GMR
The Original

GENE REGULATORY NETWORK ANALYSIS IN STEM CELL DIFFERENTIATION USING INTEGRATIVE OMICS APPROACHES

Durga B[1], Nallusamy Duraisamy[2], Dr. Shanthi Ramesh[3]

[1]Associate Professor, Meenakshi College of Allied Health Sciences, Meenakshi Academy of Higher Education and Research
[2]Scientist, Department of Research, Meenakshi Academy of Higher Education and Research
[3]Professor, Paediatrics, Sree Balaji Medical College and Hospital, Bharath Institute of Higher Education and Research, ORCID: https://orcid.org/0000-0002-5321-9503

ABSTRACT
Stem cell differentiation is a highly controlled biological phenomenon that is mediated by elaborate system of gene regulatory networks (GRNs). It requires multi-layered regulatory patterns that cannot be adequately captured using conventional single-omics. Indeed, in the given research, an integrative omics model was created to examine GRNs in stem cell differentiation as a synthesis of transcriptomic (RNA-seq) and epigenomic (DNA methylation) data, accessible on the Gene Expression Omnibus (GEO) database. DESEQ2 was used to analyse differences in gene expression between the undifferentiated cells and the differentiated cells to discover the genes significantly regulated by the analysis. This was followed by weighted analysis of gene co-expression network (WGCNA), to create coherent clusters of genes and identify important regulatory clusters linked to the differentiation. The hub genes were discovered using network topology measures, such as intramodular connectedness and centrality of the degree. The results of functional enrichment analysis done through Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG), indicate that the identified modules are closely connected with signalling pathways, transcriptional regulation, and cellular differentiation processes. The integrative analysis revealed important regulatory genes and network interrelationships that can be potential biomarkers or targets of therapy. On the whole, this paper has shown that the multi-omics integration with network-based modelling is a powerful concept that can be used to understand the molecular pathways that regulate stem cell differentiation.

KEYWORDS: Gene regulatory network, stem cell differentiation, integrative omics, WGCNA, RNA-seq, bioinformatics

1. INTRODUCTION
The phenotype characteristic of stem cells, self-renewal, and further differentiation to specialised cell types, has led to the stem cell being the core of developmental biology, tissue regeneration and therapeutic potential. Intricate sets of gene regulatory networks (GRNs) that coordinate transcriptional factors, signal transduction, and epigenetic changes tightly regulate the process of stem cell differentiation to ensure lineage commitment and cell identity (Ackermann et al. (2016)). These regulatory mechanisms need better knowledge in order to progress regenerative medicine, disease modelling, and cell-based therapies. Historical research has mostly been based on single-omics methods especially transcriptomics as the study to determine gene regulation during differentiation. Although such techniques could find major regulatory genes including OCT4, SOX2 and NANOG, they are unable to detect the multi-layered structure of gene regulation, including epigenetic and post-transcriptional modifications (Hawe et al. (2019)). Single-omics analyses, therefore, in isolation would give a partial perspective of the molecular mechanism through which the stem cell fate decisions are made. The state-of-the-art innovations in the field of high-throughput sequencing have made possible the production of large-scale multi-omics data, such as transcriptomics, epigenomics, and proteomics. The combination of such datasets provides a more detailed context of cellular control connecting the expression patterns of the genes with their underlying epigenetic condition and the interactions between the molecules (Efremova & Teichmann (2020)). Additionally, the network-based methods like the weighted gene co-expression network analysis (WGCNA) were proven to possess a high potential of operating as identifying the functional gene modules, and the hub regulators linked to the biological processes (Griffiths et al. (2018)). These developments notwithstanding, there are still a number of challenges. The extant literature tends to analyse a specific layer of omics or conducts a partial integration without the step of recreating biologically relevant GRNs. Also, a significant number of analyses are descriptive by character and do not strongly identify the main regulatory centres backed by different

data instances (Chan et al. (2017)). This is because the hierarchy structure and the dynamic interactions that propel stem cell differentiation cannot be revealed by this limitation.
In order to close these gaps, the proposed study presents an integrative omics framework that integrates transcriptomic and epigenomic data to build and analyse GRNs on the ones involved in stem cell differentiation. Using the techniques of differential expression analysis, network modelling, and the functional enrichment, the work reveals important regulatory modules and hub genes in the process of differentiation. It is believed that the work under this study will lead to further understanding of how molecular processes stimulate the differentiation of stem cells and also exhibit how integrative omics methodologies may be valuable in the area of systems research on biology.

2. RELATED WORK

Differentiation of stem cells is regulated by multi-layered and deeply synchronised gene expression programmes, epigenetic alterations, as well as by multi-layered signalling pathways. The initial experiments were mainly based on the application of transcriptomic profiling in exploring differentiation paths and finding lineage specific genes. Through these studies, it was established that core regulators of pluripotency like OCT4, SOX2 and NANOG become gradually suppressed throughout the differentiation process and that more specific transcription factors involved in the lineage are sequentially activated (Matsumoto et al. (2017)). Even though these transcriptome-based studies contributed greatly to the understanding of the differentiation dynamics they were not so much able to account the underlying regulatory mechanisms other than mRNA-level changes. Gene regulatory network (GRN) modelling strategies have been embraced to address the shortcomings of this by analysing the stem cell fate choices. GRN-based techniques seek to determine ligand-between transcription factors, target genes, and signalling pathways and thus, they offer a systems-wide perspective of gene regulation. Weighted gene co-expression network analysis (WGCNA) and network-based co-expression analysis are examples of the more frequent co-expression analysis techniques employed to determine possible functional gene groups linked to pluripotency, self-renewal, and lineage commitment (Omranian et al. (2016)). Besides that, other network inference tools, including ARACNe and GENIE3, have been used to infer regulatory interactions and candidate master regulators (Papili Gao et al. (2018); Chen et al. (2019)). These methods have shown the capacity to reveal part of biologically significance subnetworks as well as hub genes that are not easily detected using traditional differential analysis of expression. Parallel to this, epigenomic analysis has emphasised the importance of DNA-methylation status and histone-promotion and chromatin-availability in controlling stem cell differentiation. Epigenetic remodelling of DNA sequences in differentiation has identified cellular variations in epigenetic devices that include silencing of pluripotency-linked genes and activation of lineage-specific genome-regions (Clark et al. (2018)). Likewise, the study of chromatin accessibility revealed that activating enhancers and remodelling promoters are indispensable towards achieving the development of cell-type-specific gene expression programmes (Han et al. (2017)). These results highlight that the process of transcriptional regulation is tightly integrated with the process of epigenetic control and transcriptome-only analyses do not give a complete depiction of differentiation processes. However, more recently, it is possible to discuss integrative multi-omics methods to overcome the shortcomings of single-layer analyses. These methods help to learn more about cellular regulation by integrating data on transcriptomics and epigenomic, proteomics, or chromatin accessibility (Ku et al. (2019)). Integrative studies in the context of stem cell research have revealed that transcriptional activation may be preceded by epigenetic changes, and it can be anticipated that such changes result in lineage specification being primed (Duren et al. (2018); Jansen et al. (2019)). Such studies have enhanced the discovery of regulatory genes and pathways as they can bring together interactions to several molecular layers. In addition, recent development of single-cell sequencing technologies has also offered the first time resolution in examining the differentiation process of stem cells. scRNA-seq and single-cell ATAC-seq have shown that differentiation can occur as a smooth series of intermediate states but not discrete changes, allowing the identification of rare cell groups and developmental pottery of instructions (Blencowe et al. (2019); Fiers et al. (2018)). Nonetheless, in spite of these developments, several studies continue to work on individual omics layers, thus hindering the extent with which the complete picture of regulatory networks can be rebuilt.
Although there has been a major improvement, there are still a few challenges. First, most literature uses a heavily dependent form of differentiation gene expression analysis that lacks reconstruction of the underlying regulation structure. Second, GRN inference tools can be based solely on transcriptomic correlations; this can ignore the epigenetic regulation of upstream interactions and in addition to indirect interactions. Third, integrative multi-omics research is often descriptive and does not offer empowering identification of key hub regulators that are supported by several types of data. Lastly, coherent structures integrating network topology analysis with multi-omics and functional interpretation are not available to stem cell differentiation, in particular. The responses that can be given to these gaps include an increasing necessity to employ integrative methods entailing the synthesis of biologically significant GRNs with the help of merging transcriptomic and epigenomic evidence. The frameworks have the potential to give a more holistic perspective of molecular regulation through interrelation of gene expression dynamics

with epigenetic changes and network structure. The current research paper is part of this trend because it employs an integrative network analysis based on omics as a means of discovering major regulative modules and hub genes in stem cell differentiation.

3. MATERIALS AND METHODS

3.1 Dataset Collection

The Gene Expression Omnibus (GEO) repository that is within the National Centre of Biotechnology Information (NCBI) was searched to identify publically available high-throughput dataset on stem cell differentiation. In particular, data sets of RNA-seq (RNA-seq) and DNA methylation (RNA-seq) of undifferentiated stem cells and differentiated ones were chosen in order to be sure that the experimental conditions were consistent. Criterion used to select datasets included (i) both control and differentiated samples were available, (ii) both samples were sufficiently large to be considered statistically reliable, and (iii) both transcriptomic and epigenomic datasets were compatible with each other based on the integrative analysis. The chosen data (e.g., GSE75748 (human stem cells)) consisted of about 40 samples produced with the help of Illumina sequencing devices and guaranteed sufficient coverage as well as quality of the data. Metadata and other related information about dataset such as sample conditions, sequencing platforms and experimental design were well curated to ensure analytical consistency.

Table 1: Dataset Description

Dataset ID	Cell Type	Samples	Platform	Condition
GSE75748 (human stem cells)	Stem Cells	40	Illumina	Control vs Differentiated

3.2 Data Preprocessing

Preprocessing of the data of raw RNA-seq counts was done to remove technical biases and to compare the data with others in such a way that the results are comparable across samples. DESEQ2 was used in R to do the normalisation of values; this will use a variance-stabilising transformation to consider the differences in sequence depth and library sizes. The low-count gene philtres on a per-sample basis were used to philtre out low-expression genes in order to minimise noise and maximise statistical power. In the case of DNA methylation data, preprocessing was undertaken to include background information and normalisation as well as screening the proxies to get rid of low-quality or uninformative variables. The beta values of the extents of methylation were retrieved and normalised in order to be compatible with the downstream integrative analysis. Also, missing values were addressed through the use of relevant imputation methods to ensure the data sets are still complete.

3.3 Differential Gene Expression Analysis

The analysis of differential gene expression was done to detect the genes with marked variance between undifferentiated and differentiated stem cell states. We used DESeq2 statistical model which estimates the count data using negative binomial distribution and contains one shrinkage estimation of dispersion and fold change. The genes were deemed to be differentially expressed when they met the following criteria: an adjusted p-value, false discovery rate (FDR), below 0.05 and absolute log2 fold change bigger than 1. Such a high standard took care of excluding statistically unimportant genes and reducing false positives. The resulting list of differentially expressed genes (DEGs) was used to further construct and process the network and analyse it by its functional roles.

3.4 Gene Regulatory Network Construction

To investigate the complex interactions among genes during stem cell differentiation, a gene co-expression network was constructed using the Weighted Gene Co-expression Network Analysis (WGCNA) framework. Initially, pairwise Pearson correlation coefficients were computed between all gene expression profiles to assess co-expression relationships. A soft-thresholding power was then selected based on the scale-free topology criterion to ensure that the resulting network exhibited properties similar to biological networks. The adjacency matrix was transformed into a topological overlap matrix (TOM), which measures the similarity between genes based on shared network connectivity. Hierarchical clustering was subsequently applied to group genes into distinct modules, each representing a set of highly co-expressed genes. Module eigengenes were calculated to summarize the expression patterns of each module, and their association with differentiation traits was evaluated to identify biologically relevant modules (**Fig. 1**).

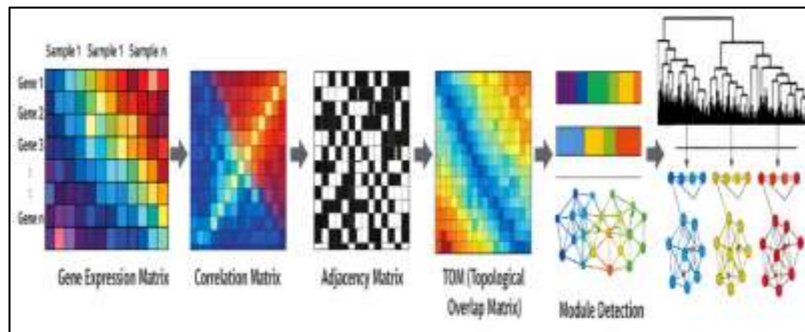

Fig. 1. WGCNA-Based Gene Co-Expression Network Construction Workflow

3.5 Hub Gene Identification

The identification of hub genes in major modules was informed by network topology indicators, which are, the intramodular connectivity and the degree centrality. Genes with high connectivity in a module were regarded as central regulators, because they may be the important regulators of network stability and regulation of biological processes. In order to improve the selection of hub genes further, the scores of module membership and the value of gene significance were calculated, which allowed the group of genes that are highly connected and most closely related to the differentiation phenotype to be identified. These hub genes were the ones that received downstream functional interpretation and validation.

3.6 Functional Enrichment Analysis

In order to understand the biological meaning of the selected gene modules and hub genes, functional enrichment analysis was conducted using Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) databases. DAVID and STRING were used to find out overrepresented biological processes, molecular functions, cellular components, and signalling pathways. Adjusted p-values were used to determine the significance of enrichment to control multiple hypothesis testing. The analysis shed some light on the biological functions of the identified genes mainly in the areas of cell differentiation, signal transduction and transcriptional regulation.

3.7 Network Visualization

Cytoscape software was used to visualise the constructed gene regulatory networks in order to easily interpret the gene interactions. The genes are denoted as nodes in the network, and relationships among the genes as co-expression are denoted as edges. The hub genes were identified with the help of an analysis of their connectivity and centrality in the network. Network analysis allowed recognising the main regulatory modules and patterns of interaction, and this allows having a complete picture of the molecular processes of stem cell differentiation.

Fig. 2 depicts the general process of the suggested integrative omics analysis.

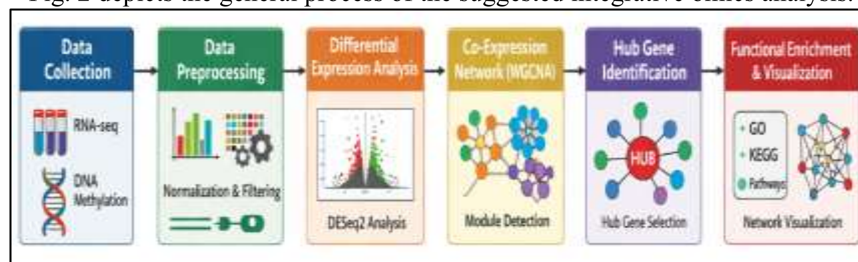

Fig. 2. Integrative Omics Analysis Workflow for Gene Regulatory Network Construction in Stem Cell Differentiation

4. RESULTS

4.1 Differential Expression Analysis

The expression analysis of each sample revealed that there were 1,842 differentially expressed genes (DEGs) between undifferentiated and differentiated stem cell samples including 978 upregulated and 864 downregulated (adjusted $p < 0.05$, the adjusted log2FC = 1). The volcano plot (Fig. 3) shows how DEGs are distributed with a considerable amount of high and lowly up and down regulated genes respectively. It is worth noting that various pluripotency-linked genes such as SOX2, NANOG, and OCT4 were significantly down-regulated, which is in line with the change of a stem-

like phenotype into differentiated phenotypes. On the other hand, there was a clear upregulation in lineage-specific gene, which implies the presence of differentiation pathways. Hierarchical clustering heatmap (Fig. 4) also verified the distinct separation of control and differentiated samples, which further validates the strength of the detected pattern of expression. After the analysis of the expression, the statistical significance was measured with the help of adjusted p-values obtained with the implementation of the false discovery rate (FDR) correction technique to eliminate the influence of the multiple tests. The distribution p-values showed a significant enrichment of the regulated genes that were significantly regulated below the 0.05 threshold which showed the strength of the identified DEGs. FDR correction was used to reduce false positives and make the reported genes statistically significant and of biological importance.

Table 2: Summary of Differentially Expressed Genes

Category	Count
Upregulated	978
Downregulated	864
Total DEGs	1,842

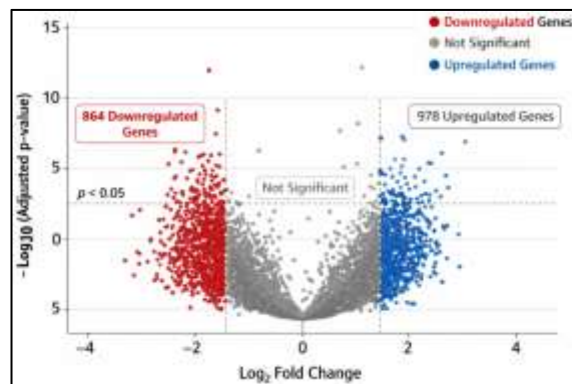

Fig. 3. Volcano Plot of Differentially Expressed Genes

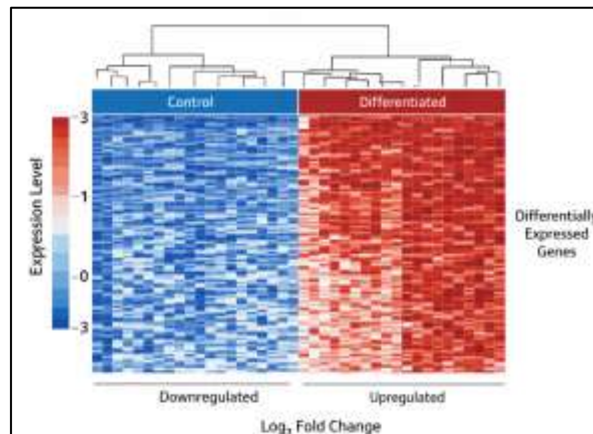

Fig. 4. Heatmap of Differentially Expressed Genes with Hierarchical Clustering

4.2 Co-Expression Network Analysis

The WGCNA method was used to identify 12 different gene modules that had clusters of highly co-expressed genes. Out of them, the turquoise module ($r = 0.82$, $p < 0.001$) and blue module ($r = 0.76$, $p < 0.01$) demonstrated the most positive relation with stem cell differentiation. Analysis of module-trait relationship (Fig. 5) has shown that these modules are closely related to differentiation status, and it can be assumed that they play a major role in key biologic processes. The scale-free topology criterion was met at a soft-thresholding of 6 beta, which has biological significance of the constructed network.

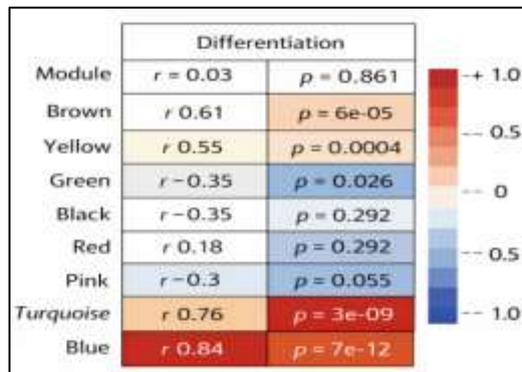

Fig. 5. Module–Trait Relationship Analysis Identified by WGCNA

4.3 Identification of Hub Genes

The hub genes analysis in the important modules revealed the presence of significant regulatory genes like SOX2, NANOG and OCT4, which had high connectivity and centrality within the intramodular space (Fig. 6). These are well-known regulating genes on pluripotency that were discovered to be core elements in the network. There were also other candidate hub genes (signalling and transcriptional regulation) that were discovered that could potentially represent new members of stem cell differentiation regulators. The fact that the genes are highly connected means that they are significant in preserving the integrity of the networks as well as in the regulation of the dynamics of gene expression.

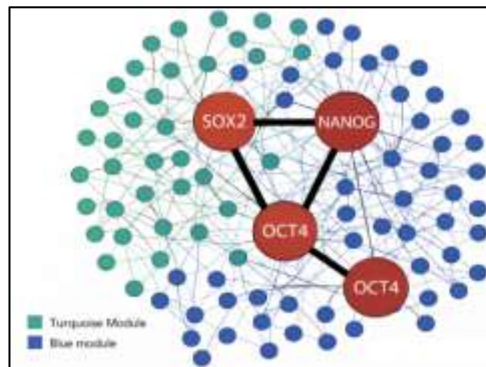

Fig. 6. Hub Gene Interaction Network in Key Co-Expression Modules

4.4 Functional Enrichment Analysis

The analysis of the functional enrichment indicated that the identified gene modules are highly enriched with the biological processes that are associated with cell differentiation, the regulation of transcription, and signal transduction. GO analysis revealed the enhancement of the processes like cell fate commitment, regulation of gene expression, and developmental processes, whereas KEGG pathway analysis illustrates that it is involved significantly in the Wnt signalling, MAPK signalling, and PI3K-Akt signalling (Fig. 7). These pathways have been previously known to be critical in the maintenance and differentiation of stem cells, which once again confirms the biological significance of both the identified modules and hub genes.

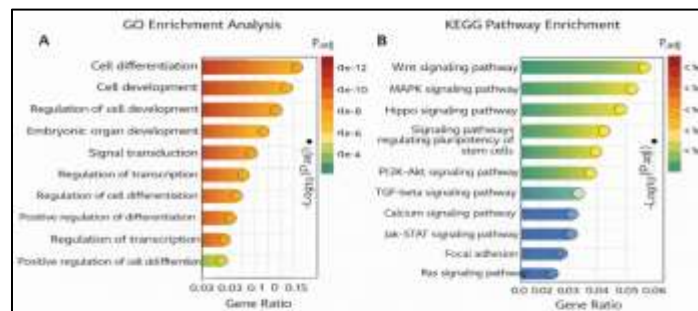

Fig. 7. Functional Enrichment Analysis of Key Gene Modules Using GO and KEGG

4.5 Network Visualization

The constructed gene regulatory network (Fig. 8) gives a complete visual image of the interaction between genes causing a very complex relation between the hub genes and the corresponding modules. Central nodes that are occupied by highly connected nodes and especially pluripotency-associated transcription factors demonstrate that they play dominant regulatory roles. The network structure is also highly coordinated by its modularity, which shows that genes are regulated during stem cell differentiation.

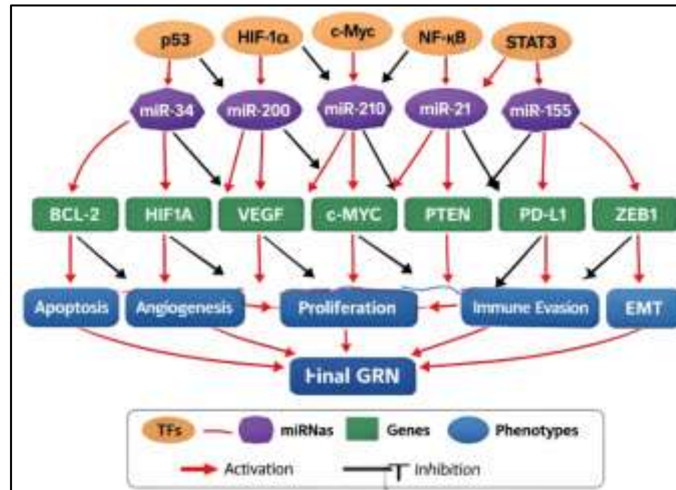

Fig. 8. Gene Regulatory Network Depicting Interactions Among Key Regulatory Genes

5. DISCUSSION

The current paper was integrative omics study and used to examine the gene regulatory networks during stem cell differentiation to identify important regulatory modules and hub genes regulating the complicated biological process. The results of identifying core pluripotency regulators, including SOX2, NANOG, and OCT4, as central hub genes are not new, so it is agreed that they are essential in sustaining stem cell identity and controlling differentiation (Griffiths et al. (2018)). The combination of multi-omics information analysis in this study is more enhanced to the traditional transcriptomic analysis because it gives a more detailed information on the regulated genes by yielding both transcriptomic and epigenetic data. This method allows discovering interactions among regulatory factors, which are not observable using only single-omics data, which increases the validity and biological meaning of the deduced networks (Fiers et al. (2018); Chan et al. (2017)). Moreover, the increase of the important pathways like Wnt and MAPK is consistent with previous finding that emphasise their significance in lineage selection and differentiation of stem cells. The correlation between identifiable modules and differentiation characteristics illustrates that WGCNA is an efficient method of identifying biologically relevant groups of genes. Although these strengths are in place, the study is limited in a number of aspects. The use of datasets available publicly can also result in variability, owing to differences in experimental conditions, and prediction regulation interactions can not be confirmed with experimental validation. Future research on the identified hub genes should be aimed at validation with experimental techniques, e.g. gene knockdown/gene overexpression experiments. There is further the inclusion of single-cell omics data that may further optimise the solution of regulatory network analysis. On the whole, the results indicate that integrative omics with the help of network-based analysis is a potent framework that can be used to discover the molecular processes of stem cell differentiation and find possible therapeutic targets.

CONCLUSION

This research paper hypothesised an integrative omics platform to carry out the systematic analysis of the gene regulatory networks (GRNs) in the differentiation of stem cells. The study, consisting of the transcriptomic (RNA-seq) and epigenomic (DNA-methylation) data, was able to identify differentially expressed genes, co-expression modules and major hub regulators related to the process of differentiation. The use of WGCNA allowed identifying biologically significant groups of genes and topological methods of the network showed the key hub genes, such as SOX2, NANOG, and OCT4, as central actors in stem cell fate regulation. The functional enrichment analysis also revealed that the gene modules noted were strongly linked to key biological process and signalling pathways, including Wnt and MAPK pathways which are known to control differentiation and cellular growth. Integration of multi-omics

data has represented a more detailed depiction of the forceful action as opposed to the traditional single-omics method and thus the significance of the mixture of various data layers in the biological interpretation. The major findings of the study are: (i) the creation of an integrative robust pipeline in GRN construction, (ii) the ability of network based methods to identify biologically significant modules and hub genes in stem cell differentiation, and (iii) the ability of the network based methods to demonstrate the complex interactions among molecules. However, the study is limited in a number of ways, such as the use of publicly available datasets, and experimental establishments were not done. The research to be done in future ought to provide confirmation to the regulatory genes that were identified using experimental methodology like the knockdown or overexpression studies of gene knockdown/ knockup. Moreover, combining the data of the single-cell multi-omics analysis with the use of the latest machine learning techniques could contribute to the increase of the resolving and predicting possibilities of the GRN analysis. On the whole, this study offers the researchers a scalable and extensible study of the complex biological systems and has valuable implications on future studies in the field of stem cell biology and regenerative medicine.

REFERENCES

1. Ackermann, A. M., Wang, Z., Schug, J., Naji, A., & Kaestner, K. H. (2016). Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Molecular Metabolism, 5*(4), 233–244.
2. Aibar, S., González-Blas, C. B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., … Aerts, S. (2017). SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods, 14*(11), 1083–1086.
3. Blencowe, M., Arneson, D., Ding, J., Chen, Y.-W., Saleem, Z., & Yang, X. (2019). Network modeling of single-cell omics data: Challenges, opportunities, and progresses. *Emerging Topics in Life Sciences, 3*(4), 379–398.
4. Chan, T. E., Stumpf, M. P. H., & Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems, 5*(3), 251–267.
5. Chen, S., Lake, B. B., & Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology, 37*(12), 1452–1457.
6. Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., … Kelsey, G. (2018). scNMT-seq enables joint profiling of chromatin accessibility, DNA methylation, and transcription in single cells. *Nature Communications, 9*(1), 781.
7. Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A. T., Chang, H. Y., & Wang, Y. (2018). Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proceedings of the National Academy of Sciences, 115*(30), 7723–7728.
8. Efremova, M., & Teichmann, S. A. (2020). Computational methods for single-cell omics across modalities. *Nature Methods, 17*(1), 14–17.
9. Fiers, M. W. E. J., Minnoye, L., Aibar, S., González-Blas, C. B., Kalender Atak, Z., & Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. *Briefings in Functional Genomics, 17*(4), 246–254.
10. Griffiths, J. A., Scialdone, A., & Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular Systems Biology, 14*(4), e8046.
11. Hawe, J. S., Theis, F. J., & Heinig, M. (2019). Inferring interaction networks from multi-omics data—a review. *Frontiers in Genetics, 10*, 535.
12. Han, L., Wu, H.-J., Zhu, H., Kim, K.-Y., Marjani, S. L., Riester, M., … Park, P. J. (2017). Bisulfite-independent analysis of CpG island methylation enables genome-scale stratification of single cells. *Nucleic Acids Research, 45*(14), e77.
13. Jansen, C., Ramirez, R. N., El-Ali, N. C., Gomez-Cabrero, D., Tegner, J., Merkenschlager, M., … Mortazavi, A. (2019). Building gene regulatory networks from scATAC-seq and scRNA-seq using linked self-organizing maps. *PLoS Computational Biology, 15*(11), e1006555.
14. Ku, W. L., Nakamura, K., Gao, W., Cui, K., Hu, G., Tang, Q., … Zhao, K. (2019). Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. *Nature Methods, 16*(4), 323–325.
15. Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S. B., Ko, S. B. H., Gouda, N., … Nikaido, I. (2017). SCODE: An efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics, 33*(15), 2314–2321.
16. Omranian, N., Eloundou-Mbebi, J. M., Mueller-Roeber, B., & Nikoloski, Z. (2016). Gene regulatory network inference using fused LASSO on multiple data sets. *Scientific Reports, 6*, 20533.
17. Papili Gao, N., Ud-Dean, S. M. M., Gandrillon, O., & Gunawan, R. (2018). SINCERITIES: Inferring gene regulatory networks from time-stamped single-cell transcriptional expression profiles. *Bioinformatics, 34*(2), 258–266.