PhUSE US Connect 2018

Paper DV10

A Picture is Worth a Thousand Words; Analyzing Clinical Data with Sequencing Data

Nalin Tikoo, Roche Sequencing Solutions, Pleasanton, USA
Hans-Peter Adams, Roche Sequencing Solutions, Potsdam, Germany
Priya Saradha, Experis IT Services US, LLC, New Jersey, USA
Aarthi Balasubramanyam, Roche Sequencing Solutions, Pleasanton, USA
Mathias Heibeck, Roche Sequencing Solutions, Potsdam, Germany

ABSTRACT

Advances in sequencing technologies have paved the path to better understanding cancer, hereditary diseases, infectious diseases, and prenatal diagnostics. This includes sequencing of germline DNA, tumor tissue DNA, and plasma DNA, i.e. liquid biopsy. Monitoring using plasma DNA compared to traditional imaging based methods has become routine clinical practice thus producing data in gigabyte to terabyte range and resulting in thousands of markers. Visualization and analysis of these data in the context of clinical data for diagnosis, prognosis, assessment of response to therapy, and outcome measures can be a challenge.

Aside from an introduction into the clinical and sequencing data, this paper will present various methods of visualization using SAS® software and R for giving informative views on the complex data. These visualizations serve the purpose of data review and analysis. Code and output snippets will be provided for subject profiling, custom survival analyses, bar charts and an RShiny application.

INTRODUCTION

According to the Precision Medicine Initiative[1], precision medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person." Visualizing this massive amount of genomic data in conjunction with the clinical data is a challenge. The purpose of this paper is to showcase a few of these data visualizations that help us make sense of the data thus facilitating the development algorithms that further help with the patient care.

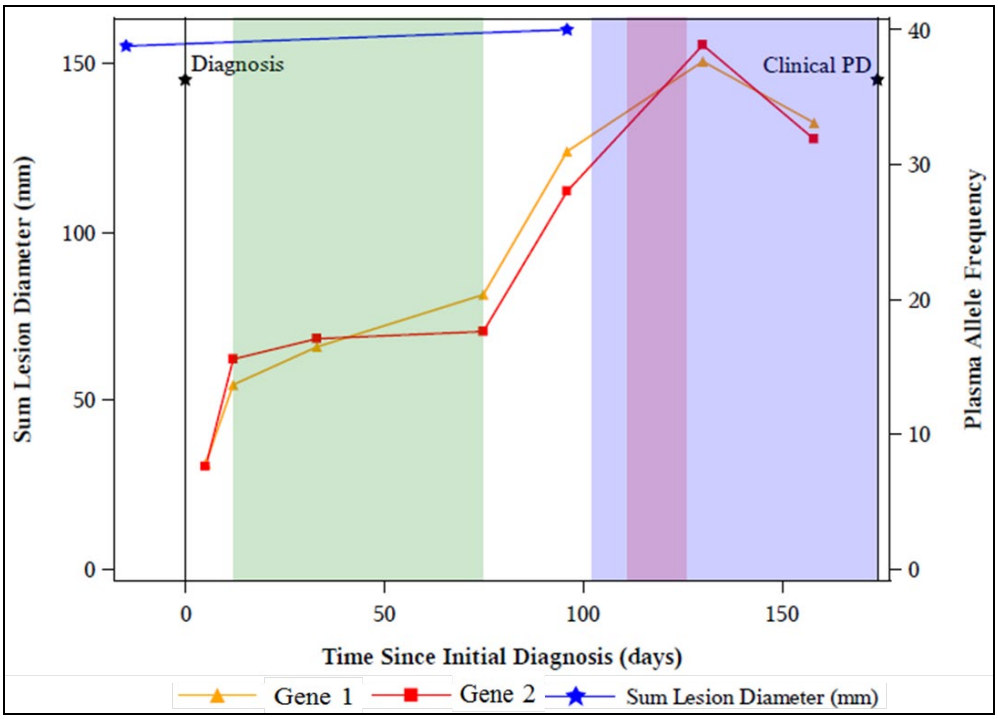
DATA

Clinical & Sequencing datasets were used for these visualizations. In terms of clinical data; demographics & baseline characteristics, Tissue & Plasma Sample collection, CT scans, treatment, treatment response measured by RECIST 1.1 and survival data were used. In tserms of sequencing data, single nucleotide variant (SNV) variant call files (VCF) were used. The structure of this dataset is 1 record per patient per sample per gene variant.

PATIENT PROFILES – DISEASE MONITORING

DESCRIPTION

The objective of this visualization is to review the patient journey from diagnosis to clinical progression; represented by vertical reference lines on the two ends. Mutation data is plotted on the right y-axis, CT Scan data on the left y-axis and the time on study on the x-axis. The treatment information is represented using the colored bands.



CODE SNIPPET

A macro was developed to loop through each subject's data; this was required to store subject specific values into macro variables to produce custom x-axis & y-axis value ranges & the lo and hi values for the treatment bands.

```
proc sgplot data=toplot_&i. noautolegend;

   series x=monthtime y=af_plasma / group=gene markers name="afp" legendlabel="Plasma Allele Frequency"
                                    datalabel=af_plasma y2axis;                                          (1)
   y2axis label='Plasma Allele Frequency' min=0 max=&maxy2axisval.;

   series x=monthtime y=sumlesdiameter / lineattrs=(pattern=solid color=blue) markers markerattrs=(symbol=starFilled color=blue)
                                         name="slesd" legendlabel="Sum Lesion Diameter (mm)";
   yaxis label='Sum Lesion Diameter (mm)' min=0 max=&maxy1axisval.;                                      (2)

   scatter x=monthtime y=gediag / markerattrs=(symbol=StarFilled color=black) datalabel=gediagc datalabelattrs=(size=10); 3a
   refline monthdiagdt / axis=x lineattrs=(color=black);

   scatter x=monthtime y=gepd / markerattrs=(symbol=StarFilled color=black) datalabel=gepdc datalabelattrs=(size=10);    3b
   refline monthpddt / axis=x lineattrs=(color=black);

   %if %sysevalf(&band1lo. ne .) %then %do;
      band y=x1 lower=&band1lo. upper=&band1hi. / fillattrs=(transparency=0.8 color=green);                 4a
   %end;

   %if %sysevalf(&band2lo. ne .) %then %do;
      band y=x2 lower=&band2lo. upper=&band2hi. / fillattrs=(transparency=0.8 color=red);                   4b
   %end;

   %if %sysevalf(&band3lo. ne .) %then %do;
      band y=x3 lower=&band3lo. upper=&band3hi. / fillattrs=(transparency=0.8 color=blue);                  4c
   %end;

   xaxis min=&minxaxisval. max=&maxxaxisval.;
   keylegend "afp" "slesd" / across=3 autoitemsize;
run;
```

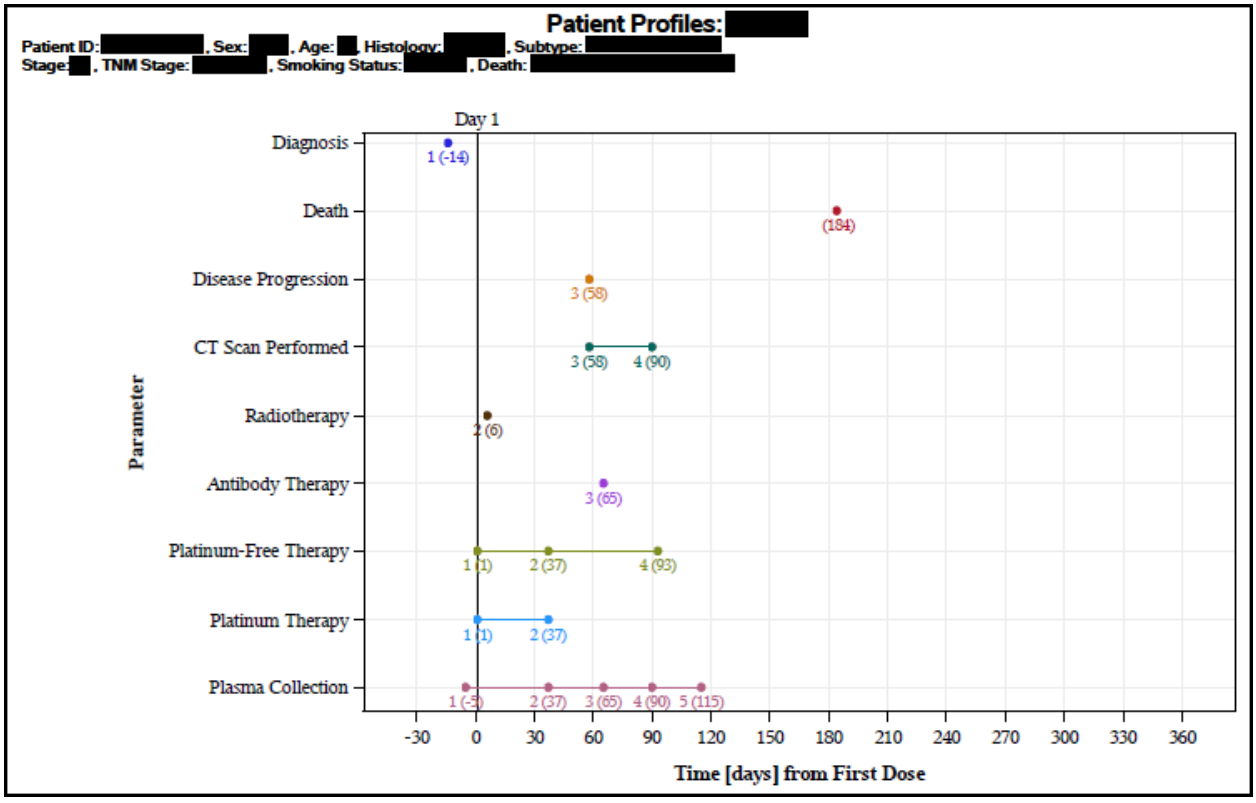
1. Right y-axis – Plasma Allele Frequency (%)
2. Left y-axis – Sum of Lesion Diameters (mm)
3. Vertical reference lines
    a. First reference line for Diagnosis Date
    b. Second reference line for Date of Disease Progression
4. Treatment bands – lower & upper bounds of the band are the number of days since diagnosis date (Day 0) to the start and end of each treatment respectively.
    a. Green band for chemotherapy
    b. Red band for radiation therapy

c. Blue band for targeted therapy

PATIENT PROFILES – DATA REVIEW FOR COMPLETENESS AND ACCURACY

DESCRIPTION

The objective of this visualization is to provide a visual of the expected events over time. This aids data management and clinical operations colleagues to query the implausible data points while helping clinicians and statisticians with developing imputation algorithms.



CODE SNIPPET

The key to this visualization is generating the dataset. The structure of the dataset needs to be 1 record per event with the number of days from first dose date to that event.

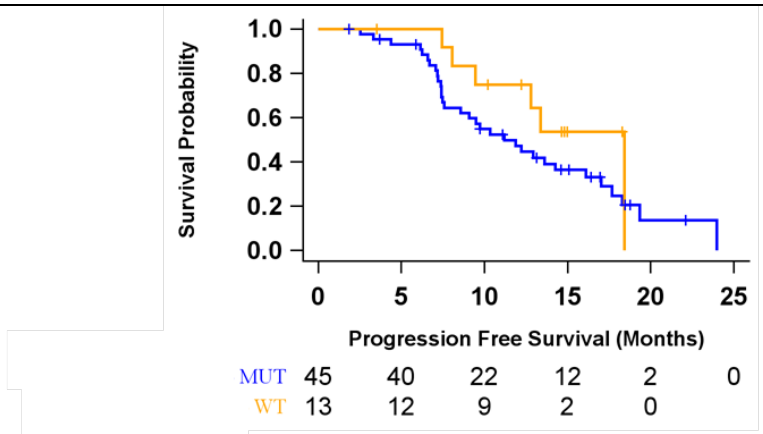
```
proc sgplot data=toplot_final_&i. noautolegend nocycleattrs;
   refline 1 / axis=x lineattrs=(thickness=1 color=black) label="Day 1";

   scatter x=aval y=param/markerattrs=(size=7px symbol=circlefilled) group=param
                          datalabel=avisit
                          datalabelpos=bottom
                          colormodel=(CX3288BD CX99D594 CXE6F598 CXFEE08B CXFC8D59 CXD53E4F);

   series x=aval y=param/group=param markers markerattrs=(symbol=circlefilled size=7px)
                         lineattrs=(pattern=solid);

   xaxis grid values=(-30 to 375 by 30);
   yaxis grid values=("Plasma Collection" "Platinum Therapy" "Platinum-Free Therapy" "Antibody Therapy"
                      "Radiotherapy" "CT Scan Performed" "Disease Progression" "Death" "Diagnosis");
run;
```

1. DATALABEL option is used to annotate the visit number and the number of days since first dose date.
2. DATALABELPOS option is used to provide the location of the annotation.

KAPLAN-MEIER CURVES

DESCRIPTION

The objective of this visualization is to present the K-M curves including relevant statistics on one PowerPoint slide. The K-M curves were customized to very specific specifications to meet publication requirements.



	Gene 1 MUT vs WT	
Description	**Gene 1 - MUT**	**Gene 1 - WT**
Number of Patients	xx	xx
Hazard Ratio & 95% CI (Relative to Strata B)	x.xxx (x.xxx, x.xxx)	
Log-Rank p-Value (Across Strata)		x.xxxx

CODE SNIPPET

Custom K-M Curve

The user guide[2] provides details on the usage of the COMPILESURVIVALTEMPLATES & PROVIDESURVIVALMACROS macros provided by SAS for producing highly customized K-M curves. The code snippets below highlight a few key customizations that we applied.

```
*** Custom Line Pattern & Colors ***;
%let GraphOpts = attrpriority=none DataContrastColors=(blue orange grey green)
                                   DataLinePatterns=(solid)
                                   DataColors=(blue orange grey green);

*** Custom Fonts ***;
%let yOptions = label="Survival Probability" shortlabel="Survival" labelattrs=(size=
                                             tickvalueattrs=(size=12pt weight=bold f
                                             linearopts=(viewmin=0 viewmax=1);
%let xOptions = shortlabel=XNAME offsetmin=.05 linearopts=(viewmax=MAXTIME tickvalue
                                               labelattrs=(size=10pt weight=bold fam
                                               tickvalueattrs=(size=12pt weight=bold

*** At Risk Table Numbers Color Size & Font Family ***;
%let AtRiskOpts = valueattrs=(size=12pt color=black family='Arial');

*** p-value table location ***;
%let InsetOpts = ;

*** Group Legend location ***;
%let LegendOpts = ;

*** Step Curve Thickness ***;
%let StepOpts = lineattrs=(thickness=2);
```

1. Customizing the colors used for the strata values

2. Customizing the text attributes for the At Risk table
3. Customizing the thickness of the step curve

One thing worth noting is that when xOptions and yOptions macro variables are used to customize xAxis and yAxis values, the default values for these macro variables need to be included in the custom call.

Printing on one slide
The paper[3] provides details on the usage of ODS POWERPOINT and ODS LAYOUT in conjunction. The ODS statements provide a variety of powerful features; some of the features that we used are described below.

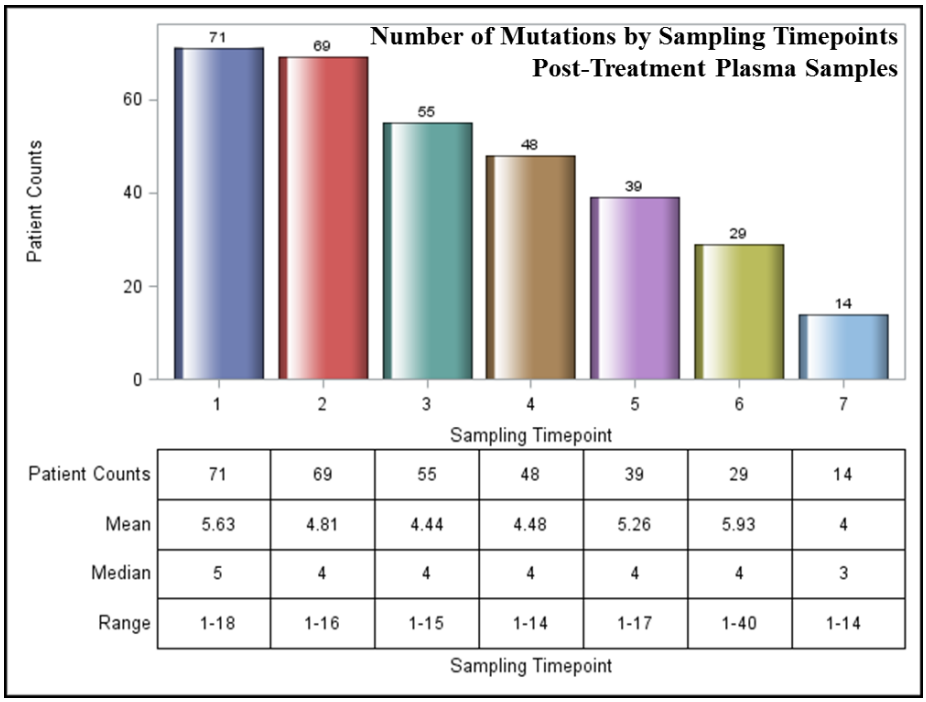
```
options nodate nonumber orientation=landscape;
ods powerpoint file="xx\xx\xx\xx\xx.pptx" nogtitle
                                         nogfootnote
                                         style=rsspptstyle
                                         dpi=300;
ods graphics on / noborder width=3.5in height=2.5in;

ods layout gridded rows=2 row_gutter=5pct;   1a
ods region;                                  1b
*** K-M Curve Code ***;
ods region;                                  1c
*** Survival Statistics Table Code ***;
ods layout end;                              1d

data _null_;
   dcl odsout obj();                         2
run;

ods layout gridded rows=2 row_gutter=5pct;
ods region;
*** K-M Curve Code ***;                      3
ods region;
*** Survival Statistics Table Code ***;
ods layout end;

ods powerpoint close;
```

1. Code for Slide#1
    a. The ODS LAYOUT GRIDDED statement starts the layout container. The ROWS= option is set to 2 so that the output is stacked together.
    b. The ODS REGION statement begins the region container that is to hold the K-M curve.
    c. The ODS REGION statement begins the region container that is to hold the statistics table.
    d. The ODS LAYOUT END statement ends the gridded layout.
2. Code for triggering a new slide.
3. Code for Slide#2

BAR CHARTS

DESCRIPTION

The objective of this visualization is to understand the mutation profile for a study over time as seen in the visual below. It helps to easily and quickly know if the number of mutations per sample is within the expected range given the parameters for the study.



	1	2	3	4	5	6	7
Patient Counts	71	69	55	48	39	29	14
Mean	5.63	4.81	4.44	4.48	5.26	5.93	4
Median	5	4	4	4	4	4	3
Range	1-18	1-16	1-15	1-14	1-17	1-40	1-14

Sampling Timepoint

CODE SNIPPET

The TEMPLATE procedure was used to produce a custom template to produce the table below the bar chart. This template was then used in the SGRENDER procedure to generate the final output.

```
proc template;
   define statgraph barblock;
      begingraph;
         layout lattice / rowweights=(.65 .35);
            layout overlay;
               barchart x=visit y=n / group=visit
                                      barlabel=true dataskin=gloss;
            endlayout;
            layout overlay / xaxisopts=(type=discrete  display=(label))
                             walldisplay=none;
               blockplot x=visit block=stats / class=desc
                                               display=(outline values label)
                                               repeatedvalues=true;
            endlayout;
         endlayout;
      endgraph;
   end;
run;
```

1. **layout lattice** statement is used to create 2 rows; first row occupying 65% of the space displays the bar chart and the second row occupying 35% of the space displays the summary statistics table.
2. **barchart** statement is used to create the bar chart.
3. **blockplot** statement is used to create the summary statistics table below the bar chart.

RSHINY APPLICATION

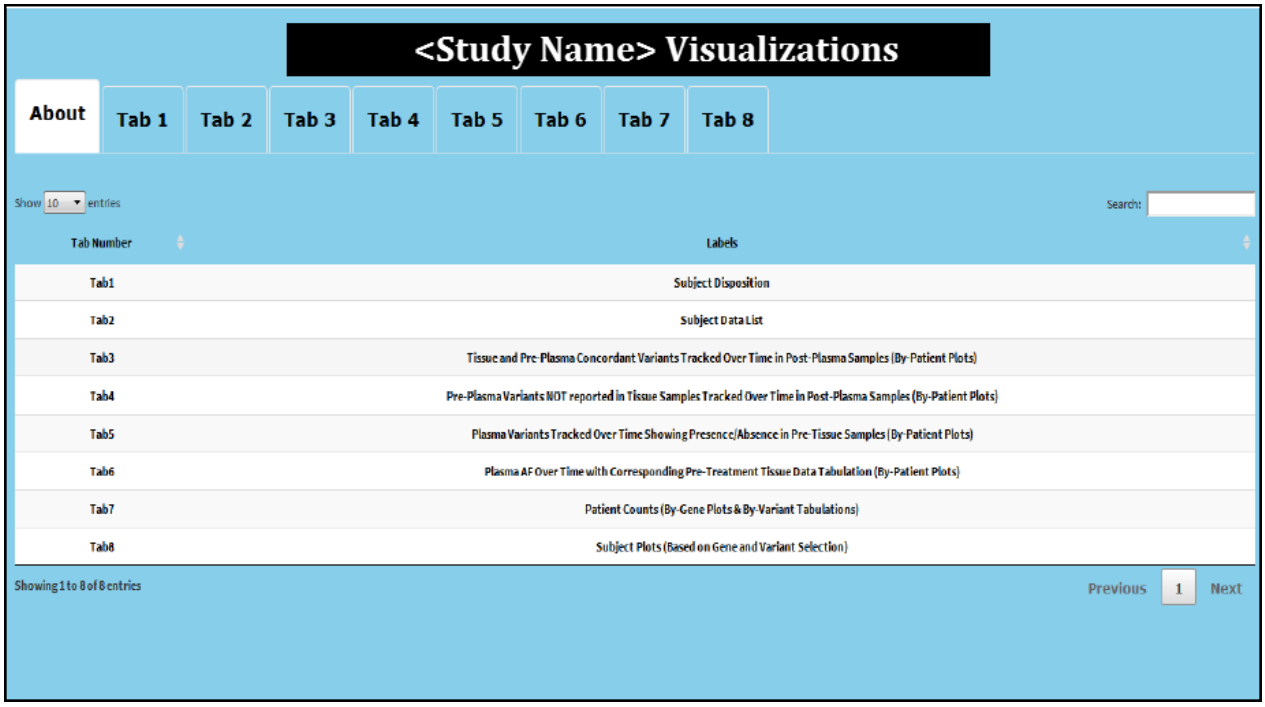
DESCRIPTION

The objective of this application was to provide the users with different visualizations to gain insights into the gene mutations for cancer patients with the help of interactive graphs and tabulations. The input data is a combination of genomic data generated by next-generation-sequencing and clinical endpoint data.

Shiny is an R package that was used to build the application. This package allows building interactive web applications straight from R. In addition to Shiny; the following R packages were used:

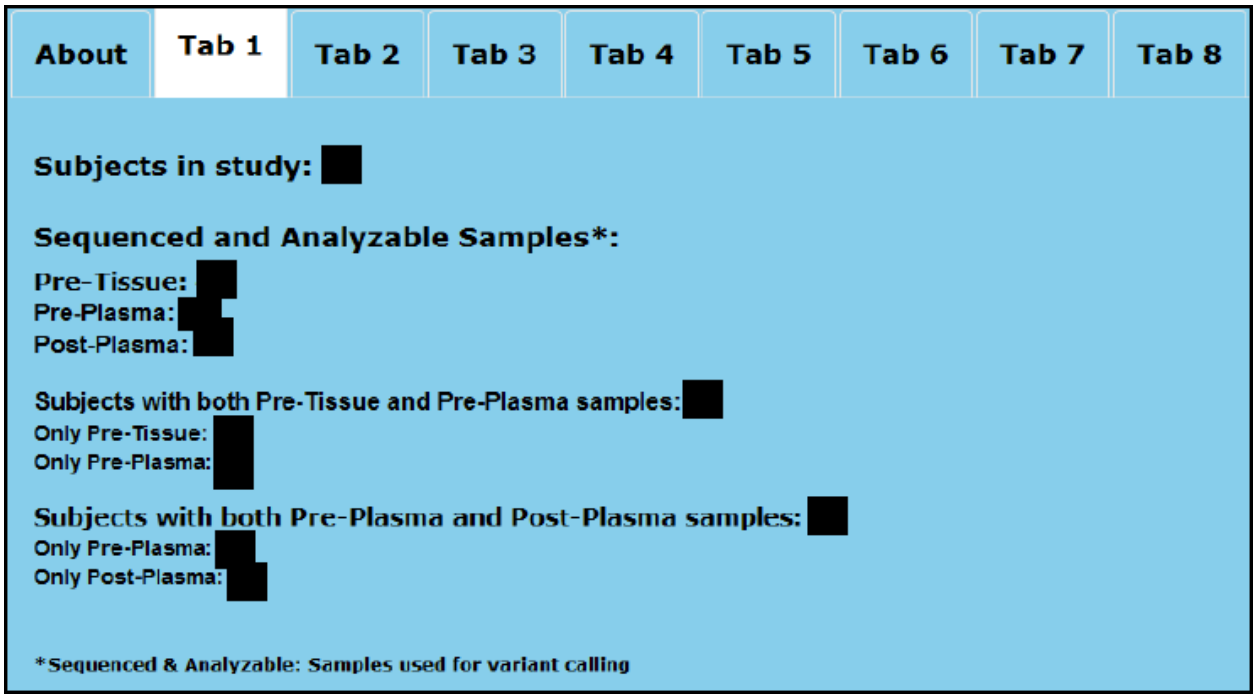
Package Name	Purpose
SHINY	Application user interface and server control
GGPLOT2	Dependency package for **PLOTLY**
PLOTLY	Interactive plots
RCOLORBREWER	Additional color components
GRDEVICES	Base color components
DT	Interactive data tables
PLYR	Required for data manipulation in R script being invoked
DPLYR	Required for data manipulation in R script being invoked
RESHAPE2	Required for data manipulation in R script being invoked
SAS7BDAT	Reading SAS files **(optional)**

The home page for the app lists the visualizations which the user can then navigate to by clicking on the corresponding tab number at the top.



TAB 1 – SUBJECT DISPOSITION

The purpose of this table is to provide a summary of the sample & subject inventory for the study.



TAB 2 – SUBJECT DATA LISTING

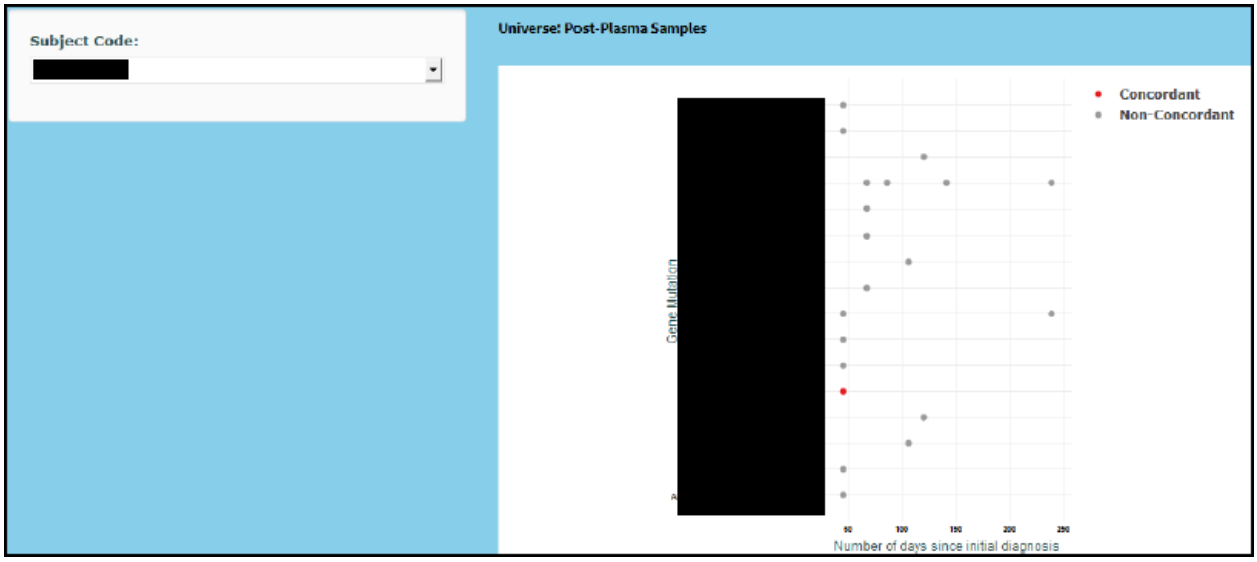
The purpose of this table is to provide a detailed data dump for all the subjects in the study.

Key Features:

- The "Subject Code" selector on the left allows the user to review data from a selected patient(s) at a time.
- The text boxes at the top of each column in the table allow for selecting specific values in that column. As an example; for the GENE column, the user could select a particular Gene of interest to look at across the entire population.
- There is also an option to download the data in an Excel file format.

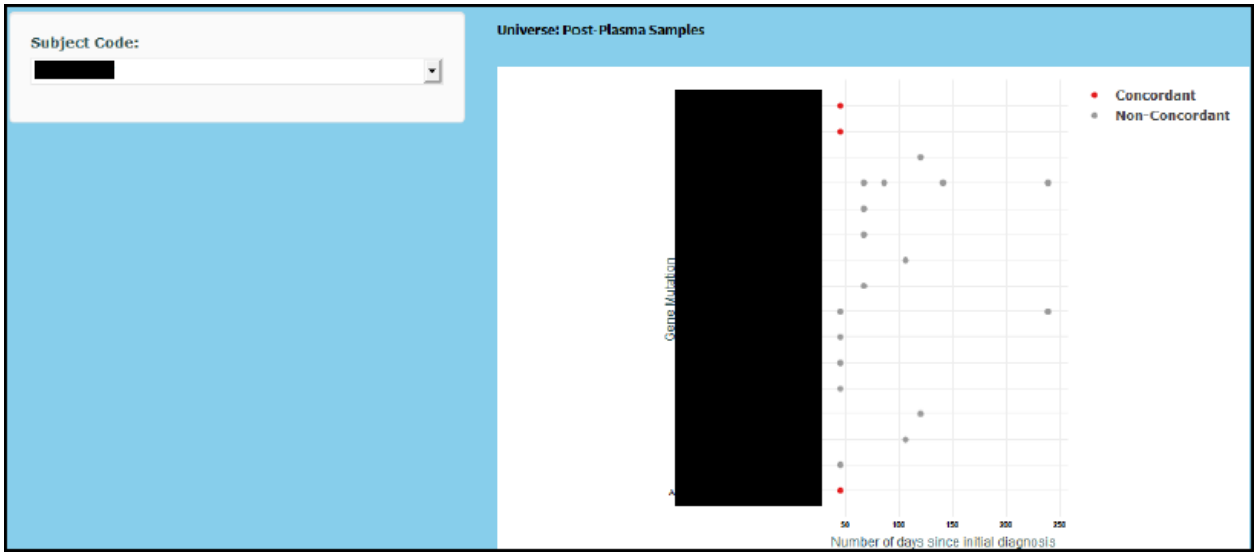
TAB 3 – CONCORDANT VARIANTS FROM PRE-TREATMENT TISSUE & PLASMA SAMPLES TRACKED OVER TIME IN POST TREATMENT PLASMA SAMPLES

The purpose of this visualization is to review & compare concordant gene mutations seen in pre-treatment samples vs post-treatment plasma samples.



TAB 4 – DISCORDANT PLASMA VARIANTS FROM PRE-TREATMENT SAMPLES TRACKED OVER TIME IN POST TREATMENT PLASMA SAMPLES

The purpose of this visualization is to review & compare discordant gene mutations seen in plasma samples; both pre & post treatment.



Only a handful of visualizations have been described in detail here. As can be seen in the list on the home page various other visualizations were developed in collaboration with stakeholders.

CONCLUSION

With the increasing volume of data being generated, data visualization has become an integral part of data analysis more than ever before. These visualizations aid in better understanding the data before adopting traditional statistical models for further analysis. The visualizations presented here are an example of that exercise.

From a programming perspective; an important aspect is to understand "fit-for-purpose" programming package. In addition to SAS® software & R described in this paper, Spotfire is another visualization package that can be used. All these are great packages & the key to figure out which one works best for your particular visualization needs.

REFERENCES

1. Understanding Genetics

    a. https://ghr.nlm.nih.gov/primer#
2. SAS Support – SAS/STAT® 13.1 User's Guide Customizing the Kaplan-Meier Survival Plot
    a. https://support.sas.com/documentation/onlinedoc/stat/131/kaplan.pdf
3. SAS Support – Add a table beneath a graph with the Graph Template Language (GTL)
    a. http://support.sas.com/kb/49/697.html
4. SAS Support – The Dynamic Duo: ODS Layout and the ODS Destination for PowerPoint
    a. http://support.sas.com/resources/papers/proceedings16/SAS5443-2016.pdf
5. SAS Documentation – BAND Statement in Proc SGPLOT
    a. http://documentation.sas.com/?docsetId=grstatproc&docsetTarget=n1t2ihc734azuvn17021ri2cl057.htm&docsetVersion=9.4&locale=en
6. Shiny
    a. https://shiny.rstudio.com/

ACKNOWLEDGMENTS

We would like to thank Saarthak Misri for his RShiny Application Development work as his 2017 Summer Intern Project at Roche Sequencing Solutions. We would also like to thank our Clinical Science & Medical Affairs leads and our Development Organization colleagues for providing valuable input during the development of these visualizations. We would also like to thank John Palma our Chief Medical Officer for his continued support for this work.

CONTACT INFORMATION

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

Author Name:	Nalin Tikoo
Company:	Roche Sequencing Solutions
Address:	4300 Hacienda Drive
City / Postcode:	Pleasanton, 94588
Work Phone:	925-353-5384
Email:	nalin.tikoo@roche.com
Web:	http://sequencing.roche.com/en.html