Coding assessment for the Liu Laboratory

Swathi Dhanasekaran dhanasekaran.s@husky.neu.edu

The following analysis was carried out in R Studio using the package maftools to discover genes with mutations associated with treatment response in a dataset of 50 .maf files containing the genomic mutations observed in a different patient's tumor, obtained by biopsy and sequenced with whole exome sequencing.

The .maf files were merged to create a maf object that was later subsetted to contain only the nonsynonymous mutations which will result in changes to the produced protein so as to consider only the biologically significant mutations. This subsetted maf object was used for subsequent analysis.

The top 15 most commonly mutated genes could be obtained from the gene summary where they are sorted based on the total mutations accrued from the highest to the lowest.

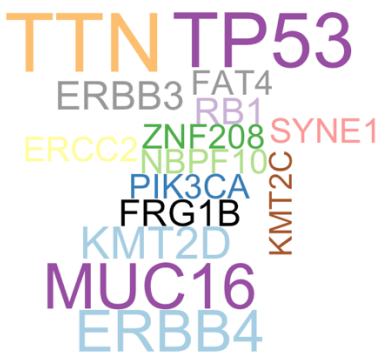

Figure: Gene Cloud of top 15 most common genes

TTN*	Encodes a large abundant protein of striated muscle Mutations in this gene are associated with familial hypertrophic cardiomyopathy 9
TP53*	Encodes a tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains Mutations in this gene are associated with a variety of human cancers, including hereditary cancers such as Li-Fraumeni syndrome
MUC16	This protein is thought to play a role in forming a barrier, protecting epithelial cells from pathogens. Products of this gene have been used as a marker for different cancers, with higher expression levels associated with poorer outcomes.
ERBB4	Encodes a protein which binds to and is activated by neuregulins and other factors and induces a variety of cellular responses including mitogenesis and differentiation. Mutations in this gene have been associated with cancer

Coding assessment for the Liu Laboratory

Swathi Dhanasekaran dhanasekaran.s@husky.neu.edu

KMT2D*	The encoded protein is part of a large protein complex called ASCOM, which has been shown to be a transcriptional regulator of the beta-globin and estrogen receptor genes. Mutations in this gene have been shown to be a cause of Kabuki syndrome
ERBB3	Amplification of this gene and/or overexpression of its protein have been reported in numerous cancers, including prostate, bladder, and breast tumors
FRG1B	This gene maps to a location 100 kb centromeric of the repeat units on chromosome 4q35 which are deleted in facioscapulohumeral muscular dystrophy (FSHD)

*Flagged genes: These genes are often non-pathogenic and passengers but are frequently mutated in most of the public exome studies.

Table: Functions of some of these genes as curated from NCBI

The associated protein changes were then extracted from the data table and stored together in a new table "top15MuInfo".

Pairwise and Groupwise Fisher tests were conducted to explore if any mutated genes are enriched in patients who either responded or not using the clinicalEnrichment(). The pairwise tests help identify significantly differently enriched genes between two groups, while the groupwise test helps identify significantly differently enriched genes between one group with the rest of the samples. However, since we have only two levels of Responses here, Responders and Non-responders, they both essentially make the same comparison.
The following is a plot of the significantly differentially mutated genes ($p < 0.05$) and their associated responsiveness. The genes along the x-axis are found to be significantly associated with response to therapy based on the high Responder sample fraction they are present in.

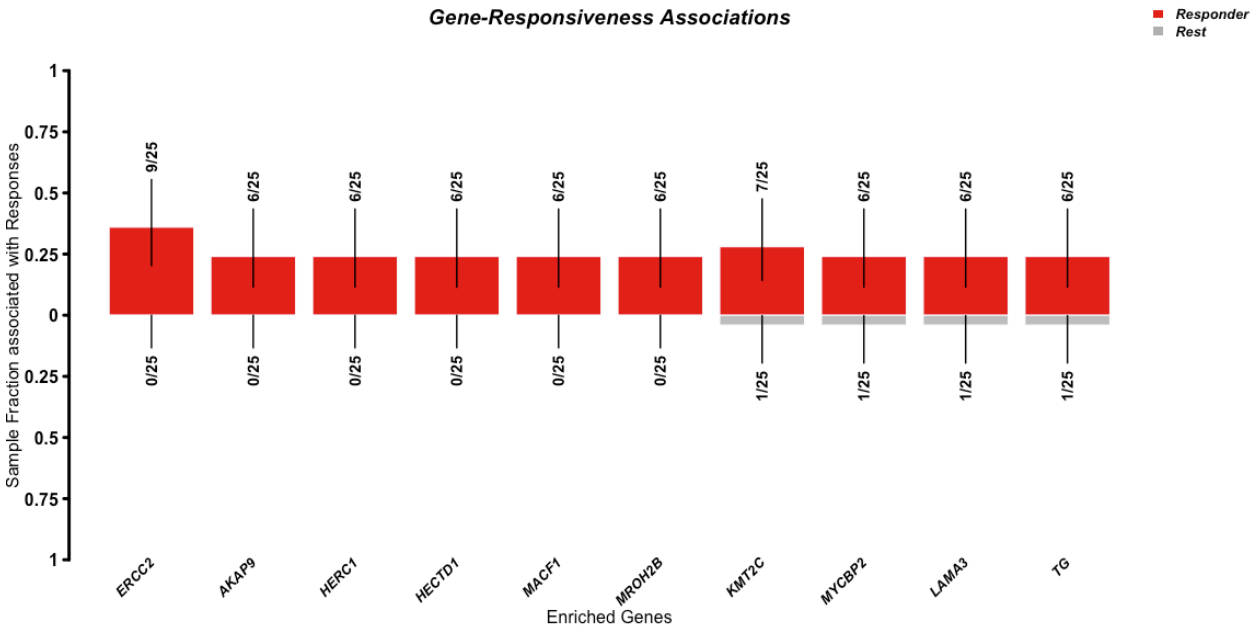

Figure: Association of enriched gene mutations with the responsiveness to treatment

A scatterplot of the significantly enriched genes and the number of mutated patients was created using ggplot(). However, it is only a plot of the absolute number of patients carrying

mutations in the enriched gene and does not give us a fair idea of the incidence of the mutation in a population.

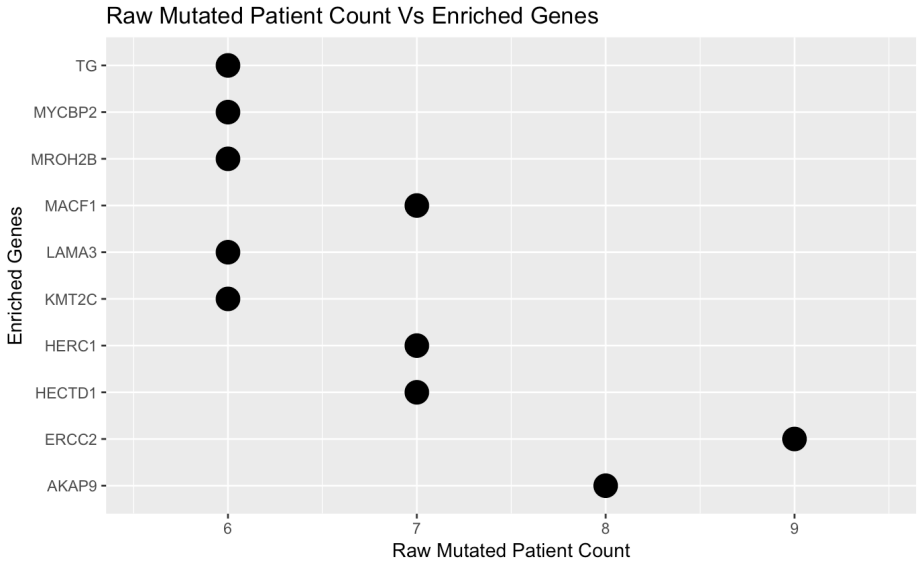

Figure: Scatterplot of number of mutated patients for each significantly enriched gene

So we make a plot of the mutation frequency, i.e, number of mutations/genome. It can give us an idea of the average incidence of mutations in a particular gene in a population. The mutation frequency can be greater than one as there can more than one mutation in a particular gene in a genome.

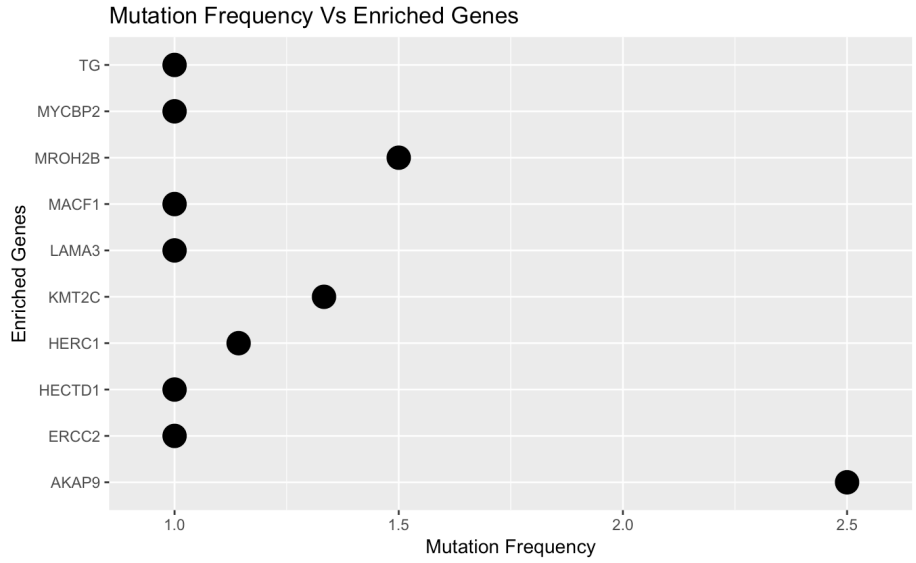

Figure: Scatterplot of mutation frequency for each significantly enriched gene

Coding assessment for the Liu Laboratory

Swathi Dhanasekaran dhanasekaran.s@husky.neu.edu

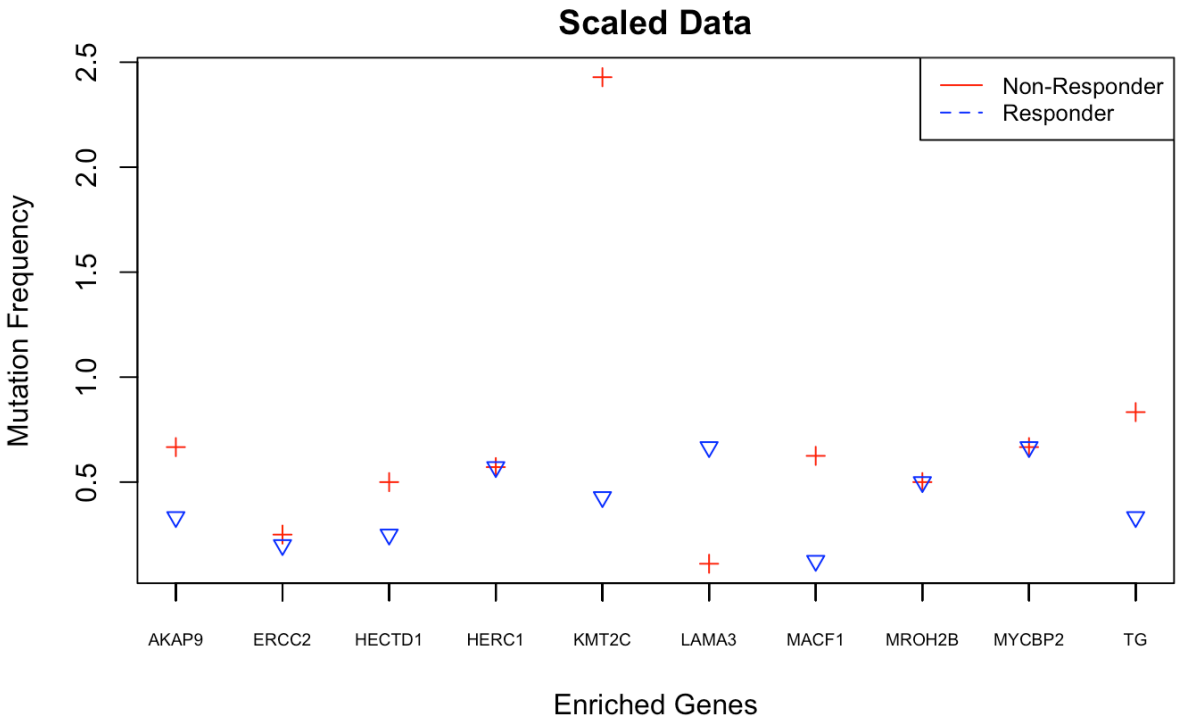

Figure: Scatterplot of mutation frequency for each significantly enriched gene in each Response level.

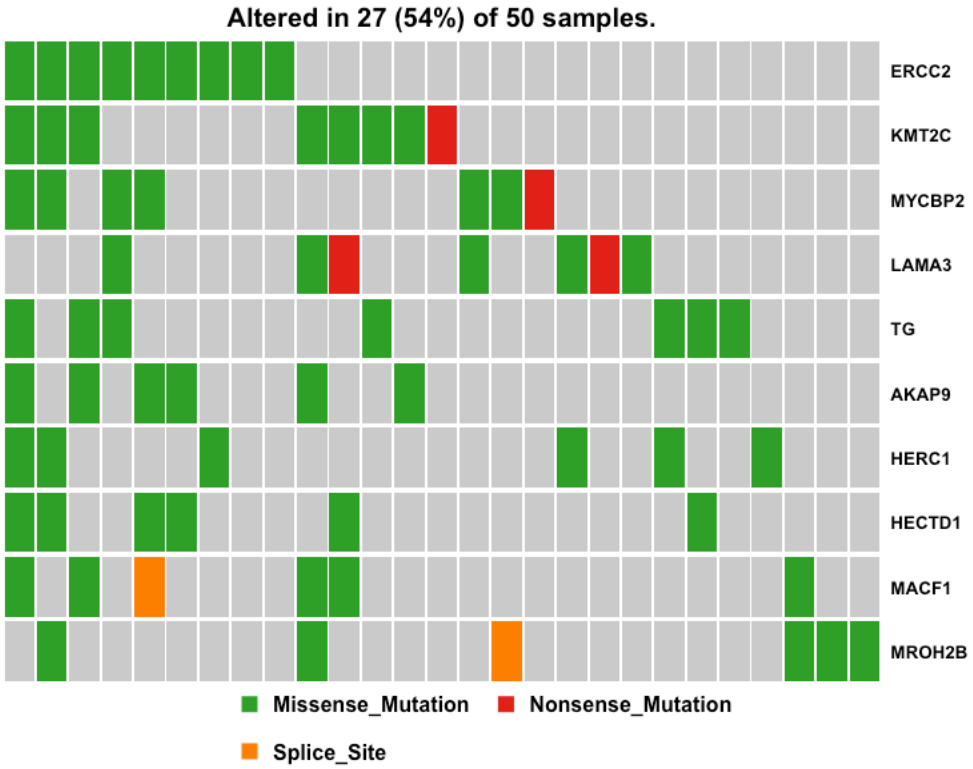

Figure: Mutation type in all patients for each significantly enriched gene

Coding assessment for the Liu Laboratory

Swathi Dhanasekaran dhanasekaran.s@husky.neu.edu

We can observe that the most significantly enriched gene, ERCC2 (p value = 0.0008154177) seems to accrue somatic, non synonymous, missense mutations in the Responders. According to data from the NCBI, the protein encoded by this gene is involved in transcription-coupled nucleotide excision repair, a mechanism to repair damage to DNA, and is an integral member of the basal transcription factor BTF2/TFIIH complex. Mutations in ERCC2 appear to confer sensitivity to treatment resulting in response to therapy. It could perhaps be used as a biomarker to assess sensitivity to the treatment under consideration. However, the mechanism of how ERCC2 confers treatment sensitivity and whether it can be validated as a biomarker has to be examined in pre-clinical studies using ERCC2 mutant animal models and cell lines.

Following this, the tumor mutational burden (TMB) of ERCC2 in the mutant samples was compared with that of the wild type samples. The TMB data, found as Nonsynonymous_mutations_per_Mb in the clinical data was extracted for both mutant and wild type samples. Prior to assessing significant differences in the means of the ERCC2 TMB in mutant and wild type data using a two-tailed Student's t-test, a test of homoscedasticity was carried out to test one of the t-test's assumptions of homogeneous variances. After confirming homoscedasticity, the two-tailed t-test was carried out. Based on the low p-value of 0.0008192 (< 0.05) we reject the null hypothesis that the mean ERCC2 TMB in mutant samples is not the same as that in the wild type samples. Then, a one-sided t-test was carried out to confirm if the ERCC2 TMB in mutant samples is higher than that in the wild type samples (p value = 0.0004096 < 0.05). Hence, based on our analysis we can conclude that ERCC2 mutations are significantly enriched in patients responding to treatment and could perhaps serve as a biomarker for treatment sensitivity.

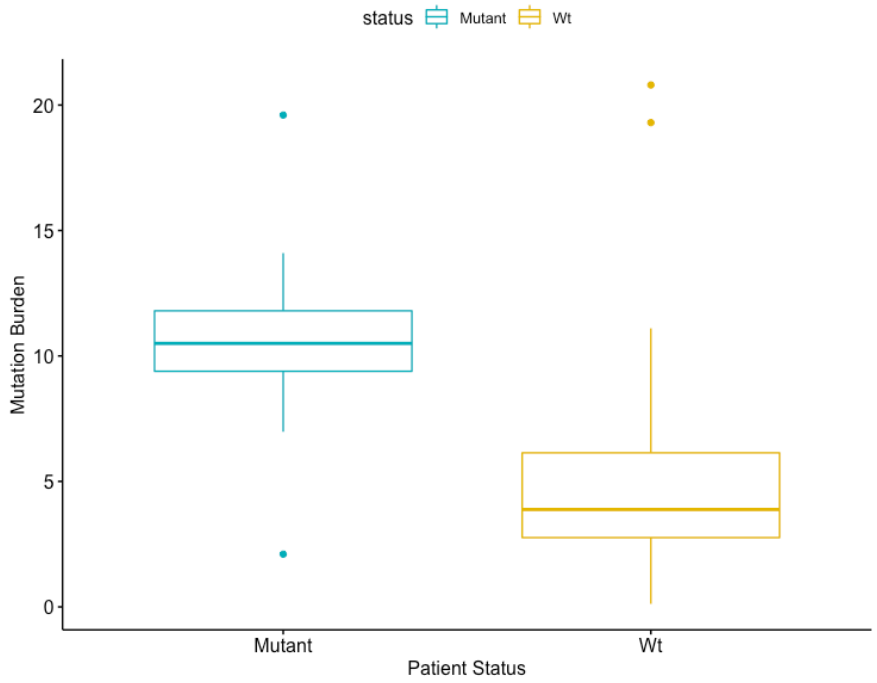

Figure. Box plot of the mean ERCC2 TMB in mutant and wild-type samples. (Data from two-tailed Student's t-test)