Differential Analysis of Cancerous and Non-Cancerous Genes in Younger and Older Female Breast Cancer Patients

Abstract:

Epidemiological studies suggest that younger women with breast cancer have poorer survival outcomes compared to older women with breast cancer, due to having larger tumors, more positive lymph nodes, and negative steroid hormone receptors (Albain et al., n.d.). We performed a differential expression analysis (DE analysis) using the Cancer Genome Altas - Breast Carcinoma (TCGA-BRCA) dataset to discern which breast cancer-associated genes are expressed in women 46 years of age or younger, and women 60 years of age or older. TCGA is a comprehensive project that achieved the feat of “molecularly characterizing over 20,000 primary cancers and matching normal samples spanning 33 cancer types” (The National Cancer Institute, 2022).

The DE analysis revealed different SCARNA, SNORA, and SNORD types of genes were upregulated in the control group. Further research of these genes revealed that they belong to a class of RNA called snoRNAs, which are responsible for cell proliferation and tumor progression in cancer (D'Souza et al., 2021). Furthermore, the most differentially expressed gene, SCARNA5, was confirmed as being connected to breast cancer-related gene expression signatures in prior studies (*Bentz et al*., 2010). Our findings may lead to the identification of genes of interest that researchers could study to create novel breast cancer treatments or utilize to improve current therapies and diagnostic techniques. Additionally, this allows for a better understanding of which genes are upregulated in older versus younger female breast cancer patients and provides insight into which BRCA genes may be inherited and cause breast cancer.

Methods:

Querying & Downloading:

The TCGA-BRCA (https://www.cancer.gov/tcga) dataset was obtained from the GDC data portal using the *GDCquery*() function, which could be used to download controlled and open-access datasets from the GDC portal (*Colaprico*, 2020). First, we downloaded a series of Bioconductor packages on RStudio, including TCGAbiolinks (*Colaprico et al.*, 2016), DESeq2 (*Love et al.*, 2014), biomaRt (*Durinck et al.*, 2005), ggplot2 (*Wickham*, 2016), grepel (*Slowikowski*, 2021), scatterD3 (*Barnier et al.*, 2021), and SummarizedExperiment (*Morgan et. al.*, 2021). These packages have functions that would allow for the downloading, viewing, filtering, and analysis of the TCGA-BRCA data. We used the *GDCquery*() function of the TCGAbiolinks package to filter the data to be downloaded based on the arguments: **TCGA project, data.category, workflow.type, data.type,** and **experimental strategy**, with stipulations set for each argument to allow for downloading the TCGA-BRCA data’s raw read counts (Bioinformatics Pipeline: MRNA Analysis - GDC Docs, n.d.-b), such as “*Transcriptome Profiling*” for data.category, “*STAR-Counts*” for workflow.type, “*Gene Expression Quantification*” for data.type, and “*RNA Seq*” for experimental strategy (Colaprico et al., 2016; Noushmehr, n.d.). We filtered the query to obtain our dataset to obtain only duplicate samples to ease the downloading process as computer memory resources limited how much data we could download. The “*Metastatic*” tumor types were also filtered from the overall dataset as we were only interested in analyzing the “*Primary Tumor*” and “*Solid Tissue Normal*” samples. This filtered data query was then downloaded in 50 file chunks using the *GDCDownload*() function (Colaprico et al., 2016) and then converted to an R object using the *GDCprepare*() function (Colaprico et al., 2016), after which it was saved to an output R file using the built-in *save*() function.

Data Preparation & Subsetting:

Before running the differential expression analysis, the data had to be further subsetted to complete its preparation as one of our key analysis variables; the “*age_at_diagnosis*” category contained NA values that required removal to be applicable in our analysis. Our dataset was additionally subsetted to include only female samples and tissue samples under the “*Primary Tumor*” and “*Solid Tissue Normal*” categories. The potentiality for male samples to adversely impact the results of our analysis led us to control for this confound, as well as metastatic tumor cases, because these cases fell outside the scope of our inquiry. The “*age_at_diagnosis*” variable was initially quantified in days but was instead manually scaled by 365.25 to more readily interpret the age of samples involved in our study. This step was also done so that the “age_at_diagnosis” could be split into categorical variables which would represent our manipulated and control groups, which were groups of females below the age of 46 and females equal to or greater than the age of 60, respectively. After this, the samples for both tissue groups were subsetted to exclude genes containing loci with a high frequency of counts below 50 for both our control and manipulated condition: the ages above 60 and the ages below or equal to 46 years old, respectively. This was performed using the *assays*() function, which was limited to the scope of our inquiry (*Morgan et al.*, 2021). Finally, the mean of the read counts was obtained for each gene using the *rowMeans*() function and compared to a threshold of 0.5, which allowed us to keep the samples that exceeded this mean loci count threshold and exclude the rest. Our data object was then subsetted using the logic mentioned above and saved to an output R file using the *save*() function.

Running the Differential Analysis:

The contents of the subsetted SummarizedExperiment object were loaded, and a comparison column that would contain the labels of our control and manipulated groups for our differential expression analysis was created by adding a column labeled "*comp*" to our object. First, we subsetted our data object only to contain the ages that met our criteria of being greater than 60 years or at most 46 years. Then an "*age.categories*" column was created and filled with the following labels: "*greaterthan60*" or "*lessthan46*" based on the age criteria that the corresponding data value in the "*age_at_diagnosis*" met. This column's results were then saved to the "*comp*" column. The "*greaterthan60*" classification was manually set as our control group, and the "*lessthan46*" classification became our manipulated group to be used as such in our differential sequence analysis. The *DESeqDataSet*() function was used to convert our data object into an input acceptable for use in the *DESEQ*() function (*Love et al*., 2014). The *results*() function was used to organize the results of our differential sequence analysis.

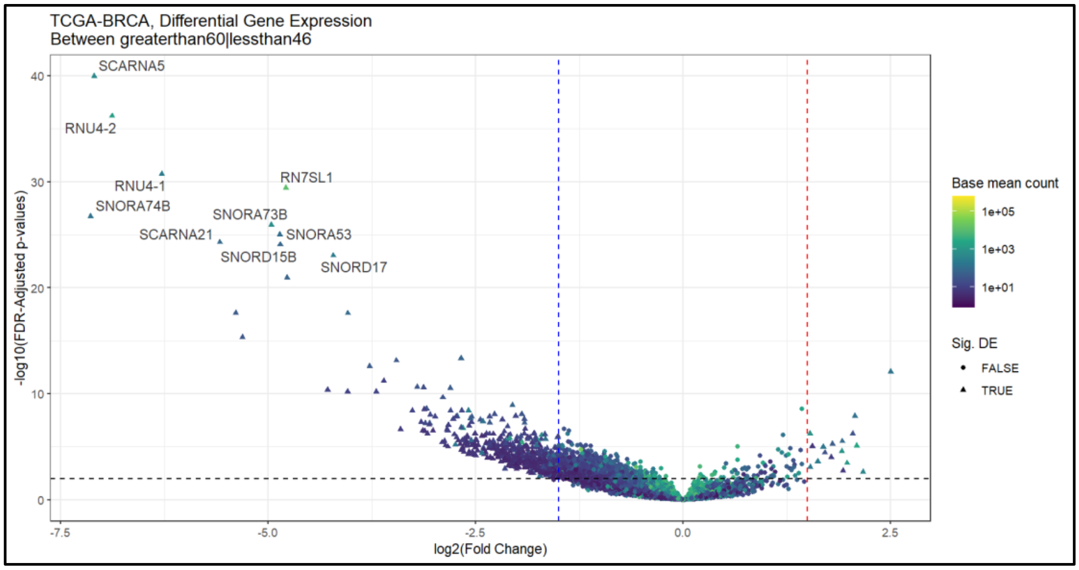
Analyzing the Differential Analysis Results & Further Visualization:

We created a volcano plot to observe the fold change in expression levels for the range of loci included in our study and the statistical significance of that change. The ggplot2 package (*Wickham*, 2016) was used to visualize the results of our analysis by plotting the log2 fold change of the genes against the -log10(adjusted p-values). These values were then compared against a minimum log fold difference and a maximum false discovery rate adjustment set to 0.01 and 1.50, respectively (*Wickham*, 2016). The genes labeled in the plot by utilizing the *ggrepel*() function correspond to the ten most differentially expressed genes between our manipulated and control groups (*Slowikowski*, 2021). A single gene distribution was created using the *ggplot*() function and filled by selecting the most differentially expressed gene from our differential expression results, accessing its information using the *rownames*() function, and subsequently ranking and scaling its loci based on how much change was observed. Finally, a principal component analysis plot of our final subsetted data sample was created using the "*scatterD3*"() package to visualize any samples in our dataset that should have been further removed from our final subsetted dataset due to extreme variance in comparison to the majority of other samples (*Antonio et al.*, 2016; *Slowikowski*, 2021; *Wickham*, 2016).

Results:

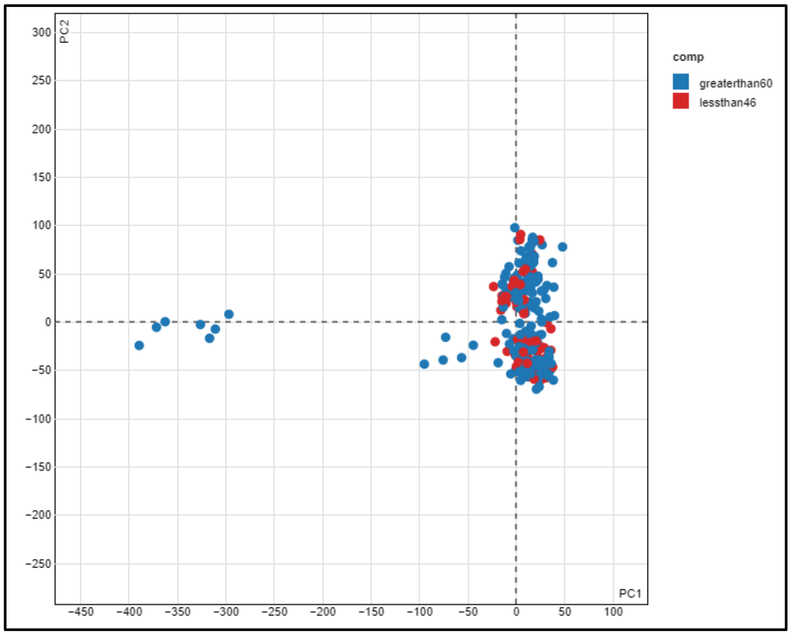
The DESeq analysis compared a control group with 102 samples to a manipulated group containing 63 samples. The control group and the manipulated group were limited to female samples and split based on age: the control treatment contained samples from women who were at least 60 years of age at the time of their breast cancer diagnosis, and the manipulated treatment was limited to women who were at most 46 years of age at their time of diagnosis. The analysis results identified 20,646 statistically significant genes compared to 10,030 statistically insignificant genes, and these associated values are tabulated in **Table 2**. Of the statistically significant genes, 4,673 genes were upregulated, and 15,973 were downregulated **Table 2**. In (**Figure 1),** the most differentially expressed genes are located towards the upper left portion of the graph and have labels with their corresponding gene name. The genes with a statistically significant p-value have a -log10(FDR-adjusted p-value) greater than the manually set cutoff of 0.01, which is represented by the horizontal dotted line in the volcano plot. The genes with a log2FoldChange exceeding the cutoffs of 0.01 and 1.50 were analyzed in **Table 2**.

***Figure 1:** Volcano plot showing the differential expression of BRCA genes in females below or equal to 46 and above the age of 60*



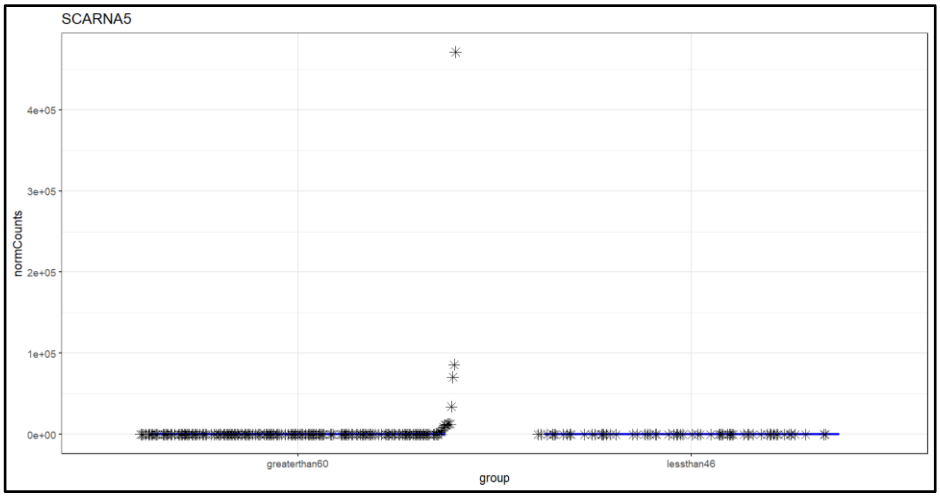
A principal component analysis (PCA) was run to visualize which components in our dataset explained the most variance in our two comparison groups (Nolan Bentley, personal communication). **Figure 2** illustrates that two principal components can explain the majority of the variance in our control and manipulated group; however, there are samples in our dataset whose variance is explained to a lesser degree by the previously mentioned components (Nolan Bentley, personal communication).

Figure 2: *PCA plot displaying the variance in our manipulated and control groups as explained by two principal components*



In (**Figure 3)**, the most differentially expressed gene between the two comparison groups is SCARNA5, and a boxplot was constructed to visualize the normalized counts of this gene in the context of each respective treatment: manipulated and control. The normalized counts act as an indicator of each comparison group's behavior regarding whether they upregulated or downregulated a particular gene.

Figure 3: *This boxplot visualizes the most differentially expressed gene from the analysis results, SCARNA5, in the sample for control: women at least 60 years of age or older, and manipulated: women aged less than 46 years*



The genes and their respective counts analyzed through the differential analysis are located in **Table 1**. The results of the analysis found that in the control group 4,673 of the significant genes are upregulated, and 15,973 of the significant genes are downregulated. **Table 2** lists the different categories of pAdj_log2 change and the genes in each category. The first TRUE indicates a significant difference in gene expression based on statistical analysis, and a second TRUE indicates that a gene is differentially expressed by a higher magnitude compared to other genes and *vice versa.*

Table 1: *This table contains the genes and samples in both conditions used for DESeq analysis after the data was subsetted*

Number of samples	After filtering
Number of genes being analyzed	30676
Number of women below or equal to the age of 46	63
Number of women above or equal to the age of 60	102

Table 2: *This table contains the counts of genes, their statistical significance, and whether they were highly upregulated or downregulated*

pAdj_log2Change	Number of Genes
TRUE_TRUE	4673
FALSE_TRUE	10030
TRUE_FALSE	15973

Analysis:

From **Table 2**, we observe that 4,673 genes are significant and upregulated, while 15,793 significant genes are downregulated. The volcano plot **(Figure 1)** labels the most upregulated genes in the control group, women sixty years or older, as being genes in the following categories: SCARNA, RNU, SNORA, SNORD, and RN. These gene categories belong to a category of RNA known as small nucleolar RNAs (snoRNAs) and are involved in RNA processing and modification. Different snoRNA host genes have snoRNAs present as introns, and these introns are spliced before transcription (*D'souza et al.*, 2021). These different snoRNA molecules go to the nucleolus or Cajal bodies in the cytoplasm, do post-transcriptional modifications to RNA and cause alternative splicing of mRNA before translation (*D'souza et al.*, 2021). SnoRNAs can act as diagnostic or prognostic markers and regulators of gene expression in breast cancer due to mutations that cause their expression patterns to be altered which increases abnormal cellular behavior that we recognize as the hallmarks of cancer: cell proliferation, cell stemness, drug resistance, disease recurrence, and activation or suppression of critical signaling pathways, and leads to tumor progression (*D'souza et al.*, 2021). The genes most highly upregulated in the control group **(Figure 1)** are snoRNAs, suggesting that the increased expression of these genes contributes substantially to the growth of primary tumors in control group individuals. In (**Figure 2**), a PCA plot revealed that within the women at least 60 years of age, there are samples that differ from the overall variation in the group as indicated by the pattern of blue dots moving further from the main cluster of samples. No samples in women no older than 46 years of age have major differences in variation.

Furthermore, individual analysis of each upregulated gene in the volcano plot reveals the role each gene plays in breast cancer. For example, the SNHG3 gene housing SNORD17 **(Figure 1)** is upregulated and promotes breast cancer cell growth (*D'Souza et al.*, 2021). The differential expression of many SNORA genes, including the SNORA74B **(Figure 1)** gene, is also seen in triple-negative breast cancer (*TNBC*) patients. TNBC is a form of breast cancer in which the typical receptors found in other forms of breast cancer are not present, which significantly limits the available treatments that doctors can use to combat this form of the disease (*Guo. et al.*, 2018; *Centers for Disease Control and Prevention*, 2022). The SCARNA5 gene is upregulated in female-to-male transgender breast cancer patients and has been associated with breast cancer-related gene expression signatures. Interestingly, SCARNA5 is the most differentially expressed gene in the control group out of all the genes included in the differential expression analysis for both treatment groups. In **(Figure 3),** the normalized count boxplot of the most differentially expressed gene, SCARNA5 is consistent with what we observed in the volcano plot; SCARNA5 is upregulated in the control group and downregulated in the manipulated group (*Bentz et al.*, 2010). Therefore, this volcano plot suggests that one effective form of targeted breast cancer therapy could impact the snoRNA regions in snoRNA host genes and seek to mediate their expression levels to comparable levels found in normal, non-tumor tissues.

Our analysis contributed to the current understanding of the types of genes expressed in breast cancer patients and a pathway that researchers could follow to create novel breast cancer treatments. A promising method could be implementing screening techniques for snoRNA expression levels in individuals predisposed to developing breast cancer or falling into groups classified as at an elevated risk of developing breast cancer. Healthcare providers should implement more frequent screening of snoRNA levels for individuals with higher risk factors than others and consider these results in tandem with already accepted methods and metrics for determining one's status related to breast cancer diagnosis and prognosis. Additionally, if a patient can, maintaining knowledge of their family history allows medical professionals to develop a plan to monitor specific genetic factors and their levels based on whether the patient is predisposed to developing breast cancer and the degree of that predisposition. An interesting follow-up would be conducting a DE analysis on young breast cancer patients' primary and normal tissues and observing what genes tend to be upregulated or downregulated. This analysis could yield good insight into which genes researchers should study to develop new targeted therapies or improve current therapy measures that address breast cancer and the circumstances under which it tends to differentially manifest in younger female breast cancer patients versus older women.

Acknowledgments

The results written about in this paper are in whole or part based upon BRCA data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

References

Colaprico, A. T. C. S. (2020, November 8). *Query GDC data*. R Package Documentation. https://rdrr.io/bioc/TCGAbiolinks/man/GDCquery.html

Breast cancer-patient version. *National Cancer Institute*. (n.d.). Retrieved April 26, 2022, from https://www.cancer.gov/types/breast

The Cancer Genome Atlas Program. *National Cancer Institute*. (n.d.). Retrieved April 26, 2022, from https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70 (2012). https://doi.org/10.1038/nature11412

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., … Perou, C. M. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, *163*(2), 506–519. doi:10.1016/j.cell.2015.09.033

Hulka, B., & Stark, A. (1995). Breast cancer: cause and prevention. *The Lancet*, *346*(8979), 883–887. https://doi.org/10.1016/s0140-6736(95)92713-1

Albain, K. S., Allred, D. C., & Clark, G. M. (1994). *Journal of the National Cancer Institute*. Google Books. https://books.google.nl/books?hl=en&lr=&id=SaVrAAAAMAAJ&oi=fnd&pg=PA35&ots=X6qDt_pZyO&sig=L95jdND6qlAb4BIw0RCX74sHE2Q&redir_esc=y#v=onepage&q&f=false

Antonio Colaprico, Tiago Chedraoui Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais Sabedot, Tathiane Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi Houtan Noushmehr. TCGAbiolinks: *An R/Bioconductor package for integrative analysis of TCGA data Nucleic Acids Research* (05 May 2016) 44 (8): e71. (doi:10.1093/nar/gkv1507)

Bentz, E. K., Pils, D., Bilban, M., Kaufmann, U., Hefler, L. A., Reinthaller, A., Singer, C. F., Huber, J. C., Horvat, R., & Tempfer, C. B. (2010). Gene expression signatures of breast tissue before and after cross-sex hormone therapy in female-to-male transsexuals. *Fertility and Sterility*, *94*(7), 2688–2696. https://doi.org/10.1016/j.fertnstert.2010.04.024

Bioinformatics Pipeline: mRNA Analysis - GDC Docs. (n.d.-a). Docs.gdc.cancer.gov; National Cancer Institute: GDC Documentation. Retrieved April 15, 2022, from https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/

BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma and Wolfgang Huber, Bioinformatics 21, 3439-3440 (2005).

Dsouza, V. L., Adiga, D., Sriharikrishnaa, S., Suresh, P. S., Chatterjee, A., & Kabekkodu, S. P. (2021). Small nucleolar RNA and its potential role in breast cancer – A comprehensive review. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1875*(1), 188501. https://doi.org/10.1016/j.bbcan.2020.188501

Guo, Y., Yu, H., Wang, J., Sheng, Q., Zhao, S., Lehmann, B. D., & Zhao, Y. (2018, January 10). *The Landscape of Small Non-Coding RNAs in Triple-Negative Breast Cancer*. MDPI. https://www.mdpi.com/2073-4425/9/1/29/htm

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016.

Johnson, R. H., Hu, P., Fan, C., & Anders, C. K. (2015). *Gene expression in "young adult type" breast cancer: a retrospective analysis.* Oncotarget, 6(15), 13688–13702. https://doi.org/10.18632/oncotarget.4051

Julien Barnier, Kent Russell, Mike Bostock, Susie Lu, Speros Kokenes and Evan Wang (2021). *scatterD3: D3 JavaScript Scatterplot from R. R package version 1.0.1.* https://CRAN.R-project.org/package=scatterD3

Kamil Slowikowski (2021). *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*. R package version 0.9.1. https://CRAN.R-project.org/package=ggrepel

Kirby, J. (n.d.-a). *TCGA-BRCA - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki* (T. Nolan, Ed.). Wiki.cancerimagingarchive.net. Retrieved April 15, 2022, from https://wiki.cancerimagingarchive.net/display/Public/TCGA-BRCA#:~:text=The%20Cancer%20Genome%20Atlas%20Breast

Love, M.I., Huber, W., Anders, S. *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology* 15(12):550 (2014)

Martin Morgan, Valerie Obenchain, Jim Hester and Hervé Pagès (2021). *SummarizedExperiment: SummarizedExperiment container. R package version* 1.24.0. https://bioconductor.org/packages/SummarizedExperiment

Noushmehr, A. C., Tiago Chedraoui Silva, Luciano Garofano, Catharina Olsen, Davide Garolini, Claudia Cava, Isabella Castiglioni, Thais Sarraf Sabedot, Tathiane Maistro Malta, Stefano Pagnotta, Michele Ceccarelli, Gianluca Bontempi, Houtan. (n.d.). *Working with TCGAbiolinks package*. Www.bioconductor.org. https://www.bioconductor.org/packages/3.3/bioc/vignettes/TCGAbiolinks/inst/doc/tcgaBiolinks.html

Principal Component Analysis for DESeq2 results. (n.da Geneious.
https://assets.geneious.com/manual/11.0/GeneiousManualsu102.html
https://assets.geneious.com/manual/11.0/GeneiousManualch11.html#GeneiousManualse51.html
https://assets.geneious.com/manual/11.0/GeneiousManualsu102.html

Triple-Negative Breast Cancer. (2022, March 9). Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/triple-negative.htm