Andrea Velazquez

Kayla Xu, Wade Boohar

QBIO490: Directed Research – Multi-omic Data Analysis

14 October 2023

# R Review Project: Evaluating the Differences Between Metastatic and Non-Metastatic Skin Cutaneous Melanoma (SKCM) Across the Genome and Transcriptome

## Part 1: Review Questions

1. What is TCGA and why is it important?
   a. The Cancer Genome Atlas, known as TCGA, is a pulicly available genomics research program from the National Cancer Institute that has data for over 20 thousand cancer and normal samples from over 30 cancer types. It is important because it provides large cancer data sets that can be studied and from which findings can be made to find groundbreaking approaches to detection, prevention, treatments, and prognosis indicators for cancer.
2. What are some strengths and weaknesses of TCGA?
   a. Strengths: It follows HIPAA, which allows patients to contribute to science by donating their samples, but without disclosing sensitive medical information publicly. TCGA protects patients data by assigning them barcodes instead of using the patients' information to identify the samples. They also get rid of identifying information that could jeopardize patients' safety.

   b. Weakness: It is not great in terms of having long term data. Meaning that there is very little to no information in the long term health of the patient; preventing us from having data about cancer recurrence, which we could use to see if certain gene mutations are more frequent in recurring cancer or not, and more.

3. What commands are used to save a file to your Github repository?
   a. git add name_of_file.type
   b. git commit -m “informative message about file upload”
   c. git push
4. What commands must be run in order to use a package in R?
   a. if(!require(PackageName)) {install.packages("PackageName")} to check if it is installed and if not, install it
   b. library(PackageName) to load the package into R
5. What commands must be run in order to use a Bioconductor package in R?
   a. (!require(BiocManager)) {install.packages("BiocManager)")}
   b. BiocManager::install(“PackageName”)
6. What is boolean indexing? What are some applications of it?
   a. When you assign a boolean value (T/F, 0/1, etc) to something based on a condition. It can be used to make a mask and filter out selected rows that comply with a given condition from a dataframe.
7. Draw a mock up of a sample df. Show an example of the following and explain what each line of code does
   a. DataFrame: Patients_df

| Pt number | Gender | Vital_status |
|---|---|---|
| 1 | Female | Alive |
| 2 | NA | Dead |
| 3 | Male | Alive |

b. An ifelse() statement: mask<- ifelse(!is.na(Patients_df$Gender), T, F). We can use this statement as a mask to filter out the patients that have NA values from the ones that don't in our dataframe by using Patient_df<- Patient_df[mask, ].

| Pt number | Gender | Vital_status |
|---|---|---|
| 1 | Female | Alive |
| 3 | Male | Alive |

c. Boolean indexing: Patients_df$isAlive<- ifelse(Patients_df$Vital_status == "Alive", T, F). This will make a new column, called "isAlive", where it will assign True is a patient's vital status is "Alive" and false otherwise. We can then use this to compare data across patients with a TRUE value in this column vs patients with a FALSE value.

| Pt number | Gender | Vital_status | isAlive |
|---|---|---|---|
| 1 | Female | Alive | TRUE |
| 2 | NA | Dead | FALSE |
| 3 | Male | Alive | TRUE |

## Part 2: Analyses

### Survival

For this analysis, we evaluate the survival of metastatic SKCM (n=368) vs non-metastatic SKCM patients (n=103). For this, we used a Kaplan-Meier plot with the occurrence of a death event based on the patient's vital status; furthermore, we used the days to last follow up as the survival time for patients that are alive, and the days to death as the survival time for patients that passed away.

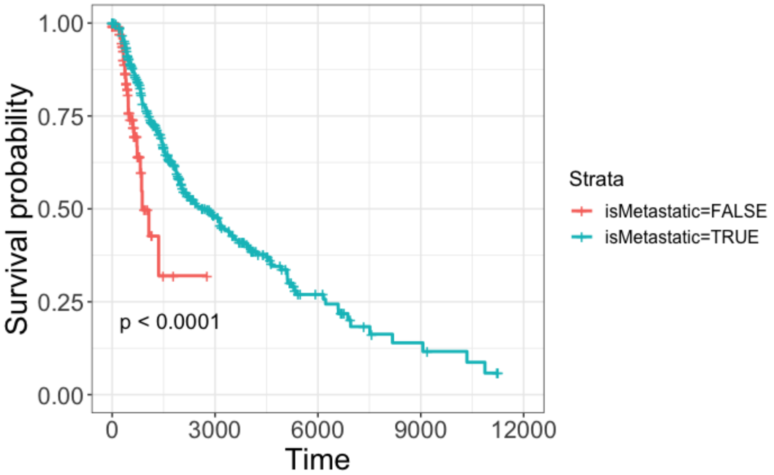

Figure 1. Survival probability of metastatic (blue) vs non-metastatic (red) SKCM patients evaluated over time in days.

In the plot obtained, we can observe the survival probability of metastatic patients significantly drop over time, with the final probability being less than 5% (0.05) after almost 12000 days. Whereas the survival probability for non-metastatic patients seems to stabilize and remain constant at a value close to 30% around the 1500 day mark and on– meaning that no additional death events occurred in the non-metastatic patient cohort after this point in time. Moreover, the p-value ($p < 0.0001$) for the survival probability over time between the two patient cohorts suggests that the difference between the survival probability over time for metastatic vs non-metastatic patients is significant. Meaning that is is unlikely to see the survival probability for metastatic and non-metastatic patients be the same. We cannot confirm that metastasis in SKCM is the sole cause of the survival probability over time being different between the two cohorts, but we can conclude that it may be a factor.

Additional literature supports our findings, suggesting that the 5-year survival rate of metastatic melanoma patients, not SKCM specific, is between 5% and 10% (Sun et al). Given that metastasis consists of a spread of the cancer to other parts in the body other than the site it originated in, it results difficult to treat. With common sites for metastatic melanomas to spread including bones, muscles, skin, and vital organs like the brain, the liver, and the lungs, metastasis can make SKCM more aggressive than it already is, significantly affecting the prognosis and the treatment options available for patients.

**Mutation differences for multiple genes**

We conducted an analysis of somatic mutations across our two patient cohorts, allowing us to observe side-by-side the differences in type and occurrence of mutations across different genes. We then selected the top 10 most-mutated genes for each cohort, from which 7 of those genes were included in both cohorts' top 10 (different order and rate in each cohort); meaning that we ended up with 13 different genes that included both cohorts' most mutated genes for our comparison. Using a co-oncoplot, we displayed them side by side, allowing us to observe the differences.

Genes TTN and MUC16 were the top mutated genes, in the same order, in both cohorts; with TTN having a 73% mutation rate for metastatic patients and a 67% mutation rate for non-metastatic patients; and MUC16 having a 67% mutation rate for metastatic patients and 62%

mutation rate for non-metastatic patients. While the mutation rates are fairly similar, the types of mutations for each of these genes across our 2 cohorts are different: the metastatic cohort showed multi-hit, missense, nonsense and splice-site mutations for TTN, while the non-metastatic cohort only showed multi-hit and missense mutations. On the other hand, the metastatic cohort showed only multi-hit and missense mutations for MUC16, whereas the non-metastatic cohort showed multi-hit, missense, and nonsense mutations. While these mutation-type differences are something to keep in mind, we cannot reach a conclusion on their significance or role in metastasis and prognosis of SKCM as more data and further analysis would be needed– our cohorts are relatively small in size for that.

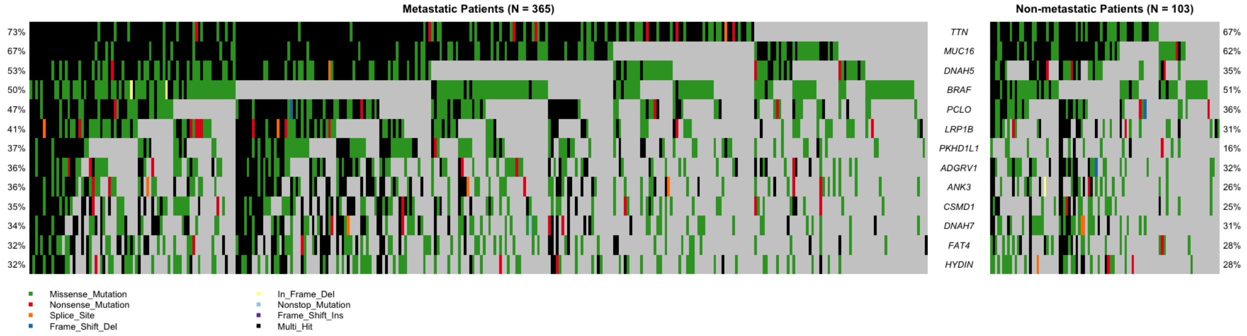

Figure 2. Mutation differences for the 13 different genes in the top 10 most-mutated genes across SKCM metastatic patients (n=365) and non-metastatic patients (n=103).

Overall, TTN and MUC16 don't tell us much about the difference in somatic mutations between metastatic and non-metastatic SKCM patients in this case because of the similarity in mutation rates across our cohorts. It is important to point out that patients (from either cohort) without mutations in TTN and MUC16 seem to have a lower number of mutations across the other 11 genes shown in our plot. This may be useful in better understanding the role of TTN and

MUC16 in other somatic mutations across SKCM cancer patients, not specific to metastasic status.

Conversely, we can look at genes with relatively different mutation rates across our cohorts to observe the differences in somatic mutation in metastatic vs non-metastatic patients. Take for example PKHD1L, which has a 37% mutation rate in the metastatic cohort and a 16% mutation rate in the non-metastatic cohort. Both cohorts show multi-hit, missense and nonsense mutations. Existing literature (Hogan et al.) suggests a link between PKHD1L and immune response; however, there is no new literature that explores the applications of this finding or further investigates it. The increased number of mutations in the metastatic cohort may further disrupt the ability of PKHD1L to contribute to the immune system working correctly and allowing it to defend itself against the cancer or work together with cancer treatments. PKHD1L may be used as a biomarker to confirm metastasis in SKCM as the mutation levels are different and having increased number of mutations may indicate metastasis; however, further analysis is required to explore this possibility.

**Mutation differences for a specific gene**

We looked at the mutation differences across metastatic and non-metastatic patients for the gene FAT4. While the mutation rates obtained from the previous analysis for FAT4 across our two cohorts are relatively similar (32% for metastatic patients and 28% for non-metastatic patients), we chose to take a deeper look into this gene because of the strong potential it has to function as a prognosis biomarker across several cancers (Mao et al.) Furthermore, recent discoveries show upregulation of FAT4 linked to metastasis suppression in non-small cell lung cancer (Ning et al.), hence prompting us to explore the possibility of that link existing in SKCM.

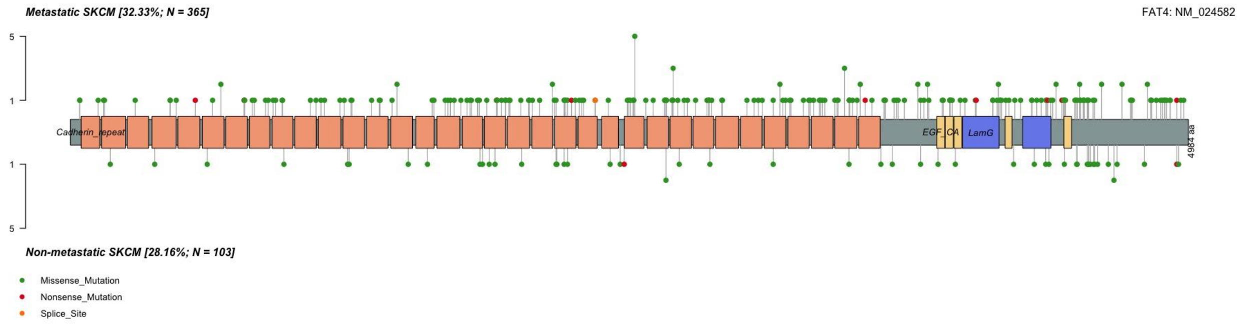

Figure 3. Mutation type and occurrences across the FAT4 gene locus for metastatic and non-metastatic SKCM patients.

The co-lollipop plot shows the mutations in the FAT4 gene loci and the amount of mutations that our cohorts have at each locus within the gene. The metastatic cohort has 226 (95% of total mutations) missense mutations, 11 nonsense mutations (4.6% of total mutations), 1 splice-site mutation (0.4% of total mutations) ; whereas the non-metastatic cohort has 50 missense (96.2% of total mutations) and 2 nonsense mutations (3.8% of total mutations), but no splice-site mutations. The makeup of the mutations seems to be pretty similar across both cohorts, so taking a deeper look at the loci in which the mutations occur might be more insightful than the makeup of the mutations in the gene across cohorts.

The metastatic cohort has relatively sparse mutations across FAT4's first 7 cadherin repeats (shown in orange), the mutations across the subsequent cadherin repeats and all other coding and non-coding parts of the gene are densely packed. The non-metastatic cohort has the most densely packed mutations in the non-coding part of the gene located close to the end of it; the mutations that occurred in the coding parts of the gene for this cohort are relatively spread out.

Additionally, the metastatic patient cohort has an increased number of mutations in loci where the non-metastatic cohort has none or very few (see cadherin repeats 6, 14, 24, 26, 30, and

33 to name some). With the most notable difference in cadherin repeat 24, where the metastatic cohort had 5 occurrences of missense mutations in one specific loci, and more across the cadherin repeat, while the non-metastatic didn't have any mutations. Looking into the connection between these mutation differences and the effect they have in differential expression of this gene may be helpful. That way we can gain insight into what types of mutations, and where in the gene locus, cause upregulation of FAT4. Allowing us to evaluate the aforementioned link between upregulation of FAT4 and metastasis suppression in SKCM, and helping us determine the potential to use FAT4 as a prognostic biomarker in metastatic SKCM.

**Co Occurrence or mutual exclusion of common mutations**

We then analyzed the co-ocurrence or mutual exclusion of common somatic mutations across each cohort. For the metastatic cohort, the somatic interactions plot obtained shows a high level of cooccurrence across the top mutated genes, and deems the coocurrence of these mutations significant. The only mutations in our plot that are not significant, and have low levels of co-ocurrence, or even some levels of mutual exclusion, are all linked to the BRAF gene. Our findings align with previous literature that shows SKCM to have a high mutation burden (Martincorena et al.), in this case exacerbated by metastasis.

On the other hand, we see more variety in the co-ocurrence levels of the top genes in the somatic interactions plot for the non-metastatic cohort. Compared to plot for the metastatic cohort, which deems all mutations (with the exception of those involving BRAF) significant and with a high level of co-ocurrence, in the non-metastatic cohort we observe some co-ocurring mutations that are not significant. Suggesting that co-ocurrence of specific mutations, like

CN10A and ZFHX4 for example, in non-metastatic patients may happen by chance rather than by an underlying connection in the disease mechanisms of SKCM.

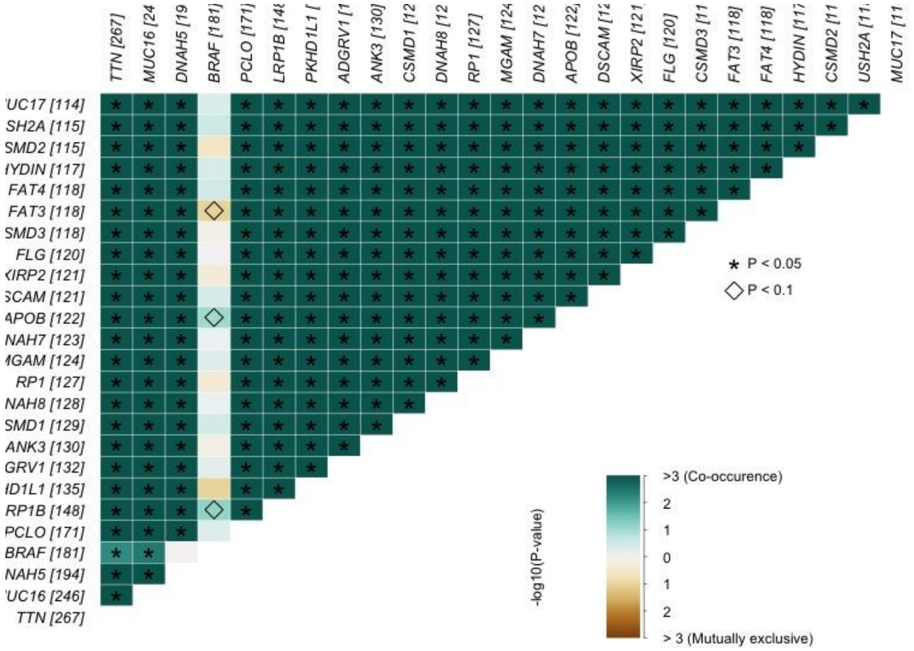

Figure 4. Somatic interactions across mutations for the top mutated genes in metastatic SKCM patients.

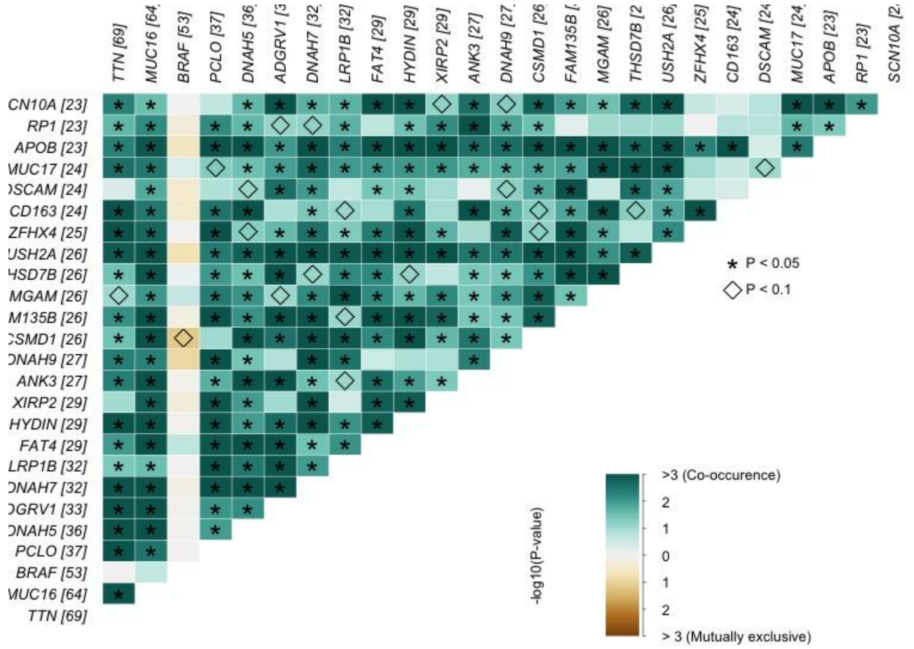

Figure 5. Somatic interactions across mutations for the top mutated genes in non-metastatic SKCM patients.

Furthermore, comparing both plots will yield us the most insight into the differences of SKCM across metastatic and non-metastatic patients. In the metastatic cohort, mutations in TTN and BRAF are co-ocurrent and significant, so is the case for BRAF and MUC16. On the non-metastatic cohort, this is not the case; both of these mutations have a low co-ocurrence and are not significant. This opens the possibility of using genetic testing for mutations in these 2 different pairs of genes (BRAF and TTN, and BRAF and MUC16) to predict the occurrence of metastasis in SKCM when detected at an early stage. This pair of mutations, along with other co-ocurring mutations could also be useful in coming up with targeted cell treatments for metastatic SKCM– depending on the treatment and the pathways it targets on a case-by-case basis. There is new literature that explores the BRAF and MUC16 pathway and its effect on tumor mutation burden, paving the road to discovery for new treatments and giving us insight into the complex mechanisms behind SKCM (Wang et al).

**Differential expression**

As seen in some of the previously conducted analyses, sometimes, it is not enough to evaluate the mutation differences across cohorts. Take for example our analysis on the FAT4 gene: though we knew the different mutations in the gene locus, we didn't know their effect in the upregulation/downregulation of the gene, which is precisely what we need to determine the potential of FAT4 as a prognostic biomarker in metastatic SKCM. For the last analysis of this work, we will conduct a differential expression analysis across our cohorts. While doing so, we will control for covariates like gender, vital status, race, and treatment effects, to be able to observe the effects of metastasis on overall gene expression in SKCM patients. We compared the fold change in expression for genes in metastatic SKCM patients compared to non-metastatic

SKCM patients. Leading us to some observations; for the purpose of this work, only the most notable will be mentioned.

Metastatic SKCM patients have significant downregulation of some genes. One of them being TGM1, which was deemed to be “downregulated in metastatic melanoma tissue in comparison to primary melanoma tissue” (Li et al.) The gene that resulted to be the most downregulated, meaning it had the largest negative fold change compared to expression in non-metastatic patients, was DEFB4A. This gene was recently demeed a novel biomarker for early stage metastasis in tongue cancer (Lee et al.); given the significant downregulation of this gene in metastatic vs non-metastatic SKCM samples, the potential of it serving as a biomarker to detect early stage metastasis in SKCM could be explored with further research.

Similarly, there is upregulation in some genes as well. One of those genes being C7. This gene has been linked to poor prognosis in breast cancer, suggesting it aids the progression of the cancer, and also acts as a contributing factor for bone metastasis– patients with a higher expression of C7 have shown to have a shorter time window between their breast cancer diagnosis and the onset of bone metastasis. Additionally, breast cancer patients that show upregulation on the C7 gene have poorer outcomes when using taxane-based and anthracycline-based chemotherapy as a treatment (Zhang et al). We could explore the possibility of the effects of upregulation of C7 in breast cancer transposing to SKCM. We could also explore the possibility of using C7 as a biomarker for treatment options in SKCM patients. The most upregulated gene in this differential expression analysis was RN7SK43, which is actually a pseudogene. Upregulation of this pseudogene has been connected to a positive feedback loop with m6A proteins, which contribute to progression and metastasis of cancers across various types of cancers (Xu et al.).

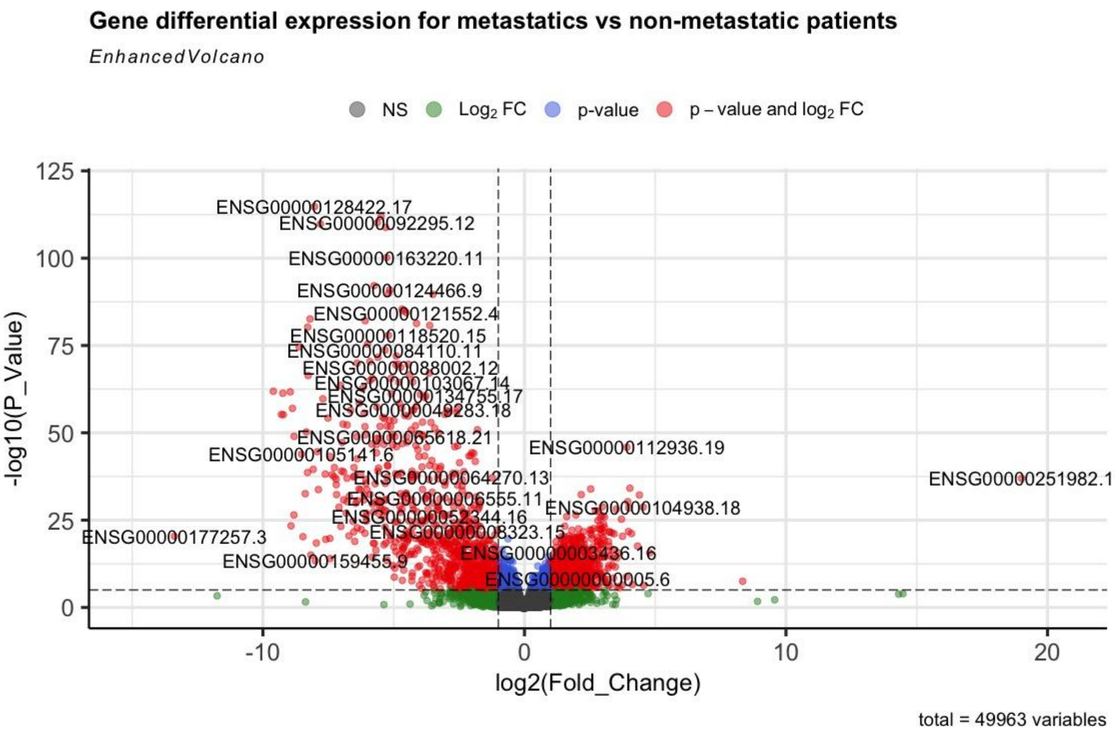

Figure 6. Differential expression for genes in metastatic SKCM patients compared to non-metastatic SKCM patients. Significant gene expression differences are shown in red, non-significant (p-value > 0.05) shown in green, no notable difference in gene expression shown in blue, and no notable difference + no significant change in expression shown in gray.

# Works Cited

Hogan, Marie C., et al. "PKHDL1, a Homolog of the Autosomal Recessive Polycystic Kidney Disease Gene, Encodes a Receptor with Inducible T Lymphocyte Expression." *PubMed*, 15 Apr. 2003, pubmed.ncbi.nlm.nih.gov/12620974/. Accessed 14 Oct. 2023.

Lee, Doh Y., et al. "The Expression of Defensin-Associated Genes May Be Correlated With Lymph Node Metastasis of Early-Stage Tongue Cancer." *PubMed*, 16 Nov. 2022, pubmed.ncbi.nlm.nih.gov/36097842/. Accessed 14 Oct. 2023.

Li, Kenzhu, et al. "Identification of Keratinocyte Differentiation-Involved Genes for Metastatic Melanoma by Gene Expression Profiles." *Publishing Open Access Research Journals & Papers | Hindawi*, 28 Dec. 2021, www.hindawi.com/journals/cmmm/2021/9652768/. Accessed 14 Oct. 2023.

Mao, Weili, et al. "A pan-cancer analysis of FAT atypical cadherin 4 (FAT4) in human tumors." *PubMed Central (PMC)*, 16 Aug. 2022, pubmed.ncbi.nlm.nih.gov/36051999/. Accessed 14 Oct. 2023.

Martincorena, Iñigo, et al. "High burden and pervasive positive selection of somatic mutations in normal human skin." *Science*, 22 May 2015, www.science.org/doi/10.1126/science.aaa6806. Accessed 14 Oct. 2023.

National Cancer Institute. "Metastatic Cancer: When Cancer Spreads." *Cancer types: Metastatic Cancer*, 10 Nov. 2020, www.cancer.gov/types/metastatic-cancer#what. Accessed 14 Oct. 2023.

Ning, Yue, et al. "Increased expression of FAT4 suppress metastasis of lung adenocarcinoma through regulating MAPK pathway and associated with immune cells infiltration." *Cancer Medicine*, 30 June 2022, onlinelibrary.wiley.com/doi/full/10.1002/cam4.4977. Accessed 14 Oct. 2023.

Roosan, Moom R., et al. "Evaluation of Somatic Mutations in Solid Metastatic Pan-Cancer Patients." *MDPI*, 3 June 2021, www.mdpi.com/2072-6694/13/11/2776. Accessed 14 Oct. 2023.

Sun, Ledong, et al. "Identification of Long Non-coding and Messenger RNAs Differentially Expressed Between Primary and Metastatic Melanoma." *PubMed Central (PMC)*, 4 Apr. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6459964/. Accessed 14 Oct. 2023.

Wang, Zi, et al. "Effect of MUC16 Mutations on Tumor Mutation Burden and Its Potential Prognostic Significance for Cutaneous Melanoma." *PubMed Central (PMC)*, 15 Feb. 2022, www.ncbi.nlm.nih.gov/pmc/articles/PMC8902552/. Accessed 14 Oct. 2023.

Xu, Xin, et al. "A positive feedback circuit between RN7SK snRNA and m6A readers is essential for tumorigenesis." *Molecular Therapy*, www.cell.com/molecular-therapy-family/molecular-therapy/pdf/S1525-0016(22)00717-1.pdf. Accessed 14 Oct. 2023.

Zhang, Huikun, et al. "High Expression of Complement Component C7 Indicates Poor Prognosis of Breast Cancer and Is Insensitive to Taxane-Anthracycline Chemotherapy." *PubMed Central (PMC)*, 24 Sept. 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC8497743/. Accessed 14 Oct. 2023.

Zhu, Yanan, et al. "The potential role of m6A reader YTHDF1 as diagnostic biomarker and the signaling pathways in tumorigenesis and metastasis in pan-cancer." *Nature*, 28 Jan. 2023, doi.org/10.1038/s41420-023-01321-4. Accessed 14 Oct. 2023.

Zou, Sheng, et al. "Mutations in the TTN Gene Are a Prognostic Factor for Patients with Lung Squamous Cell Carcinomas." *PubMed Central (PMC)*, www.ncbi.nlm.nih.gov/pmc/articles/PMC8742622/. Accessed 14 Oct. 2023.