Journal of High School Science

Analysis of single cell RNA sequencing data reveals pre-metastatic oncogenes within lung cancer tumor cells

Wang J[1]

Submitted: September 11, 2022, Revised: version 1, January 15, 2023
Accepted: January 28, 2023

Abstract

In the field of cancer studies, tumor malignancy marked by metastasis is a problematic effect of cancer that accounts for the majority of cancer deaths. Understanding genes relevant to metastasis is necessary for treatment development and further studies on metastasis. Using the computational part of the process of single cell RNA sequencing (scRNA-seq), information on oncogene expression can be identified and analyzed. In this paper, we provide a demonstration of usage of such computational analysis on an existing data set of non-small lung cancer samples. Through this analysis, we identified metastasis oncogenes within the sample indicating that scRNA-seq is a useful tool to detect and provide information about metastasizing cells in a cancer patient. This can inform treatment choices to enable the best likelihood of remission.

Keywords

Genetics, Tumor, Cancer, Metastasis, Epithelial-mesenchymal transition (EMT), Single cell RNA sequencing (scRNA-seq), N-cadherin, E-cadherin, Decorin

[1]Corresponding author: Jiyao (Peter) Wang, Cushing Academy, 39 School St, Ashburnham MA 01430, USA, wangjiyao@outlook.com

Introduction

According to the National Cancer Association, cancer is a disease caused by invasive cells characterized by their uncontrolled growth (1). For a tumor, a lump of cells that perform mitosis in an abnormal fashion to become malignant and thus cancerous, it has been proposed that 6 different capabilities, or hallmarks, must be acquired through a series of evolutionary processes driven by genomic instability or the increased frequency of genetic mutations either by external factors (such as radiation or carcinogens) or internal causes (such as the degradation of telomeres) resulting in the DNA's increased susceptibility to mutations (2). The six proposed hallmarks are: the abilities to sustain proliferative signaling, evasion of growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion/metastasis (2). The maintenance of proliferative signaling involves disruption of homeostasis and manipulation of the production of pro-mitosis growth signals. Evasion of growth suppressors is achieved through circumvention of growth limitations presented by tumor suppressors usually through deactivating the corresponding genes, intervention, disruption, and redirection of pathways that leads to apoptosis; cell death triggered by danger signals, which in this case would be oncogenic. Replicative immortality is enabled by the removal of restrictions to the number of mitosis cycles cells can commit, which in normal cases would result in the cell entering an irreversible state of senescence - non reproducing state - and eventual apoptosis through increased production of telomerase that restores chromosomal ends after mitosis-induced damage to enable replicative immortality. Angiogenesis occurs through the creation of blood vessels to supply tumors with nutrients and oxygen as well as to dispel waste and CO_2. The final process of metastasis results in the detachment and migration of tumor cells to other somatic sites (3). Although all the above hallmarks hold great significance in cancer studies, metastasis is the leading cause of cancer-associated death and is thus is the focus in this study.

Metastasis occurs through a cascade of steps: the detachment of epithelial cells from the intercellular adhesive complex, entrance into blood vessels and the lymphatic system, transportation to other sites within the body, the escape from blood vessels, and finally colonization. The first step occurs through a process called epithelial-mesenchymal transition (EMT) (2). While EMT, in its natural form, is a biological process used for embryo development (type 1) and tissue regeneration during wound healing (type 2), it also occurs abnormally during cancer development leading to metastasis and invasiveness (type 3) (3, 4). During EMT, epithelial cells that serve as the outer layer of organs transition into mesenchymal cells. The newly formed mesenchymal cells can detach from the intercellular complex, often through the reduction of E-cadherin production, marked by the down-regulation of the CDH1 gene repressed by certain transcriptional factors commonly belonging to the Snail family, along with that of other similar functional genes that have an important role in cell to cell binding and adhesion (5). In contrast, it is commonly observed that N-cadherin, and adhesive molecules produced by CDH2 gene that favor migration and are usually expressed in neuron cells, stromal cells and endothelial cells, are upregulated during EMT (2, 5). The anterior-

posterior polarity, asymmetry between front and back end of cells, and elongated shape of mesenchymal cells also promote migration through decreased cell adhesion (6). Additionally, an observed increase in cell plasticity, the ability of cells to engage in phenotypic changes in the absence of mutation, is associated with EMT allowing the mesenchymal cell to evade both natural defense mechanisms of the body and develop resistance to therapy drugs in response to environmental signals (2, 7, 8, 9). This increased plasticity also plays a role in the final colonization of invasive cells: it allows some invading mesenchymal cells to transition back to a differentiated form through a reverse process called Mesenchymal-epithelial transition (MET), in preparation for the establishment of the secondary tumor colony (2, 7, 10).

Research into this field suggests that EMT does not exist in a binary fashion; in fact, most of the potentially invasive tumor cells have completed only a partial level of EMT and are largely heterogeneous in terms of their degree of EMT (2, 10, 11). The fact that metastasis occurs with a low efficiency (less than 0.1% of disseminated tumor cells successfully establish distant colonies (12) suggests that only a small subset of tumor cells express the genes that give them the potential to create viable descendants for metastasis. Thus, those descendants must possess most or all genes that are required for the successful initiation of metastasis with assistance from selective pressure (7). The selective pressures involved in metastasis extend beyond dissociation and migration. For the purpose of successfully establishing a distant tumor colony after extravasation from the blood vessels and lymphatic systems, it is suggested that the migrating cells must be stem cells or similar cells with the unique ability to self-renew (produce more stem or stem-like cells) and grant descendant cells with the phenotypes of the mother cell through differentiation (6, 12, 13). Those specific cells, also called cancer stem cells (CSCs), have been suggested by recent studies to be quite rare and are necessary for the initiation of the new tumor colony; additionally, a portion of their stem cell characteristics are acquired during EMT (6, 10, 12, 13). It has also been proposed that the small subset of tumor cells with metastatic potential might correspond to cancer-stem cells or cells that can produce cancer-stem cells offspring (6, 10, 12, 13). Additionally, to gain the required nutrients and oxygen in the seed site, the migrating cells must express genes required for angiogenesis (2, 7). To more efficiently gain energy for tumor growth, it is proposed that genes expressed for more efficient tumor metabolism, specifically aerobic glycolysis, are needed (2).

In the current context, metastasis accounts for 90% of deaths from cancer (7). Part of the reason for this low survival and cure rate stems from the fact that much information about metastasis remains unknown and unstudied. Mass heterogeneity exists within cancer cells; in addition, important phenotypes along with expressed genes gained from EMT must be present in order for metastasis to be successful. scRNA-seq could be utilized to identify metastatic cells or ones with the potential to metastasize within a sample of lung cancer cells by performing statistical analysis and identifying EMT facilitating overexpressed genes.

Materials and Methods

In previous research, *Meta Analysis of Gene Expression signatures defining the epithelial to Mesenchymal Transition During Cancer Progression* (6), the researchers analyzed data from 18 different Gene expression studies and identified 130 genes expressed most frequently during EMT (Table 1). The genes are presented in groups categorized according to their specific purpose in EMT and the gain in invasion capabilities. Using R language programming, we discovered a cluster of cells associated with a high risk of future metastasis based on those genes.

Table 1: Up-regulated and down-regulated genes frequently observed in the 18 studies analyzed in *Meta Analysis of Gene Expression signatures defining the epithelial to Mesenchymal Transition* categorized into their corresponding functions related to EMT and metastasis (6)

	Upregulated	Downregulated
Cell adhesion and migration	ADAM12, CDH11, CDH2, COL1A1, COL3A1, COL5A1, COL6A1, COL6A3, CTGF, CYP1B1, DLC1, FBLN1, FBLN5, FGF2, FGFR1, FN1, HAS2, LUM, MMP2, MYL9, NID2, NR2F1, NRP1, PLAT, PPAP2B, PRKCA, RECK, SERPINE1, SERPINE2, SPOCK1, TGM2, TNFAIP6, TPM1, VCAN, WNT5A	CD24, CDH1, CXADR, CXCL16, DSG3, ELF3, EPCAM, EPHA, JUP, MPZL2, OVOL2, PLXNB1, S100P, SLC7A5, SYK
Development, cell differentiation and proliferation	CDKN2C, EMP3, FBN1, IGFBP3, IL1R1, LTBP1, MME, PMP22, PTGER2, PTX3, SRGN, SULF1, SYNE1, TAGLN, TUBA1A, VIM, ZEB1	ABLIM1, ADRB2, ALDH1A3, ANK3, BIK CA2, CTSL2, FGFR2, FGFR3, FST, GJB3, IFI30, IL18, KLK7, KRT15, KRT17, LSR, MAP7, MBP, OCLN, PKP2, PPL, PRSS8, RAPGEF5, SPINT1
Angiogenesis and wound healing	DCN, LOX, TFPI	*no gene with a major classification**
Metabolism	ABCA1, GALNT10, SLC22A4	GPX3, SLC27A2, SMPDL3B, SORL1, ST6GALNAC2
Others or unclassified	C5orf13, CDK14, EML1, FSTL1, LTBP2, MAP1B, RGS4, SYT11, TMEM158	AGR2, C10orf10, CDS1, FAM169A, FXYD3, KLK10, LAD1, MTUS1, PLS1, PRRG4, RHOD, SERPINB1, SLPI, TMEM30B, TPD52L1, TSPAN1, ZHX2, ZNF165

Categories have been chosen according to the GO classifications of the enrichment tools. Genes may be present in more than one category.
*see Table S3 for more information.
doi:10.1371/journal.pone.0051136.t002

The sample for analysis was a library of gene expression data from a sample of cells sequenced and made available by 10X Genomics. As the process of physical sequencing of the cells, involving the standard steps of isolation, lysing, amplification, next generation sequencing, and barcoding has been already performed by the 10X RNA Sequencing Platform, we proceeded to computational analysis (14, 15, 16).

The original data was obtained from Discovery Life Sciences and accessed through 10X Genomics (16). The sample included 33,000 cells from a patient diagnosed with non-small cell lung cancer (NSCLC). NSCLC is a type of cancer where malignant tumor(s) is present within the lungs (17). As the tumor formed in NSCLC is malignant, it is plausible to suggest that the primary tumor has metastasized or at least possesses the potential to metastasize. Smoking is the most common cause of NSCLC and current treatments are inadequate to provide remission (17). Lung cancer is the second most common cause of cancer with the first being breast cancer; about 236,740 adults are diagnosed with lung cancer every year in the US and it is estimated that 82% of lung cancer diagnoses fall into the NSCLC subtype (18).

All data in this paper was analyzed using Seurat. Seurat is a R package that performs statistical analysis on single cell RNA data. In our process, we used the weighted nearest neighbor (WNN) analysis of Seurat V4 to multimodally group the cells based on similarities in gene expression (19). Statistical analysis techniques used in this study, including dimensionality reduction, JackStraw, and heatmap generation, are all built into the Seurat V4 package (16).

The first step of computation analysis (Figure 1) was to perform a cleaning of the data. Generally, cells that express too few genes or too many mitochondrial genes are considered to be dying. In addition, cells that are expressing an excessive gene count would signify cell doublets or multiplets (cases arising when two or more cells are read as one and stored in the same barcode). To improve accuracy and avoid interference in later analysis, these poor quality cells were excluded from further analysis (14, 20).

Step 1

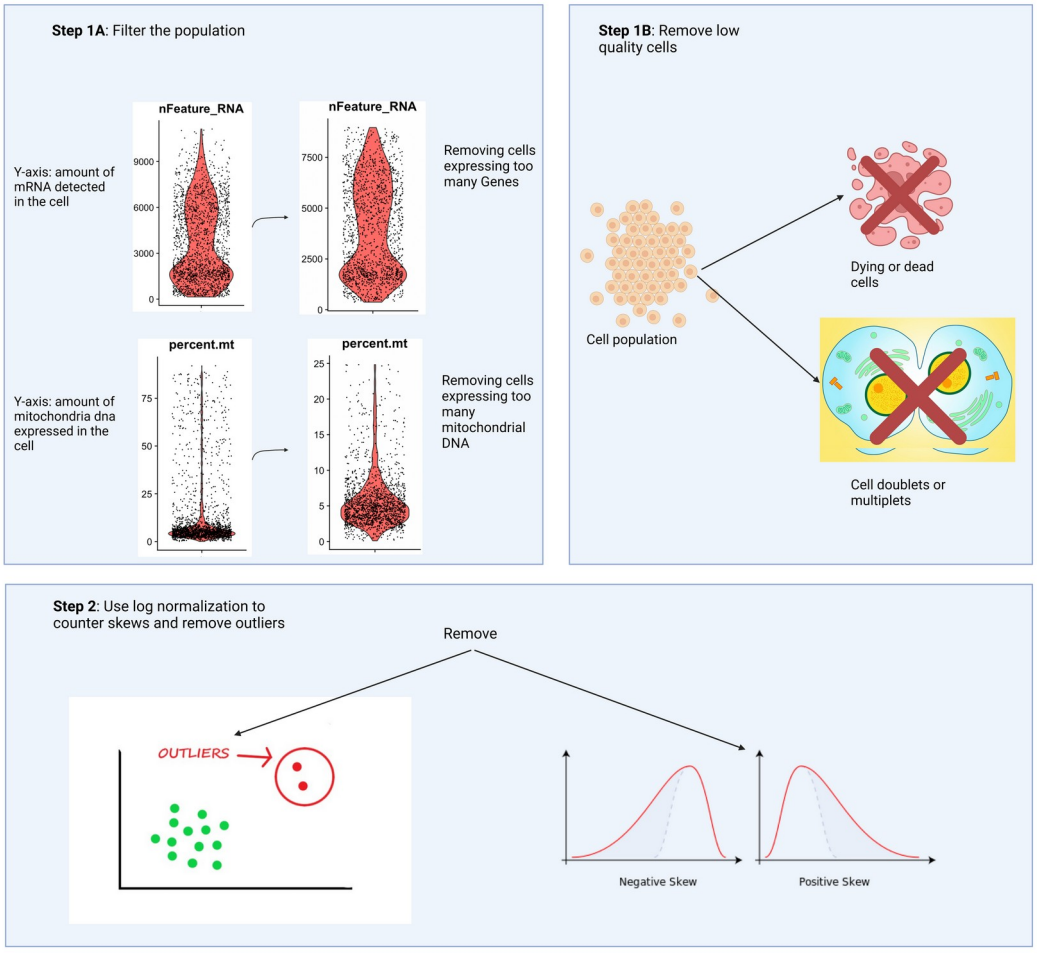

Figure 1: Excluding poor quality cells from further analysis.

Step 2

It is crucial that the data must be normalized through counteraction of skews and removal of outliers (Figure 1). This step is performed to avoid bias and noise that originated during sequencing. Because single-cell RNA sequencing involves collecting information on the expressed genes in every individual cell, the low input of mRNA from single cells requires amplification to increase accuracy. However, depending on where on the RNA strand the DNA polymerase binds, certain segments would be favored and duplicated in a higher amount while others would be copied less or even completely missed, thereby creating outliers and skews that are present within the sample data (14, 20, 21).

Steps 3 and 4

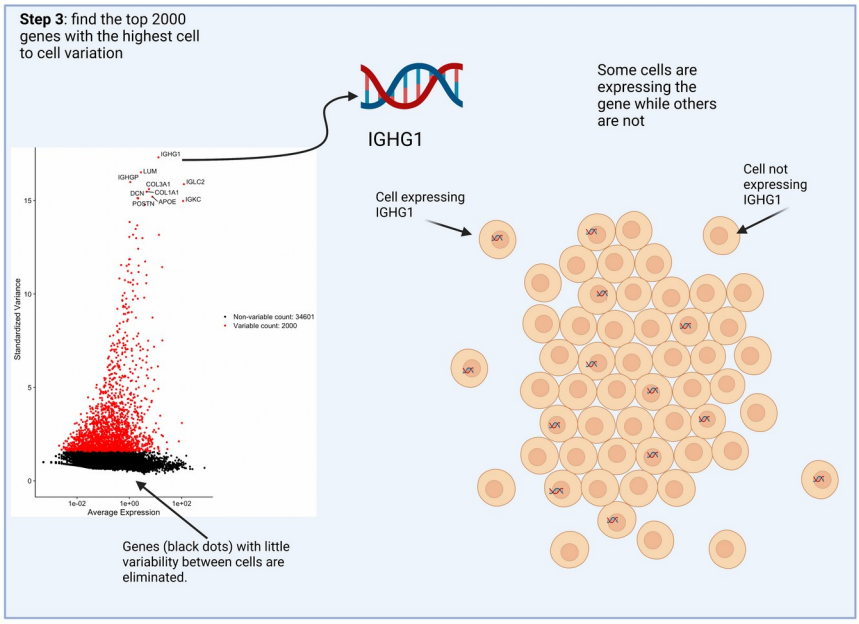

Figure 2: Selecting the most variable genes expressed in the sample

Next, we selected the most variable genes expressed in this sample and eliminated those that were either expressed in very little or close to the entirety of the cells (Figure 2). The IGHG1 gene produces a protein crucial for the activation of B cell responses, thus it is mostly expressed in immune cells and is highly upregulated in those cells. Thus, the system identified it as the gene with the most cell to cell variability (Figure 2) (21). This process was performed for the purpose of displaying important gene expression differences between non-small groups of cells instead of features or lack of features shared between the entire collection. Using the mathematical algorithm principal component (PC) analysis (PCA), the numerous dimensions of heterogeneity arising from the comparison of the expression of thousands of genes selected as having high variance can be reduced and easily visualized (14, 20). The reduced dimension appears in the form of PCA groups in increasing order of

corresponding P-values that correlate to their significance in explaining the cell to cell variability (Figure 3). A P-value is an important measurement in statistical analysis that indicates the probability that the results are due to chance alone. While a high P-value indicates a high probability that the variability is due to chance alone, a low P-value suggests that the variability may be due to real effects.

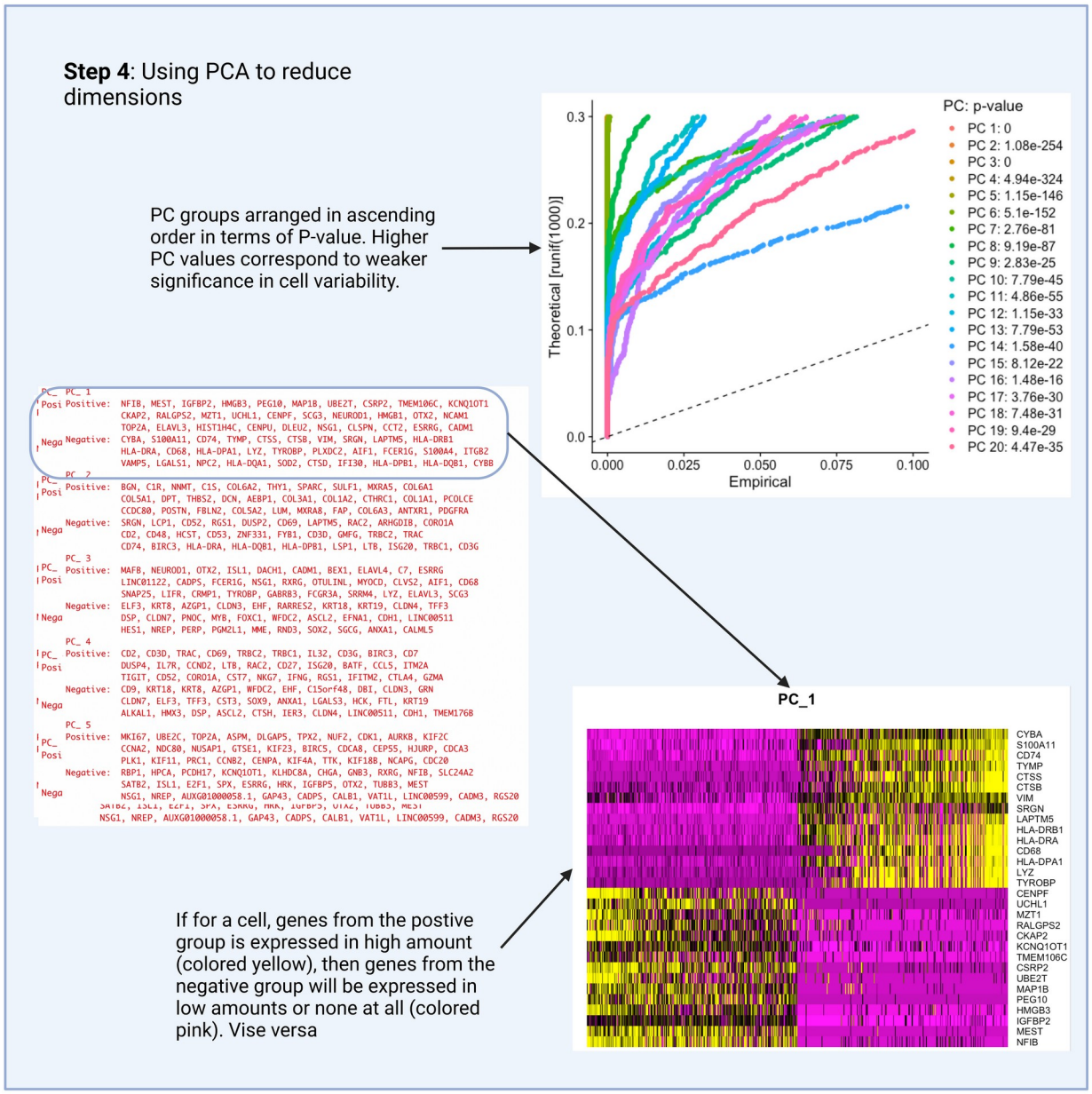

Figure 3: Using Principal Component Analysis to group genes in increasing order of corresponding P-values that account for their significance in explaining the cell to cell variability

Step 5

We took the first 15 PCs (Figure 4) selected with a P-value cutoff point of 1.0e-20. This was done in conjunction with the Jackstraw (Figure 4) and elbow plot information (Figure 5) of the PCs; the data indicated that all the PCs were statistically significant (P-value < 0.05), and the "elbow" or inflection point at 15 (see Figure 5) was selected to ensure consideration of as many significant factors as possible. The cells were subsequently clustered into 12 subpopulations to reflect heterogeneity, using the k-nearest neighbors algorithm. Specifically, the cells were plotted in different locations on a UMAP according to the PCA list. We then ran the findNeighbors() function with K-value set to a default of 20. Next, centroids (one for every cluster) were randomly placed in the clusters until the sum of the cluster's cell's distance to the centroid was minimized (20). We set the resolution parameter of the findClusters() function to 0.5 which resolved 15 distinct clusters within the cell sample. Clusters were visualized on the UMAP such that the distance between two clusters reflected their difference in gene expression, with a longer distance implying a larger difference in gene expression.

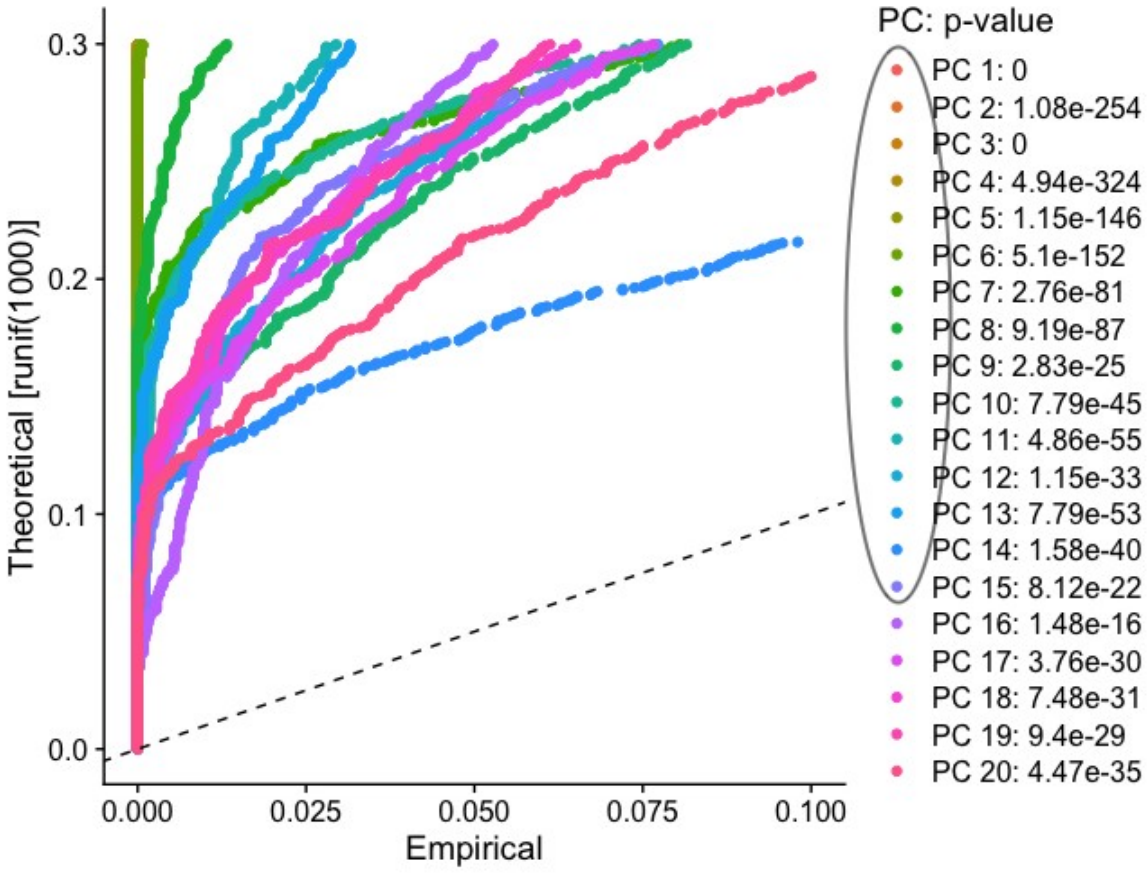

Figure 4: Selection of statistically significant principal components.

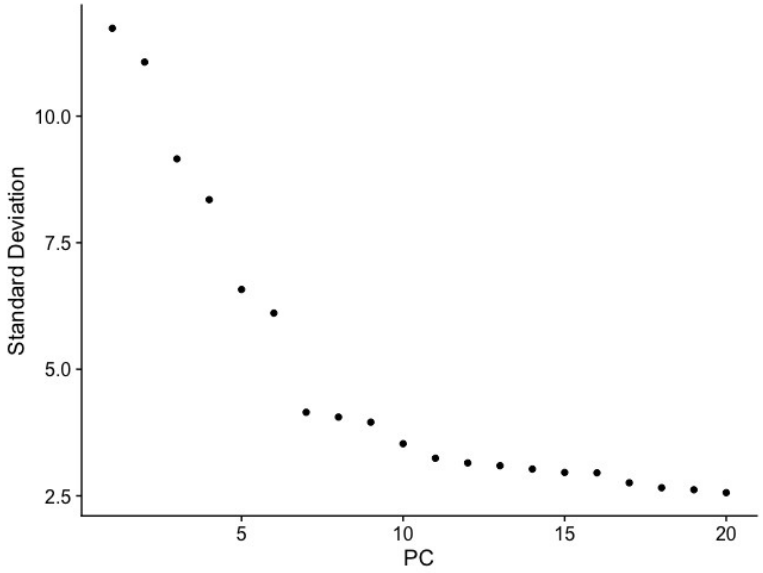

Figure 5: The 'elbow' or inflection point to determine the number of principal components to be selected.

Steps 6 and 7

Before we could identify our target; *viz*. tumor cells with the potential to metastasize, it was important to label immune cell clusters, (which were not the target for this study), for clarity and to avoid confusion. The clusters representing immune cell types, including plasma cells, B cells, monocytes, and T-cells were found by looking for the expression of five genes characteristic of those immune cells (Figure 6). In the final step, all the 130 genes found in the above study frequently up-regulated or down-regulated during EMT were input to ascertain expression or non-expression. In addition, the genes CD44 and PRRX1, which associate with CSCs capabilities, were also input. The results are displayed in feature graphs (Figure 6). If a particular cluster displayed high or low expression of pro-metastasis or anti-metastasis genes respectively, then that cluster would be identified as a group of cells going through metastasis or having the potential to do so. In the final step, we looked for specific genes in the sample and identified clusters representing immune cells, tumor cells, and normal lung cells.

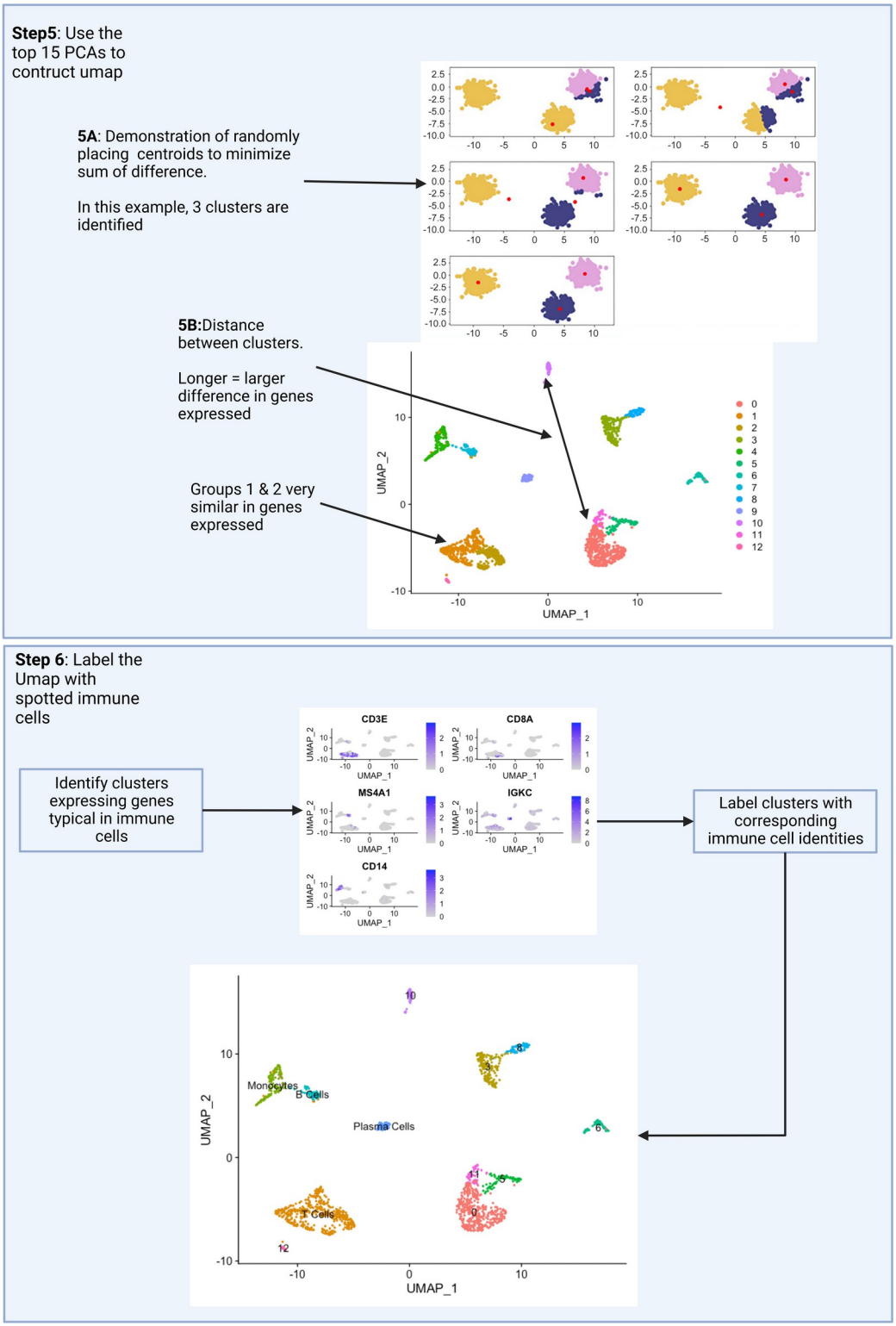

Figure 6: Identifying immune cell clusters.

Results

By inputting genes relating to metastasis, cluster 10 was found to significantly over-express most pro-EMT genes and down-regulate anti-EMT genes. Based on the evidence, this cluster could contain high amounts of pre-metastatic cells. Recent studies suggest that only a small subset of cells in a tumor can potentially activate EMT. This was consistent with the results which showed that only cells in cluster 10 expressed and silenced pro-EMT and anti-EMT genes respectively amidst a total of 33,000 cells in the sample.

Discussion

In Figure 7, a few feature maps of specific genes are displayed. All of those genes correspond to particular purposes associated with EMT. They are categorized into three groups, each facilitating different aspects of metastasis, including cell adhesion and migration, creation of CSCs for colony establishment and angiogenesis. All of those three capabilities could have been acquired by invasive cancer cells by EMT.

In the first category, all the genes - CDH2, CDH1, and CDH11- produce different types of cadherin molecules that serve as cell surface molecules that are critical for establishing and maintaining cell to cell adhesion (5) . Together with other molecules, cadherins bind cells together into a tissue environment called the extracellular matrix (ECM) (23). CDH1 produces E-cadherin which is useful in maintaining epithelial ECMs. CDH2 and CDH11 produce N-cadherin and cadherin 11 respectively; N-cadherin is expressed by neuron cells and endothelial cells while cadherin 11 is a significant marker of osteoblasts (cells involved in bone development and remodeling); all of which are types of mesenchymal cells (5, 24, 25, 26, 27). With regards to metastasis, increased expression of N-cadherin and cadherin 11 would very likely promote cell migration through facilitating bulk mesenchymal migration, while decreased amount of E-cadherin would possibly result in reduced adhesion between epithelial cells thus allowing invasive cells to detach (5, 26). The feature maps relating to “cell adhesion and migration” show that in cluster 10, CDH11 is highly expressed while CDH1 shows little to no expression (Figure 7). However, CDH2 is expressed at a low level. Based on this evidence, it is plausible to suggest that the tumor cells in cluster 10 have developed the ability to break free from the epithelial ECM of the primary tumor and are in the process of developing migration abilities.

In the second category, the genes ZEB1, CD44, and PRRX1 are frequently observed to be upregulated in CSCs and normal stem cells or stem-like cells (28). ZEB1 has been found in poorly differentiated pancreatic carcinoma cells and suppresses inhibitors of stem cell characteristics, CD44 is naturally present in stem cells of the mammary glands, and PRRX1 is naturally present during embryo development (28). In the above feature maps, it is evident that there is a moderate expression of ZEB1 and CD44 gene and a high expression of PRRX1 gene present in cluster 10. This evidence suggests that the cells in cluster 10 could be CSCs or are in the process of developing CSC characteristics during EMT. Once the cell has successfully migrated to distant body sites, the invasive CSCs may be able to perform MET and seed a new colony (7, 28, 29).

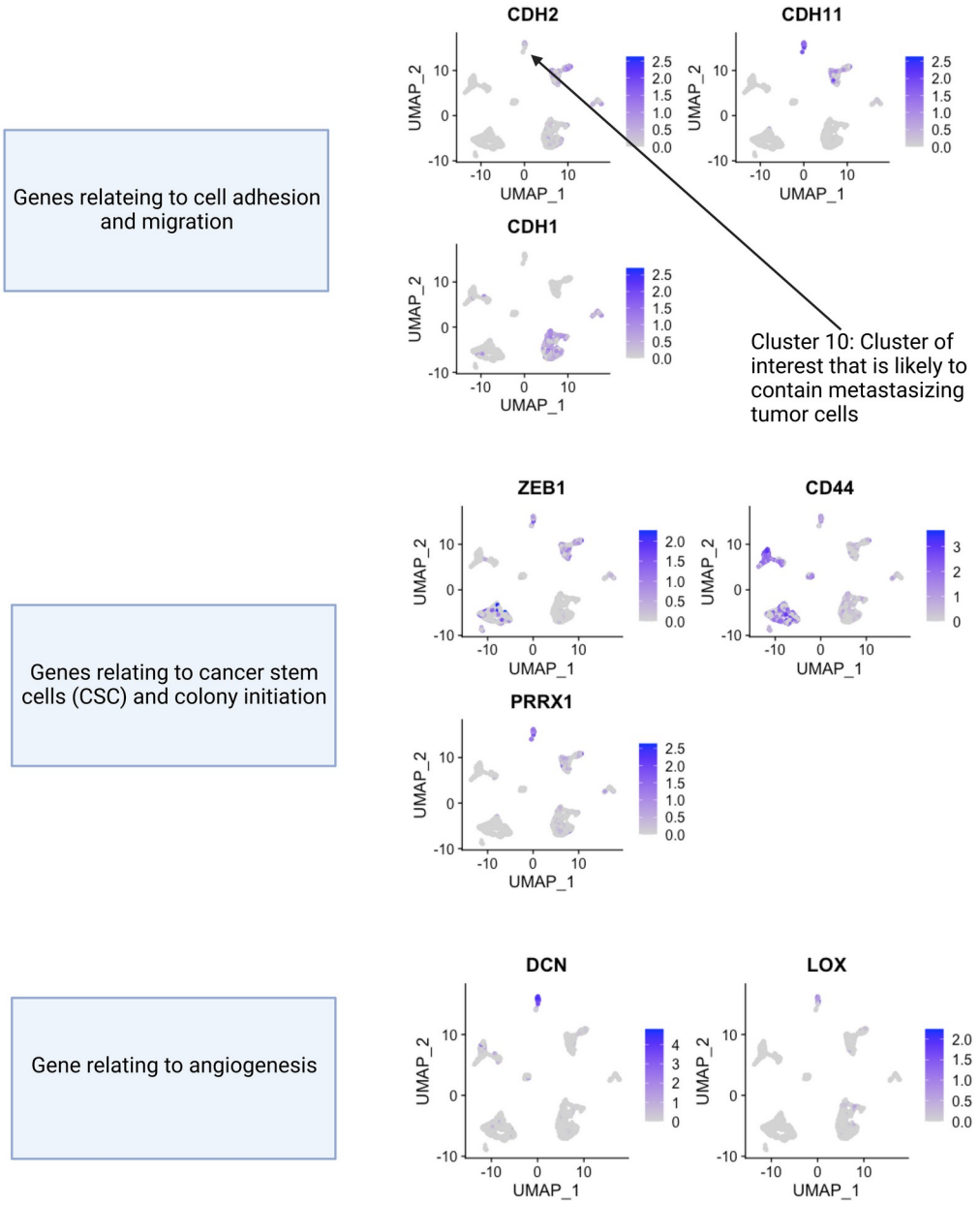

Figure 7: Feature maps of specific genes.

Finally, out of the three genes presented in the "Angiogenesis" category, only DCN was found to be highly expressed in cluster 10. DCN encodes an ECM protein called decorin which both promotes and inhibits angiogenesis (30). In its natural environment, decorin is imperative for angiogenesis as it serves to facilitate the binding of endothelial cells; the cells lining the interior of vessel tubes; to collagen I, creates a template for collagen structure, and instructs vessel wall creation (30). However, with regards to a tumor environment, decorin, especially when DCN is overly expressed and creates excessive accounts of decorin, works against angiogenesis, partly through degrading collagen IV (a molecule important for blood vessel creation and stability) and decreasing vascular escalation (30, 31). Thus, overexpression of DCN slows cancer development and should oppose metastasis. In addition, the LOX gene which promotes angiogenesis through upregulation of the VEGF is expressed in moderate amounts (32). Based on the degrees of expression of DCN and LOX, it is reasonable to infer that the metastasizing cell is either in the process of developing angiogenesis capabilities or has already acquired them. If the former is true, the metastasizing cell would most likely enter dormancy after migration and maintain its dormant state until the completion of its development (2, 7).

In a larger context, many more genes than the ones' mentioned above are participants in metastasis and EMT. The above context serves as an example of a post-computational analysis process in which expression of metastasis and EMT related genes can be identified in the sample and which can be used for prediction of potential tumor metastasis events. In addition, such information on gene expression within an individual's tumor could facilitate the development of personalized targeted therapy, depending on which genes are involved. Furthermore, tumor study through scRNA-seq could add a new dimension to cancer research by revealing genes that are involved in different cancers and those that correspond with particular metastasis patterns.

While we have identified cells that have the potential to metastasize, demonstrating that scRNA-seq can be used to identify metastasis genes, certain deficiencies are present within this study especially within the sample data used. Because the sample came from a freely available online source, much information about the donor is not revealed. Context about the patient's age, sex, time of cancer, and prior treatments are not readily available. This information is greatly relevant for our study as it informs us more about the current state of the tumor: whether it has already metastasized and whether it has already been weakened by prior treatments. Knowing the origin of this cancer manifestation (smoking, radiation, poor diet, etc.) could also help us identify specific mutated genes. Additionally, this data originated from one person developing one type of cancer, which is insufficient for it to be generalized to other individuals and other cancer types. Deficiencies also exist within scRNA-seq and data analysis. As the mRNA were collected at one specific moment, genes that are normally expressed but not at that specific time period would be missed. Furthermore, the ranges we set in the first step to remove poor-quality could have unintentionally eliminated tumor cells with differentially expressed genes.

Conclusion

In this paper, we identified pre-metastasis cells, responsible for cancer relapse, using scRNA-seq. Despite the deficiencies, scRNA-seq is a powerful and promising tool for biological analysis. As scRNA-seq is a fairly new technology that has been devised to yield an in depth analysis of every individual cell in a sample for the sake of identifying heterogeneity, constant efforts are being made to improve its accuracy, reduce its cost, and perfect its algorithms. As such scRNA-seq, with its continuous procedural improvements, possesses great potential for cancer studies and other medical innovations.

There are various potential and in-use applications of scRNA-seq to cancer treatments. For example, scRNA-seq is already used to some degree to design targeted therapy strategies. On the other hand, scRNA-seq can also be used to address a major deficiency of targeted therapy: the existence of treatment resistant tumor cells that can arise due to mutations during mitosis and subpopulations in the form of secondary tumor sites created by metastasis. It has been demonstrated that the usage of scRNA-seq can be applied to study heterogeneity in gene expression within Renal Cell Carcinomas (33). By studying variations in gene expression, the subpopulation's gene differential expression in key signaling pathways can be identified. The correct combination of drugs that target those pathways can then be administered (33).

An additional use of scRNA-seq in cancer treatment is the prediction of future metastasis. The evolutionary journey that a tumor cell takes to reach metastasis involves the gain of multiple capabilities supported by the upregulation of certain pro-metastasis genes and the downregulation of tumor suppressor genes (2, 7). Thus, knowing which genes are being up or down regulated in the primary tumor periphery can help researchers to some degree determine where the tumor is in the evolutionary timeline and about how much more time may be remaining until metastasis occurs. For instance, the colonization of a secondary site is usually preceded by the preconditioning of the new environment (2, 7). A major mechanism for such action is to use the body's natural hematopoietic cells (undifferentiated cells from the bone marrow) by upregulating VEGFR1, CD133, CD34, and c-kit genes (7). Thus, if these genes are found to be upregulated in certain hematopoietic cell samples, it could signify a later point in the timeline where a secondary tumor colony may be in the process of forming, by performing a Mesenchymal to Epithelial transaction (MET); in this case metastasis is near complete. In contrast, if only genes related to detachment from the intracellular complex are found to be overexpressed, then the tumor colony would most likely be in the early stages of developing metastasis and distant colonization may have yet to occur. In another instance where migrating cells are found by scRNA-seq to be expressing the majority of the genes needed for metastasis, except those for angiogenesis, this may be indicative of the existence of potentially dormant tumor colonies suggesting a delay in metastasis (7). This temporal information can help determine the most effective type of treatment to administer depending on how long the patient was diagnosed with the cancer.

These examples are reflective of the significant potential of scRNA-seq in metastasis prediction; an area in cancer research and treatment that has defied medical advances. With further future advancement, it is hoped that not only will the spatio-temporal regime of metastasis be identified and predicted for all tumors; but also where the likely site(s) of such metastasis may occur. There appears no doubt that this technique will continue to be utilized by researchers to unravel more determining genetic markers of metastasis and cancer.

Data

The entire code and all the feature/heat maps generated can be found in the following link: https://github.com/PeterW23/Metastasis-scRNA-seq-.git

Link to the cell samples: https://www.10xgenomics.com/resources/datasets/20-k-mixture-of-nsclc-dt-cs-from-7-donors-3-v-3-1-3-1-standard-6-1-0

Abbreviations

CDH2: Cadherin
IGHG1: Immunoglobulin heavy constant gamma 1
CD44: Cluster of differentiation 44, Homing cell adhesion molecule, Phagocytic glycoprotein-1
PRRX1: Paired related homeobox 1
CDH1: Cadherin-1
CDH11: Cadherin-11
ZEB1: Zinc finger E-box binding homeobox 1
DCN: Decorin
LOX: Lysyl oxidase
VEGF: Vascular endothelial growth factor
CD133: Cluster of differentiation 133, Prominin-1
CD34: Cluster of differentiation 34, Hematopoietic progenitor cell antigen
c-kit: Receptor tyrosine kinase

References

1. *What Is Cancer? - NCI* (nciglobal,ncienterprise). (2007, September 17). [CgvArticle]. https://www.cancer.gov/about-cancer/understanding/what-is-cancer

2. Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

3. Sinha, T. (2018). Tumors: Benign and Malignant. *Cancer Therapy & Oncology International Journal*, *10*(3). https://doi.org/10.19080/CTOIJ.2018.10.555790

4. Tiwari, N., Gheldof, A., Tatari, M., & Christofori, G. (2012). EMT as the ultimate survival mechanism of cancer cells. *Seminars in Cancer Biology*, *22*(3), 194–207. https://doi.org/10.1016/j.semcancer.2012.02.013

5. Loh, C.-Y., Chai, J. Y., Tang, T. F., Wong, W. F., Sethi, G., Shanmugam, M. K., Chong, P. P., & Looi, C. Y. (2019). The E-Cadherin and N-Cadherin Switch in Epithelial-to-Mesenchymal Transition: Signaling, Therapeutic Implications, and Challenges. *Cells*, *8*(10), 1118. https://doi.org/10.3390/cells8101118

6. Gröger, C., Grubinger, M., Waldhör, T., Vierlinger, K., & Mikulits, W. (2012). Meta-Analysis of Gene Expression Signatures Defining the Epithelial to Mesenchymal Transition during Cancer Progression. *PloS One*, *7*, e51136. https://doi.org/10.1371/journal.pone.0051136

7. Gupta, G. P., & Massagué, J. (2006). Cancer Metastasis: Building a Framework. *Cell*, *127*(4), 679–695. https://doi.org/10.1016/j.cell.2006.11.001

8. *JCI - Biomarkers for epithelial-mesenchymal transitions*. (n.d.). Retrieved August 18, 2022, from https://www.jci.org/articles/view/36183

9. Shen, S., & Clairambault, J. (2020). Cell plasticity in cancer cell populations. *F1000Research*, *9*, F1000 Faculty Rev-635. https://doi.org/10.12688/f1000research.24803.1

10. Chaffer, C. L., San Juan, B. P., Lim, E., & Weinberg, R. A. (2016). EMT, cell plasticity and metastasis. *Cancer and Metastasis Reviews*, *35*(4), 645–654. https://doi.org/10.1007/s10555-016-9648-7

11. Pastushenko, I., & Blanpain, C. (2019). EMT Transition States during Tumor Progression and Metastasis. *Trends in Cell Biology*, *29*(3), 212–226. https://doi.org/10.1016/j.tcb.2018.12.001

12. van Zijl, F., Krupitza, G., & Mikulits, W. (2011). Initial steps of metastasis: Cell invasion and endothelial transmigration. *Mutation Research/Reviews in Mutation Research*, *728*(1), 23–34. https://doi.org/10.1016/j.mrrev.2011.05.002

13. Thiery, J. P., Acloque, H., Huang, R. Y. J., & Nieto, M. A. (2009). Epithelial-Mesenchymal Transitions in Development and Disease. *Cell*, *139*(5), 871–890. https://doi.org/10.1016/j.cell.2009.11.007

14. Bacher, R., & Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*, *17*(1), 63. https://doi.org/10.1186/s13059-016-0927-y

15. Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, *9*(1), 75. https://doi.org/10.1186/s13073-017-0467-4

16. *Home Page - 10X Genomics. (n.d.). Retrieved January 30, 2023, from https://www.10xgenomics.com/resources/datasets/20-k-mixture-of-nsclc-dt-cs-from-7-donors-3-v-3-1-3-1-standard-6-1-0*

17. *Non-Small Cell Lung Cancer Treatment (PDQ®)–Patient Version—NCI* (nciglobal,ncienterprise). (2022, May 23). [PdqCancerInfoSummary]. https://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq

18. *Lung Cancer—Non-Small Cell—Statistics*. (2012, June 25). Cancer.Net. https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/statistics

19. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048

20. *Seurat—Guided Clustering Tutorial*. (n.d.). Retrieved August 18, 2022, from https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

21. Lytal, N., Ran, D., & An, L. (2020). Normalization Methods on Single-Cell RNA-seq Data: An Empirical Survey. *Frontiers in Genetics*, *11*. https://www.frontiersin.org/articles/10.3389/fgene.2020.00041

22. *Ighg1 gene—Genecards | ighg1 protein | ighg1 antibody*. (n.d.). Retrieved August 27, 2022, from https://www.genecards.org/cgi-bin/carddisp.pl?gene=IGHG1

23. *Definition of extracellular matrix—NCI Dictionary of Cancer Terms—NCI* (nciglobal,ncienterprise). (2011, February 2). [NciAppModulePage]. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/extracellular-matrix

24. Alimperti, S., & Andreadis, S. T. (2015). CDH2 and CDH11 act as regulators of stem cell fate decisions. *Stem Cell Research*, *14*(3), 270–282. https://doi.org/10.1016/j.scr.2015.02.002

25. Database, G. C. H. G. (n.d.). *CDH11*. GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. Retrieved August 27, 2022, from https://www.genecards.org/cgi-bin/carddisp.pl?gene=CDH11

26. Hazan, R. B., Qiao, R., Keren, R., Badano, I., & Suyama, K. (2004). Cadherin Switch in Tumor Progression. *Annals of the New York Academy of Sciences*, *1014*(1), 155–163. https://doi.org/10.1196/annals.1294.016

27. *Osteoblast—An overview | ScienceDirect Topics*. (n.d.). Retrieved August 18, 2022, from https://www.sciencedirect.com/topics/medicine-and-dentistry/osteoblast

28. Nieto, M. A., Huang, R. Y.-J., Jackson, R. A., & Thiery, J. P. (2016). EMT: 2016. *Cell*, *166*(1), 21–45. https://doi.org/10.1016/j.cell.2016.06.028

29. Mani, S. A., Guo, W., Liao, M.-J., Eaton, E. Ng., Ayyanan, A., Zhou, A. Y., et. al. (2008). The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. *Cell*, *133*(4), 704–715. https://doi.org/10.1016/j.cell.2008.03.027

30. Zhang, W., Ge, Y., Cheng, Q., Zhang, Q., Fang, L., & Zheng, J. (2018). Decorin is a pivotal effector in the extracellular matrix and tumour microenvironment. *Oncotarget*, *9*(4), 5480–5491. https://doi.org/10.18632/oncotarget.23869

31. Gross, S. J., Webb, A. M., Peterlin, A. D., Durrant, J. R., Judson, R. J., Raza, Q., Kitajewski, J. K., & Kushner, E. J. (2021). Notch regulates vascular collagen IV basement membrane through modulation of lysyl hydroxylase 3 trafficking. *Angiogenesis*, *24*(4), 789–805. https://doi.org/10.1007/s10456-021-09791-9

32. Baker, A.-M., Bird, D., Welti, J. C., Gourlaouen, M., Lang, G., Murray, G. I., Reynolds, A. R., Cox, T. R., & Erler, J. T. (2013). Lysyl oxidase plays a critical role in endothelial cell stimulation to drive tumor angiogenesis. *Cancer Research*, *73*(2), 583–594. https://doi.org/10.1158/0008-5472.CAN-12-2447

33. Kim, K.-T., Lee, H. W., Lee, H.-O., Song, H. J., Jeong, D. E., Shin, S., et. al. (2016). Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. *Genome Biology*, *17*(1), 80. https://doi.org/10.1186/s13059-016-0945-9

34. Ribatti, D., Tamma, R., & Annese, T. (2020a). Epithelial-Mesenchymal Transition in Cancer: A Historical Overview. *Translational Oncology*, *13*(6), 100773. https://doi.org/10.1016/j.tranon.2020.100773

35. *Frontiers | Single-Cell Transcriptomics Bioinformatics and Computational Challenges*. (n.d.). Retrieved August 18, 2022, from https://www.frontiersin.org/articles/10.3389/fgene.2016.00163/full

36. https://www.10xgenomics.com/resources/datasets/20-k-mixture-of-nsclc-dt-cs-from-7-donors-3-v-3-1-3-1-standard-6-1-0

37. Woodhouse, E. C., Chuaqui, R. F., & Liotta, L. A. (1997). General mechanisms of metastasis. *Cancer*, *80*(S8), 1529–1537. https://doi.org/10.1002/(SICI)1097-0142(19971015)80:8+<1529::AID-CNCR2>3.0.CO;2-F

38. Ribatti, D., Tamma, R., & Annese, T. (2020b). Epithelial-Mesenchymal Transition in Cancer: A Historical Overview. *Translational Oncology*, *13*(6), 100773. https://doi.org/10.1016/j.tranon.2020.100773

39. *Knowing all about Outliers in Machine Learning*. (n.d.). Retrieved August 18, 2022, from https://datascience.foundation/sciencewhitepaper/knowing-all-about-outliers-in-machine-learning

40. Hermans, R. (2008). *English: Diagrams illustrating negative and positive skew. (Created with Inkscape, an Open Source software, and based on the previous PNG version en:File:Skew.png with the text removed.)*. Own work; transferred from en.wikipedia by Rodolfo Hermans (Godot). https://commons.wikimedia.org/wiki/File:Negative_and_positive_skew_diagrams_%28English%29.svg

41. *What Is Cytokinesis?* (2019, January 21). Science ABC. https://www.scienceabc.com/pure-sciences/what-is-cytokinesis.html

42. *Introduction to k-means Clustering | The Unsupervised Learning Workshop*. (n.d.). Retrieved August 18, 2022, from https://subscription.packtpub.com/book/data/9781800200708/1/ch01lvl1sec05/introduction-to-k-means-clustering

43. *Cancer Statistics—NCI* (nciglobal,ncienterprise). (2015, April 2). [CgvArticle]. https://www.cancer.gov/about-cancer/understanding/statistics