

Group 3 Variant_Analysis

2025-07-10

```
#get working directory  
getwd() #" /Users/sequencingplatform/Documents/Linux_Basics_to_Mastery_training/Module-15"
```

```
## [1] "/Users/sequencingplatform/Documents/Linux_Basics_to_Mastery_training/Module-15"
```

```
#load libraries  
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'  
  
## The following objects are masked from 'package:stats':  
##  
##   filter, lag  
  
## The following objects are masked from 'package:base':  
##  
##   intersect, setdiff, setequal, union
```

```
library(openxlsx)  
library(janitor)
```

```
##  
## Attaching package: 'janitor'  
  
## The following objects are masked from 'package:stats':  
##  
##   chisq.test, fisher.test
```

```
library(stringi)  
library(stringr)  
library(ggplot2)  
library(data.table)
```

```
##  
## Attaching package: 'data.table'  
  
## The following objects are masked from 'package:dplyr':  
##  
##   between, first, last
```

```
#load hq_allvariants.tsv
data_allvariants <- fread("vcf/hq_allvariants.tsv")
```

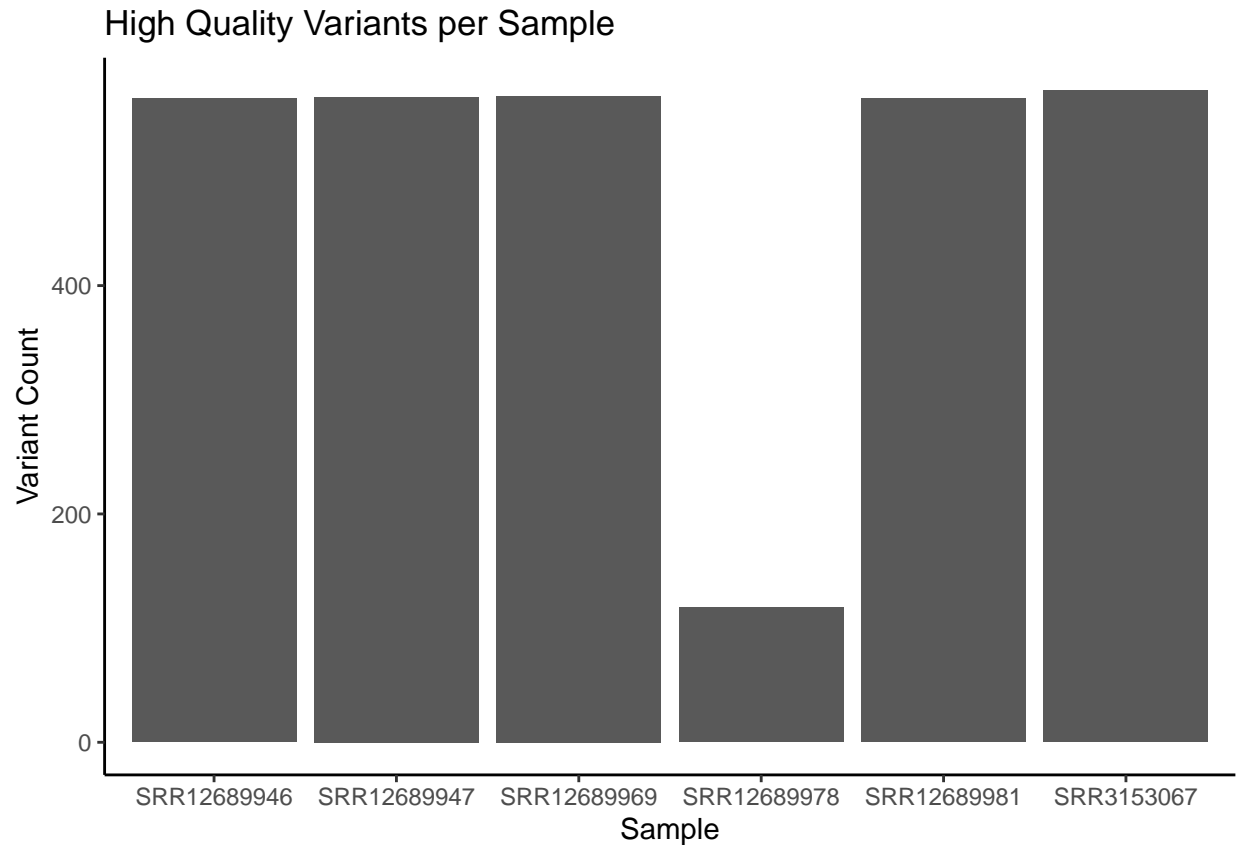
```
#Calculate and plot the total number of high quality variants per sample
```

```
high_qual_vars <- data_allvariants %>%
  group_by(Sample) %>%
  summarise(count = n(), .groups = "drop")

high_qual_vars
```

```
## # A tibble: 6 x 2
##   Sample      count
##   <chr>      <int>
## 1 SRR12689946    564
## 2 SRR12689947    565
## 3 SRR12689969    566
## 4 SRR12689978    118
## 5 SRR12689981    564
## 6 SRR3153067     571
```

```
plot1 <- high_qual_vars %>%
  ggplot(aes(x = Sample, y = count)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "High Quality Variants per Sample",
       x = "Sample",
       y = "Variant Count") +
  theme_classic()
plot1
```



```
#Merge REF and ALT column to a new column mutation
data_allvariants <- data_allvariants %>%
  mutate(mutation = paste0(REF,ALT))
```

```
#Determine the type of each variant (Transition:A<->C, OR G<->C, or Transversion) and add them to a new
is_transition <- function(ref, alt) {
  (ref == "A" & alt == "G") |
  (ref == "G" & alt == "A") |
  (ref == "C" & alt == "T") |
  (ref == "T" & alt == "C")
}
```

```
data_allvariants <- data_allvariants %>%
  filter(nchar(REF) == 1 & nchar(ALT) == 1) %>% # keep only SNPs
  mutate(Type = ifelse(is_transition(REF, ALT), "Transition", "Transversion"))
```

```
#Analyze and plot the distribution of variant types per sample both count and proportions
Var_dist <- data_allvariants %>%
  na.omit() %>%
  group_by(Sample, Type) %>%
  summarise(count = n(), .groups = "drop") %>%
  mutate(prop = count/sum(count)*100)
```

```
Var_dist
```

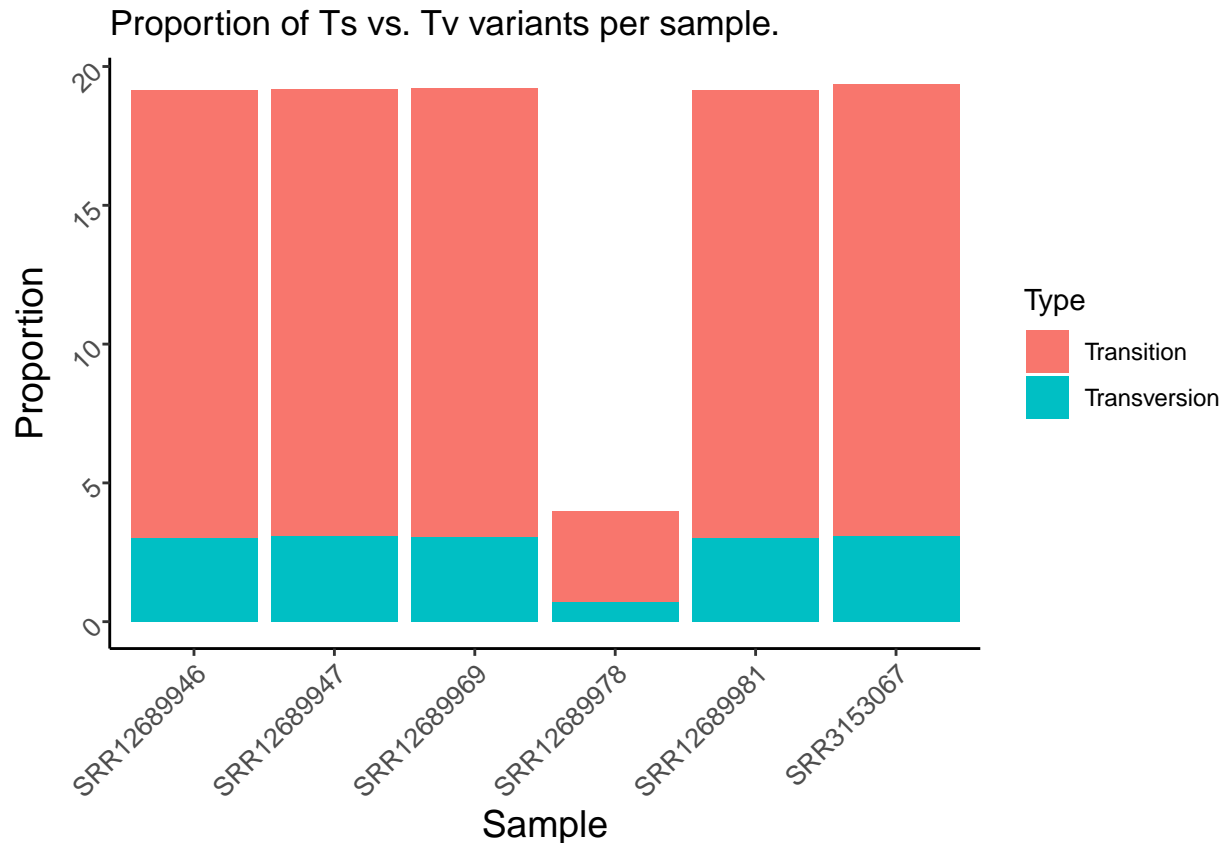
```
## # A tibble: 12 x 4
```

##	Sample	Type	count	prop
##	<chr>	<chr>	<int>	<dbl>
## 1	SRR12689946	Transition	475	16.1
## 2	SRR12689946	Transversion	89	3.02
## 3	SRR12689947	Transition	474	16.1
## 4	SRR12689947	Transversion	91	3.09
## 5	SRR12689969	Transition	476	16.2
## 6	SRR12689969	Transversion	90	3.05
## 7	SRR12689978	Transition	96	3.26
## 8	SRR12689978	Transversion	21	0.713
## 9	SRR12689981	Transition	475	16.1
## 10	SRR12689981	Transversion	89	3.02
## 11	SRR3153067	Transition	479	16.3
## 12	SRR3153067	Transversion	91	3.09

```

plot2 <- Var_dist %>%
  ggplot(aes(x = Sample, prop, fill = Type)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Proportion of Ts vs. Tv variants per sample.",
       x = "Sample",
       y = "Proportion")
) +
theme_classic() +
theme(
  axis.title = element_text(size = 14),
  axis.text = element_text(size = 10, angle = 45, hjust = 1)
)
plot2

```



```
#Identify common variants across samples or unique variants

## Create a unique identifier for each variant
data_allvariants$variant_id <- paste(data_allvariants$CHROM,
                                     data_allvariants$POS,
                                     data_allvariants$REF,
                                     data_allvariants$ALT,
                                     sep = "_")

## Keep one row per sample per variant (remove duplicates)
variant_sample_table <- data_allvariants %>%
  distinct(Sample, variant_id)

##Count how many samples each variant appears in
variant_counts <- variant_sample_table %>%
  group_by(variant_id) %>%
  summarise(sample_count = n(), .groups = "drop")

## Identify unique variants (appear in only one sample)
unique_variants <- variant_counts %>%
  filter(sample_count == 1)

## Get total number of samples
total_samples <- n_distinct(data_allvariants$Sample)

## Identify common variants (appear in all samples)
```

```
common_variants <- variant_counts %>%
  filter(sample_count == total_samples)
common_variants
```

```
## # A tibble: 114 x 2
##   variant_id      sample_count
##   <chr>          <int>
## 1 NC_002549.1_10208_C_T        6
## 2 NC_002549.1_10566_T_C        6
## 3 NC_002549.1_10569_C_A        6
## 4 NC_002549.1_10575_T_C        6
## 5 NC_002549.1_10602_A_T        6
## 6 NC_002549.1_10624_T_C        6
## 7 NC_002549.1_10979_T_A        6
## 8 NC_002549.1_11043_A_G        6
## 9 NC_002549.1_11308_C_T        6
## 10 NC_002549.1_12096_A_G       6
## # i 104 more rows
```

```
## Join back to see which sample each unique variant belongs to
unique_variants_with_samples <- unique_variants %>%
  inner_join(variant_sample_table, by = "variant_id")
unique_variants_with_samples
```

```
## # A tibble: 22 x 3
##   variant_id      sample_count Sample
##   <chr>          <int> <chr>
## 1 NC_002549.1_11817_T_C        1 SRR3153067
## 2 NC_002549.1_12780_C_T        1 SRR3153067
## 3 NC_002549.1_12910_A_G        1 SRR3153067
## 4 NC_002549.1_12996_C_A        1 SRR3153067
## 5 NC_002549.1_14449_C_T        1 SRR12689946
## 6 NC_002549.1_16514_G_A        1 SRR3153067
## 7 NC_002549.1_16600_A_G        1 SRR3153067
## 8 NC_002549.1_18467_C_T        1 SRR12689981
## 9 NC_002549.1_18622_A_C        1 SRR12689947
## 10 NC_002549.1_2309_C_A        1 SRR12689947
## # i 12 more rows
```