Aim and Overall Strategy

The aim of this analysis is to systematically investigate sex-specific transcriptional differences in human bone marrow cells using single-cell RNA sequencing (scRNA-seq) data from Sun et al. (Immunity 2024), with a particular focus on how these differences relate to BCG vaccination status and longitudinal time points (t0 vs td90). To achieve this, a computational pipeline is proposed that integrates raw data processing, rigorous quality control, batch correction, cell-type annotation, differential gene expression, network-level inference, pathway activity analysis, and trajectory modeling.

Each of the step is explained in brief below:

Data Processing, Quality Control, and Integration

Raw sequencing data can be retrieved using the SRA Toolkit, which provides a reliable and reproducible mechanism for downloading large-scale sequencing datasets and generating FASTQ files. These reads can be processed using Cell Ranger, which performs demultiplexing, alignment, and UMI-based quantification to produce standardized gene-cell count matrices suitable for downstream single-cell analysis. Individual 10X Genomics datasets (count/barcode/feature/metadata) could subsequently be imported into R using Seurat (ReadMtx) and converted into sample-specific Seurat objects, thereby preserving metadata required for sex, batch, and condition-specific analyses.

To enable joint analysis across multiple samples and experimental timepoints, datasets should be merged and subjected to quality control (QC). Standard QC metrics, including mitochondrial (percent.mt), ribosomal (percent_ribo), and haemoglobin (percent_hb) gene percentages, nCount_RNA, nFeature_RNA, and log10GenesPerUMI, should be calculated to identify and remove low-quality cells. Normalization can be performed using either CLR normalization or SCTransform to stabilize variance across cells and mitigate technical noise. Batch effects arising from sequencing runs or sample processing can be addressed by integrating with RPCA-based anchors in Seurat or by refining with Harmony to minimize residual technical variation. Cell types can be annotated using SingleR (celldex::HumanPrimaryCellAtlasData) to assess cell-type identity and stratify cell-level identity relevant to the downstream analysis. Dimension reduction methods (UMAP) help visualize the data distribution across gender and condition-specific groupings. Finally, saving the processed object as an H5Seurat file ensures reproducibility and enables reloading for downstream analyses.

Differential Expression and Network-Level Analysis

Following preprocessing and integration, differential gene expression analysis using (DESeq2) can be conducted by explicitly assigning cell identities to biological sex and applying nonparametric statistical tests (Wilcoxon rank-sum test) to identify gender specific genes. Visualization with volcano plots using the EnhancedVolcano package

enables rapid assessment of statistical significance (log2 fold change, p-value) of the obtained markers.

To move beyond single-gene comparisons and capture coordinated transcriptional programs, high-dimensional weighted gene co-expression network analysis (hdWGCNA) can be applied. To improve signal-to-noise ratio and reduce cell-level stochasticity while preserving underlying biological structure, meta-cells could be constructed (MetacellsByGroups), and co-expression modules could be identified using a scale-free network (soft thresholding) framework. Module eigengenes can then be correlated with Traits, which in our case are sex, vaccination status, and timepoint, to reveal higher-order gene programs associated with sex-specific immune and hematopoietic states. Hub gene analysis (ModuleConnectivity, PlotKMEs) could further highlight regulatory drivers within sex-associated modules, providing mechanistic insight that complements differential expression results.

Pathway Activity and Temporal Dynamics

To functionally interpret sex-specific transcriptional differences, Gene Set Variation Analysis (GSVA) can be applied using curated Hallmark pathways. By transforming gene-level expression into pathway-level enrichment scores using a Gaussian kernel (kcdf = "Gaussian"), GSVA reduces noise from individual genes and enhances the detection of coordinated biological programs that may differ between male and female bone marrow cells. Differential pathway activity testing could further link sex differences to immune, metabolic, and signaling pathways to trained immunity and bone marrow function.

Finally, trajectory analysis (pseudotime) using Monocle3 can introduce a temporal dimension to the analysis by modeling transcriptional state transitions across baseline (t0) and post-vaccination (td90) timepoints. By integrating Seurat-derived embeddings and cluster annotations, cells can be ordered along pseudotime to identify genes and programs that change dynamically over time. Mapping pseudotime back onto the Seurat object enables integrated visualization, which helps compare gender-specific and different experimental conditions.

Summary

Together, this multi-layered framework can provide an in-depth strategy for identifying gender-specific differences in bone marrow scRNA-seq data, accounting for the dimension of experimental state. By integrating gene-level, network-level, pathway-level, and temporal analyses with adequate QC checks, the pipeline can help ensure that observed differences reflect true biological variations.