Single-Cell Transcriptomic Analysis of Zebrafish Pigment Cells

Diogo Esteves[1][0000−0001−5741−5686] and David Henriques[2][0000−0002−9477−292X]

[1] Department of Informatics, University of Minho, Portugal
pg28935@uminho.pt
[2] CSIC IIM, Vigo, Spain
davidh@iim.csic.es

Abstract. Zebrafish (*Danio rerio*) is a model organism for understanding disease processes and vertebrate development. Their distinct coloring patterns provide valuable insights into genetic regulation and cellular differentiation. This study employs single-cell transcriptomics to analyze the complexity of pigment cell formation in zebrafish, providing new insight into cellular heterogeneity and gene expression. I hope to elucidate the mechanisms involved in pigment cell development by using cutting-edge methods like scRNA-seq and sophisticated analytical approaches such as UMAP and PCA, with a focus on the role of transcription factors and genetic networks. This complete approach is essential for the study of genetic variation and cellular development in this well-studied model species.

1 Introduction

Pigmentation of *Danio rerio* cells could offer valuable insights into genetic and developmental processes. *D. rerio* is a key model organism in biology. Understanding its genetic regulation and cellular differentiation may be visually determined by its distinct pigmentation patterns, which are regulated by several types of pigment cells. Single-cell transcriptomics provides a comprehensive view of cellular heterogeneity and gene expression patterns, beyond standard bulk mRNA sequencing methodologies that fail to capture the cellular complexity and dynamic nature of live tissues [8].

The zebrafish is commonly employed in genetic and developmental biology due to a variety of advantages, including transparent embryos, rapid development, and the capacity to generate a high number of offspring. These features make in vivo imaging and high-throughput genetic screening easier. Zebrafish are an ideal model for researching vertebrate development and disease mechanisms because of their genetic similarity to humans [4,11].

In zebrafish, neural crest (NCs) cells generate pigment cells: melanophores (black), xanthophores (yellow), and iridophores (iridescent). These pigment cells are not only necessary for the striping pattern of zebrafish, but also serve as a model to understand processes of cell differentiation, migration, and survival [11].

The precise arrangement and interaction of these cells is essential for the development of the pigment pattern, making zebrafish an ideal vehicle for researching pigment cell biology. Single-cell RNA sequencing (scRNA-seq) has transformed the study of transcriptomics by allowing gene expression analysis and research at the level of individual cells. This approach provides insights into cellular heterogeneity and gene expression dynamics, allowing for the identification of diverse cell types and states within a tissue. ScRNA-seq has greatly improved our understanding of developmental processes and disease mechanisms at the cellular level [10].

The 10x Genomics Chromium system, have increased the accuracy, precision, and "output" of single-cell studies. These technologies enable the profiling of thousands to millions of cells in a single experiment, resulting in a thorough understanding of cellular diversity and function. Integrating scRNA-seq data with other omics data, such as epigenomics and proteomics, enables a more comprehensive understanding of cellular regulation and interaction networks [16].

Pigmentation in vertebrates is controlled by a complex balance of genetic and environmental variables. Pigmentation in zebrafish is regulated by multiple genes and signaling pathways, including transcription factors like MITF (Microphthalmia-associated transcription factor) and receptors like KIT (stem cell factor receptor). Pigment cells rely on these for development, survival, and function. Understanding the genetic basis of pigmentation provides knowledge in several biological processes, including cell differentiation, development, and illness [7].

Objective

The primary objective of this study is to analyze the transcriptome of zebrafish pigment cells using single-cell RNA sequencing (scRNA-seq). The goal was to elucidate the genetic and molecular factors involved in pigment cell development by employing advanced analytical approaches such as UMAP (Uniform Manifold Approximation and Projection) and PCA (Principal Component Analysis). The study aimed to understand the role of transcription factors and genetic networks in regulating pigment cell differentiation and function. For the proposed schedule before some challenges appeared due to time and equipment limitations. It lewd to multiple changes and attempts in our objective approach, in powerful time-consuming script runs, more explanations in further sections.

2 Background

The zebrafish (*Danio rerio*) is an important model organism in biomedical research because of its genetic closeness to humans and practical advantages, including rapid development, transparency throughout embryonic stages, and high fertility. These traits help investigate developmental processes, genetic roles, and disease models [4].

Zebrafish embryos' transparency enables real-time monitoring of developmental processes, making them ideal for studying cellular dynamics and interactions. Furthermore, zebrafish's rapid development cycle, in which major organ systems arise within days, allows for quick turnaround for experimental experiments, making it a very efficient model organism [11,9].

Zebrafish pigment cells, originating from the NCs, are an excellent model for investigating cell differentiation and pattern generation. The NCs is a transient, multipotent, migrating cell population unique to vertebrates, originating a large variety of cell types, including neurons, glia, and pigment cells [17].

These NCs cells travel throughout the organism and develop into many cell types, including pigment cells, which are required for the zebrafish's distinctive striping pattern. This migration and differentiation process is tightly controlled and involves a complex interaction of genetic and environmental factors[11].

Pigmentation is regulated by a complex network of genes. Key transcription factors, such as MITF (Microphthalmia-associated transcription factor), help melanophores develop and operate by controlling genes involved in melanin synthesis and pigment cell survival. KIT, a receptor tyrosine kinase, promotes pigment cell migration and survival [7,6].

Mutations or alterations in these pathways can induce pigmentation abnormalities and lead to a variety of pigmentation illnesses, highlighting the importance of these components in normal pigment cell growth and function. For example, mutations in the MITF gene can cause Waardenburg syndrome, which is characterized by pigmentary abnormalities and hearing loss [7,5].

Single-cell RNA sequencing (scRNA-seq) revolutionized our understanding of cellular diversity and gene expression dynamics. Unlike bulk RNA sequencing, which averages gene expression patterns over a large number of cells, scRNA-seq offers gene expression data at the individual cell level, revealing tissue heterogeneity. This method enables researchers to detect unusual cell types, reconstruct developmental trajectories, and comprehend the regulatory mechanisms that drive cell differentiation [10]. The ability to study gene expression at the single-cell level has given researchers new insight into the complexities of cellular processes as well as the molecular underpinnings of development and illness.

The introduction of high-throughput scRNA-seq technologies, such as the 10x Genomics Chromium platform, has allowed the profiling of tens of thousands of cells in a single experiment. These improvements have changed our knowledge in complex tissues and developmental processes by recording the transcriptomes of a large number of cells with high resolution and throughput [16,9]. This high-throughput capability is critical for researching tissues such as zebrafish pigment cells, where knowing the diversity and interconnections of various cell types is critical for deciphering pattern formation mechanisms[6,1,2].

Using scRNA-seq on zebrafish pigment cells has revealed important information on the genetic pathways regulating their growth. Studies have identified multiple cell states and developmental pathways within the pigment cell lineage, emphasizing the cells' complexity and adaptability. Single-cell investigations have indicated that zebrafish pigment cells arise directly from persistent

multipotent progenitors, and their differentiation involves dynamic changes in gene expression [17]. These findings highlight the significance of studying the temporal dynamics of gene expression during pigment cell formation.

Integrating scRNA-seq data with other omics data, such as epigenomics and proteomics, yields a more complete picture of cellular function and control. These integrative approaches aid in understanding the complicated networks of gene regulation, signaling pathways, and cellular connections. Combining data from many sources allows researchers to acquire a fuller knowledge of the variables affecting cell differentiation and function [16,6]. For example, combining transcriptome and chromatin accessibility data can indicate how changes in chromatin state affect gene expression and cell fate decisions.

Research into the genetic and molecular basis of zebrafish pigmentation has also revealed various signaling pathways that are essential for pigment cell development. These include the Wnt, FGF, and Notch signaling pathways, which have been demonstrated to influence cell proliferation, differentiation, and survival[1]. Disruptions in these pathways can cause problems in pigment cell growth and pattern creation, offering more information about the complex regulatory networks that control these processes[2].

3 Materials and Methods

The main challenge addressed in this study was to analyze the complexity of pigment cell formation in zebrafish using scRNA-seq data. The analysis aimed to identify key genes, transcription factors, and signaling pathways involved in pigment cell development. Additionally, the study sought to understand the cellular heterogeneity and gene expression patterns within the pigment cell population.

This analysis was performed using the R programming language within RStudio. The environment setup included the following key libraries and tools:
R: The primary programming language used for data analysis.
RStudio: An IDE for R that provides a user-friendly interface for writing scripts, visualizing data, and managing projects.
Monocle: An R package for analyzing single-cell transcriptomics data.
Slingshot: An R package for clustering single-cell transcriptomics data.
Seurat: An R package for single-cell transcriptomics data analysis.

Data Source

This study used data from the GEO series GSE202639, which includes single-cell RNA-seq data from developing zebrafish embryos. This collection contains single-cell transcriptome data from 1812 individually resolved growing zebrafish embryos, with 19 time points and 23 genetic alterations totaling 3.2 million cells. This dataset, called ZSCAPE (Zebrafish Single Cell Atlas of Perturbed Embryos), has generated several research publications, revealing its potential for further detailed analysis [14,3].

Data Preprocessing

The expression matrix, cell metadata, and gene metadata were loaded and validated. To ensure data integrity, the expression matrix underwent dimensional accuracy checks. This step is critical for confirming that the data dimensions align with the expected number of cells and genes.

Dataset	Dimensions	Description
Expression Matrix	20000 x 3200000	Raw counts of gene expression
Cell Metadata	3200000 x 10	Quality control metrics for each cell
Gene Metadata	20000 x 5	Gene annotations and information

Table 1: Summary of Loaded Data

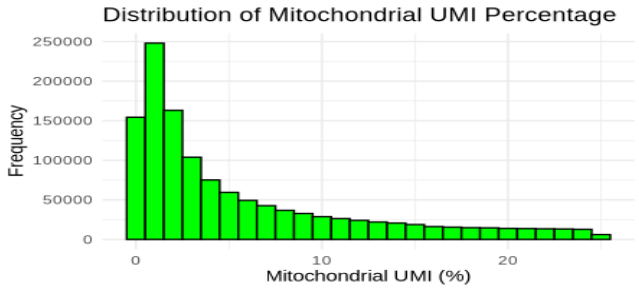


Fig. 1: Distribution of Mitochondrial UMI Percentage: This figure shows the distribution of mitochondrial UMI percentages across all cells. High mitochondrial content can indicate cell stress or damage, and cells with more than 5% mitochondrial content were filtered out to improve data quality.

Quality Control (QC): To assess the quality of each cell QC metrics were calculated. Cells with high mitochondrial content (more than 5%) or low library size (fewer than 500 unique molecular identifiers, UMIs) were filtered out. Genes expressed in fewer than 10 cells were also removed to eliminate noise from lowly expressed genes [12]. This rigorous QC process helps in maintaining the integrity of the data and ensures that subsequent analyses are based on reliable information.

Normalization and Identification of Highly Variable Genes: To stabilize the variance across the dataset, raw counts were normalized using a log transformation. The **scran** package was utilized to identify highly variable genes, which are crucial for distinguishing different cell types. For further analysis, the top

2000 most variable genes were selected. Identifying highly variable genes is essential for focusing on the most informative features of the dataset and disregarding genes that have low or no variability under these conditions [16].

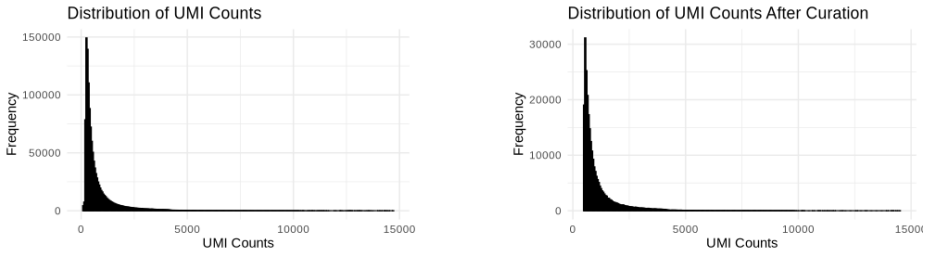


Fig. 2: Distribution of UMI Counts After Curation: This figure displays the distribution of UMI counts per cell after quality control and normalization. It shows that most cells have a reasonable number of UMIs, indicating good library complexity and quality after curation

Dimensionality Reduction and Clustering

Principal Component Analysis (PCA): PCA was performed on the highly variable genes to reduce the dimensionality of the data. PCA identifies principal components that capture the most significant variations in the dataset. This step is crucial for simplifying the data while retaining its essential patterns [10].

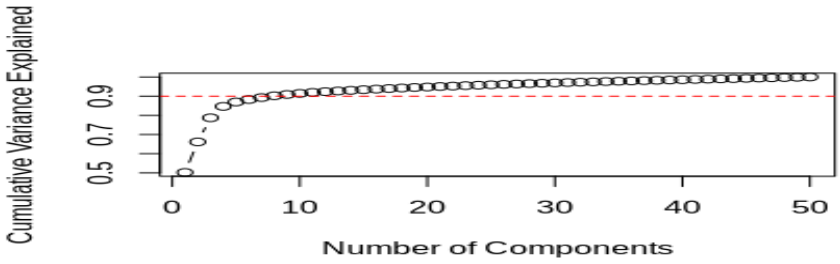


Fig. 3: Cumulative Variance Explained by PCA Components: This figure shows the cumulative variance explained by the first 50 principal components. It helps to determine the number of components to retain for further analysis by showing the proportion of variance captured by each component.

Uniform Manifold Approximation and Projection (UMAP): UMAP was used to visualize the data in two dimensions. UMAP provides a clearer separation of distinct cell populations compared to PCA, making it easier to identify clusters of similar cells [13].

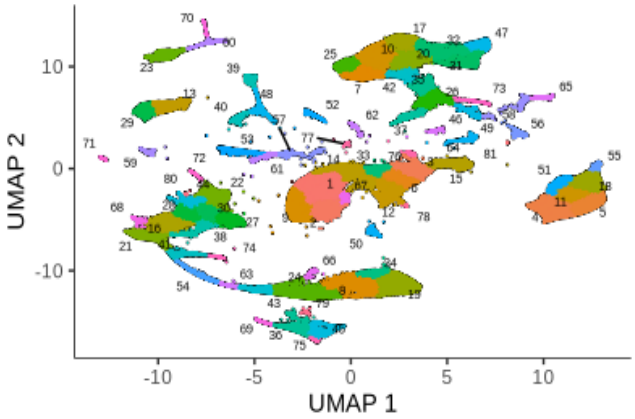
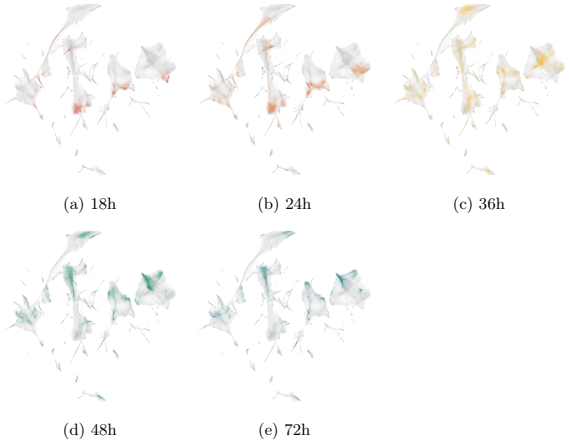


Fig. 4: UMAP Visualization of Zebrafish Pigment Cells: This UMAP plot shows the distribution of zebrafish pigment cells in a two-dimensional space, with each point representing a single cell. The colors represent different clusters, indicating distinct cell populations.

Clustering Analysis: Clustering analysis was conducted to identify distinct cell populations. Clusters were determined based on the principal components from PCA and the UMAP projections. This step involved using algorithms such as Louvain or Leiden for community detection in the data, which help in grouping cells with similar expression profiles [15].

Functional Analysis and Gene Expression

Marker Genes: For each cluster to characterize the cell types. The expression of key genes, including *mitfa*, *kit*, and *csf1ra*, was visualized to validate their roles in pigment cell development. Marker genes are critical for defining the identity and function of each cell cluster [4].



(a) 18h (b) 24h (c) 36h

(d) 48h (e) 72h

Fig. 5: UMAP visualizations of zebrafish pigment cells over time (18h-72h), showing the clustering of cells at different developmental stages. Each cluster represents a group of cells with similar gene expression profiles, indicating different cell types or states. These could generate a time-lapse analysis.

4 Results & Discussion

Quality control was an essential first step to ensure that only high-quality cells were included in the analysis. In Fig. 1 and 2, the distribution of mitochondrial UMI percentages across the cells is show cells with a high mitochondrial content (>5%). These were filtered out, as they often indicate stressed or dying cells, which could introduce bias into the analysis. This step ensured that the dataset only contained viable cells, providing a reliable foundation for downstream analyses as shown in Fig 2 after curation graph.

These distributions highlight the effectiveness of the curation process, ensuring consistent and comparable data across all cells. After this normalization is crucial for accurate downstream analysis, enabling the identification of the most variable and biologically significant genes.

Principal Component Analysis (PCA) was employed to reduce the dimensionality of the data. Figure 3 illustrates the cumulative variance explained by the PCA components, with the first few components capturing most of the variance in the dataset. This step was essential for simplifying the dataset while retaining the key variations. By reducing the complexity of the data, PCA facilitated the identification of major patterns and trends, which are critical for understanding the underlying biological processes.

Uniform Manifold Approximation and Projection (UMAP) was used for visualization, as shown in Figure 4. The UMAP plot reveals distinct clusters of cells based on their gene expression profiles, indicating the presence of different cell types or states within the pigment cell population. This visualization is particularly valuable for identifying subpopulations of cells that may have specific

roles in pigment cell development or function. The clustering analysis, using algorithms such as Louvain or Leiden, confirmed these distinct cell populations, providing a detailed map of cellular diversity.

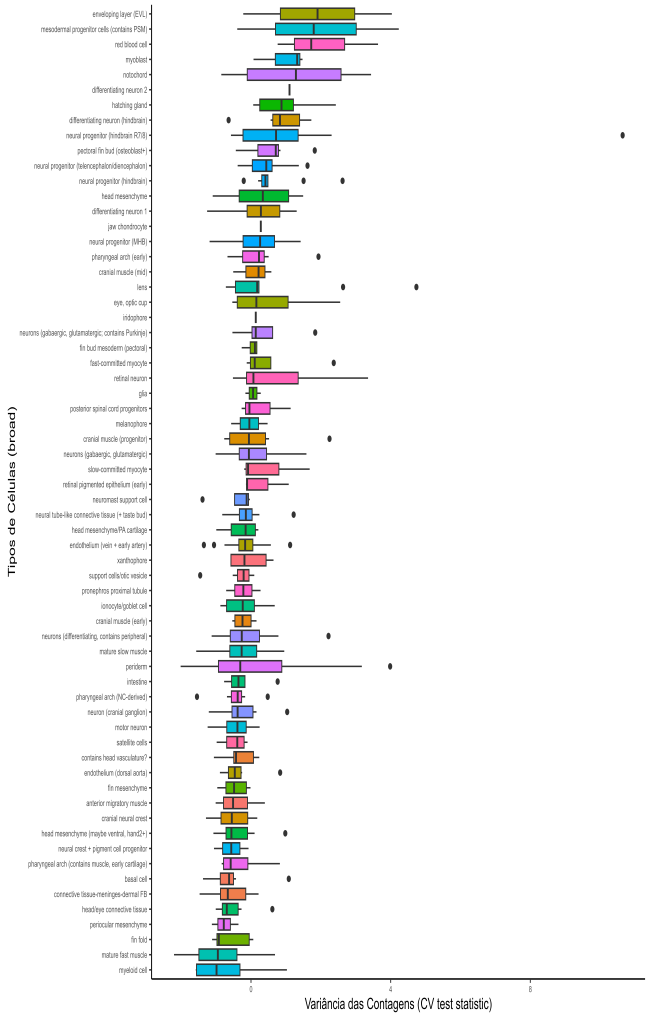


Fig. 6: Box plot of total variance in gene expression. This box plot summarizes the overall variance in gene expression levels across different cell types, providing insights into the heterogeneity within the pigment cell population.

The UMAP visualizations over different time points in Figure 5, were generated from pigment cells positive expressions and demonstrate the clustering of cells at various timepoints. These allow for time-lapse visualizations show

how cell populations evolve over time, providing insights into the dynamics of pigment cell development. For instance, the clustering patterns observed at 18h, 24h, 36h, 48h, and 72h indicate shifts in gene expression as cells differentiate and mature. This temporal analysis is crucial for understanding the progression and regulation of pigment cell development, highlighting key stages where significant changes occur.

High-throughput single-cell transcriptomics has revealed previously unknown details on the cellular diversity of tissues in various organisms. According to Kulkarni *et al.* (2019), single-cell RNA sequencing provides an unbiased insight into the cellular makeup of complex tissues, addressing the stochastic nature of gene expression often overlooked in bulk tissue studies [8].

Trajectory inference is a crucial aspect of single-cell RNA-seq analysis. Qiu *et al.* (2017) presented Monocle 2, an algorithm utilizing reversed graph embedding to express numerous fate decisions in an unsupervised manner, effectively resolving complex single-cell trajectories [13]. This method has been instrumental in studies of blood formation and other biological processes.

Patterson & Parichy's (2019) examination of zebrafish pigment pattern creation revealed the arrangement and interaction of melanophores, xanthophores, and iridophores for zebrafish stripe formation. Genetic factors such as transcription factors *mitfa*, *tfec*, and receptor tyrosine kinases *kita* and *ltka* are essential for pigment cell differentiation [11].

Saunders *et al.* (2023) conducted a comprehensive single-cell transcriptome investigation of developing zebrafish embryos, resulting in a high-resolution map of cell type compositions and developmental trajectories. Their dataset, which includes 1812 embryos and 3.2 million cells, allows for the estimation of cell type abundance variation and discovery of perturbation-dependent variances in cell type composition. This robust dataset provides a framework for understanding the genetic connections and developmental processes of zebrafish pigment cells [14].

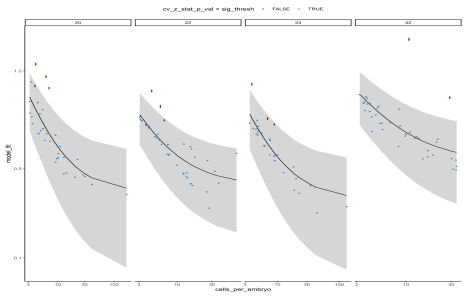


Fig. 7: Variance of gene expression over different time points. This plot shows the variance in gene expression for various genes across different developmental stages, highlighting the dynamic nature of gene regulation during pigment cell development.

The previous figures in this section illustrate the key findings from replicating Saunders *et al.* (2023) processed code, datasets and tables and the last image was generated by me after isolating a small subsample of transcription factors over time, and it was very memory intensive as well.

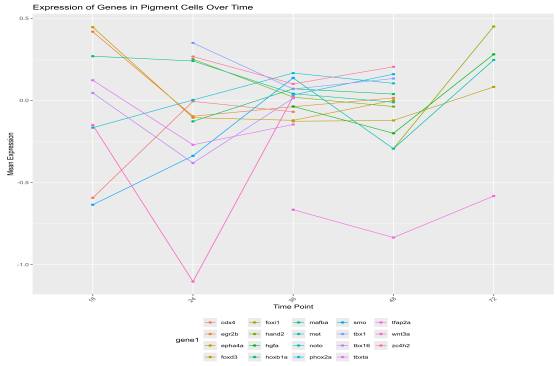


Fig. 8: Expression of transcription factors in pigment cells over time. This graph shows the mean expression levels of various genes involved in pigment cell development across different time points, highlighting dynamic changes in gene expression.

In Fig.8 we can see that a few transcription factors stand out for their peculiar expression behavior over time. These would be great targets for future single cell transcriptomics analysis.

5 Conclusion

This study utilized high-quality single-cell RNA-seq data to understand the zebrafish pigment cell transcriptome better. Initial analysis steps, including data processing and normalization, were completed, but the full analysis faced limitations due to computational constraints. Nonetheless, the study provided valuable preliminary insights into the genetic regulation of pigmentation and highlighted the importance of adequate computational resources for single-cell RNA-seq analysis. Future work could involve experimental validation of these findings and the integration of additional datasets to create a more comprehensive perspective.

6 Challenges & Future Work

Given these challenges, only a limited number of results were obtained, including a few visualizations of QC metrics, UMI counts, PCA, and UMAP.

Despite the challenges and limitations, the analysis provided some preliminary insights into the cellular heterogeneity and gene expression patterns of zebrafish pigment cells. The quality control steps ensured that only high-quality cells were included in the analysis, and the normalization process allowed for accurate comparison of gene expression levels across cells.

The single-cell transcriptome investigation of zebrafish pigment cells aimed to reveal cellular heterogeneity and gene expression patterns, focusing on key regulators of pigment cell development.

Several challenges appeared:

Computational Challenges: The large data volume required significant computational resources, leading to memory limitations and long script execution times.

Resource Limitations: Despite remote access to a better computer in the final weeks, script processing times remained long (2 to 24 hours). The dataset was not ideal for available resources, as single-cell RNA sequencing analyses require extensive computational power and memory.

To address these challenges:

Enhanced Resources: Use a high-performance computing cluster.

Optimized Scripts: Refine scripts for better performance.

Extended Timeframe: Allocate more time for analysis.

Experimental Validation: Validate findings experimentally.

Additional Datasets: Incorporate more datasets for a comprehensive understanding of zebrafish pigment cell development.

In conclusion, the study highlighted the complexities and demands of single-cell RNA sequencing. Future work with enhanced resources and optimized approaches is essential to achieve the study's objectives.

Website

To consult detailed information about this project, R scripts used, or other detailed information, visit this website.

Acknowledgements

I am grateful to David Henriques of CSIC IIM in Vigo and Professor Miguel Rocha of the University of Minho in Braga for the chance and collaboration on this research project. I also want to thank my family, friends, and colleagues for their useful advices and support over the first year of my Masters degree in Bioinformatics, as this is my final course during this academic year.

References

1. Budi, E.H., Patterson, L.B., Parichy, D.M.: Embryonic requirements for erbb signaling in neural crest development and adult pigment pattern formation. Development (2008)
2. Camargo-Sosa, K., Colanesi, S., Müller, J., Schulte-Merker, S., Stemple, D., Patton, E.E., Kelsh, R.N.: Endothelin receptor aa regulates proliferation and differentiation of erb-dependent pigment progenitors in zebrafish. PLoS genetics **15**(2), e1007941 (2019)
3. Dorrity, M.W., Saunders, L.M., Duran, M., Srivatsan, S.R., Barkan, E., Jackson, D.L., Trapnell, C.: Proteostasis governs differential temperature sensitivity across embryonic cell types. Cell **186**(23), 5015–5027 (2023)
4. Howard IV, A.G., Baker, P.A., Ibarra-García-Padilla, R., Moore, J.A., Rivas, L.J., Tallman, J.J., Singleton, E.W., Westheimer, J.L., Corteguera, J.A., Uribe, R.A.: An atlas of neural crest lineages along the posterior developing zebrafish at single-cell resolution. Elife **10**, e60005 (2021)
5. Huang, S., Song, J., He, C., Cai, X., Yuan, K., Mei, L., Feng, Y.: Genetic insights, disease mechanisms, and biological therapeutics for waardenburg syndrome. Gene therapy **29**(9), 479–497 (2022)
6. Jang, H.S., Chen, Y., Ge, J., Wilkening, A.N., Hou, Y., Lee, H.J., Choi, Y.R., Lowdon, R.F., Xing, X., Li, D., et al.: Epigenetic dynamics shaping melanophore and iridophore cell fate in zebrafish. Genome biology **22**, 1–18 (2021)
7. Kenny, C., Dilshat, R., Seberg, H.E., Van Otterloo, E., Bonde, G., Helverson, A., Franke, C.M., Steingrímsson, E., Cornell, R.A.: Tfap2 paralogs facilitate chromatin access for mitf at pigmentation and cell proliferation genes. PLoS genetics **18**(5), e1010207 (2022)
8. Kulkarni, A., Anderson, A.G., Merullo, D.P., Konopka, G.: Beyond bulk: a review of single cell transcriptomics methodologies and applications. Current opinion in biotechnology **58**, 129–136 (2019)
9. Lencer, E., Prekeris, R., Artinger, K.B.: Single-cell rna analysis identifies premigratory neural crest cells expressing markers of differentiated derivatives. Elife **10**, e66078 (2021)
10. Nayak, R., Hasija, Y.: A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines. Genomics **113**(2), 606–619 (2021)
11. Patterson, L.B., Parichy, D.M.: Zebrafish pigment pattern formation: insights into the development and evolution of adult form. Annual Review of Genetics **53**, 505–530 (2019)
12. Phipson, B., Sim, C.B., Porrello, E.R., Hewitt, A.W., Powell, J., Oshlack, A.: propeller: testing for differences in cell type proportions in single cell data. Bioinformatics **38**(20), 4720–4726 (2022)
13. Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., Trapnell, C.: Reversed graph embedding resolves complex single-cell trajectories. Nature methods **14**(10), 979–982 (2017)
14. Saunders, L.M., Srivatsan, S.R., Duran, M., Dorrity, M.W., Ewing, B., Linbo, T.H., Shendure, J., Raible, D.W., Moens, C.B., Kimelman, D., et al.: Embryo-scale reverse genetics at single-cell resolution. Nature **623**(7988), 782–791 (2023)
15. Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., Dudoit, S.: Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC genomics **19**, 1–16 (2018)

16. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R.: Comprehensive integration of single-cell data. cell **177**(7), 1888–1902 (2019)
17. Subkhankulova, T., Camargo Sosa, K., Uroshlev, L.A., Nikaido, M., Shriever, N., Kasianov, A.S., Yang, X., Rodrigues, F.S., Carney, T.J., Bavister, G., Schwetlick, H., Dawes, J.H.P., Rocco, A., Makeev, V.J., Kelsh, R.N.: Zebrafish pigment cells develop directly from persistent highly multipotent progenitors. Nature Communications **14**(1), 1258 (2023)