REVIEW

JBMR®

# The Unmixing Problem: A Guide to Applying Single-Cell RNA Sequencing to Bone

Matthew B Greenblatt,[1,2] Noriaki Ono,[3] Ugur M Ayturk,[4] Shawon Debnath,[1] and Sarfaraz Lalani[1]

[1]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA
[2]Research Division, Hospital for Special Surgery, New York, NY, USA
[3]University of Michigan School of Dentistry, Ann Arbor, MI, USA
[4]Musculoskeletal Integrity Program, Hospital for Special Surgery, New York, NY, USA

## ABSTRACT

Bone is composed of a complex mixture of many dynamic cell types. Flow cytometry and in vivo lineage tracing have offered early progress toward deconvoluting this heterogeneous mixture of cells into functionally well-defined populations suitable for further studies. Single-cell sequencing is poised as a key complementary technique to better understand the cellular basis of bone metabolism and development. However, single-cell sequencing approaches still have important limitations, including transcriptional effects of cell isolation and sparse sampling of the transcriptome, that must be considered during experimental design and analysis to harness the power of this approach. Accounting for these limitations requires a deep knowledge of the tissue under study. Therefore, with the emergence of accessible tools for conducting and analyzing single-cell RNA sequencing (scRNA-seq) experiments, bone biologists will be ideal leaders in the application of scRNA-seq to the skeleton. Here we provide an overview of the steps involved with a single-cell sequencing analysis of bone, focusing on practical considerations needed for a successful study. © 2019 American Society for Bone and Mineral Research.

**KEY WORDS:** SINGLE-CELL RNA SEQUENCING; MESENCHYMAL STEM CELLS; OSTEOBLASTS

## Introduction

How many discrete populations of mesenchymal cells exist in bone? What is the differentiation hierarchy among these populations, and is this linear or more complex and plastic? How do external stimuli shape the dynamics of these populations to impact bone formation? Understanding the cellular basis of bone formation requires clarity on each of these points and is therefore among the highest priorities for skeletal biology. However, for many years, progress on these issues has been hampered by limitations inherent in traditional methods to identify, isolate, or otherwise study skeletal cells, which produce highly heterogeneous pools containing many mesenchymal cell types. For example, traditional bone marrow stromal or calvarial osteoblast cultures are typically composed of an extremely heterogeneous mixture of cells.[1] This leads to an inability to assign phenotypes observed in vitro to discrete cell populations, and changes in cellular composition either between experimental and control groups or over the course of the culture experiment may confound experimental interpretation. Moreover, unappreciated differences in the cellular composition of heterogeneous skeletal mesenchymal cultures are likely to be a major contributor to problems with experimental reproducibility among labs.[2] Some of these issues potentially extend to in vivo studies using single markers to identify populations of interest, as many available markers capture not one but multiple cell populations.[3,4] In addition to these methodological factors that confound progress in understanding the cellular basis of bone formation, mesenchymal biology has inherent features that make resolving discrete populations of mesenchymal cells and determining their hierarchy challenging. Mesenchymal cells are notorious for displaying a high degree of plasticity and phenotypic instability in culture that complicate in vitro analysis of skeletal populations. Examples of this include the propensity of chondrocyte cultures to dedifferentiate in culture, the high degree of plasticity displayed by mesenchymal cells relative to other tissues, and the inability of some populations to survive in vitro in the absence of stimulation.[5–7]

To solve this issue of heterogeneity confounding our understanding of the cellular composition of bone, examination of other fields that have addressed similar questions can suggest successful strategies. In particular, immunology has identified an extensive range of discrete cell types and has assigned functions and molecular identities to each of these populations.[8,9] Perhaps the major factor facilitating this success has been an early adoption of single-cell analyses throughout the field, in this case predominantly flow cytometry, that allowed for identification and subsequent study

Received in original form February 27, 2019; revised form May 23, 2019; accepted May 25, 2019. Accepted manuscript online May 31, 2019.
Address correspondence to: Matthew B Greenblatt, MD, PhD, Department of Pathology and Laboratory Medicine, Weill Cornell Medicine LC929a, 1300 York Avenue, New York, NY, USA. E-mail: Mag3003@med.cornell.edu
Journal of Bone and Mineral Research, Vol. 34, No. 7, July 2019, pp. 1207–1219
DOI: 10.1002/jbmr.3802

of discrete cellular populations.[10] Thus, single-cell approaches, when used in concert with supporting in vivo and ex vivo functional studies, are likely to be key in facilitating skeletal biology to reach a similarly detailed understanding of the cellular compartment of our organ of interest. However, efforts to adapt these approaches to the particular challenges of bone, including issues related to cellular isolation of bone cells, will be needed.

Although flow cytometry remains an indispensable technique because of its ability to prospectively isolate defined cellular populations for further study, complementary approaches in the form of single-cell RNA sequencing (scRNA-seq) have flourished over approximately the past 5 years. Over this time, scRNA-seq has moved from being a technique restricted to a handful of technology-focused research groups to becoming a truly "ready for prime time" approach that is accessible to a wide range of investigators primarily focusing on biological questions and not methodology. The increasing availability of powerful single-cell technologies offers an unprecedented ability to deconvolute mixed populations of cells and identify new discrete cellular populations contributing to bone physiology. As an example of this promise, scRNA-seq studies of lung tissue have identified a novel cell type that is the major cell expressing the CFTR channel in airway epithelium, demonstrating the ability of scRNA-seq to provide substantial insights into the cellular basis of physiology and disease.[11,12] At the same time, these technologies still have important limitations as discussed below that must be taken into account. Here, we aim to provide a practical overview of application of this family of technologies to skeletal biology, including suggestions for investigators who are looking to add single-cell sequencing to their experimental toolbox.

## Planning a scRNA-seq Study of Bone

Although single-cell RNA-seq technologies are constantly evolving, two classes of approaches to capturing single cells have become available: 1) methods that rely on index sorting by FACS to achieve single-cell capture, and 2) techniques that utilize microfluidics-based capture of cells into droplets in an emulsion (Fig. 1, Table 1). The initially emerging index-based techniques (also known as Smart-seq) rely on sorting single cells into individual wells of 96- or 384-well plates and processing their transcriptomes into RNA-seq libraries while they are physically separated.[13,14] This method allows the sequencing of entire transcripts; however, the per cell cost of this technique tends to be considerably higher than that of microfluidics-based methods and scaling to large total numbers of captured cells can be limited by sorting rate. Other index sorting methods include MARS-seq and CEL-seq/CEL-seq2.[15–17] Index sorting methods are therefore advantageous in "deep-sequencing" experiments, wherein the aim is to characterize the individual transcriptomes of a small group of cells thoroughly, rather than interrogating the diversity of a highly heterogeneous pool of cells and identifying rare populations. Conversely, recently developed droplet/microfluidics-based methods, including Drop-seq, inDrop, and the commercial 10× Genomics platform, allow large numbers of cells (10,000 to 100,000) to be captured and processed in a rapid fashion.[18–20] Briefly, single cells are captured inside aqueous droplets emulsified in oil with beads ligated to cell-indexing primers. The cell membrane and nucleus are lysed to release the mRNA from each cell to mark each of its transcripts with a unique barcode (also known as unique molecular identifier, or UMI). These barcodes are later utilized in verifying the uniqueness of each mRNA molecule and thereby eliminate bias associated with repeat sampling of transcripts that are highly amplified during library preparation. An additional droplet-specific barcode separates the collective sequence output into cell-specific bins. Droplet-based methods have gained tremendous popularity since they were pioneered in 2015,[19,20] as they allow convenient processing of tens of thousands of cells and can therefore quantify the cellular heterogeneity of highly complex tissues, such as the retina or bone marrow, while facilitating the discovery of previously unrecognized cell populations. However, unlike the SMART-seq method, current droplet-based single-cell RNA-seq relies on 3′ biased sequencing of each mRNA molecule. Furthermore, while all current scRNA-seq methods are limited to largely capturing only a fraction of highly expressed mRNAs from each cell, Drop-seq captures fewer transcripts per cell than Smart-seq, which can exacerbate analytic challenges created by sparse transcriptome sampling. Detailed analysis of the relative strengths of specific methods is available in recent methodologic comparison studies.[18,21]

The per cell sequencing depth is an important experimental design parameter largely determined by the scRNA-seq method utilized. Earlier single-cell RNA-seq studies in the field of neuroscience suggest that 50,000 or even fewer reads per cell might be sufficient to define distinct cell populations.[22,23] The aforementioned sequencing depth might be expected to lead to the detection of 1000 to 3000 genes per cell population. However, the definition of this minimum threshold needed for robust separation of cellular populations is highly dependent on the complexity of the tissue at hand, as well as the objective of the experiment; much deeper sequencing may be necessary for data saturation to detect rare cell types. Additionally, it should be noted that estimates suggest that a mammalian cell expresses approximately 12,000 genes across 100,000 or more mRNA molecules, which suggests that most scRNA-seq methods are capable of sampling only a fraction of the total transcriptome, and this fraction is weighted toward genes with the highest expression. This poses a challenge for scRNA-seq to monitor transcriptional response or classify cell types in a sample, as many genes of interest, including cell type–defining transcription factors, show lower degrees of expression. The deeper transcriptome sampling offered by bulk RNA-sequencing on purified cell types can aid in overcoming this challenge when used in parallel. Bulk RNA sequencing can be particularly useful for discovery of genes of interest that can then be queried in scRNA-seq data sets. Thus, scRNA-seq and bulk RNA sequencing should be viewed as complementary rather than competing techniques.

In addition to choosing a suitable scRNA-seq method, another key step in planning a scRNA-seq experiment is to consider exactly which cells will be submitted for sequencing, which subset of cells within this population represent the population of interest, and how these cells will be isolated. Consideration of these points is needed to ensure that the population of interest will be sufficiently represented to allow for robust downstream analysis. Wherever feasible, performing scRNA-seq on cells isolated directly from bone without intervening culture is advised, as certain stem cell and progenitor populations may be lost upon culture, even when "basal" culture conditions are utilized.[3] An important decision when planning the cell isolation is whether cells will be subjected to FACS before

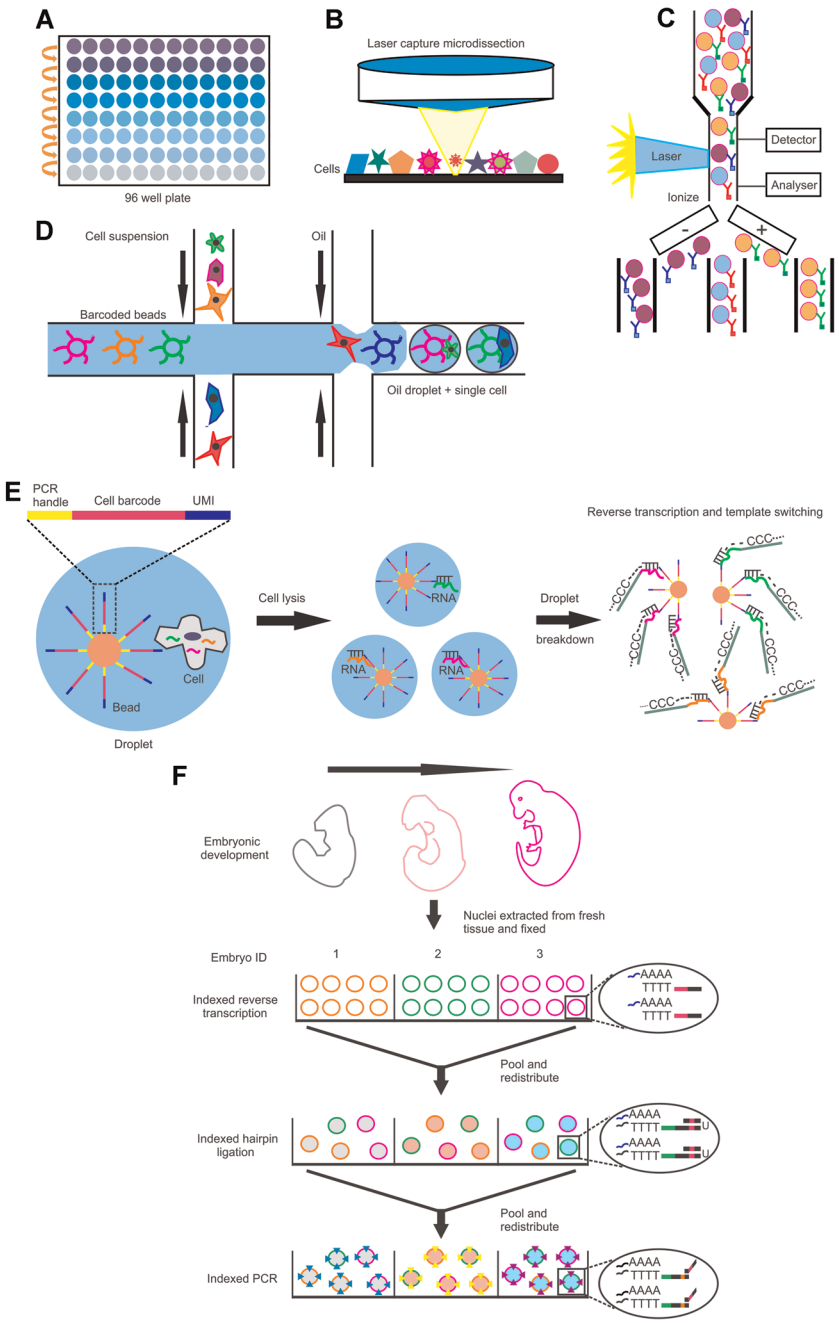


**Fig. 1.** Techniques for single-cell isolation and library generation. (*A*) Limiting dilution method to isolate single cells. (*B*) Laser capture microdissection (LCM) to isolate single cells from biological samples. (*C*) FACS-based isolation of specific cell types based on fluorescent marker proteins. (*D*) Microfluidic technology for capturing single cells as used in Drop-seq and other methods. Drop-seq allows transcriptional profiling of thousands of single cells by encapsulating cells in nanoliter droplets along with uniquely barcoding beads. It reveals transcriptionally distinct cell populations present in complex biological tissue creating a molecular atlas of gene expression. (*E*) A schematic representation of droplet-based library preparation. Individual cells are captured into droplets with microparticles that contain barcoded primers (beads). The primers on all beads contain a common sequence (PCR handle) for PCR amplification, cell barcode, and different unique molecular identifiers (UMIs) that allow mRNA transcripts to be digitally counted. Cells are lysed and the mRNAs are reverse transcribed into cDNAs, creating a set of beads called single-cell transcriptomes attached to microparticles (STAMPs) followed by cDNA amplification. Template switching is used to introduce a PCR handle downstream of the synthesized cDNA. (*F*) Schematic representation of single-cell combinatorial-indexing RNA-sequencing analysis 3 (sci-RNA-seq3). The technique involves a combinatorial indexing method that labels transcriptomes of single cell or nuclei. Nuclei from fresh tissue sample are extracted and fixed. A molecular index is applied to the mRNA from each cell followed by in situ reverse transcription incorporating a barcode bearing a polythymidine primer with a UMI. Cells are pooled and redistributed. For sci-RNA-seq3, hairpin ligation is performed for the third level of indexing. PCR primers target the barcoded polythymidine primer resulting in PCR amplicons to capture the 3′ ends of transcripts and these primers introduce a second barcode specific to each well of the PCR plate. Amplicons are pooled and sequenced, creating a 3′-tag digital gene expression profile. [Color figure can be viewed at wileyonlinelibrary.com]

**Table 1.** A Comparison of scRNA Sequencing Methods

| Methods | Region | UMI | System of isolation | cDNA amplification | Library construction | No. of genes detected/ cell | Cost ($)/cell |
|---|---|---|---|---|---|---|---|
| MARS seq (2014) | 3′ end | Yes, 8 bp UMI | FACS | IVT | RNA fragmentation; adaptor ligation | 4763 lowest sensitivity | ~1.3 |
| SCRB seq (2014) | 3′ end | Yes, 10 bp UMI | FACS | PCR | Tagmentation; 3′ enrichment | 7906 | ~2 |
| Smart-seq/C1 (2014) | Full length | None | Fluidigm C1 | PCR | Tagmentation | 7572 | ~25 |
| Smart-seq2 (2013, 2014) | Full length | None | FACS | PCR | Tagmentation | 9138 highest sensitivity | ~30 (commercial) |
| Drop seq (2015) | 3′ end | Yes, 8 bp UMI | Droplets | PCR | Tagmentation; 3′ enrichment | 4811 lowest sensitivity | ~0.1 |
| inDrop Seq (2015) | 3′ end | Yes, 6 bp UMI | Hydrogel-based droplets | IVT | RNA fragmentation; reverse transcription | — | ~0.06 |
| CEL-seq2/C1 (2016) | 3′ end | Yes, 5 bp UMI | Fluidigm C1 | IVT | RNA fragmentation; reverse transcription | 7536 | ~9 |

UMI = unique molecular identifier; FACS = fluorescence activated cell sorting; IVT = in vitro transcription; PCR = polymerase chain reaction.
Summary parameters for the listed scRNA-seq methods are provided. Date of publication is indicated. Cost per cell is in US dollars. Tagmentation is a library preparation reaction, in which a transposase cuts a double-stranded DNA and inserts the linker sequences required for sequencing.

single-cell sequencing or if cells will be directly utilized after enzymatic digestion. FACS has the advantage of clearing doublets, dead cells, and debris from the sample, and FACS allows for direct assessment of population frequencies present in the specimen, which can provide a key reference point to guide setting parameters during later analysis. FACS can also facilitate restricting the cells sequenced to only a small group of interest, thereby increasing the relative representation of these groups and avoiding expense associated with sequencing large numbers of cells not relevant to the study. This can be critical for bone biology, as, without additional enrichment steps, mesenchymal cells are often outnumbered by hematopoietic cells in most specimen types. Hematopoietic depletion strategies include negative selection on the pan-hematopoietic marker CD45 (gene symbol PTPRC), though due to the weak expression of CD45 on certain erythroid lineage cells additional negative selection with erythroid markers such as CD71 (transferrin receptor, gene symbol TFRC), glycophorin A (CD235a, gene symbol GYPA), or Ter119, a mouse-specific antibody clone recognizing Ly76, may be necessary for comprehensive removal of hematopoietic cells.[24] Endothelial cells can be depleted based on negative selection on the pan-endothelial marker CD31 (gene symbol PECAM1). Notably, capture of some number of unwanted cell populations is inevitable even with FACS and these cells must be accounted for during analysis.[3,25] Additionally, use of cell type–specific fluorescent reporters, such as a GFP variant driven by a reporter active in osteoblasts or cre-based lineage tracing methods, can allow for positive selection of populations of interest.

When an index sorting method is employed, the surface immunophenotype of each cell can be linked with that particular cell's transcriptome. This linking of surface immunophenotype to transcriptome is a key advantage, as it offers a solution to perhaps the biggest drawback of scRNA-seq studies: the inability to prospectively isolate populations of interest identified in these studies. Without prospective isolation of populations discovered with scRNA-seq, this approach is largely limited to being descriptive, as no functional studies can be performed. This limitation stems from the transcripts defining a population of interest rarely being either cell surface markers suitable for FACS or having an associated genetic reporter. Linked flow cytometry and transcriptome data can identify the surface immunophenotype corresponding to a cluster of interest, allowing for prospective isolation of this population in subsequent experiments. Notably, staining cells with antibodies containing nucleic acid barcodes as in the CITE-seq technique or the commercial version, TotalSeq, may allow for similar advantages as index sorting methods in terms of linking surface immunophenotype to cell clusters of interest.[26] Disadvantages of adding a pre-sequencing FACS step include the additional time and experimental complexity added. Additionally, FACS itself can have a negative effect on cell viability, though optimization of nozzle sizes and flow rates can minimize these effects; a larger nozzle size with a slower flow rate generally offers a better outcome.[3] Regardless of whether FACS is used, optimization of enzymatic digestion conditions for cell viability and yield is critical, as in our experience this represents the most common point of experimental failure in scRNA-seq studies. In keeping with this, comparison of in vivo transcriptional profiles on fixed cells to cells undergoing a typical isolation protocol in muscle suggests that enzymatic digestion is the major step that can disrupt the in vivo transcriptional profile.[27] This emphasizes the importance of minimizing the duration and harshness of enzymatic digestion, and we note that brief digestion protocols can provide robust yields of mesenchymal cells, particularly in younger mice.[3] Alternatively, there are several approaches designed to circumvent isolation-associated artifacts, including in vivo fixation before cell isolation, though fixation can negatively impact cell isolation efficiency and RNA quality.[27,28] In another approach, transgenic expression of Toxoplasma gondii uracil phosphoribosyltransferase (UPRT) only in cell types of interest allows for selective labeling, capture, and bulk sequencing of transcripts only from this cell type after whole-tissue RNA extraction.[28] However, this method requires a suitably specific promoter that allows targeting UPRT expression only to the cell type of interest, and experience to date with cre lines suggests that such a promoter may be elusive in the skeletal system.

Recent work on muscle comparing post-isolation to "in vivo" transcriptional profiles provides insight into the likely scope and degree of the impact of tissue digestion and isolation; these procedures induce an immediate early stress response and loss of quiescence that may occur predominantly in subpopulations of cells.[27–29] Particular care is warranted in assessing whether such isolation-associated activating transcriptional changes may either drive cell clustering or confound efforts to assess the transcriptional response to environmental or genetic perturbations.

In addition to the isolation procedure potentially influencing the transcriptome, some types of bone cells of high importance, such as osteocytes, can be challenging to dissociate into single-cell suspension, though there are examples of success.[30,31] Similarly, some large cell types, such as mature multinucleated osteoclasts may become physically disrupted or otherwise lost during FACS or microfluidics steps. Thus, all cellular isolation methods will necessarily introduce bias in the frequencies of cell types present relative to the in vivo tissue, with some cell types underrepresented or absent. These biases will largely be determined by the enzymatic digestion protocol employed. For these cell types with challenging isolation requirements, epigenetic readouts such as single-cell ATAC-seq may offer a more stable method to assess cell state in the face of harsh isolation procedures and can be multiplexed with scRNA-seq using recent methods, though it is noted that epigenetic features can also be impacted by cell isolation protocols.[27,32,33] The field of neuroscience in particular has had to deal with challenges in disassociating their tissue of interest into a single-cell suspension and has found success in instead performing sequencing of single nuclei through a family of microfluidics or flow cytometry–based approaches.[34–37] Because these nuclei are generally easier to isolate than intact cells and many of these methods can be used on fixed tissue, single nuclear sequencing methods may have utility in allowing robust analysis of particularly hard to isolate populations such as osteocytes. We note that recent advances employing a combinatorial barcoding labeling of nuclei have resulted in methods, such as Sci-RNA-Seq3, displaying very high throughput and favorable per cell sequencing costs.[38,39]

Before starting sequencing, a validation step is needed to ensure that single cells are being captured by the methodology employed. To clarify terminology relevant to this validation, most scRNA-seq methods have a proxy term to refer to individual cells in the analysis, as the occurrence of either cellular doublets or beads/wells containing only free-floating "background" RNAs and no cells break the expectation that each data point represents one cell. These cell equivalent terms include "STAMPs" (single-cell transcriptomes attached to microparticles) for microfluidics particle capture–based methods such as drop-seq or just "wells" for index sorting–based methods.[20] For all methods, validation of the doublet rate and cell capture rate is essential. For index sorting methods, this is fairly straightforward and consists of a test sort where each well is examined visually after sorting, primarily to assess the rate of empty wells, as the use of doublet gates makes issues with post-sort doublets rare. For microfluidics/droplet-based methods, this commonly takes the form of a species mixing experiment, where typically human and mouse cells are mixed together at different numbers and the rate of STAMPs containing transcripts from both species is assessed.[20] Higher-input numbers result in capture of larger numbers of STAMPs but result in higher doublet rates, and these two factors need to be balanced in pre-experimental validation.

Lastly, the limitations inherent in scRNA-seq approaches suggest that this technique is best reserved for questions specifically requiring this approach and that robust consideration of alternatives is warranted. For instance, examination of gene expression signatures or transcriptional responses to genetic or environmental perturbations in known populations of defined, isolatable cell types would be best accomplished by bulk RNA-sequencing rather than scRNA-seq. Additionally, similar to the gene expression atlases built with bulk RNA-sequencing, tissue atlases utilizing single-cell sequencing of large numbers of cells have recently become available and often include analysis of skeletal mesenchyme.[40–42] As these resources continue to expand, they may offer a route to answer selected questions utilizing existing data, especially for questions focusing on basal identity and gene expression in skeletal cells in the absence of specific stimuli.

## Analysis of scRNA-seq Data From Skeletal Specimens

Perhaps the step that requires the greatest effort in a scRNA-seq experiment is not cell isolation or sequencing but rather data analysis. Fortunately, scRNA-seq analysis approaches have been evolving at least as rapidly as the sequencing methods themselves, leading to a wide range of options, which notably include several very accessible tools that facilitate bone biologists with no prior computational training to conduct this analysis themselves[43–45] (Table 2). Regardless of the software used, the analysis process typically involves four key steps: approaches to account for technical artifacts/data cleaning, dimensionality reduction, clustering, and post-clustering examination of gene expression (Fig. 2). To summarize each of these in order, scRNA-seq is subject to characteristic confounding by covariates that must be addressed during the early stages of analysis. These include batch effects, the relative content of mitochondrial and ribosomal RNA, the total number of transcripts collected from each cell equivalent, or the cell cycle stage of each cell. Often, the effects of these covariates can be large relative to the biologic variation of interest, necessitating understanding and subsequently addressing their impact. Methods to address these include filtering out outlier cells, downsampling of populations with higher per cell transcriptional sampling than the rest of the specimen, or regression to remove the portion of the signal driven by these covariates. However, the impact of "regressing out" these covariates should be carefully and manually assessed in the final analysis, as some of these covariates may be unequally present in cell clusters of interest, leading to regression potentially masking the true biologic signal associated with these populations. It is worth emphasizing that this and nearly every other step of this analysis process will be ideally subject to iterative tweaking of analysis parameters and observing whether these tweaks help to recapitulate expected biology present in the sample. In this respect, scRNA-seq analysis is intensely informed by one's knowledge of the relevant underlying biology and is best conducted by investigators with a deep familiarity with this biology, though support of institutional cores and consulting bioinformaticians can be critical.

Next, most analysis platforms engage in some kind of dimensionality reduction and clustering. Dimensionality reduction often takes the form of principal component analysis

**Table 2.** Examples of scRNA-seq Analysis Pipelines

| Pipeline | Year | Programming language | Dimensionality reduction | Strategy |
|---|---|---|---|---|
| Monocle | 2014 | R | ICA, MST | Differential expression |
| SCUBA | 2014 | Matlab | *t*-SNE | Principle curve |
| Waterfall | 2015 | R | PCA, k-means, MST | Cell clustering |
| Wishbone | 2016 | Python | PCA, diffusion maps | Ensemble |
| TSCAN | 2016 | R | PCA | MST clusters |
| StemID | 2016 | R | PCA, ICA | Cell clustering |
| Slingshot | 2017 | R | Any | Cluster-based MST |
| scTDA | 2017 | Python | Any (MDS, ICA, *t*-SNE) | Topology-based differential expression |
| Velocyto | 2018 | R, Python | PCA | Cell clustering |
| Monocle 3 | 2019 | R | *t*-SNE or UMAP | Louvain clustering |

ICA = independent component analysis; MST = minimal spanning tree; *t*-SNE = *t* distributed stochastic neighbor embedding; PCA = principal component analysis; MDS = multidimensional scaling; UMAP = uniform manifold approximation and projection.

A number of analysis pipelines focus on inferring the differentiation trajectory of populations present in scRNA-seq data, including Monocle,[57] SCUBA,[95] Waterfall,[96] Wishbone,[97] TSCAN,[98] Slingshot,[99] scTDA,[100] and Monocle 3.[40] Velocyto focuses on inferring future gene expression profiles of each cell via analysis of unspliced transcripts.[56] StemID focuses on identification of rare outlier populations.[55]

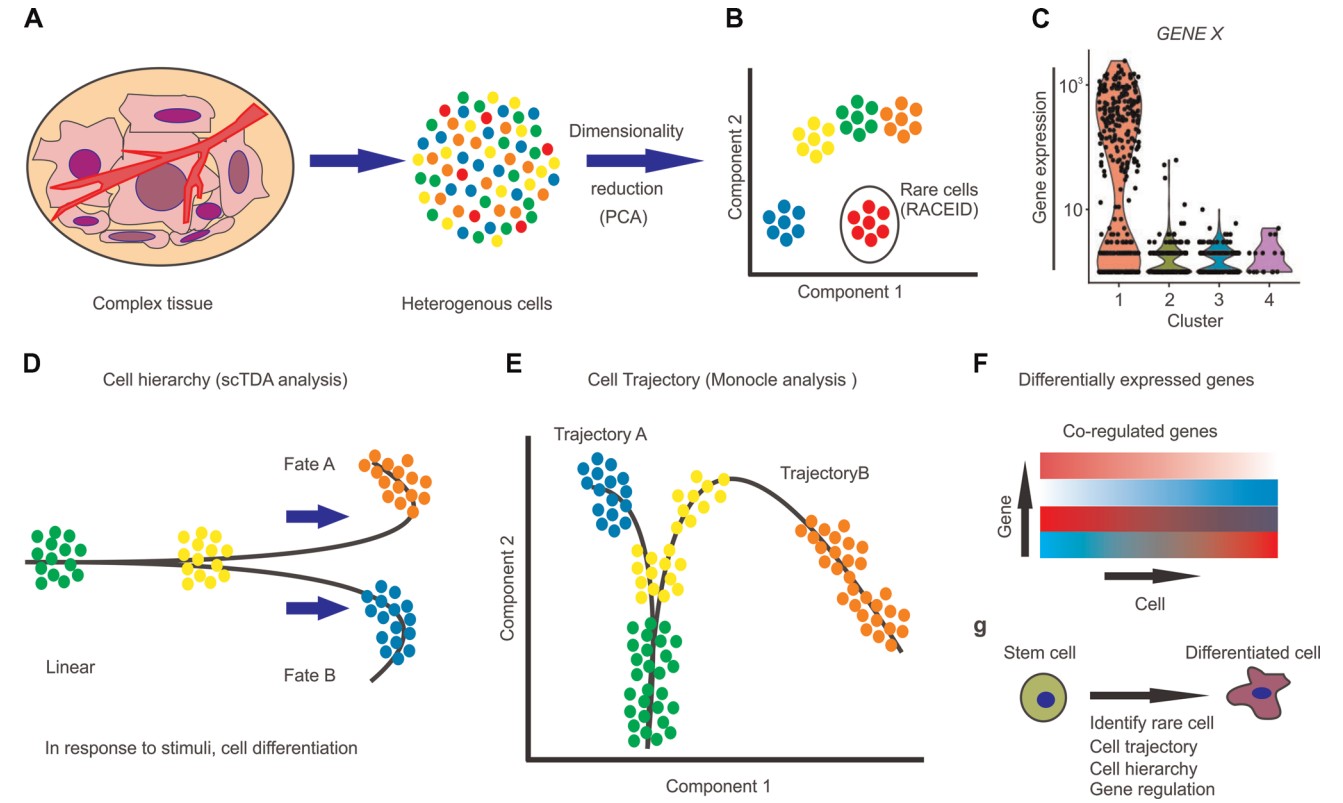


**Fig. 2.** Application of scRNA sequencing to decode biological complexity. (*A*, *B*) Single-cell analysis captures transcriptional profile of individual cells and can deconvolute populations present in suspension of mixed cell types. Principal component analysis (PCA) is a linear dimensionality reduction method and can be used to identify different cell clusters present in heterogeneous cell populations (*B*). *t*-SNE (*t*-distributed stochastic neighbor embedding) is a nonlinear dimensionality reduction method commonly used to display different cell clusters. (*C*) A violin plot is a density plot that can be used to determine the expression of a gene across different cell clusters. Dots represent individual cells. (*D*–*G*) Different types of scRNA analysis pipelines can infer cell commitment/hierarchy (*D*), cell trajectory (*E*), decode gene expression patterns (*F*), or stem cell differentiation (*G*). Dots represent the location of individual cells on the differentiation trajectory. scTDA (single-cell topological data analysis) is a topology-based computational algorithm that can be used to infer cell hierarchy and differentiation. Asynchronized cells represent different instantaneous time points along cell trajectories. scTDA resolves asynchrony and reconstructs a dynamic, continuous cell trajectory pathway (*D*). Monocle is an unsupervised algorithm that infers cellular differentiation trajectories occurring across the time surrogate pseudotime (*E*). [Color figure can be viewed at wileyonlinelibrary.com]

(PCA), which simplifies the complex variation present in the sample by identifying covariant transcripts and grouping these together in principal components (PCs). For instance, osteocalcin (*BGLAP*) and other transcripts highly expressed in osteoblasts, such as *COL1A1* and bone sialoprotein (*IBSP*) and others, may be grouped together, a set of genes called metagenes, into a principal component reflecting osteoblast identity. Often, it is instructive to manually examine the genes comprising each of the PCs to see what aspect of mesenchymal biology is being captured. Depending on the specimen preparation method, early PCs will likely be dominated by the signature of erythroid cells or leukocytes given their broad differences in gene expression in comparison to skeletal mesenchyme. Some PCs may largely correspond to the covariates discussed above, and visualizing these covariates across the top few principal components can be a helpful method to understand their impact on downstream analyses. For instance, cell cycle often drives one or more of the early PCs, and observing disappearance of this PC can be helpful in ensuring that regression or other approaches have accounted for cell cycle effects. After PC generation, users will commonly select which of these PCs to use to cluster the data using *k* means clustering or another method. Computational methods such as a Jackstraw plot can help illustrate how likely each principal component is likely to have been observed by chance and thereby help guide selection of which PCs can aid in guiding a biologically meaningful clustering of the data. However, perhaps the most useful method is to iteratively conduct the analysis with different numbers of PCs and empirically observe how these choices impact populations of interest, using populations that correspond to osteoblasts, chondrocytes, or other clearly delineated mesenchymal populations as "landmarks" to aid in evaluating how expected populations segregate as an internal control for the correctness of the clustering.

The last step in scRNA-seq analysis is to display the clusters and understand which cellular populations are represented by analyzing both the genes defining each cluster and also the expression of genes of interest that classically define known populations, such as osteocalcin transcripts defining mature osteoblasts. Clusters are typically represented using *t* distributed stochastic neighbor embedding (*t*-SNE), a dimensionality reduction data visualization algorithm,[46] or more recently, uniform manifold approximation and projection (UMAP).[47] Notably, *t*-SNE employs several user-defined parameters that can have dramatic effects on the end output, meaning that caution is required to avoid overinterpretation of features such as cluster size or distance that may reflect these user-defined parameters more than the underlying data (https://distill.pub/2016/misread-tsne/).[48]

With the ability to characterize the transcriptomes of individual cells, one intriguing question is whether gene expression changes in distinct cell populations can be detected after pharmacologic, genetic, or environmental perturbations on the skeleton. Although this is an exciting possibility, there are multiple important challenges to consider before attempting comparative gene expression profiling experiments with scRNA-seq. First, currently available methodologies capture only a small percentage of the transcripts (roughly 5% to 15%) present in each cell. As a result of this sparse and stochastic sampling, gene expression data may be difficult to interpret for genes with mid to low expression levels, as many of these genes may show apparent "dropout" of expression of these transcripts within each cluster due to that transcript not being sampled in that particular cell. As a result, the apparent absence of a gene of interest in a cluster must be interpreted with caution, as it may simply represent that the transcript is expressed at a level below the high threshold needed for detection. These issues can be further complicated if differences in transcript sampling among cellular populations lead to different detection thresholds in each cluster. Computational strategies to address this issue include MAGIC (Markov affinity-based graph imputation of cells), which infers values for gene expression data missing due to sampling issues in each cell based on gene expression in similar cells.[49] Alternatively, where feasible, cellular isolation followed by bulk RNA-sequencing offers perhaps the most straightforward method to experimentally validate gene expression changes observed by scRNA-seq. Second, an equally important consideration is to ensure proper definition of distinct clusters that accurately represent the cellular diversity of skeletal tissues at hand: As they descend from similar lineages and exhibit functional similarities, distinct mesenchymal cell populations co-express several genes at high levels, and their transcriptomes in a single-cell RNA-seq data set can resemble each other, leading to coclustering of very distinct cell populations. Therefore, a thorough evaluation of each cell cluster in order to exclude methodological artifacts is essential during data analysis. Recently developed feature-barcoding techniques such as CITE-seq and TotalSeq show little dropout and can help overcome this issue and verify cell identity through correlation of membrane-bound protein markers and transcriptional output.[26] A third and perhaps more obvious challenge is to ensure that the transcriptomes of cells are not significantly altered by the cell isolation process. Although there have been concerns that FACS can perturb gene expression, published validation studies in non-bone tissues show minimal effects on gene expression with optimized protocols.[50–52] As also discussed above, cell isolation–induced biases or artifacts can be particularly difficult to exclude when the goal of the experiment is to characterize the effects of environmental changes (such as dietary intake or mechanical loading) in the absence of an internal control (such as a genetic mutation blocking this response).

Despite the potential complexity of the scRNA-seq analysis pipeline, an increasing number of software tools are available, and several of these are designed to be accessible to investigators with no prior computational biology training. Notably, Seurat has online tutorials designed to get new users started with scRNA-seq analysis (https://satijalab.org/seurat/) and has several tools to help with regression or filtering-based approaches to account for covariates.[53] In addition to the basic analysis pipeline described here, a number of analytic tools have been designed to focus on answering specialized questions (Table 2). One of these, RaceID, focuses on identifying outlier cells relative to each of the clusters and thereby attempts to identify rare, sparsely sampled populations that may be of biologic interest.[54] Combination of RaceID approaches with identification of computational features of stemness, including high transcriptional entropy and interconnectedness of the population in an inferred differentiation trajectory, has been used for de novo computational identification of stem cell populations.[55]

Another set of analysis tools focuses on inferring the relationships among the populations defined during the clustering step, often by defining a series of edges or lines

that connect these populations into a tree or trajectory through additional dimensionality reduction. These connections are typically inferred on the principle that changes in gene expression as cells differentiate tend to be parsimonious, involving minimal changes during each differentiation event. For example, a series of cells differentiating along an osteoblast differentiation pathway are likely to retain many elements of the transcriptional character of osteoblasts during this process and therefore be more transcriptionally similar to each other than to unrelated mesenchymal lineages. In a common form of this analysis, construction of a minimum spanning tree, algorithms seek to connect all of the cell clusters with a "tree" that minimizes total sum of the "distances" of these connections across a space representing gene expression. Notably these kinds analyses makes the assumption that all of the cell types present in the sample share a lineage relationship, and for some types of specimens such as those including both endosteal and periosteal mesenchymal cells, this assumption may be false.[3] Thus, these approaches are greatly enhanced when used in conjunction with positive selection for a genetically encoded lineage tracing marker to provide assurance that the cells under analysis do share a lineage relationship. In an alternative method to infer cellular differentiation trajectories, a recent approach measures RNA velocity, or the rate of change in the expression of a gene through the ratio of unspliced to mature transcripts.[56] This can in turn be used to infer the future expression profile of cells and predict impending transitions among cell types. One of the most widely used tools for this kind of analysis is Monocle.[40,57,58] After dimensionality reduction and clustering, Monocle performs minimum spanning tree analysis to connect each cell cluster, finds the longest path along this tree, and then orders these clusters according to an inferred timeline of differentiation. Because this timeline does not refer to actual measured time, it is instead termed "pseudotime." Proof of concept of this approach includes demonstrating that Monocle 2 can reconstruct known hematopoietic lineage trees from single-cell data. Notably, Monocle is able to accept sequential data drawn from multiple time points, making it particularly suitable for reconstructing in vitro cellular differentiation pathways from multiple sampled cultures of asynchronously differentiating cells or an analysis of the differentiation of cells in a fracture callus over time.

In scRNA-seq studies, some tissue types appear to show robust separation by clustering, such as different lineages of immune cells, while other tissue types display less robust separation due to broadly shared gene expression programs, intermediate cell states, or other causes. Studies to date suggest that skeletal mesenchyme may fall more in the latter than the former category, so tools that focus on resolving closely related populations may be useful in skeletal studies.[3] One clustering algorithm, biSNE (biclustering on stochastic neighbor embedding), aims to enforce a more robust separation of populations and thereby delineate between distinct but related cell populations and may thereby be useful for separating distinct mesenchymal subpopulations.[35]

## Validating scRNA-seq Results

As the technical and analytic issues described above can lead to the identification of spurious cellular populations, validation of populations identified by scRNA-seq with a complementary method should be considered a key component of any complete scRNA-seq study. When the genes defining the cluster of interest include cell-surface markers, flow cytometry offers a straightforward validation path. However, clusters may lack defining cell surface markers, and even when putative cluster-defining cell surface markers are identified, suitable antibody reagents may not be available. Furthermore, the overall weak correlation between mRNA and protein abundance for many genes may frequently prevent this approach, resulting in distinct sets of genes serving as the most robust markers of a given cell type when using RNA versus protein-based detection methods.[59,60] An improved interpretation of scRNA-seq results in the context of skeletal biology can be achieved by revealing the spatial identity of the identified cell populations. This is typically done by testing identified cell type–specific markers on histological sections. This in situ corroboration of flow cytometry–based transcriptional profiling is challenging because of its high technical sensitivity, and discrepancies between scRNA-seq and histological analyses are often encountered. It is important to note that some RNAs and proteins can significantly lose their integrity and antigenicity, respectively, during routine histological procedures. A protocol maintaining tissue samples as much in native conditions as possible would be ideal. However, it is practically impossible because of the inherent structural hardness of bone tissues requiring complex tissue preparations, such as extended fixation and decalcification. Where relevant, analysis of minimally fixed and decalcified tissue, such as embryonic or neonatal bones using frozen sections, can help minimize the impact of these issues.

Major approaches validating expression of identified markers in an effort to confirm the presence of populations identified by scRNA-seq approaches include immunohistochemistry (IHC), in situ hybridization (ISH), and use of genetically engineered reporter lines, particularly in knock-in reporter mice if available. The success of IHC-based validation entirely relies on the quality of antibodies available, and antibodies that work in other tissues sometimes do not work well on bone sections. Additional steps are often required, including antigen retrieval and signal amplification, depending on how tissue samples have been prepared. Moreover, genes encoding proteins released into the milieu or the circulation, such as cytokines and hormones, may have staining patterns irrelevant to their cellular origin. Considering all these variables, ISH is often a more straightforward and indeed preferable method to validate expression of marker genes identified by scRNA-seq analyses. Historically, ISH was a technique most widely used for embryology. However, with the advent of a high-sensitivity ISH approach such as RNAscope technology (Advanced Cell Diagnostics [ACDBio], Newark, CA, USA), its application has been significantly expanded. This technology utilizes double Z probes (18~25 bp each, designed up to 20 probes) that increases the specificity and sensitivity of the hybridization, followed by explosive amplification of the signal. Probes for the vast majority of genes are readily available from the supplier. Although the applicability of ISH has been substantially expanded, detecting genes that are expressed only at a low level can still be challenging. It can be particularly the case for adult bones, due in part to the need for deep decalcification. The third option, use of transgenic reporter lines widely available in mice, is indeed a reliable and reproducible way to validate expression of identified marker genes in situ. Fluorescent proteins that are stable during complex tissue

preparations, including eGFP, eYFP, tdTomato, and mCherry, are used to visualize cells of interest by recapitulating endogenous gene expression. Typically, a cassette encoding a fluorescent protein is inserted into the endogenous locus or the transgene so that its expression is regulated by the promoter and enhancer of the gene of interest. This transgenic reporter-based approach is highly versatile and facilitates downstream analyses because tagged cells can be readily isolated as live cells on flow cytometry and cell sorting. A large collection of transgenic reporter mouse lines are available from the public repositories, including the Jackson Laboratory (Bar Harbor, ME, USA; www.jax.org), MMRRC (www.mmrrc.org), and the GENSAT (www.gensat.org). A combination of these approaches should be utilized to validate and reveal the anatomic distribution of cell populations identified by scRNA-seq analyses.

## Spatially Annotated scRNA-seq Approaches

For bone biologists, the specific location of a given cell within a complex microenvironment provides important clues as to that cell's identity and function. Spatial information also has an advantage over transcriptional data in that it is not subject to constant fluctuation due to cellular plasticity or phenomena such as transcriptional bursting.[61–63] Two cells with an identical transcriptome may perform discrete functions, depending on their neighboring cells or matrices in which they are embedded. This critical piece of information is permanently lost upon cell dissociation, an inevitable step to prepare cells for above discussed scRNA-seq procedures. For most investigators, the most straightforward method to annotate the anatomic location of cellular clusters emerging from scRNA-seq studies is to manually localize the expression of cluster-defining genes using immunohistochemistry, immunofluorescence, or in situ hybridization as discussed above. Methods have been reported to aid in the determination of the minimal gene set needed to spatially resolve a given set of cell clusters.[64] However, this manual approach to spatial annotation is dependent on the existence of suitable staining reagents and can be infeasible to scale when large numbers of target genes need to be stained to resolve the clusters detected. Although methods have been reported to allow for arbitrary scaling of RNA hybridization or in situ RNA sequencing-based transcript detection,[65–67] another approach is to perform transcriptional profiling directly on histological sections. This approach has been particularly developed for neuroscience research, in which cell dissociation–based transcriptional profiling is impractical. Here, we briefly mention two particular methods introduced recently, spatial transcriptomics[68] and STARmap (spatially resolved amplicon mapping).[69] The former method, spatial transcriptomics, is the first approach to spatially resolve RNA-seq data in individual tissue sections. In this method, spatially barcoded oligo(dT) primers are attached to the surface of microscope slides, enabling a genomewide analysis. The resolution of the original spatial transcriptomics approaches were 30 μm; however, a more recently developed Slide-seq approach offers a 10 μm resolution, therefore making it feasible to capture single cells in brain tissue.[70] The latter method, STARmap, utilized an improved FISH (fluorescent in situ hybridization) approach using SNAIL (specific amplification of nucleic acid via intramolecular ligation) probes and hydrogels, which can map somewhere between 160 and 1020 genes. Another technique, MERFISH (multiplexed error-robust FISH) provides a similar ability to spatially measure gene expression for approximately 100 to 1000 genes through the use multiplexed FISH probes whose barcodes are read out over successive rounds of hybridization.[66] Although this number of genes is sufficient to discover new clusters, this approach can potentially hamper the discovery of new genes regulating an important biological process, suggesting a role for complementary approaches providing deep transcriptome coverage, such as bulk RNA sequencing. Another important limitation is that the resolution of current spatially resolved approaches is approximately 10 to 30 μm; thus, it may capture groups of adjoining cells, rather than single cells, especially in regions of high cellular density such as in bone marrow or periosteum. A scRNA-seq approach fully integrating spatial information of single cells could be particularly attractive to correlate with strain maps or other spatially resolved mechanical parameters during loading to generate an advanced understanding of the transcriptional response to bone loading. A challenge in applying these methodologies to bone is that many skeletal cells are extremely compact and arranged in a highly intricate manner. For example, in marrow space, mesenchymal cells are intertwined with hematopoietic cells and endothelial cells. Therefore, applicability of such spatial transcriptional profiling approach will need to be determined on a case-by-case basis. Lastly, all of these approaches for spatial annotation require the production of high-quality tissue sections, and recently popularized methods for cutting and transfer of unfixed, nondecalcified sections of adult bone can help overcome hurdles in specimen preparation.[71]

## Sample Insights From scRNA-seq Studies in Bone

While most of the impact of scRNA-seq studies on bone biology will doubtless come from future studies, early examples of scRNA-seq studies in bone provide proof of the utility of this approach. Chan and colleagues published key papers describing skeletal stem cells in young bones, first in mice termed mouse skeletal stem cells (mSSCs)[72] and, more recently, in humans termed human skeletal stem cells (hSSCs).[24] In the first study, they identified nonhematopoietic/endothelial AlphaV(CD51)$^{+}$Thy1(CD90)$^{-}$CD105$^{-}$CD200$^{+}$ cells isolated from the perinatal growth plate as self-renewing multipotent skeletal stem cell populations, using extensive in vitro and transplantation assays. They also showed clonal cell populations within the growth plate using multicolor lineage-tracing experiments. In this study, single-cell sequencing was used to characterize small numbers of these stem cells and their derivative populations. In a second study, they conducted a scRNA-seq analysis of the microdissected human fetal growth plate and found that cells in the late prehypertrophic zone and the hypertrophic zone express human orthologs of mSSC-specific genes, therefore suggesting that hSSCs reside in these layers of the growth plate.

There is a line of evidence that prehypertrophic and hypertrophic chondrocytes represent transient cell types that are destined to undergo apoptosis or transdifferentiate into osteoblasts. In contrast, the resting zone of the growth plate has been shown to contain stemlike cells, originally in rabbits based on transplantation studies.[73] More recently, the existence of skeletal stem cells within the resting zone has been demonstrated based on more definitive lineage-tracing experiments in mice.[6] Cells in other layers of the growth plate, such as

proliferating, prehypertrophic, and hypertrophic layers, do not self-renew and rapidly disappear from the growth plate.[6] Potential fates of these non-resting cells include apoptosis in the hypertrophic layer and transdifferentiation into osteoblasts[74,75] or bone marrow stromal cells,[76] as indicated by a series of lineage-tracing experiments in mice. In light of this literature, more sophisticated approaches that can analyze and purify these SSCs in their native environment are desirable. A recent scRNA-seq analysis of mouse neonatal growth plate discovered a novel population of chondrocytes corresponding to borderline chondrocytes, which was previously described by histological analysis.[77] A follow-up lineage-tracing experiment further demonstrated that these chondrocytes behave as transient mesenchymal precursor cells. Thus, multiple groups have used scRNA-seq and other complementary techniques to uncover cellular heterogeneity and discrete functionality of distinct growth plate chondrocyte subpopulations.

Debnath and colleagues recently reported a novel population of Cathepsin K (CTSK)-labeled periosteal stem cells (PSCs). Unlike the mesenchymal stem cells present in the endosteal compartment, these PSCs do not express markers associated with mesenchymal cells capable of supporting hematopoiesis, such as LEPR or CD146, and can only mediate intramembranous ossification at baseline.[78,79] Because the identification of these cells was made through hypothesis-driven FACS gating and transplantation experiments, single-cell analysis was utilized as a parallel method to identify CTSK-labeled stem cell populations to see if this parallel approach converged on nominating the same population as stem cells. CEL-SEQ2 was performed on mesenchymal CTSK+ periosteal cells, which allowed for collecting the full surface immunophenotype for each cell during index sorting and subsequently linking this surface immunophenotype with that same cell's RNA expression data. Analysis of CTSK-cre-labeled periosteal cells showed clustering into four groups: group 1 was defined by expression of *Sox9* and *Col2a1*; group 2 expressed osteoblast markers such as *Bglap* and *Alpl*; group 3 expressed *Ly6a* (Sca1); and a small group 4 was characterized by high expression of alpha-smooth muscle actin (*Acta2*). Almost all of the cells identified as periosteal stem cells on the basis of surface immunophenotype fell into this group 1 expressing *Sox9* and *Col2a1*. Why would an intramembranous-specialized stem cell express transcripts classically associated with chondrocytes?[80] Previously, pulse-chase lineage tracing studies have identified both Sox9 and Col2a1 was labeling long-lived osteoblast progenitors in both the endosteal and periosteal compartments.[76] Yamashiro and colleagues also reported a set of periosteal cells in both long bone and calvarium that display robust expression of Sox9.[81] Similarly, ablation of the ability of Sox9-expressing cells to give rise to osteoblasts via deletion of osterix with Sox9-cre resulted in severe impairments in both intramembranous and endochondral bone formation.[82,83] Thus, although Sox9 and Col2a1 are chondrocyte markers, they additionally serve as markers of mesenchymal stem cells. In particular, their expression in both intramembranous specialized periosteal stem cells and their endochondral-specialized endosteal counterparts suggests that Sox9 and Col2a1 are core components of the transcriptional signature shared by multiple populations of skeletal stem cells.

Group 2 within the pool of periosteal CTSK-labeled cells was defined by high expression of osteoblast markers such as *Bglap* and *Alpl*, offering support for parallel immunohistochemical and transplantation studies identifying that PSCs give rise to osteoblasts. Further analysis of gene expression in this cluster also identifies robust *Ifitm5* expression, suggesting that *Ifitm5* may have general utility in identifying osteoblasts in scRNA-seq studies.[84–86] Regarding the cluster of CTSK-positive cells expressing *Acta2*, it has been previously reported that *Acta2* is a marker of pericytes and myofibroblasts that display osteogenic capacity.[87,88] During fracture, periosteal mesenchymal cells labeled with an inducible Acta2-cre undergo expansion and can differentiate into osteogenic and chondrogenic lineages.[88,89] Within the pool of CTSK cre-labeled periosteal mesenchyme, *Acta2* + cells are distinct from FACS-defined PSCs, and PSCs sit at the apex of their differentiation hierarchy in heterotopic transplantation studies.[3] Similarly, this population of *Acta2* + cells are distinct from the group 2 cells expressing osteoblast markers. Taken together with prior lineage tracing studies of *Acta2* + cells, this suggests a model whereby the CTSK-lineage subset of *Acta2* + cells are an intermediate progenitor linking PSCs to mature osteoblasts, and the transition of *Acta2* + cells to osteoblasts may be dynamically regulated in response to injury. However, further direct transplantation studies will be needed to test this model and establish the hierarchy of PSCs relative to *Acta2* + cells.

This CEL-SEQ2 study of CTSK-labeled periosteal mesenchymal cells was further analyzed using Monocle to observe if the cellular differentiation hierarchy computationally inferred from cells directly isolated from the native bone environment was consistent with the differentiation hierarchy experimentally determined by heterotopic transplantation studies.[3] Unsupervised analysis placed a vast majority of the CTSK-lineage cells along an unbranched linear differentiation trajectory. Consistent with PSCs giving rise to THY1 + cells after transplantation, FACS-identified PSCs were present at the root of this trajectory and the cells expressing later mesenchymal markers such as THY1 and SCA1 were present at the end of the trajectory. Thus, scRNA-seq studies can provide insights into differentiation hierarchy in minimally manipulated native systems that complement the limitations of heterotopic transplantation studies of defined populations.

Consistent with the above, analysis of genes that were differentially expressed during this inferred differentiation trajectory demonstrated the early expression and subsequent downregulation of markers associated with early mesenchymal progenitors including *Col2a1* and *Sox9*. Genes showing low early expression and subsequent upregulation included the later-stage mesenchymal markers such as *Thy1*, *Postn*, CD34, and *Ly6a* encoding SCA1. Interestingly, *Bmp2* was noted very early in the differentiation trajectory and was also detected in bulk sequencing studies of PSCs. To put this into context, during earliest stages of skeletal development, BMP signaling is necessary to initiate Sox9 expression and chondrogenesis in early limb bud mesenchymal condensations, suggesting that BMP2 is a key inducer of skeletal stem cells.[90] Consistent with this, BMP2 is able to expand skeletal stem cells in vitro and induce skeletal stem cells de novo in soft tissues,[72] and BMP2 is necessary for bone formation and to initiate fracture healing.[91–93] Taken together with the expression of BMP2 directly in skeletal stem cell populations in this data set, this suggests that BMP2 may be involved in an autocrine loop to maintain skeletal stem cell pools and that external signals tuning BMP2 expression within skeletal stem cells may be critical determinants of the size of this stem cell pool. This model is consistent with a recent study finding that periosteal BMP2 controls functions specific to periosteal physiology, such as the radial expansion of bone.[94] We speculate that this putative autocrine model for stem cell self-regulation may be advantageous

in a system such as the skeleton that maintains separate pools of stem cells in distinct anatomic compartments,[3] as it could facilitate separate regulation of the sizes of each of these distinct stem cell pools in a manner not possible with an externally expressed systemic signal. The finding that endosteal bone formation is relatively preserved with deletion of BMP2 despite substantial periosteal defects suggests that these autocrine factors regulating stem cell pool size or function are likely to be "compartmentalized" with effects limited to specific subsets of skeletal stem cells and, accordingly, specific anatomic regions within bone.

## Conclusions

Perhaps the key challenge in advancing our understanding of the cellular basis of bone metabolism is "unmixing" the many heterogeneous cell types present in bone to resolve discrete homogenous populations. scRNA-seq shows promise as an important part of the experimental toolbox needed to address this issue. However, tapping into this potential requires extensive pre-experimental planning to select specimens, cell isolation methods, scRNA-seq methods, and analysis tools tailored to the experimental question. Even after completing a scRNA-seq study, post-sequencing validation approaches are needed to exclude that any key cell populations identified represent analytic artifacts. Because each step along this path is heavily informed by knowledge of the underlying biology, bone biologists are best positioned to lead these advances, and indeed a proliferation of accessible sequencing and analysis tools make completion of scRNA-seq studies accessible even for groups with no specialized experience.

## Disclosures

All authors state that they have no conflicts of interest.

## Acknowledgments

MBG holds award DP5OD021351 from Office of the Director of the NIH, a Career Award for Medical Scientists from the Burroughs Welcome Foundation, a Basil O'Connor Award from the March of Dimes, and a Pershing Square Sohn Prize for Young Investigators in Cancer Research. NO holds award R01DE026666 from NIH/NIDCR. This content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

Authors' roles: MBG, UMA, NO, SD, and SL contributed to authoring this manuscript. SD and SL created the tables.

## References

1. Chang MK, Raggatt L-J, Alexander KA, et al. Osteal tissue macrophages are intercalated throughout human and mouse bone lining tissues and regulate osteoblast function in vitro and in vivo. J Immunol. 2008;181(2):1232–44.
2. Manolagas SC, Kronenberg HM. Reproducibility of results in preclinical studies: a perspective from the bone field. J Bone Miner Res. 2014;29(10):2131–40.
3. Debnath S, Yallowitz AR, McCormick J, et al. Discovery of a periosteal stem cell mediating intramembranous bone formation. Nature. 2018;562(7725):133–9.
4. Zhang J, Link DC. Targeting of mesenchymal stromal cells by cre-recombinase transgenes commonly used to target osteoblast lineage cells. J Bone Miner Res. 2016;31(11):2001–7.
5. Zelzer E, McLean W, Ng Y-S, et al. Skeletal defects in VEGF(120/120) mice reveal multiple roles for VEGF in skeletogenesis. Development. 2002;129(8):1893–904.
6. Mizuhashi K, Ono W, Matsushita Y, et al. Resting zone of the growth plate houses a unique class of skeletal stem cells. Nature. 2018;563(7730):254–8.
7. Omatsu Y, Sugiyama T, Kohara H, et al. The essential functions of adipo-osteogenic progenitors as the hematopoietic stem and progenitor cell niche. Immunity. 2010;33(3):387–99.
8. Cooper MD. Exploring lymphocyte differentiation pathways. Immunol Rev. 2002;185:175–85.
9. Good RA. Cellular immunology in a historical perspective. Immunol Rev. 2002;185:136–58.
10. Herzenberg LA, Parks D, Sahaf B, Perez O, Roederer M, Herzenberg LA. The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. Clin Chem. 2002;48(10):1819–27.
11. Plasschaert LW, Žilionis R, Choo-Wing R, et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. Nature. 2018;560(7718):377–81.
12. Montoro DT, Haber AL, Biton M, et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. Nature. 2018;560(7718):319–24.
13. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9(1):171–81.
14. Ramsköld D, Luo S, Wang Y-C, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30(8):777–82.
15. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012;2(3):666–73.
16. Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014;343(6172):776–9.
17. Hashimshony T, Senderovich N, Avital G, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 2016;17:77.
18. Zhang X, Li T, Liu F, et al. Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. Mol Cell. 2019;73(1):130–142.e5.
19. Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–201.
20. Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161(5):1202–14.
21. Ziegenhain C, Vieth B, Parekh S, et al. Comparative analysis of single-cell RNA sequencing methods. Mol Cell. 2017;65(4):631–43.e4.
22. Shekhar K, Lapan SW, Whitney IE, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell. 2016;166(5):1308–23.e30.
23. Zheng GXY, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
24. Chan CKF, Gulati GS, Sinha R, et al. Identification of the human skeletal stem cell. Cell. 2018;175(1):43–56.e21.
25. Takahashi A, Nagata M, Gupta A, et al. Autocrine regulation of mesenchymal progenitor cell fates orchestrates tooth eruption. Proc Natl Acad Sci U S A. 2019;116(2):575–80.
26. Stoeckius M, Hafemeister C, Stephenson W, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14(9):865–8.

27. Machado L, Esteves de Lima J, Fabre O, et al. In situ fixation redefines quiescence and early activation of skeletal muscle stem cells. Cell Rep. 2017;21(7):1982–93.
28. van Velthoven CTJ, de Morree A, Egner IM, Brett JO, Rando TA. Transcriptional profiling of quiescent muscle stem cells in vivo. Cell Rep. 2017;21(7):1994–2004.
29. van den Brink SC, Sage F, Vértesy Á, et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat Methods. 2017;14(10):935–6.
30. Tanaka K-I, Xue Y, Nguyen-Yamamoto L, et al. FAM210A is a novel determinant of bone and muscle structure and strength. Proc Natl Acad Sci U S A. 2018;115(16):E3759–68.
31. Shah KM, Stern MM, Stern AR, Pathak JL, Bravenboer N, Bakker AD. Osteocyte isolation and culture methods. Bonekey Rep. 2016; 5:838.
32. Buenrostro JD, Wu B, Litzenburger UM, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523(7561):486–90.
33. Cao J, Cusanovich DA, Ramani V, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018;361(6409):1380–5.
34. Lake BB, Ai R, Kaeser GE, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science. 2016;352(6293):1586–90.
35. Habib N, Li Y, Heidenreich M, et al. Div-Seq: single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. Science. 2016;353(6302):925–8.
36. Habib N, Avraham-Davidi I, Basu A, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods. 2017;14(10):955–8.
37. Grindberg RV, Yee-Greenbaum JL, McConnell MJ, et al. RNA-sequencing from single nuclei. Proc Natl Acad Sci U S A. 2013;110(49):19802–7.
38. Rosenberg AB, Roco CM, Muscat RA, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018;360(6385):176–82.
39. Cao J, Packer JS, Ramani V, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017;357(6352):661–7.
40. Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566(7745): 496–502.
41. Tabula Muris Consortium, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018;562(7727): 367–72.
42. Tikhonova AN, Dolgalev I, Hu H, et al. The bone marrow microenvironment at single-cell resolution. Nature. 2019; 569(7755):222–8.
43. Rostom R, Svensson V, Teichmann SA, Kar G. Computational approaches for interpreting scRNA-seq data. FEBS Lett. 2017; 591(15):2213–25.
44. Zappia L, Phipson B, Oshlack A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput Biol. 2018;14(6):e1006245.
45. Freytag S, Tian L, Lönnstedt I, Ng M, Bahlo M. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. F1000Research. 2018;7:1297.
46. van der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
47. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. E-pub 2018. https://doi.org/10.1038/nbt.4314.
48. Wattenberg M, Viégas F, Johnson I How to use t-SNE effectively [Internet]. Distill. 2016. Available at: https://distill.pub/2016/misread-tsne/
49. Azizi E, Carr AJ, Plitas G, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell. 2018;174(5):1293–1308.e36.
50. Bhagwat N, Dulmage K, Pletcher CH, et al. An integrated flow cytometry-based platform for isolation and molecular characterization of circulating tumor single cells and clusters. Sci Rep. 2018;8(1):5035.
51. Richardson GM, Lannigan J, Macara IG. Does FACS perturb gene expression? Cytometry A. 2015;87(2):166–75.
52. Beliakova-Bethell N, Massanella M, White C, et al. The effect of cell subset isolation method on gene expression in leukocytes. Cytometry A. 2014;85(1):94–104.
53. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20.
54. Grün D, Lyubimova A, Kester L, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015; 525(7568):251–5.
55. Grün D, Muraro MJ, Boisset J-C, et al. De novo prediction of stem cell identity using single-cell transcriptome data. Cell Stem Cell. 2016;19(2):266–77.
56. La Manno G, Soldatov R, Zeisel A, et al. RNA velocity of single cells. Nature. 2018;560(7719):494–8.
57. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4): 381–6.
58. Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017;14(10): 979–82.
59. Kosti I, Jain N, Aran D, Butte AJ, Sirota M. Cross-tissue analysis of gene and protein expression in normal and cancer tissues. Sci Rep. 2016;6:24799.
60. Perl K, Ushakov K, Pozniak Y, et al. Reduced changes in protein compared to mRNA levels across non-proliferating tissues. BMC Genomics. 2017;18(1):305.
61. Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. Mammalian genes are transcribed with widely different bursting kinetics. Science. 2011;332(6028):472–4.
62. Bahar Halpern K, Tanami S, Landen S, et al. Bursty gene expression in the intact mammalian liver. Mol Cell. 2015;58(1):147–56.
63. Chubb JR, Trcek T, Shenoy SM, Singer RH. Transcriptional pulsing of a developmental gene. Curr Biol. 2006;16(10):1018–25.
64. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502.
65. Strell C, Hilscher MM, Laxman N, et al. Placing RNA in context and space—methods for spatially resolved transcriptomics. FEBS J. 2019;286(8):1468–81.
66. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015;348(6233):aaa6090.
67. Lee JH, Daugharthy ER, Scheiman J, et al. Highly multiplexed subcellular RNA sequencing in situ. Science. 2014;343(6177):1360–3.
68. Ståhl PL, Salmén F, Vickovic S, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353(6294):78–82.
69. Wang X, Allen WE, Wright MA, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018;361(6400).
70. Rodriques SG, Stickels RR, Goeva A, et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019;363(6434):1463–7.
71. Kawamoto T, Kawamoto K. Preparation of thin frozen sections from nonfixed and undecalcified hard tissues using Kawamot's film method (2012). Methods Mol Biol. 2014;1130:149–64.
72. Chan CKF, Seo EY, Chen JY, et al. Identification and specification of the mouse skeletal stem cell. Cell. 2015;160(1–2):285–98.
73. Abad V, Meyers JL, Weise M, et al. The role of the resting zone in growth plate chondrogenesis. Endocrinology. 2002;143(5):1851–7.
74. Yang L, Tsang KY, Tang HC, Chan D, Cheah KSE. Hypertrophic chondrocytes can become osteoblasts and osteocytes in endochondral bone formation. Proc Natl Acad Sci U S A. 2014; 111(33):12097–102.

75. Zhou X, von der Mark K, Henry S, Norton W, Adams H, de Crombrugghe B. Chondrocytes transdifferentiate into osteoblasts in endochondral bone during development, postnatal growth and fracture healing in mice. PLoS Genet. 2014;10(12): e1004820.
76. Ono N, Ono W, Nagasawa T, Kronenberg HM. A subset of chondrogenic cells provides early mesenchymal progenitors in growing bones. Nat Cell Biol. 2014;16(12):1157–67.
77. Mizuhashi K, Nagata M, Matsushita Y, Ono W, Ono N. Growth plate borderline chondrocytes behave as transient mesenchymal precursor cells. Bone Miner Res. E-pub 2019. https://doi.org/10.1002/jbmr.3719.
78. Zhou BO, Yue R, Murphy MM, Peyer JG, Morrison SJ. Leptin-receptor-expressing mesenchymal stromal cells represent the main source of bone formed by adult bone marrow. Cell Stem Cell. 2014;15(2):154–68.
79. Sacchetti B, Funari A, Michienzi S, et al. Self-renewing osteoprogenitors in bone marrow sinusoids can organize a hematopoietic microenvironment. Cell. 2007;131(2):324–36.
80. Bi W, Deng JM, Zhang Z, Behringer RR, de Crombrugghe B. Sox9 is required for cartilage formation. Nat Genet. 1999;22(1): 85–9.
81. Yamashiro T, Wang X-P, Li Z, et al. Possible roles of Runx1 and Sox9 in incipient intramembranous ossification. J Bone Miner Res. 2004;19(10):1671–7.
82. Akiyama H, Kim J-E, Nakashima K, et al. Osteo-chondroprogenitor cells are derived from Sox9 expressing precursors. Proc Natl Acad Sci U S A. 2005;102(41):14665–70.
83. Nakashima K, Zhou X, Kunkel G, et al. The novel zinc finger-containing transcription factor osterix is required for osteoblast differentiation and bone formation. Cell. 2002;108(1):17–29.
84. Moffatt P, Gaumond M-H, Salois P, et al. Bril: a novel bone-specific modulator of mineralization. J Bone Miner Res. 2008;23(9): 1497–508.
85. Cho T-J, Lee K-E, Lee S-K, et al. A single recurrent mutation in the 5′-UTR of IFITM5 causes osteogenesis imperfecta type V. Am J Hum Genet. 2012;91(2):343–8.
86. Semler O, Garbes L, Keupp K, et al. A mutation in the 5′-UTR of IFITM5 creates an in-frame start codon and causes autosomal-dominant osteogenesis imperfecta type V with hyperplastic callus. Am J Hum Genet. 2012;91(2):349–57.
87. Grcevic D, Pejda S, Matthews BG, et al. In vivo fate mapping identifies mesenchymal progenitor cells. Stem Cells. 2012;30(2):187–96.
88. Kalajzic Z, Li H, Wang L-P, et al. Use of an alpha-smooth muscle actin GFP reporter to identify an osteoprogenitor population. Bone. 2008;43(3):501–10.
89. Matthews BG, Grcevic D, Wang L, et al. Analysis of αSMA-labeled progenitor cell commitment identifies notch signaling as an important pathway in fracture healing. J Bone Miner Res. 2014;29(5):1283–94.
90. Rosen V. BMP2 signaling in bone development and repair. Cytokine Growth Factor Rev. 2009;20(5–6):475–80.
91. Tsuji K, Bandyopadhyay A, Harfe BD, et al. BMP2 activity, although dispensable for bone formation, is required for the initiation of fracture healing. Nat Genet. 2006;38(12):1424–9.
92. Salazar VS, Gamer LW, Rosen V. BMP signalling in skeletal development, disease and repair. Nat Rev Endocrinol. 2016;12(4): 203–21.
93. McBride SH, McKenzie JA, Bedrick BS, et al. Long bone structure and strength depend on BMP2 from osteoblasts and osteocytes, but not vascular endothelial cells. PLoS One. 2014;9(5):e96862.
94. Salazar VS, Capelo LP, Cantù C, et al. Reactivation of a developmental Bmp2 signaling center is required for therapeutic control of the murine periosteal niche. Elife. 2019;8.
95. Marco E, Karp RL, Guo G, et al. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proc Natl Acad Sci U S A. 2014;111(52):E5643–50.
96. Shin J, Berg DA, Zhu Y, et al. Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. Cell Stem Cell. 2015;17(3):360–72.
97. Setty M, Tadmor MD, Reich-Zeliger S, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol. 2016;34(6):637–45.
98. Ji Z, Ji H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res. 2016;44(13):e117.
99. Street K, Risso D, Fletcher RB, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19(1):477.
100. Rizvi AH, Camara PG, Kandror EK, et al. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. Nat Biotechnol. 2017;35(6):551–60.