Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210862

Utilizing Shared Big Data to Identify Liver Cancer Dedifferentiation Markers

Kirill BORZIAK[a,1] and Joseph FINKELSTEIN[a]
[a] *Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, New York 10029 USA*

Abstract. Big data reanalysis has the potential to generate novel comparative analyses which aim to generate novel hypotheses and knowledge. However, this approach is underutilized in the realm of cancer research, particularly for cancer stem cells (CSCs). CSCs are a rare subset of tumor cells, which dedifferentiate from healthy adult cells, and have the potential for self-renewal and treatment resistance. This analysis utilizes two publically available single-cell RNA-seq datasets of liver cancer and adult liver cell types to demonstrate how reanalysis of big data can lead to valuable new discoveries. We identify 519 differentially expressed genes between liver CSCs and healthy liver cell types. Here we report the potential novel liver CSC dedifferentiation factor, Msh Homeobox 2, which was significantly upregulated in liver CSCs by 1.36 fold (p-value < 1E-10). These findings have the potential to further advance our knowledge of genes governing the formation of CSCs.

Keywords. Liver, cancer stem cells, single cell transcriptomics, dedifferentiation

1. Introduction

The improvements in high-throughput genomics through the funding of National Institutes of Health and other agencies are driving ever increasing volumes of digital data to be available in the public domain [1]. As the number of peer-reviewed publications that utilize previously deposited data continues to grow, the importance of shared big data in catalyzing knowledge discovery and predictive analytics continues to be further illustrated. Specifically, cancer stem cells (CSC) are a promising target for research which utilizes publically available single cell sequencing archives, due to the needs of isolating CSCs from the surroundings [2]. As one of the leading frontiers of oncology, the study of CSCs provides an interesting insight into the causes of therapeutic resistance, metastasis, and tumor recurrence [3]. Although much work has been done on identifying how the similarities between CSCs and stem cells promote their abilities to self-renew, less is known about how CSCs arise from terminally differentiated cell types and about the key triggers that initiate their stemness [3].

Although a lot research has been dedicate to the origins of CSCs, it is still not entirely clear how adult terminally differentiated cell types are able to transform into cancer cells [3]. In particular, it has been previously shown that adult terminally-differentiated hepatocytes undergo dedifferentiation which then leads to hepatocellular carcinoma [4]. The exact process of dedifferentiation which leads to the establishment of a CSC

[1] Corresponding Author, Kirill Borziak, Icahn School of Medicine, One Gustave L. Levy Place, New York, New York 10029 USA; E-mail: Kirill.Borziak@mountsinai.org.

population, however, remains to be elucidated. To help answer this question, we undertook a reanalysis of publically available single-cell RNA-seq datasets from primary liver cancer samples and healthy adult samples [5, 6]. Due to the deadliness of liver cancer and the difficulty of detecting and treating primary liver cancers, it is important to understand how treatment of liver CSCs can be utilized to promote efficacy of treatment [7]. Using the cost-effective approach of reanalyzing publically available datasets, we examined the differences in expression between liver CSCs and healthy adult liver cell types, with a focus on understanding the dedifferentiation process.

2. Methods

Expression data was obtained for the liver cancer [5] and fetal and adult healthy liver [6] studies. For the purposes of our reanalysis study, gene expression data count matrices were utilized as the starting point. This consideration was done due to the unavailability of the raw data from the liver cancer study, as it is currently still under embargo. As the two studies utilized similar approaches to sequencing, assembly, and gene calling, we do not expect any systematic issues in the gene expression profiles, which cannot be controlled for using the stringent normalization we employed.

The following studies were utilized for liver cancer cell profiles (GSE125449) [5] and healthy liver cell profiles (GSE130473) [6]. Pre-processing normalization steps included filtering out low coverage samples, low coverage genes, and non-protein coding genes. Samples with fewer than 1000 total reads were excluded. In addition to excluding genes with 0 reads in all remaining samples, we excluded all non-protein coding genes.

The normalization and differential expression analysis was performed using the edgeR [8] R package, using the gold standard methodology. To further account for possible systematic differences in mRNA detection between the two datasets, we utilized a batch effect correction in the analysis. Study type was also included in the design matrix as an additional variable. The Bonferroni multiple testing correction was used to control the false discovery rate.

Gene Ontology (GO) analysis was done using the DAVID 6.8 [9] Functional Annotation Tool, using Benjamini multiple testing correction.

3. Results

Using previously published single-cell RNA-seq data for liver cancer [5] and fetal and adult healthy liver [6], we reanalyzed 2434 single-cell samples across 18,263 protein-coding genes. To examine the differences between adult liver cells and liver CSCs, we preformed differential expression analysis between two types of adult liver cells (CD235a-/EpCAM+/ASGPR1+ and CD235a-/EpCAM+) and the liver CSCs. The adult liver cells totaled 444 samples, and the liver CSCs totaled 1990 samples. Of the 18,263 protein coding genes included in the differential expression analysis, we identified 519 genes that were differentially expressed between liver CSCs and adult liver cell types. Of these, 134 were significantly higher expressed in the liver CSCs, while 385 protein coding genes were significantly higher expressed in the adult liver cell types.

Based on Gene Ontology analysis, genes with significantly higher expression in liver CSCs were significantly enriched in several GO terms that are hallmarks of higher rates of cell division seen in cancer cells. This includes structural constituent of ribosome (p-

value = 8.9E-27), translation initiation (p-value = 9.8E-24), and rRNA processing (p-value = 1.1E-20). Additionally, liver CSCs had significant enrichments of mitochondrial related GO terms. These include mitochondrial respiratory chain complex I (p-value = 3.1E-5), NADH dehydrogenase (ubiquinone) activity (p-value = 3.2E-5), and ATP biosynthetic process (p-value = 2.9E-3). CSCs have been previously shown to be more reliant on ATP oxidative metabolism relative to other cancer cell types [10]. Potentially of most interest is the observed enrichment of over-expressed genes that are involved in the extracellular vesicle (p-value = 6.2E-12). It has recently been shown that extracellular vesicles are important factors in driving cell dedifferentiation [11, 12]. Similarly, we also see significant enrichment of ncRNA processing (p-value = 3.2E-14), supporting recent evidence that ncRNAs are able to drive dedifferentiation phenotypes [13].

On the other hand, liver CSCs showed significant decreases in proteins involved in normal liver function. This includes organic acid metabolic process (p-value = 8.1E-18), carboxylic acid metabolic process (p-value = 1.2E-17), lipid metabolic process (p-value = 4.3E-7), and drug metabolic process (p-value = 5.4E-5).

Next we examined if previously identified dedifferentiation protein were more abundant in the liver CSCs, and known differentiation factors more abundant in normal liver cell types. As expected, Hepatocyte Nuclear Factor 4 Alpha (HNF4A), which acts as the primary differentiation factor of liver cell types [14], was significantly higher expressed in healthy liver cell types (2.96X, p-value = 4.43E-5). We found Transforming Growth Factor β1 (TGFB1) to be significantly upregulated in liver CSCs relative to adult differentiated liver cell types (4.74X, p-value = 1.46E-104). This result further confirms the strong implication of TFGB1 in driving the mesenchymal/stemness phenotype observed in hepatocellular carcinomas [15]. Additionally, Msh Homeobox 2 (MSX-2) was significantly upregulated in liver CSCs (1.36X, p-value = 1.99E-17). Although MSX-2 has been previously implicated in dedifferentiation of myotubules [16], MSX-2 has not previously been reported as a potential dedifferentiation factor of liver CSCs.

4. Discussion

With the diversity of single-cell next generation sequencing available from cancer studies, we can begin asking novel questions beyond the scope of the original researchers. Cell gene expression profiles provide us with an important insight into transition to cancerous cells, particularly how dedifferentiation plays a role in generating CSCs. The use of public big data is critical to this aim. Specifically, we aimed to understand how the expression profiles of CSCs compare to healthy adult liver cell types in order to better understand the dedifferentiation capabilities of CSCs, using two publically available single-cell RNA-seq datasets.

We examined the differences in expression profiles of adult liver cells and the liver CSCs, revealing 519 differentially expressed genes. Among these, we see significant upregulation of genes involved in translation, extracellular vesicle proteins, and ncRNA processing in liver CSCs. This is mirrored by significant downregulation of normal liver metabolic proteins. Additionally, we see downregulation of a key liver differentiation factor, HNF4A, and upregulation of dedifferentiation factors, TGFB1 and MSX-2. These results provide further indication of the importance of ncRNAs and TGFB1 in promoting dedifferentiation in CSCs [13, 15]. Further, we report the first evidence of the importance of MSX-2 in the dedifferentiation of liver CSCs. Our work concurs with previous results of liver CSC studies on the importance of dedifferentiation factors [17].

These results provide an insight into cancer biology made possible by utilizing publically available big data. Our results provide a unique insight into the process of dedifferentiation in forming liver CSCs. In particular, revealing novel potential dedifferentiation factors, such as MSX-2.

5. Conclusions

Our analysis presents the power and utility of reanalyzing publically available single-cell RNA-seq datasets to ask novel biomedical questions. Focusing on identifying potential dedifferentiation factors that promote the generation of liver CSCs from adult liver cell types, we have identified increased expression of ncRNA processing and a potentially novel dedifferentiation factor, MSX-2, acting in liver CSCs. Our work demonstrates the value of shared big data to catalyze knowledge discovery and predictive analytics.

References

[1] Toga AW, Foster I, Kesselman C, Madduri R, Chard K, Deutsch EW, Price ND, Glusman G, Heavner BD, Dinov ID, Ames J. Big biomedical data as the key resource for discovery science. Journal of the American Medical Informatics Association. 2015 Nov 1;22(6):1126-31.
[2] Navin NE. The first five years of single-cell cancer genomics and beyond. Genome Res. 2015;25(10):1499-507.
[3] Batlle E, Clevers H. Cancer stem cells revisited. Nat Med. 2017;23(10):1124-34.
[4] Mu X, Español-Suñer R, Mederacke I, Affò S, Manco R, Sempoux C, et al. Hepatocellular carcinoma originates from hepatocytes and not from the progenitor/biliary compartment. J Clin Invest. 2015;125(10):3891-903.
[5] Ma L, Hernandez MO, Zhao Y, Mehta M, Tran B, Kelly M, et al. Tumor Cell Biodiversity Drives Microenvironmental Reprogramming in Liver Cancer. Cancer Cell. 2019;36(4):418-30.e6.
[6] Segal JM, Kent D, Wesche DJ, Ng SS, Serra M, Oulès B, et al. Single cell analysis of human foetal liver captures the transcriptional profile of hepatobiliary hybrid progenitors. Nat Commun. 2019;10(1):3350.
[7] Liu CY, Chen KF, Chen PJ. Treatment of Liver Cancer. Cold Spring Harb Perspect Med. 2015;5(9):a021535.
[8] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.
[9] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.
[10] Zhu X, Chen HH, Gao CY, Zhang XX, Jiang JX, Zhang Y, et al. Energy metabolism in cancer stem cells. World J Stem Cells. 2020;12(6):448-61.
[11] Sandiford OA, Donnelly RJ, El-Far MH, Burgmeyer LM, Sinha G, Pamarthi SH, et al. Mesenchymal Stem Cell-Secreted Extracellular Vesicles Instruct Stepwise Dedifferentiation of Breast Cancer Cells into Dormancy at the Bone Marrow Perivascular Region. Cancer Res. 2021;81(6):1567-82.
[12] Sun Z, Wang L, Dong L, Wang X. Emerging role of exosome signalling in maintaining cancer stem cell dynamic equilibrium. J Cell Mol Med. 2018;22(8):3719-28.
[13] Huang T, Alvarez A, Hu B, Cheng SY. Noncoding RNAs in cancer and cancer stem cells. Chin J Cancer. 2013;32(11):582-93.
[14] Martínez-Jiménez CP, Gómez-Lechón MJ, Castell JV, Jover R. Underexpressed coactivators PGC1alpha and SRC1 impair hepatocyte nuclear factor 4 alpha function and promote dedifferentiation in human hepatoma cells. J Biol Chem. 2006;281(40):29840-9.
[15] Fabregat I, Caballero-Díaz D. Transforming Growth Factor-β-Induced Cell Plasticity in Liver Fibrosis and Hepatocarcinogenesis. Front Oncol. 2018;8:357.
[16] Yilmaz A, Engeler R, Constantinescu S, Kokkaliaris KD, Dimitrakopoulos C, Schroeder T, et al. Ectopic expression of Msx2 in mammalian myotubes recapitulates aspects of amphibian muscle dedifferentiation. Stem Cell Res. 2015;15(3):542-53.
[17] Borziak K, Finkelstein J. Identification of Liver Cancer Stem Cell Stemness Markers Using a Comparative Analysis of Public Data Sets. Stem Cells Cloning. 2021;14:9-17.