Rishee Venkatesh

QBIO 490

10 March 2023

Introduction

The TCGA – The Cancer Genome Atlas – is a publicly available multi-omic dataset, consisting of genomic data from over 20,000 samples of 33 cancer types. TCGA Data consists of clinical data, mutation data (in MAF format), RNA count data, and protein data through the Clinical Proteomic Tumor Analysis Consortium.

Approximately 12.9% of women born today will develop breast cancer at some point in their lives, according to current incidence rates (NCI). Aside from risk factors such as mutations to the BRCA1 and BRCA2 tumor suppressor genes, one of the most prominent risk factors attributed to breast cancer is age. Most breast cancers are known to be diagnosed after the age of 50, with the median age of diagnosis in the US as 62 (American Cancer Society). This study analyzes the association of age with specific types of mutations as well as genes that are mutated, with patients classified as either younger or older by a median age of 58.

While no significant association of a particular type of mutation was found with age, the study further supported that PIK3CA mutation associates with older age (Kalinsky et. al), and that SCARNA7 was significantly upregulated in older populations, while CSN2 was significantly downregulated.

Methods

The R packages for Bioconductor, TCGA, maftools (for the coOncoplot), survival and survminer (for the Kaplan-Meier plot), ggplot2, DESeq2 (for the differential expression analysis), and EnhancedVolcano (for the Enhanced Volcano Plot) were loaded into the R

environment. TCGA data was first queried into the RStudio environment from the GDC (Genomic Data Commons) with the accession code TCGA-BRCA. The study used TCGA Clinical Data, TCGA MAF Data, and RNASeq data, saved to individual data frames (for RNASeq data, as a 3D dataframe). Unavailable NA values in the data were removed using Boolean masking, from the Age variable in the clinical data frame and the Age Category variable in the RNA Clinical and Counts data. Boolean masking was also applied to remove irrelevant negative infinity values from the Survival Time variable in the clinical data frame prior to the creation of the Kaplan-Meier plot, to subset data by median age as either 'Young' or 'Old' before creating the coOncoplot, and to pre-filter the gene data from RNA Counts, where total sum of the gene across all patients was less than ten, making it insignificant. A box-plot was first created to measure the association between age and type of mutation. A Kaplan-Meier plot was then created to identify survival rates for each mutation type, by categorizing patients by vital status and using the Surv, survfit and survplot functions in the survival and survminer libraries. A coOncoplot was then created by sorting the patients by age – those under or of the median age of 58 were considered 'Young,' and those over the age of 58 were considered 'Old.' The subsetMaf function, used to summarize, visualize and analyze MAF data, was then applied to the young and old datasets to sort the data for the coOncoplot. Then, a ddsSeq was performed after converting the age category in the RNA Clinical dataframe into a factor datatype and organizing the data. A Volcano Plot was then created using the ddsSeq data.

Results

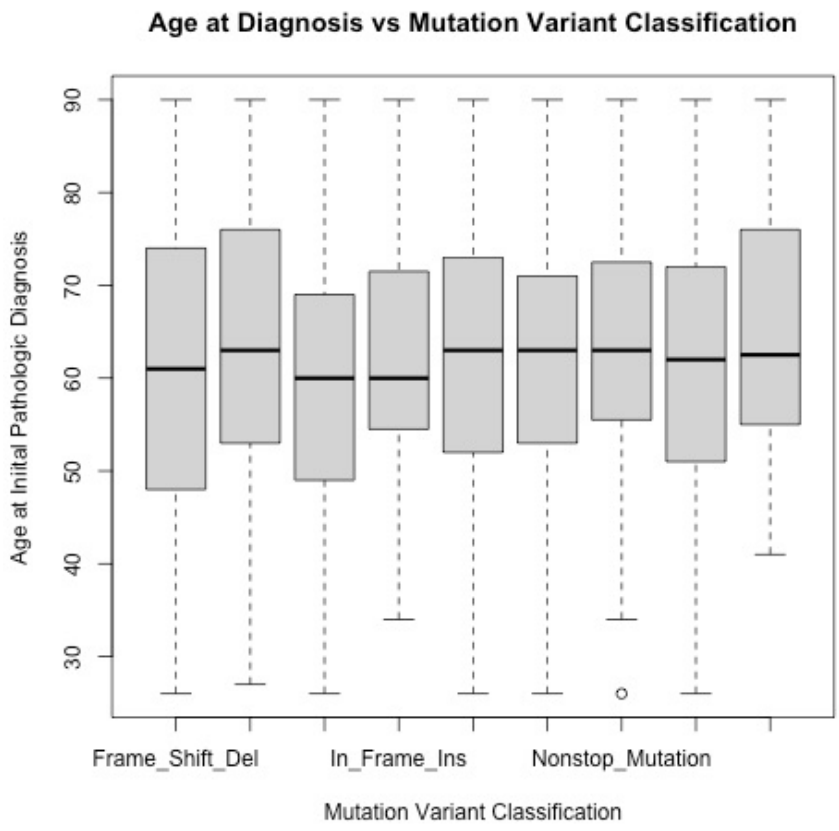

Figure 1: Box-Plot showing Age at Initial Pathological Diagnosis versus the type of mutation variant, sorted into nine types – frameshift insertion & deletion, in-frame insertion & deletion, missense, nonsense, nonstop, splice-site, and translation-site. There is no strong correlation between age and type of mutation.

As shown by the box-plot (Figure 1), among the nine given mutation types, all had a similar median age at diagnosis; insertion frameshift mutations had a larger third quartile, with the 75th percentile value reaching approximately 75, while the other mutation types had 75th percentile values under or equal to the age of 70. Deletion frameshift mutations had a larger first quartile, with the 25th percentile reaching approximately 45, while the other mutation types had 25th percentile values of age 50 or above. Range was also similar for most of the mutation types; nonstop mutation type had the smallest range from age ~35 to ~80, along with translation start site (age ~40 to ~90) and In Frame Insertion (age ~35 to ~90).

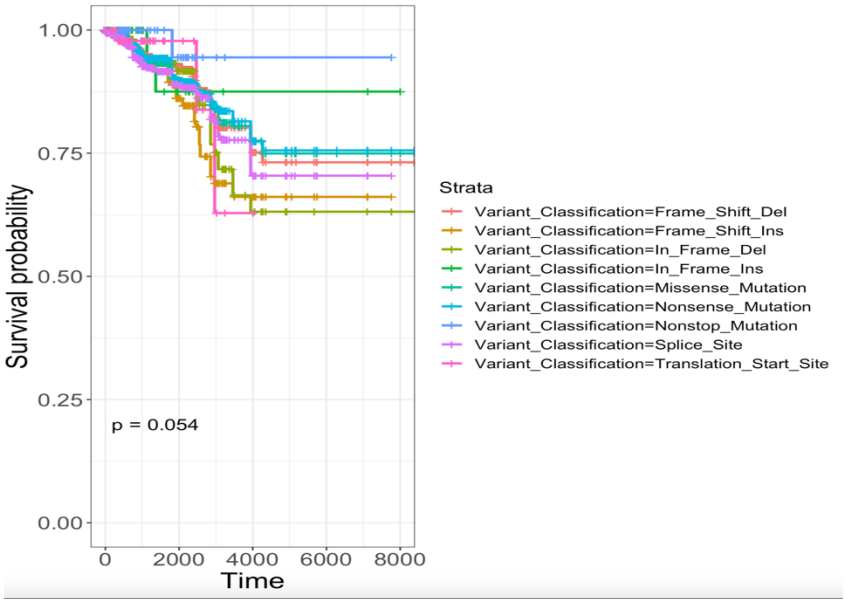

Figure 2: Kaplan-Meier survival plot showing the survival probability for each type of mutation, with a p-value of 0.054 demonstrating the statistical significance of the results. The mutation type with the highest survival probability was the non-stop variant, at approximately 93%, and the lowest was the in-frame deletion and frameshift insertion at ~62.5%.

Figure 2, the Kaplan-Meier plot, displayed the survival probability for patients with each mutation type. The mutation type with the highest survival probability displayed was the Non-stop variant (~0.93), followed by the In-frame insertion (~0.87), nonsense mutation and missense mutation (~0.75), frameshift deletion (~0.7), splice site (~0.67), frameshift insertion (~0.65), and lastly, translation start site and in-frame deletion as the mutation types with the lowest survival probability (~0.625).

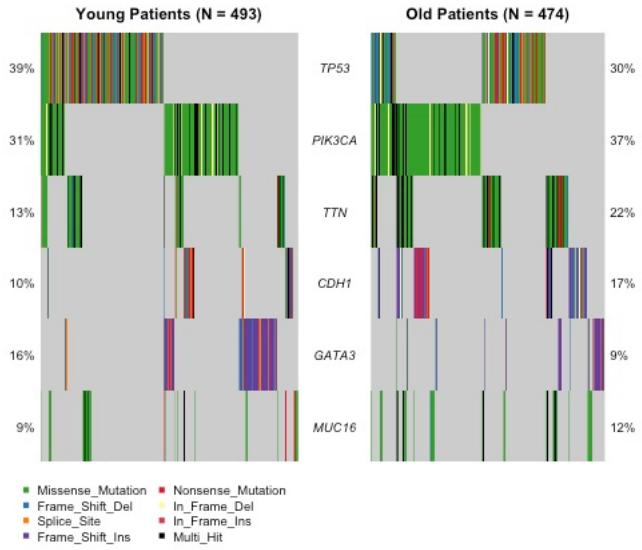

Figure 3: Co-Oncoplot showing the distribution of the most frequently mutated genes and their mutation types across age groups of 'Young' and 'Old' split across the median age of 58. TP53 and PIK3CA are the most mutated genes shown; there are more TP53 mutations in younger patients than older, and there are more PIK3CA mutations in older patients than younger.

Figure 3, the co-Oncoplot, shows the distribution of most mutated genes and mutation types among young and old patients. TP53 is the most mutated in both age groups – TP53 mutations consist of 39% of mutations in young patients compared to older patients (30%); GATA3 also has a higher mutation percentage in younger people (16% vs. 9%). For PIK3CA, TTN, CDH1, and MUC16, older patients have higher percentages of mutation. (37% vs. 31%, 22% vs. 13%, 17% vs. 10%, 12% vs. 9% respectively).

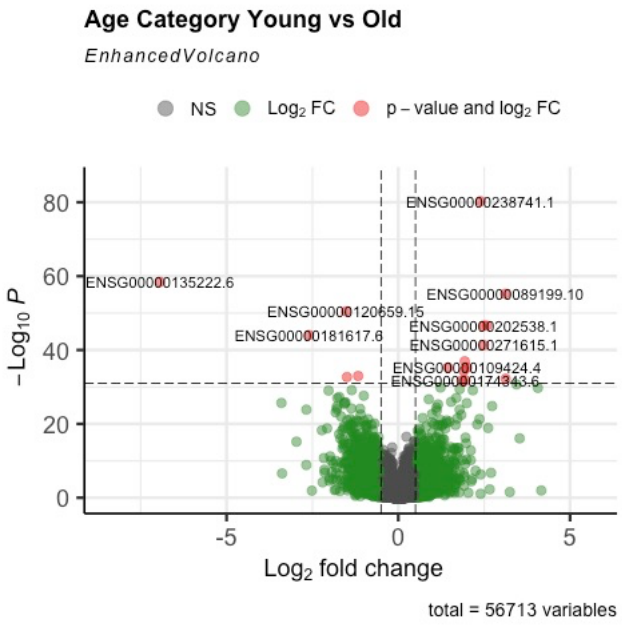

Figure 4: Enhanced Volcano Plot showing upregulated and downregulated genes in older patients compared to younger patients. The gene IDs present in the chart were cross-referenced with the significant upregulated and downregulated data frames to identify which gene was overexpressed or underexpressed.

	gene_name	gene_id	log2foldchange	pvalue	padj	-log10(padj)
ENSG00000135222.6	CSN2	ENSG00000135222.6	-6.934765	3.492842e-59	7.216036e-55	54.141701
ENSG00000171209.3	CSN3	ENSG00000171209.3	-3.402970	2.136697e-26	2.675339e-23	22.572621
ENSG00000234124.6	CSN1S2AP	ENSG00000234124.6	-3.384910	2.532826e-07	7.534475e-06	5.122947
ENSG00000167531.6	LALBA	ENSG00000167531.6	-2.963995	6.638271e-16	1.931597e-13	12.714084
ENSG00000196228.3	SULT1C3	ENSG00000196228.3	-2.679336	1.370178e-09	8.915653e-08	7.049847
ENSG00000126545.14	CSN1S1	ENSG00000126545.14	-2.676603	1.360822e-24	1.302127e-21	20.885347

Figure 5: Top 5 downregulated genes from the data frame down_sig_results after differential expression was run, viewed in RStudio with the command head(down_sig_results). As shown, CSN2 is the most significantly downregulated gene with an adjusted p-value of ~7.2e-55.

	gene_name	gene_id	log2foldchange	pvalue	padj	-log10(padj)
ENSG00000238741.1	SCARNA7	ENSG00000238741.1	2.397201	5.249697e-81	2.169122e-76	75.663716
ENSG00000089199.10	CHGB	ENSG00000089199.10	3.114464	5.791073e-56	7.976045e-52	51.098212
ENSG00000252010.1	SCARNA5	ENSG00000252010.1	2.579118	2.408878e-47	1.658874e-43	42.780187
ENSG00000202538.1	RNU4-2	ENSG00000202538.1	2.487174	2.394778e-47	1.658874e-43	42.780187
ENSG00000271615.1	ACTG1P22	ENSG00000271615.1	2.493356	4.587472e-42	2.369372e-38	37.625367

Figure 6: Top 5 upregulated genes from the data frame up_sig_results after differential expression was run, viewed in RStudio with the command View(up_sig_results). As shown, SCARNA7 is the most significantly upregulated gene with an adjusted p-value of ~2.17e-76.

The Enhanced Volcano Plot, as well as the upregulation and downregulation table results

(Figures 4, 5, 6) show genes that are either upregulated or downregulated in the older subset

compared to the younger subset. The plot shows gene ID ENSG00000238741.1, corresponding to the gene SCARNA7, is significantly upregulated in older populations versus younger populations, with a p-adjusted value of 2.17e-76. Other upregulated genes shown in the plot are CHGB (p-adjusted value 5.8e-56), RNU4-2 (p-adjusted value 1.66e-47), ACTG1P22 (p-adjusted value 2.37e-42), UCP1 (p-adjusted value 2.1e-32), and CHRNA9 (p-adjusted value 4.1e-32). Genes that are downregulated in older patients compared to younger patients include CSN2 (p-adjusted value 7.2e-55), FDCSP (p-adjusted value 5.33e-45), and TNFSF11 (p-adjusted value 3.51e-47).

Discussion

One of the goals of the study was to find a correlation between type of mutation and age: whether an age group had a higher incidence of a specific mutation type. However, the data from the box-plot shows that all mutation types are almost equally highly incidental in older age groups, and less present in younger age groups. This is undoubtedly because of a bias in the data, with most of the data samples present in the TCGA being of older ages, likely due to the fact that most breast cancer patients are in fact older in age. However, the fact that all median ages were ~60 and most 25^{th} percentile values were above age 50 still supports the higher incidence of breast cancer in older populations over younger populations. The median age calculated in the study was 58 and the separation of patients as young and old was based on this number - so it does not provide as much insight into patients considered ‘traditionally younger,’ age 40s and below.

Figure 2, the Kaplan-Meier plot showing the probability of survival, showed no pattern of having a set of similar mutations (insertions versus deletions, frameshift versus in-frame) having a higher or lower probability of survival; for example, frameshift insertions had a lower survival

probability than frameshift deletions, whereas in contrast, in-frame insertions had a higher probability of survival than in-frame deletions. The figure does, however, provide insight on which type of mutation could potentially be more fatal. Further research into causation for each type of mutation may provide insight on treatment, screening, diagnosis, and prevention of breast cancer. Ultimately, however, there is less of a correlation between just the type of mutation and fatality in itself, as the specific gene being mutated would factor in.

The data from the co-Oncoplot corresponds with other studies (Pereira et. al., Chavarri-Guerra et. al.) that show that TP53 and PIK3CA, both tumor-suppressor genes, are often mutated in cases of breast cancer – with the TP53 percentage within 5% of the Pereira et. al study (35%). It supports that PIK3CA (40.1% of mutations in Pereira et. al) mutation is significantly associated with older age groups (79% of sample in Kalinsky et. al. with PIK3CA mutation were post-menopausal). The study also showed that TP53 mutations were 9% more common in younger patients than older patients – though this goes against the data showing that breast cancer risk increases with age, germline TP53 mutations are known to have earlier ages of onset (median age of diagnosis is 34), and there is an 85% chance that a female patient with a TP53 mutation will develop breast cancer by age 60 (Schon and Tischkowitz). Figures 4 – 6, representing the DESeq2, support that SCARNA7 was significantly upregulated – also shown in Lin et. al. to be overexpressed and highly methylated, and not silence gene expression. CSN2, shown to be downregulated in older patients, is shown to be a biomarker for breast cancer and a potential tumor suppressor gene, with reduced expression (Zhu et. al).

It is already widely known that increasing age is associated with increasing risk of cancer. However, as society faces a shift towards living longer and faces an aging crisis (WHO), and a dramatic rise in early-onset cancers continues to occur with a variety of risk factors (Ugai

et. al.), it is important to understand further links between age and mutations that can cause cancer, and what factors are associated with those mutations (if they are epigenetic, resulting from lifestyle habits, etc.) in order to be better informed about prevention.

References

Schon, K., & Tischkowitz, M. (2018). Clinical implications of germline mutations in breast cancer: TP53. *Breast cancer research and treatment*, *167*(2), 417–423. https://doi.org/10.1007/s10549-017-4531-y

Kalinsky, K., Jacks, L. M., Heguy, A., Patil, S., Drobnjak, M., Bhanot, U. K., ... & Moynahan, M. E. (2009). PIK3CA Mutation Associates with Improved Outcome in Breast CancerPIK3CA Mutation and Improved Outcome in Breast Cancer. *Clinical cancer research*, *15*(16), 5049-5059.

Chavarri-Guerra, Y., Hendricks, C. B., Brown, S., Marcum, C., Hander, M., Segota, Z. E., Hake, C., Sand, S., Slavin, T. P., Hurria, A., Soto-Perez-de-Celis, E., Nehoray, B., Blankstein, K. B., Blazer, K. R., Weitzel, J. N., & Clinical Cancer Genomics Community Research Network (2019). The Burden of Breast Cancer Predisposition Variants Across The Age Spectrum Among 10 000 Patients. *Journal of the American Geriatrics Society*, *67*(5), 884–888. https://doi.org/10.1111/jgs.15937

Pereira, B., Chin, S. F., Rueda, O. M., Vollan, H. K., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S. J., Tsui, D. W., Liu, B., Dawson, S. J., Abraham, J., Northen, H., Peden, J. F., Mukherjee, A., Turashvili, G., Green, A. R., McKinney, S., … Caldas, C. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature communications*, *7*, 11479. https://doi.org/10.1038/ncomms11479*Breast cancer risk in American women*. National Cancer Institute. (n.d.). Retrieved March 22, 2023, from https://www.cancer.gov/types/breast/risk-fact-sheet#r1

American Cancer Society. (n.d.). *Breast cancer statistics: How common is breast cancer?* Breast Cancer Statistics | How Common Is Breast Cancer? Retrieved March 22, 2023, from https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

Lin, I. H., Chen, D. T., Chang, Y. F., Lee, Y. L., Su, C. H., Cheng, C., Tsai, Y. C., Ng, S. C., Chen, H. T., Lee, M. C., Chen, H. W., Suen, S. H., Chen, Y. C., Liu, T. T., Chang, C. H., & Hsu, M. T. (2015). Hierarchical clustering of breast cancer methylomes revealed differentially methylated and expressed breast cancer genes. *PloS one*, *10*(2), e0118453. https://doi.org/10.1371/journal.pone.0118453

Amjad, E., Asnaashari, S., Sokouti, B., & Dastmalchi, S. (2020). Systems biology comprehensive analysis on breast cancer for identification of key gene modules and genes associated with TNM-based clinical stages. *Scientific reports*, *10*(1), 10816. https://doi.org/10.1038/s41598-020-67643-w

Ugai, T., Sasamoto, N., Lee, H.-Y., Ando, M., Song, M., Tamimi, R. M., Kawachi, I., Campbell, P. T., Giovannucci, E. L., Weiderpass, E., Rebbeck, T. R. & Ogino, S. (2022). Is early-onset cancer an emerging global epidemic? Current evidence and future implications. *Nature Reviews Clinical Oncology*, *19*(10), 656–673. https://doi.org/10.1038/s41571-022-00672-8

M Steverson. (2022, October 1). *Ageing and health*. World Health Organization. Retrieved March 22, 2023, from https://www.who.int/news-room/fact-sheets/detail/ageing-and-health