Paper DV10

Visualizing the Complexities of Genomic Datasets

Vibha Raghavan, Hoffmann-La Roche Limited, Mississauga, Canada

ABSTRACT

The wealth of information contained in genomic datasets has the potential to drastically change the landscape of therapeutics and clinical trials. However, as the volume of data increases, so does the challenge of being able to mine these datasets for medically relevant insights. In the last few years, many visualization tools and R packages have been made available that are both user friendly and aid in analyzing and interpreting genomic data, including whole exome sequencing and copy number information, in meaningful ways. Their utility ranges from focused to broader applications, such as visualization of specific mutations within a gene to visualization of mutational landscapes of entire cohorts. Moreover, it is also possible to use these tools to visually compare your genomics data to other publicly available data. In this paper, I discuss select visualization tools and R packages used to interpret data in the ever-growing world of genomics.

INTRODUCTION

The advent of next-generation sequencing (NGS) technologies has revolutionized the field of genomics, which includes various types of data that provide information on alterations within a genome. Single nucleotide variants occur when there is a change in a nucleotide in the DNA sequence, while copy number alterations occur when there are changes to the structure of the chromosome which result in a gain or loss of sections of the DNA. In addition, RNA-seq is a method that captures the transcriptome profile of a biological specimen. However, as the influx of data increases, so does the challenge of being able to mine this data in meaningful ways. Data visualization tools offer a useful means to gain preliminary medical relevant insights from genomics datasets. In this paper, I will discuss three widely tools to mine and visualize multidimensional genomics data.

MAFTOOLS

Maftools is a user-friendly R Bioconductor package developed to analyze and visualize data generated by NGS [1]. It uses this data to perform common statistical and computational analyses in cancer genomics studies [1]. The *Maftools* package includes visualizations, such as, being able to view the mutational landscape of multiple patients, comparison of two cohorts by creating oncoplots and forest plots, viewing the location of mutations within a gene using lollipop plots, and mutational signature plots, to name a few [1] (Fig. 1). The input for this package is a Mutation Annotation Format (MAF) file, which stores information on somatic variants in the patients and can be generated using numerous variant annotation tools that are publicly available. In this paper, I go through some important functions of *Maftools* that can be applied to gain a better understanding of the genomic landscape of patients in clinical trials.

Fig. 1 Overview of the Maftools package pipeline

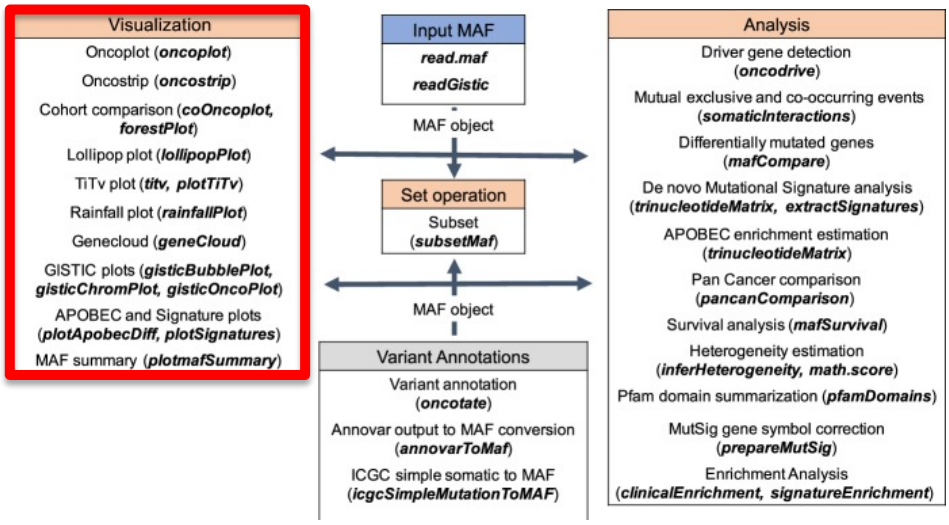

Figure credit [1]

SUMMARY VISUALIZATION

One of the first steps to characterize the genomic landscape of a cohort of patients is to get a visual summary of the cohort itself. The *plotmafSummary* function creates several plots that display key information about the patients (Fig. 2) [1]. It provides the user with bar plots showing the prevalence of the variant classification and type, as well as the distribution of variants within each sample. This can be used to identify patients with higher mutation frequencies and investigate them in further detail. Moreover, it identifies the top ten mutated genes in the cohort. If there is a certain gene of interest that appears in that list, it may prove to be an important biomarker for the disease under investigation. Hence, the *plotmafSummary* function is a useful visualization tool that provides the researcher with an overview of the cohort and, consequently, a starting point for more detailed biomarker analyses.

Fig. 2 Output from the plotmafSummary function providing the overall genomic landscape of a cohort of patients using the Maftools package

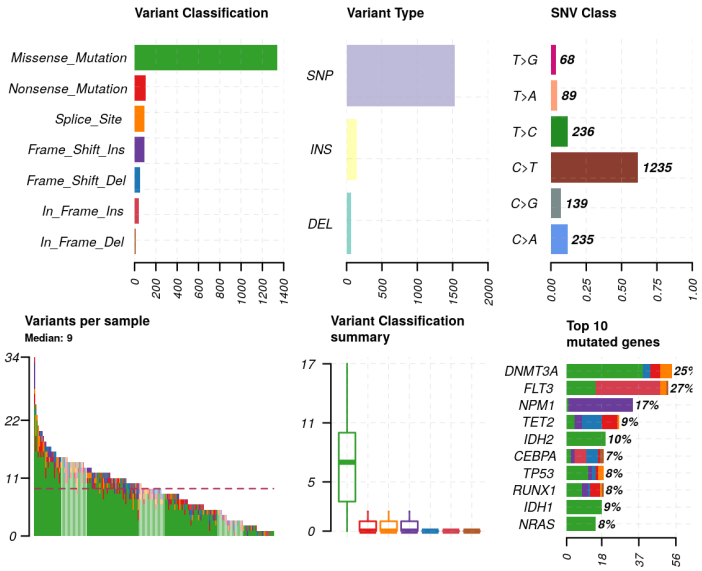

Figure credit [1]

ONCOPLOT

Oncoplots can be used to visualize the mutational landscape of patients using a list of specific genes of interest. Fig. 3 [1]. show an output of the *oncoplot* function, where each column is a patient and each row is a gene of interest, which is sorted by the mutation rate of that gene in the cohort. Each coloured cell indicates that the patient has a mutation in that gene and it is colour-coded by the type of mutation, which is depicted in the legend on the bottom left of the plot. In the example presented in Fig. 3, these genes are the top 10 mutated genes in the cohort. This specific oncoplot can be interpreted as follows – *FLT3* is the most frequently mutated gene in the cohort, following by *DNMT3A* and *NPM1*. *NPM1* is primarily composed of frameshift insertion type mutations and those mutations are mutually exclusive with mutations in *IDH2*. These observations could lead to questions such as – What role does *FLT3* play in these patients? Why are the mutations in *NPM1* and *IDH2* mutually exclusive? Could these top 10 genes be affecting certain pathways that may lead to these patients acquiring this disease?

Fig. 3 Oncoplot generated by the Maftools package showing the mutations present within a cohort of patients

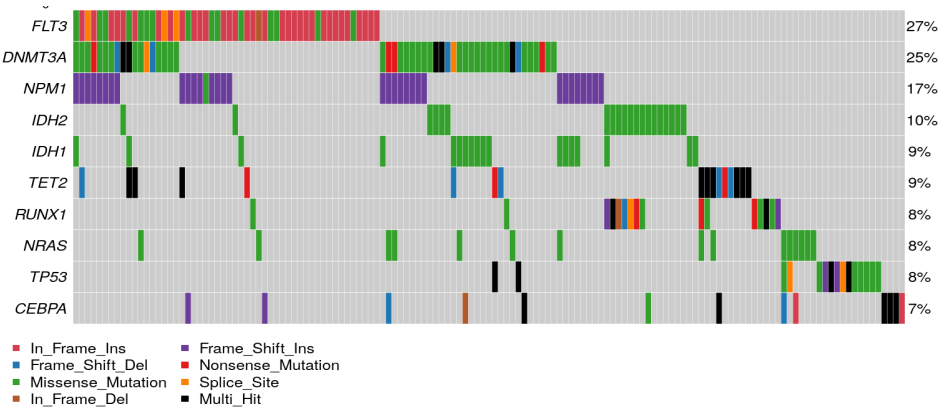

Figure credit [1]

COHORT COMPARISON

While an oncoplot can open doors to investigating numerous insights within a cohort in further detail, Fig. 4 illustrates an example of a *coOncoplot* function, which can be used to compare two different oncoplots [1]. Such a plot is very useful for visualizing comparisons, such as, patients in two different arms of a study, primary vs. relapse patients and, also, pre-treatment vs. post-treatment patients. The co-oncoplots in Fig. 4 show that while mutations in *FLT3* are more prevalent in the primary cohort, mutations in *PML* and *RARA* are more prevalent in the relapse cohort. Such observations can raise important questions such as - Could the *PML* and *RARA* mutations be the cause of relapse in these patients and make them resistant to the drug?

Fig. 4 Comparing the mutational landscape of two different cohorts by using the coOncoplot function from the Maftools package

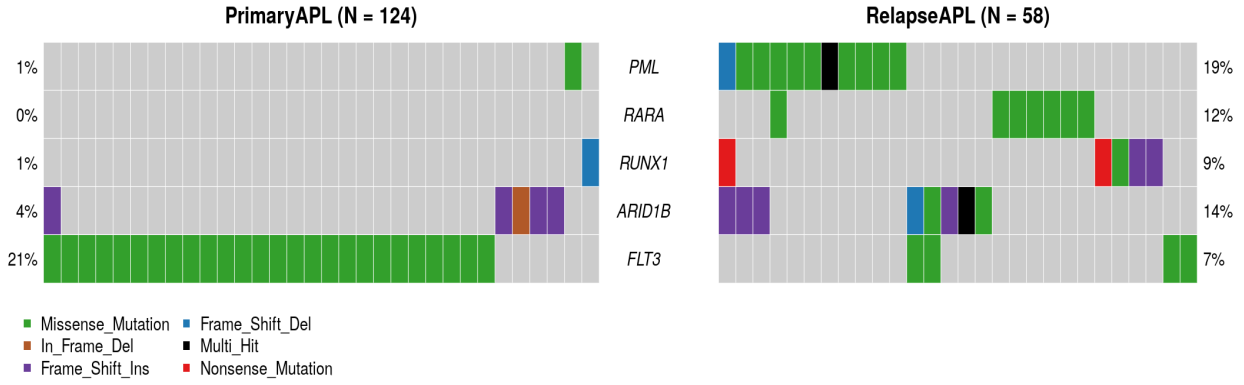

Figure credit [1]

These results can also be visualized using the *forestplot* function, as shown in Fig. 5 [1]. This function creates a forest plot which allows the user to compare the frequency of such mutations between two cohorts, while also taking into account the statistical significance of the results.

Fig. 5 Forest plot showing the statistics which result from comparing two cohorts of patients using the Maftools package

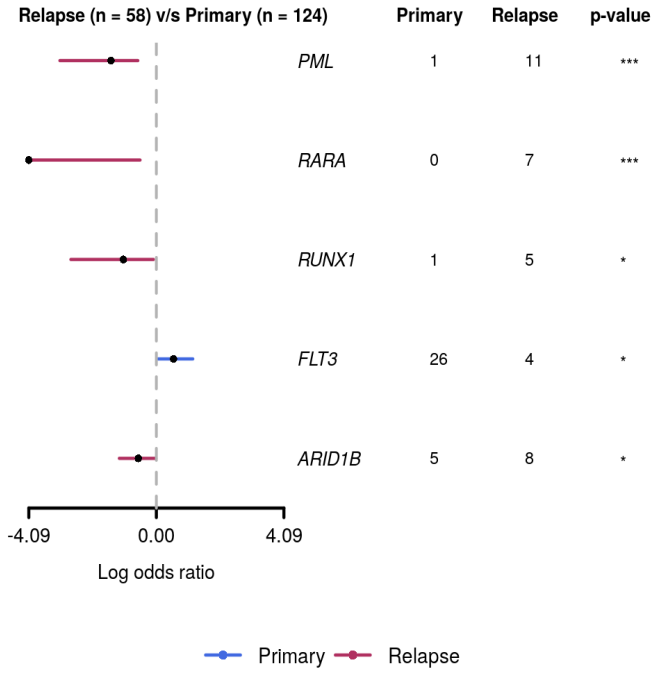

Figure credit [1]

COMPLEXHEATMAPS

Another R Bioconductor package used for visualizing high dimensional genomics data is *ComplexHeatmaps*. This is an excellent tool to create multiple parallel heatmaps along with multiple annotation graphics to provide a complete overview of the dataset [2]. It can be used to view genomic alterations of an entire cohort, expression profiles from RNA-seq data, and relationships between methylation, gene expression, and gene-related information [2].

GENE EXPRESSION MATRIX

Heatmap is a useful tool to visualize gene expression data. Each row represents a gene, each column represents a patient, and each cell in the heatmap is a visual representation of the relative expression of that gene in each patient. Using *ComplexHeatmaps*, we have the ability to add more information to a standard gene expression heatmap by including columns for various categories, such as gene type, base mean, gene length, etc. Fig. 6 is an illustration of a heatmap produced using the *Heatmap*, *HeatmapAnnotation*, and *rowAnnotation* functions offered in this R package [2]. The large heatmap on the left represents the gene expression values in the patients, while there are also single column heatmaps on the right that provide information on the base mean expression, gene length, and the gene type. The genes are grouped into five clusters from k-mean clustering, as shown by the coloured blocks on the very left of the heatmap [2]. Moreover, the two plots above the base mean and gene type columns indicate the distributions and statistics of the data points in the five clusters [2]. The relationships between the different plots and heatmaps in the same visual give the user useful overviews of the data that would otherwise need to be manually searched, such as which clusters of genes to further investigate, which sets of patients are clustered together based on their gene expression values, and which genes have a high or low base mean value across the different patients in the cohort, amongst others.

Fig. 6 Heatmap showing the mRNA expression of patients using the ComplexHeatmaps package

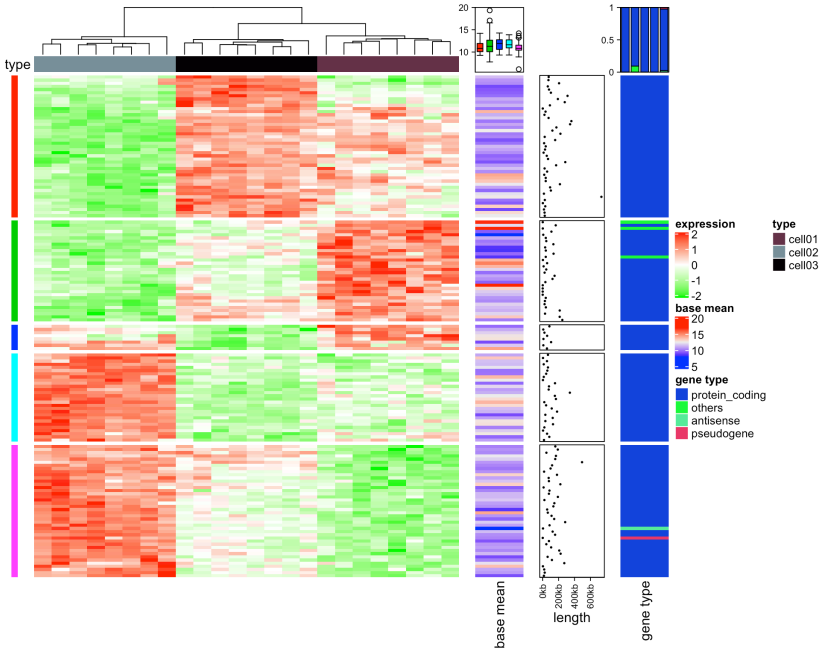

Figure credit [2]

CREATING DETAILED ONCOPRINTS

ComplexHeatmaps provides an exceptional approach to creating enhanced OncoPrints, which is another way to visualize genomic alterations across patients in the form of heatmaps [2]. Fig. 7 depicts a heatmap generated by the *oncoPrint* function, based on copy number and single nucleotide variations. The genes in the OncoPrint are grouped into low and high amplification (AMP) events. Green cells indicate the presence of a mutation, while the red and blue cells indicate the presence of an amplification and deletion, respectively. Relevant patient information, such as age, gender, and disease stage, are shown at the bottom of the large heatmap. The heatmap on the right represents the distribution of the genes in different pathways related to the disease. An OncoPrint can provide immediate insights into which genes have the highest frequency of alterations, which pathways are affected by those genes, are gender/age/stage enriched by alterations in certain genes, why are certain patients enriched in high AMP events, etc. The answers to these questions can, in turn, provide a better understanding of the disease biology and the landscape of alterations within patients.

Fig. 7 OncoPrint generated by the ComplexHeatmaps package

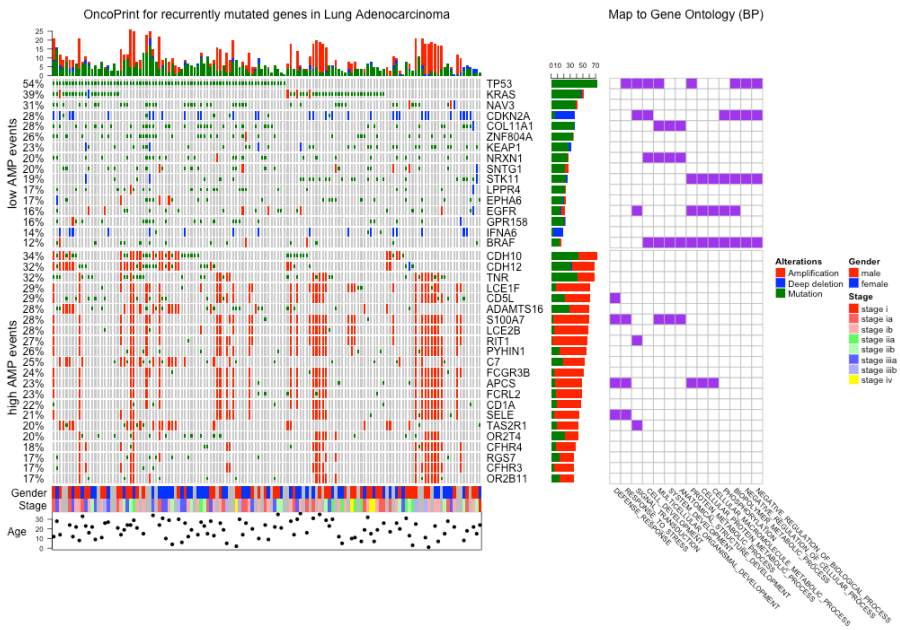

Figure credit [2]

CBIOPORTAL FOR CANCER GENOMICS

cBioPortal for Cancer Genomics is a web resource for accessing, analyzing, and viewing cancer genomics data from previously established studies [3]. It provides a graphic summary of genomic alterations in various genes across multiple patients, scatter plots of protein abundance versus mRNA expression for select genes, Kaplan-Meir plots for datasets which include survival data, and a visualization of networks that are altered in cancers [3]. These visualizations can be used to extrapolate information that can then be used to compare results to other studies. Users can start using the resource by either querying a single cancer study or multiple cancer studies [3]. Fig. 8 shows an example of the input required for querying a single cancer study. After selecting the study of interest, the user should select the genomic profiles, the patient/case set, and enter a list of genes of interest. Once the query is submitted, the user is presented with various tabs that have outputs related to the study and genes of interest.

Fig. 8 Example of input required for querying a single cancer study in cBioPortal

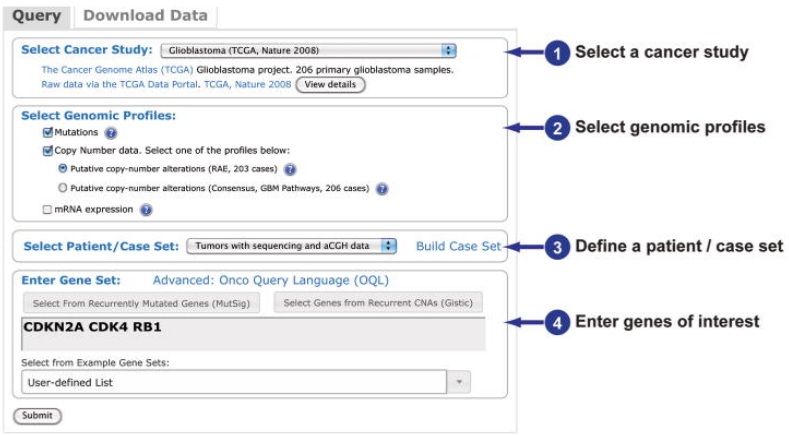

Figure credit [3]

One of the tabs provides information on mutual exclusivity for the genes entered in the query page. Fig. 9 shows that genes that alter RB signal in glioblastoma have a tendency to be mutually exclusive, while also showing which pairs are significant based on Fisher's exact test [3]. In this case, even though *CDK4* and *RB1* have the strongest tendency towards mutual exclusivity, it is not a statistically significant pair. However, the other two pairs that have some tendency towards mutual exclusivity are statistically significant [3].

Fig. 9 Tab showing mutual exclusivity across genes of interest in cBioPortal

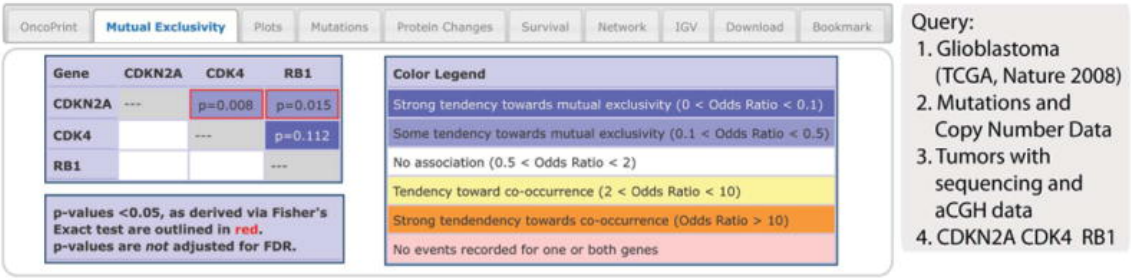

Figure credit [3]

Another tab plots the relationships between different genomic platforms. Fig. 10 shows two plots depicting the relationship between mRNA expression and copy number, as well as protein level and mRNA expression. The plot on the left can be interpreted as the gene *ERBB2* having higher mRNA expression in patients who are also amplified for that gene [3]. Meanwhile, the plot on the right suggests that *ERBB2* protein levels tend to be positively correlated with the gene expression levels [3].

Fig. 10 Plots depicting relationships between genomics and proteomics platforms in cBioPortal

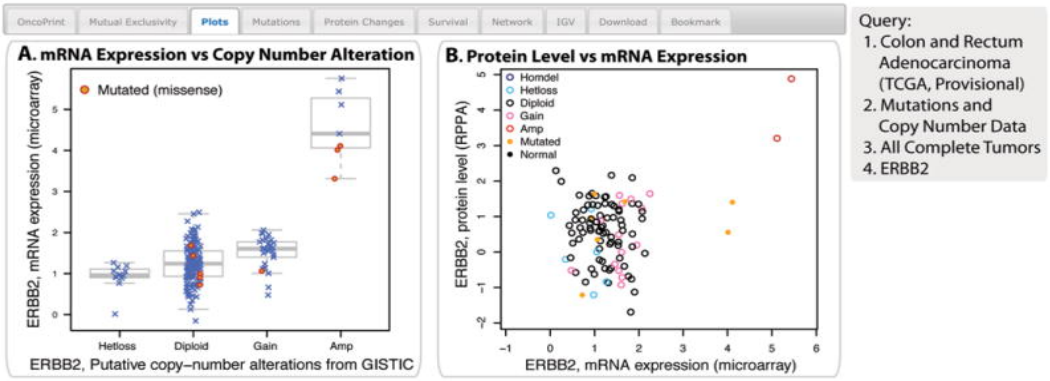

Figure credit [3]

Yet another tab plots Kaplan-Meier curves illustrating the results of survival analysis of the query population. Fig. 11 plots two survival curves – one showing the overall survival (left), and the other showing the disease-free survival of ovarian cancer patients with and without *BRCA1* or *BRCA2* (right) [3]. The red curves include patients that have a mutation in one of those two genes, while the blue curve shows patients that have no mutations in either of those genes.

Fig. 11 Survival analysis using Kaplan-Meier curves in cBioPortal

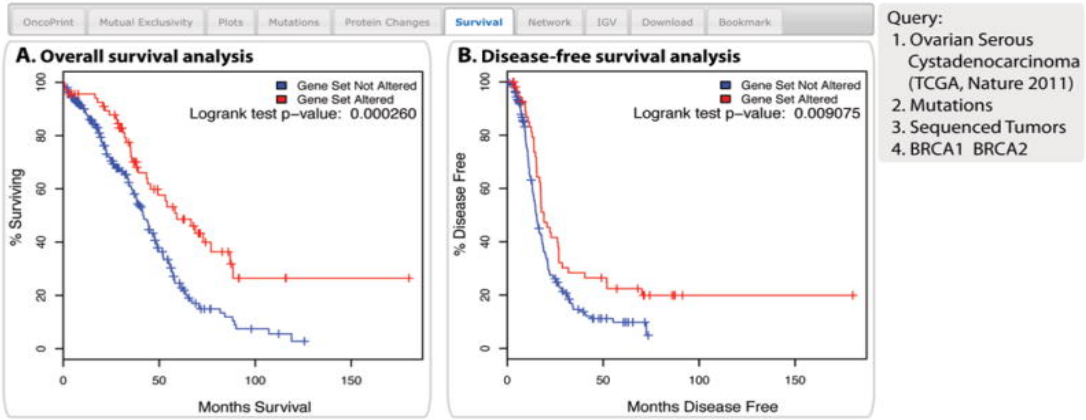

Figure credit [3]

CONCLUSION

These visualization tools provide an excellent approach to visualizing complex and multi-dimensional genomic datasets and gaining preliminary insights that can be leveraged for more detailed downstream investigations. They are user-friendly and are easily accessible. With these tools we can continue to explore the world of genomics and thereby change the future of therapeutics for the better.

REFERENCES

1. Mayakonda et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Research. 2018; 28(11):1747-1756.
2. Gu et al. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016; 32(18):2847-9. doi:10.1093/bioinformatics/btw313.
3. Gao et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science Signaling. 2013; 6(269):pl1. doi:10.1126/scisignal.2004088.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Vibha Raghavan
Hoffmann-La Roche Limited
7070 Mississauga Road
Mississauga, ON, L5N 5M8
Email: vibha.raghavan@roche.com