Castellanos, A., Welz, P.S., Caballero, J.M., et al. (2017). Aged Stem Cells Reprogram Their Daily Rhythmic Functions to Adapt to Stress. Cell *170*, 678–692.

Sousa-Victor, P., Gutarra, S., García-Prat, L., Rodriguez-Ubreva, J., Ortet, L., Ruiz-Bonilla, V., Jardí, M., Ballestar, E., González, S., Serrano, A.L., et al. (2014). Geriatric muscle stem cells switch reversible quiescence into senescence. Nature *506*, 316–321.

Stearns-Reider, K.M., D'Amore, A., Beezhold, K., Rothrauff, B., Cavalli, L., Wagner, W.R., Vorp, D.A., Tsamis, A., Shinde, S., Zhang, C., et al. (2017). Aging of the skeletal muscle extracellular matrix drives a stem cell fibrogenic conversion. Aging Cell *16*, 518–528.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Sun, D., Luo, M., Jeong, M., Rodriguez, B., Xia, Z., Hannah, R., Wang, H., Le, T., Faull, K.F., Chen, R., et al. (2014). Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal. Cell Stem Cell *14*, 673–688.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol *32*, 381–386.

Waltman, L., and Van Eck, N.J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. Eur. Phys. J. B *86*.

Wojciechowicz, K., Gledhill, K., Ambler, C.A., Manning, C.B., and Jahoda, C.A. (2013). Development of the mouse dermal adipose layer occurs independently of subcutaneous adipose tissue and is marked by restricted early expression of FABP4. PLoS One 8. https://doi.org/10.1371/journal.pone.0059811.

Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, j.F., Schmidt, H.K., Yellapantula, V., Miller, C.A., et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med *20*, 1472–1478.

Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics *31*, 1974–1980.

Yang, H., Adam, R.C., Ge, Y., Hua, Z.L., and Fuchs, E. (2017). Epithelial-mesenchymal micro-niches govern stem cell lineage choices. Cell *169*, 483–496.

Zhang, B., Tsai, P.C., Gonzalez-Celeiro, M., Chung, O., Boumard, B., Perdigoto, C.N., Ezhkova, E., and Hsu, Y.C. (2016). Hair follicles' transit-amplifying cells govern concurrent dermal adipocyte production through sonic hedgehog. Genes Dev. *30*, 2325–2338.

Zhou, B.O., Yu, H., Yue, R., Zhao, Z., Rios, J.J., Naveiras, O., and Morrison, S.J. (2017). Bone marrow adipocytes promote the regeneration of stem cells and haematopoiesis by secreting SCF. Nat. Cell Biol. *19*, 891–903.

Please cite this article in press as: Salzer et al., Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging, Cell (2018), https://doi.org/10.1016/j.cell.2018.10.012

STAR★METHODS

KEY RESOURCE TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Biotin anti-mouse CD326 (Ep-Cam) antibody (Clone G8.8)	BioLegend	Cat# 118203, RRID:AB_1134174
Rat Anti-CD31 Monoclonal Antibody, Biotin Conjugated, Clone MEC 13.3	BD Biosciences	Cat# 553371, RRID:AB_394817
PerCP/Cy5.5 anti-mouse CD31 antibody (Clone MEC13.3)	BioLegend	Cat# 102522, RRID:AB_2566761
Rat Anti-CD24 Monoclonal Antibody, Biotin Conjugated, Clone M1/69	BD Biosciences	Cat# 553260, RRID:AB_394739
CD45 Monoclonal Antibody (30-F11), Biotin, eBioscience	Thermo Fisher Scientific	Cat# 13-0451-81, RRID:AB_466445
Rat Anti-CD117 Monoclonal Antibody, Biotin Conjugated, Clone 2B8	BD Biosciences	Cat# 553353, RRID:AB_394804
PE anti-mouse CD140a antibody	BioLegend	Cat# 135906, RRID:AB_1953269
CD140a (PDGFRA) Monoclonal Antibody (APA5), APC, eBioscience	Thermo Fisher Scientific	Cat# 17-1401-81, RRID:AB_529482
CD34 Monoclonal Antibody (RAM34), FITC, eBioscience	Thermo Fisher Scientific	Cat# 11-0341-82, RRID:AB_465021
CD26 Monoclonal Antibody (H194-112 (H194-112.12.4.4.5)), PerCP-Cyanine5.5, eBioscience	Thermo Fisher Scientific	Cat# 45-0261-80, RRID:AB_1548741
Brilliant Violet 605 anti-mouse Ly-6A/E (Sca-1) antibody	BioLegend	Cat# 108133, RRID:AB_2562275
Dlk-Pref-1 (24-11)-PE Monoclonal Antibody	MBL International	Cat# D187-5, RRID:AB_1520822
Streptavidin-Allophycocyanin antibody	BD Biosciences	Cat# 554067, RRID:AB_10050396
Streptavidin APC-eFluor 780 100 ug antibody	Thermo Fisher Scientific	Cat# 47-4317-82, RRID:AB_10366688
PPARgamma (H-100) antibody	Santa Cruz Biotechnology	Cat# sc-7196, RRID:AB_654710
FABP4 antibody	Proteintech Group	Cat# 15872-1-AP, RRID:AB_2102440
RFP Tag Monoclonal Antibody (RF5R)	Thermo Fisher Scientific	Cat# MA5-15257, RRID:AB_10999796
Keratin 14 Polyclonal Chicken Antibody, Purified	Covance Research Products Inc	Cat# SIG-3476-100, RRID:AB_10718041
Vimentin antibody	Fitzgerald Industries International	Cat# 20R-VP004, RRID:AB_1289477
BCAT1 Monoclonal Antibody (OTI3F5)	Thermo Fisher Scientific	Cat# MA5-25892, RRID:AB_2722890
PPAR delta antibody	Abcam	Cat# ab23673, RRID:AB_2165902
Mouse PDGF R alpha Affinity Purified Polyclonal Ab antibody	R and D Systems	Cat# AF1062, RRID:AB_2236897
Mouse Anti-Actin, alpha-Smooth Muscle Monoclonal Antibody, Unconjugated, Clone 1A4	Sigma-Aldrich	Cat# A5228, RRID:AB_262054
Donkey anti-Rabbit IgG Secondary Antibody, Alexa Fluor 488	Thermo Fisher Scientific	Cat# R37118, RRID:AB_2556546
Donkey anti-Mouse IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 568	Thermo Fisher Scientific	Cat# A10037, RRID:AB_2534013
Goat anti-Chicken IgY (H+L) Secondary Antibody, Alexa Fluor 647	Thermo Fisher Scientific	Cat# A-21449, RRID:AB_2535866
Goat polyclonal Secondary Antibody to Guinea pig IgG - H&L (Cy3), pre-adsorbed	Abcam	Cat# ab102370, RRID:AB_10711466

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Collagenase 1	Sigma	Cat#C0130
Fluorescence mounting medium	Dako	S302380-2
EnVision FLEX antibody diluent	Dako	K800621-2
Paraformaldehyde	Electron Microscopy Sciences	Cat#15710
10% neutral buffered saline	Sigma	Cat#HT501128-4l
Hematoxilin	Dako	S302084-2
Donkey serum	Sigma	Cat#D9663-10ML
DAPI	Sigma	Cat#D9542-10MG
Dnase1	Sigma	Cat#DN25-1G
TRIzol	ThermoFisher Scientific	Cat#A33250
Chelex	Bio-rad	Cat#142-2842
SYBR green Master	Roche	Cat#04707516001
Collagen 1	Corning	Cat#354236
4-Hydroxytamoxifen	Sigma	Cat#H6278
Critical Commercial Assays		
PureLink Quick PCR Purification Kit	ThermoFisher Scientific	Cat#K310002
WTA2 Complete Whole Transcriptome Amplification Kit	Sigma	Cat#WTA2-50RXN
Quiagen RNeasy Mini columns	Quiagen	Cat#74106
RNACleanXP Kit	Beckman coulter	Cat#A63987
MG-430 PM Array Strip Kit	ThermoFisher Scientific	Cat#901570
GeneAtlas Fluidics Station	Affymetrix, ThermoFisher Scientific	P/N: 00-0377
Genechip Human Mapping 250K Nsp Assay Kit	ThermoFisher Scientific	P/N: 900766
GeneAtlas Imaging Station	Affymetrix, ThermoFisher Scientific	P/N: 00-0376
SuperScript II	ThermoFisher Scientific	Cat#18064014
KAPA Hifi Hotstart ReadyMix	Kappa Biosystems	Cat#KK2501
Agencourt Ampure XP beads	Beckman coulter	Cat#A63881
Nextera XT DNA Library preparation kit (96 samples)	Illumina	Cat#FC-131-1096
Bioanalyzer High Sensitivity DNA Kit	Agilent Technologies	Cat#5067-4626
Deposited Data		
Single-cell RNA-sequencing data	This paper	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111136
Raw and analyzed microarray data	This paper	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE110983
Experimental Models: Organisms/Strains		
Mouse: C57BL/6J Wild-type strain	The Jackson Laboratory	JAX:00664
Mouse: Lrig1-EGFP-IRES-CreERT2	Page et al., 2013	N/A
Mouse: Rosa26-CAG-STOPflox-tdTomato	The Jackson Laboratory	JAX:007905
Oligonucleotides		
qPCR primer	This paper	Table S1 sheet 20
Software and Algorithms		
RMA – affy package	(Irizarry et al., 2018)	https://bioconductor.org/packages/release/bioc/html/affy.html
limma – R Bioconductor package	(Smyth, 2004)	https://bioconductor.org/packages/release/bioc/html/limma.html
Gaga – R Bioconductor package	(Rossell, 2009)	https://bioconductor.org/packages/release/bioc/html/gaga.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GSEA	(Subramanian et al., 2005)	http://software.broadinstitute.org/gsea/index.jsp; RRID:SCR_003199
Genomatix	Genomatix GmbH	www.genomatix.de; RRID:SCR_008036
David version 6.8	(Huang, Sherman and Lempicki, 2009)	https://david.ncifcrf.gov/; RRID:SCR_001881
Fiji v2.0.0-rc-14/1.49 g	(Schindelin et al., 2012)	https://imagej.net/Fiji
Illustrator CS6	Adobe	RRID:SCR_010279
GraphPad Prism6	Pad Software	RRID:SCR_002798
BD FACS Diva Software	BD Biosciences	RRID:SCR_001456
R 3.5.1	R Development Core Team (2011)	http://www.R-project.org
Seurat Package 2.1.0	(Butler et al., 2018)	https://satijalab.org/seurat/; RRID:SCR_016341
FastQC	(Andrews, 2010)	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/; RRID:SCR_014583
GEMTools 1.7.0 suite	(Marco-Sola et al., 2012)	http://gemtools.github.io; RRID:SCR_001259
NDP.view 2	N/A	https://www.hamamatsu.com/eu/en/product/type/U12388-01/index.html
Monocle 2.6.1	(Trapnell et al., 2014) (Qiu et al., 2017)	http://cole-trapnell-lab.github.io/monocle-release/
Other		
Rodent diet Caloric-Restricted	Harlan	TD.120686
Rodent Control diet for Caloric- Restricted diet	Harlan	TD.120685
Rodent High-fat diet	Harlan	TD.06414
Standard rodent diet	SDS Special Diets Services	RM1 (P) 801151

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Salvador Aznar-Benitah (salvador.aznar-benitah@irbbarcelona.org).

EXPERIMENTAL MODELS AND SUBJECT DETAILS

Lrig1-EGFP-IRES-CreERT2 (Page et al., 2013), Rosa26-CAG-STOPflox-tdTomato (JAX 007905), and wild-type C57BL/6 mice were bred and aged at the animal facilities of the Barcelona Science Park and fed standard rodent chow (RM1 (P) 801151 for maintenance, RM3 (P) 801700 for breedings) unless otherwise indicated. For some experiments indicated below, C57BL/6J mice were purchased from Charles River. The Catalan Government approved the work protocols, in accordance with applicable legislation. All experiments were performed with female mice, unless otherwise indicated.

Mice

Aging study

To perform the aging study, four two-month-old mice (young) and four 18-month-old mice (old) aged in the Barcelona Science Park were used. For sample collection, one old and one young mouse were sacrificed per day at 10 am.

For bioinformatic analysis of the transcriptome of old versus young dermal fibroblasts the expression data of the samples from this experiment was combined with that of the ‘‘Normal Diet’’ samples from the Caloric Restriction study (see ‘‘quantification and statistical analysis’’).

Caloric restriction study

To perform the caloric restriction study, female retired breeders from the C57BL/6J strain were purchased from Charles Rivers at the age of 6 to 7 months. At the age of 9 months, mice were then either fed a normal diet *ad libitum* for 9 months (ND group), a normal diet for 2 months followed by 30% caloric restriction for 7 months (CR group), or a 30% caloric restriction for 7 months followed by a normal diet for 2 months (CR-ND group). The normal diet (Harlan TD.120686) contained 18.8% protein, 7.3% fatty acids, 55.1% carbohydrates and 3.6 kcal/g. The caloric restriction diet (Harlan TD.120685) contained 32.9% protein, 12.7% fatty acids, 31.9% carbohydrates and 3.7 kcal/g. During the CR periods, mice were fed with an automatic feeder (Fishmate F14) that provided food in 3 portions during the most active phase of the mice (at 8 pm [time of lights-off], 11 pm, and 2 am). Prior to subjecting mice to 30% CR, and prior to re-feeding them *ad libitum* with ND, mice had a 3-week adaptation period during which food was reduced or increased, respectively, by 10% each week. Mice were weighed weekly. Young mice (4-week-old) were purchased from Charles Rivers and fed ND *ad libitum* for 4 weeks. For sample collection, 3 mice from different groups were sacrificed each day at 9 am. Each group contained five mice.

High-fat diet study

For the high-fat diet study, 2-month-old female C57BL/6J mice were purchased from Charles River and fed *ad libitum* either a high-fat diet (Harlan TD. 06414) or a control diet (standard chow, RM1 (P) 801151) for 6 months. The high-fat diet contained 23.5% protein, 34.3% fatty acids, 27.3% carbohydrates and 5.1 kcal/g. Mice were weighed every two weeks. For sample collection, all mice were sacrificed the same day at 8 am. Each group contained four mice.

Newborn fibroblast subpopulation study

For the transcriptome analysis of papillary, reticular, and pro-adipogenic fibroblasts isolated from newborn mice, three female littermates at the age of P1.5- 2.5 were used.

Fibroblast aging in vivo compared to culture

For the aging study comparing *in vivo* with cultured fibroblasts four two-month-old and four 18-month-old female mice bred and aged at the at the animal facilities of the Barcelona Science Park were used.

Aging study in males

For the aging study in males four two-month-old and four 18-month-old male C57/Bl6 mice bred and aged at the animal facilities of the Barcelona Science Park were used.

Single-cell RNA-sequencing

Single-cell RNA-seq of dermal fibroblasts was performed using female mice at P1.5-2.5 (newborn), two months (young) and 18 months (old). For each biological replicate, one newborn, one young and one old mouse were processed in parallel. Importantly, the newborn mouse used in Replicate 1 was 12-18 hr older than that used in Replicate 2. The mice were bred and aged at the animal facilities of the Barcelona Science Park.

Lineage tracing

To perform lineage tracing of papillary fibroblasts we crossed Lrig1-EGFP-IRES-CreERT2 mice with Rosa26-CAG-STOPflox-tdTomato mice. Heterozygous Lrig1-EGFP-IRES-CreERT2/ Rosa26-CAG-STOPflox-tdTomato mice were treated at P3-P4 with 1.5mg 4-hydroxytamoxifen (Sigma, H6278) dissolved in 100ul Acetone. To dissolve 4-hydroxytamoxifen in Acetone, the solution was heated to 40-50°C and vortexed. To avoid degradation, the 4-hydroxytamoxifen solution was always prepared right before treatment.

METHOD DETAILS

FACS sorting

Mice were sacrificed between 8 am and 10 am. Skin was shaved and then dissected, anagen patches were removed, and the subdermis was scraped off with a scalpel. The tissue was then chopped with a McILWAIN tissue chopper until skin pieces were smaller than 0.5 mm in diameter. Skin pieces were then digested in 20 mL DMEM-Ca (Cat#21068028) containing 10% FBS (GIBCO, Cat#10270) and 2.5 mg/ml collagenase 1 (Sigma, C0130) at 37°C for 1 hr (females) or 2 hr (males) on the shaker, after which 5 μg/ml DNase 1 (Sigma, AMPD1) was added, and the solution was incubated for 15 min at 37°C without shaking. Liberated cells were then sequentially filtered through a 100μm filter and 40μm filter (SPL life sciences). Centrifugation steps were carried out at 300 g for 10 min at 4°C. Cells were then stained with antibodies (as stated below) in approx. 500μl staining buffer per skin. Staining buffer contained EMEM (Lonza) and 15% chelated FBS (Bio-rad, Cat#142-2842). To isolate fibroblasts for the experiments of aging, caloric restriction, HFD, *in vivo* versus culture and aging in males, cells were stained with the following antibodies: CD45-biotin (1:200), CD117-biotin (1:100), EpCam-biotin (1:200), CD24-biotin (1:100), SA-APC (1:500), CD31-PerCP/Cy5.5 (1:200), CD34-FITC (1:50), and Pdgfrα-PE (1:200). To isolate fibroblasts for single-cell RNA-sequencing, the same FACS staining protocol was used but without using CD24 antibody for lineage selection, to avoid negative selection of fibroblasts with pro-adipogenic characteristics (Berry and Rodeheffer, 2013). To isolate the different fibroblast subpopulations from newborn skin, cells were stained with CD45-biotin (1:200), CD117-biotin (1:100), EpCam-biotin (1:200), CD31-biotin (1:200), SA-APC-Cy7 (1:400), CD34-FITC (1:50), Pdgfrα-APC (1:100), Dlk1-PE (1:1.5 in chelated FBS), CD26-PerCp-Cy5.5 (1:50), and Sca1-BV605 (1:300). To assess viability, cells were stained with 0.5μg/ml DAPI (Sigma, D9542) in PBS with 2% chelated FBS immediately prior to sorting. Fibroblasts were sorted using a FACS ARIA Fusion instrument.

Fibroblast culture

For the transcriptome study comparing *in vivo* with cultured fibroblasts, 5000 Lin- Pdgfrα+ CD34+ fibroblasts from four old (18-month-old) and four young (2-month-old) female mice were directly sorted into lysis buffer (20 mM DTT, 10 mM Tris-HCl pH 7.4, 0.5% SDS, 0.5 μg/μl proteinase K) and another 5000 cells were FACS-sorted on one Collagen 1 (Corning, Cat# 354236)-coated 4cm^2- well containing Fibroblast medium (DMEM, 10% FBS, 1x Glutamax, 1x Pyruvate, 1x Pen/Strep, 1x Amphotericin B (GIBCO)). Cells were grown under physiological oxygen conditions (37C, 3% O2, 10% CO2) and the medium was changed every second day. After 10 days of culture, cells were detached from the surface using 0.25% Trypsin/EDTA (GIBCO) and 10.000 cells were reseeded on a 4cm^2 –well. After 3 more days of culture, when cells reached 70%–80% confluency, cells were detached from the surface using trypsin, washed with PBS 2% chelated FBS, stained with Dapi (0.5μg/ml), 5.000 Dapi- cells were FACS sorted into lysis buffer and processed together with the *in vivo* samples for microarray analysis as described below. For the clonogenic assay 1000, 3000 and 8000 fibroblasts isolated from young and old mice were directly sorted onto one 4cm2 - well. To assess clonogenic potential of cells after one passage, 1000, 2000 and 3000 fibroblasts were directly sorted onto 4cm2 - well. Cells were grown until separate clones started to merge, fixed in 10% NBF for 30 min and stained with 0.01% crystal violet (Sigma) / 1% NBF in PBS over night. Stained cell clones were then washed with ddH2O and dried.

Microarrays

To perform the aging study, RNA was isolated using TRIzol (ThermoFisher Scientific, A33250) followed by purification via the Quiagen RNeasy Mini columns (Cat#74106). 100μl Chlorophorm were added to 500μl TRIzol, incubated for 5 min at RT and centrifuged for 15min at max. speed. The aqueous phase was mixed with 3.5V RLT buffer (provided by RNeasy kit) and 2.5V 100% EtOH and added to the RNeasy columns. The subsequent steps were performed according to the manufacturer's protocol. cDNA library preparation and amplification were performed from 25 ng total RNA using WTA2 (Sigma-Aldrich), with 18 cycles of amplification. For the other experiments, cells were directly sorted into lysis buffer (20 mM DTT, 10 mM Tris-HCl pH 7.4, 0.5% SDS, 0.5 μg/μl proteinase K) and digested for 15 min at 65°C. RNA was isolated using magnetic beads (RNACleanXP, Cat#A63987). cDNA Library preparation and amplification were performed using the WTA2 kit (Sigma). For the transcriptome analysis of fibroblast subpopulations from newborn, 4.500 cells were sorted and cDNA was amplified for 21 cycles. For the caloric restriction and high-fat diet study, 10.000 cells were sorted and cDNA was amplified for 20 cycles. And for both the aging study in males and the aging study comparing *in-vivo* with cultured fibroblasts, 5.000 fibroblasts were sorted and cDNA was amplified for 21 cycles and 23 cycles, respectively.

Samples were then purified using PureLink Quick PCR Purification Kit (Thermo Fisher Scientific). cDNA from each sample (8μg) was subsequently fragmented by DNaseI and biotinylated by terminal transferase obtained from GeneChip Mapping 10K v2 Assay Kit (Affymetrix). Hybridization mixtures were prepared according to the Gene Atlas protocol (Affymetrix). Each sample target was hybridized to a Mouse Genome 430 PM array. After hybridization for 16 hr at 45°C, samples were washed and stained in the GeneAtlas Fluidics Station (Affymetrix). Arrays were scanned in a GeneAtlas Imaging Station (Affymetrix). All processing was performed according to manufacturer's recommendations. CEL files were generated from DAT files using Affymetrix Command Console software. Microarray processing was performed at IRB Barcelona Functional Genomics Core Facility.

Single-cell RNA-sequencing

We performed two independent single-cell RNA-sequencing experiments of Pdgfrα+ CD34+ fibroblasts isolated from newborn (P1.5-P2.5), young (2-months-old) and old (18-months-old) mice. The two Experiments or Replicates are termed Replicate 1 and Replicate 2. The newborn mouse used for Replicate 2 was 12-18 hr younger than that used for Replicate 1.

Full-length single-cell RNA-seq libraries were prepared using the Smart-seq2 protocol (Picelli et al., 2013) with minor modifications. Briefly, freshly harvested single cells were sorted into 96-well plates containing the lysis buffer (0.2% Triton-100). Reverse transcription was performed using SuperScript II (ThermoFisher Scientific) in the presence of oligo-dT30VN, template-switching oligonucleotides, and betaine. cDNA was amplified using the KAPA Hifi Hotstart ReadyMix (Kappa Biosystems) and ISPCR primer, with 23 cycles of amplification. Following purification with Agencourt Ampure XP beads (Beckmann Coulter), product size distribution and quantity were assessed on a Bioanalyzer using a High Sensitivity DNA Kit (Agilent Technologies). A total of 200 ng of the amplified cDNA was fragmented using Nextera XT (Illumina) and amplified with indexed Nextera PCR primers. Products were purified twice with Agencourt Ampure XP beads and quantified again using a Bioanalyzer High Sensitivity DNA Kit. Pooled sequencing of Nextera libraries was carried out using a HiSeq2000 (Illumina) to an average sequencing depth of 1.2 million reads per cell in experiment 1, and 1.4 million reads per cell in experiment 2. Sequencing was carried out as paired-end (PE75) reads with library indexes corresponding to cell barcodes. After sequencing, libraries were inspected with the FastQC suite (Andrews, 2010) to assess the quality of the reads. Reads were then demultiplexed according to the cell barcodes and mapped on the mouse reference genome (Gencode release M15, assembly GRCm38 (Mudge and Harrow, 2015)) with the RNA pipeline of the GEMTools 1.7.0 suite (Marco-Sola et al., 2012) using default parameters (6% of mismatches, minimum of 80% matched bases, and minimum quality threshold of > 26). Low-quality cells were filtered out based on the distribution of the number of non-zero count genes per cell (minimum number of genes detected), to remove cells with less than two median absolute deviations (MAD) with respect to the median. Genes expressed in less than five cells, poorly annotated genes (GM identifiers), and histone gene clusters were also discarded from the dataset. The final dataset included 385 cells and 11,327 genes for experiment 1, and 287 cells and 11,303 genes for experiment 2. Data analysis was performed in R, version 3.4.2.

qRT-PCR

For qPCR validation of differentially expressed genes between old and young fibroblasts, we used the amplified cDNA of three biological replicates. Gene expression was quantified by quantitative real-time PCR using Sybrgreen Mastermix (Roche, Cat#04707516001) and the specific primer listed in Table S1 sheet 20. A LightCycler 480 instrument (Roche) was used. Three biological replicates were used in each assay. The expression of each gene was normalized to the housekeeping gene B2M.

Immunofluorescence

Mouse telogen hind back skin (a 1cm2 area 0.5-1cm away from the tail) was fixed in 10% NBF (Sigma, Cat#HT501128) for 4 hr at room temperature, or for 24 hr at 4°C, and then processed for embedding in paraffin blocks. Antigen retrieval was performed for 20 min at 97°C with citrate (pH 6) or Tris (pH 9) on 5-micron tissue sections as stated below.

For frozen sections, mouse back skin was fixed for 24 hr at 4°C in 4% Paraformaldehyde (Electron Microscopy Sciences, Cat#15710), incubated in 30% Sucrose in PBS for 24hr at 4°C, quickly rinsed in PBS, dried and frozen on dry ice in O.C.T Compound (Tissue-Tek, Cat#4583). 10-micron tissue sections were prepared using a cryostat, sections were dried for 30 min at RT and washed for 30 min in PBS to remove the O.C.T.

For some immunostainings sections were permeabilized in PBS 0.05% Triton or 0.1% Tween for 10min as stated below. Subsequently, sections were blocked in PBS containing 10% donkey serum (Sigma, Cat#D9663) and 2% BSA (Sigma) for 1 hr at room temperature. For primary antibodies raised in mouse, sections were additionally blocked with mouse-on-mouse blocking reagent (MKB-2213, Vector Laboratories). Primary antibody incubation was done overnight at 4°C in DAKO Envision Flex Antibody diluent (Cat# K8006). The following conditions were used for the different antibodies for paraffin sections (antibody/ permeabilization/ AG retrieval/ dilution): PPARγ H-100/ 0.05% Triton/ Tris (pH9)/ 1:150—Fabp4/ 0.1% Tween/ citrate (pH6)/ 1:200—RFP-dTomato/0.1% Tween/ citrate (pH6)/ 1:400 (adult) or 1:100 (newborn)—Keratin 14/ 0.1% Tween or none/ citrate (pH6)/ 1:1000—Vimentin/ 0.1% Tween or none/ citrate (pH6); 1:400—Bcat1/ none/ citrate (pH6)/ 1:100—Pparδ / 0.1% Tween / citrate (pH6)/ 1:200. The following conditions were used for the different antibodies for OCT sections (antibody/ permeabilization/ dilution): Pdgfrα/ none/ 1:50—CD34-Biotin/ 0.1% Tween or none/ 1:50—α-smooth muscle actin/ 0.1% Tween/ 1:200.

Secondary antibody incubation was done at room temperature for 2 hr. Secondary antibodies used were anti-donkey rabbit Alexa Fluor 488, donkey anti-mouse Alexa Fluor 568, goat anti-chicken Alexa Fluor 647 (concentration 1:500) and donkey anti-guineapig Cy3 (concentration 1:200). Nuclei were stained with DAPI (0.5μg/ml in PBS), and sections were mounted in DAKO fluorescent mounting medium (Cat#S3023). Washes were done with PBS.

Microscopy and image analysis

Fluorescence pictures of four biological replicates were acquired using a Leica TCS SP5 confocal microscope (63 × /1.40 oil objective or 40 x /1.25 oil objective at 1024 × 1024 pixel resolution) and processed using the Fiji v2.0.0-rc-14/1.49 g software (ImageJ).

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification of skin morphology features

Skin samples were taken from telogen hind backskin (a 1cm2 area 0.5-1cm away from the tail), sectioned at 6μm and stained with Hematoxylin/Eosin (Dako, Cat#S202084). Brightfield images were acquired with a NanoZoomer-2.0 HT C9600 scanner (Hamamatsu, Photonics, France) equipped with a 20X objective. Images were visualized with a gamma correction set at 1.8 in the image control panel of the NDP.view 2 U123888-01 software (Hamamatsu, Photonics, France). Dermal, subdermal, and subcutaneous thicknesses were measured using the "Ruler" tool from the NDP.view2 software. At least 10 measurements were taken from each replicate, and the means of the different measurements were calculated. Regions of low structural integrity were excluded from the analysis. Four biological replicates and two technical replicates (e.g., two pieces of the same skin that were embedded separately) were used for the analysis. Statistical analysis was performed using Prism 6 software (Pad Software). Experimental groups were compared by unpaired t test. Error bars represent ± standard deviation.

qRT-PCR analysis

Statistical analyses were performed using Prism 6 software (Pad Software). Experimental groups were compared by unpaired t test using the ddct values (log-transformed fold changes). Error bars represent ± standard deviation.

Microarrays

Quality control and normalization

Microarray samples from each experiment were processed separately using packages *affy* and *affyPLM* from R (R Foundation for Statistical Computing, Vienna, no date; Gautier et al., 2004; Gentleman et al., 2004). Raw CEL files were normalized using RMA background correction and summarization (Irizarry et al., 2018). Technical metrics described by Eklund AC and Szallasi Z ((Eklund and Szallasi, 2008) were computed and recorded as additional features for each sample. Standard quality controls were performed in order to identify abnormal samples and relevant sources of technical variability (Gentleman, R., Carey, V., Huber, W., Irizarry, R., Dudoit, 2005) regarding: a) spatial artifacts in the hybridization process (scan images and pseudo-images from probe level models); b) intensity dependences of differences between chips (MvA plots); c) RNA quality (RNA digest plot); d) global intensity levels (boxplot of perfect-match log-intensity distributions before and after normalization and RLE plots); e) anomalous intensity profile compared to the rest of samples (NUSE plots, Principal Component Analyses). No samples were excluded according to the results of these quality control checks. Chip probesets were annotated using the information provided by Affymetrix (https://www.affymetrix.com/analysis/index.affx).

Please cite this article in press as: Salzer et al., Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging, Cell (2018), https://doi.org/10.1016/j.cell.2018.10.012

Differential expression and visualization

Group comparisons in microarray experiments were performed using a linear model with empirical shrinkage (Smyth, 2004) as implemented in the limma R package (Ritchie et al., 2015). Relevant sources of technical variability identified in the quality control process were included in these models as covariates (aging, HFD, *in vivo* vs. culture, male datasets: scanning batch; CR dataset: scanning batch and Eklund metrics; newborn cell population dataset: biological replicates). Adjustment by multiple contrasts was performed by the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). For age group comparisons, samples from the aging dataset, and normal diet samples from the CR dataset, were reprocessed and analyzed using scanning batch as confounding factor. This "combined aging dataset" was used for all analyses involving the transcriptional changes in old compared to young fibroblasts, unless otherwise stated. For each experiment, samples were displayed in the first principal components after *a priori* correction by the relevant technical variables. For such corrections, a linear model was fitted gene-wise in which the groups of interest were also included as explanatory variables. Regarding the HFD dataset one ctrl diet sample, which did not cluster with the other samples from its biological group, was excluded when performing downstream analyses.

Biological significance analysis

Pathway enrichment was assessed through the pre-ranked version of Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) and through Gene Ontology Analysis using David 6.8 Software (GO) (Huang, Sherman and Lempicki, 2009). GSEA was applied to the rankings defined by the t-statistic of the differential expression analysis described in the previous section. Gene sets for analyses were derived from the KEGG pathway database (Ogata et al., 1999), from those annotated under Gene Ontology (GO) (Ashburner et al., 2000) terms as collected in the *org.Mm.eg.db* R package (Carlson, 2016) or from the Hallmark collection (Liberzon et al., 2015) after retrieval from the MsigDB (Liberzon et al., 2011) and translation to mouse homologous using *biomaRt* (Durinck et al., 2009). GSEA was performed on each of these gene set collections separately. Expression data were summarized to gene level using probesets with the highest median absolute deviation within each gene.

For Gene Ontology analyses Affymetrix IDs were used as input that showed a FC > 1.5 and an FDR < 0.05 in the combined aging dataset and male aging dataset (Figure 1G, Table S1 and Figure S5E, Table S2, respectively) and a FC > 2 and an FDR < 0.01 for the fibroblast subpopulation comparisons (Figures 3C–3E, Table S4). To identify categories that are changed with age, both up- and downregulated genes at a given threshold were used as an input. To perform GO of genes differentially expressed between fibroblast clusters identified by single-cell RNA-sequencing, marker genes were used as an input identified with the MAST method (see below) that are significant at a p value < 0.005 in either replicate 1 or replicate 2 (Figure 5E, Table S5 sheet 18,19). Signal transduction pathway association analysis was performed with Genomatix (http://www.genomatix.de (2016)), using the Gene Ranker package.

For visualization purposes, the most relevant gene sets were represented at the gene level using an unsupervised approach: for each gene set, a hierarchical clustering was performed on the most variable features within each gene according to the median absolute deviation across all samples. Before that, technical-corrected gene expression values were previously centered and scaled across samples. Ward's criterion was used as agglomeration method while Euclidean and correlation distance was selected as distance measure for samples and genes, respectively. Clustering results were shown in a heatmap in which the intensity of the colors represented the expression intensity from blue (low expression) to red color (high expression).

Global transcriptomic comparison

Transcriptomic differences found in each of the microarray experiments were summarized at the gene set level using the differential expression results as follows: In the CR dataset probesets were selected that showed a FC > 1.25 and an FDR < 0.01 for the Old CR versus Old ND comparison (131 probesets, 99 genes) and a FC > 1.25 and a p value < 0.001 for the Old CR-ND versus ND comparison (131 probesets, 98 genes). In the HFD dataset probesets were selected that showed a FC > 1.5 and an FDR < 0.05 (227 probesets, 188 genes). In the newborn fibroblast subpopulation dataset, a GaGa analysis (Rossell, 2009) was performed at the probeset level in order to identify genes that were over- or under-expressed in each cell population as compared to the rest of populations. Subsequently, for the CD26+ papillary fibroblasts probesets were selected that showed a FC > 3 (585 probesets, 324 genes), for the Dlk1+ reticular fibroblasts that showed a FC > 2 (131 probesets, 86 genes), for the Sca1+ pro-adipogenic fibroblasts that showed a FC > 2 (129 probesets, 83 genes) and for the Sca1+/Dlk1+ pro-adipogenic fibroblasts that showed a FC > 1.5 (207 probesets, 134 genes) compared to each other fibroblast population. Thresholds were chosen dependent on the adjusted p value, the fold change and the number of genes within each signature, such that each signature contained at least 100 genes. Importantly, using different thresholds to define a signature did not affect the downstream result. Gene sets were summarized as gene expression signatures and compared across sample groups in the rest of the experiments. For summarization at signature level, a *Z* score was computed for each gene and sample; these were then averaged across all genes included in the gene set, and the resulting score was then centered and scaled across samples. Genes over- and under-expressed were combined in the same signature after changing the sign of the latter prior to averaging them across samples. These scores were computed in each dataset separately after *a priori* correction by technical effects as described in previous sections. Signature comparisons between groups were carried out using linear regression in which these technical variables were included as covariates.

Single-cell RNA-sequencing analysis

Normalization and marker identification

Gene expression levels for each cell were normalized by the total expression, multiplied by a scale factor (10,000), and log-transformed. Batches were then regressed out, and scaled *Z* scored residuals of the model were used as normalized expression values.

We defined the 10% most variable genes based on their average expression and dispersion as highly variable genes (HVG). We reduced the dimensionality of the data by performing principle component analysis (PCA) on HVG. To find fibroblast subpopulations, clustering was performed on PCA scores using significant PCs assigned by a randomization approach proposed by Chung and Storey ("jack straw") (Chung and Storey, 2015; Macosko et al., 2015) and the amount of variance explained by them (Figures S7D–S7M). In detail, we visualized the standard deviation of each PC on an elbowplot and drew a cut-off where the curve became asymptotic to the x axis (Figure S7D, E, H, I). Also, in order to determine the significance of each PC a resampling test was performed by randomly permuting a subset of that data (1%) and repeating PCA (Figure S7F, G, J, K). We repeated this procedure for 100 times for the first 10 PCs and selected the significant ones. For both replicates (1 and 2), the first 7 PCs were selected for clustering. To cluster cells, a K-nearest neighbor (KNN) graph constructed on a Euclidean distance matrix in PCA space was calculated and then converted to a shared nearest neighbor (SNN) graph, in order to find highly interconnected communities of cells (Xu and Su, 2015). Cells were then clustered using the Louvain method to maximize modularity (Waltman and Van Eck, 2013). To display data, the *t*-distributed stochastic neighbor embedding (t-SNE) was applied to cell loadings of selected PCs, and the cluster assignments from the graph-based clustering were used. All analyses described in this section were performed using Seurat R package version 2.1.0.

To find marker genes specific for each cluster, a GLM-based method for single-cell differential expression analysis (MAST) was used; it accounts for the bimodality of single-cell data by jointly modeling rates of expression and positive average expression values before combining both models to infer changes in expression levels (Finak et al., 2015). MAST was run within an implementation of the Seurat package with default parameters. The full list of cluster markers is provided in Table S5.

Trajectory analysis

Trajectory analysis was performed using Monocle version 2.6.1. (Trapnell et al., 2014; Qiu et al., 2017). After creating a Monocle object using "negbinomial.size()" distribution, the analysis was performed on HVG selected from previous steps. Dimensionality of data was reduced using the "DDTree" method, and the seven PCs that entered the clustering step were also used to create the trajectories. Finally, the cluster annotations were projected on the inferred trajectories, to better visualize the aging-related processes.

Comparison of cell clusters to NB fibroblast subpopulations

We compared the signature of NB fibroblast subpopulations identified by GaGa to the fibroblast clusters by averaging the expression of each gene set in each cell and plotting the distribution of these average expression points of all cells of a cluster in a boxplot (Figures 5D and 6G).

Distance of cell clusters on PC1

To assess the distance of cell clusters on PC1, we calculated the average distance for each pair of cells from opposite clusters (average linkage) for old and for young excluding old cells assigned to a "young" cluster and young cells assigned to an "old" cluster (Figure 6B).

Average gene-based distance between cell clusters

For each HVG, we calculated the average distance between cell pairs of two clusters in young and old (Figure 6C).

So ∀x; x ∈ HVG, if we denote Y1 and Y2 as cluster 1 and 2 of young cells respectively, M as number of cells in Y1 and N as number of cells in Y2, then the vector of pairwise cell distances of gene x between Y1 and Y2 cells will be:

$$For\ i\ in\{1,\ldots,M\}\ and\ j\ in\{1,\ldots,N\}, d_x = \sqrt{(x_{Y1i} - x_{Y2j})^2}$$

Then, we calculated the arithmetic mean of the d_x as mean distance per gene denoted as μ_x. We repeated the same procedure for Old clusters.

Coefficient of variation

For each HVG, we calculated the coefficient of variation of the distances of that gene between cells of two clusters in young and old. The coefficient of variation (CV) is calculated as standard deviation divided by the mean distance (Figure 6D).

Standard deviation of the distances per gene

For each HVG, we calculated the standard deviation of the distances of that gene between cells of two clusters in young and old (cluster markers are highlighted in blue). As denoted above, ∀x; x ∈ HVG, we calculated the standard deviation of pairwise distances of x as $\sigma_x = SD(d_x)$.

Coefficient of variation (Figure 6D)

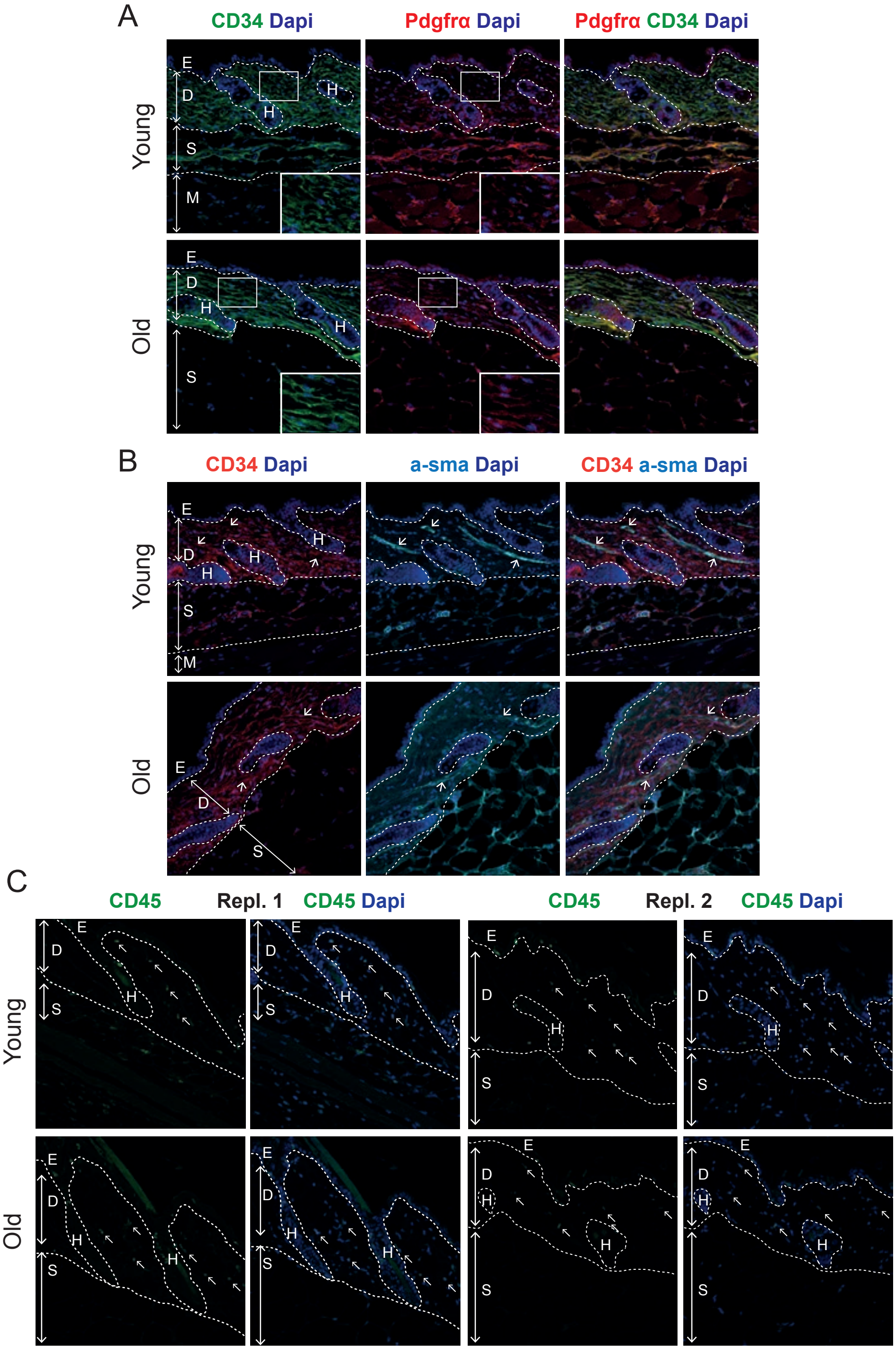
Finally, we calculated the coefficient of variation of pairwise distances for each gene as:

$$CV_x = \frac{\sigma_x}{\mu_x} * 100$$

DATA AND SOFTWARE AVAILABILITY

All array expression data files and single-cell RNA-seq files were uploaded to the NCBI GEO database (GSE110983, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE110983 and GSE111136, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111136).

Supplemental Figures



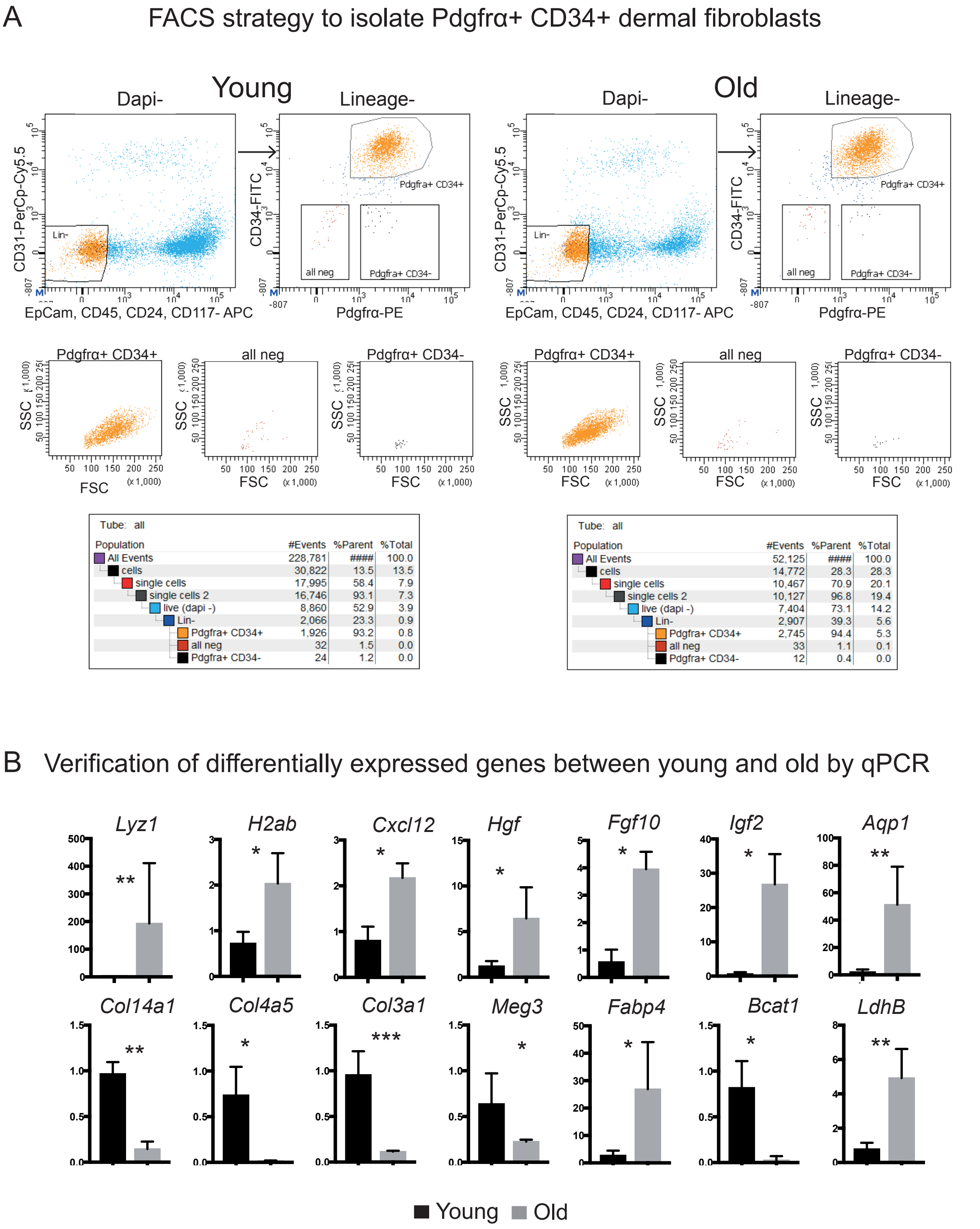
(legend on next page)

Figure S1. Expression of the Fibroblast Marker CD34 and Pdgfrα, the Smooth Muscle Cell Marker α-Smooth Muscle Actin and the Immune Cell Marker CD45 across Old and Young Murine Dermis, Related to Figure 1

(A) Immunohistochemistry of Pdgfrα and CD34 on young and old skin sections. Both Pdgfrα and CD34 are expressed in cells throughout the dermis in young skin (upper panel) and old skin (lower panel).

(B) Immunohistochemistry of CD34 and α- smooth muscle actin on young and old skin sections. The α-smooth muscle actin + arrector pili muscle (indicated with arrows) is negative for CD34 in young skin (upper panel) and old skin (lower panel).

(C) CD45 staining in young and old skin. We did not observe a notable increase in CD45+ immune cells in old dermis. E, epidermis; H, hair follicle; D, dermis; S, subdermis.



Young

Tube: all

Population	#Events	%Parent	%Total
All Events	228,781	####	100.0
cells	30,822	13.5	13.5
single cells	17,995	58.4	7.9
single cells 2	16,746	93.1	7.3
live (dapi -)	8,860	52.9	3.9
Lin-	2,066	23.3	0.9
Pdgfra+ CD34+	1,926	93.2	0.8
all neg	32	1.5	0.0
Pdgfra+ CD34-	24	1.2	0.0

Old

Tube: all

Population	#Events	%Parent	%Total
All Events	52,125	####	100.0
cells	14,772	28.3	28.3
single cells	10,467	70.9	20.1
single cells 2	10,127	96.8	19.4
live (dapi -)	7,404	73.1	14.2
Lin-	2,907	39.3	5.6
Pdgfra+ CD34+	2,745	94.4	5.3
all neg	33	1.1	0.1
Pdgfra+ CD34-	12	0.4	0.0

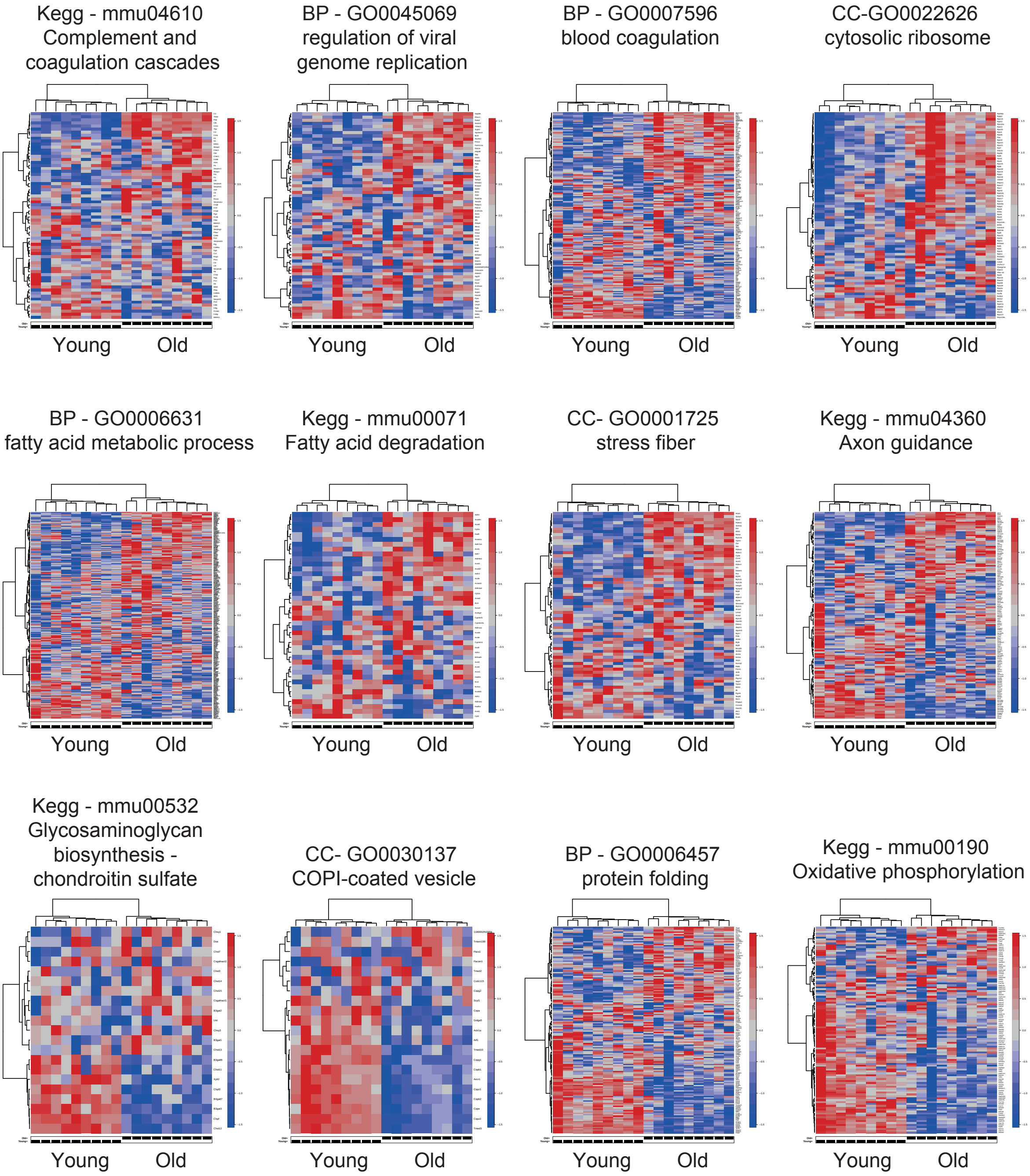
(legend on next page)

Figure S2. FACS Plots of Skin Cells from Young and Old Mice and qPCR Validation of Differentially Expressed Genes Identified by Transcriptome Analysis, Related to Figure 1

(A) FACS strategy to isolate dermal fibroblasts. Cells were first selected based on the FSC and SSC, potential duplexes were excluded that had a high FSW or SSW and dead cells were excluded that were Dapi+. We further excluded CD31+ endothelial cells, EpCam+ epithelial cells, CD45+ immune cells, CD117+ melanocytes and CD24+ pre-adipocytes. Lineage- cells were plotted for Pdgfrα and CD34. > 90% of Lineage- cells were positive for CD34 and Pdgfrα. Cells or Events that were negative for CD34 or both fibroblast marker had a lower FSC profile indicating that they are different cell types (lower panel).

(B) Verification of the differential expression of selected genes between young and old fibroblasts by quantitative PCR using three biological replicates. Error bars represent mean ± SD.

Unsupervised clustering of young and old fibroblast samples based on gene ontology signatures

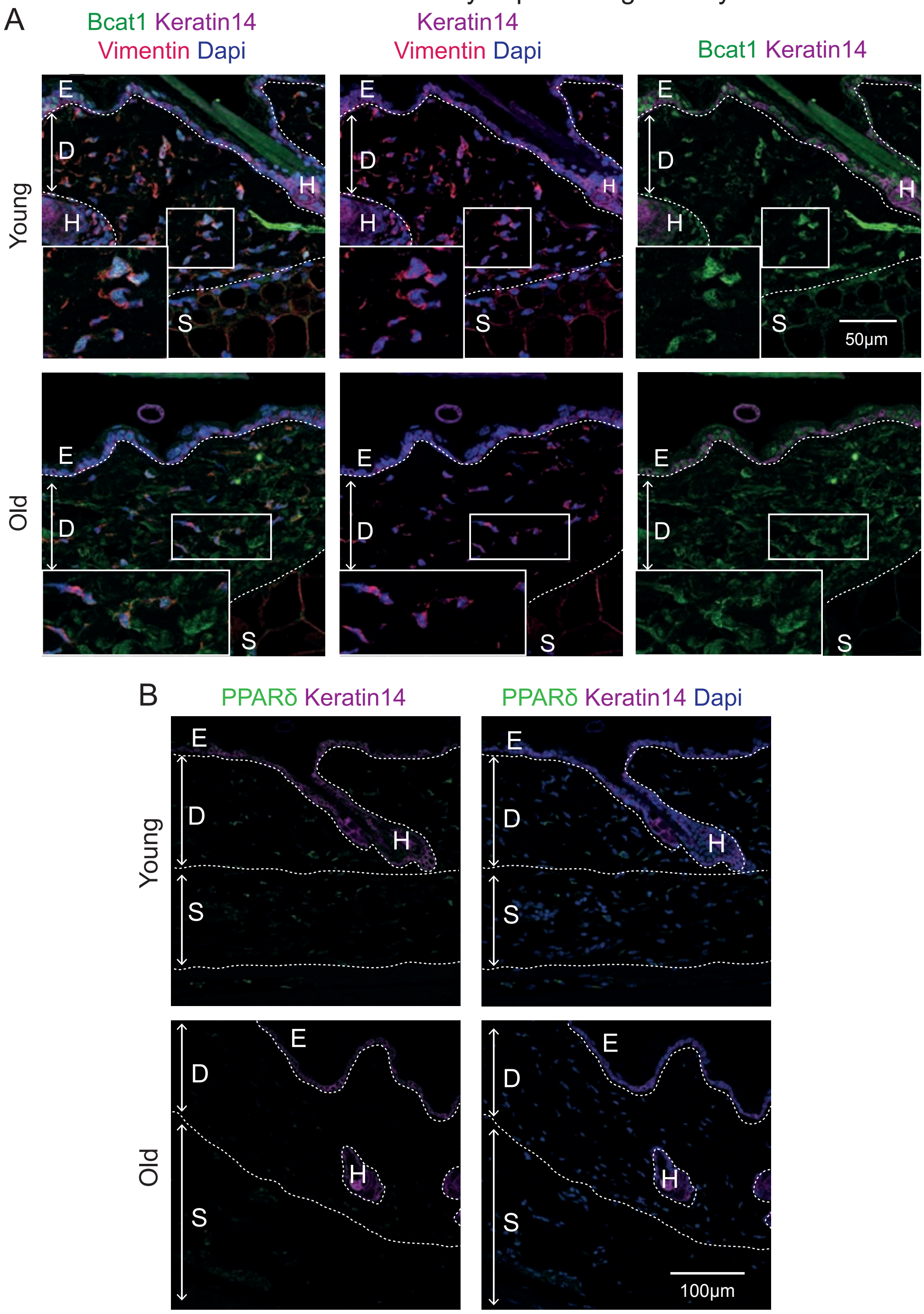


(legend on next page)

Figure S3. Unsupervised Clustering of Young and Old Fibroblast Samples Isolated from Female Mice Based on Gene Ontology Signatures, Related to Figure 1

Clustering results are shown in a heatmap in which the intensity of the colors represents the expression intensity from blue (low expression) to red color (high expression). The respective GO category is indicated above each heatmap.

Verification of differentially expressed genes by IHC

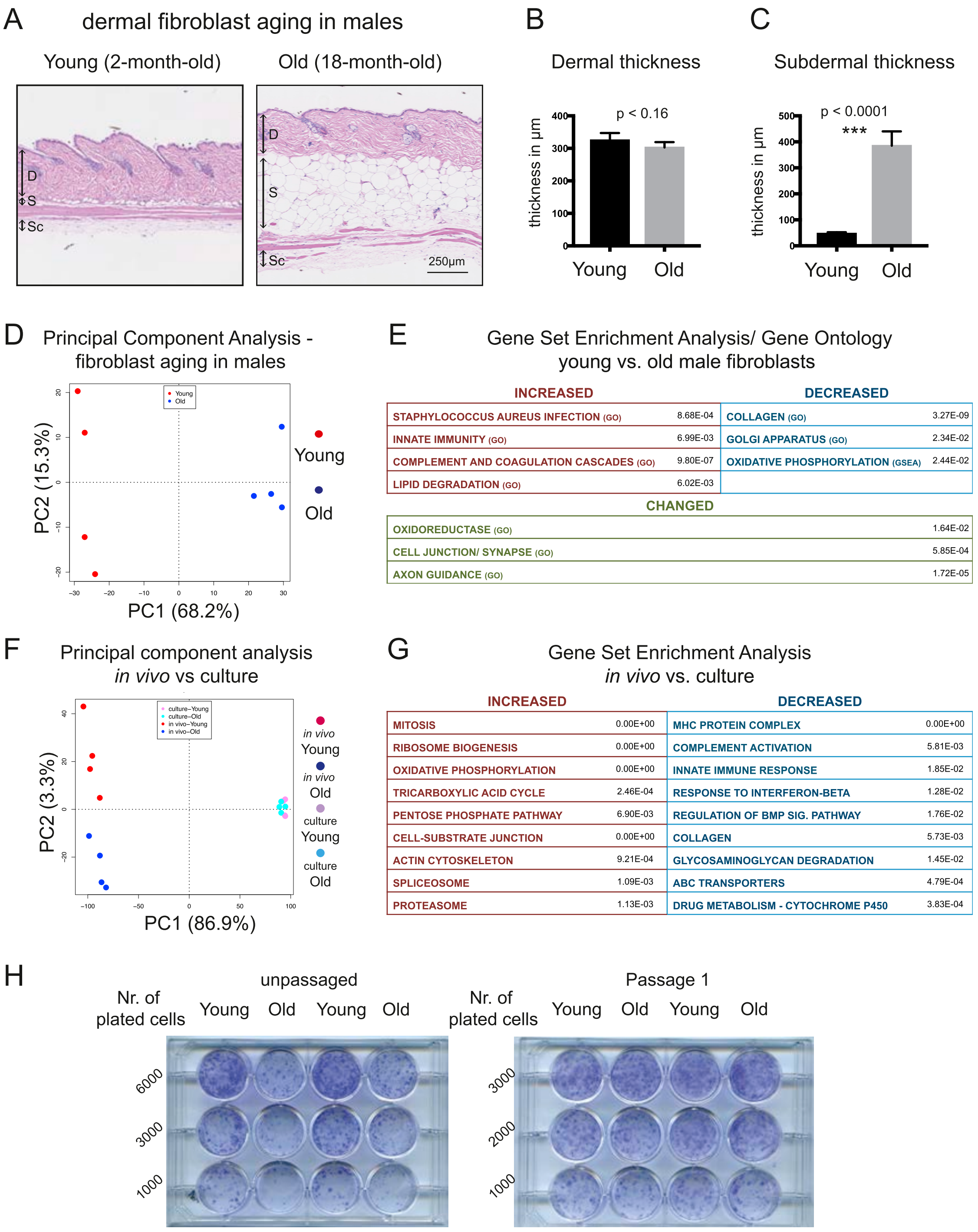


(legend on next page)

Figure S4. Verification of Differentially Expressed Genes between Old and Young Dermal Fibroblasts by Immunohistochemistry, Related to Figure 1

(A) Immunohistochemistry of Bcat1 (old versus young: FC = −24, FDR = 6ˆ-10, green), Keratin 14 (epidermal cells, pink), Vimentin (stromal cells, red) and Dapi (nuclei, blue). Bcat1 immunoreactivity is observed in young, but barely detected in old dermal fibroblasts.

(B) Immunohistochemistry of Pparδ (old versus young: FC = −1.47, FDR = 0.01, green), Keratin 14 (epidermal cells, pink), and Dapi (nuclei, blue). Pparδ immunoreactivity is higher in young than in old dermal fibroblasts. E, epidermis; D, dermis; S, subdermis; H, hair follicle



E — Gene Set Enrichment Analysis/ Gene Ontology young vs. old male fibroblasts

INCREASED		DECREASED	
STAPHYLOCOCCUS AUREUS INFECTION (GO)	8.68E-04	COLLAGEN (GO)	3.27E-09
INNATE IMMUNITY (GO)	6.99E-03	GOLGI APPARATUS (GO)	2.34E-02
COMPLEMENT AND COAGULATION CASCADES (GO)	9.80E-07	OXIDATIVE PHOSPHORYLATION (GSEA)	2.44E-02
LIPID DEGRADATION (GO)	6.02E-03		

CHANGED	
OXIDOREDUCTASE (GO)	1.64E-02
CELL JUNCTION/ SYNAPSE (GO)	5.85E-04
AXON GUIDANCE (GO)	1.72E-05

G — Gene Set Enrichment Analysis *in vivo* vs. culture

INCREASED		DECREASED	
MITOSIS	0.00E+00	MHC PROTEIN COMPLEX	0.00E+00
RIBOSOME BIOGENESIS	0.00E+00	COMPLEMENT ACTIVATION	5.81E-03
OXIDATIVE PHOSPHORYLATION	0.00E+00	INNATE IMMUNE RESPONSE	1.85E-02
TRICARBOXYLIC ACID CYCLE	2.46E-04	RESPONSE TO INTERFERON-BETA	1.28E-02
PENTOSE PHOSPHATE PATHWAY	6.90E-03	REGULATION OF BMP SIG. PATHWAY	1.76E-02
CELL-SUBSTRATE JUNCTION	0.00E+00	COLLAGEN	5.73E-03
ACTIN CYTOSKELETON	9.21E-04	GLYCOSAMINOGLYCAN DEGRADATION	1.45E-02
SPLICEOSOME	1.09E-03	ABC TRANSPORTERS	4.79E-04
PROTEASOME	1.13E-03	DRUG METABOLISM - CYTOCHROME P450	3.83E-04

(legend on next page)

Figure S5. Age-Related Changes in Dermal Fibroblasts Are Similar in Male Compared to Female Mice, but Are Lost in Cell Culture, Related to Figure 1

(A) Hematoxylin and eosin (H&E)-stained sections of young (2-month-old) and old (18-month-old) male murine skin. D, dermis; S, subdermis; Sc, subcutis.

(B and C) Quantification of dermal thickness (B) and subdermal thickness (C) at different ages using H&E-stained murine skin sections. Data are represented as mean ± SD.

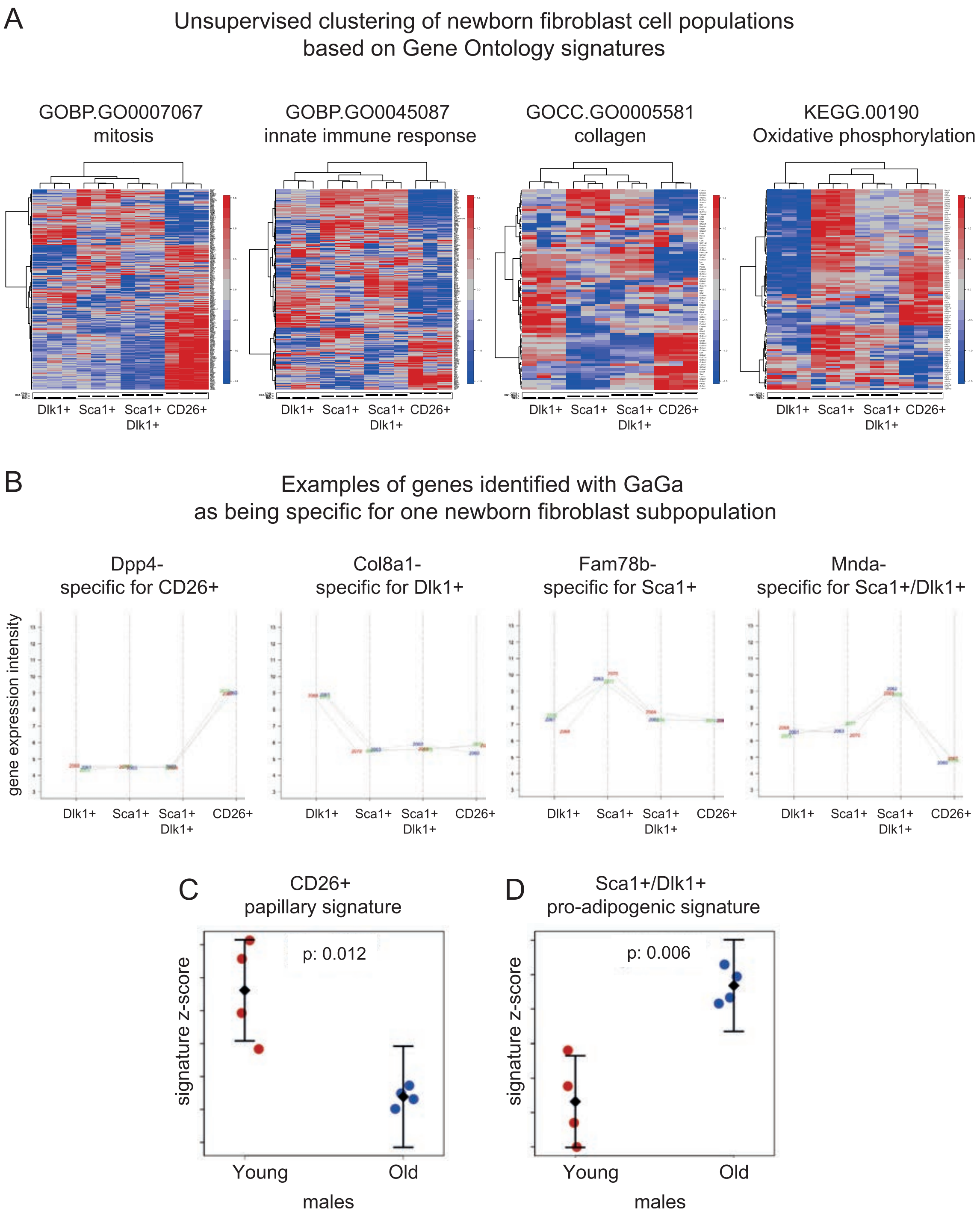
(D) Principal component analysis (PCA) of 4 young and 4 old dermal fibroblast samples isolated from male mice, according to their transcriptome.

(E) Gene set enrichment analysis (GSEA) and gene ontology (GO) analysis of genes differentially expressed between young and old male dermal fibroblasts. Categories derived from GO analysis using David 6.8 are marked with "(GO)," categories derived from GSEA are indicated with "(GSEA)." For the GSEA categories the FDR is indicated; for the GO categories, the p value is indicated.

(F) PCA analysis of the transcriptomes of old and young dermal fibroblasts directly lysed and processed after FACS isolation (*in vivo*), and after one passage in culture (culture). Gene expression changes that characterize dermal fibroblast aging *in vivo* are lost when fibroblasts are placed in culture.

(G) GSEA comparing fibroblasts *in vivo* with fibroblasts (isolated from the same mice) after one passage in culture.

(H) Clonogenic assay of young and old dermal fibroblasts at Passage 0 (directly after isolation) and after one passage.



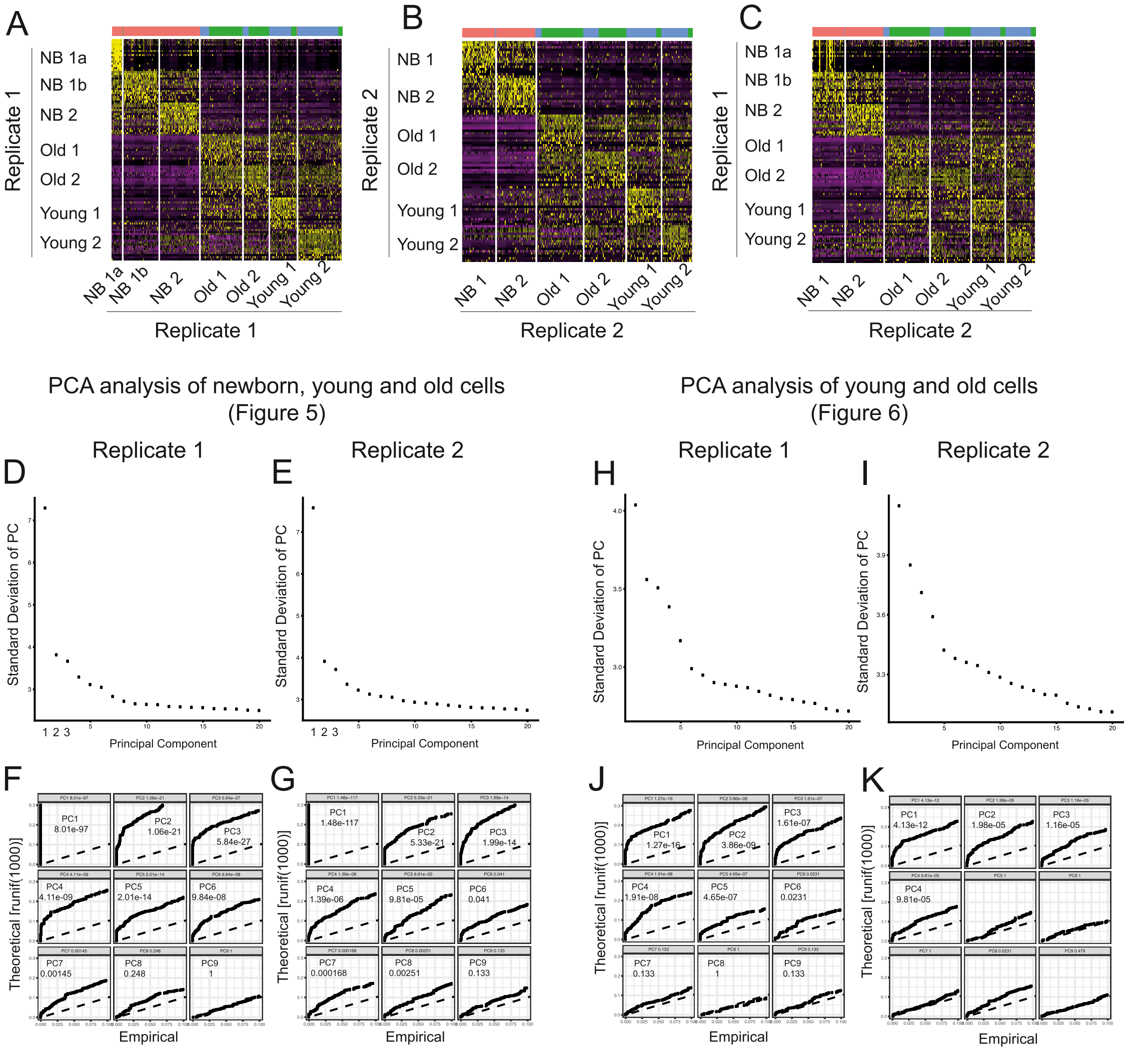
(legend on next page)

Figure S6. Identification of Different Types of Newborn Fibroblast Signatures, Related to Figure 3

(A) Unsupervised clustering of newborn fibroblast subpopulations based on gene ontology signatures. Clustering results are shown in a heatmap in which the intensity of the colors represents the expression intensity from blue (low expression) to red color (high expression). The respective GO category is indicated above each heatmap.

(B) Examples of the expression profiles of genes selected by the GaGa algorithm across the four fibroblast subpopulations.

(C and D) The gene expression profile of old male dermal fibroblasts negatively associates with the signature of CD26+ papillary fibroblasts (B) and positively associates with the signature of Sca1+ Dlk1+ pro-adipogenic fibroblasts (C). Expression values of genes specific for a fibroblast subpopulation were summarized (signature *Z* score) and compared across old and young samples. Each dot represents one old or young sample. Error bars represent 95% confidence interval.



L % variance explained by each Principal Component

PC	**1**	**2**	**3**	**4**	**5**
Repl. 1	25.8	7.08	6.53	5.26	4.71
Repl. 2	24.5	6.54	5.91	4.84	4.45
PC	**6**	**7**	**8**	**9**	**10**
Repl. 1	4.51	3.89	3.59	3.43	3.39
Repl. 2	4.18	4.04	3.99	3.78	3.59
PC	**11**	**12**	**13**	**14**	**15**
Repl. 1	3.38	3.27	3.25	3.22	3.2
Repl. 2	3.63	3.58	3.51	3.45	3.38
PC	**16**	**17**	**18**	**19**	**20**
Repl. 1	3.14	3.13	3.11	3.05	3.03
Repl. 2	3.36	3.35	3.29	3.28	3.22

M % variance explained by each Principal Component

PC	**1**	**2**	**3**	**4**	**5**
Repl. 1	8.91	6.93	6.72	6.26	5.49
Repl. 2	7.58	6.53	6.07	5.68	5.16
PC	**6**	**7**	**8**	**9**	**10**
Repl. 1	4.89	4.75	4.60	4.56	4.53
Repl. 2	5.04	4.98	4.93	4.83	4.76
PC	**11**	**12**	**13**	**14**	**15**
Repl. 1	4.50	4.42	4.34	4.28	4.26
Repl. 2	4.67	4.62	4.57	4.51	4.50
PC	**16**	**17**	**18**	**19**	**20**
Repl. 1	4.21	4.19	4.08	4.04	4.04
Repl. 2	4.39	4.34	4.31	4.27	4.27

(legend on next page)

Figure S7. Statistical Background of Single-Cell RNA-Seq Data Analysis, Related to Figure 5 and 6

(A–C) Heatmap showing transcriptome similarities between single fibroblasts isolated from newborn, young, and old mice from two independent single-cell RNA-seq experiments termed Replicate 1 and Replicate 2. The expression levels of the 15 most significant markers identified for each cluster (y axis) are plotted for each cell (x axis). Markers identified in Replicate 1 are plotted against each cell from Replicate 1 in (A), those identified in Replicate 2 are plotted against each cell from Replicate 2 in (B), and those identified in Replicate 1 are plotted against each cell from Replicate 2 in (C).

(D and E) Elbowplot showing the standard deviation of each Principal Component (PC) when performing Principal Component Analysis (PCA) on newborn, young, and old cells in Replicate 1 (D) and Replicate 2 (E).

(F and G) Plots showing the significance of the variation of each PC after performing a resampling test by randomly permuting a subset of the data (1%) and repeating PCA (JackStraw method) for the first 10 PCs for Replicate 1 (F) and Replicate 2 (G) when performing PCA on newborn, young, and old cells. PC1- PC7 were used for clustering.

(H and I) Elbowplot showing the standard deviation of each PC when performing PCA on young and old cells in Replicate 1 (H) and Replicate 2 (I).

(J and K) Plots showing the significance of the variation of each PC after performing a resampling test by randomly permuting a subset of the data (1%) and repeating PCA (JackStraw method) for the first 10 PCs for Replicate 1 (J) and Replicate 2 (K) when performing PCA on young and old cells.

(L and M) Table showing the % of variance explained by each PC when performing PCA on newborn, young, and old cells (L), and young and old cells (M).

Chapter III

Benchmarking single cell RNA-seq methods

A variety of scRNA-seq protocols have been developed and their utility proven in single-cell transcriptome analysis of complex and dynamic tissues. The available protocols vary in the efficiency of RNA molecule capture, resulting in differences in sequencing library complexity and sensitivity in identifying transcripts and genes. However, there has been no systematic testing of how their performance varies between cell types, and how this affects the resolution of cell phenotyping in complex samples. In this paper, to extend current efforts to compare the molecule capture efficiency of scRNA-seq protocols, we have systematically evaluated the power of these techniques to describe tissue complexity, and their suitability for creating a cell atlas. In this project, we performed a multi-center benchmarking study to compare scRNA-seq protocols using a unified reference sample resource. By analyzing human peripheral blood and mouse colon tissue, we have covered a broad range of cell types and states, in order to represent common scenarios in cell atlas projects. We have also added spike-in cell lines to allow us to assess batch effects, and have combined different species to pool samples into a single reference. We performed a comprehensive comparative analysis of 13 different scRNA-seq protocols, representing the most commonly used methods. We applied a wide range of different quality control metrics to evaluate datasets from different perspectives, and to test their suitability for producing a reproducible, integrative and predictive reference cell atlas. Mereu & Lafzi *et al.*, Accepted in Nature Biotechnology; Preprint available in BioRxiv [79]

Benchmarking Single-Cell RNA Sequencing Protocols for Cell Atlas Projects

Elisabetta Mereu[1,+], Atefeh Lafzi[1,+], Catia Moutinho[1], Christoph Ziegenhain[2], Davis J. MacCarthy[3,4,5], Adrian Alvarez[6], Eduard Batlle[6,7,8], Sagar[9], Dominic Grün[9], Julia K. Lau[10], Stéphane C. Boutet[10], Chad Sanada[11], Aik Ooi[11], Robert C. Jones[12], Kelly Kaihara[13], Chris Brampton[13], Yasha Talaga[13], Yohei Sasagawa[14], Kaori Tanaka[14], Tetsutaro Hayashi[14], Itoshi Nikaido[14,15], Cornelius Fischer[16], Sascha Sauer[16], Timo Trefzer[17], Christian Conrad[17], Xian Adiconis[18,19], Lan T. Nguyen[18], Aviv Regev[18, 20, 21], Joshua Z. Levin[18,19], Swati Parekh[22], Aleksandar Janjic[23], Lucas E. Wange[23], Johannes W. Bagnoli[23], Wolfgang Enard[23], Marta Gut[1], Rickard Sandberg[2], Ivo Gut[1,24], Oliver Stegle[3,4,25], Holger Heyn[1,24,*]

[1] CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
[2] Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden
[3] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
[4] European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany
[5] St Vincent's Institute of Medical Research, Fitzroy, Victoria 3065, Australia.
[6] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology, Barcelona, Spain.
[7] ICREA, Barcelona, Spain.
[8] Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Barcelona, Spain.
[9] Max-Planck-Institute of Immunobiology and Epigenetics, Freiburg, Germany
[10] 10x Genomics, Pleasanton, CA, USA
[11] Fluidigm Corporation, South San Francisco, USA
[12] Department of Bioengineering, Stanford University, Stanford, CA, USA
[13] Bio-Rad, Hercules, CA, USA
[14] Laboratory for Bioinformatics Research RIKEN Center for Biosystems, Dynamics Research, Saitama, Japan
[15] Bioinformatics Course, Master's/Doctoral Program in Life Science, Innovation (T-LSI), School of Integrative and Global Majors (SIGMA), University of Tsukuba, Wako, Saitama, Japan
[16] Max Delbrück Center for Molecular Medicine (BIMSB/BIH), Berlin, Germany
[17] Digital Health Center, Berlin Institute of Health (BIH), Charité-Universitätsmedizin Berlin, Berlin, Germany
[18] Klarman Cell Observatory, Broad Institute of MIT & Harvard, Cambridge, MA, USA
[19] Stanley Center for Psychiatric Research, Broad Institute of MIT & Harvard, Cambridge, MA, USA
[20] Koch Institute of Integrative Cancer Research, Cambridge, MA, USA
[21] Howard Hughes Medical Institute, Dept. of Biology, MIT, Cambridge, MA, USA
[22] Max-Planck-Institute for Biology of Ageing, Cologne, Germany
[23] Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians-University, Martinsried, Germany
[24] Universitat Pompeu Fabra (UPF), Barcelona, Spain
[25] Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany

[+] These authors contributed equally

* Correspondence should be addressed to HH (holger.heyn@cnag.crg.eu)

Corresponding author:
Holger Heyn
Single Cell Genomics Group
Centro Nacional de Análisis Genómico (CNAG-CRG)
Parc Científic de Barcelona – Torre I, Baldiri Reixac, 4
08028 Barcelona, Spain
Phone: +34.934020286
Email: holger.heyn@cnag.crg.eu

Keywords
Single-cell genomics, RNA sequencing, Transcriptomics, Benchmarking, Cell atlas

Abstract

Single-cell RNA sequencing (scRNA-seq) is the leading technique for charting the molecular properties of individual cells. The latest methods are scalable to thousands of cells, enabling in-depth characterization of sample composition without prior knowledge. However, there are important differences between scRNA-seq techniques, and it remains unclear which are the most suitable protocols for drawing cell atlases of tissues, organs and organisms. We have generated benchmark datasets to systematically evaluate techniques in terms of their power to comprehensively describe cell types and states. We performed a multi-center study comparing 13 commonly used single-cell and single-nucleus RNA-seq protocols using a highly heterogeneous reference sample resource. Comparative and integrative analysis at cell type and state level revealed marked differences in protocol performance, highlighting a series of key features for cell atlas projects. These should be considered when defining guidelines and standards for international consortia, such as the Human Cell Atlas project.

Single-cell genomics provides an unprecedented view of the cellular makeup of complex and dynamic systems. Single-cell transcriptomics approaches in particular have led the technological advances that allow unbiased charting of cell phenotypes[1]. The latest improvements in single-cell RNA sequencing (scRNA-seq) allow these technologies to scale to thousands of cells per experiment, providing comprehensive profiling of cellular composition[2,3]. This has led to the identification of novel cell types and the fine-grained description of cell plasticity in dynamic systems, such as development[4,5]. The latest large-scale efforts are attempting to produce cellular maps of entire cell lineages, organs and organism[6,7], with probably the most notable effort being the initiation of the Human Cell Atlas (HCA) project[8]. To comprehensively chart the cellular composition of the human body, the HCA project conducts phenotyping at the single-cell level. It will advance our understanding of tissue function and serve as a reference to pinpoint variation in healthy and disease contexts. In addition to methods that capture the spatial organization of tissues[9,10], the main approach to create a first draft human cell atlas is scRNA-seq-based transcriptome analysis of dissociated cells, in which tissues are disaggregated and individual cells are captured by cell sorting or using microfluidic systems[1]. In sequential processing steps, the RNA is reverse transcribed to cDNA, amplified and processed to sequencing-ready libraries. Continuous technological development has improved the scale, accuracy and sensitivity of the initial scRNA-seq methods, and now allows us to create tailored experimental designs by selecting from a plethora of different scRNA-seq protocols. However, there are marked differences between these methods, and it is still not clear which are the best protocols for drawing a cell atlas.

Experience from other large-scale consortium efforts has shown that neglecting benchmarking, standardization and quality control at the beginning can lead to major problems later on in the project, when investigators are attempting to exploit the results[11]. The overall success of any project depends critically on bringing the work of different consortium partners up to a high common standard. Thus, before launching into large-scale data collection efforts for the HCA and similar projects, it is important to conduct a comprehensive comparison of available single-cell profiling techniques.

In this paper, to extend current efforts to compare the molecule capture efficiency of scRNA-seq methods[12,13], we have systematically evaluated the power of these techniques to describe tissue complexity, and their suitability for building a cell atlas. We performed a multi-center benchmarking study to compare the most common scRNA-seq protocols using a unified reference sample resource. By analyzing human peripheral blood and mouse colon tissue, we have covered a broad range of cell types and states, in order to represent common scenarios in cell atlas projects. We have also added spike-in cell lines to allow us to assess sample composition, and have combined different species to pool samples into a single reference. We performed a comprehensive comparative analysis of 13 different scRNA-seq protocols, representing the most commonly used methods. We applied a wide range of different quality control metrics to evaluate datasets from different perspectives, and to test their suitability for producing a reproducible, integrative and predictive reference cell atlas.

Results

Reference sample and experimental design.

A variety of scRNA-seq methods have been developed, and their utility proven, in single-cell transcriptome analysis of complex and dynamic tissues. The available protocols vary in the efficiency of RNA molecule capture, resulting in differences in sequencing library complexity and sensitivity to identify transcripts and genes[12–14]. However, there has been no systematic testing of how their performance varies between cell types, and how this affects the resolution of cellular phenotyping of complex samples. To address this problem, we benchmarked current scRNA-seq protocols to inform the methodology selection process of cell atlas projects. Ideally, methods should a) be accurate and free of technical biases, b) be applicable across distinct cell properties, c) fully disclose tissue heterogeneity, including subtle differences in cell states, d) produce reproducible expression profiles, e) comprehensively detect population markers, f) be integrable with other methods, and g) have predictive value with cells mapping confidently to a reference atlas.

To perform a systematic comparison of scRNA-seq methods for cell atlas projects, we created a reference sample containing: i) a high degree of cell type heterogeneity with various frequencies, ii) closely related subpopulations with subtle differences in gene expression, iii) a defined cell composition with trackable markers, and iv) cells from different species. For this study, we selected human peripheral blood mononuclear cells (PBMC) and mouse colon, which are tissue types with highly heterogeneous cell populations, as determined by previous single-cell sequencing studies[15,16]. In addition to the well-defined cell types, both tissues contain cells in transition states that present subtle transcriptional differences. These tissues also have a wide range of cell sizes and RNA contents, which are key parameters that affect performance in cell capture and library preparation. Interrogating tissues from different species allowed us to pool samples and exclude cell doublets. In addition to the intra-sample complexity, the spiked-in cell lines enabled the identification of batch effects and biases introduced during cell capture and library preparation. We added cell lines with distinct fluorescent markers that allowed us to track them during sample preparation.

Specifically, the reference sample contained (% viable cells): PBMC (60%, human), colon (30%, mouse), HEK293T (6%, RFP labelled human cell line), NIH3T3 (3%, GFP labelled mouse cells) and MDCK (1%, TurboFP650 labelled dog cells) (**Figure 1**). To reduce variability due to technical effects during library preparation, the reference sample was prepared in a single batch, distributed into aliquots of 250,000 cells, and cryopreserved. We have previously shown that cryopreservation is suitable for single-cell transcriptomics studies of these tissue types[17]. For cell capture and library preparation, the thawed samples underwent FACS separation to remove damaged cells and physical doublets, except for the single-nucleus experiment.

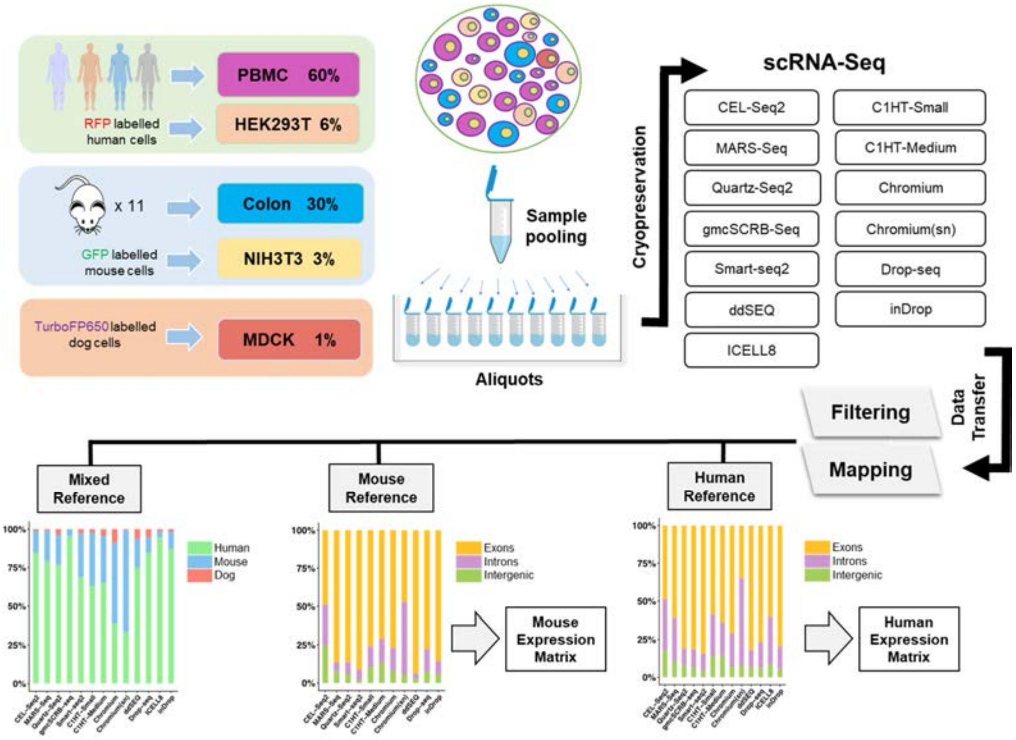


Figure 1. Overview of the experimental design and data processing.
The reference sample consists of human PBMC (60%) and HEK293T (6%), mouse colon (30%) and NIH3T3 (3%) and dog MDCK (1%). The sample was prepared in one single batch, cryopreserved and sequenced by 13 different sc/snRNA-seq methods. Sequences were uniformly mapped to a joint human, mouse and canine reference and then separately to produce gene expression counts for each sequencing method.

A reference dataset for benchmarking experimental and computational protocols.

To obtain sufficient sensitivity to capture low-frequency cell types and subtle differences in cell state, we profiled ~3,000 cells with each scRNA-seq method. In total, we produced datasets for 5 microtiter plate-based methods and 7 microfluidics systems, including cell-capture technologies based on droplets (4), nanowells (1) and integrated fluidic circuits (IFC), to capture small (1) and medium (1) sized cells (**Figure 1** and **Table S1**). We also included experiments to produce single-nucleus RNA sequencing (snRNA-seq) libraries (1), and an experimental variant that profiled >50,000 cells to produce a reference of our complex sample. The unified sample resource and standardized sample preparation (**Online Methods**) were designed to largely eliminate sampling effects, and allow the systematic comparison of scRNA-seq protocol performance.

To compare the different technologies, and to create a resource for the benchmarking and development of computational tools (e.g. batch effect correction, data integration and annotation), all datasets were processed in a uniform manner. Therefore, we designed a streamlined primary data processing pipeline tailored to the peculiarities of the reference sample (**Online Methods**). Briefly, raw sequencing reads were mapped to a joint human, mouse and canine reference genome and separately to their respective references to produce gene count matrices for subsequent analysis (data resource openly available). Consistent with the design of the reference sample, we detected most cells as human (63-95%) or mouse (4-34%; **Figure 1**). Notably, we observed a higher fraction of mouse colon cells in the single-nucleus sequencing dataset (Chromium (sn)). This could result from damaging the more fragile colon cells during sample preparation and resulting in proportionally fewer colon cells when selecting for cell viability. Indeed, when we skipped the viability selection step in the single-cell Chromium experiment as done in the single-nucleus experiment, we observed the same shift in composition towards mouse cells, suggesting that cell viability staining excludes cells that are amenable for scRNA-seq. Consequently, replacing viability staining with a thorough *in silico* quality filtering in cell atlas experiments might better conserve the composition of the original tissue. The canine cells, spiked-in at a low concentration, were detected by all protocols (1-9%) except gmcSCRB-seq. Furthermore, the different methods showed notable differences in mapping statistics between different genomic locations (**Figure 1**). As expected, due to the presence of unprocessed RNA in the nucleus, the snRNA-seq experiment detected the highest proportion of introns, although several scRNA-seq protocols also showed high frequencies of intronic and intergenic mappings.

Molecule capture efficiency and library complexity

We produced reference datasets by analyzing 30,807 human and 19,749 mouse cells (Chromium V2; **Figure 2a-c**). The higher cell number allowed us to annotate the major cell types in our reference sample, and to extract population-specific markers (**Table S2**). Noteworthy, the reference samples solely provided the basis to assign cell identities and gene sets and was not utilized to quantify the methods' performance. This strategy ensured that the choice of technology to derive the reference was not influencing downstream analyses. Indeed, cell clustering and reference-based cell annotation showed high agreement (average 80%; **Online Methods**) and only cells with consistent annotations were used subsequently for comparative analysis at cell type level. Notably, the PBMCs (human) and colon cells (mouse) represented two largely different scenarios. While the differentiated PBMCs clearly separated into subpopulations (e.g. T/B-cells, monocytes, **Figure 2b** and **Supplementary Fig. 1a, 2a-d**), colon cells were ordered as a continuum of cell states that differentiate from intestinal stem cells into the main functional units of the colon (i.e. absorptive enterocytes and secretory cells, **Figure 2c** and **Supplementary Figs. 1b, 3a-d**). After identifying major subpopulations and their respective markers in our reference sample, we clustered the cells of each sc/snRNA-seq protocol and annotated cell types using *matchSCore2* (**Online Methods**). This algorithm allows a gene marker-based projection of single cells (cell-by-cell) onto a reference sample and, thus, the identification of cell types in our datasets (**Supplementary Fig. 4** and **5**).

To compare mRNA capture efficiencies among protocols we downsampled the sequencing reads per cell to a common depth and step-wise reduced fractions (100% to 25%). Library complexity was determined separately for largely homogenous cell types with markedly different cell properties and function, namely human HEK293T cells, monocytes and B-cells (**Figure 2d,e**), and mouse colon secretory and transit-amplifying (TA) cells (**Supplementary Fig. 6a,b**). We observed large differences in the number of detected genes between the protocols, with consistent trends across cell types and gene quantification strategies (**Supplementary Fig. 6c**). Notably, some protocols, such as Smart-seq2 and Chromium V2, performed better with higher RNA quantities (HEK293T) compared to lower starting amounts (monocytes and B-cells), suggesting an input-sensitive optimum. Consistent with the variable library complexity, the protocols presented large differences in drop-out probabilities (**Figure 2f**), with Quartz-seq2, Chromium V2 and CEL-seq2 showing consistently lower probability.

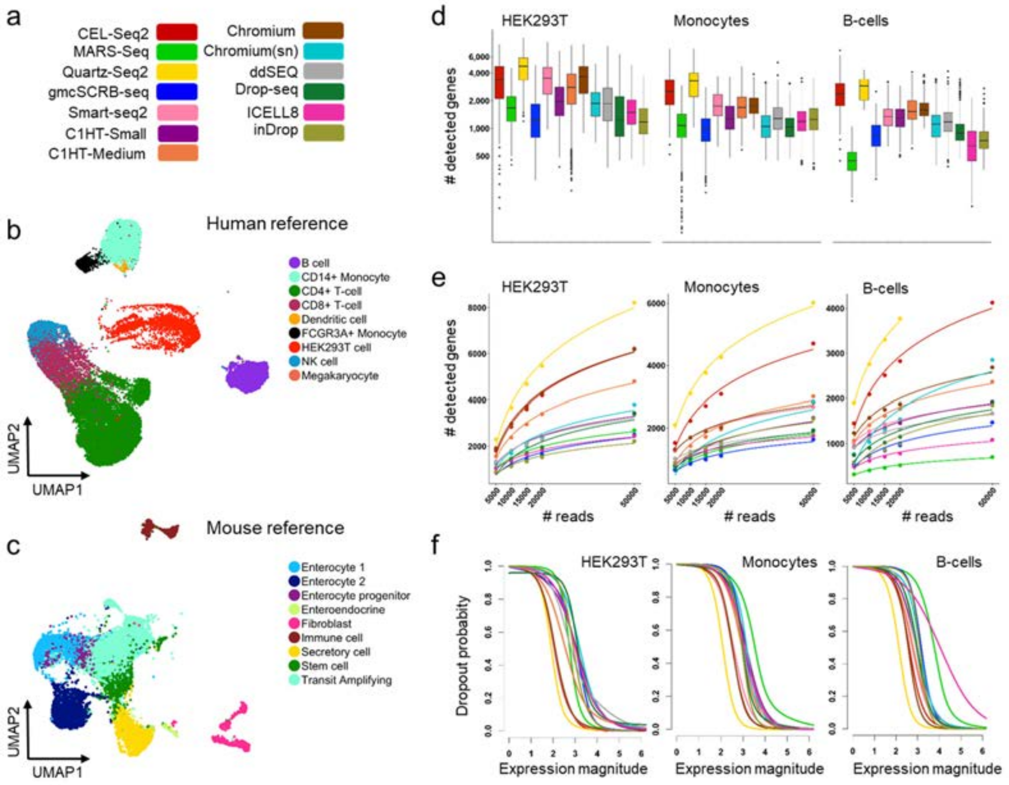


Figure 2. Comparison of 13 sc/snRNA-seq methods.
a. Color legend of sc/snRNA-seq protocols. **b.** UMAP of 30,807 cells from the human reference sample (Chromium) colored by cell type annotation. **c.** UMAP of 19,749 cells from the mouse reference (Chromium) colored by cell type annotation. **d.** Boxplots comparing the number of genes detected across protocols, in downsampled (20K) HEK293T cells, monocytes and B-cells. Cell identities were defined by combining the clustering of each dataset and cell projection onto the reference. **e.** Number of detected genes at step-wise downsampled sequencing depths. Points represent the average number of detected genes as a fraction of all cells of the corresponding cell type at the corresponding sequencing depth. **f.** Dropout probabilities as a function of expression magnitude, for each protocol and cell type, calculated on downsampled data (20K).

Technical effects and information content.

We further assessed the magnitude of technical biases, and the methods' ability to describe cell populations. To quantify the technical variation within and across protocols, we selected highly variable genes (HVG) across all datasets, and plotted the variation in the main principle components (PC; **Figure 3a**). Using the downsampled data for HEK293T cells, monocytes and B-cells, we observed a strong protocol-specific signature, with the main source of variability being the number of genes detected per cell (**Figure 3b**). Nevertheless, PC analysis also showed a mixing of the data points for cells from different methods, suggesting generally conserved information content across the methods. Data from snRNA-seq did not show notable outliers, indicating conserved representation of the transcriptome between the cytoplasm and nucleus. The technical effects were also visible when using t-distributed stochastic neighbor embedding (tSNE) as non-linear dimensionality reduction method (**Supplementary Fig. 7**). By contrast, the methods largely mixed when the analysis was restricted to cell type-specific marker genes, suggesting a conserved cell identity profile across techniques (**Supplementary Fig. 8**).

Next, we quantified the similarities in information content of the protocols. Again, we used the downsampled datasets and calculated the correlation between methods in average transcript counts across multiple cells, thus compensating for the sparsity of single-cell transcriptome data. For the three human cell types, we observed a broad spectrum of correlation between technologies, with generally lower correlation for smaller cell types (**Figure 3c**). Here, the Chromium snRNA-seq protocol displayed a notable outlier, possibly driven by a decreased correlation of immature transcripts (intronic counts; **Figure 1**). Restricting the correlation analysis to population-specific marker genes, we observed less variation between techniques (Pearson's r, 0.5-0.7), which underlines the fact that the expression of these markers is largely conserved between the methods (**Supplementary Fig. 9**).

To further test the suitability of protocols to describe cell types, we determined their sensitivity to detect population specific expression signatures, and found that they had remarkably variable power to detect marker genes (**Figure 3d,e**). Although most of the marker genes were detected by all technologies (>83% of genes), the magnitude of detection varied substantially. Quartz-seq2 and Smart-seq2 showed high expression levels for all cell type signatures, indicating that they have higher power for cell type identification. Since marker genes are particularly important for data interpretation (e.g. annotation), low marker detection levels could severely limit the interpretation of poorly explored tissues, or when trying to identify subtle differences between subpopulations.

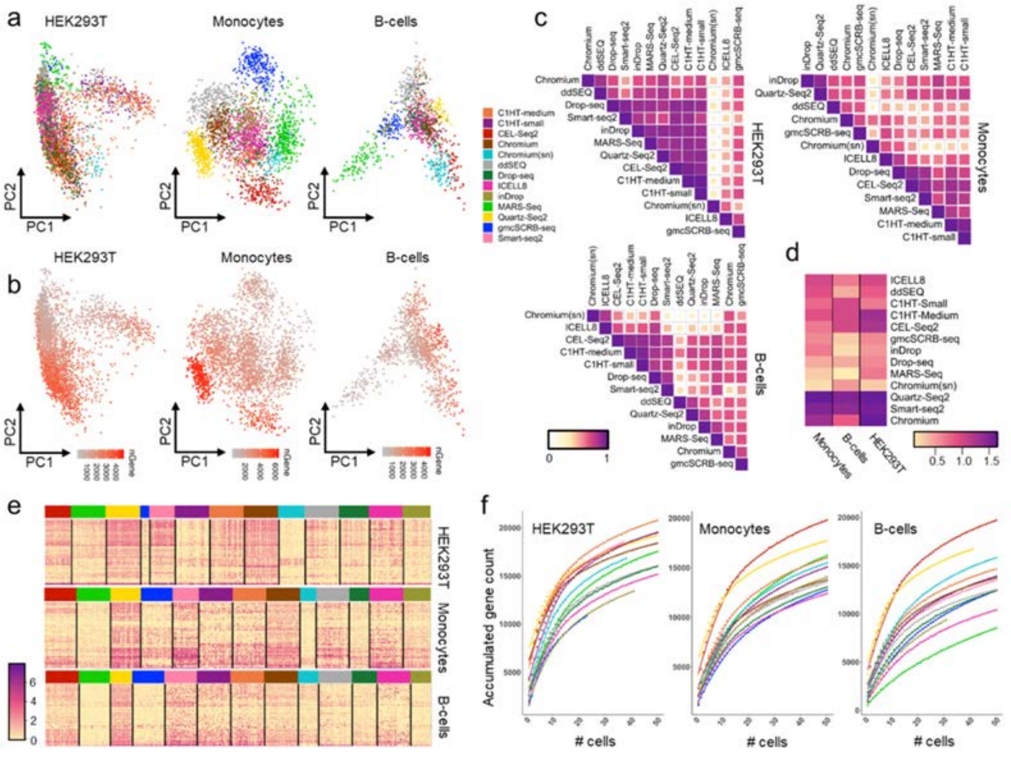


Figure 3. Similarity measures of sc/snRNA sequencing methods.

a,b. PCA analysis on downsampled data (20K) using highly variable genes between protocols, separated into HEK293T cells, monocytes and B-cells, and color-coded by protocol (**a**) and number of detected genes per cell (**b**). **c.** Pearson correlation plots across protocols using expression of common genes. For a fair comparison, cells were downsampled to the same number for each method. Protocols are ordered by agglomerative hierarchical clustering. **d.** Heatmap representing average log expression values of downsampled (20K) HEK293T cell, monocyte, and B-cell reference markers per protocol. **e.** Heatmap representing the log expression values of HEK293T cell, monocyte and B-cell reference markers on downsampled data (20K). **f.** Cumulative gene counts per protocol as the average of 100 randomly sampled HEK293T cells, monocytes and B-cells separately on downsampled data (20K).

The methods also detected vastly different total numbers of genes when accumulating transcript information over multiple cells, with strong positive outliers observed for the smaller cell types (**Figure 3f**). In particular, CEL-seq2 and Quartz-seq2 identified many more genes than other methods. Intriguingly, CEL-seq2 outperformed all other methods by detecting many weakly expressed genes; genes detected specifically by CEL-seq2 had significantly lower expression than the common genes detected by Quartz-seq2 ($p<2.2e-16$). The greater sensitivity to weakly expressed genes makes this protocol particularly suitable for describing cell populations in detail, an important prerequisite for creating a comprehensive cell atlas and functional interpretation.

To further illustrate the power of the different protocols to chart the heterogeneity of complex samples, we clustered and plotted downsampled datasets in two-dimensional space (**Figure 4a**) and then calculated the cluster accuracy and Average Silhouette Width (ASW[18], **Figure 4b**), a commonly used measure for assessing the quality of data partitioning into communities. Consistent with the assumption that library complexity and sensitive marker detection provides greater power to describe complexity, methods that performed well for these two attributes showed better separation of subpopulations, greater ASW and cluster accuracy. This is illustrated in the monocytes, for which accurate clustering protocols separated the major subpopulations (CD14+ and FCGR3A+), while methods with low ASW did not distinguish between them. Similarly, several methods were able to distinguish between CD8+ and NK cells, while others were not.

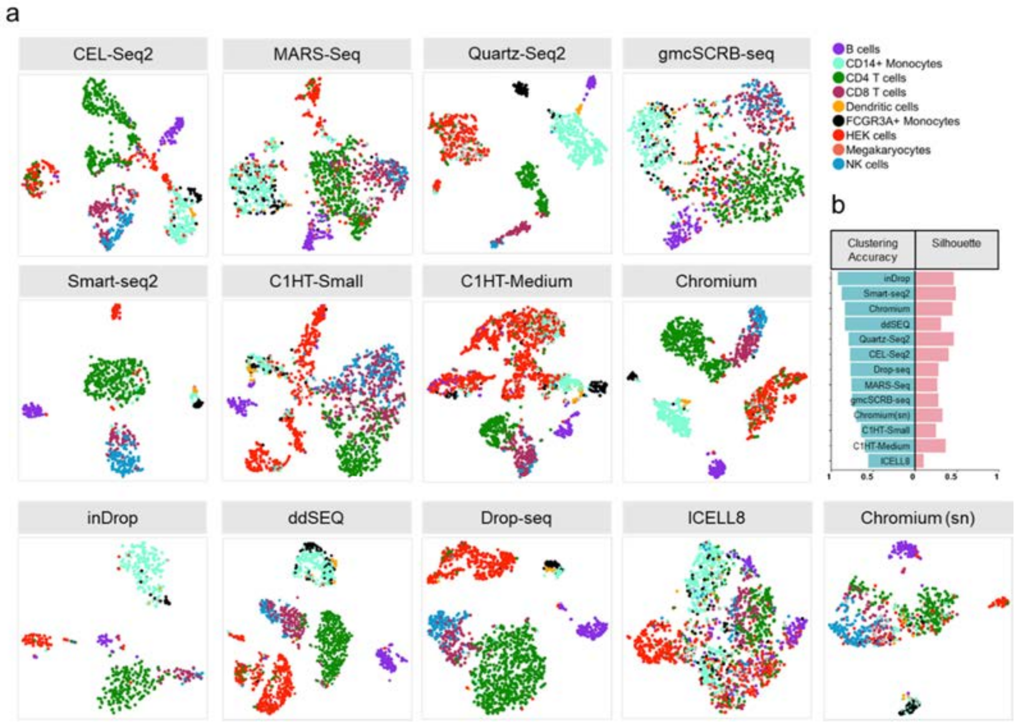


Figure 4. Clustering analysis of 13 sc/snRNA-seq methods on downsampled datasets (20K). a. T-SNE visualizations of unsupervised clustering in human samples from 13 different methods. Each dataset was analyzed separately after downsampling to 20K reads per cell. Cells are colored by cell type inferred by

matchSCore2 before downsampling. Cells that did not achieve a probability score of 0.5 for any cell type were considered unclassified. **b.** Clustering accuracy and Average Silhouette Width for clusters in each protocol.

Joint analysis across datasets

A common scenario for cell atlas projects is that data are produced at different sites using different scRNA-seq protocols. However, the final atlas is created from a combination of datasets, which requires that the technologies used are compatible. To assess how suitable it is to combine the results from our protocols into a joint analysis, we used downsampled human and mouse datasets to produce a joint quantification matrix for all techniques[19]. Importantly, single cells grouped themselves by cell type, suggesting that cell phenotypes are the main driver of heterogeneity in the joint datasets (**Figure 5a-d** and **Supplementary Fig. 10a,b**). Indeed, the combined data showed a clear separation of cell states (e.g. T-cell and enterocyte subpopulations) and rarer cell types, such as dendritic cells. However, within these populations there were some differences between the methods, indicating the presence of technical effects that could not be entirely removed during the merging step (**Figure 5e-f** and **Supplementary Fig.10c,d**). To formally assess the capacity of the methods to be joined, we calculated the degree to which technologies mix in the merged datasets (**Figure 5g,h**). Intriguingly, the methods' suitability to be combined was not directly correlated with their power to discriminate between cell types. Thus, while a well-performing method might result in a high-resolution cell atlas maps, it could perform poorly in a consortium-driven project that includes different data sources. Moreover, when integrating further downsampled datasets, we observed a drop in mixing ability, although the cell type separation was largely conserved (**Supplementary Fig. 10e**). Consequently, quality standard guidelines for consortia might define minimum coverage thresholds to ensure the subsequent option of data merging.

Cell atlas datasets will serve as a reference for annotating cell types and states in future experiments. Therefore, we assessed cells' ability to be projected onto our reference sample (**Figure 2b,c**). We used the population signature model defined by *matchSCore2* and evaluated the protocols based on their cell-by-cell mapping probability, which reflects the confidence of cell annotation (**Supplementary Fig. 11a-c**). Although there were some differences in the protocols' projection probabilities and a potential bias due to the selection of the reference protocol, a confident annotation was observed for most cells with inDrop and ddSEQ reporting the highest probabilities. Notably, high probability scores were also observed in further downsampled datasets (**Supplementary Fig. 11b**). This has practical consequences, as data derived from less well performing methods (from a cell atlas perspective) or from poorly sequenced experiments could be identifiable and thus suitable for specific analysis types, such as tissue composition profiling.

Conclusion

Systematic benchmarking of available technologies is a crucial prerequisite for large-scale projects. Here, we evaluated scRNA-seq protocols for their power to produce a cellular map of complex tissues. Our reference sample simulated common scenarios in cell atlas projects, including differentiated cell types and dynamic cell states. We defined the strengths and weaknesses of key features that are relevant for cell atlas studies, such as comprehensiveness, integratability, and predictive value. The methods revealed a broad spectrum of performance, which should be considered when defining guidelines and standards for international consortia (**Figure 6**). In addition, cell atlas projects need to consider other protocol-specific features, such as cost-effectiveness and scalability, in their decision making process towards large-scale datasets. It is equally important to benchmark computational pipelines for data analysis and interpretation[20–22]. We envision that the datasets provided by our study will serve as a valuable resource for the single-cell community to develop and evaluate novel strategies towards an informative and interpretable cell atlas.

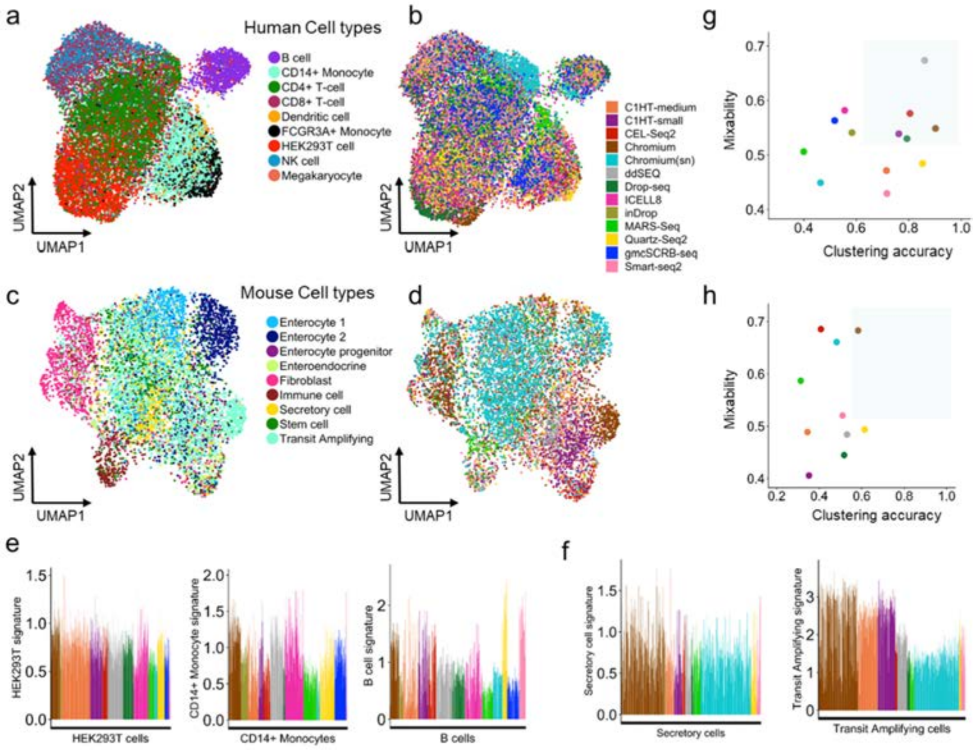


Figure 5. Integration of sc/snRNA-seq methods. UMAP visualization of cells after integrating technologies for human (**a,b**) and mouse (**c,d**) datasets. Cells are colored by cell type (**a,c**) and sc/snRNA-seq protocol (**b,d**). **e,f.** Barplots showing normalized and method-corrected (integrated) expression scores of cell-type-specific signatures for human HEK293T cells, monocytes, B-cells (**e**) and mouse secretory and transit-amplifying cells (**f**). Bars represent cells and colors methods. **g,h.** Evaluation of method integratability in human (**g**) and mouse (**h**). Protocols are compared according to their ability to group cell types into clusters (after integration) and mix with other technologies within the same clusters. Points are colored by sequencing method. The gray area shows the optimal trade-off between the two properties.

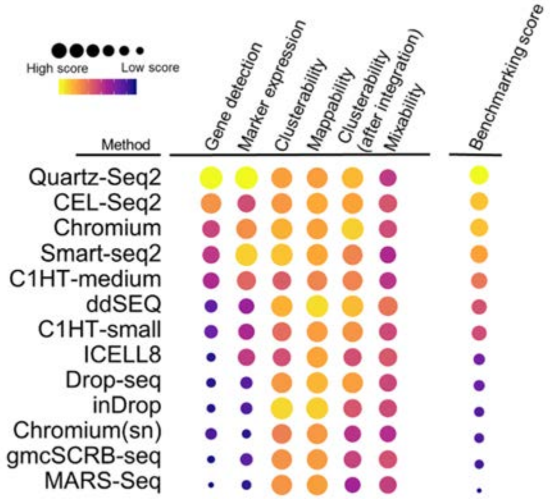


Figure 6. Benchmarking summary of 13 sc/snRNA-seq methods. Methods are scored by key analytical metrics, characterizing protocols according to their ability to recapitulate the original structure of complex tissues, and their suitability for cell atlas projects. The methods are ordered by their overall benchmarking score, which is computed by averaging the scores across metrics.

Addendum

This study complements a study entitled "Systematic comparative analysis of single cell RNA-sequencing method" by Ding et al., which applied a complementary design (BIORXIV/2019/632216).

Acknowledgement

This project has been made possible in part by grant number 2018-182827 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. HH is a Miguel Servet (CP14/00229) researcher funded by the Spanish Institute of Health Carlos III (ISCIII). CM is supported by an AECC postdoctoral fellowship. This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. H2020-MSCA-ITN-2015-675752 (Singek), and the Ministerio de Ciencia, Innovación y Universidades (SAF2017-89109-P; AEI/FEDER, UE). S. is supported by the DFG (GR4980) the Behrens-Weise-Foundation. D.G. and S. are supported by the Max Planck Society. CZ is supported by EMBO through long-term fellowship ALTF 673-2017. The single-nucleus RNA-Seq data were generated with support from the National Institute of Allergy and Infectious Diseases grant U24AI118672, the Manton Foundation, and the Klarman Cell Observatory (AR). We thank ThePaperMill for critical reading and scientific editing services and the Eukaryotic Single Cell Genomics Facility at Scilifelab (Stockholm, Sweden) for support. Core funding is from the ISCIII and the Generalitat de Catalunya.

Ethical Statement

This study was approved by the Parc de Salut MAR Research Ethics Committee (reference number: 2017/7585/I) to Dr. Holger Heyn. We adhered to ethical and legal protection guidelines for human subjects, including informed consent.

Author's Contributions

HH designed the study. EM and AL performed all data analyses. CM, AA and EB prepared the reference sample. CZ, DJM, SP and OS supported the data analysis. MG and IG provided technical and sequencing support. All other authors provided sequencing-ready single-cell libraries or sequencing raw data. HH, EM and AL wrote the manuscript with contributions from the co-authors. All authors read and approved the final manuscript.

Conflicts of Interest

AR is a co-founder and equity holder of Celsius Therapeutics, and an SAB member of ThermoFisher Scientific and Syros Pharmaceuticals. AR is a co-inventor on patent applications to numerous advances in single-cell genomics, including droplet based sequencing technologies, as in PCT/US2015/0949178, and methods for expression and analysis, as in PCT/US2016/059233 and PCT/US2016/059239. KK, CB and YT are employed by Bio-Rad Laboratories. JKL and SCB are employees and shareholders at 10x Genomics. All other authors declare no conflicts of interest associated with this manuscript. CS and AO are employed by Fluidigm.

Data Availability

All raw sequencing data will be freely available through the Human Cell Atlas Data Coordination Portal (DCP). The code for *matchSCore2* is available under https://github.com/elimereu/matchSCore2.

References

1. Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc.* **13**, 2742–2757 (2018).
2. Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat. Rev. Genet.* **18**, 345–361 (2017).
3. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
4. Karaiskos, N. *et al.* The Drosophila embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
5. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
6. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
7. Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* eaaq1723 (2018). doi:10.1126/science.aaq1723
8. Regev, A. *et al.* Science Forum: The Human Cell Atlas. *eLife* **6**, e27041 (2017).
9. Moffitt, J. R. *et al.* High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11046–11051 (2016).
10. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell *in situ* RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
11. Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* **6**, 10001 (2015).
12. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* **65**, 631-643.e4 (2017).
13. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
14. Tung, P.-Y. *et al.* Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).
15. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
16. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
17. Guillaumet-Adkins, A. *et al.* Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biol.* **18**, 45 (2017).
18. Azuaje, F. A cluster validity framework for genome expression data. *Bioinforma. Oxf. Engl.* **18**, 319–320 (2002).
19. Lin, Y. *et al.* scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc. Natl. Acad. Sci.* 201820006 (2019). doi:10.1073/pnas.1820006116
20. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
21. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
22. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv* 276907 (2018). doi:10.1101/276907
23. Sasagawa, Y. *et al.* Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* **19**, (2018).
24. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187–1201 (2015).
25. Goldstein, L. D. *et al.* Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics* **18**, 519 (2017).
26. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
27. Isolation of Nuclei for Single Cell RNA Sequencing - Demonstrated Protocol - Sample Prep - Single Cell Gene Expression - Official 10x Genomics Support. Available at: https://support.10xgenomics.com/single-cell-gene-expression/sample-prep/doc/demonstrated-protocol-isolation-of-nuclei-for-single-cell-rna-sequencing. (Accessed: 25th March 2019)
28. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).

29. Herman, J. S., Sagar, null & Grün, D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* **15**, 379–386 (2018).
30. Sagar, null, Herman, J. S., Pospisilik, J. A. & Grün, D. High-Throughput Single-Cell RNA Sequencing and Data Analysis. *Methods Mol. Biol. Clifton NJ* **1766**, 257–283 (2018).
31. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
32. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
33. Barriga, F. M. *et al.* Mex3a Marks a Slowly Dividing Subpopulation of Lgr5+ Intestinal Stem Cells. *Cell Stem Cell* **20**, 801-816.e7 (2017).
34. Bagnoli, J. W. *et al.* Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 2937 (2018).
35. Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinforma. Oxf. Engl.* **28**, 2520–2522 (2012).
36. SingleCellExperiment: S4 Classes for Single Cell Data version 1.4.1 from Bioconductor. Available at: https://rdrr.io/bioc/SingleCellExperiment/. (Accessed: 26th March 2019)
37. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *GigaScience* **7**, (2018).
38. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).
39. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma. Oxf. Engl.* **30**, 923–930 (2014).
40. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
41. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).
42. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).

Online Methods

Reference sample

Cell Lines

NIH3T3-GFP, MDCK-TurboFP650 and HEK293-RFP were cultured at 37ºC in an atmosphere of 5% (v/v) carbon dioxide in Dulbecco's Modified Eagle's Medium, supplemented with 10% (w/v) fetal bovine serum, 100 U penicillin, and 100 µg/L streptomycin (Invitrogen). On the reference sample preparation day, the culture medium was removed and the cells washed with 1X PBS. Afterwards, cells were trypsinized (trypsin 100X), pelleted at 800 x g for 5min, washed in 1X PBS, re-suspended in PBS-EDTA (2mM) and stored on ice.

Mouse Colon Tissue

The colon from eleven mice (7x*LGR5/GFP* and 4WT) was dissected and removed. For single-cell separation the colons were treated separately. The colon was sliced, opened and washed twice in cold 1X HBSS. It was then placed on a petri plate on ice and minced with razor blades until disintegration. The minced tissue was transferred to a 15 ml tube containing 5 ml of 1X HBSS and 83 µl of collagenase IV (final concentration 166 U/ml). The solution was incubated for 15 min at 37ºC (vortexed for 10 sec every 5 min). To inactivate the collagenase IV, 1 ml of FBS was added and vortexed for 10 seconds. The solution was filtered through a 70 µm nylon mesh (changed when clogged). Finally, all samples were combined, cells pelleted for 5 min at 400 g at 4ºC. The supernatant was removed, and the cells resuspended in 20 ml of 1X HBSS and stored on ice.

Isolation of peripheral blood mononuclear cells (PBMC)

Whole blood was obtained from four donors (two female, two male). The extracted blood was collected in Heparin tubes (GP supplies) and processed immediately. For each donor, PBMCs were isolated according to the manufacturer's instructions for FICOLL extraction (pluriSelect). Briefly, blood from two Heparin tubes (approximately 8 ml) was combined, diluted in 1X PBS and carefully added to a 50 ml tube containing 15 ml FICOLL. The tubes were centrifuged for 30 min at 500 g (minimum acceleration and deceleration). The interphase was carefully collected and diluted with 1X PBS + EDTA (2mM). Following a second centrifugation, the supernatant was discarded and the pellet resuspended in 2 ml of 1X PBS + EDTA (2mM) and stored on ice.

Preparation of the reference sample

Cell counting was performed using an automated cell counter (TC20™ Automated Cell Counter, Bio-Rad Laboratories). The reference sample was calculated to include human PBMC (60%), mouse colon (30%), HEK293T (6%, RFP labelled human cell line), NIH3T3 (3%, GFP labelled mouse cells) and MDCK (1%, TurboFP650 labelled dog cells). To adjust for cell integrity loss during sample processing, we measured the viability during cell counting, and accounted for an expected viability loss after cryopreservation (10% for cell lines and PBMC; 50% for colon[17]). All single cell solutions were combined in the proportions mentioned above and diluted to 250,000 viable cells per 0.5 ml. For cryopreservation, 0.5 ml of cell suspension was aliquoted into cryotubes and gently mixed with a freezing solution (final concentration 10% DMSO; 10% heat-inactivated FBS). Cells were then frozen by gradually decreasing the temperature (1ºC/min) to –80ºC (cryopreserved), and stored in liquid nitrogen. MARS-Seq and Smart-Seq2 experiments were performed to validate sample quality and composition before distributing aliquots to the partners.

Sample processing instructions

This cryopreserved reference sample forms the basis for systematic comparison of scRNA-seq techniques. The sample consists of two complex tissues (human PBMC and mouse colon) and three cell lines (HEK293-RFP, NIH3T3-GFP, MDCK-Turbo650). The primary PBMC and the colon cells account for around 90% of the living (DAPI negative) sample content, and the cell lines the remaining 10% (6% HEK293-RFP, 3% NIH3T3-GFP, <1% MDCK-Turbo650). Each cryo-vial contains ~250,000 living cells, sufficient to sort a minimum of 4 x 384-well plates or to isolate >3000 cells (microfluidic systems), and should be stored at -80ºC upon arrival.

The sample preparation aims to be standardized for all methods to allow comparison of the performance of library preparation. FACS isolation should be performed before sample processing to exclude damaged/dead cells (DAPI positive). Moreover, we aim to simulate the exclusion of unwanted cell types by excluding NIH3T3 (GFP positive) cells during FACS isolation. For the remaining sample, FACS gates should be set to exclude debris, cell fragments and doublets (Appendix: screen shots provided). Proportions of intact (DAPI negative) and fluorescence labeled cells (RFP, GFP and TurboFP650) should be recorded, and, if possible, cells should be index-sorted (for microtiter plates).

NOTE: The cryopreserved samples consists of approximately 30-40% intact (DAPI negative) cells. We recommend FACS isolation of DAPI negative cells before single-cell capture.
NOTE: We provide cryo-vials of PBMC and fluorescence-labeled cell lines to facilitate gate-setting for debris exclusion, and to define the degree of compensation. Please set gates to include blood and larger cells as indicated in the Appendixes.
NOTE: One HCA reference vial is sufficient to fill 4x 384-well plates.
NOTE: FACS isolation into plates should be performed at low speed (below 100 cells/sec) to avoid loss of the sample.
NOTE: To simulate the exclusion of cell types, GFP-labeled NIH3T3 cells should be excluded from the final single-cell selection.

Sample thawing instructions

- Remove sample from -80ºC and process immediately
- De-freeze in water bath (37ºC) with continuous agitation until material is almost thawed
- Transfer to 15 ml Falcon using a 1000 ul tip (wide-bored or cut tip) without mixing by pipetting
- Add drop-wise 1000 ul of pre-warmed (37ºC) Hibernate-A while gently swirling the sample
- Let sample rest for 1 min
- Add drop-wise 2000 ul of pre-warmed (37ºC) Hibernate-A while gently swirling the sample
- Let sample rest for 1 min
- Add drop-wise 2000 ul of pre-warmed (37ºC) Hibernate-A while gently swirling the sample
- Let sample rest for 1 min
- Add 3000 ul pre-warmed (37ºC) Hibernate-A
- Invert Falcon 3 times
- Let sample rest for 1 min
- Add 5000 ul pre-warmed (37ºC) Hibernate-A
- Invert Falcon 6 times
- Let sample rest for 1 min
- Centrifuge sample at 400 g for 5 min at 4ºC (pellet clearly visible)
- Remove supernatant until 500 ul supernatant remains in tube
- Resuspend the pellet by gentle pipetting
- Add 3500 ul of 1X PBS + 2mM EDTA
- Store sample on ice until processing
- Filter cells through a nylon mesh into FACS tubes (2 tubes with 2 ml sample)
- Add 3 ul DAPI
- Mix gently
- Store on ice
- **Exclude DAPI and GFP positive cells during sorting**
- Use index sorting for RFP and TurboFP650 (optional)

Single-cell RNA sequencing library preparation

Quartz-Seq2[23]

We isolated single-cells into 384-well PCR plates from cell suspension using a MoFlo Astrios EQ (Beckman Coulter) cell sorter. The cell sorter was equipped with a 100-µm nozzle and a custom-made splash-guard. In total, we analyzed 3,072 wells corresponding to eight 384-well PCR plates. Sequence library preparation of Quartz-Seq2 was performed as described previously[23] with the following modifications. For lysis buffer, we used 768 kinds of RT primers corresponding to v3.2A and v3.2B. We prepared two sets of the 384-well PCR plate with lysis buffer containing no ERCC spike-in RNA. We added 1 µl of RT premix (2X Thermopol buffer, 1.25 units/µL SuperScript III, 0.1375 units/µL RNasin plus) to 1 µl of lysis buffer for each well. After cell barcoding, we collected cDNA solution into one well reservoir from two sets of 384-well plates, which corresponded to 768 wells. For cDNA purification and concentration, we used four purification columns for cDNA solution from two 384-well PCR plates. In the PCR step, we amplified cDNA for ten cycles. In an additional purification step for amplified cDNA, we added 26 µl (0.65X) of resuspended AMPure XP Beads to the cDNA solution. We obtained amplified cDNA of 32.6 ± 6.8 ng (n = 4) from the 768 wells. We sequenced the Quartz-Seq2 sequence library with a NextSeq 500/550 High Output v2 Kit. The BCL files obtained were converted to FASTQ files using bcl2fastq2 (v2.17.1.14) with demultiplexing pool barcodes. Each FASTQ file was split into single FASTQ files for each cell barcode using a custom script (https://github.com/rikenbit/demultiplexer_quartz-seq2, DOI: 10.5281/zenodo.2585429).

inDrop System (1CellBio)[24]

Cells were isolated using an Aria3Fusion (BD Bioscience) cell sorter with a 100µm nozzle and a flow rate of 6-7. The workflow was carried out using the inDrop instrument and the inDrop single cell RNA-seq kit (Cat. No. 20196, 1CellBio) according to the manufacturer's protocols. Microfluidic chips were prepared by silanization, and barcode labeled hydrogel microspheres (BHMs) were prepared shortly before cell capture, according to protocol (version v2.0., 1CellBio website). Droplet-making oil, single-cell suspension (200 cells/µL), and freshly prepared RT/lysis buffer were loaded onto the chip for droplet generation, according to the inDrop protocol for single-cell encapsulation and reverse transcription (version 2.1., 1CellBio website). An emulsion corresponding to ~4000 droplets was collected in a cooled tube and irradiated with UV light to release the photo-cleavable barcoding oligos from the BHMs. cDNA synthesis proceeded within the droplets, and the emulsion was subsequently split into equal volumes in such a way as to not exceed ~2000 droplets per reaction tube. After de-emulsification, cDNA contained in the aqueous phase was stored at -80°C. The RT product was further processed according to the InDrop library preparation protocol (version 1.2. 1CellBio website). The cDNA was fragmented by ExoI/HinfI and purified by AMPure XP beads. Second strand synthesis was conducted using NEB second-strand synthesis module (Cat. no. E6111S, NEB). In vitro-transcription was conducted using HiScribe T7 High Yield RNA Synthesis kit (cat. no. E2040S, NEB). Amplified RNA was then fragmented, and the fragments used in a second reverse transcription reaction with random hexamers to convert the sample back into DNA and to add a read primer-binding site to each molecule. Hybrid molecules of RNA and DNA were cleaned up using AMPure beads and amplified by PCR. Final libraries were sequenced using HiSeq4000 and NextSeq (Illumina).

ICELL8 SMARTer Single-Cell System (Takara Bio)[25]

Hoechst 33342 and propidium iodide co-stained single-cell suspension (20 cells/µL) was distributed in eight wells of a 384-well source plate (Cat. No. 640018, Takara) and dispensed into a barcoded SMARTer ICELL8 3' DE Chip (Cat. No. 640143, Takara) using an ICELL8 MultiSample NanoDispenser (MSND, Takara). 4 chips were used to target ~3000 single cells. Nanowells were imaged using the ICELL8 Imaging Station (Takara). After imaging, the chip was sealed, placed in a pre-cooled freezing chamber, and stored at −80 °C. CellSelect software was used to identify each nanowell that contained a single cell. These nanowells were then selected for subsequent targeted deposition of a 50 nL/nanowell RT-PCR reaction solution from the

SMARTer ICELL8 3' DE Reagent Kit (Cat. No. 640167, Takara) using the MSND. After RT and amplification in a Chip Cycler, barcoded cDNA products from nanowells were pooled together using the SMARTer ICELL8 Collection Kit (Cat. No. 640048, Takara). cDNA was concentrated using the Zymo DNA Clean & Concentrator kit (Cat. No. D4013, Zymo Research), and purified using AMPure XP beads. cDNA was then used to construct Nextera XT (Illumina) DNA libraries, followed by 0.6X AMPure XP bead purification. Library quantification and size distribution was done using Qubit, KAPA Library Quantification and Agilent TapeStation. Final libraries were sequenced using HiSeq4000 and NextSeq500 (Illumina).

Drop-Seq (Dolomite)[26]

Single-cell RNA experiments were performed using the scRNA system with P-Pumps and a scRNA-chip (100µm channel width) from Dolomite Bio (Royston, UK). Encapsulation was conducted according to the manufacturer's instructions, and library construction was completed according to the published DropSeq protocol[26]. Briefly, polyT-barcoded beads (MACOSKO-2011-10; ChemGenes) were loaded at a concentration of 600/µl, and cells at a concentration of 450/µl. The pumps were operated at a flowrate of 30 µl/min for beads and cell suspension (PBS+2mM EDTA), and at 200 µl/min for oil (QX200™ Droplet Generation Oil for EvaGreen; BioRad). After encapsulation, cell lysis, and hybridization of RNA to the beads, droplets were broken using PFO (Sigma-Aldrich) and aliquots of a maximum of 90000 beads were collected. Reverse transcription was performed in a 200µl volume with Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific) and 2.5 µM TSO-primer (AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG; Qiagen) at room temperature for 30 min, followed by 90 min at 42°C. After exonuclease treatment (ExoI; New England Biolabs) at 47°C in 200 µl, to digest the unbound primer, cDNA was amplified by PCR using HiFi HotStart mix (Kapa Biosystems) and amplification primer (AAGCAGTGGTATCAACGCAGAGT; Qiagen) in batches of 4000 beads in a volume of 50 µl (95°C - 3min; 4 cycles: 98°C - 20s, 65°C - 45s, 72°C - 3min; 9 cycles: 98°C - 20s, 67°C - 20s, 72°C - 3min; 72°C - 5min). Libraries were generated using the Nextera XT library Kit (Illumina) in five pooled PCR samples with 600 pg of cDNA and a custom P5-primer (AATGATACGGCGACCACCGAGATCTACACGCCTGTCCGCGGAAGCAGTTGGTATC AACGCAGAGT*A*C; Qiagen). Final library QC was conducted using the BioAnalyzer High Sensitivity DNA Chip (Agilent Technologies). For sequencing on an Illumina HiSeq2500 V4, we used a custom read 1 primer (GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC; Qiagen).

Chromium V2 (10X Genomics): Single-cell RNA sequencing[15]

Two cell preparations were conducted on two different days: one to prepare 2 libraries for sequencing at high read depth, and one to prepare 8 libraries at low read depth. To prepare the libraries for high read depth, one frozen vial of a Human Cell Atlas reference sample was thawed and prepared as described. At the end of this protocol, the cells were resuspended in PBS with 2 mM EDTA. Since cells showed clumping and low viability, they were centrifuged 3 times at 150 g for 10 min at room temperature, and resuspended in 50% PBS, 2mM EDTA and 50% Iscove's Modified Dulbecco Medium (IMDM, ATCC) supplemented with 10% FBS and filtered through a 40µm FlowMi cell strainer (Sigma-Aldrich) to remove cell aggregates and large cell debris. At the final count before loading, the cell suspension demonstrated a viability of 60%. To prepare the libraries for low read depth, two frozen vials of a the reference sample were thawed and prepared as described in an updated version of the HCA Benchmark protocol. At the end of this protocol, the cells were resuspended in IMDM, 10% FBS and 1mM EDTA, and filtered through a 40-µm FlowMi cell strainer to remove cell aggregates and large cell debris. At the final count before loading, the cell suspension demonstrated a viability of 65%. The cells were not processed using FACS isolation, but run directly on the 10x Chromium system (10x Genomics, Pleasanton, CA, USA).

Cells were mixed with single-cell master mix, and the resulting cell suspensions were loaded on a 10x Chromium system to generate 2 libraries at 5,000 cells each and 5 libraries at 10,000 cells each. The single-cell libraries were generated using 10x Chromium Single Cell gene expression

V2 reagent kits according to the manufacturer's instructions (Chromium single cell 3' reagents kits v2 user guide). Single cell 3' RNA-seq libraries were quantified using an Agilent Bioanalyzer with a high sensitivity chip (Agilent), and a Kapa DNA quantification kit for Illumina platforms (Kapa Biosystems). The libraries were pooled according to the target cell number loaded. Sequencing libraries were loaded at 200 pM on an Illumina NovaSeq6000 with Novaseq S2 Reagent Kit (100 cycles) using the following read lengths: 26 bp Read1, 8 bp I7 Index and 91 bp Read2. The 2 libraries of 5,000 cells and the 8 libraries of 10,000 cells were sequenced at 250,000 and 25,000 reads per cell, respectively.

Chromium V2 (10X Genomics): Single-nucleus RNA sequencing

We isolated nuclei from the cell suspension using a protocol provided by 10x Genomics[27]. We counted the nuclei using a Countess II (Thermo Fisher Scientific). We made an aliquot containing ~11,000 nuclei in a volume of 33.8 µL in RB buffer (1x PBS, 1% BSA, and 0.2U/µl RNaseIn (TaKaRa)) as sample A, and stained the rest of the nuclei suspension with Vybrant DyeCycle Violet Stain (Thermo Fisher Scientific) at a concentration of 10 µM. We used a MoFlo Astrios EQ cell sorter (Beckman Coulter) and set fluorescence activated cell sorting (FACS) gating on forward scatter plot, side scatter plot and on fluorescent channels to pick Violet-positive (for nuclei), while excluding debris and doublets. We used a 100 µm nozzle to sort 20,000 nuclei into 20 µl RB buffer as sample B. After sorting, we measured the volume of B with a pipette, spun it at 500 g for 5 min at 4ºC, and then carefully removed part of the supernatant to leave ~40µl. We resuspended B by gentle pipetting 40 times.

Immediately after nuclei isolation, we loaded sample A into one channel of a Chromium Single Cell 3' Chip (10x Genomics, PN-120236), and then processed it through the Chromium Controller to generate GEMs (Gel Beads in Emulsion). We then loaded 33.8 µL of B 25 minutes later after sorting and centrifugation, as described above, into one channel of a second chip, and processed it in the same way as the first chip. We prepared RNA-Seq libraries for both samples in parallel with the Chromium Single Cell 3' Library & Gel Bead Kit V2 (10x Genomics, PN-120237), according to the manufacturer's protocol. We pooled the 2 samples based on molar concentrations and sequenced them on a NextSeq500 instrument (Illumina).

Smart-seq2[28]

Smart-seq2 libraries were prepared at half the volume, as described previously[28], with minor modifications. In brief, 2 µl of lysis buffer containing 0.1 % Triton X-100 (Sigma-Aldrich), 1 U/µl RNase inhibitor (Takara), 2.5 mM dNTPs (Thermo Fisher) and 2 µM oligo-dT primer (5′–AAGCAGTGGTATCAACGCAGAGTACT30VN-3′; IDT) were dispensed into each well of a 384-well plate (4titude). Lysis plates were stored at -20°C until cell sorting, after which single-cell lysates were kept at -80 °C. Before reverse transcription, cell lysates were denatured at 72 °C for 3 min and immediately placed on ice. The RT reaction was performed in a 5 µl total volume, with final reagent concentrations of 1x Superscript first-strand buffer (Thermo Fisher), 5 mM DTT (Thermo Fisher), 1 M Betaine (Sigma-Aldrich), 9 mM MgCl2 (Sigma-Aldrich), 1 U/µl RNase inhibitor (Takara), 1 µM LNA template-switching oligo (5′-AAGCAGTGGTATCAACGCAGAGTACATrGrG+G-3′; Exiqon), and 10 U/µl Superscript II RT enzyme. Next, pre-amplification PCR was performed for 22 cycles at final concentrations of 1x KAPA HiFi HotStart ReadyMix (Roche) and 0.08 µM ISPCR primer (5′-AAGCAGTGGTATCAACGCAGAGT-3′; IDT) in a total reaction volume of 11 µl. The cDNA was cleaned up by adding 10 µl of SPRI beads (19.5 % PEG, 1 M NaCl, 1 mM EDTA, 0.01 % IGEPAL CA-630), washing twice with 20 µl 80 % ethanol, and eluting in 10 µl H_2O. The cDNA concentration was measured for all wells using Picogreen dsDNA assay (Thermo Fisher), and diluted to 200 pg/µl using a Mantis liquid handler (Formulatrix). Next, 1 µl of cDNA was used as input for the Nextera XT library preparation kit (Illumina) at 1/5 volume, according to the manufacturer's instructions. During the 12 cycles library PCR, custom i7 and i5 indexing primers (IDT) were added at 0.5 µM each. Finally, 5 µl of library per well were pooled, cleaned and concentrated using SPRI beads (19.5 % PEG; see above). Final libraries were sequenced using HiSeq2500 V4 (Illumina).

CEL-Seq2[29,30]

Single-cell RNA sequencing was performed using a modified version of the mCEL-Seq2 protocol, an automated and miniaturized version of CEL-Seq2, on a Mosquito nanoliter-scale liquid-handling robot (TTP LabTech)[29,31]. Briefly, cells were sorted into 384-well plates (Bio-Rad) containing 240 nl of lysis buffer containing polyT primers and 1.2 µl of mineral oil (Sigma-Aldrich). Sorted plates were centrifuged at 2200 g for several minutes at 4°C, snap-frozen in liquid nitrogen and stored at −80°C until processing. 160nL of reverse transcription reaction mix and 2.2 µl of second strand reaction mix were used to convert RNA into cDNA. cDNA from 96-cells were pooled together before clean up and *in vitro* transcription, generating 4 libraries from one 384-well plate. During all purification steps, including the library cleanup, we used 0.8 µl of AMPure/RNAClean XP beads (Beckman Coulter) per 1 µl of sample. Sixteen libraries with 96 cells each (one of the libraries contained 30,000 RNA molecules from ERCC spike-in mix per cell) were sequenced on an Illumina HiSeq3000 sequencing system (pair-end multiplexing run).

MARS-Seq[32]

To construct single-cell libraries from poly(A)-tailed RNA, we used massively parallel single-cell RNA sequencing (MARS-Seq)[32]. Briefly, single cells were FACS-isolated into 384-well plates containing lysis buffer (0.2% Triton X-100 (Sigma-Aldrich); RNase inhibitor (Invitrogen)) and reverse-transcription (RT) primers. Single-cell lysates were denatured and immediately placed on ice. The RT reaction mix, containing SuperScript III reverse transcriptase (Invitrogen), was added to each sample. After RT, the cDNA was pooled using an automated pipeline (epMotion, Eppendorf). Unbound primers were eliminated by incubating the cDNA with exonuclease I (NEB). A second stage of pooling was performed through cleanup with SPRI magnetic beads (Beckman Coulter). Subsequently, pooled cDNAs were converted into double-stranded DNA using the Second Strand Synthesis enzyme (NEB), followed by clean-up and linear amplification by T7 *in vitro* transcription overnight. The DNA template was then removed by Turbo DNase I (Ambion), and the RNA purified using SPRI beads. Amplified RNA was chemically fragmented using Zn2+ (Ambion), and then purified using SPRI beads. The fragmented RNA was ligated with ligation primers containing a pool barcode and partial Illumina Read1 sequencing adapter using T4 RNA ligase I (NEB). The ligated products were reverse-transcribed using the Affinity Script RT enzyme (Agilent Technologies) and a primer complementary to the ligated adapter, partial Read1. The cDNA was purified using SPRI beads. Libraries were completed by a PCR step using the KAPA Hifi Hotstart ReadyMix (Kapa Biosystems) and a forward primer containing the Illumina P5-Read1 sequence, and a reverse primer containing the P7-Read2 sequence. The final library was purified using SPRI beads to remove excess primers. Library concentration and molecular size were determined with a High Sensitivity DNA Chip (Agilent Technologies). Multiplexed pools were run on Illumina HiSeq2500 Rapid flow cells (Illumina).

C1 High-Throughput (HT-IFC)[33]

Cells were sorted into 15-ml tubes containing 7 ml of PBS with 5% FBS, using a Sony SH800 Cell Sorter. Cells were concentrated by centrifugation at 350 g for 5 minutes at 4ºC. The supernatant was removed, and cells were counted and diluted to 900 cells/ul for the Fluidigm C1 HT Small-Cell Integrated Fluidic Circuits (IFCs), and 450 cells/ul for the Fluidigm C1 HT Medium-Cell IFCs. A total of eight small-cell and seven medium-cell IFCs were used to generate cDNA on the Fluidigm C1 System. cDNA generation and the subsequent preparation of sequencing libraries were performed according to the recommended Fluidigm C1 HT protocols. Enrichment Primers from the Fluidigm reagent kit were replaced with NEBNext i5xx primers from NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1 & 2) (New England BioLabs), to enable library pooling. Libraries from fifteen IFCs were pooled and sequenced on the NovaSeq6000 system (Illumina) in two runs on the S2 flow cell.

ddSEQ (Bio-Rad)

Flow cytometry analysis and cell sorting were performed on the S3e Cell Sorter using ProSort Software (Bio-Rad Laboratories, #12007058) for acquisition and sorting. Viable cells were sorted into 1x PBS with + 0.1% BSA and kept at 4°C until scRNA-Seq. Cell concentration of sorted cells was determined using the TC20 Automated Cell Counter (Bio-Rad Laboratories, #1450102) and adjusted to a final concentration of 2,500 cells/ul. Cells were then prepared for single-cell sequencing using the Illumina Bio-Rad SureCell WTA 3' Library Prep Kit for the ddSEQ (Illumuina, #20014280). Cells were loaded onto ddSEQ cartridges and processed in the ddSEQ Single-Cell Isolator (Bio-Rad Laboratories, #12004336) to isolate and barcode single cells in droplets. First-strand cDNA synthesis occurred in droplets, which were then disrupted for second strand cDNA synthesis in bulk. Libraries were prepared according to manufacturer's instructions and then sequenced on the NextSeq500 system (Illumina).

gmcSCRB-seq[34]

Cells were sorted and processed using the alternative lysis (Guanidin) condition (gmcSCRB-seq) as described suitable for PBMCs in Bagnoli et al (2018). Briefly, single cells ("3 drops" purity mode) were sorted into 96-well DNA LoBind plates (Eppendorf) containing 5 µl lysis buffer using a Sony SH800 sorter (Sony Biotechnology; 100 µm chip). Lysis buffer consisted of 5 M guanidine hydrochloride (Sigma-Aldrich), 1% 2-mercaptoethanol (Sigma-Aldrich) and a 1:500 dilution of Phusion HF buffer (New England Biolabs). Samples were processed in six batches, with one batch of two plates and five batches of six plates. Each well was cleaned up using SPRI beads and resuspended in 4 µl H2O (Invitrogen) and a mix of 5 µl reverse transcription master mix, consisting of 20 units Maxima H- enzyme (Thermo Fisher), 2 × Maxima H- Buffer (Thermo Fisher), 2 mM each dNTPs (Thermo Fisher), 4 µM template-switching oligo (IDT), and 15% PEG 8000 (Sigma-Aldrich). For libraries containing ERCCs, 30,000 molecules of ERCC spike-in Mix 1 (Ambion) was used and the H2O (Invitrogen) was adjusted accordingly. After the addition of 1 µl 2 µM barcoded oligo-dT primer (E3V6NEXT, IDT), cDNA synthesis and template switching was performed for 90 min at 42 °C. Barcoded cDNA was then pooled in 2 ml DNA LoBind tubes (Eppendorf) and cleaned up using SPRI beads. Purified cDNA was eluted in 17 µl and residual primers digested with Exonuclease I (Thermo Fisher) for 20 min at 37 °C. After heat inactivation for 10 min at 80 °C, 30 µl PCR master mix consisting of 1.25 U Terra direct polymerase (Clontech) 1.66 × Terra direct buffer and 0.33 µM SINGV6 primer (IDT) was added. PCR was cycled as given: 3 min at 98 °C for initial denaturation followed by 19 cycles of 15 s at 98 °C, 30 s at 65 °C, 4 min at 68 °C. Final elongation was performed for 10 min at 72 °C. Batch 4 was erroneously denatured for 10 min due to a cycler error, but left in as we consider such errors as possible batch variation errors.

Following pre-amplification, all samples were purified using SPRI beads at a ratio of 1:0.8 with a final elution in 10 µl of H2O (Invitrogen). The cDNA was then quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher). Size distributions were checked using high-sensitivity DNA Fragment Analyzer kits (AATI) and high-sensitivity DNA Bioanalyzer kits (Agilent). As the samples had large primer peaks, they were purified a second time using SPRI beads at a ratio of 1:0.8 and then pre-amplified for an additional 3 cycles, as above. The cDNA was then purified and reanalyzed as above. Samples passing the quantity and quality controls were used to construct Nextera XT libraries from 0.8 ng of pre-amplified cDNA. During library PCR, 3′ ends were enriched with a custom P5 primer (P5NEXTPT5, IDT). Libraries were pooled and size-selected using 2% E-Gel Agarose EX Gels (Life Technologies), cut out in the range of 300–800 bp, and extracted using the MinElute Kit (Qiagen) according to manufacturer's recommendations. Libraries were paired-end sequenced on high output flow cells of an Illumina HiSeq 1500 instrument. Sixteen bases were sequenced with the first read to obtain cellular and molecular barcodes and 50 bases were sequenced in the second read into the cDNA fragment. An additional 8 base i7 barcode read was done to allow multiplexing.

Data analysis

Primary data preprocessing

FASTQ files for each technique were collected and processed in a unified manner. We developed a snakemake[35] workflow that streamlines all steps, including read filtering and mapping, quantification, downsampling and species deconvolution, and provides a Single Cell Experiment Object[36] output with detailed metadata. We used zUMIs[37], a single-cell processing tool compatible with all major scRNA-Seq protocols for filtering, mapping and quantification, ensuring comparable primary data processing between all methods. First, we discarded low-quality reads (barcodes and UMI sequences with more than 1 base below the Phred quality threshold of 20) and removed barcodes with less than 100 reads.

For techniques with known barcodes, we provided zUMIs with these barcode sequences, and used the automatic barcode detection function to detect the sequenced cells for other techniques. Next, cDNA reads were mapped to the human GRCh38, mouse GRCm38, and a human-mouse-dog mixed (for species level doublet detection) reference genomes using STAR[38]. Reads were then assigned to exonic and intronic features using featureCounts[39] and counted using the default parameters of zUMIs for human-only, mouse-only and mixed bam-files, separately. The output expression matrix of reads mapping to both exonic and intronic regions was selected for the downstream analysis. Of note, we included intronic counts in the expression quantification to improve gene detection and to enable a comparison with the snRNA-seq derived dataset. To deconvolute species, detect doublets and low quality cells, the mixed-species mapped data was used. Cells for which >70% of the reads mapped to only one species were assigned to the corresponding species. The remaining cells (those for which <70% of the reads mapped to only one species) were removed from the downstream analysis. Finally, for each technique, a *human* and *mouse* Single Cell Experiment object was created by combining the expression matrix and the metadata.

For subsequent data analysis, we discarded cells with <10,000 total number of reads as well as the cells having <65% of the reads mapped to their reference genome. Cells in the 95th percentile of the number of genes/cell and those having <25% mitochondrial gene content were included in the downstream analyses. Genes that were expressed in less than five cells were removed.

Clustering

Filtering, normalization, selection of highly variable genes (HVG), and clustering of cells were performed using the Seurat[40] package (version 2.3.4). The read counts were log-normalized for each cell using the natural logarithm of 1 + counts multiplied by a scale factor (10,000). To avoid spurious correlations, the library sizes were regressed out, and the genes were scaled and centered. The scaled Z-score values were then used as normalized gene measurement input for clustering and for visualizing differences in expression between cell clusters. We selected HVGs by evaluating the relationship between gene dispersion (y.cutoff = 0.5) and the log mean expression. The clustering procedure projects cells onto a reduced dimensional space, and then groups them into subpopulations by computing a shared-nearest-neighbour (SNN) based on the Euclidean distance (finding highly interconnected communities). The algorithm is a variant of the Louvain method, which uses a resolution parameter to determine the number of clusters.

In this step, the dimension of the subspace was set to the number of significant principal components (PC) based on the distribution of the PC standard deviations and by inspecting the ElbowPlot graph. The number of clusters was aligned to the expected biological variability, and cluster identities were assigned using previously described gene markers. T-SNE and UMAP were used to visualize the clustering distribution of cells. Cluster-specific markers were then identified using the Wilcoxon rank-sum test.

Trajectory analysis and pseudo-ordering of cells was performed using the Monocle[41] package (version 2.8.0) with the previously identified HVGs. Monocle works with the raw data and allows to specify the family distribution of gene measurements, which was set to a negative binomial, as defined in the family function from the VGAM package. As for the clustering, the expression space was reduced before ordering cells using the DDRTree algorithm. To validate cell populations, and for cell type identification and annotation, we used pseudotime ordering of single cells derived from the mouse colon.

Sample deconvolution and annotation

To identify and annotate cell types and states, we analyzed the individual single-cell experiments separately, taking advantage of the original sequencing depth. Gene expression counts were log-normalized to identify HVGs, as input to compute cell-to-cell distances and graph-based clustering (see Clustering). Cell clusters were visualized in two-dimensional space using t-SNE and UMAP, and then annotated by examining previously described cell population marker genes. All methods were able to recapitulate most cell types in both human and mouse samples, although in different proportions and resolutions.

In human samples, the T-cell marker CD3 was used to differentiate T-cells from other populations. While the CD4 T-cells cluster was clearly identifiable (with non-overlapping expression of markers), CD8 T-cells and Natural Killer (NK) were often intermixed. Monocytes were the second most abundant cell type, including subpopulations of CD14 and FCGR3A monocytes. High levels of CD79A and CD79B allowed the clear identification of B-cells. HEK293T cells generally fell into the same cluster, separate from blood subpopulations. They were clearly identifiable by the high number of detected genes (up to six-fold higher than PBMC populations). However, there was a correlation between the expression profiles of immune cells, leading in some instances to mixtures of PBMCs and HEK293T cells.

With few exceptions (Chromium), significantly fewer cells mapped to the mouse genome (half that of human cells, on average), leading to poorer clustering performance. However, the expected subpopulation composition of the colon was maintained overall. A small set of putative intestinal stem cells (Lgr5 and Smoc2 expression) were close (in transcriptional space) to rapidly proliferating transit amplifying (TA) cells (showing high ribosomal genes). Secretory cells (e.g. Muc2, Tff3, Agr2) resulted in a well-defined cluster. Enterocytes were more heterogeneous and ordered along their grade of lineage commitment. Notably, in some experiments two distinct clusters of enterocytes were identified, as well as a very small group of enterocyte progenitors. In addition to colon cells, fibroblasts and immune-cells were detected in all samples.

Reference datasets

To compare the efficiency of scRNA-seq protocols in describing the structure of a mixed population, we produced a reference dataset with 30,807 human and 19,749 mouse cells. Cells were clustered and annotated as described above. Due to the high number of cells, major cell types were clustered and clearly identifiable using population marker genes (**Supplementary Fig.2a-b**). However, to improve cell-to-cell annotations, we combined clustering with additional analyses. To annotate human blood cells, we used *matchSCore2* (see Methods) using an annotated set of 2700 PBMCs[15] as reference (**Supplementary Fig.2c-d**). We used cluster-specific markers of annotated populations as input to create a multinomial logistic model according to the *matchSCore2* algorithm. For each unknown cell, we assigned probability values for any possible cell identity, and the most likely identity was used for the classification (where this probability was >0.5; otherwise the cell was considered unclassified). Cell identities inferred by *matchSCore2* were highly consistent with clusters, with agreement ranging from 96% for CD4 T-cells to 100% for B-cells. Cell-by-cell prediction helped to identify smaller cell subsets, such as FCGR3A monocytes, dendritic cells and megakaryocytes. For all clusters, 17% of the cells remained unclassified (**Supplementary Fig.2c**). Half of these were previously annotated as HEK293T cells, which split into three different clusters because they varied in number of genes (**Supplementary Fig.2d**). Cells with fewer genes (cluster HEK293T cell2 and partially HEK293T cell3) were classified as CD4+ T-cells, although these did not show expression of any of the key blood markers. For the purposes of subsequent analysis, we removed the *unclear* cluster, representing 1% of the total number of cells, as well as the unclassified cells (except cells in HEK293T clusters). To further validate annotations, we assigned a score to each cell, corresponding to the overall expression of cell type signatures from the list of the top 100 computational markers (**Supplementary Fig.2d**). Transcriptional signatures revealed a set of cells from the HEK293T cell1 and HEK293T cell2 clusters showing high scores (>0.5, range 0-

1) for multiple signatures. We considered these as potential doublets, and removed them. The remaining cells were then used to compute an unbiased set of cell-type specific markers.
In the case of the mouse reference sample, we used clustering to dissect the colon subpopulation structure (excluding immune cells and fibroblasts). The largest cluster was formed by immature enterocytes (**Supplementary Fig.3a-b**). Other clusters included similar proportions of mature enterocytes, secretory cells, transit-amplifying cells and other undifferentiated cells. To refine annotations of immature cells, we ordered cells by intermediate states and projected them along a trajectory (see Clustering). The trajectory analysis (**Supplementary Fig.3c-d**) revealed 9 different states, ranging from intestinal stem cells and transit-amplifying cells (expressing high levels of Lgr5, Smoc2, Top2a) to enterocytes (Slc26a3, Saa1). Based on the pseudo-ordering and expression levels of previously described markers, states were merged into four major groups (**Supplementary Fig.3d**). For annotation, we labeled these four groups as Intestinal Stem cells (ISC), Transit Amplifying cells (TA), Enterocyte progenitors (Epr), and Enterocyte (E). We combined this finer-grained annotation with the remaining cell types, and then computed population-specific gene markers for training the reference model.

MatchSCore2

To systematically compare cell types from the analysis of different methods, we used *matchSCore2*, a mathematical framework for classifying cell types based on reference data (https://github.com/elimereu/matchSCore2). The reference data consists of a matrix of gene expression counts in individual cells whose identity is known. The following preliminary steps were applied before training the model:

- *Normalization of the reference data*
  Gene expression counts are log-normalized for each cell using the natural logarithm of 1 + counts per 10,000. Genes are then scaled and centered using the ScaleData function in the Seurat package.
- *Definition of signatures and their relative scores*
  For each of the identity classes in the reference data, positive markers were computed using the Wilcoxon Rank sum test. The top 100 ranked markers in each class were used as the signature for that class. To each cell, we assigned a vector x=(x_1,..,x_n) of signature scores, where n is the number of classes in the reference sample. The *i*-th signature score is computed as follow:

$$\sum_{j \text{ in } J} z_j$$

  where J is the set of genes in signature *i*, and z_j represents the z-score of gene *j*.

Statistical model

We proposed a supervised multinomial logistic regression model to explicitly infer cell identities. The model learns by training with the reference dataset, where n cell types and relatively ranked markers are defined. We assume that the distribution of signature scores is preserved, independent of which technology is used. The notion behind this model is that the random variable X=($X_{1,\dots,}X_n$), where X_i is the score in signature *i* across all cells*,* follows a multinomial distribution *M*(*s*= X_1+..+ X_n, π=($\pi_{1,\dots,}$ π_n)), where π_i represents the proportion of the *i-th* cell type in the training set. Training and test sets were created by subsampling the reference into two datasets, maintaining the original proportions of cell types in both sets. The model was trained by using the *multinom* function from the *nnet* R package (*decay*=1e-04, *maxit*=500). To improve the convergence of the model function, X_i variables were scaled to the interval [0,1].

Cell Classification

For each cell, model predictions consisted of a set of probability values per identity class, and the highest probability was used to annotate the cell if it was >0.5; otherwise the cell resulted unclassified.

Model accuracy

To evaluate the fitted model using our reference datasets, we assessed the prediction accuracy in the test set, which was around 0.9 for human and 0.85 for mouse reference. We further assessed *matchSCore2* classifications in datasets from other sequencing methods by looking at the agreement between clusters and classification. Notably, the resulting average agreement was of 80% (range: from 58% in gmcSCRB-seq to 92% in Quartz-Seq2), while the rate of unclassified cells was less than 2%.

Downsampling

To decide on a common downsampling threshold for sequencing depth per cell, we inspected the distribution of the total number of reads per cell for each technique, and chose the lowest first quartile (fixed to 20,000 reads/cell). We then performed stepwise downsampling (25%, 50% and 75%) using the zUMIs downsampling function. We omitted cells that did not achieve the required minimum depth.

Estimation of dropout probabilities

We investigated the impact of dropout events in HEK293T, monocytes and B-cells extracted for each technique on downsampled data (20,000 reads/cell). For datasets with >50 cells from the selected populations, we randomly sampled 50 cells to eliminate the effect of differing cell number. The dropout probability was computed using SCDE R package[42]. SCDE models the measurements of each cell as a mixture of a negative binomial process to account for the correlation between amplification and detection of a transcript and its abundance, and a Poisson process to account for the background signal. We then used estimated individual error models for each cell as a function of expression magnitude to compute dropout probabilities using SCDE's scde.failure.probability function. Next, we calculated the average estimated dropout probability for each cell type and technique. To integrate dropout measures into the final benchmarking score, we calculated the Area Under the Curve (AUC) of the expression prior and failure probabilities (Figure 2f). We expected that protocols that result in fewer dropouts would have lower AUC.

Cumulative number of genes

The cumulative number of detected genes in downsampled data was calculated separately for each cell type. For cell types with >50 cells annotated, we randomly selected 50 cells and calculated the average number of detected genes per cell after 100 permutations over n sampled cells, where n is an increasing sequence of integers from 1 to 50.

Silhouette scores

To measure the strength of the clusters, we calculated the Average Silhouette Width (ASW)[18]. The downsampled data (20,000 reads/cell) were clustered by Seurat[40], using graph based clustering with the first 8 principle components and resolution of 0.6. We then computed an average Silhouette width for the clusters using an Euclidean distance matrix (based on principle components 1 to 8). We report the average Silhouette width for each technique separately.

Dataset merging

Dataset integration across studies is one of the most challenging analyses. It is important to assess which scRNA-seq methods integrate best, while conserving biological variability. To integrate datasets, we used the R package *scMerge*[19], which uses a set of genes with stable expression levels across different cell types. Also, creating pseudo-replicates across datasets allows to estimate and correct for undesired sources of variability. To avoid differences due to sequencing depth, we combined downsampled count matrixes using the *sce_cbind* function, which includes the union of genes from different batches. After computing the set of highly variable genes using log-normalized gene measurements, we then apply the *scmerge* function to align data from different experiments. Following integration, cells were clustered using normalized gene expression levels and HVG computed using scMerge. We used UMAP plots color-coded by clusters and cell types to visualize and annotate clusters with the greatest agreement between cell types and clustering.

Clustering accuracy

To determine the clusterability of methods to identify cell types, we measured the probability of cells to be clustered with cells of the same type. Let C_k , $k \in \{1, \ldots, N\}$ the cluster of cells corresponding to a unique cell type (based on the highest agreement between clusters and cell types), and T_j , $j \in \{1, \ldots, S\}$ the set of different cell types, where $C \subseteq T$. For each cell type T_j, we compute the proportion p_{jk} of T_j cells that cluster in their correct cluster C_k. We define the cell-type separation accuracy as the average of these proportions.

Mixability

To account for the level of mixing of each technology, we used kBet[20] to quantify batch effects by measuring the rejection rate of a Pearson's χ^2 test for random neighborhoods. To make a fair comparison, kBet was applied to the common cell types separately by subsampling batches to the minimum number of cells in each cell type. Due to the reduced number of cells, the option heuristic was set to 'False', and the testSize was increased to ensure a minimum number of cells. Mixability was calculated by averaging cell type specific rejection rates.

Benchmarking score

To create an overall benchmarking score with which to compare technologies, we considered six key metrics: gene detection, overall level of expression in transcriptional signatures, cluster accuracy, classification probability, cluster accuracy after integration, and mixability. Each metric was scaled to the interval [0,1], then in order to equalize the weight of each metric score, the harmonic mean across these metrics was calculated to obtain the final Benchmarking scores. Gene detection, overall expression in cell type signatures, and classification probabilities were computed separately for B-cells, HEK293T cells and monocytes, and then aggregated by the arithmetic mean across cell types.

Supplementary Material

Supplementary Figure legends 1-11.

Supplementary Figures 1-11.

Supplementary Table 1.

Supplementary Fig. 1. Gene expression levels of selected marker genes.
UMAP visualization of normalized expression levels for selected marker genes of the most common PBMC (**a**) and colon (**b**) populations. Maps are shown for CD4+ T-cell markers IL7R and CD4 (expressed also in monocytes), the CD8+ T-cell marker CD8A, the B-cell marker CD79A, NK cell markers GNLY and NKG7, and monocyte-specific markers LYZ, CD14 and FCGR3A. In (**b**) maps are shown for markers of Intestinal Stem cell and proliferation (Smoc2, Miki67 and Top2a), secretory markers (Muc2, Agr2 and Tff3), enteroendocrine cell markers (Chga and Chgb), and enterocyte markers (Slc26a3, Car1 and Fabp2).

Supplementary Fig. 2. Identifying PBMC cell types using unsupervised clustering and classification. a. UMAP visualization of 38,195 human PBMC and HEK293T human cells coloured according to their assignment to clusters. Cluster labels are defined by examining the expression levels of known markers. **b.** Heatmap indicating the relative expression and gene detection rates for most common PBMC marker genes. **c.** UMAP vizualization of PBMC and HEK293T cells colour coded by cell classification inferred by matchSCore2. 17% of cells were unclassified and were removed from the analysis. **d.** UMAP visualization of cells showing the number of genes per cell, and scores for transcriptional signatures obtained by computing cell-type-specific markers (*lightgray*: low-score, *blue*: high score).

Supplementary Fig. 3. Identifying colon cell types by unsupervised clustering and trajectory analysis. a. UMAP visualization of 17,558 mouse colon cells. Cells are coloured by their assignment to clusters. Annotations are defined by examining the expression of known markers and differentially expressed genes (DEG). **b.** Heatmap of top DEG per cluster. Key markers of common colon cell populations are shown. **c.** Trajectory and pseudotime analysis of 8716 immature enterocytes (IE) showing the transition from intestinal stem cells (ISC) to enterocytes. Trajectories with the relative expression of known markers are shown (yellow: low, gray: mid, blue: high). **d.** (Top) Ordered cells are grouped into four different states according to their differentiation stage: intestinal stem cell (ISC), transit amplifying (TA), enterocyte progenitor (Epr), Enterocytes (E). (Bottom) UMAP visualization of IE cells coloured according to the four resulting states.

Supplementary Fig. 4. Clustering analysis of 13 sc/snRNA-seq methods. T-SNE visualizations of unsupervised clustering in human samples from 13 different methods. Each dataset is analyzed separately by taking advange of its original sequencing depth. Cells are coloured by cell type inferred by *matchSCore2*. Cells that did not reach a probability score of 0.5 for any cell type were considered unclassified.

Supplementary Fig. 5. Clustering analysis of 11 sc/snRNA-seq methods. T-SNE visualizations of the unsupervised clustering in mouse samples from 11 different methods. Each dataset is analyzed separately by taking advange of its original depth. Cells are coloured according to cell type inferred by *matchSCore2*. Cells that did not reach a probability score of 0.5 for any cell types were considered unclassified.

Supplementary Fig. 6. Comparison of 13 scRNA sequencing methods in mouse data.
a. Boxplots comparing the number of detected genes across protocols on downsampled data (20K), in mouse secretory and transit-apmlifying cells. Cell identities were defined by cell projection onto the reference. **b.** Number of genes detected at step-wise downsampled sequencing depths. Points represent the average number of genes detected for all cells of the corresponding cell type at the corresponding sequencing depth. **c.** Boxplots comparing the number of detected genes from countification of reads mapping to only Exonic regions, across protocols on downsampled data (20K), in Human HEK293T cells, monocytes and B-cells.

Supplementary Fig. 7. T-SNE representation of human cell types using highly variable genes.
a,b. T-SNE representation (calculated on first 8 principle components) on downsampled data (20K) using highly variable genes across protocols, separated by HEK293T cells, monocytes and B-cells and color coded by protocols (**a**) or the number of detected genes per cell (**b**).

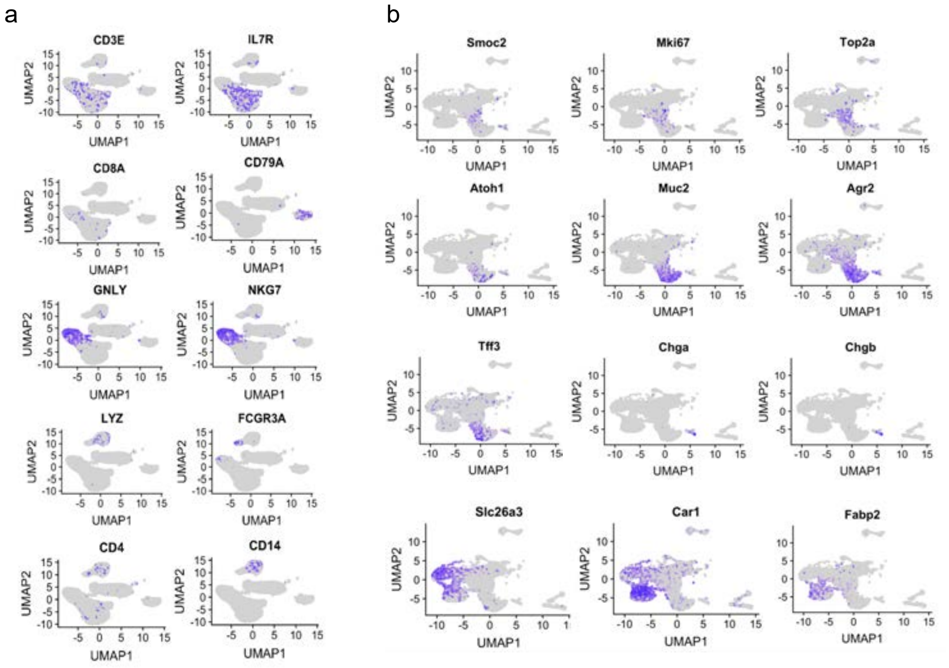
Supplementary Fig. 8. PCA representation of Human cell types using cell type markers
a,b. PCA analysis on downsampled data (20K) for HEK293T cells, monocytes and B-cells separately using the corresponding cell type's reference markers and color coded by protocols (**a**) or number of detected genes per cell (**b**).

Supplementary Fig. 9. Gene expression correlations across 13 sc/snRNA-seq methods. Pearson correlation plots between protocols using gene expression of cell-type-specific signatures for HEK293T cells (**a**), monocytes (**b**) and B-cells (**c**). Cells are ordered by agglomerative hierarchical clustering.

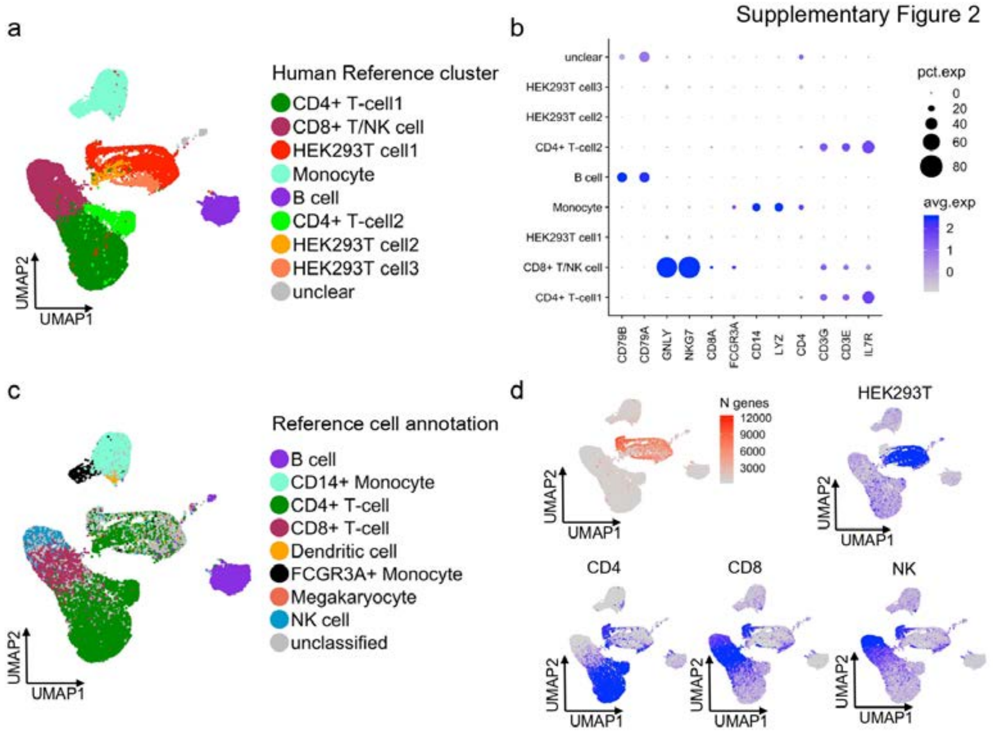
Supplementary Fig. 10. Analysis of integrated methods. a,b. UMAP visualization of clusters after the integration of technologies for human (**a**) and mouse (**b**) datasets. Cluster annotations are assigned on the basis of the most frequent cell type. **c,d.** Barplots showing normalized and method-corrected (integrated) expression scores in cell type specific signatures for CD4+ and CD8+ T-cells (**c**) and enterocytes 1, enterocytes 2 and intestinal stem cells (**d**). Bars are coloured by method. **e.** Evaluation of method integrability. Protocols are compared in their ability to group cell types into clusters (after integration) and to mix with other technologies within same clusters. Point sizes are indicating the level of downsampled sequencing depths. Dotted lines connect points from the same technology, highlighting the drop of integratability at lower depth. Points are coloured by sequencing method.

Supplementary Fig. 11. Comparison of mappability scores across technologies. Boxplots displaying minimum, 1st, 2nd, 3rd quantiles and maximum probabilities values (scores) obtained by *matchSCore2* in classifying most common cell types in human (**a,b**) and mouse (**c**) samples. B-cells, HEK293T cells and CD14+ monocytes are shown with data downsampled to 20K (**a**) and 10K (**b**) sequencing reads.

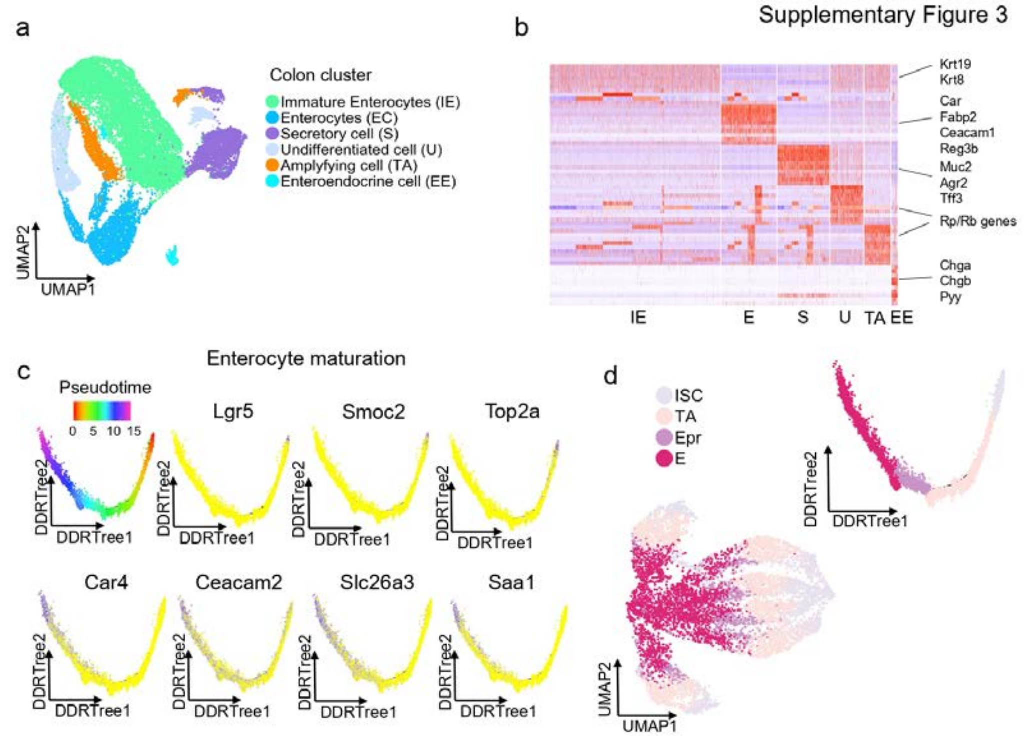
Supplementary Figure 1



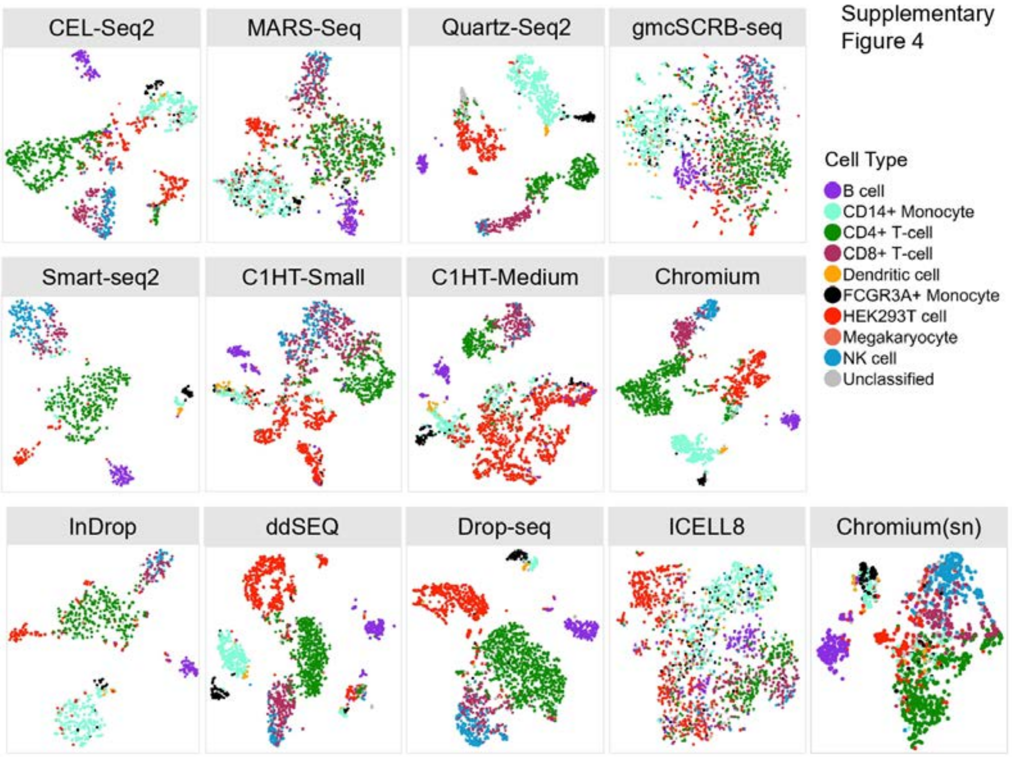
Supplementary Figure 2



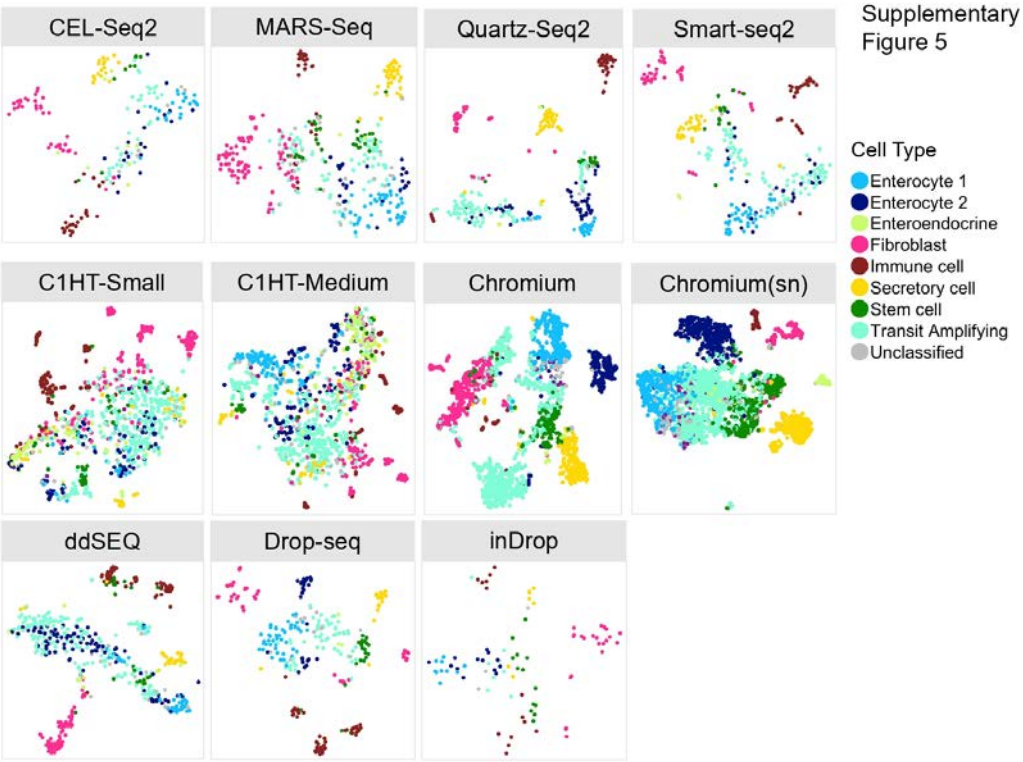
Supplementary Figure 3



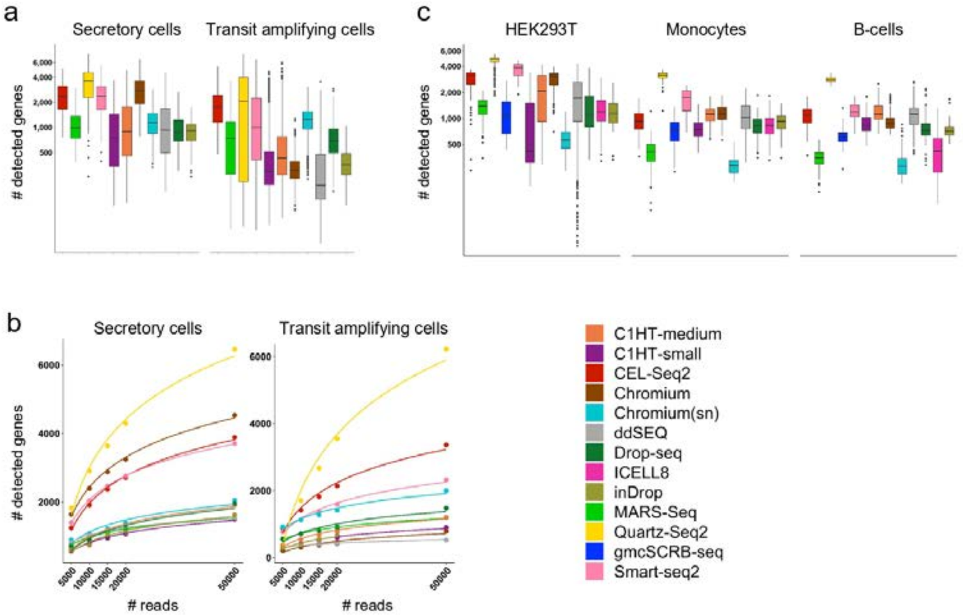
Supplementary Figure 4



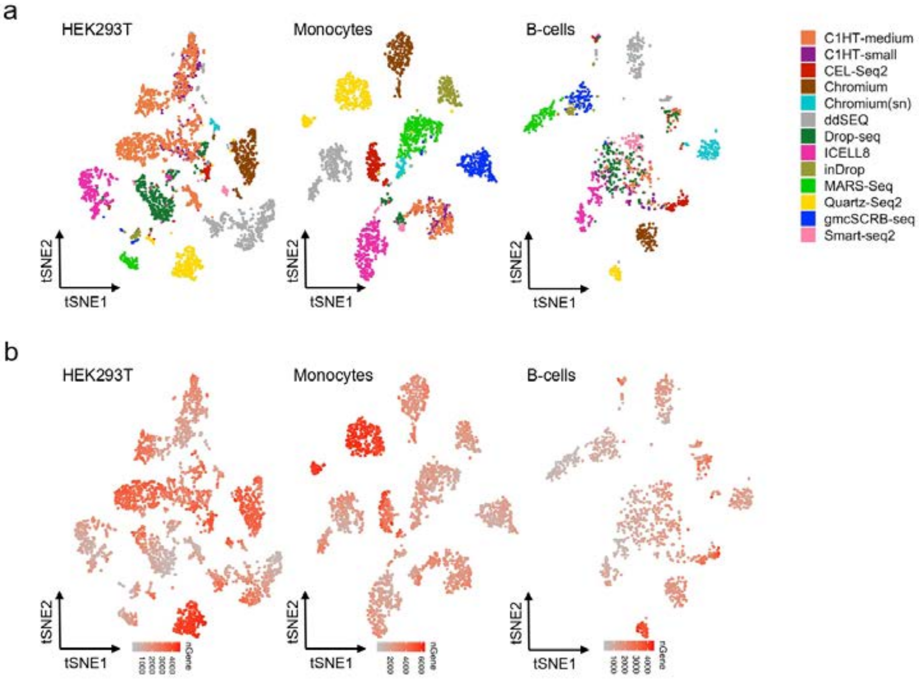
Supplementary Figure 5



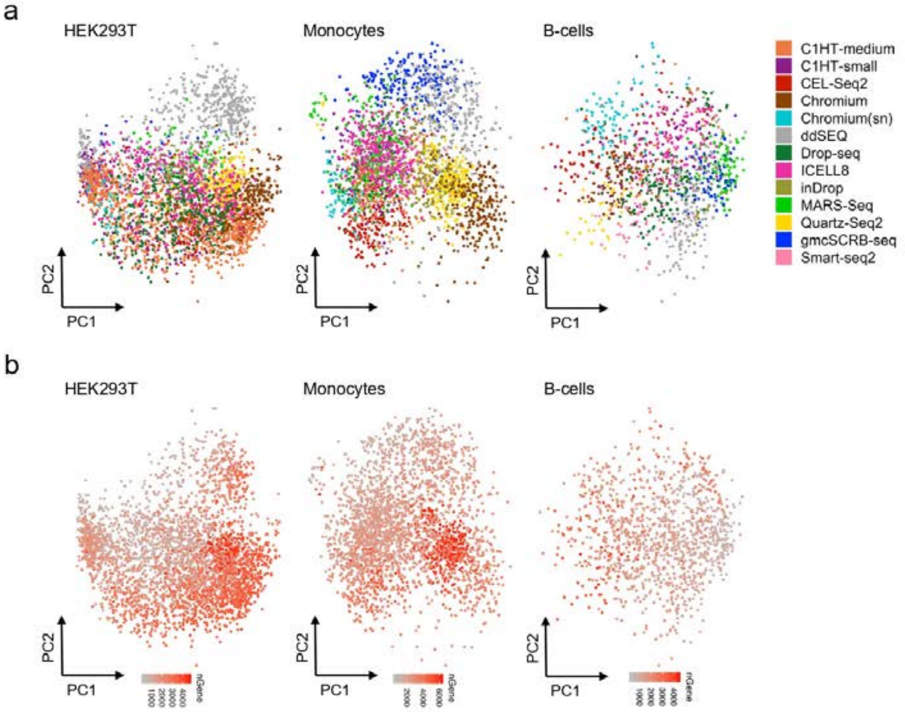
Supplementary Figure 6



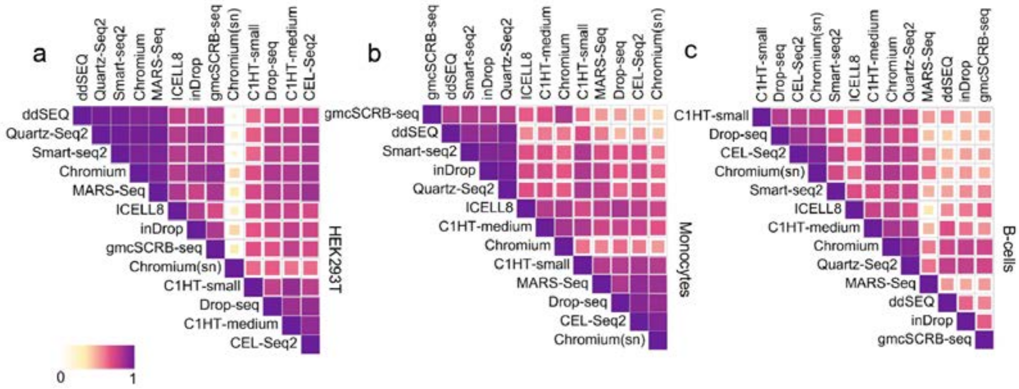
Supplementary Figure 7



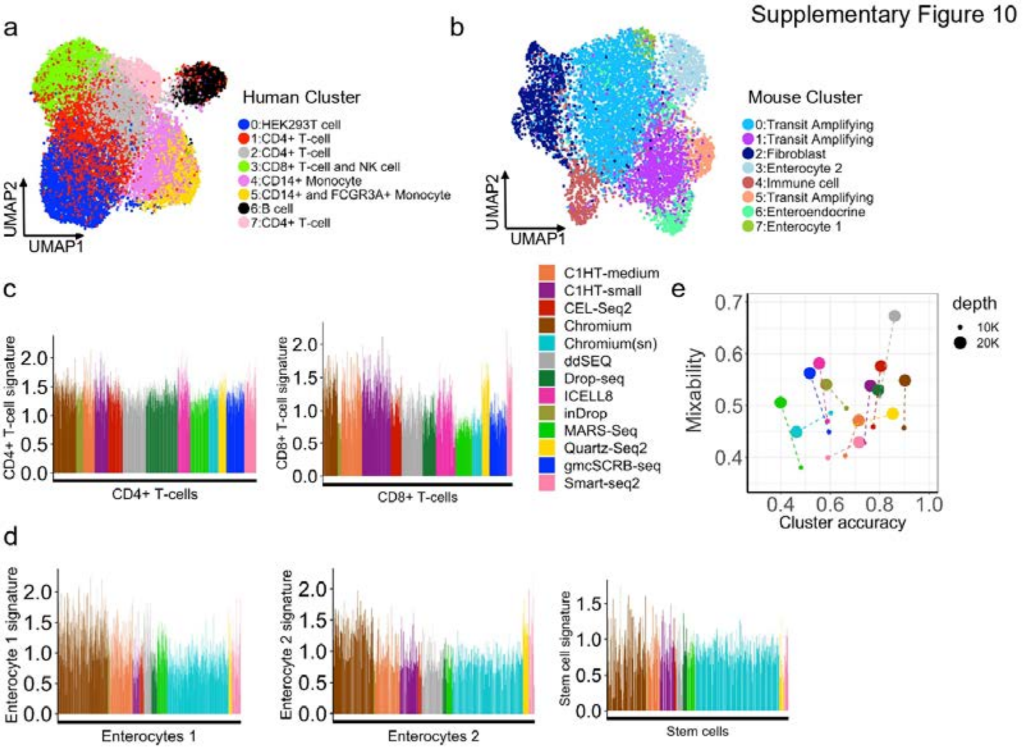
Supplementary Figure 8



Supplementary Figure 9



Supplementary Figure 10



Supplementary Figure 11

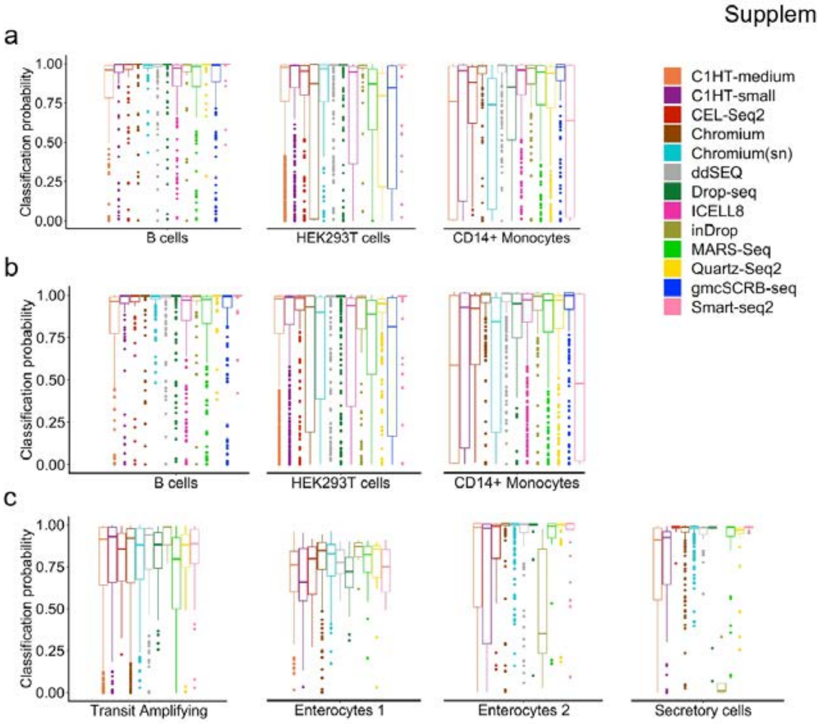


Table S1: Single-cell RNA sequencing methods and protocol specific features.

Method	Provider	Capture format	Cell loading	Single-cell indexing	Molecule identifier	cDNA amplification	Transcript coverage	Ref.
C1HT	Fluidigm	IFC	Trapping	OligoT primer	UMI	PCR	3´-end	1
CEL-Seq2		Plate	FACS	OligoT primer	UMI	IVT	3´-end	2,3
Chromium	10x Genomics	Droplets	Poisson	OligoT beads	UMI	PCR	3´-end	4
ddSEQ	Bio-Rad	Droplets	Double Poisson	OligoT beads	UMI	PCR	3´-end	5
Drop-seq	Dolomite	Droplets	Double Poisson	OligoT beads	UMI	PCR	3´-end	6
ICELL8	Takara Bio	Nanowells	Poisson	OligoT probes	UMI	PCR	3´-end	7
inDrops	1CellBio	Droplets	Poisson	OligoT beads	UMI	IVT	3´-end	8
gmcSCRB-seq		Plate	FACS	OligoT primer	UMI	PCR	3´-end	9
MARS-Seq		Plate	FACS	OligoT primer	UMI	IVT	3´-end	10
Quartz-Seq2		Plate	FACS	OligoT primer	UMI	PCR	3´-end	11
Smart-seq2		Plate	FACS	Tagmentation	N/A	PCR	Full-length	12

Table S1 References

1. Barriga, F. M. *et al.* Mex3a Marks a Slowly Dividing Subpopulation of Lgr5+ Intestinal Stem Cells. *Cell Stem Cell* **20**, 801-816.e7 (2017).
2. Herman, J. S., Sagar, null & Grün, D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* **15**, 379–386 (2018).
3. Sagar, null, Herman, J. S., Pospisilik, J. A. & Grün, D. High-Throughput Single-Cell RNA Sequencing and Data Analysis. *Methods Mol. Biol. Clifton NJ* **1766**, 257–283 (2018).
4. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
5. Sarkar, A. *et al.* Efficient Generation of CA3 Neurons from Human Pluripotent Stem Cells Enables Modeling of Hippocampal Connectivity In Vitro. *Cell Stem Cell* **22**, 684-697.e9 (2018).
6. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
7. Goldstein, L. D. *et al.* Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics* **18**, 519 (2017).
8. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187–1201 (2015).
9. Bagnoli, J. W. *et al.* Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 2937 (2018).
10. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
11. Sasagawa, Y. *et al.* Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* **19**, (2018).
12. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).

Chapter IV

Improving data quality of low performing protocols

As we demonstrated in the previous study [79], scRNA-seq protocols generate data of different quality. To computationally address such variance in quality output, we developed a computational approach that enhances data quality of low performing protocols by learning from best performing protocols. To this end, we calculated the latent space of the joint high and low quality datasets using bottleneck layer of a variational autoencoder (VAE), and applying vector arithmetic to generate a "transformation vector" that represents the average of differences between two datasets in the latent space. Later, by encoding low quality datasets into a latent space, adding the transformation vector and decoding it back to the high dimensional space, we were able to enhance the quality of the dataset. As an application example, we trained our model on data from the Mouse Cell Atlas (Tabula Muris consortium) and demonstrated the improvement in data quality of thymus cells. We decreased the level of dropout events, enhanced the expression level of known markers and reduced the variance and noise in the expression of expected markers from low quality dataset. The manuscript for this study is under preparation.

Bioinformatics, 2020, 0–0
doi: 10.1093/bioinformatics/
Advance Access Publication Date: DD Month YYYY
Application Note

Gene expression

scAutoTransfer: Improving the scRNA-seq data quality by learning from high-quality datasets

Atefeh Lafzi[1], Fabian j. Theis[3, 4, 5] and Holger Heyn[1, 2 ,*]

[1] CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain, [2] Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3] Helmholtz Zentrum München – German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany, [4] School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany, [5] Department of Mathematics, Technical University of Munich, Munich, Germany

*To whom correspondence should be addressed.

Associate Editor:

Received on; revised on; accepted on

Abstract

Motivation: Recent scRNA-seq benchmarking studies have demonstrated that scRNA-seq protocols generate data of different quality. Low-quality datasets suffer from lower library complexity and resolution, causing problems in downstream data analysis. To address such variance in quality output, we propose a computational approach to enhance data quality of low-performing protocols by learning from best-performing protocols using Variational Autoencoders (VAE).
Results: We trained a model on paired SMART-seq2 and 10x Genomics Chromium data from the Mouse Cell Atlas (Tabula Muris consortium) and demonstrated the improvement in data quality after correction using VAE. Data enhancing decreased the level of dropout events, enhanced the expression level of gene markers and reduced the variance and noise of low-quality datasets.
Availability: The pipeline and the codes available at https://github.com/ati-lz/scAutoTransfer
Contact: holger.hey@cnag.crg.eu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has provided the opportunity to increase the resolution of insights into biological systems. With technical advances, throughput and quality increased, considering the number of profiled cells and detected genes, respectively. Nevertheless, there is a variety of methods with different protocols providing data with different qualities. Recent benchmarking studies have measured the performance of available scRNA-seq protocols in great detail and provided a handbook to choose the best approaches to target biological questions [1], [2].

The scRNA-seq data analysis is challenged by so called dropout events, a high number of zeros in the expression matrix due to stochastic RNA losses or technical limitations, such as inefficiency of reverse transcription [3]. Several computational methods have been proposed to impute existing data [4]–[8]. These methods correct dataset without having a ground truth of high-quality data, subsequently called blind imputation. In this manuscript, we propose to improve the quality of existing low-quality datasets by learning from high-quality versions. We are proposing to use Variational Autoencoders (VAE) [9], a class of deep generative models, combined with linear extrapolation of the latent space using vector arithmetic. Differences between low- and high-quality data are captured in latent space after encoding, converted to a transformation vector, which is added to the low-quality manifold in the latent space, to generate a transformed version of the low-quality data after decoding. This model can be used to transform any out-of-sample data and improve the quality of the scRNA-seq analysis in general. Lotfollahi et al. [10] used such approach to predict perturbation responses and batch correction without improving the data quality. The use of Autoencoders to denoise scRNA-seq data has been previously suggested [11], which included blind imputing without learning from a ground truth. Here, we combine these ideas and propose a new application of the technique to transform and improve scRNA-seq data based on a higher quality dataset.

2 Materials and Methods

In order to demonstrate the application of this approach in improving the data quality, we analyzed datasets from the Tabula Muris consortium [12], which contain paired 10X Genomics and SMART-seq2 data for 20 mouse organs and tissues. Generating paired SMARTseq2-10X Genomics scRNAseq profiles has become a common practice in cell atlases (MCA, Allen Brain Atlas) since SMART-seq2 provides high resolution data at low-throughput due to the plate-based design and higher price. On the other hand, 10X Genomics provides high-throughput cell profiling at lower cost with the expense of losing resolution. Using these two data types together provides complementary information types. Our model transfers the quality of 10X Genomics data to SMART-seq2 data regardless of cell type and intracellular differences using VAEs and vector arithmetic. We selected 6 tissues (bladder, liver, lung, mammary gland, thymus and tongue) from Tabula Muris, paired datasets that contain sufficient cell numbers , to train our VAE. We excluded the thymus as our out-of-sample, a tissue showing heterogeneity in the combined latent space. To implement the model, we utilized scGen [10], a single-cell generator VAE model with an architecture adapted for scRNA-seq data (for more details on the model refer to [10]).

The model is based on the conditional distribution $P(x_i|z_i, p_i)$, which assumes that each cell *i* with expression profile x_i comes from a low-dimensional representation z_i in condition p_i; conditions are connectedto the two protocols SMART-seq2 and 10X Genomics. In this approach, VAE is used to model $P(x_i|z_i, p_i)$ in its dependence on z_i and vector arithmetic in latent space of VAE to model the dependence on p_i .

We excluded the thymus SMART-seq2 data from our training set (Fig 1.a) to apply the transformation learnt from other cell types. We predict the latent representation of SMART-seq2 version of thymus using the equation below:

$$\hat{z}_{i,\ thymus,\ p=SMARTseq2} = z_{i,\ thymus,\ p=10XGenomics} + \delta$$

In this equation, $\hat{z}_{i,\ thymus,\ p=SMARTseq2}$ and $z_{i,\ thymus,\ p=10XGenomics}$ denote the latent representation of thymus cells in each condition. δ is the transformation vector, the difference vector of means between cells in the training set in two corresponding conditions. After transformation in the latent space, $\hat{z}_{i,\ thymus,\ p=SMARTseq2}$ is mapped back to high-dimensional gene expression space using the generator network estimated while training the VAE.

This model is based on the assumption of the linearity of transformations in latent space, which has been discussed comprehensively in scGEN paper [10]. On the other hand, we also ran DCA (Deep Count Autoencoder) [11] on only 10X thymus data to compare its performance to our approach. The three datasets SMART-seq2, 10X Genomics and Transformed 10X were separately clustered in order to measure the information quality and impact on downstream analysis. We used Seurat R Package [13], [14] pipeline, to normalize the data, chose the top 2000 highly variable genes, selected the informative principle components and plotted UMAP representation of the data.

3 Results

We demonstrate an application of VAE in transforming scRNA-seq data between protocols. Exemplarily, we transformed mouse thymus 10x Genomics data to SMART-seq2 quality by learning from different mouse tissues (bladder, liver, lung, mammary gland, thymus and tongue) (Fig 1.A). Transformed 10x Genomics data using VAE and denoised data using DCA [11] was compared to the original dataset based on their quality and correlation with the high-quality SMART-seq2 data. The results showed, even though DCA-denoised data has significantly decreased dropout events (Fig 1.B), the correlation of gene expression with the SMART-seq2 data was very low; showing a high number of true negatives due to the blind imputation approach. On the other hand, the VAE-transformed data, showed a higher number of detected genes compared to original 10x Genomics data (low dropout events) as well as high correlation with the gold-standard SMART-seq2 data (Fig 2.C). Embedding both original and transformed 10x Genomics data together with the original SMART-seq2 data into the Principal Component (PC) space supported the quality improvements with the transformed data falling closer to the SMART-seq2 in a low dimensionality space (Fig 1.D, Supplementary Fig. to show the distance maybe). Also, the expression of top differentially expressed genes in SMART-seq2 showed a similar pattern in the transformed 10x Genomics data (Fig 1.E).

We further compared the quality of the predicted data to the original datasets by clustering each dataset separately and comparing the heterogeneity and quality of the obtain clusters. We clustered the three datasets separately (Fig 2.A) and matched the clusters between datasets by calculating the Jaccard Index of top 100 markers of each cluster (*matchSCore2* [1]) (Fig. 2.B). We detected all clusters of SMART-seq2 dataset (A, B, C, D, E) in original and transformed 10x Genomics data. In addition, we detected four well-defined clusters unique to the 10x Genomics datasets (A-2, B-2, C-2, 10X-NEW).

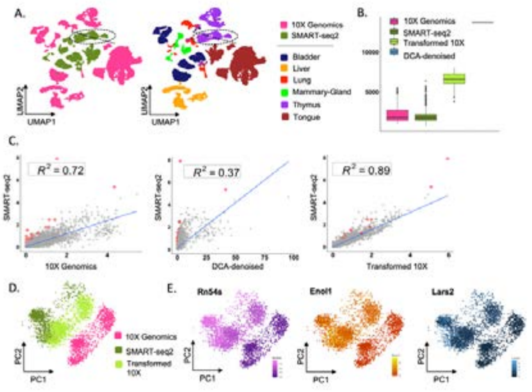


Fig. 1. A: UMAP representation of Tabula Muris SMART-seq2 and 10X Genomics selected tissue data color coded by the protocol and tissue. Here the Thymus cells from SMART-seq2 data has been put out for out-of-sample prediction. B: Distribution of total number of detected genes per thymus cell in each dataset. C: Pearson correlation plots of average expression per gene. R2 shows squared Pearson correlation. D: Principle component plot representing the first two principle components. E: The expression level of top DE genes on PCA plot.

To assess the quality of the markers of common clusters calculated from the SMART-seq2 data in the original and transformed datasets, we calculated the squared Pearson correlation (R2) of the average marker expression between SMART-seq2 and the 10x Genomics datasets. The results show that also at the cluster level, the transformation led to an improved correlation with the SMART-seq2 data (Fig 2.C). Finally, we assessed the stability and noise of markers for the 10x Genomics specific clusters. Comparing the variance and coefficient of variation of each gene among cells of each cluster showed that markers of transformed data have lower variance and a lower coefficient of variation, i-e- lower noise (Fig 2.D).

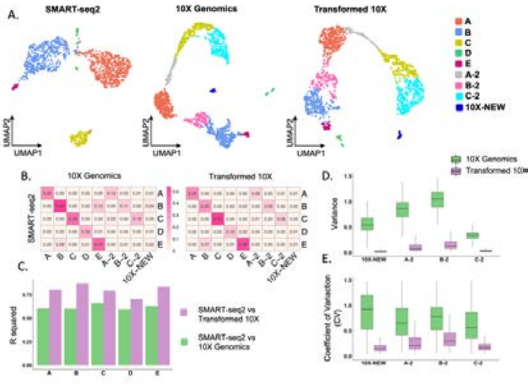


Fig. 2. Comparing the quality of transformed data. A:UMAP representation of clustering results in each of the datasets. B: Cluster matching heatmap based on highest Jaccard Index. C: R2 measures of gene based correlation between SMART-seq2 cluster markers and the two 10X dataset clusters. D: Boxplots showing the gene-based variance of 10X specific clusters. E: Boxplots showing the gene-based coefficient of variation in10X specific clusters.

4 Discussion and conclusion

In this manuscript, we proposed an approach to improve scRNA-seq data from low-performing protocols to the quality of high-performing methods. We demonstrated that combining VAE with vector arithmetic allows learning the transformation between two data types regardless of the cell type origin. As an example, we transformed 10x Genomics to SMART-seq2 data by training a model on multiple mouse tissues. We can envision other valuable application scenarios, for example the upgrading of data derived from older versions of the widely used 10x Genomics platform (e.g Chromium V2 and V3).

Acknowledgements

We thank all the members of Theis group at the Institute of Computational Biology (ICB) of HelmholtzZentrum munchen, especially Mohammad Lotfollahi and Malte Lucken for their valuable advices and sharing knowledge.

Funding

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. H2020-MSCA-ITN-2015-675752 (SINGEK). H.H. is a Miguel Servet (CP14/00229) researcher funded by the Spanish Institute of Health Carlos III (ISCIII).

Conflict of Interest: none declared.

References

[1] E. Mereu *et al.*, “Benchmarking Single-Cell RNA Sequencing Protocols for Cell Atlas Projects,” *bioRxiv*, p. 630087, 2019.

[2] J. Ding, “Systematic comparative analysis of single cell RNA-sequencing methods Jiarui,” *bioRxiv*, 2019.

[3] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, “Bayesian approach to single-cell differential expression analysis,” *Nat. Methods*, vol. 11, no. 7, pp. 740–742, 2014.

[4] T. Peng, Q. Zhu, P. Yin, and K. Tan, “SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data,” *Genome Biol.*, vol. 20, no. 1, p. 88, Dec. 2019.

[5] D. van Dijk *et al.*, “Recovering Gene Interactions from Single-Cell Data Using Data Diffusion,” *Cell*, vol. 174, no. 3, pp. 716-729.e27, Jul. 2018.

[6] W. V. Li and J. J. Li, “An accurate and robust imputation method scImpute for single-cell RNA-seq data,” *Nat. Commun.*, vol. 9, no. 1, pp. 1–9, Dec. 2018.

[7] K. Van den Berge *et al.*, “Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications,” *Genome Biol.*, vol. 19, no. 1, p. 24, Dec. 2018.

[8] A. T. L. Lun, K. Bach, and J. C. Marioni, “Pooling across cells to normalize single-cell RNA sequencing data with many zero counts,” *Genome Biol.*, vol. 17, no. 1, p. 75, Dec. 2016.

[9] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *2nd*

Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc., no. Ml, pp. 1–14, 2014.

[10] M. Lotfollahi, F. A. Wolf, and F. J. Theis, “scGen predicts single-cell perturbation responses,” *Nat. Methods*, vol. 16, no. 8, pp. 715–721, 2019.

[11] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, “Single-cell RNA-seq denoising using a deep count autoencoder,” *Nat. Commun.*, vol. 10, no. 1, pp. 1–14, 2019.

[12] N. Schaum *et al.*, “Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris,” *Nature*, vol. 562, no. 7727, pp. 367–372, 2018.

[13] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, “Integrating single-cell transcriptomic data across different conditions, technologies, and species,” *Nat. Biotechnol.*, vol. 36, no. 5, pp. 411–420, Jun. 2018.

[14] T. Stuart *et al.*, “Comprehensive Integration of Single-Cell Data,” *Cell*, pp. 1–15, 2019.

Discussion

In this thesis, we created a comprehensive view of single cell RNA-seq experimental, data generation and analysis strategies. We discussed the challenges and provided guidelines and solutions for individual problems. Starting from the initial steps of experimental design, we discussed that thoroughly planned and conducted sample preparation is critical to preserve cellular and RNA integrity to allow unbiased representation of sample composition. We provided guidelines about the correct steps for choosing suitable library preparation protocols for experiments. We explained in detail protocol differences and similarities, the amount of information they provide, the trade-off between more cells and lower resolution when going toward high-throughput techniques, the sequencing depth and budget restrain, in order to plan informative scRNA-seq experiments to target the specific biological questions. We did not stop with the experimental design but further discussed the correct sequence of analytical approaches from pre-processing to normalisation and more downstream analysis, leading to a better understanding of the information content. Analysis and interpretation of single-cell transcriptomes is enabled by a wealth of computational methods specifically tailored to answer biological questions in a hypothesis-free manner or guided by previous knowledge. We demonstrated the value such scRNA-seq experiments by detecting dermal fibroblasts changes in aging. In this study, we showed that old dermal fibroblasts acquire more transcriptional noise and adipogenic traits. They lose production and secretion of Extra Cellular Matrix (ECM) components yet concomitantly upregulate the expression of genes involved in inflammation, lipid metabolism, and adipogenesis.

As member of Standard and Technology Working Group (STWG) of Human Cell Atlas (HCA) Consortium, we performed a comprehensive benchmarking study of scRNA-seq protocols for cell atlas projects. We designed a complex sample with different cell types from small to big, homogeneous to heterogeneous, intermediate to differentiated cells to simulate scenarios that can occur in a cell atlas project. We performed 13 different scRNA-seq protocols on this sample and processed the data in a controlled way to reduce the bias that can come from any other source of variation but library preparation protocol. Our comparative study demonstrated the differences between protocols in many different aspects from library complexity and droupout probability, to clusterability and mixability with other datasets.

With the benchmarking study, we provided an idea of the performance level of different protocols from different point of views. As a follow up on this study, we provided a deep learning method to enhance the quality of the data coming from poor-performing protocols by learning from the data coming from high-performing protocol. Using Autoencoders and latent space arithmetic, we developed a framework in which the data is encoded into latent space, a transformation vector is calculated and added to the out of sample part of the data, and finally the transformed data is decoded back to the high dimensional space.

Single-cell transcriptomics technologies are advancing very rapidly. With the technological improvement, studies are getting more ambitious in terms of the number of profiled cells as well as extracting information beyond a transcriptional profile. Although the trade-off between high-throughput techniques and decreased molecular capture rate is still a challenge, some commercial systems like 10X Genomics Chromium, have significantly improved their product performance. Recent versions of the Chromium, namely V3 and V3.1 have shown an increased sensitivity in molecular capture and gene detection. Low molecule capture affects sparsity of the gene expression matrix at higher levels in downstream analysis. Even though different computational methods have been proposed to solve this *in silico,* from modelling technical zeros with probabilistic models to machine learning approaches for data reconstruction [80], this is still an important concern especially when dealing with large numbers of cells.

One of the main goals of scRNA-seq is to capture transcriptional differences between subpopulation of cells within a heterogenous sample. So far, most methods have focused on dissecting the heterogeneity first – using different clustering algorithms – and then comparing average expression between clusters [77,81]. Although this approach have provided insights by defining transcriptional markers of the heterogeneity within the sample, it does not uniformly outperform classical bulk methods [82]. Also, there is a level of uncertainty in current clustering analysis and cell type assignment, which the subsequent differential expression analysis rely on. Hence, other options of differential expression analysis at the single-cell level need to be explored. In this regard, some attention has been given to more general patterns of differential expression, such as changes in trajectory along pseudotime [83,84], changes in variability that account for mean expression confounding [85] or more generally changes in distributions [86].

Annotating inferred subpopulation of cells is another challenge in the field. Despite many methodological improvements, it is still a common practice to manually annotate clusters based on previous defined marker signatures or curated cell type marker sets. However, this labor-intensive process may generate biased and irreproducible results, and challenges comparisons between different datasets [87]. Dealing with novel, rare and unstudied cell populations, amplifies the problem since there are no public resources available to refer to. GO (Gene Ontology) enrichment analysis on cluster markers can provide ideas about functional characteristics of subpopulations, but they mostly explain general pathways and lack detail for a solid annotation at single cell level. Also, there is an inherent incompleteness in GO terms as they represent the current state of knowledge, while single cell studies aim to explore unknown biology. To overcome these challenges, we need i) a reliable, comprehensive reference which is tailored to provide high-resolution cellular functions at the single cell level, and ii) algorithms and tools to automatically label cells and clusters in scRNA-seq data. To address the first requirement, various tissue and species atlases are being created such as Planaria [88,89], Mouse [90,91] and Human Cell Atlases [92]. For the methodological requirement, there have been efforts to automatically classify cells based on a reference atlas [87,93–95]. These recent and future advances will enable more precise, robust and reproducible annotations.

Although we have discussed batch effect problems in scRNA-seq experiments and practical solutions, it remains a big problem when batches are confounded with different biological samples or when the library preparation method does not allow a mixed sample design. Cell hashing has been introduced by the Satija group to allow sample multiplexing and super-loading of microfluidic systems for batch effect and significant cost reduction, respectively [96]. This technique is based on oligo-tagged antibodies that uniquely label cells from distinct samples. By sequencing these tags alongside the cellular transcriptome, we can assign each cell to its original sample, to robustly identify sample identity and cross-sample multiplets. An arising topic in scRNA-seq, which is the next level of batch correction problem, is dataset alignment and merging. With more datasets being generated on tissues in different conditions, perturbation states, or even simply different laboratories, researchers are trying to capture the communalities and differences between single-cell datasets from a broader perspective. Algorithms that had been initially developed to correct replicate-level batch effects, improved to solve higher degree differences. Datasets can now be merged using the common cell types and states while keeping the condition-specific subpopulations unique [97–99]. Although these merging algorithms provide opportunities to deal with multiple datasets from different experiments and individual or patients, they have caused confusion for data analysts. There is a wide range of data integration methods that expands from correcting classic replicate-level batches by simply regressing out the batch covariate, to more sophisticated techniques that aim to integrate even single-cell data from different modalities (e.g. scATAC-seq, spatial transcriptomics). Each of this techniques corrects for differences between datasets at different level and using a stringent method to integrate datasets that harbour subtle differences may hide biological signals. In conclusion, there is a strong need for guidelines for method selection for single-cell datasets integration.

Another challenge that remains is to distinguish between cell doublets and cells that are in an intermediate cellular state. As intermediate cell states show transient transcriptional profile of two distinct cell subpopulation, their profiles are challenging to separate from cell doublets or multiplets. Cell hashing can help to address this problem to some extent to identify intra-sample doublets [96]. Also computational approaches have been developed aiming to uncover doublets by comparing the distance of each cell's transcriptomic profile to that of *in silico* generated "synthetic doublets" [100,101], though these techniques are not completely solving this problem.

After defining single cell transcriptome profiles [3], it did not take too long for the development of techniques to profile genome of single cells and to detect mutations and Copy Number Variations (CNVs) [102,103]. Techniques have been developed to extract both genome and transcriptome information of individual cells simultaneously, such as G&T-seq [104] or an alternative method developed by Han *et al.* [105]. Later, studies started to provide solutions to detect different types of information, such as chromatin accessibility with scATAC-seq [106], scChIP-seq [107], single cell small-RNA sequencing [108]. Generally, bulk sequencing approaches are being updatedto higher resolution with the hope of collecting more precise and accurate knowledge. Integrating and aligning the data from different omics presents a new challenge for the computational biologists.

So far, all the techniques we had discussed in this thesis were based on tissue dissociation, leading to a loss of tissue coordinates of the cells. However, the physical location of cells within the tissue is a key determinant of its molecular identity and function. Hence, several methods were developed for obtaining whole genome measurements while accounting for the spatial localization of cells. Recent state-of-the-art spatial transcriptomics techniques are based on two main approaches: 1) Imaging-based methods, which started with development of single molecule fluorescence *in situ* hybridization (smFISH) [109]. This technique acquires spatial transcript quantification *in situ* by using libraries of multiple short oligonucleotide probes, each labelled with a single fluorophore. Next, by specific accumulation of these fluorescent probes on the target mRNA, individual transcripts can be visualized as diffraction-limited spots by fluorescence microscopy. One of the most important limitations of smFISH is the small number of transcripts that can be identified simultaneously due to the limited number of fluorophores that are suitable for parallel use [110]. There are different approaches developed to address this limitation either by combinatorial labelling [111,112], sequential hybridization[113], multiplexed error-robust fluorescence in situ hybridization (MERFISH) [114,115], cyclinc–ouroboros smFISH (osmFISH) [116] and seqFISH+ [117]. Increasing the number of targets and converting image-based methods to be able to measure transcription in high throughput, brings in computational challenges specially in image processing. Methods that run sequential rounds of imaging will generate large amount of image data, which will be difficult to store, analyse and merge. 2) *In situ* sequencing is another approach that enables an unbiased census of all RNA molecules while preserving localization. This approach replaces the flow cell with the original tissue of interest and then utilizes *in situ* cDNA synthesis, cDNA amplification and cross-linking. Ke *et al*. first introduced this technique using padlock probes to initiate targeted cDNA synthesis in situ [118]. Lee *et al*., further developed fluorescent *in situ* RNA sequencing (FISSEQ) by generating 150K short 30 bp reads that were mapped to 8100 genes in fibroblasts [119]. They applied their technique on intact tissues like drosophila embryos and mouse brain sections [120]. In 2016, Stahl et al., developed another spatial method that uses a glass slide that spatially tags mRNA before library assembly [121]. In this technique, each polyA-capturing probe contains of a positional barcode, unique molecular identifier (UMI), and library adaptor sequences; spatial information is preserved by the positional barcodes. The technique was commercialized in 2017 by the company Spatial Transcriptomics, which since beginning of 2019 belongs to 10X Genomics Inc. These techniques are state-of-the-art in the field and are quickly being improved in resolution, throughput and data quality. Regardless of the technique, high-quality spatial transcriptomics data will soon provide new perspectives to explore inter-cellular and intra-cellular system. It will provide information about how cells are interacting within tissues to achieve their concerted function, while monitoring mRNA localization and cellular polarity at intra-cellular level [122]. As for scRNA-seq, there is uncertainty in the measurements from spatial techniques which can propagate to downstream analysis. A way to tackle this uncertainty in image-based spatial measurements, is to extract additional information such as cell shape, size and subcellular spatial distribution of transcripts to guide the clustering and classification process.

In summary, although single cell genomics with all its applications has significantly helped in our understanding of biological systems, it is still at its early stages and there are important challenges that needs to be addressed in order to get the best out of it.

Conclusion

Single-cell genomics has revolutionized the research of the last decade in cellular biology by providing a way to study cells individually from different perspectives. Despite being a hot topic, going for single-cell experiments in any biological study without a clear understanding of the type of information it provides, and an informed experimental design, may lead to huge disappointment and waste of resources. Hence, a comprehensive understanding of the technology, its application and characteristics, as well as guidelines for experimental design at the wet- and dry-lab is required. In this thesis, we provided detailed guidelines of the major steps of a scRNA-seq experiment and data analysis, and demonstrated an example of its power in solving a biological question by choosing the adequate protocol and analysis approaches. We also showed that scRNA-seq protocols differ significantly in quality of the data they produce, which is an important point to take into account when designing an experiment.

Bibliography

1. Komin N, Skupin A. ScienceDirect Systems Biology How to address cellular heterogeneity by distribution biology. *Curr Opin Syst Biol*. 2017;3:154-160. doi:10.1016/j.coisb.2017.05.010
2. Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. 2008;5(7):1-8. doi:10.1038/NMETH.1226
3. Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell. 2009;6(5). doi:10.1038/NMETH.1315
4. Eberwine J, Finnell R, Zettel M, Coleman P. Analysis of gene expression in single live neurons. 1992;89(April):3010-3014.
5. Sheng M, Cummings J, Roldan LA, Jan YN, Jan LY. Changing subunit composition of heteromeric NMDA receptors during development of rat cortex. *Nature*. 1994;368(6467):144-147. doi:10.1038/368144a0
6. Tietjen I, Rihel JM, Cao Y, et al. Single-Cell Transcriptional Analysis of Neuronal Progenitors 3380 Central Expressway. 2003;38(Figure 1):161-175.
7. Kurimoto K, Yabuta Y, Ohinata Y, et al. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. 2006;34(5). doi:10.1093/nar/gkl050
8. Esumi S, Wu S, Yanagawa Y, Obata K, Sugimoto Y, Tamamaki N. Method for single-cell microarray analysis and application to gene-expression profiling of GABAergic neuron progenitors. 2008;60:439-451. doi:10.1016/j.neures.2007.12.011
9. Kamme F, Salunga R, Yu J, et al. Single-Cell Microarray Analysis in Hippocampus CA1 : Demonstration and Validation of Cellular Heterogeneity. 2003;23(9):3607-3615.
10. Tang F, Barbacioru C, Bao S, et al. Resource Tracing the Derivation of Embryonic Stem Cells from the Inner Cell Mass by Single-Cell RNA-Seq Analysis. *Stem Cell*. 2010;6(5):468-478. doi:10.1016/j.stem.2010.03.015
11. Ramsköld D, Luo S, Wang Y, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. 2012;30(8). doi:10.1038/nbt.2282
12. Guo G, Huss M, Tong GQ, et al. Resource Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Dev Cell*. 2010;18(4):675-685. doi:10.1016/j.devcel.2010.02.012

13. Islam S, Kja U, Moliner A, Zajac P, Fan J, Linnarsson S. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. 2011:1160-1167. doi:10.1101/gr.110882.110.1160

14. Zeisel A, Linnarsson S. Ac knowled gme nts. 2015:0-5.

15. Hashimshony T, Wagner F, Sher N, Yanai I. by Multiplexed Linear Amplification. *CellReports*. 2012;2(3):666-673. doi:10.1016/j.celrep.2012.08.003

16. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods*. 2013;10(11):1096-1100. doi:10.1038/nmeth.2639

17. Picelli S, Sagasser S. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. 2014:2033-2040. doi:10.1101/gr.177881.114.Freely

18. Jaitin DA, Amit I. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. 2014;(February).

19. Muraro MJ, Dharmadhikari G, Gru D, Carlotti F, Dharmadhikari G, Gru D. A Single-Cell Transcriptome Atlas of the Human Pancreas. 2016:385-394. doi:10.1016/j.cels.2016.09.002

20. Hashimshony T, Senderovich N, Avital G, et al. CEL-Seq2 : sensitive highly-multiplexed. *Genome Biol*. 2016:1-7. doi:10.1186/s13059-016-0938-8

21. Mazutis L, Gilbert J, Ung WL, Weitz DA, Griffiths AD, Heyman JA. Single-cell analysis and sorting using droplet-based microfluidics. 2013:54-56. doi:10.1038/nprot.2013.046

22. Klein AM, Mazutis L, Weitz DA, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells Resource Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*. 2015;161(5):1187-1201. doi:10.1016/j.cell.2015.04.044

23. Macosko EZ, Basu A, Regev A, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets Resource Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202-1214. doi:10.1016/j.cell.2015.05.002

24. Zheng GXY, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. 2017. doi:10.1038/ncomms14049

25. Bose S, Wan Z, Carr A, et al. Scalable microfluidics for single-cell RNA printing and sequencing. *Genome Biol*. 2015:1-16. doi:10.1186/s13059-015-0684-3

26. Gierahn TM, Ii MHW, Hughes TK, et al. RNA sequencing of single cells at high throughput. 2017;14(4). doi:10.1038/nmeth.4179

27. Sasagawa Y, Nikaido I, Hayashi T, et al. Quartz-Seq : a highly reproducible and sensitive single-cell RNA sequencing method , reveals non- genetic gene-expression heterogeneity. 2013:1-17.

28. Sasagawa Y, Danno H, Takada H, et al. Quartz-Seq2 : a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. 2018:1-24.

29. Rosenberg AB, Roco CM, Muscat RA, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science (80-)*. 2018;360(6385):176-182. doi:10.1126/science.aam8999

30. Cao J, Packer JS, Ramani V, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (80-)*. 2017;357(6352):661-667. doi:10.1126/science.aam8940

31. Habib N, Li Y, Heidenreich M, Swiech L. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. 2016;353(6302).

32. Lake BB, Ai R, Kaeser GE, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. 2016;352(6293):352-357.

33. Guillaumet-adkins A, Rodríguez-esteban G, Mereu E, et al. Single-cell transcriptome conservation in cryopreserved cells and tissues. 2017:1-15. doi:10.1186/s13059-017-1171-9

34. Lafzi A, Moutinho C, Picelli S, Heyn H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat Protoc*. 2018;13(12). doi:10.1038/s41596-018-0073-y

35. Islam S, Zeisel A, Joost S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. 2014;11(1). doi:10.1038/nmeth.2772

36. Tirosh I, Suvà ML. Deciphering Human Tumor Biology by Single-Cell Expression Profiling. *Annu Rev Cancer Biol*. 2019;3(1):151-166. doi:10.1146/annurev-cancerbio-030518-055609

37. Giustacchini A, Thongjuea S, Barkas N, et al. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat Publ Gr*. 2017;23(6):692-702. doi:10.1038/nm.4336

38. Stubbington MJT, Lönnberg T, Proserpio V, et al. T cell fate and clonality inference from single-cell transcriptomes. 2016;13(4). doi:10.1038/nmeth.3800

39. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single- seq data. *Nat Rev Genet*. 2019;20(May):273-282. doi:10.1038/s41576-018-0088-9

40. Andrews TS, Hemberg M. Gene expression M3Drop : dropout-based feature selection for scRNASeq. 2019;35(December 2018):2865-2867. doi:10.1093/bioinformatics/bty1044

41. Baran-gale J, Chandra T, Kirschner K. Experimental design for single-cell RNA sequencing. 2017;17(November 2017):233-239. doi:10.1093/bfgp/elx035

42. G FG, Onana CA. High Dimensional Data Visualization : Advances and Challenges. 2017;162(10):23-27.

43. Kurasova O, Marcinkeviˇ V, Medvedev V, Rapeˇ A. Strategies for Big Data Clustering. 2014. doi:10.1109/ICTAI.2014.115

44. Chenna P. Comparative Study of Dimension Reduction Approaches With Respect to Visualization in 3-Dimensional Space. 2016.

45. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*. 1933;24(6):417-441. doi:10.1037/h0071325

46. Maaten L Van Der, Hinton G. Visualizing Data using t-SNE. 2008;9:2579-2605.

47. Malte D L, Theis FJ. Current best practices in single-cell RNA-seq analysis : a tutorial. 2019. doi:10.15252/msb.20188746

48. Mcinnes L, Healy J, Melville J. UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction arXiv : 1802 . 03426v2 [stat . ML] 6 Dec 2018. 2018.

49. Altschuler SJ, Wu LF. Essay Cellular Heterogeneity : Do Differences Make a Difference ? 2010:559-563. doi:10.1016/j.cell.2010.04.033

50. Hartuv E, Shamir R. Clustering algorithm based on graph connectivity. *Inf Process Lett*. 2000;76(4-6):175-181. doi:10.1016/S0020-0190(00)00142-3

51. Xu C, Su Z. Gene expression Identification of cell types from single-cell transcriptomes using a novel clustering method. 2015;31(February):1974-1980. doi:10.1093/bioinformatics/btv088

52. Levine JH, Simonds EF, Bendall SC, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis Resource Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*. 2015;162(1):184-197. doi:10.1016/j.cell.2015.05.047

53. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. 2008;10008. doi:10.1088/1742-5468/2008/10/P10008

54. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data Resource Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031

55. Wolf FA, Angerer P, Theis FJ. Open Access S CANPY : large-scale single-cell gene expression data analysis. 2018:1-5.

56. Duò A, Robinson MD, Soneson C. systematic performance evaluation of clustering methods for single-cell RNA-seq data [version 2 ; referees : 2 approved] Referee Status : 2018;(0). doi:10.12688/f1000research.15666.1

57. Tian L, Dong X, Freytag S, et al. experiments. *Nat Methods*. 2019;16(June). doi:10.1038/s41592-019-0425-8

58. Magwene PM, Lizardi P, Kim J. Reconstructing the temporal ordering of biological samples using microarray data. 2003;19(7):842-850. doi:10.1093/bioinformatics/btg081

59. Oja E, Hyva A. Independent component analysis : algorithms and applications. 2000;13:411-430.

60. Trapnell C, Cacchiarelli D, Grimsby J, et al. letters The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014;32(4):381-386. doi:10.1038/nbt.2859

61. Ji Z, Ji H. TSCAN : Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. 2016;44(13):1-17. doi:10.1093/nar/gkw430

62. Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-cell trajectories. 2017;14(10). doi:10.1038/nmeth.4402

63. Street K, Risso D, Fletcher RB, et al. Slingshot : cell lineage and pseudotime inference for single-cell transcriptomics. 2018:1-16.

64. Bendall SC, Davis KL, Amir ED, et al. Resource Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell*. 2014;157(3):714-725. doi:10.1016/j.cell.2014.04.005

65. Coifman RR, Lafon S, Lee AB, et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data : Diffusion maps. 2005;(10).

66. Haghverdi L, Buettner F, Theis FJ. Gene expression Diffusion maps for high-dimensional single-cell analysis of differentiation data. 2015;31(May):2989-

2998. doi:10.1093/bioinformatics/btv325

67. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. 2016;13(10). doi:10.1038/nmeth.3971

68. Reid JE, Wernisch L. Gene expression Pseudotime estimation : deconfounding single cell time series. 2016;32(June):2973-2980. doi:10.1093/bioinformatics/btw372

69. Lönnberg T, Svensson V, James KR, et al. Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves T H 1 / T FH fate bifurcation in malaria. 2017;2192(March):1-12.

70. Campbell KR, Yau C, Gitter A. Probabilistic modeling of bifurcations in single-cell gene expression data using a Bayesian mixture of factor analyzers [version 1 ; peer review : 2 approved]. 2017;19:1-17.

71. Monge G. Prof . Cayley on Mongers. 1781.

72. Schiebinger G, Shu J, Tabaka M, Jaenisch R, Regev A, Lander ES. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Resource Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*. 2019;176(4):928-943.e22. doi:10.1016/j.cell.2019.01.006

73. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2 , a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. 2014;9(6):1-12. doi:10.1371/journal.pone.0098679

74. Furlan A, Kamenev D, Petersen J, Sunadome K, Memic F. Multipotent peripheral glial cells generate neuroendocrine cells of the adrenal medulla. 2017;3753. doi:10.1126/science.aal3753

75. La Manno G, Soldatov R, Zeisel A, et al. RNA velocity of single cells. *Nature*. 2018. doi:10.1038/s41586-018-0414-6

76. Mayer J, Robert-Moreno A, Sharpe J, Swoger J. Attenuation artifacts in light sheet fluorescence microscopy corrected by OPTiSPIM. *Light Sci Appl*. 2018;7(1). doi:10.1038/s41377-018-0068-z

77. Finak G, Mcdavid A, Yajima M, et al. MAST : a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol*. 2015:1-13. doi:10.1186/s13059-015-0844-5

78. Salzer MC, Lafzi A, Berenguer-Llergo A, et al. Identity Noise and Adipogenic

Traits Characterize Dermal Fibroblast Aging. *Cell*. 2018;175(6). doi:10.1016/j.cell.2018.10.012

79. Mereu E, Lafzi A, Moutinho C, et al. Benchmarking Single-Cell RNA Sequencing Protocols for Cell Atlas Projects. *bioRxiv*. 2019:630087. doi:10.1101/630087

80. Lähnemann D, Köster J, Szczurek E, et al. Eleven grand challenges in single-cell data science. *Genome Biol*. 2020;21(1):31. doi:10.1186/s13059-020-1926-6

81. Kharchenko P V., Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. *Nat Methods*. 2014;11(7):740-742. doi:10.1038/nmeth.2967

82. Van den Berge K, Perraudeau F, Soneson C, et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol*. 2018;19(1):24. doi:10.1186/s13059-018-1406-4

83. Berge K Van den, Bézieux HR de, Street K, et al. Trajectory-based differential expression analysis for single-cell sequencing data. *bioRxiv*. May 2019:623397. doi:10.1101/623397

84. Campbell KR, Yau C. Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. *Nat Commun*. 2018;9(1):1-12. doi:10.1038/s41467-018-04696-6

85. Eling N, Richard AC, Richardson S, Marioni JC, Vallejos CA. Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. *Cell Syst*. 2018;7(3):284-294.e12. doi:10.1016/j.cels.2018.06.011

86. Korthauer KD, Chu L-F, Newton MA, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol*. 2016;17(1):222. doi:10.1186/s13059-016-1077-y

87. Cao Y, Wang X, Peng G. SCSA: a cell type annotation tool for single-cell RNA-seq data. *bioRxiv*. December 2019:2019.12.22.886481. doi:10.1101/2019.12.22.886481

88. Fincher CT, Wurtzel O, de Hoog T, Kravarik KM, Reddien PW. Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science (80-)*. 2018;360(6391). doi:10.1126/science.aaq1736

89. Plass M, Solana J, Alexander Wolf F, et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science (80-)*.

2018;360(6391). doi:10.1126/science.aaq1723

90. Schaum N, Karkanias J, Neff NF, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*. 2018;562(7727):367-372. doi:10.1038/s41586-018-0590-4

91. Han X, Wang R, Zhou Y, et al. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*. 2018;172(5):1091-1107.e17. doi:10.1016/j.cell.2018.02.001

92. Regev A, Teichmann S, Lander E, et al. The Human Cell Atlas. *bioRxiv*. May 2017:121202. doi:10.1101/121202

93. Mieth B, Hockley JRF, Görnitz N, et al. Using transfer learning from prior reference knowledge to improve the clustering of single-cell RNA-Seq data. *Sci Rep*. 2019;9(1):1-14. doi:10.1038/s41598-019-56911-z

94. Ianevski A, Giri AK, Aittokallio T. Fully-automated cell-type identification with specific markers extracted from single-cell transcriptomic data. doi:10.1101/812131

95. Hou R, Denisenko E, Forrest ARR. scMatch: a single-cell gene expression profile annotation tool using reference datasets. Kelso J, ed. *Bioinformatics*. 2019;35(22):4688-4695. doi:10.1093/bioinformatics/btz292

96. Stoeckius M, Zheng S, Houck-loomis B, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. 2018:1-12.

97. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16(12):1289-1296. doi:10.1038/s41592-019-0619-0

98. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol*. 2018;36(5):421-427. doi:10.1038/nbt.4091

99. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031

100. Depasquale EAK, Schnell DJ, Camp P Van, Grimes HL, Singh H, Salomonis N. DoubletDecon : Deconvoluting Doublets from Single- Resource DoubletDecon : Deconvoluting Doublets from Single-Cell RNA-Sequencing Data. 2019:1718-1727. doi:10.1016/j.celrep.2019.09.082

101. Mcginnis CS, Murrow LM, Gartner ZJ, Mcginnis CS, Murrow LM, Gartner ZJ. DoubletFinder : Doublet Detection in Single-Cell RNA Sequencing Data Using

Artificial Nearest Neighbors Brief Report DoubletFinder : Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst*. 2019;8(4):329-337.e4. doi:10.1016/j.cels.2019.03.003

102. Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. *Nature*. 2011;472(7341):90-95. doi:10.1038/nature09807

103. Xu X, Hou Y, Yin X, et al. Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor. *Cell*. 2012;148(5):886-895. doi:10.1016/j.cell.2012.02.025

104. Macaulay IC, Haerty W, Kumar P, et al. G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods*. 2015;12(6):519-522. doi:10.1038/nmeth.3370

105. Han A, Glanville J, Hansmann L, Davis MM. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat Biotechnol*. 2014;32(7):684-692. doi:10.1038/nbt.2938

106. Buenrostro JD, Wu B, Litzenburger UM, et al. of regulatory variation. 2015. doi:10.1038/nature14590

107. Rotem A, Ram O, Shoresh N, et al. Articles Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol*. 2015;33(11). doi:10.1038/nbt.3383

108. Hagemann-jensen M, Faridani OR. Small-seq for single-cell small-RNA sequencing. *Nat Protoc*. 2018;13(October). doi:10.1038/s41596-018-0049-y

109. Raj A, Bogaard P Van Den, Rifkin SA, Oudenaarden A Van, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. 2008;5(10):877-879. doi:10.1038/NMETH.1253

110. Moor AE, Itzkovitz S. ScienceDirect Spatial transcriptomics : paving the way for tissue-level systems biology. *Curr Opin Biotechnol*. 2017;46:126-133. doi:10.1016/j.copbio.2017.02.004

111. Lubeck E, Cai L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. 2012;9(7). doi:10.1038/nmeth.2069

112. Levsky JM, Shenoy SM, Pezo RC. Single-Cell Gene Expression Profiling. 2002;297(August):836-841.

113. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization MutationTaster2 : mutation prediction for the deep-sequencing age. *Nat Publ Gr*. 2014. doi:10.1038/nmeth.2892

114. Moffitt JR, Zhuang X. *RNA Imaging with Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH)*. Vol 572. 1st ed. Elsevier Inc.; 2016. doi:10.1016/bs.mie.2016.03.020

115. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. 2015;1363(2014):1360-1363. doi:10.1126/science.aaa6090

116. Codeluppi S, Borm LE, Zeisel A, et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods*. 2018;15(November). doi:10.1038/s41592-018-0175-z

117. Eng CHL, Lawson M, Zhu Q, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*. 2019;568(7751):235-239. doi:10.1038/s41586-019-1049-y

118. Ke R, Mignardi M, Pacureanu A, et al. In situ sequencing for RNA analysis in preserved tissue and cells. 2013;10(9). doi:10.1038/nmeth.2563

119. Lee JH, Daugharthy ER, Scheiman J, et al. Sequencing in Situ. 2014;(March):1360-1364.

120. Lee JH, Daugharthy ER, Scheiman J, et al. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. 2015. doi:10.1038/nprot.2014.191

121. Stahl PL, Salmen F, Vickovic S, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. 2016;353(6294).

122. Moor AE, Golan M, Massasa EE, et al. Global mRNA polarization regulates translation efficiency in the intestinal epithelium. *Science (80-)*. 2017;357(6357):1299-1303. doi:10.1126/science.aan2399