Single Cell Omics for Musculoskeletal Research Workshop

**Type of Workshop:** Scientific and Technology Workshop

**Sponsors:** The Big Data Workgroup of the IFMRS.

**Significance and Purpose:**
The ability to analyze individual cells within a tissue or cell population is transforming biology and medicine by allowing for the identification of new cell types and lineages that are present during normal and disease states. This workshop will feature speakers who are using cutting edge technologies to study the epigenome, transcriptome and/or proteome of musculoskeletal tissues at the single cell level. Speakers will review the strengths and limitations of current single cell technologies (scRNA-seq, Cy-ToF, and sci-ATAC-seq) as compared to bulk technologies and summarize how these technologies are advancing our understanding of skeletal development and disease.

**Educational need:**
Single cell omics have burst into the field and are being applied to all tissues, including those in the musculoskeletal system. The technologies are relatively easy to perform but bioinformatics analyses of large datasets require more expertise. This workshop will cover key aspects of single cell omics experiments from experimental design through bioinformatics. Strengths and weaknesses of available approaches will be discussed.

**Learning objectives:**
1. Understand how single-cell RNA sequencing can reveal and define cell lineages during embryonic limb development.
2. Describe how ATAC seq and RNA-seq can be combined to understand chromatic structure and epigenomic regulation of gene expression during cartilage development and skeletal disease.
3. Describe how proteomics techniques such as Cy-ToF are used to understand the heterogeneity of cellular populations in healthy and diseased skeletal tissues.

**Results and outcomes:**
It is expected that these presentations will increase understanding of these novel tools and how to access to datasets, as well as pave the way for their integration into our basic and translational research.

**Therapeutic areas to be discussed:** none

**Learning outcome levels (based on Moore's Level outcomes):**
*Level 3B Learning: Procedural Knowledge (Knows How)*
Clinical relevance: This workshop will increase understanding and prediction of patient heterogeneity and stratification of patient populations.

Targeted audience: Basic and physician scientists interested in the cell biology of musculoskeletal system development, repair and disease will find this workshop of interest.

Benefit to meeting attendees for this workshop: Attendees will leave with an increased understanding of single cell omics technologies, discover how they may incorporate the technologies into their own research projects, and learn how to access and interpret datasets.

Abstract: In recent years, tools for analysis of the transcriptome, epigenome and proteome at the individual cell level have advanced by leaps and bounds. "Big data" generated from single cell "omics" brings about a paradigm shift to the field as it is relatively easy to collect, but not as easy to analyze, store and interpret. The ever-increasing publications in single cell omics suggest that these analyses are both dynamic and impactful. In fact, these technologies are our latest tools for understanding complex diseases and tissue physiology because they provide single cell resolution of genes and proteins. These emerging tools are underutilized by our musculoskeletal community. Therefore, it is of paramount significance to educate ORS members on how single cell technologies can be used to answer research questions and advance translation. This workshop will feature three outstanding speakers from across the nation who will cover three different yet complementary aspects of single cell "omics". Dr. Chia-Lung Wu, PhD (Washington University in St. Louis, ORS member) will summarize how scRNA-seq analysis can unravel the cellular heterogeneity and trajectories of lineage specification during embryonic limb development. Dr. Terence Capellini, PhD (Harvard University, ORS member) will speak about how ATAC-seq and RNA seq data can be combined to understand how genes are regulated during cartilage development and disease. Finally, Dr. Nidhi Bhutani, PhD (Stanford University, ORS member) will highlight novel high-content single-cell analysis (CyTOF, mass cytometry) to reveal the heterogeneity of cellular populations in healthy and diseased skeletal tissues. It is expected that these presentations will not only increase understanding of these novel tools and access to these datasets but also pave the way for their integration into our basic and translational research.

Please indicate appropriate keyword(s): Bone, Cartilage, Fracture repair, Meniscus, Skeletal Muscle, Spine, Tendon
Using Single Cell RNA Sequencing to Determine Cellular Heterogeneity and Trajectories of Lineage Specification

Chia-Lung Wu¹,³, PhD, Amanda Dicks¹,²,³, and Farshid Guilak¹,³, PhD
¹Orthopaedic Surgery, ²Biomedical Engineering, Washington University in Saint Louis
³Shriners Hospitals for Children – St. Louis

Abstract. The recent advance of next-generation sequencing (NGS) including bulk RNA sequencing (bulk RNA-seq) provides a high-throughput approach to investigate diverse, complex biological systems. Transcriptomic studies using bulk RNA-seq are based on the assumption that all the cells within a given tissue are homogeneous and thus they share similar gene expression profile at the given time or status. However, such an assumption ignores the stochastic nature of gene expression. This stochastic gene expression can lead to heterogenous cellular composition in a tissue and often plays a critical role in driving cell fate decision during development. Therefore, recent NGS-based studies started to characterize transcriptomic profiles at the resolution of an individual cell by using single-cell RNA sequencing (scRNA-seq). In this presentation, I will introduce how scRNA-seq can be applied to 1) reveal cellular heterogeneity, 2) delineate differentiation trajectories, as well as construct 3) heterogenous cellular signaling models and 4) gene regulatory networks (GRNs) (Fig 1). Several commonly used R packages for bioinformatic analysis of scRNA-seq data to accomplished these goals will be discussed.

Experimental design. Human induced pluripotent stem cells (hiPSCs) were induced into chondrogenic differentiation according to our previously established protocol. Cells were harvested at various time points for scRNA-seq. Single cells were captured using the Chromium Controller (10x Genomics) platform, and sequencing was performed by Illumina HiSeq2500. Note that there are several other platforms which have been developed for this purpose and while the advantages/disadvantages of each platform are beyond the scope of this presentation, they have been extensively reviewed¹. Sequencing reads were processed by Cell Ranger (10x Genomics software). Briefly, reads were aligned to the GRCh38 (version 90) for genome annotation, demultiplexing, barcode filtering, and gene quantification. Gene barcode matrices for each sample were generated by counting the number of unique molecular identifiers (UMIs) for a given gene (as row) in an individual cell (as column). For each sample, ~2,500 cells were captured in our case.

Determine cellular heterogeneity by Seurat. To perform quality control and access the heterogenous composition of cell populations, gene barcode matrices were input into the Seurat R package (https://satijalab.org/seurat/)². We then removed the low-quality cells characterized by less than 200 or more than 7,000 detected genes or if the mitochondrial gene content was more than 5%. Note that the cutoff criteria will need to be adjusted as gene expression levels could be cell type-dependent. Genes that were detected in less than 3 cells were also filtered out. Next, to reduce the variance introduced by “unwanted” sources, we regressed out variation in gene

Fig 1. Representative applications that can be accomplished by scRNA-seq and bioinformatic analyses.
expression driven by cell cycle stages and mitochondrial gene expression with the `vars.to.regress` argument in the function `ScaleData` in Seurat. Dimensionality reduction on the data was then performed by computing the significant principal components on highly variable genes. We then performed unsupervised clustering by using the `FindClusters` function in Seurat with the resolution argument set to 0.6 to visualize clusters in a t-distributed Stochastic Neighbor Embedding (tSNE) plot (Fig 2). Differentially expressed genes (DEGs) among each cell cluster were determined using the `FindAllMarkers` function. DEGs expressed in at least 25% cells within the cluster and with a fold change of more than 0.25 in log scale were considered to be marker genes of the cluster. To determine the biological functions of the marker genes from a given cluster, we performed Gene Ontology (GO) enrichment analysis by using the DAVID Gene Functional Classification Tool (http://david.abcc.ncifcrf.gov). By comparing these unique biological GO terms with existing RNA-seq datasets and the literature, we were able to annotate cell clusters.

**Determine differentiation trajectory by Monocle.** We used Monocle2 R package (http://cole-trapnell-lab.github.io/monocle-release/) to reconstruct differentiation trajectories by computing and ordering the sequence of gene expression changes of the cells collected from different time points in an unsupervised manner. First, scRNA-seq datasets from different timepoints were merged into one single object using the `MergeSeurat` function in Seurat. The merged matrix was then converted into a Monocle object using the `importCDS` and `newCellDataSet` functions in Monocle. We then identified a set of DEGs between the cells collected at the beginning of the process to those at the end using the `differentialGeneTest` function with argument `qval < 0.01`. The dimensions of the dataset were then reduced using the `DDRTree` method. Next, we used the `orderCells` function to order the cells based on the selected DEGs and then visualized the trajectory of the cells with the `plot_cell_trajectory` function in Monocle. The temporal expression of the gene of interest was visualized using the `plot_genes_in_pseudotime` function in Monocle (Fig 3).

**Heterogenous multicellular signaling models by Seurat and Circlize.** To investigate the ligand-receptor interaction in heterogenous multicellular signaling systems, we used a list comprising of 2,557 human ligand–receptor pairs curated by the Database of Ligand-Receptor Partners (DLRP), IUPHAR, and Human Plasma Membrane Receptome (HPMR). We first quantified the percentage of the cells (i.e., neural cells, melanocytes, and chondrocytes) that expressed a specific WNT ligand and its associated frizzled (FZD) receptors using scRNA-seq datasets. To ensure the ligand and receptors are uniquely expressed, we required a fold change in expression of more than 0.25 in log scale (Fig 4A). We then used the Circlize R package to visualize the directions of the signaling in the cell type based on ligand-receptor pairs (Fig 4B).
Reconstruct GRNs by WGCNA. We used Weighted Gene Co-expression Network Analysis (WGCNA) to reconstruct GRNs and identify their associated hub genes that regulate cell differentiation. First, the dataset of interest (e.g., a given timepoint) created in Seurat was converted into a plain matrix for a given gene (column) in an individual cell (row). The dataset was then cleaned by removing cells with too many missing values using the goodSamplesGenes function. Next, we used the pickSoftThreshold function in WGCNA to determine the proper soft-thresholding power ($\beta$) that fits the criterion of approximate scale-free topology of the network. An adjacency matrix was then built with a soft-thresholding power of 8 in our study. Hierarchical clustering and GRNs were constructed by using the blockwiseModules function with arguments TOMType set to "unsigned", networkType set to "signed", and mergeCutHeight set to 0.25. Modules containing genes that were highly associated with each other were identified in this process. Gene lists of interesting modules were extracted and submitted to DAVID for GO term analysis to retrieve their biological processes and molecular functions. We then identified transcription factors (TFs) and TF regulators from the genes based on the GO terms in molecular functions. Based on the highest weight (i.e., high correlation coefficient) connected to a given TF or TF regulator, we selected the top 100 genes. Finally, the GRN based on these TFs and TF regulators underwent cluster analysis using community cluster and was then visualized using Cytoscape. Hub genes for each GRN were identified as genes with high weight (summed correlation coefficients), high degree (summed connectivity; i.e., total numbers genes connected to this specific gene), and high betweenness centrality measure of the network (Fig 5).

References:
Exploring the Regulatory Control of Cartilage Development to Understand Skeletal Disease
Terence D. Capellini, PhD
Department of Human Evolutionary Biology, Harvard University, Cambridge, MA

INTRODUCTION: In the past decade, the rapid rise of functional genomics methods has helped to foster insights into how the skeleton is built and how such underlies skeletal disease. In the past five years, the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) has made it possible to epigenetically profile open versus closed states of chromatin across the genome in specific cell types involved in forming and maintaining the skeleton (e.g., chondrocytes, osteoblasts). While this method, along with a more recent application of it in the single cell sphere, has great potential to find important regulatory regions in the genome as well as to characterize the epigenome of skeletal cell subtypes, there remain important methodological and computational obstacles that need to be addressed. In this workshop, I address some of these issues in the context of our ATAC-seq studies on chondrocytes acquired in bulk from both mouse and human developmental skeletal samples.

METHODS: We first used ATAC-seq to epigenetically profile chondrocytes extracted in vivo from mouse E15.5 and stage-matched human E54 long bones elements (N=3 per tissue/time-point). We also performed ATAC-seq on developing brain tissues in order to help identify regions that are unique to individual bones and bone-ends (e.g., the distal femur). We next used a method called the Irreproducible Discovery Rate (IDR) to find peaks or regions that are consistently accessible between biological replicates per tissue per timepoint and species. We then explored the regulatory and evolutionary potential of called peak sets using comparative bioinformatics and evolutionary analyses.

RESULTS: Our studies found for each species tens of thousands of open chromatin or accessible regions that are shared between chondrocytes extracted from different anatomical locations (i.e., from the developing proximal and distal bone ends of the skeleton). Yet, our analyses also discovered hundreds to thousands of bone-end-specific chondrocyte open chromatin regions (e.g., regions uniquely accessible in the distal femur as compared to the proximal femur or proximal tibia). Our use of IDR on each species tissue replicates caused unfiltered (pre-IDR) peak sets to be reduced by 50-80%, resulting in much smaller conservative peak sets per tissue. Our comparisons of ATAC-seq peaks between stage- and anatomical-site-matched mouse and human samples reveals generally little overlap (~25-45%) between species. Our evolutionary analyses on human/mouse sequences reveal that open chromatin regions unique to distinct bone ends (e.g., the distal femur as compared to the proximal femur) display substantial evidence of ancient natural selection and modern human sequence constraint, reflecting functional demands of joint sites during the evolution of human bipedalism.

DISCUSSION: Our studies identified both general chondrocyte as well as bone-end-specific chondrocyte regulatory profiles with important ramifications to understanding the development of bone-end-specific anatomy and disease risk. Our use of IDR on each set and the resultant substantial decrease in peak set sizes for each bone-end reveals the importance of understanding and controlling for biological noise among sample datasets, and in-so-doing has important ramifications for applying this method in the single cell omics sphere. Our cross-species comparisons between human and mouse samples reveal substantial divergence in accessibility patterns for all bone ends examined, likely reflecting, among other factors, stark differences in genomic usage in building the skeleton over evolutionary time. Finally, our evolutionary analyses highlight the importance of examining open chromatin sequences in non-standard ways in order to understand how the functions of these sequences (i.e., during chondrocyte and skeletal development), as well as risk variants within them, have been shaped by long-acting evolutionary forces.

SIGNIFICANCE: Insights from ATAC-seq studies performed on chondrocytes acquired in bulk from developing mouse and human bone-ends have the potential to inform and guide future single cell ATAC-seq studies.
Aging or injury contribute towards skeletal diseases such as osteoarthritis (OA), osteoporosis, and sarcopenia. Single-cell studies have the great potential of identifying cellular subpopulations that work in a synchronized fashion to regenerate and repair damaged tissue during normal homeostasis. In addition, such studies can elucidate how these processes break down in disease as well as identify interpatient variability in the cellular subpopulations that drive the disease. Mass cytometry by time-of-flight, or cyTOF is a mass-spectrometry based high dimensional method for single-cell detection of isotope labeled antibodies that can map as many as 40-120 protein epitopes for a single cell. This high-dimensional method for proteomic analyses complements single cell transcriptomic and epigenetic techniques. In our recent studies (Science Advances, in press), we have profiled healthy and OA cartilage samples using mass cytometry to establish a single-cell atlas for cartilage. Although cartilage is a relatively simple tissue, with a single cell type being encapsulated in its secreted extracellular matrix (ECM), the variable degree of degeneration associated with each OA patient suggests that understanding this tissue at a single cell level can provide novel insights into the onset and progression of pathology. Our single cell atlas has revealed distinct chondrocyte progenitor and inflammation modulating subpopulations in OA cartilage. These rare populations include an inflammation amplifying (Inf-A) population, whose inhibition decreased inflammation, and an inflammation dampening (Inf-D) population which is resistant to inflammation. A pharmacological strategy targeting Inf-A and Inf-D cells significantly decreased inflammation in OA chondrocytes, highlighting the relevance of small cellular sub-populations in OA. Additionally, the single cell data allowed us to stratify OA patients in three groups that are distinguished by the relative proportions of inflammatory to regenerative cells. Single cell proteomic studies therefore have a great potential to provide a deeper understanding of healthy and diseased skeletal tissues as well as identify the patient-to-patient differences that may be critical in determining precision medicine approaches for effective therapies.

**Figure 1** Schematic outlining the approach for high dimensional, single cell profiling of chondrocytes (normal or OA) by mass cytometry, cyTOF