Characterizing NSD3 Amplification in Lung Cancer

David Dilworth and Dalia Barsyte-Lovejoy

2018-04-15

Objective

NSD3 (WHSC1L1) is amplified in ~5% of Non-Small Cell Lung Cancer patients(**cBioPortal**: Cerami et al. Cancer Discov. 2012 and Gao et al. Sci. Signal. 2013). However, the implications of this event on the formation and progression of the disease are unclear. While NSD3 may be a driver of lung cancer, it is also plausible that this loci is simply amplified at a higher frequencies in the context of cancer-associated genomic instability. To dive deeper into this question I will use The Cancer Genome Atlas (TCGA) lung cancer data-sets to look for associations between NSD3 amplification and mutational status as well as gene expression profiles. This data has been generated by the TCGA Research Network: http://cancergenome.nih.gov/. I hypothesize that if NSD3 amplification is a driving force in a subset of lung tumors, these samples will share similar gene expression profiles and exhibit higher expression levels of NSD3. Here, I am using FirebrowserR (Deng M., et al. Database. 2017 - PMID:28062517), an R client for Broad Institute's Firehose Web API, which allows TCGA data processed by the Firehose Pipeline to be directly imported into R for analysis.

Code & Observations

Pacakage Requirements

```
# We will use several R Packages for this analysis, which are loaded below
# and author acknowledgement shown;

require(FirebrowseR) # Deng M., et al. Database. 2017 - PMID:28062517
require(maftools)    # Mayakonda, A. and H.P. Koeffler. bioRxiv, 2016 -
                     # doi:http://dx.doi.org/10.1101/052662
require(rms)         #  Frank E Harrell Jr (2018)
require(SummarizedExperiment) # Morgan M, Obenchain V, Hester J, and Pagès H (2017)
require(RColorBrewer)# Erich Neuwirth (2014)
require(ggridges)    # Claus O. Wilke
require(ggbeeswarm)  # Erik Clarke and Scott Sherrill-Mix (2017)
require(tidyverse)   # Hadley Wickham
require(survminer)   # Alboukadel Kassambara and Marcin Kosinski (2018)
```

Identification of TCGA Cohort Code and Available Lung Cancer Datasets

Here, the cohort code for lung cancer is extracted and stored. We can see that there are two data sets for lung cancer, LUAD (Lung Adenocarinoma) and LUSC (Lung Squamous Cell Carcinoma). These two subtypes make up the majority of non-small cell lung cancer and are primarily differentiated by cell type of origin. Adenocarinoma stems from epithelial cells that line the larger airways, while squamous cell carcinoma derive from peripheral small airways. Differential diagnosis of NSCLC subtype is important, as treatment regimens differ.

```
# Identify TCGA cohorts containing "lung"" in the description
```

```r
cancer <- "lung"

cohort  <- Metadata.Cohorts(format = "csv") %>%
             filter(str_detect(description, regex(cancer, ignore_case = T))) %>%
             .$cohort

paste("TCGA cohort code for",  cancer, "is :", cohort)
```

```
## [1] "TCGA cohort code for lung is : LUAD"
## [2] "TCGA cohort code for lung is : LUSC"
```

Next, sample counts for each sample type can be obtained to ensure that when we download different data types we are retrieving all available data.

```r
# Retrieve sample counts for identified cohorts.

sample_count <- list()
for (i in 1:length(cohort)) {
  temp_count <- as.data.frame(Metadata.Counts(cohort = cohort[i])) %>%
                    dplyr::select(Counts.date, Counts.bcr,
                                  Counts.clinical, Counts.maf,
                                  Counts.mrnaseq, Counts.cn) %>%
                    setNames(., c("Date", "Biospecimen", "Clinical",
                                  "Mutations", "Gene Expression", "Copy Number"))
  sample_count[[i]] <- temp_count
}

print(sample_count)
```

```
## [[1]]
##                           Date Biospecimen Clinical Mutations
## 1 Thu, 28 Jan 2016 00:00:00 GMT         585      522       230
##   Gene Expression Copy Number
## 1             515         516
##
## [[2]]
##                           Date Biospecimen Clinical Mutations
## 1 Thu, 28 Jan 2016 00:00:00 GMT         504      504       178
##   Gene Expression Copy Number
## 1             501         501
```

NSD3 Gene Expression in Lung Cancer

Retrieving NSD3 (WHSC1L1) Expression Data

The following code will download the gene expression data for NSD3 (WHSCL1L1) and save the results. The page size is set to the sample number obtained in the previous step.

```r
# Download and save gene expression data for the gene "WHSC1L1".

gene <- "WHSC1L1"

if(file.exists("TCGA_GeneExpression.csv")) {
  gene_exp <- read.csv("TCGA_GeneExpression.csv")
} else {
```

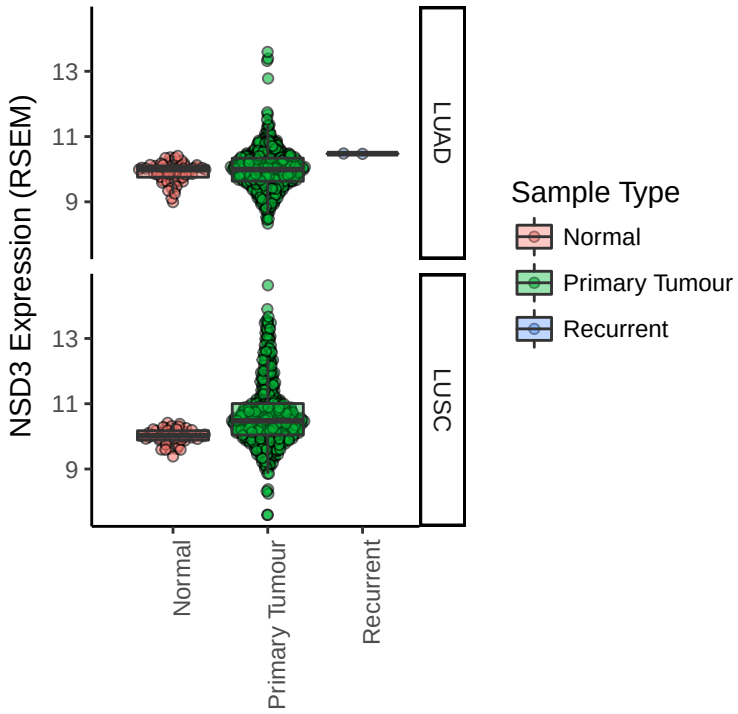
```
gene_exp <- list()
for(i in 1:length(cohort)) {
    temp <-   Samples.mRNASeq(format = "csv",
                              gene = gene,
                              cohort = cohort[i],
                              page_size = sample_count[[i]][1,5])
    temp$HistoType <- rep(cohort[i], nrow(temp))
    gene_exp[[i]] <- temp}

gene_exp <- bind_rows(gene_exp)
}
if (!file.exists("TCGA_GeneExpression.csv")) {
write.csv(gene_exp, file = "TCGA_GeneExpression.csv")}

# Replace tumour code with readable description.

gene_exp$sample_type <- str_replace(gene_exp$sample_type, "NT", "Normal")
gene_exp$sample_type <- str_replace(gene_exp$sample_type, "TP", "Primary Tumour")
gene_exp$sample_type <- str_replace(gene_exp$sample_type, "TR", "Recurrent")
gene_exp$sample_type <- factor(gene_exp$sample_type, levels = c("Normal", "Primary Tumour",
                                                                "Recurrent"))
```

Plotting NSD3 Expression

This next bit of code will plot NSD3 (WHSC1L1) expression for all samples within the selected TCGA cohorts by sample type.

Figure.1 - Plot of NSD3 Expression in Normal, Primary Tumor, and Recurrent Tumor Samples.

```
ggplot(gene_exp, aes(sample_type, expression_log2)) +
      geom_quasirandom(aes(fill = sample_type), pch = 21, alpha = 0.6, dodge.width = 1) +
      geom_boxplot(aes(fill = sample_type), pch = 21, alpha = 0.4, outlier.shape = NA) +
      theme_classic() +
      theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
      ylab("NSD3 Expression (RSEM)") +
      xlab("") +
      guides(fill=guide_legend(title="Sample Type")) +
      facet_grid(cohort ~ .)
```



Observation: Primary lung tumor samples display a broader distribution of NSD3 expression relative to normal, trending towards overexpression (Figure 1). This is more prominent for the LUSC cohort. There are relatively few samples associated with recurrence, therefore we should exclude these from further analysis. Importantly, not all primary tumor samples have a matched normal sample. Thus, we can next look at expression for tumor samples with a matched normal sample to confirm the increased expression levels observed are unlikely to be due to normal biological variation in NSD3 expression. The code below will filter gene expression values to only include those with matched normal samples and plot the results as a box and density plot.

Figure.2 NSD3 Expression Levels in Lung Cancer Samples with Matched Normal

```
# Filter expression data for samples with matched normal pair & plot.

exp_matched <- gene_exp %>%
  filter(sample_type == "Normal") %>%
  semi_join(gene_exp, ., by = "tcga_participant_barcode") %>%
  filter(sample_type == "Normal" | sample_type =="Primary Tumour") %>%
  as_tibble()

# Plot as point and boxplot.

ggplot(exp_matched, aes(sample_type, expression_log2)) +
  geom_quasirandom(aes(fill = sample_type), pch = 21, alpha = 0.6, dodge.width = 1) +
  geom_boxplot(aes(fill = sample_type), pch = 21, alpha = 0.4, outlier.shape = NA) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ylab("NSD3 Expression (RSEM)") +
  xlab("") +
  guides(fill=guide_legend(title="Sample Type")) +
  facet_grid(HistoType ~ .)
```

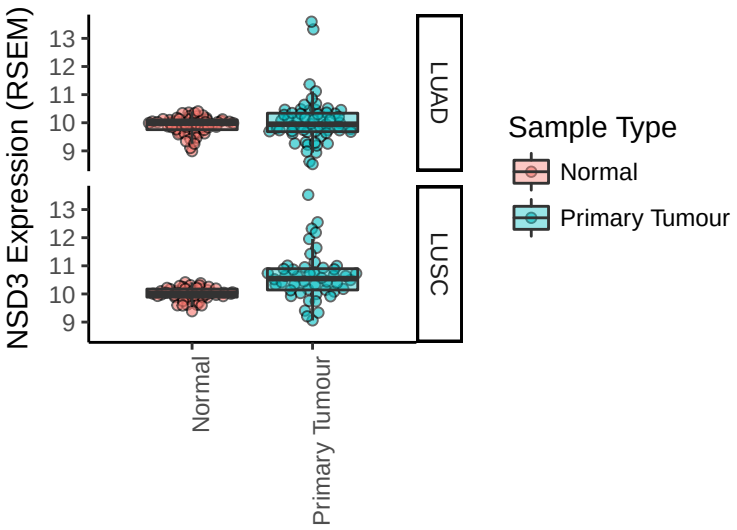
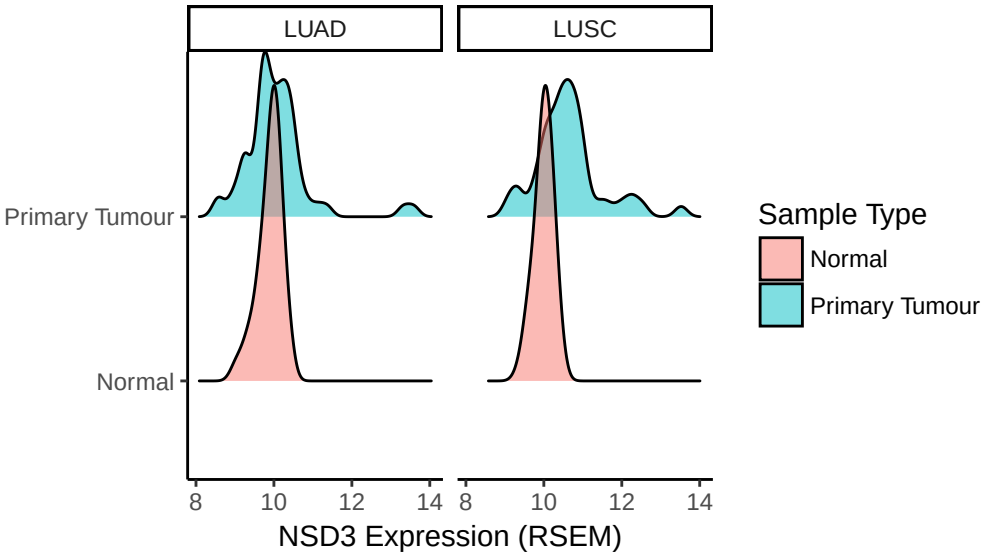


Figure.3 Density Plots of NSD3 Expression Levels in Lung Cancer Samples with Matched Normal

```
# Plot as density ridges.

ggplot(exp_matched, aes(x = expression_log2, y = sample_type)) +
  geom_density_ridges(aes(fill = sample_type), alpha = 0.5) +
  theme_classic() +
  xlab("NSD3 Expression (RSEM)") +
  ylab("") +
  guides(fill=guide_legend(title="Sample Type")) +
  facet_grid(. ~ HistoType)
```

```
## Picking joint bandwidth of 0.149
```

```
## Picking joint bandwidth of 0.161
```



Observations: With this analysis, we do observe a broader distribution of NSD3 expression in tumor samples relative to matched normal controls (Figure 2-3). This indicates that the overexpression of NSD3 in

lung cancer is unlikely to be due to biological variation alone.

NSD3 Amplification in TCGA Lung Cancer Data

Retrieving NSD3 (WHSC1L1) Copy Number Data

Next, we can download gene level copy number data to look at NSD3 amplification in lung cancer. From this data we can identify patient tumor samples that have an amplification score greater than the high-confidence threshold for amplification (>2) and test if these samples show any differences in NSD3 expression levels, patient outcomes, or mutational profiles.

```
# Download and save copy number data for "WHSC1L1"

if(file.exists("TCGA_CopyNumber.csv")) {
  gene_cn <- read.csv(file = "TCGA_CopyNumber.csv")
} else {
  gene_cn <- list()
  for(i in 1:length(cohort)){
    temp <-  Analyses.CopyNumber.Genes.All(format = "csv",
                                           gene = gene,
                                           cohort = cohort[i],
                                           page_size = sample_count[[i]][1,6])
    gene_cn[[i]] <- temp }
  gene_cn <- bind_rows(gene_cn)
  write.csv(gene_cn, file = "TCGA_CopyNumber.csv")
}
```

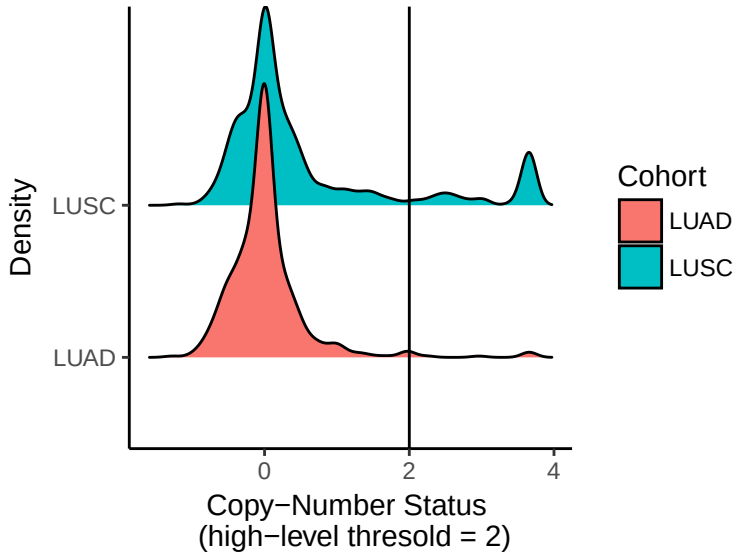
Plotting NSD3 Copy Number Status

The code below draws a density plot of the copy-number status. The high-level threshold is indicated with a vertical line at 2.

Figure.4 Density Plot of NSD3 Copy Number Scores Across Lung Cancer Samples

```
# Plot density of NSD3 Gistic scores by lung cohort.

ggplot(gene_cn, aes(all_copy_number, cohort)) +
        geom_density_ridges(aes(fill = cohort)) +
        geom_vline(xintercept = 2) +
        theme_classic() +
        xlab("Copy-Number Status \n (high-level thresold = 2)") +
        ylab("Density") +
        guides(fill=guide_legend(title="Cohort"))
```

```
## Picking joint bandwidth of 0.105
```



Amplification Frequency Across Lung Cohorts

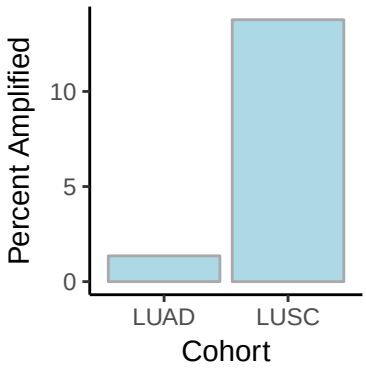
The following code calculates the percentage of patients with copy-number status greater than the high-level threshold then plots a bar graph depicting the relative numbers.

Figure.5 NSD3 Amplification Frequency in LUAD and LUSC

```
# Calculate Percent Amplified

Percent_AMP <- gene_cn %>%
                  group_by(cohort) %>%
                  summarize(nAMP = sum(all_copy_number >2),
                            nTOTAL = n(),
                            Percent = (nAMP / nTOTAL)*100)
print(Percent_AMP)
```

```
## # A tibble: 2 x 4
##   cohort  nAMP nTOTAL Percent
##   <fct>  <int>  <int>   <dbl>
## 1 LUAD       7    516    1.36
## 2 LUSC      69    501   13.8
```

```
# Plot bar graph of amplification frequency.

ggplot(Percent_AMP, aes(cohort, Percent)) +
  geom_bar(color = "darkgrey", fill = "lightblue", stat = "identity") +
  theme_classic() +
  xlab("Cohort") +
  ylab("Percent Amplified")
```



Observations. From this data, we can see that NSD3 amplification occurs more frequently in squamous cell lung carcinoma compared to adenocarcinoma, at 13.8% and 1.4% of samples respectively (Figure 4-5).

Combining Copy Number & Expression Data

```
# Call amplified samples and merge with expression data.

gene_cn_exp <- gene_cn %>%
                  mutate(amplified = if_else(all_copy_number >= 2,
                                             "Amplified",
                                             "Non-Amplified")) %>%
                  dplyr::select(c("tcga_participant_barcode",
                                  "all_copy_number",
                                  "amplified")) %>%
                  inner_join(gene_exp, by = "tcga_participant_barcode") %>%
                  filter(sample_type == "Primary Tumour")
```

Figure.6 NSD3 Expression by Amplification Status and Cohort

```
# Plot point & boxplot of NSD3 expression levels by amplification status

ggplot(gene_cn_exp, aes(amplified, expression_log2)) +
  geom_quasirandom(aes(fill = amplified), pch = 21, alpha = 0.6, dodge.width = 1) +
  geom_boxplot(aes(fill = amplified), pch = 21, alpha = 0.4, outlier.shape = NA) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ylab(paste(gene, "Expression (RSEM)")) +
  xlab("") +
  guides(fill=guide_legend(title="Sample Type")) +
  facet_grid(HistoType ~ .)
```

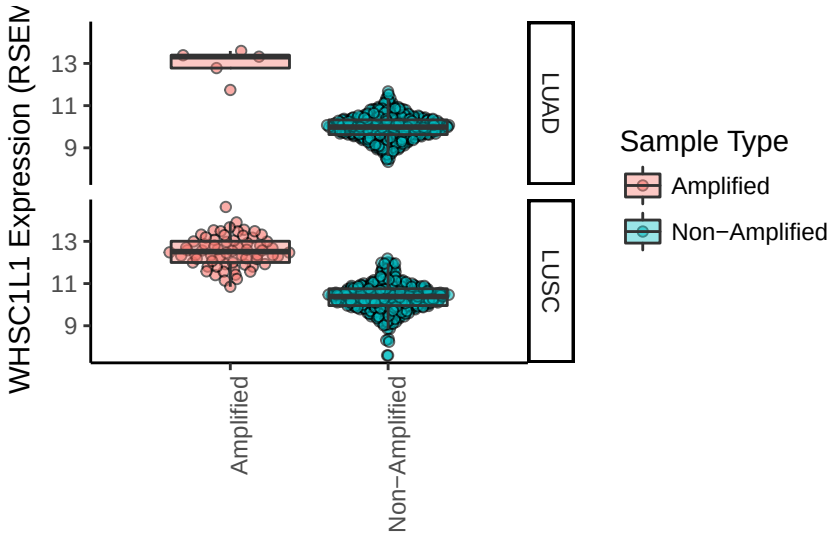
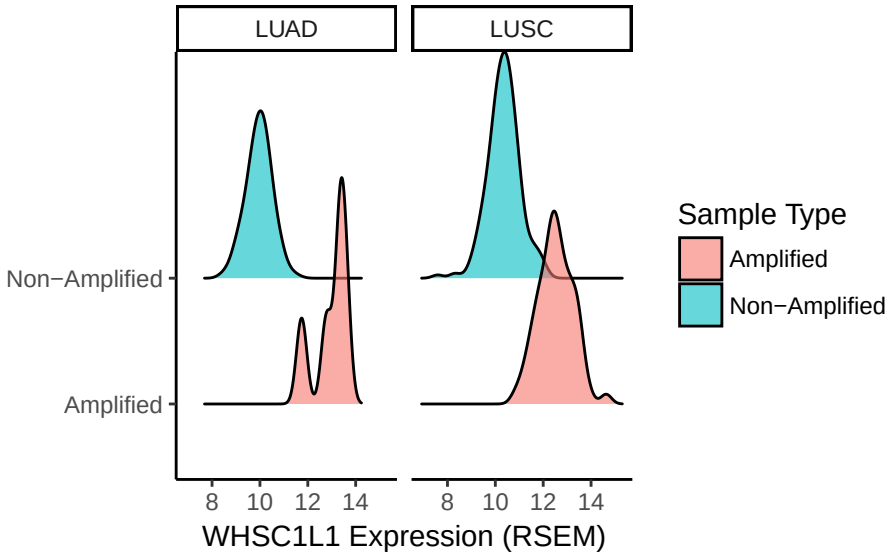


Figure.7 NSD3 Expression Across Samples by Amplification Status

```
# Plot density ridges

ggplot(gene_cn_exp, aes(expression_log2, amplified)) +
  geom_density_ridges(aes(fill = amplified), alpha = 0.6) +
  theme_classic() +
  xlab(paste(gene, "Expression (RSEM)")) +
  ylab("") +
  guides(fill=guide_legend(title="Sample Type")) +
  facet_grid(. ~ HistoType)
```

```
## Picking joint bandwidth of 0.217
```

```
## Picking joint bandwidth of 0.225
```



Observations. From Figure 6-7, we can see that NSD3 amplification does indeed result in increased gene expression, suggesting that this event has the potential to be functional in driving the disease.

Integrating Clinical Metadata & Expression

Now that we have identified patient samples that have a high probability of NSD3 amplifications, we can use the clinical metadata to evaluate any differences in survival between groups.

```
# Download and save clincial data.

if(file.exists("TCGA_ClinicalData.csv")) {
 clinical <- read.csv(file = "TCGA_ClinicalData.csv")
} else {
clinical <- list()
for(i in 1:length(cohort)) {
  temp <- Samples.Clinical(format = "csv",
                           cohort = cohort[i],
                           page_size = sample_count[[i]]$Clinical)
  clinical[[i]] <- temp
}
clinical <- bind_rows(clinical)
write.csv(clinical, file = "TCGA_ClinicalData.csv")
}
```

Survival Plots

To plot survival, we will first need to determine time to event, for the deceased this is days to death and for the living days to last followup. This data can then be used with the R packages survive and survminer to evaluate differences in survival associated with either NSD3 amplification or high/low expression.

```
# Determine time to event.

for (i in seq_along(clinical$vital_status)) {
  if (clinical$vital_status[i] == "alive") {
      clinical$time[i] <- clinical$days_to_last_followup[i]
  } else {
      clinical$time[i] <- clinical$days_to_death[i]
  }
}

# Add gene expression and copy number data to clinical

clinical <- gene_cn_exp %>%
            select(tcga_participant_barcode, amplified, expression_log2) %>%
            right_join(clinical, by = "tcga_participant_barcode") %>%
            drop_na(amplified) %>%
            filter(cohort == "LUSC")


# Calculate gene expression quartiles.

clinical <- within(clinical,
                   quartile <- cut(expression_log2,
                                   quantile(expression_log2, probs=0:4/4),
                                   include.lowest=TRUE,
                                   labels = FALSE))
```

Figure 8 Survival Curves for Patients with a High-confidence NSD3 Amplification

```r
# Generate survival plots.

km.by.gain <- npsurv(Surv(time, vital_status == "dead") ~ amplified,
                     data = clinical, conf.type = "log-log")

ggsurvplot(km.by.gain, data = clinical, size = 1,
           palette = c("#E7B800", "#2E9FDF"),
           conf.int = TRUE,
           pval = TRUE,
           risk.table = TRUE,
           risk.table.col = "strata",
           legend.labs = c("Amplified", "Non-Amplified"),
           risk.table.height = 0.25,
           ggtheme = theme_bw())
```

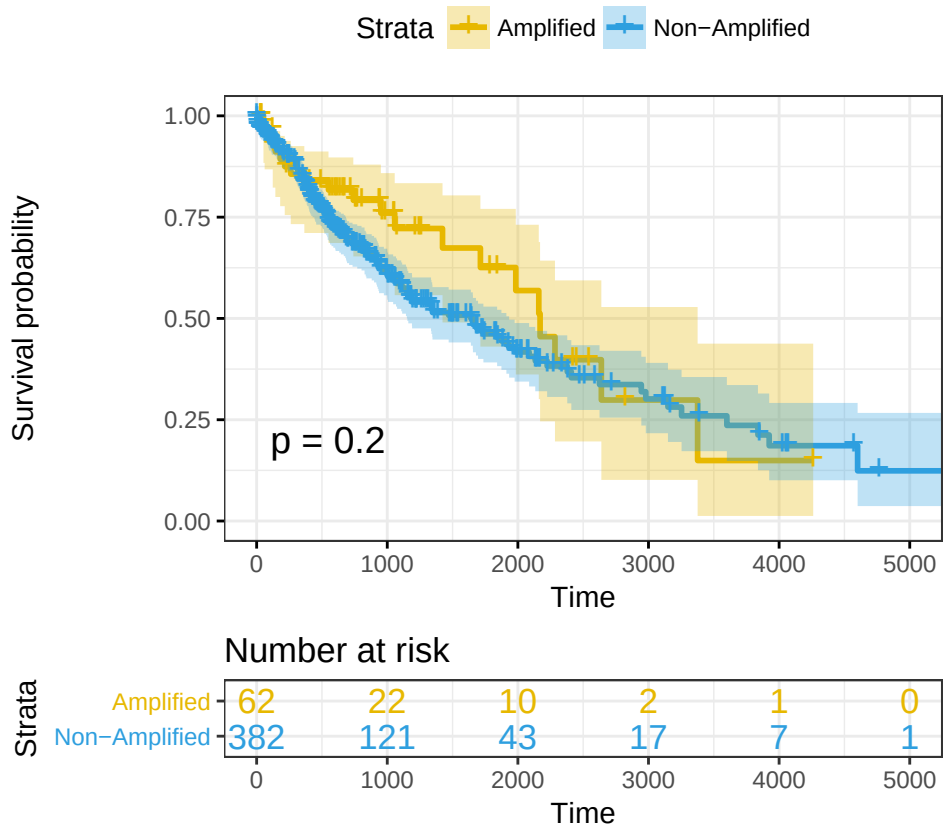
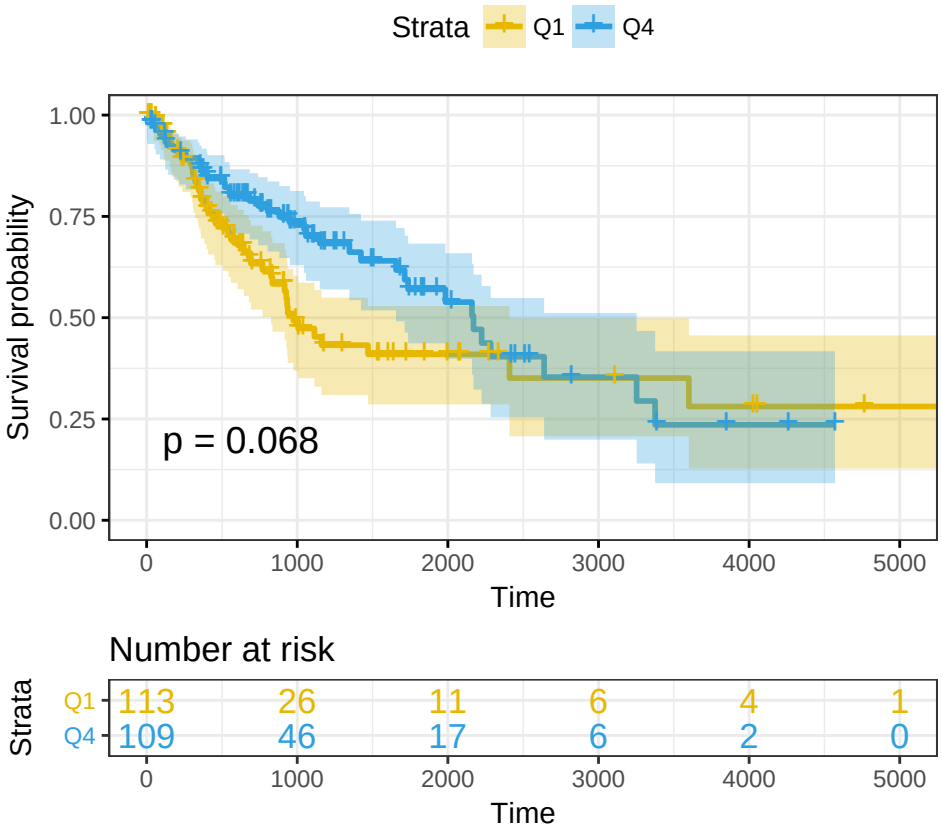


Figure 9 Survival by NSD3 Expression Quartile (Q1 and Q4)

```r
# Filter for NSD3 Expression quartiles

clin_1q4q <- filter(clinical, quartile == 1 | quartile == 4)

# Generate survival plots.

km.by.quartile <- npsurv(Surv(time, vital_status == "dead") ~ quartile,
                     data = clin_1q4q, conf.type = "log-log")

ggsurvplot(km.by.quartile, data = clin_1q4q, size = 1,
```

```
          palette = c("#E7B800", "#2E9FDF"),
          conf.int = TRUE,
          pval = TRUE,
          risk.table = TRUE,
          risk.table.col = "strata",
          legend.labs = c("Q1", "Q4"),
          risk.table.height = 0.25,
          ggtheme = theme_bw())
```



Observation. There is not a significant difference in survival outcomes for patients with either increased NSD3 expression or NSD3 amplification (Figure 8-9). In fact, in the short term, patients with increased expression or amplification of NSD3 appear to have better outcomes.

Analysis of Mutational Status

Next we look at mutational status of NSD3 amplified patient samples. From this analysis, we may be able to identify mutations that are enriched in the NSD3 amplified group. This information could potentially tell us about a the mutational landscape that either promotes or results from NSD3 amplification. The analysis will rely heavily upon the R package maftools (Mayakonda, A. and H.P. Koeffler. bioRxiv, 2016 - doi:http://dx.doi.org/10.1101/052662), which is an excellent resource for working in mutation annotation files.

Download Mutation Data

```
# Download and Save LUSC MAF. This is a large file and may take some time to download.

if (file.exists("lusc.maf.csv")) {
  lusc.maf <- read.csv("lusc.maf.csv")
} else {
    pages <- c(2:500)
    lusc.maf <- Analyses.Mutation.MAF(format = "csv",
                              cohort = "LUSC",
                              page = 1,
                              column = "all")
    for (page_num in pages) {
      read_in <- Analyses.Mutation.MAF(format = "csv",
                                    cohort = "LUSC",
                                    page = page_num,
                                    column = "all")
      colnames(read_in) <- colnames(lusc.maf)
      lusc.maf <- rbind(lusc.maf, read_in) }
}

if (!file.exists("lusc.maf.csv")) {
  write.csv(lusc.maf, file = "lusc.maf.csv")
}


# Read data into maftools - clinical data needs to include Tumor_Sample_Barcode

clinical <- clinical %>%
  mutate(Tumor_Sample_Barcode = tcga_participant_barcode)

lusc.maf <- lusc.maf %>%
  mutate(Tumor_Sample_Barcode = str_sub(Tumor_Sample_Barcode, 1, 12))

lusc.maf.in <- read.maf(maf = lusc.maf, clinical = clinical)
```

```
## NOTE: Non MAF specific values in Variant_Classification column:
## [1] "Start_Codon_Del" "Stop_Codon_Ins"
## silent variants: 15571
##                            ID     N
##  1:                  Samples   178
##  2:                    3'UTR    27
##  3:                  5'Flank    82
##  4:                    5'UTR    15
##  5:    De_novo_Start_InFrame     3
##  6: De_novo_Start_OutOfFrame     8
##  7:                      IGR   192
##  8:                   Intron   159
##  9:                      RNA   278
## 10:                   Silent 14803
## 11:          Start_Codon_Del     3
## 12:           Stop_Codon_Ins     1
```

```
## Summarizing..
```

	ID	summary	Mean	Median
1:	NCBI_Build	37	NA	NA
2:	Center	1	NA	NA
3:	Samples	176	NA	NA
4:	nGenes	13245	NA	NA
5:	Frame_Shift_Del	506	2.875	2.0
6:	Frame_Shift_Ins	111	0.631	0.0
7:	In_Frame_Del	43	0.244	0.0
8:	In_Frame_Ins	3	0.017	0.0
9:	Missense_Mutation	40587	230.608	187.5
10:	Nonsense_Mutation	3588	20.386	15.0
11:	Nonstop_Mutation	60	0.341	0.0
12:	Splice_Site	2368	13.455	11.0
13:	total	47266	268.557	221.0

```
## Gene Summary..
```

	Hugo_Symbol	Frame_Shift_Del	Frame_Shift_Ins	In_Frame_Del
1:	TP53	12	2	2
2:	TTN	7	0	1
3:	CSMD3	4	0	0
4:	MUC16	1	0	0
5:	RYR2	0	0	0

13241:	ZSCAN29	0	0	0
13242:	ZUFSP	0	0	0
13243:	ZW10	0	0	0
13244:	ZXDB	0	0	0
13245:	ZYX	0	0	0

	In_Frame_Ins	Missense_Mutation	Nonsense_Mutation	Nonstop_Mutation
1:	1	96	20	0
2:	0	264	19	0
3:	0	96	15	0
4:	0	125	11	0
5:	0	105	4	0

13241:	0	1	0	0
13242:	0	1	0	0
13243:	0	1	0	0
13244:	0	1	0	0
13245:	0	1	0	0

	Splice_Site	total	MutatedSamples	AlteredSamples
1:	17	150	145	145
2:	5	296	126	126
3:	6	121	81	81
4:	3	140	77	77
5:	7	116	76	76

13241:	0	1	1	1
13242:	0	1	1	1
13243:	0	1	1	1
13244:	0	1	1	1
13245:	0	1	1	1

```
## NOTE: Possible FLAGS among top ten genes:
## [1] "TTN"   "MUC16" "USH2A" "SYNE1"
## Checking clinical data..
## Annotation missing for below samples in MAF
##  [1] "TCGA-18-3406" "TCGA-18-3407" "TCGA-18-3408" "TCGA-18-3409"
##  [5] "TCGA-18-3410" "TCGA-18-3411" "TCGA-18-3412" "TCGA-18-3414"
##  [9] "TCGA-18-3415" "TCGA-18-3416" "TCGA-18-3417" "TCGA-18-3419"
## [13] "TCGA-18-3421" "TCGA-18-4083" "TCGA-18-4086" "TCGA-18-4721"
## [17] "TCGA-18-5592" "TCGA-18-5595" "TCGA-21-1070" "TCGA-21-1071"
## [21] "TCGA-21-1076" "TCGA-21-1077" "TCGA-21-1078" "TCGA-21-1081"
## [25] "TCGA-21-5782" "TCGA-21-5784" "TCGA-21-5786" "TCGA-21-5787"
## [29] "TCGA-22-0944" "TCGA-22-1002" "TCGA-22-1011" "TCGA-22-1012"
## [33] "TCGA-22-1016" "TCGA-22-4591"
## Done !
```

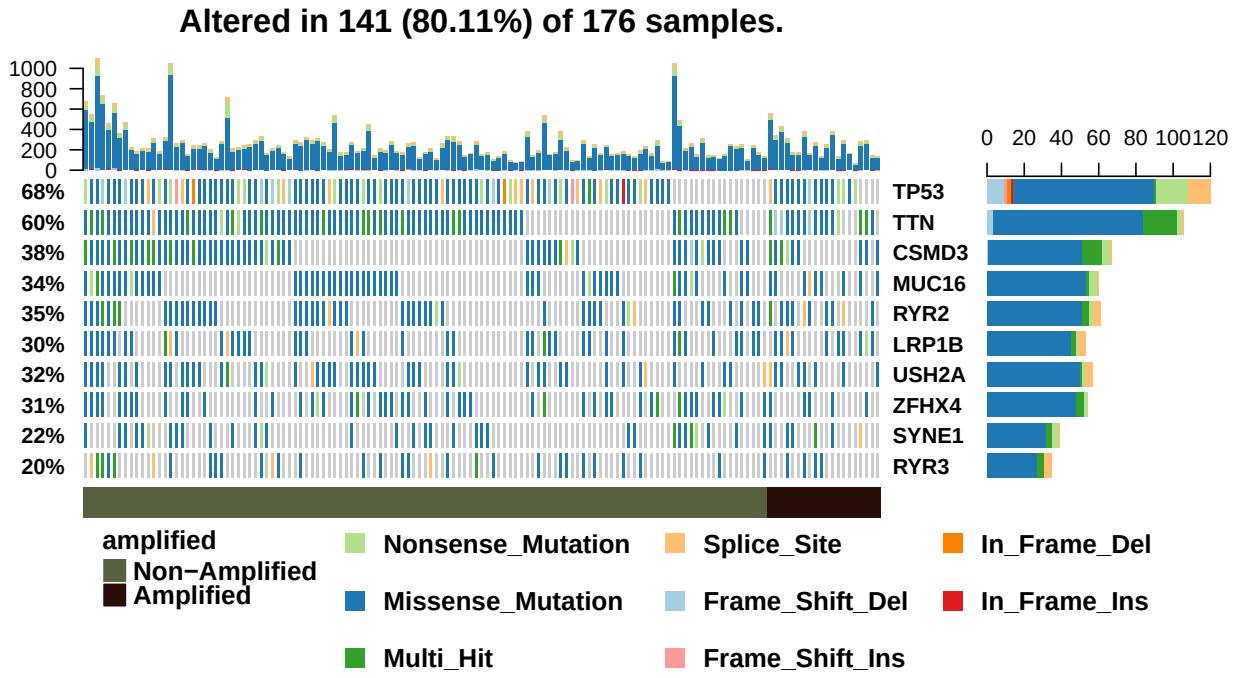
```r
# Set colours for oncoplot

col <- brewer.pal(n = 8, name = 'Paired')
names(col) <- c('Frame_Shift_Del','Missense_Mutation',
                'Nonsense_Mutation', 'Multi_Hit',
                'Frame_Shift_Ins','In_Frame_Ins',
                'Splice_Site', 'In_Frame_Del')
```

Plotting Mutation Data

Here, we use built-in plotting functions from maftools to look at the mutational status of patient samples.

Figure 10. Top 10 Mutated Genes Across LUSC Cohort and Grouped by NSD3-Amplification Status

```r
# maftools oncoplot to display top 10 mutated genes across LUSC samples

oncoplot(maf = lusc.maf.in, colors = col,
         clinicalFeatures = "amplified",
         sortByAnnotation = TRUE, top = 10)
```



Next, we can look more specifically at mutated genes enriched in patients with amplified NSD3.

Figure 11 Mutations Enriched in NSD3-Amplified Samples

```
# Use maftools to identify mutated genes enrichment by amplification status - p less than 0.01

amp.ce <- clinicalEnrichment(maf = lusc.maf.in, clinicalFeature = "amplified")
```

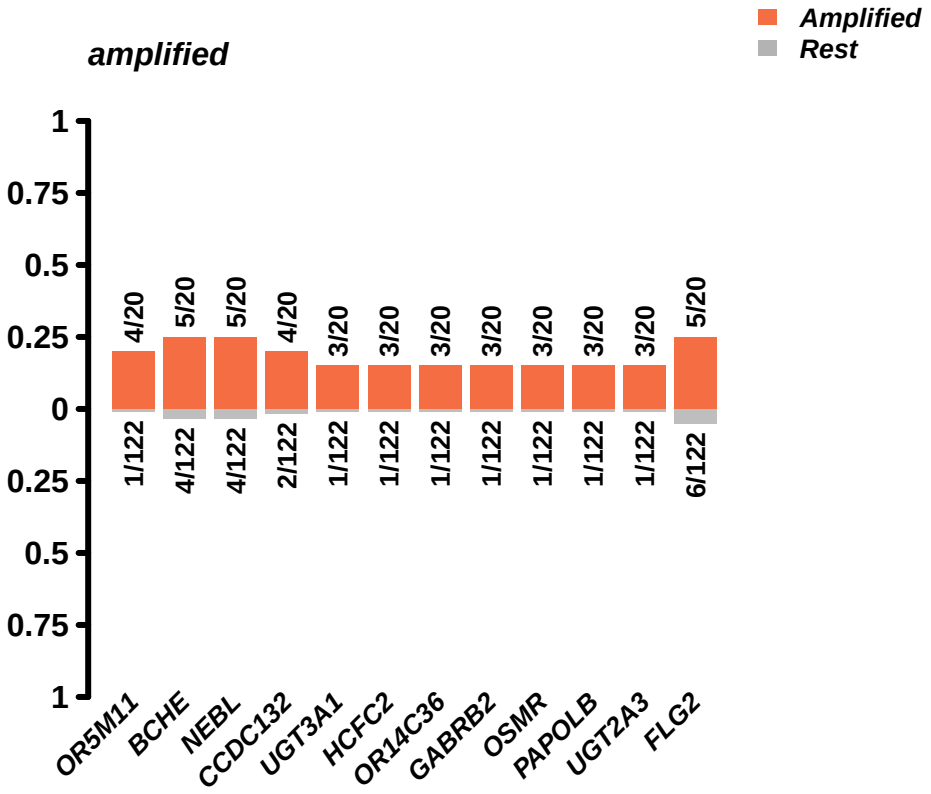
```
## Sample size per factor in amplified:

##
##     Amplified Non-Amplified
##            20           122
```

```
print(amp.ce$groupwise_comparision %>% filter(p_value < 0.01))
```

```
##    Hugo_Symbol    Group1 Group2 n_mutated_group1 n_mutated_group2
## 1       OR5M11 Amplified   Rest          4 of 20         1 of 122
## 2         BCHE Amplified   Rest          5 of 20         4 of 122
## 3         NEBL Amplified   Rest          5 of 20         4 of 122
## 4      CCDC132 Amplified   Rest          4 of 20         2 of 122
## 5       UGT3A1 Amplified   Rest          3 of 20         1 of 122
## 6        HCFC2 Amplified   Rest          3 of 20         1 of 122
## 7      OR14C36 Amplified   Rest          3 of 20         1 of 122
## 8       GABRB2 Amplified   Rest          3 of 20         1 of 122
## 9         OSMR Amplified   Rest          3 of 20         1 of 122
## 10      PAPOLB Amplified   Rest          3 of 20         1 of 122
## 11      UGT2A3 Amplified   Rest          3 of 20         1 of 122
## 12        FLG2 Amplified   Rest          5 of 20         6 of 122
##        p_value OR_low   OR_high fdr
## 1  0.001353786      0 0.2838071   1
## 2  0.002967631      0 0.4310186   1
## 3  0.002967631      0 0.4310186   1
## 4  0.003684024      0 0.4007113   1
```

```
## 5  0.008865377      0 0.4671807   1
## 6  0.008865377      0 0.4671807   1
## 7  0.008865377      0 0.4671807   1
## 8  0.008865377      0 0.4671807   1
## 9  0.008865377      0 0.4671807   1
## 10 0.008865377      0 0.4671807   1
## 11 0.008865377      0 0.4671807   1
## 12 0.008960651      0 0.5908272   1
```

```
plotEnrichmentResults(enrich_res = amp.ce, pVal = 0.01)
```



Observations. From this analysis, we can see that mutation of the tumour supressor p53 is one the most common events across both NSD3 amplified and non-amplified lung patient samples (Figure 10). Looking at only mutated genes that are enriched in NSD3 amplified samples, we identify 12 genes with a p value below 0.01 (Figure 11). I ran a quick GO analysis on the 12 genes identified as enriched, which identified the flavonoid biosynthetic processes as the only enriched term based on UDP glucuronosyltransferases UGT2A3 and UGT3A1. Interestingly, these two factors may be important for metabolizing polyaromatic hydrocarbons (PAHs) associated with tobacco smoking (Bushley RT, et al. 2011 - PMID:21164388). In the absence of these enzymes and other factors that promote genomic stability gene amplification events may become more prevalent, as seen with NSD3. Based on the fact that the overall mutational profile of NSD3 amplified samples is relatively similar to those that do not display amplification of NSD3 may indicate that this is a later event in the progression of the disease and not an initilizing genomic lesion.

ExpID-020