frequency_analysis_CNV

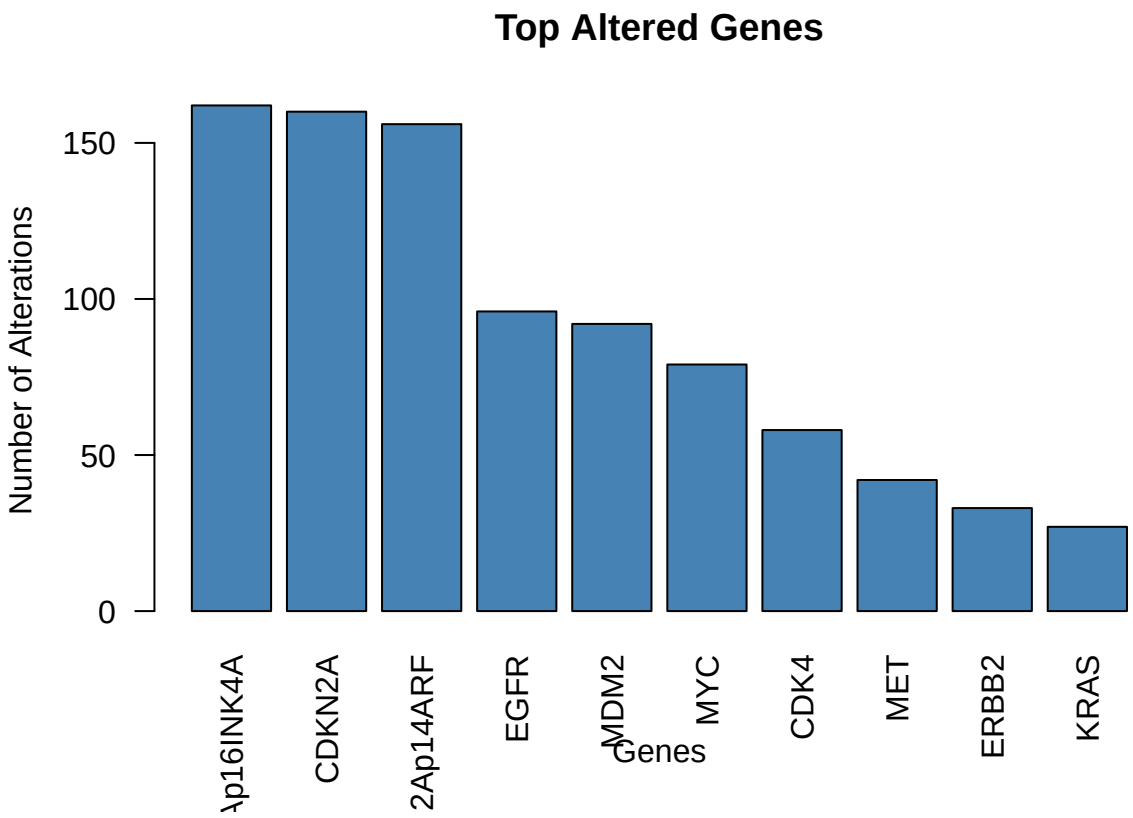
charvi khanna

2025-09-25

```r
# Load CNA data
data_cna <- read.table("C:/Users/Charvi Khanna/Documents/data_cna.txt",
                       header = TRUE,
                       sep = "\t",
                       row.names = 1,
                       stringsAsFactors = FALSE)

# Peek first few rows
head(data_cna[, 1:5])
```

```
##        HP.0006047.T03.IM6 P.0011107.T01.IM5 P.0053219.T02.IM6 P.0025284.T01.IM6
## TAP1                    0                  0                  0                  0
## ERRFI1                  0                  0                  0                  0
## STK19                   0                  0                  0                  0
## CRKL                    0                  0                  0                  0
## SCG5                    0                  0                  0                  0
## STK11                   0                  0                  0                  0
##        P.0026141.T01.IM6
## TAP1                   0
## ERRFI1                 0
## STK19                  0
## CRKL                   0
## SCG5                   0
## STK11                  0
```

```r
# Gene & sample counts
gene_counts <- rowSums(data_cna != 0)
sample_counts <- colSums(data_cna != 0)

# Top genes
top_genes <- names(head(sort(gene_counts, decreasing = TRUE), 10))

# Barplot of top genes
barplot(gene_counts[top_genes],
        names.arg = top_genes,
        las = 2,
        col = "steelblue",
        main = "Top Altered Genes",
        ylab = "Number of Alterations",
        xlab = "Genes")
```

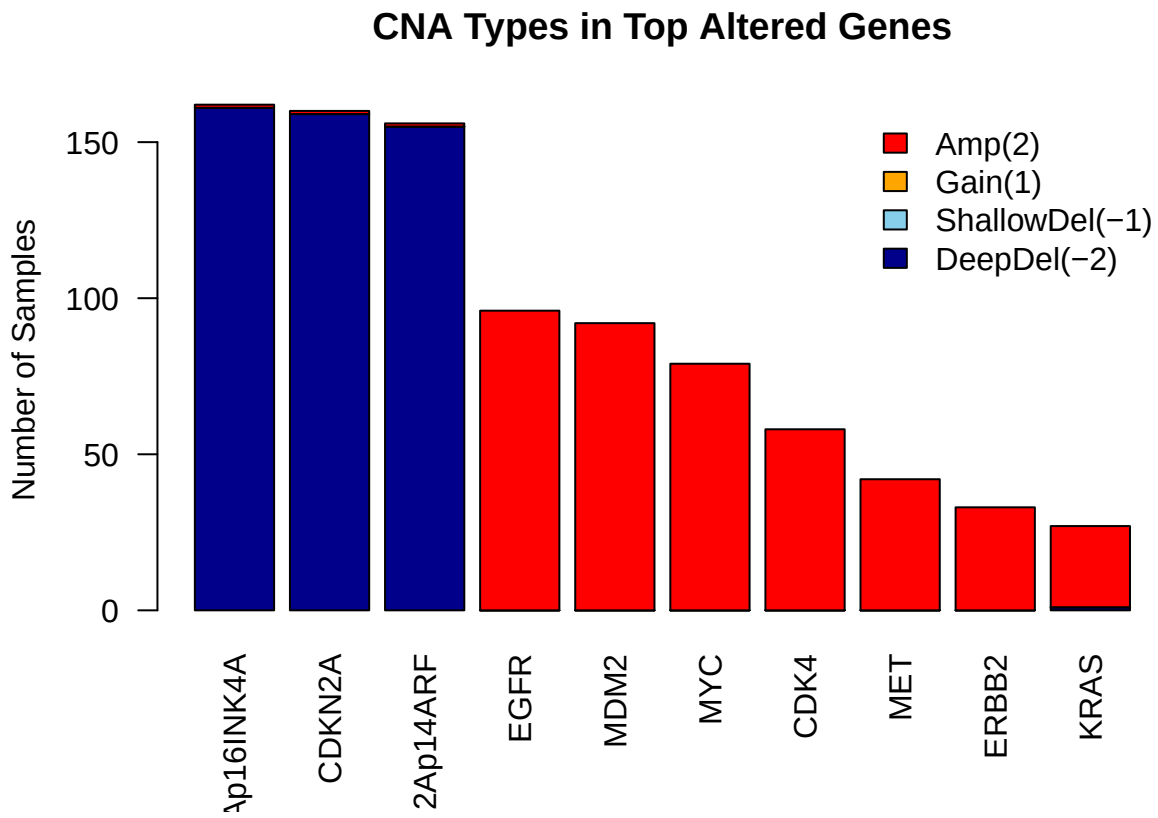



```
# Subset CNA matrix to top genes
sub_cna <- data_cna[top_genes, ]

# Count alteration types per gene
alteration_summary <- as.data.frame(t(apply(sub_cna, 1, function(x) table(factor(x, levels=c(-2, -1, 0,
colnames(alteration_summary) <- c("DeepDel(-2)", "ShallowDel(-1)", "Neutral(0)", "Gain(1)", "Amp(2)")

alteration_summary
```

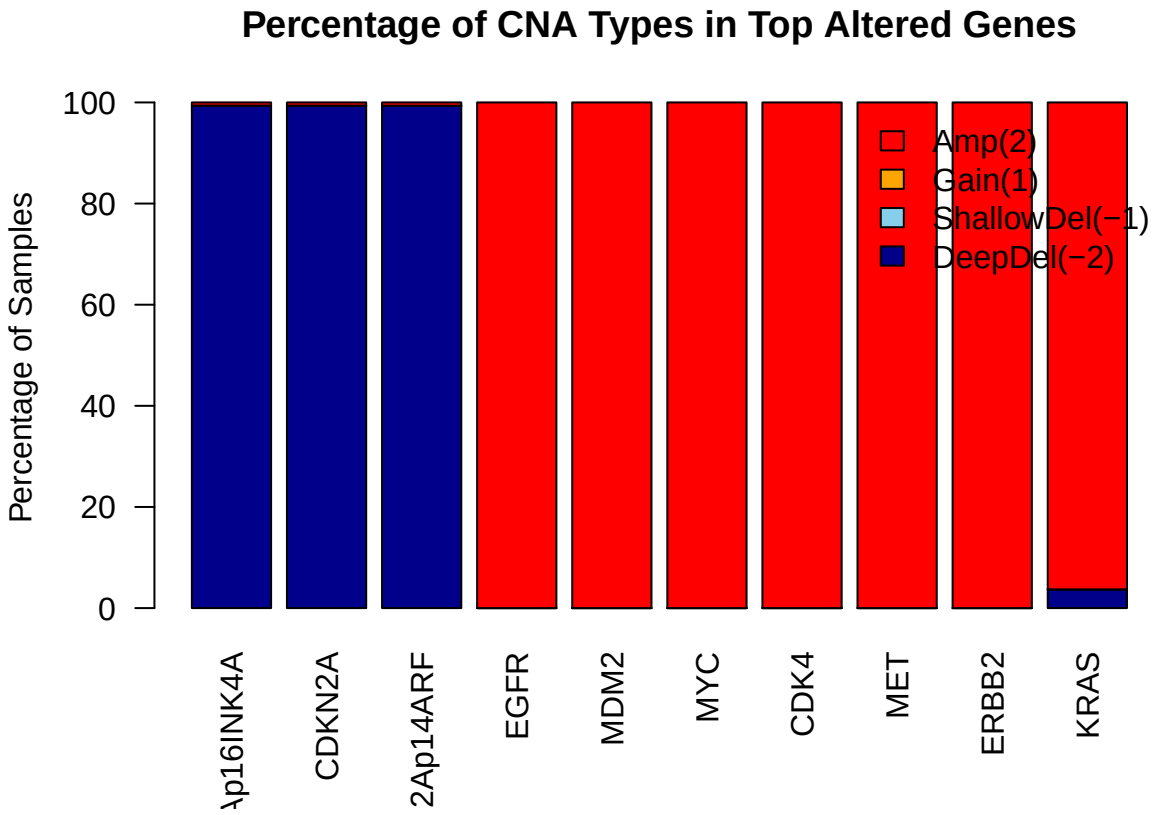
```
##                DeepDel(-2) ShallowDel(-1) Neutral(0) Gain(1) Amp(2)
## CDKN2Ap16INK4A         161              0       1095       0      1
## CDKN2A                 159              0       1097       0      1
## CDKN2Ap14ARF           155              0       1101       0      1
## EGFR                     0              0       1161       0     96
## MDM2                     0              0       1165       0     92
## MYC                      0              0       1178       0     79
## CDK4                     0              0       1199       0     58
## MET                      0              0       1215       0     42
## ERBB2                    0              0       1224       0     33
## KRAS                     1              0       1230       0     26
```

```
# Remove neutral for stacked barplot
alteration_summary_no0 <- alteration_summary[, c("DeepDel(-2)", "ShallowDel(-1)", "Gain(1)", "Amp(2)")]

# Stacked barplot counts
```

```r
barplot(t(as.matrix(alteration_summary_no0)),
        beside = FALSE,
        col = c("darkblue", "skyblue", "orange", "red"),
        legend.text = TRUE,
        args.legend = list(x="topright", bty="n"),
        main = "CNA Types in Top Altered Genes",
        ylab = "Number of Samples",
        las = 2)
```




```r
# Stacked barplot percentages
alteration_percent <- t(apply(alteration_summary_no0, 1, function(x) (x / sum(x)) * 100))
barplot(t(alteration_percent),
        beside = FALSE,
        col = c("darkblue", "skyblue", "orange", "red"),
        legend.text = TRUE,
        args.legend = list(x="topright", bty="n"),
        main = "Percentage of CNA Types in Top Altered Genes",
        ylab = "Percentage of Samples",
        las = 2)
```




```
# Binarize & compute co-occurrence
binary_cna <- ifelse(sub_cna != 0, 1, 0)
co_occur <- cor(t(binary_cna))
co_occur
```

```
##                 CDKN2Ap16INK4A      CDKN2A CDKN2Ap14ARF        EGFR       MDM2
## CDKN2Ap16INK4A     1.00000000  0.99290164   0.97862925  0.03243012 0.02865402
## CDKN2A             0.99290164  1.00000000   0.98562557  0.03397515 0.03014778
## CDKN2Ap14ARF       0.97862925  0.98562557   1.00000000  0.02803574 0.03318861
## EGFR               0.03243012  0.03397515   0.02803574  1.00000000 0.09170418
## MDM2               0.02865402  0.03014778   0.03318861  0.09170418 1.00000000
## MYC                0.07649171  0.07813436   0.08148532  0.11066863 0.09084643
## CDK4              -0.02800992 -0.02710878  -0.02528141  0.06525514 0.52058453
## MET                0.04739177  0.04853112   0.05085451  0.12987930 0.10072292
## ERBB2              0.10019241  0.10150857   0.10419960 -0.02848140 0.03027504
## KRAS               0.10678320  0.10805782   0.09402192 -0.02194295 0.10584785
##                          MYC         CDK4         MET       ERBB2         KRAS
## CDKN2Ap16INK4A  0.07649171 -0.028009919  0.04739177  0.10019241  0.106783203
## CDKN2A          0.07813436 -0.027108778  0.04853112  0.10150857  0.108057821
## CDKN2Ap14ARF    0.08148532 -0.025281413  0.05085451  0.10419960  0.094021920
## EGFR            0.11066863  0.065255139  0.12987930 -0.02848140 -0.021942954
## MDM2            0.09084643  0.520584529  0.10072292  0.03027504  0.105847850
## MYC             1.00000000 -0.010081173  0.06129498  0.03948762  0.074685573
## CDK4           -0.01008117  1.000000000  0.10681335 -0.03611359 -0.006429844
## MET             0.06129498  0.106813354  1.00000000 -0.00284138 -0.027546462
## ERBB2           0.03948762 -0.036113590 -0.00284138  1.00000000 -0.024327394
```

```
## KRAS            0.07468557 -0.006429844 -0.02754646 -0.02432739  1.000000000
```

```r
library(pheatmap)

pheatmap(co_occur,
         color = colorRampPalette(c("blue", "white", "red"))(50),
         cluster_rows = TRUE,
         cluster_cols = TRUE,
         main = "Co-occurrence of Top Altered Genes")
```