Write an instruction document for the single-cell dataset for beginners to read.

a) What's the single-cell data?
Single-cell data are datasets that include expression level information for individual cells from a biological sample. In a single-cell dataset, each row represents a gene, and each column is a cell. The single-cell datasets are in the form of a data matrix, with the rows as the features and the columns as the individual independent observations. The values in the count matrix represent the counts for each of the genes corresponding to each cell.

b) Summarize descriptive statistics of real single-cell data.
Perhaps the most significant descriptive statistics of a real single-cell dataset is that most genes will contain zero counts. To save memory and speed, these zero counts are represented by a "." in the dataset. For the Peripheral Blood Mononuclear Cells (pbmc) dataset, the sparse matrix takes up 29905192 bytes of memory, whereas the normal dense matrix where all count values including zeros are stored takes up 709591472 bytes of memory. There are 2700 columns/cells in the dataset, and there are 32378 rows/genes. We could perform colSums() on the dataset to obtain the sum of a column/cell, and perform rowSums() on the dataset to obtain the sum of a row/gene. By performing rowSums()==0 on the dataset, we can see that the majority of the results are values close to 2700; that is, when you randomly select a row/gene to observe, most of the count values will be 0. And if you perform colSums() ==0 on the dataset, the majority of the results are between 31000 to 32000; that is , when you randomly select a column/cell to observe, most of the count values will also be 0. We could then plot the histogram of counts_per_cell, counts_per_gene, genes_per_cell, and cells_per_gene. We may observe that these histograms are mostly skewed to the right. We could perform the log(x+1)-transition on these histograms, resulting in a normal distribution. There is also a total of 88392600 counts in the pbmc dataset, the sum of those counts is 6390631, and the mean is 0.07229826. The maximum count value in the dataset is 419. We could then create the Seurat object with the raw, non-normalized data. The Seurat object serves as a collection of Assay and DimReduc objects, representing expression data and dimensionality reductions of the expression data. The Seurat object has a total of 12 slots and two columns, named nCount_RNA and nFeature_RNA respectively. We could then use the subset function to filter out low-quality cells based on user-defined criteria. After removing unwanted cells from the dataset, we use a global-scaling normalization method. The third pre-processing step is to detect highly variable genes across the single cells, and Seurat focuses on these cells for downstream analysis.

c) Summarize the read materials and package.
The article from https://www.embopress.org/doi/full/10.15252/msb.20188746 titled "Current best practices in single-cell RNA-seq analysis" details the steps of a typical single-cell RNA-seq analysis. Figure 1 gives a schematic of a typical single-cell RNA-

seq analysis workflow. In broad terms, raw sequencing data are processed and aligned to give count matrices, which represent the start of the workflow. The count data undergo pre-processing and downstream analysis. The first step in pre-processing is called Quality Control (QC). Cell Quality control is commonly performed based on 3 QC covariates: the number of counts per barcode (count depth), the number of genes per barcode, and the fraction of counts from mitochondrial genes per barcode. The distributions of these QC covariates are examined for outlier peaks that are filtered out by thresholding.
The next step in pre-processing is Normalization. When gene expression is compared between cells based on count data, any difference may have arisen solely due to sampling effects. Normalization addresses this issue by scaling count data to obtain correct relative gene expression abundances between cells. After normalization, data matrices are typically log(x+1) transformed. The next step is data correction and integration. Data correction targets further technical and biological covariates such as batch, dropout, or cell cycle effects. Data integration is the integration of data from multiple experiments. Moving on, feature selection is next performed. In this step, the dataset is filtered to keep only genes that are "informative" of the variability in the data. After that, the dimensions of these single-cell expression matrices can be further reduced by dimensionality algorithms, and this step is called dimensionality reduction. Now preprocessing is finished. Moving on, methods we call downstream analysis are used to extract biological insights and describe the underlying biological system. These descriptions are obtained by fitting interpretable models to the data. Downstream analysis is also divided into cell and gene level approaches, whereas cell-level analysis approaches are again subdivided into cluster and trajectory analysis branches. We then perform clustering and compositional data analysis. Compositional data analysis revolves around the proportion of cells that fall into each cell-identity cluster.

The article titled "Gene regulatory network inference: an introduction survey" (Paper 1) summarizes the current strategies to construct gene regulatory networks. Different models and strategies can be used to construct GRNs, though each method has both advantages and limitations. No strategy is perfect. According to the article, popular strategies to construct GRN include measurement techniques (Microarray technology, Next Generation Sequence (NGS) technology), data-driven methods (Correlation Networks, Mutual Information Networks), regression-based methods, probabilistic models (Gaussian Graphical Models, Bayesian Networks), Dynamic Networks (Dynamic Bayesian Networks, Differential Equation Methods), and Multi-Network Models (Time-Varying Models).

d) Summarize the basic pipeline (pre-processing) of single-cell data analysis.
The basic pipeline (pre-processing) of single-cell data analysis is as follows:
    1) Quality Control --- removing unwanted cells (such as broken cells and doublets) from the dataset and selecting cells for further analysis

2) Normalization --- Normalizing the data by a global-scaling normalization method. When gene expression is compared between cells based on count data, any difference may have arisen solely due to sampling effects. Normalization addresses this issue by scaling count data to obtain correct relative gene expression abundances between cells, therefore reducing sampling variability
3) Scaling the data – Shifts the expression of each gene, so the mean expression across cells is 0 and the variance across cells is 1. This gives equal weight for each cell in downstream analysis
4) Data correction and integration. Data correction targets further technical and biological covariates such as batch effects, dropout, or cell cycle (biological process) effects. Data integration is the integration of data from multiple experiments.
5) Feature Selection --- identification of highly variable features for usage in downstream analysis, the dataset is filtered to keep only genes that are "informative" of the variability in the data.
6) Dimensionality reduction --- using dimensionality algorithms to further reduce the dimensions of the single-cell expression matrices. Use linear or nonlinear methods to reduce the number of genes in the dataset to produce a much lower dimensional dataset.