```
library(readr)
library(dplyr)
library(gtools)
library(ggplot2)
library(scales)

set.seed(42)

if (!dir.exists("figures")) dir.create("figures")
if (!dir.exists("results")) dir.create("results")
```

Read data

Read the number of times each of l codons were mutated across patients with each disease d,

$$y_d = (y_{d1}, y_{d2}, \ldots, y_{dl}).$$

```
(df <- filter(read_csv("data.csv", show_col_types=F),
              AML>0 | MDS>0 | `AML-MRC`>0))  # Disregard sites with no mutations
```

```
## # A tibble: 121 x 4
##     site   AML   MDS `AML-MRC`
##    <dbl> <dbl> <dbl>     <dbl>
##  1     4     0     1         0
##  2     9     1     0         0
##  3    11     1     0         0
##  4    23     1     0         0
##  5    39     0     0         1
##  6    46     0     1         0
##  7    47     0     1         0
##  8    48     0     1         0
##  9    54     1     0         0
## 10    72     1     0         0
## # i 111 more rows
```

```
(n <- apply(df[,2:4], 2, sum))  # Compute the sample size of each disease
```

```
##     AML     MDS AML-MRC
##     411     286     113
```

```
(l <- nrow(df))  # The number of sites considered
```

```
## [1] 121
```

```
combos <- list(c("AML", "MDS"), c("MDS", "AML-MRC"), c("AML", "AML-MRC"))
N <- 10^5  # Set the simulation size
```

Sample posterior proportions of mutations

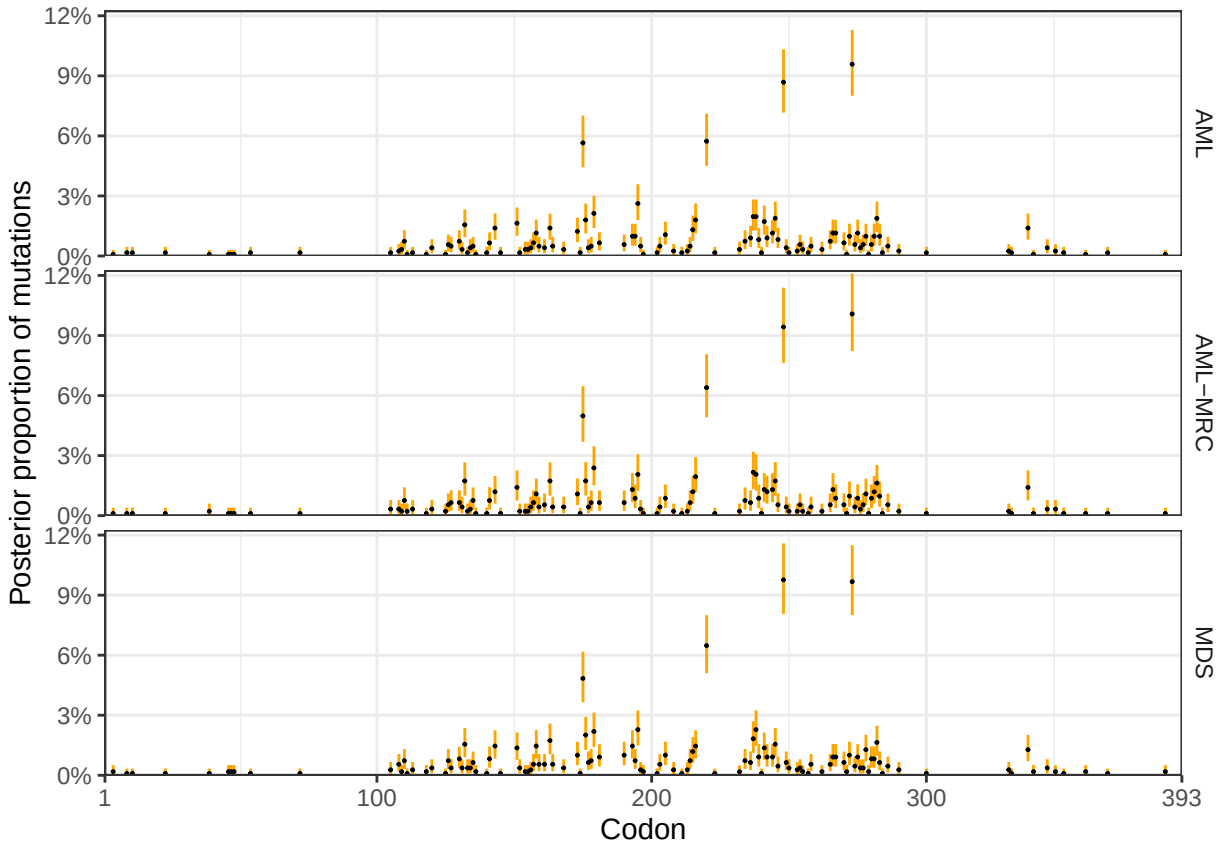
We will use the counts across all diseases to set an empirical prior $\theta_d \overset{\text{iid}}{\sim} \text{Dirichlet}(\alpha)$ over the relative probabilities of mutation at each codon, where $\alpha_i = \sum_d y_{di}$.

If we assume $y_d \sim \text{Multinomial}(\sum y_d, \theta_d)$, then the posterior $\theta_d | y_d \sim \text{Dirichlet}(\alpha + y_d)$.

```
# Sample the posterior
prior <- apply(df[2:4], 1, sum)
posts <- list()
for (di in 2:4)
  posts[[names(df)[di]]] <- rdirichlet(N, prior + df[[di]])
```

```
# Compute posterior distribution statistics
thetas <- list(disease=rep(names(posts), each=1),
               site=rep(df$site, 3))
for (d in names(posts)) {
  thetas[["mean"]] <- c(thetas[["mean"]], apply(posts[[d]], 2, mean))
  thetas[["q025"]] <- c(thetas[["q025"]], apply(posts[[d]], 2, quantile, probs=0.025))
  thetas[["q975"]] <- c(thetas[["q975"]], apply(posts[[d]], 2, quantile, probs=0.975))
}
theta_df <- as_tibble(thetas)
write_csv(theta_df, "results/proportions_blood.csv")
```

```
# Visualize inferred mutation proportions
ggplot(theta_df, aes(x=site)) +
  geom_segment(aes(xend=site, y=q025, yend=q975), color="orange") +
  geom_point(aes(y=mean), size=0.2) +
  facet_grid(rows=vars(disease)) +
  scale_x_continuous(breaks=c(1,100,200,300,393), limits=c(1,393), expand=c(0,0)) +
  scale_y_continuous(labels=percent_format(),
                     limits=c(0, max(theta_df$q975)+0.0015), expand=c(0,0)) +
  xlab("Codon") +
  ylab("Posterior proportion of mutations") +
  theme_bw() +
  theme(strip.placement="outside", strip.background=element_blank(),
        panel.grid.minor.y=element_blank())
```




```
ggsave("figures/proportions_blood.pdf", width=7, height=4)
```

Sample posterior differences in mutation proportions between diseases

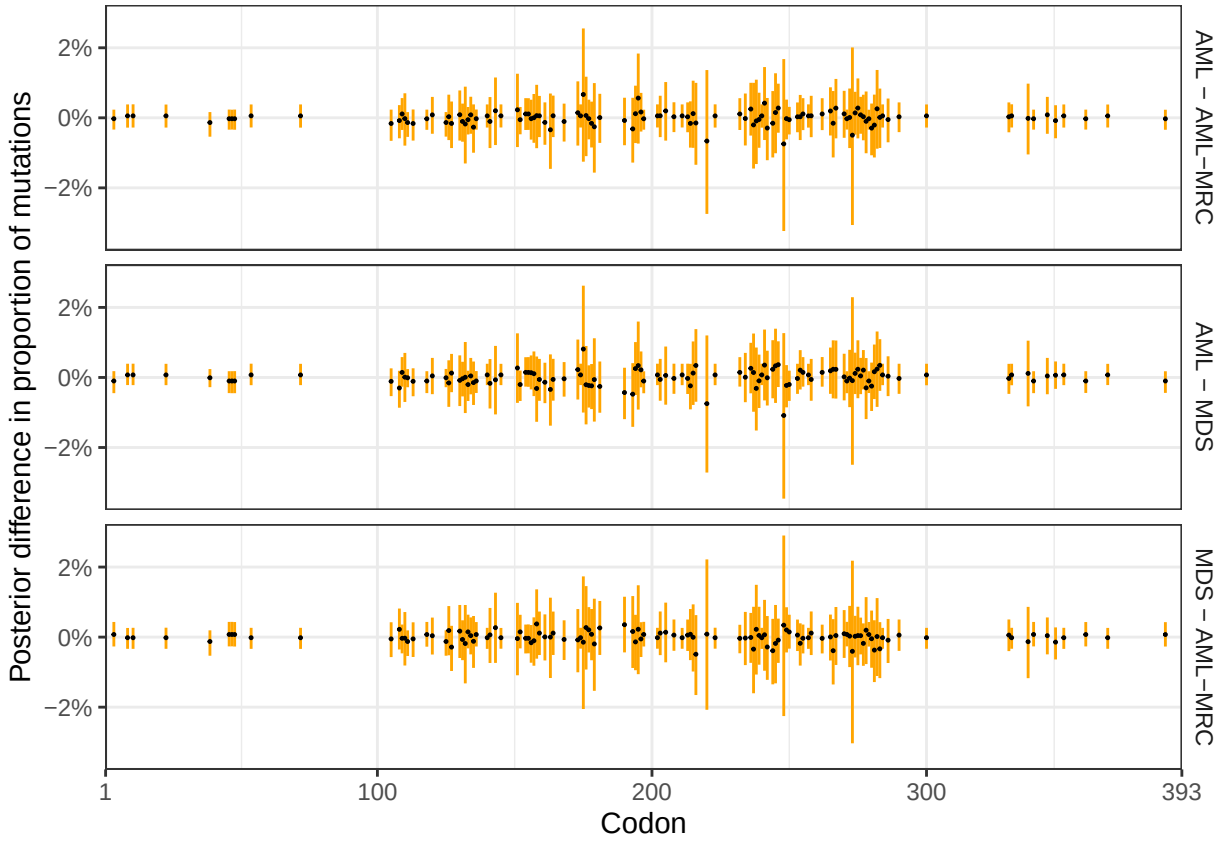
From the posterior we can sample $(\theta_d|y_d) - (\theta_{d'}|y_{d'})$, the difference between proportions of mutations at each codon for each pair of diseases d and d'.

```
diff_df <- c()
for (combo in combos) {
  # Sample the posterior proportion differences between diseases
  diff <- posts[[combo[1]]] - posts[[combo[2]]]

  # Collect statistics of the differences
  diff_df <- rbind(diff_df,
                   data.frame(combo=paste(combo[1], combo[2], sep=" - "),
                              site=df$site,
                              mean=apply(diff, 2, mean),
                              q025=apply(diff, 2, quantile, probs=0.025),
                              q975=apply(diff, 2, quantile, probs=0.975)))
}
diff_df <- as_tibble(diff_df)
write_csv(diff_df, "results/differences_blood.csv")
```

```
ggplot(diff_df, aes(x=site)) +
  geom_segment(aes(xend=site, y=q025, yend=q975), color="orange") +
```

```
  geom_point(aes(y=mean), size=0.2) +
  facet_grid(rows=vars(combo)) +
  scale_x_continuous(breaks=c(1,100,200,300,393), limits=c(1,393), expand=c(0,0)) +
  scale_y_continuous(labels=percent_format()) +
  xlab("Codon") +
  ylab("Posterior difference in proportion of mutations") +
  theme_bw() +
  theme(strip.placement="outside", strip.background=element_blank(),
        panel.grid.minor.y=element_blank())
```




```
ggsave("figures/differences_blood.pdf", width=7, height=4)

# The number of positions whose 95% central credible interval excludes zero
with(diff_df, sum(0<q025 | 0>q975))
```

```
## [1] 0
```

Compare myeloid neoplasm mutation proportions with ISB-CGC

```
# Pool the blood data
df$blood <- apply(df[2:4], 1, sum)
```

Read ISB-CGC data

> "For variants in exons, codon number at which the variant is located (1-393). If a variant spans more than one codon, (e.g. tandem variant or deletion of several bases) only the first (5') codon is entered. For variants in introns, 0 is entered." https://tp53.isb-cgc.org/help#MUT_id'

```
isb_codon_counts <- table(read_csv("TumorVariantDownload_r20.csv")$Codon_number)
isb <- c()
for (i in df$site) {
  if (as.character(i) %in% names(isb_codon_counts))
    isb <- c(isb, isb_codon_counts[[as.character(i)]])
  else
    isb <- c(isb, 0)
}
df$isb <- isb
(df)
```

```
## # A tibble: 121 x 6
##     site   AML   MDS `AML-MRC` blood   isb
##    <dbl> <dbl> <dbl>     <dbl> <dbl> <dbl>
##  1     4     0     1         0     1     0
##  2     9     1     0         0     1     0
##  3    11     1     0         0     1    12
##  4    23     1     0         0     1     0
##  5    39     0     0         1     1     2
##  6    46     0     1         0     1    15
##  7    47     0     1         0     1    16
##  8    48     0     1         0     1     5
##  9    54     1     0         0     1    11
## 10    72     1     0         0     1    17
## # i 111 more rows
```

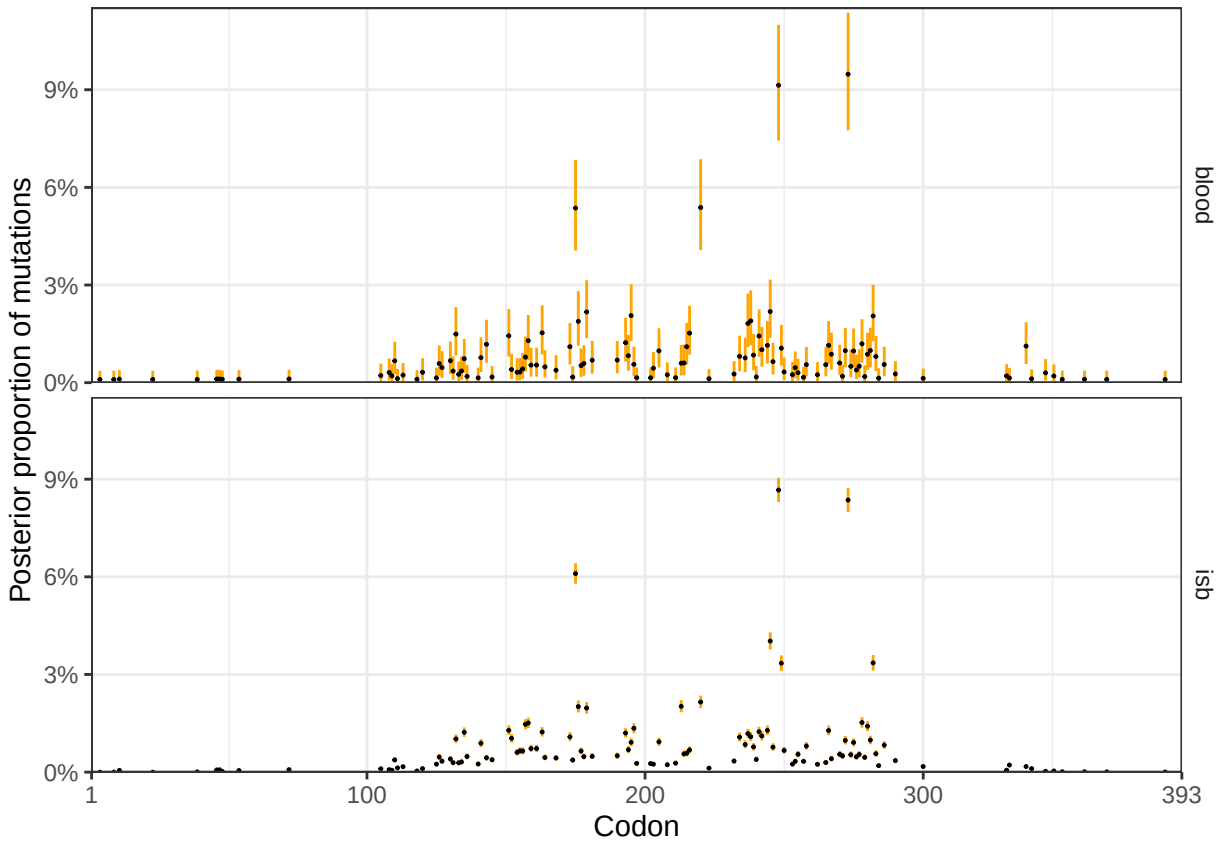
Sample posterior proportions of mutations

We will use the number of mutations at each codon observed in ISB-CGC to construct a prior θ_{blood} over the pooled myeloid neoplasm data. The prior is weighted such that $\sum \theta_{\text{blood}} = 200$. We will infer ISB-CGC proportions under a prior of $\theta_{\text{ISB}} = 0$.

```
posts <- list()
posts[["isb"]] <- rdirichlet(N, rep(0, l) + df$isb)
posts[["blood"]] <- rdirichlet(N, df$isb/sum(df$isb) * 200 + df$blood)
```

```
# Compute posterior distribution statistics
thetas <- list(disease=rep(names(posts), each=l),
               site=rep(df$site, 2))
for (d in names(posts)) {
  thetas[["mean"]] <- c(thetas[["mean"]], apply(posts[[d]], 2, mean))
  thetas[["q025"]] <- c(thetas[["q025"]], apply(posts[[d]], 2, quantile, probs=0.025))
  thetas[["q975"]] <- c(thetas[["q975"]], apply(posts[[d]], 2, quantile, probs=0.975))
}
theta_df <- as_tibble(thetas)
write_csv(theta_df, "results/proportions_blood_ISB.csv")
```

```
# Visualize inferred mutation proportions
ggplot(theta_df, aes(x=site)) +
  geom_segment(aes(xend=site, y=q025, yend=q975), color="orange") +
  geom_point(aes(y=mean), size=0.2) +
  facet_grid(rows=vars(disease)) +
  scale_x_continuous(breaks=c(1,100,200,300,393), limits=c(1,393), expand=c(0,0)) +
  scale_y_continuous(labels=percent_format(),
                     limits=c(0, max(theta_df$q975)+0.0015), expand=c(0,0)) +
  xlab("Codon") +
  ylab("Posterior proportion of mutations") +
  theme_bw() +
  theme(strip.placement="outside", strip.background=element_blank(),
        panel.grid.minor.y=element_blank())
```



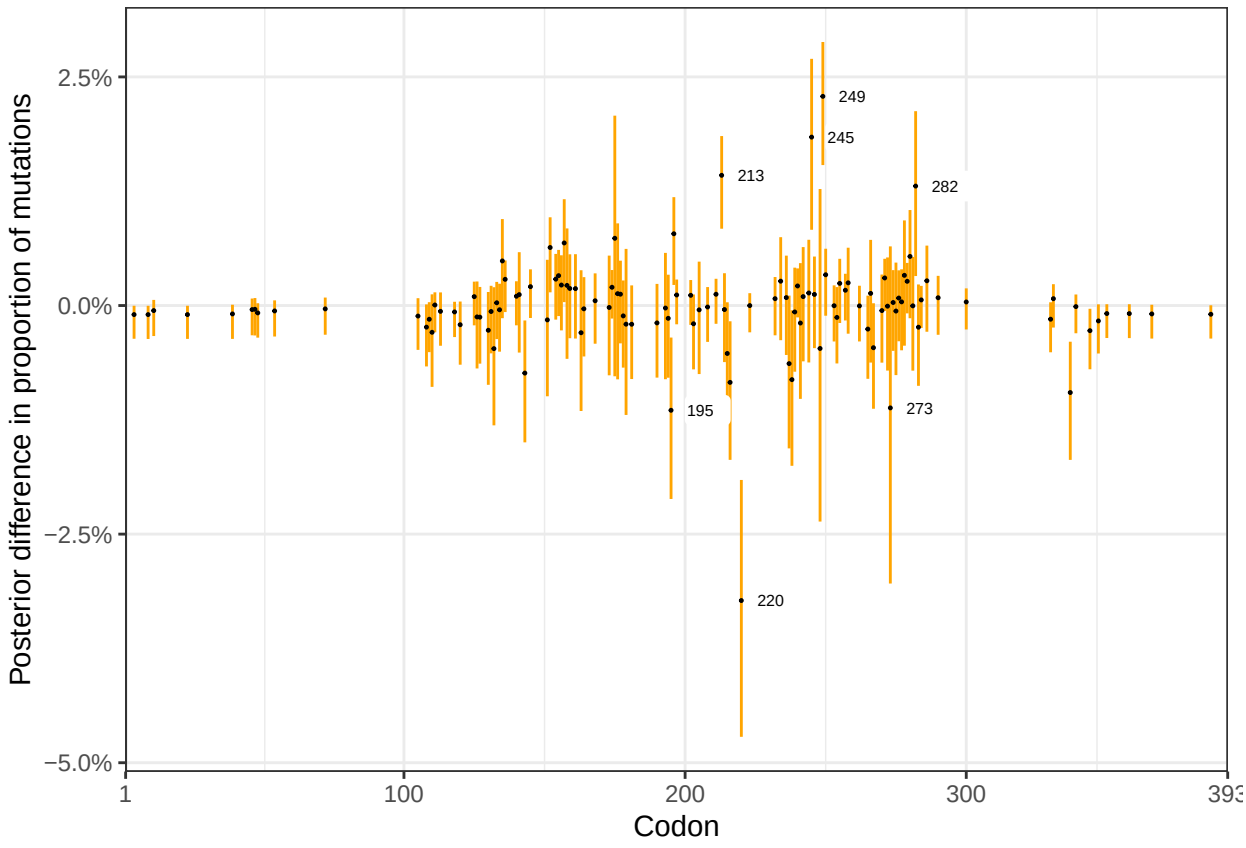

```
ggsave("figures/proportions_blood_ISB.pdf", width=7, height=4)
```

Sample posterior differences in mutation proportions

```
diff_df <- c()
combos <- list(c("isb", "blood"))
for (combo in combos) {
  # Sample the posterior proportion differences between diseases
  diff <- posts[[combo[1]]] - posts[[combo[2]]]
```

```
  # Collect statistics of the differences
  diff_df <- rbind(diff_df,
                   data.frame(combo=paste(combo[1], combo[2], sep=" - "),
                              site=df$site,
                              mean=apply(diff, 2, mean),
                              q025=apply(diff, 2, quantile, probs=0.025),
                              q975=apply(diff, 2, quantile, probs=0.975)))
}
diff_df <- as_tibble(diff_df)
write_csv(diff_df, "results/differences_blood_ISB.csv")
```

```
ggplot(diff_df, aes(x=site)) +
  geom_segment(aes(xend=site, y=q025, yend=q975), color="orange") +
  geom_point(aes(y=mean), size=0.2) +
  geom_label(data = . %>% filter(abs(mean)>.01), aes(y=mean, label=site), hjust=-0.3, size=2.2, label.s
  scale_x_continuous(breaks=c(1,100,200,300,393), limits=c(1,393), expand=c(0,0)) +
  scale_y_continuous(labels=percent_format()) +
  xlab("Codon") +
  ylab("Posterior difference in proportion of mutations") +
  theme_bw() +
  theme(strip.placement="outside", strip.background=element_blank(),
        panel.grid.minor.y=element_blank())
```




```
ggsave("figures/differences_blood_ISB.pdf", width=7, height=4)
```

```r
# The number of positions whose 95% central credible interval excludes zero
with(diff_df, sum(0<q025 | 0>q975))
```

```
## [1] 17
```