Correlation between EGFR mutation status and somatic SNV mutations on proteins, and its difference between smoking status

1. Background

Our data, which consists of lung cancer patients in East Asia, is characterized by predominant EGFR mutations. Known mutation of EGFR gene are L858R mutation and exon 19 deletion.

Among these patients, 337 SNV mutated peptides corresponding to 319 proteins were identified, in which among variant isoforms of cancer driver genes were identified. TP53BP1, RNF213, and KRAS mutations are in the top ranking genes.

The goal of this project is to find out the correlation between EGFR mutation type and SNV mutation type, and it's difference among different smoking status.

2. Exploring Data

2.1. Importing Data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(readr)
```

```
# importing characteristics and clinical data of TW lung cancer patients
clinical_patient <- read_csv("clinical_patient.csv")
```

```
## Rows: 103 Columns: 9

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): ID, Proteome_Batch, Gender, Smoking Status, Histology Type, Stage, ...
## dbl (1): Age
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# importing list of translated somatic SNV mutations on proteins
snv_info <- read_csv("snv_info.csv")
```

```
## Rows: 377 Columns: 15
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): Acession number_mutation, Protein, Chromosome, Reference_Allele, Al...
## dbl (8): No. of patients, positition, WES_depth, WES_ALF, RNAseq_T_depth, RN...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Let's find out what column names were used in these datasets.

```
colnames(clinical_patient)
```

```
## [1] "ID"                    "Proteome_Batch"        "Gender"
## [4] "Age"                   "Smoking Status"        "Histology Type"
## [7] "Stage"                 "EGFR_Status"           "Primary Tumor Location"
```

```
colnames(snv_info)
```

```
##  [1] "Acession number_mutation" "Protein"
##  [3] "No. of patients"          "Chromosome"
##  [5] "positition"               "Reference_Allele"
##  [7] "Alteration"               "Patient ID"
##  [9] "Batch"                    "WES_depth"
## [11] "WES_ALF"                  "RNAseq_T_depth"
## [13] "RNAseq_T_AF"              "RNAseq_N_depth"
## [15] "RNAseq_N_AF"
```

2.2. Modifying Data

Right now, our snv_info data is providing the reference allele and alternative allele at a seperate column.

```
snv_info %>%
  select(Protein, Chromosome, Reference_Allele, Alteration)
```

```
## # A tibble: 377 x 4
##    Protein Chromosome Reference_Allele Alteration
##    <chr>   <chr>      <chr>            <chr>
##  1 ABCF3   chr3       G                A
##  2 ACAA1   chr3       T                C
##  3 ACAN    chr15      A                G
```

```
##  4 ACAT2   chr6       A                G
##  5 ACIN1   chr14      A                G
##  6 ACIN1   chr14      T                C
##  7 ACOT2   chr14      A                G
##  8 ACSL1   chr4       C                T
##  9 ACSL5   chr10      G                A
## 10 ACTN2   chr1       G                A
## # ... with 367 more rows
```

Let's make a new column, 'snv_type' where we can check the nucleotide change at once.

```
snv_info <- snv_info %>%
mutate(snv_type = paste(as.character(.$Reference_Allele), '>', as.character(.$Alteration)))

snv_info %>% select(snv_type)
```

```
## # A tibble: 377 x 1
##    snv_type
##    <chr>
##  1 G > A
##  2 T > C
##  3 A > G
##  4 A > G
##  5 A > G
##  6 T > C
##  7 A > G
##  8 C > T
##  9 G > A
## 10 G > A
## # ... with 367 more rows
```

We are also going to add a new column named 'top_cancer_driver', indicating the peptides that are isoforms of top rank cancer driver genes, mentioned above.

```
snv_info <- snv_info %>%
  mutate(top_cancer_driver = ifelse(Protein %in% c('TP53BP1', 'KRAS', 'RNF213'), 'Y', 'N'))
```

Now let's merge our two datasets together! We will merge by patient ID.

```
snv_info <- snv_info %>% rename(ID = `Patient ID`)

merged_ds <- merge(clinical_patient, snv_info, by = 'ID')
head(merged_ds)
```

```
##     ID Proteome_Batch Gender Age Smoking Status Histology Type Stage
## 1 P002          B01-2   Male  74       Nonsmoke            ADC    IB
## 2 P002          B01-2   Male  74       Nonsmoke            ADC    IB
## 3 P002          B01-2   Male  74       Nonsmoke            ADC    IB
## 4 P007          B02-3   Male  67       Nonsmoke            ADC   IIA
## 5 P009          B03-1 Female  54       Nonsmoke            ADC   IIA
## 6 P009          B03-1 Female  54       Nonsmoke            ADC   IIA
##   EGFR_Status Primary Tumor Location Acession number_mutation Protein
```

```
## 1      others                    LUL           NP_071766_E31Q   TOR3A
## 2      others                    LUL          NP_004542_D190N  NDUFS3
## 3      others                    LUL           NP_613258_A84P   H2AFY
## 4          WT                    RLL          NP_005737_L394V   NAMPT
## 5       L858R                    LLL         NP_000248_E1401Q    MYH7
## 6       L858R                    LLL          NP_003117_E851Q   SPTA1
##   No. of patients Chromosome posititon Reference_Allele Alteration Batch
## 1               1       chr1  179051354                G           C B01-2
## 2               1      chr11   47603961                G           A B01-2
## 3               1       chr5  134705755                C           G B01-2
## 4               1       chr7  105894860                G           C B02-3
## 5               1      chr14   23886864                C           G B03-1
## 6               2       chr1  158631113                C           G B03-1
##   WES_depth    WES_ALF RNAseq_T_depth RNAseq_T_AF RNAseq_N_depth RNAseq_N_AF
## 1       118 0.06779661              2      0.0000              0           0
## 2       162 0.09259259             61      0.0000            171           0
## 3       157 0.11464968            270      0.1185            475           0
## 4       244 0.17622951           2368      0.1981            333           0
## 5       124 0.04032258              0      0.0000              0           0
## 6       213 0.05164319              0      0.0000              0           0
##   snv_type top_cancer_driver
## 1    G > C                N
## 2    G > A                N
## 3    C > G                N
## 4    G > C                N
## 5    C > G                N
## 6    C > G                N
```

3. Visualizing Data

3.1. Creating our main plot

What we want to know is the ratio of SNV type to each EGFR status.

To get this information, we will create a new dataset called 'snv_percentage', containing the snv rate of each snv type within each EGFR status, divided by Smoking Status.

```
snv_percentage <- merged_ds %>%
  group_by(`Smoking Status`, EGFR_Status) %>%
  count(snv_type) %>%
  summarize(snv_type = snv_type, snv_rate = n/sum(n))
```

```
## `summarise()` has grouped output by 'Smoking Status', 'EGFR_Status'. You can override using the `.gr
```

```
snv_percentage
```

```
## # A tibble: 53 x 4
## # Groups:   Smoking Status, EGFR_Status [8]
##    `Smoking Status` EGFR_Status snv_type snv_rate
##    <chr>            <chr>       <chr>       <dbl>
##  1 Current_Smoker   WT          C > A      0.75
##  2 Current_Smoker   WT          C > T      0.25
```

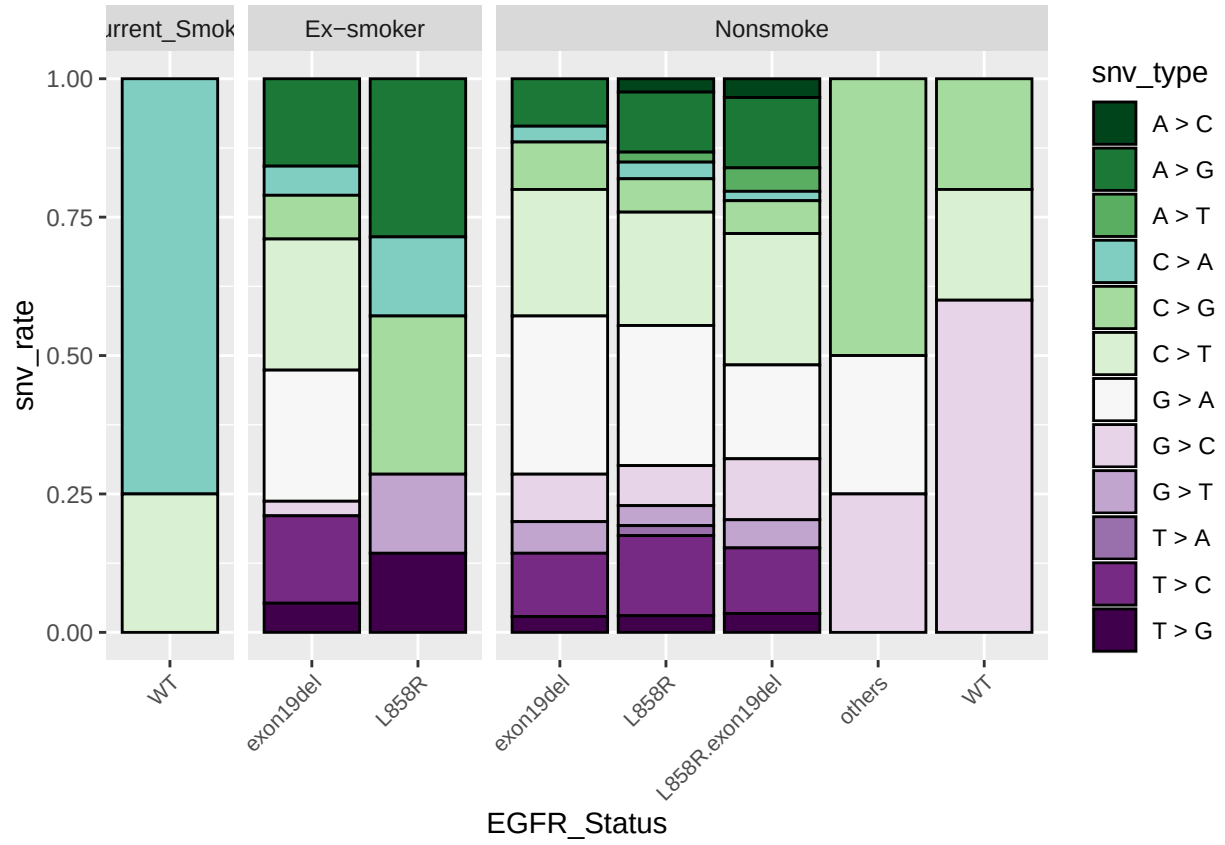
```
##  3 Ex-smoker        exon19del   A > G      0.158
##  4 Ex-smoker        exon19del   C > A      0.0526
##  5 Ex-smoker        exon19del   C > G      0.0789
##  6 Ex-smoker        exon19del   C > T      0.237
##  7 Ex-smoker        exon19del   G > A      0.237
##  8 Ex-smoker        exon19del   G > C      0.0263
##  9 Ex-smoker        exon19del   T > C      0.158
## 10 Ex-smoker        exon19del   T > G      0.0526
## # ... with 43 more rows
```

Now let's visualize our data! We will be using a bar plot.

```
pal = c('#00441b','#1b7837', '#5aae61','#80cdc1',
        '#a6dba0','#d9f0d3' , '#f7f7f7','#e7d4e8',
        '#c2a5cf', '#9970ab' ,'#762a83','#40004b')

snv_bar <- snv_percentage %>%
  ggplot(aes(EGFR_Status, snv_rate, fill = snv_type))+
    geom_bar(stat = 'identity', color = 'black')+
    facet_grid(~ `Smoking Status`, scale = 'free', space = 'free_x')+
  scale_fill_manual(values = pal)+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1, size = 8))

snv_bar
```



3.2. Highlighting top rank cancer drivers

We want to highlight our top cancer driver genes, since there is a high chance that their SNV mutation is actually related to EGFR mutation and patients' smoking status.

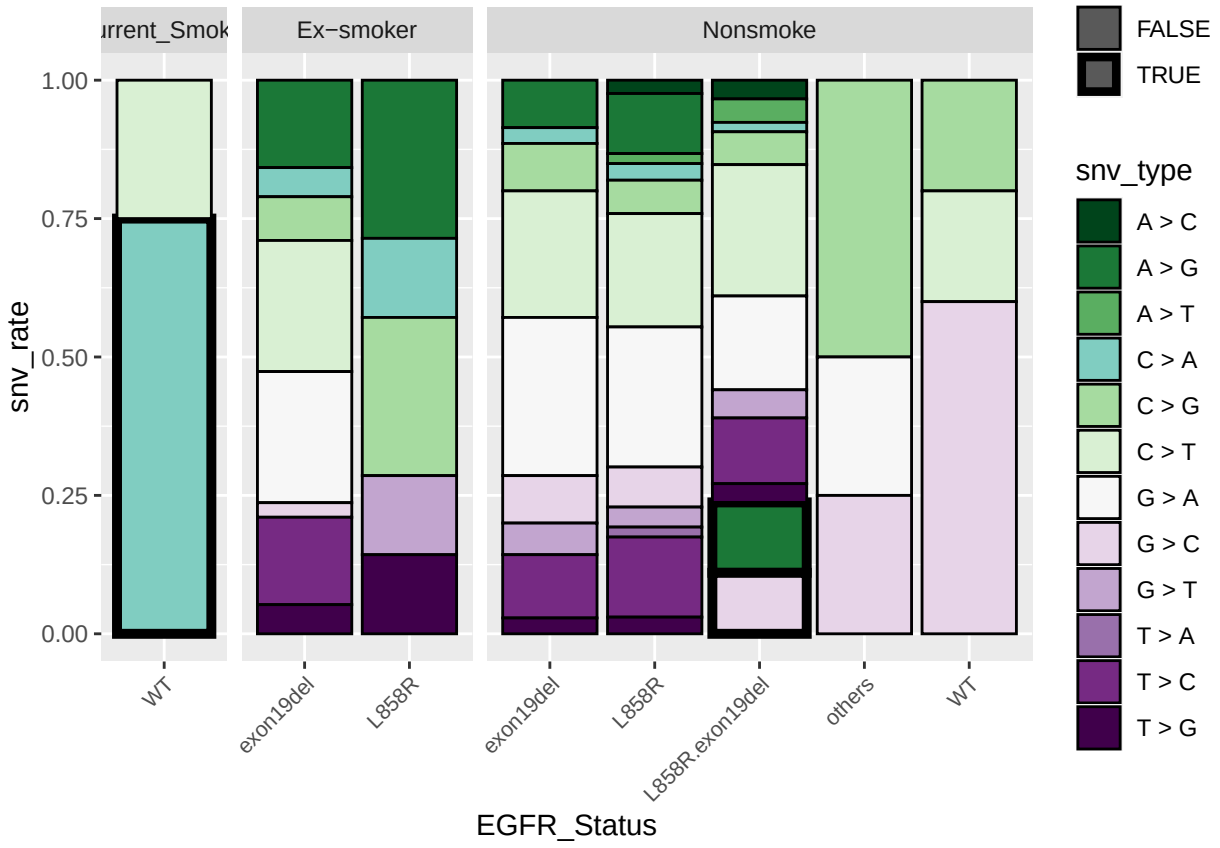
First, let's check which categories they belong to

```
merged_ds %>%
  filter(top_cancer_driver == 'Y') %>%
  select(ID, Protein, `Smoking Status`, EGFR_Status, snv_type)
```

```
##     ID Protein Smoking Status     EGFR_Status snv_type
## 1 P051  RNF213       Nonsmoke L858R.exon19del    G > C
## 2 P051  RNF213       Nonsmoke L858R.exon19del    A > G
## 3 P051 TP53BP1       Nonsmoke L858R.exon19del    G > C
## 4 P061    KRAS Current_Smoker              WT    C > A
```

We have four peptides that are isoforms of top cancer driver genes, which we will now highlight within our plot.

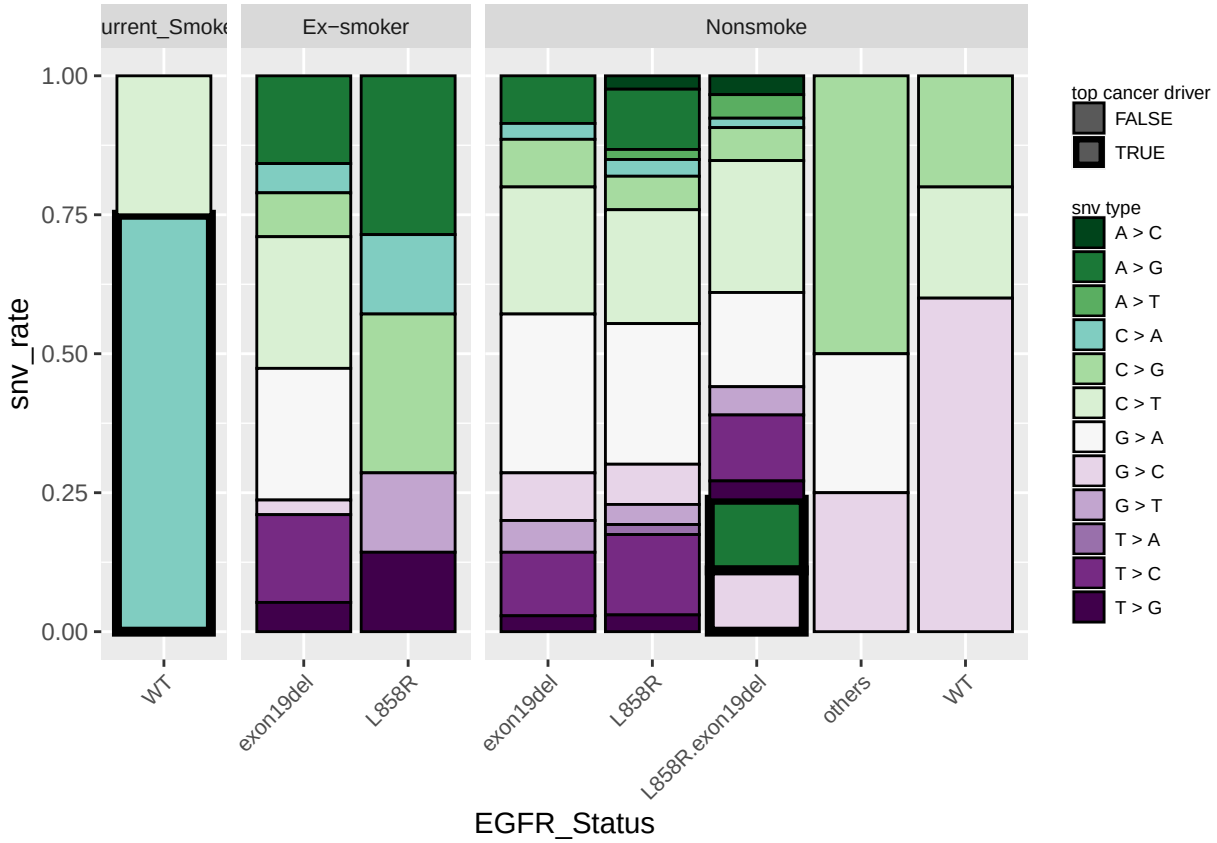
```
snv_percentage %>%
  mutate(tcd = (`Smoking Status` == 'Nonsmoke' &
                       EGFR_Status == 'L858R.exon19del' & snv_type %in% c("G > C", "A > G")) |
                      (`Smoking Status` == 'Current_Smoker' & EGFR_Status == "WT" &
                      snv_type == "C > A"))%>%
   ggplot(aes(EGFR_Status, snv_rate, fill = snv_type))+
   geom_bar(aes(size = tcd, col = tcd),
            stat = 'identity', color = 'black', position = "fill") +
   facet_grid(~ `Smoking Status`, scale = 'free', space = 'free_x') +
   scale_fill_manual(values = pal) +
  scale_size_manual(values = c(0.5, 1.75)) +
   theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1, size = 8))
```



3.3. Modifying our plot

First, we want both of our legends to be shown clearly.

```
snv_percentage %>%
  mutate(tcd = (`Smoking Status` == 'Nonsmoke' &
                         EGFR_Status == 'L858R.exon19del' & snv_type %in% c("G > C", "A > G")) |
                        (`Smoking Status` == 'Current_Smoker' & EGFR_Status == "WT" &
                        snv_type == "C > A"))%>%
   ggplot(aes(EGFR_Status, snv_rate, fill = snv_type))+
   geom_bar(aes(size = tcd, col = tcd),
            stat = 'identity', color = 'black', position = "fill") +
   facet_grid(~ `Smoking Status`, scale = 'free', space = 'free_x') +
  theme(strip.text.x = element_text(size = 7.5))+
   scale_fill_manual(values = pal) +
  scale_size_manual(values = c(0.5, 1.75)) +
   theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1, size = 8))+
  labs(size = 'top cancer driver', fill = 'snv type')+
  theme(legend.key.size = unit(0.45, 'cm'),
    legend.spacing.y = unit(0, "cm"),
        legend.title = element_text(size = 7),
        legend.text = element_text(size = 7))
```



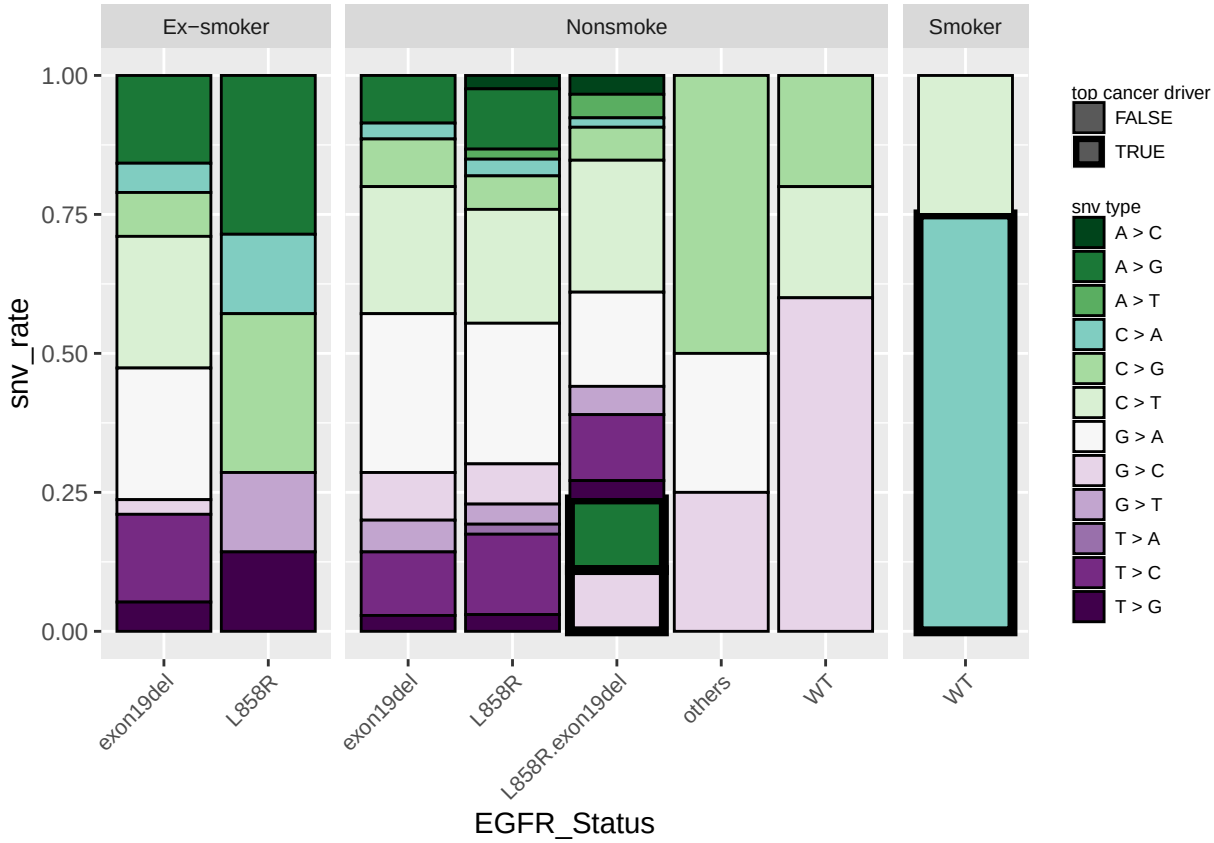
Finally, we can't see the full 'current_smoker' label, so we will change our 'Current_Smoker' status to 'Smoker'.

```
snv_percentage$`Smoking Status` <-
  gsub('Current_Smoker', 'Smoker',
       snv_percentage$`Smoking Status`)
```

Now let's produce our final plot!

```
snv_percentage %>%
  mutate(tcd = (`Smoking Status` == 'Nonsmoke' &
                        EGFR_Status == 'L858R.exon19del' & snv_type %in% c("G > C", "A > G")) |
                       (`Smoking Status` == 'Smoker' & EGFR_Status == "WT" &
                       snv_type == "C > A"))%>%
   ggplot(aes(EGFR_Status, snv_rate, fill = snv_type))+
   geom_bar(aes(size = tcd, col = tcd),
            stat = 'identity', color = 'black', position = "fill") +
   facet_grid(~ `Smoking Status`, scale = 'free', space = 'free_x') +
  theme(strip.text.x = element_text(size = 7.5))+
   scale_fill_manual(values = pal) +
  scale_size_manual(values = c(0.5, 1.75)) +
   theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1, size = 8))+
  labs(size = 'top cancer driver', fill = 'snv type')+
  theme(legend.key.size = unit(0.45, 'cm'),
    legend.spacing.y = unit(0, "cm"),
```

```
        legend.title = element_text(size = 7),
        legend.text = element_text(size = 7))
```



4. Discussion

The research indicated that 'C > T' was the most common in this cohort, which is quite consistent with our plot, especially among 'non' and 'ex' smokers with exon 19 deletion and L858R mutation at EGFR. Another prevalent SNV type among 'non' and 'ex' smokers with known EGFR mutation type is 'G > A'. Interestingly, there aren't any 'C > T' or 'G > A' found in 'Ex-smokers' with L858R mutation, so this can be a matter of further research.

Among 'Smokers', 'C > A', which is known to be smoking - related, accounts for the majority.

All of the mutations from our 'top cancer driver genes' come from only 2 patients, so it might be difficult to derive meaningful results just from this plot. However, it can be used to support studies finding pathogenic functions of SNV types in lung cancer, especially in correlation with EGFR mutation.