Single cell transcriptome of mESC data

LifeBytes Consulting

Authors: LifeBytes Consulting Pty Ltd

Contact: contact@lifebytes.com.au

scRNA pipeline white paper

"Single cell analysis is gaining momentum in the scientific community, as it offers cellular resolution that can help decipher complex biological systems. This increase in biological resolution leads to a corresponding increase in analytical complexity, demanding high-end bioinformatics expertise. To help researchers better navigate the complex and ever-changing field of scRNA-seq biology, we at **LifeBytes** have a dedicated team that is constantly exploring new developments in this field, in order to test and identify the optimal approaches necessary to create an effective scRNA-seq data analysis pipeline. In this paper we showcase our single cell Bioinformatics solution using an scRNA-seq test case, demonstrating our expertise both in data analytics and biological interpretation".

Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized the study of complex biological systems, offering cell-by-cell insights not available from traditional bulk-RNA sequencing. This platform yields superior biological resolution, but that resolution comes at the cost of increased technical noise and data complexity. Traditional RNA-seq data analysis methods cannot handle these technical challenges, and it is instead vital that dedicated single-cell methods are applied to effectively address these platform-specific challenges and to exploit these gains in cellular resolution to attain meaningful biological insights.

scRNA sequencing permits for direct cell-by-cell comparisons within a heterogeneous cell population allowing, for example, the identification of gene expression patterns specific to malignant cells within a complex tumour mass. scRNA-seq data can be widely applied to identify novel cell subtypes, to characterize differentiation processes, to assign cells to their current cell cycle phases, or to identify highly variably expressed genes that may drive broader variation across a mixed cell population. These novel insights can only arise from carefully constructed bespoke scRNA-seq analysis methods.

Recently, the improvements in techniques that capture and sequence the 3' end of uniquely barcoded RNA molecules has vastly improved scRNA-seq data quality by allowing for quantification of unamplified cDNA sequences, avoiding potential biases that can appear in datasets stemming from sequence amplification. The vast majority of currently used approaches introduce a UMI to each RNA molecule and then sequence only their 3' ends, allowing for cost effective sequencing of thousands of mRNA molecules simultaneously. However, in single cell RNA sequencing it is more difficult to reliably capture transcripts for cDNA production due to low starting RNA quantities, increasing the frequency of drop-out events wherein transcripts for a particular gene are ultimately not captured, yielding false negative results. Dedicated steps are necessary to compensate for the noise introduced by this technical complexity. Globally, researchers are developing myriad approaches to analyse scRNA data. These approaches can result in very different interpretations of the underlying data and, importantly, some of these approaches are not well supported beyond the initial release. Rigorous testing is often needed to select appropriate tools and evaluate the suitability of a given approach to a scRNA-seq experiment.

scRNA analysis workflows at LifeBytes

Unlike analytical techniques associated with other well established high-throughput sequencing approaches, scRNA-seq analyses requires expertise in programming languages, particularly R, to effectively handle data output. Development of an experienced bioinformatics support team with a carefully crafted analytical pipeline is essential to handle these data effortlessly and reproducibly. Some sequencing platforms offer their own software packages for data analysis, but these are usually hamstrung by proprietary algorithms that prohibit analytical customization and often mask the ways in which data are being manipulated. Beyond basic analyses, these packages are generally insufficient and other publically available tools are necessary to complete the analyses. To alleviate these constraints and provide complete bioinformatics support for scRNA-seq analysis, we have tested a range of tools and developed a standardized pipeline that can handle analysis of data from many different types of UMI-based scRNA-seq.

The general computational workflow of scRNA-seq analyses involves three key analytical stages:

1. **Debarcoding** – Differentiating barcodes and generating a gene-cell count matrix.
2. **Dimensionality reduction** – Clustering and cell type/sub-population identification.
3. **Dynamic analysis** – Data/project specific analysis of key readouts of interest, such as cellular heterogeneity, cell-cycle staging, pseudotime analysis, etc.

The outputs of these different stages of the analysis pipeline are illustrated in Figure 1.

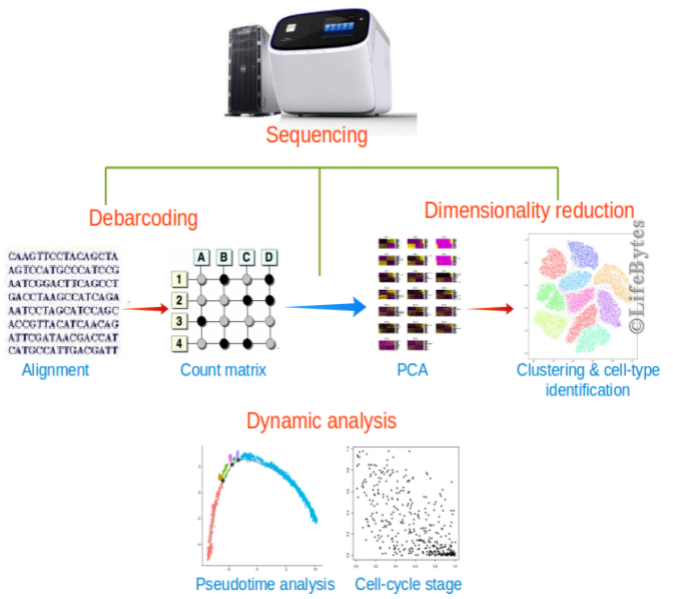

Figure 1. A basic scRNA analysis workflow with representative plots for each step.

Debarcoding

The preliminary analysis produces a count matrix based on the raw sequencing reads. Much as is the case for bulk RNA analysis, scRNA-seq also necessitates careful QC processing and alignment of these raw reads to facilitate downstream analysis. The most important aspect of this step is the accurate identification of true cellular barcodes and UMIs. Any sequencing- or PCR-related errors within these barcoded stretches can complicate such identification. Indeed, in short barcodes (usually 6-36bp), one or two sequencing errors can alter their specificity resulting in combination of reads from multiple cells, altering the count matrix output. To overcome this challenge it is essential to model the distribution of these barcodes in order to identify any errors.

Many approaches have been developed to extract reads based on the cell and UMI barcodes. Most sequencing platforms provide software suites to carry out this step, however their results are restricted to use within their proprietary suite and cannot be easily exported. For instance, the 10X platform provides the CellRanger kit for this analysis, but the extracted reads cannot be used outside of their pipeline for any downstream analyses. Other open-source tools handle these barcodes using a network-based approach that accounts for sequencing and PCR induced errors, thus increasing the read count yield. A network has to be built to account for all possible unique barcodes, and it must be able to error correct those barcodes to fit an expected minimum number of cells. When the expected number of cells is clear, as in a well/droplet-based approach where the cell number is specified beforehand, it is a straightforward process for any tool to identify true barcodes and the reads associated to them. When the number of cells is not previously defined, however, true barcode estimation has to be unsupervised and must be able to both account for errors and accurately estimate the number of cells. For illustrative purposes, we have downloaded the 10X chromium data of E18 mouse cells from their public repository – a dataset expected to have ~10,000 cells. Accurate

extraction tools should be able to determine the minimum number of expected cells and identify reads accordingly. Our pipeline was able to identify 10270 cells with error-corrected barcodes for downstream analysis in this dataset (Figure 2).

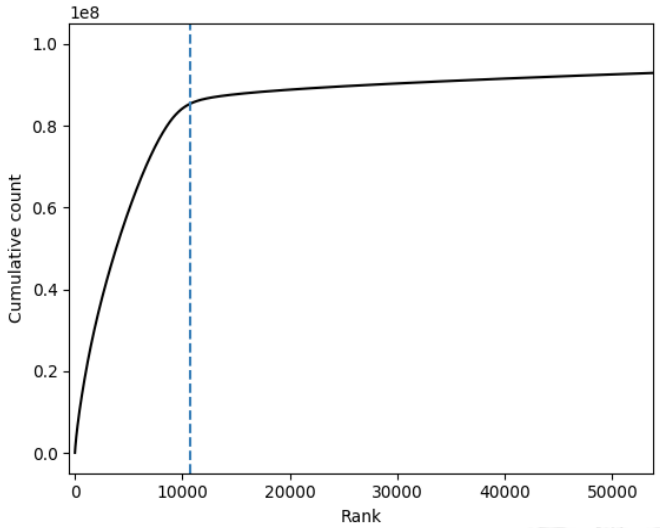

Figure 2. A knee plot which displays the estimated number of unique barcodes and identifies the saturation point, after which the cumulative barcode counts become saturated. In this plot of 10X chromium data, the knee point is at 10,270 cells, which is close to the ~10,000 cells reported in the 10X repository.

As alignment tools are well established for NGS data, alignment remains a straightforward step in scRNA-seq analyses. Even so, choosing appropriate aligners and post-processing steps (based on the data type and/or aim of the study), can significantly alter the results. For example, UMI data can be potentially used for effectively de-duplicating PCR reads, which is often a difficult step even in bulk RNA sequencing[1]. However, many pipelines and publications seem to ignore these PCR duplicates which can lead to over-estimation of read/UMI counts. Following alignment, UMI counting for each cell and for the gene list will result in a gene-cell count matrix, where each row is a gene/transcript and each column is a cell, with UMI counts for each of them. This is the basic input file that can be used for any downstream scRNA-seq analyses.

DATASET

To illustrate advanced and customized analysis modules, we have used datasets from a study of mouse Embryonic Stem Cells (mESC). In this study, mES cells were collected after being stimulated with retinoic acid for different periods (2h, 6h, 12h, 24h, 36h, 48h, 60h, 72h & 96h) of time to induce differentiation. Collected cells were then sequenced using a single cell sequencing SCRB-seq protocol (Semrau et al, 2017, GSE79578).

Dimensionality reduction

Deriving biologically meaningful insights from the count matrix generated above is an important and yet complex step of scRNA-seq data analysis. Outliers and low quality cells must be removed as they can bias subsequent interpretations. Three common quality control steps involve assessing the library size (the number of UMIs), the number of expressed genes, and the percentage of mitochondrial contamination in each library. Cells with relatively low UMI counts are considered to be of low quality as their RNA has not been efficiently captured. Similarly, any cells with low numbers of expressed genes are likely to be of poor quality as their diverse transcripts haves not

been successfully captured. Likewise, genes that are expressed in very few cells can also be considered to be outliers[2]. Such differences can be efficiently spotted from by plotting the number of features and UMIs per cell (Figure 3).

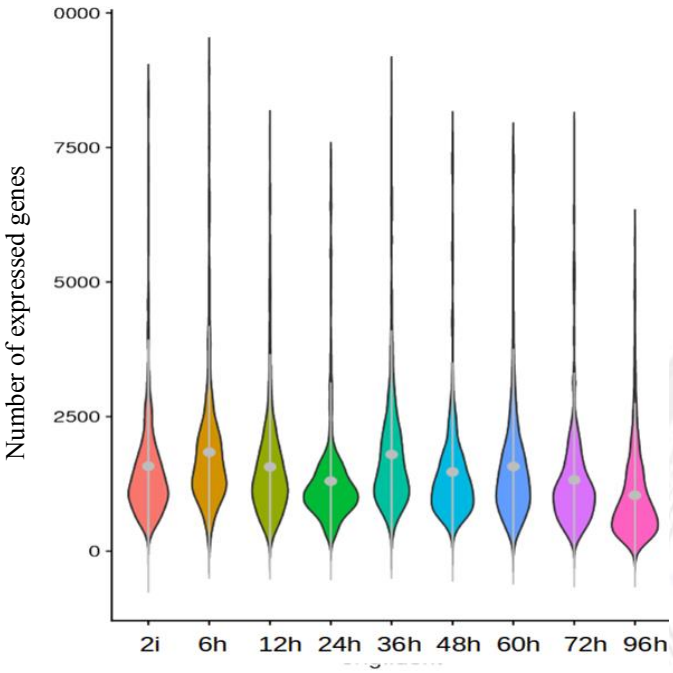

Figure 3. A basic violin plot displaying the number of genes expressed at each time-point to illustrate the expression distribution over time.

An expression matrix can contain millions of cells, making it extremely difficult to understand the underlying patterns. Reducing the number of dimensions by grouping cells with similar expression patterns is required to refine downstream analyses. Principal component analysis (PCA) is one of the most widely used methods for such dimensionality reduction[3,4]. Although effective in certain contexts, the relatively high number of undetected genes per cell can produce many zeroes in the count matrix, challenging traditional PCA or clustering analyses. Several other PCA-dependent methods have been developed to handle these "zero inflated" datasets[5]. The key step in this process is identifying the optimal number of Principal Components (PCs), which requires testing PCA with a default number of dimensions and analysing the PC scores to select significant PCs. For the mESC data, we have selected the first four PCs based on the Elbow plot shown in Figure 4.

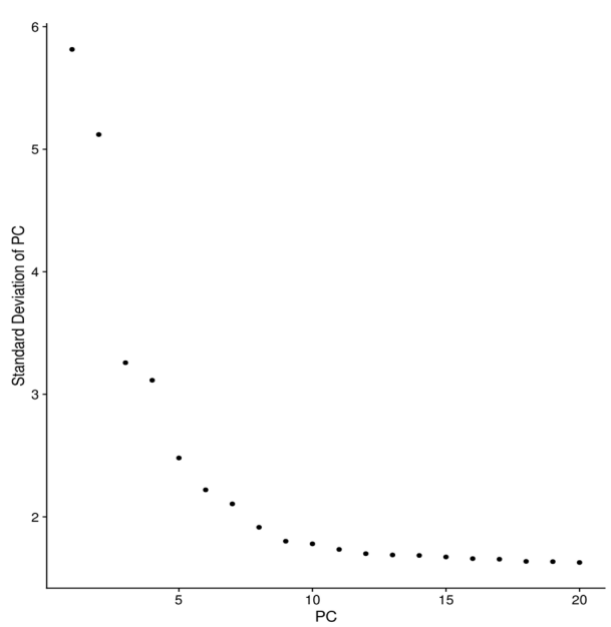

Figure 4. Elbow plot illustrating the PC score (standard deviation of PCs) to efficiently highlight the significant PCs for mESC data. The first four PCs appear to be significant, with the first two being highly significant.

Following dimensionality reduction, clustering of cells to group them based on their digital gene expression patterns allows us to identify the sub-populations present in a heterogeneous cell mixture. For clustering, the *t*-stochastic neighbour embedding (*t*-SNE) method, a non-linear clustering approach, is widely used to identify

different clusters/subpopulations[6]. For the example mESC data, each sample represents expression values at different time-points during the mESC differentiation process, and as a result the specific time of collection of cells can lead to similar gene expression across cells at any given time point, causing clustering based on time of collection as shown in Figure 5.

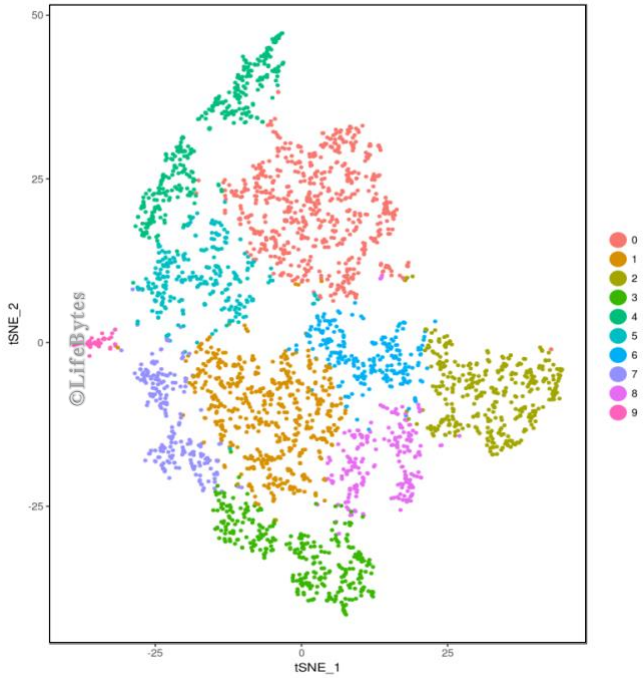

Figure 5. t-SNE plot with clustering of cells of mESC data. As the data represent expression values at different time-points, cells seem to be clustered based on their time of collection, with the number of clusters being equivalent to the number of time-points.

Hence it is necessary that cell-cycle specific expression patterns be omitted from the analysis in order to generate accurate cell-type specific clusters.

Dynamic analysis

For this sample dataset consisting of different time-points, the clustering approach must be modelled based on the different time-points of capture. By identifying cell-type specific marker genes, we were able to assign the cell types to clusters accordingly (Figure 6). We were able to identify two cell types in this population: extraembryonic endoderm (XEN) cells and neuro-ectodermal (NEU) cells. There were additionally many ambiguous cells, wherein mixed expression of different markers was observed, as well as a few cells without a clear cell identity requiring further scrutiny and an enhanced list of gene markers. This finding is consistent with the results in the original publication[7].

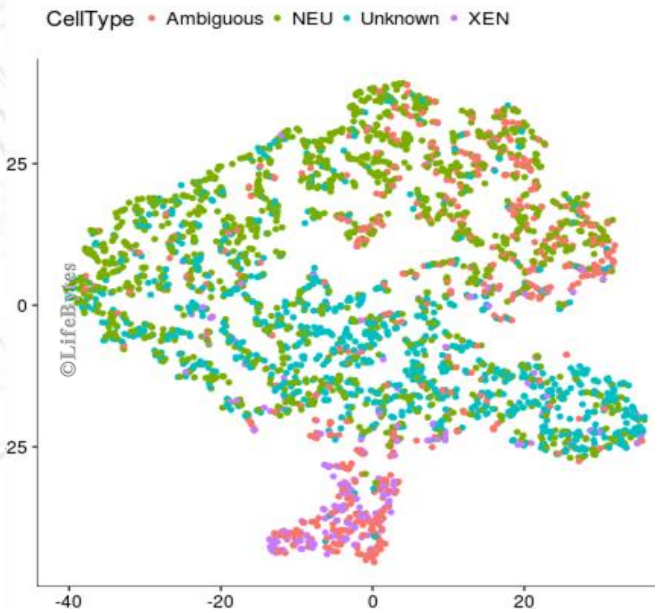

Figure 6. t-SNE plot showing clustering of cells in the mESC dataset modelled based on time-points the cells are captured. Using cell specific marker genes, two cell types – extraembryonic endoderm (XEN) cells [violet] and neuro-ectodermal (referred as NEU) cells [green] - were identified. Ambiguous [pink] and unidentified [blue] cells were also observed, requiring further investigation using additional cell-specific genetic markers.

Pseudotime analysis

As these data represent a time series, it is imperative to sort/order the cells according to their time points of collection and cluster them as time series data. Pseudotime analysis of these data can reveal the gene expression patterns underlying differentiation of these cells over time so that the point at which cells truly begin to differentiate into distinct populations can be identified. The pseudotime clustering plot reveals that over the time, cells are differentiating into two distinct cell types as established above (the XEN and NEU lineages, shown in Figure 7). Additional small branched clusters were also evident in this analysis which could suggest the existence of additional subtypes or early commitment of cells to a particular lineage, requiring further investigation.

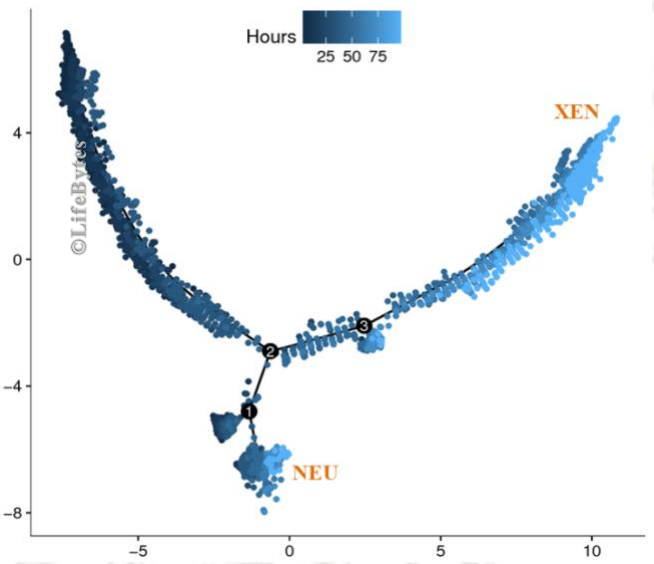

Figure 7. Pseudotime trajectory plot with clustering of cells of mESC data, where cells are sorted and clustered based on their time-point. Cells are coloured based on the time-point from dark blue (2h) to light blue (96h). Around 48h – 60h, the cells seem to commit to a particular lineage, as the trajectory splits into two branches.

Two main branches are observed in the trajectory plot. The branching seems to happen between 48h-60h, and it reflects the commitment of cells to a particular lineage. NEU cells seem to commit and differentiate at early time points based on this plot. This pattern is also observed upon direct examination of marker gene expression patterns, which similarly change over time as illustrated in Figure 8. In the expression plot, it is evident that NEU marker gene expression is higher at intermediate time-points and decreases over the time, whereas XEN marker gene expression steadily increases over the time.

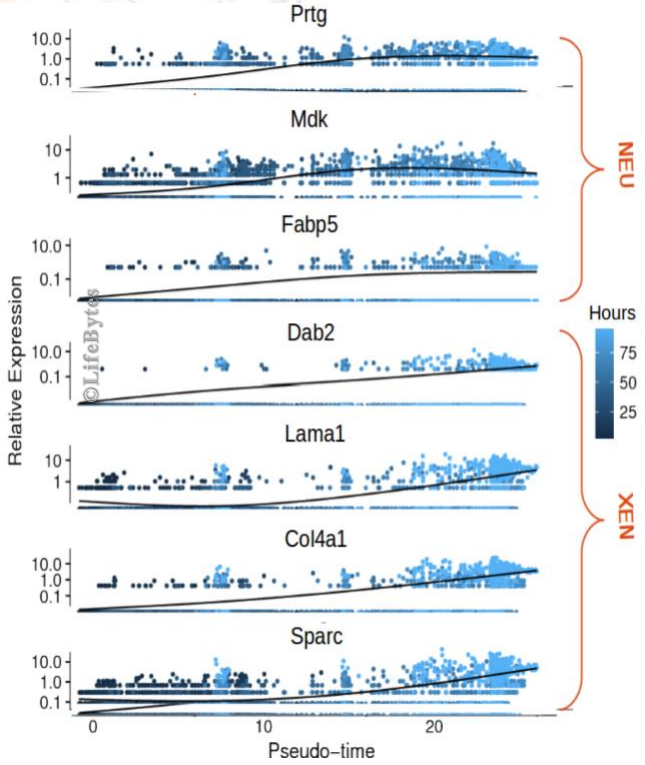

Figure 8. Marker genes expression trajectory plot over time. NEU lineage markers seem to be more highly expressed at intermediate time points (~48h-60h), with subsequent expression decreasing over time. XEN lineage marker genes exhibit steadily increasing expression over time, with higher accumulation towards the final time point.

Cell-cycle stage analysis

In the context of scRNA-seq experiments, the transcriptome data itself can be used to identify the cell cycle stages of individual cells. During the differentiation process, different cells are at different stages of this cycle, and may exhibit different expression patterns even from cells of the same type. Based on the expression of cell-cycle specific marker genes, each cell is scored for G1 and G2M phases. A cell is assigned to be in the G1 phase if the G1 score is above 0.5 and the G2M score is less than the G1 score. Cells with a G2M score higher than 0.5 and a G1 score less than the G2M score are assigned to be in the G2M phase. If both the scores are below 0.5, then the cells are assigned to the S phase[8]. Based on these scores, we have identified the number of cells in these different phases across all times in the mESC data (Figure 9).

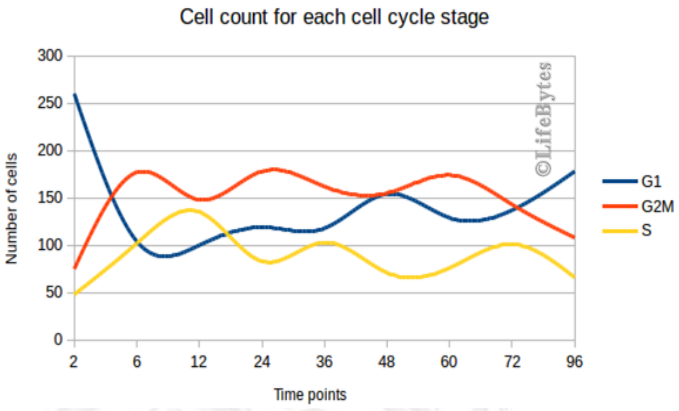

Figure 9. A simple line plot illustrating the numbers of cells at different cell-cycle stages over the course of time.

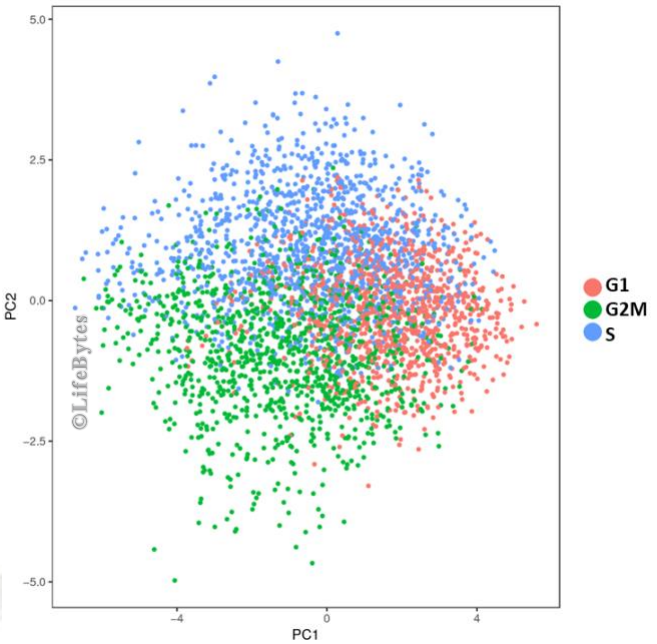

Figure 10. A PCA plot illustration of cells clustered based on their respective cell-cycle stage assignment.

From this analysis it is clear that most cells are initially in the G1 phase, but that as the differentiation process occurs, the G1 frequency reduces, and the G2M and S phase frequencies increase. Clustering of cells based on their cell-cycle stage displays the pattern of cells with different cell-cycle stages (Figure 10).

CONCLUSION

Due to dynamic nature of analytical developments and lack of a well-established pipeline, scRNA-seq analyses demand advanced bioinformatics support. Using this test case analysis, we have highlighted our ability to handle scRNA-seq analyses, taking raw data and producing robust results that are ready to publish. We are constantly exploring novel innovations in this field in order to ensure that our clients receive cutting edge Bioinformatics services. We can handle any types of single cell RNA data provided by your organization efficiently and confidentially. We look forward to the opportunity to work with your organization to enable innovation in single cell-omics.

References

1. Fu, Y., Wu, P.-H., Beane, T., Zamore, P.D. & Weng, Z. *bioRxiv* 251892 (2018).doi:10.1101/251892
2. Bloomberg, J.H., Ji, H., Lun, A.T.L., Mccarthy, D.J. & Marioni, J.C. (2016).doi:10.12688/f1000research.9501.1
3. Poirion, O.B., Zhu, X., Ching, T. & Garmire, L. *Frontiers in Genetics* **7**, (2016).
4. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. *Nature Biotechnology* **33**, 495–502 (2015).
5. Pierson, E. & Yau, C. *Genome Biol.* **16**, 241 (2015).
6. Macosko, E.Z. et al. *Cell* **161**, 1202–1214 (2015).
7. Semrau, S. et al. *Nature Communications* **8**, (2017).
8. Scialdone, A. et al. *Methods* **85**, 54–61 (2015).