Selectome_v_Petrov_branch_lengths

sadie

9/20/2019

```r
library(readr)
library(dplyr)
library(ggplot2)
library(stringr)
library(plotly)
library(tidyr)
```

```r
selectome.dat <- read_csv("~/Downloads/SELECTOME_SRV_nr_v02.csv")

petrov.dat <- read_csv("~/Downloads/PETROV_SRV_nr_v02.csv")

schultz.dat <- read_csv("~/Downloads/IMMUNE_SRV_nr_v02.csv")
```

put branch length estimates for each model in own column for each branch

Lengthx1 - the branch length according to Standard MG94

Lengthx2 - the branch length according to the double hit MG94

Lengthx3 - the branch length according to the triple hit MG94

```r
attach(selectome.dat)
temp.2 <- selectome.dat %>% separate_rows(., `Branch Attributes - MG94 with double and triple instantane

temp.2 <- temp.2 %>% separate(`Branch Attributes - MG94 with double and triple instantaneous substituti

temp.2 <- temp.2 %>% separate(`Branch Attributes - MG94 with double instantaneous substitutions`, c("Bra

temp.2 <- temp.2 %>% separate(`Branch Attributes - Standard MG94`, c("Branchx1", "Lengthx1"), sep = ":"


temp.2$Branch <- temp.2$Branchx3 %>% str_remove_all("'") %>% str_remove("\\{") %>% str_remove("\\}") %>%


temp.2$Lengthx1 %>% str_remove_all("'") %>% str_remove("\\{") %>% str_remove("\\}") %>% str_trim() -> te

temp.2$Lengthx2 %>% str_remove_all("'") %>% str_remove("\\{") %>% str_remove("\\}") %>% str_trim() -> te

temp.2$Lengthx3 %>% str_remove_all("'") %>% str_remove("\\{") %>% str_remove("\\}") %>% str_trim() -> te

sel.dat.bl <- temp.2
```

put branch length estimates for each model in own column for each branch

Lengthx1 - the branch length according to Standard MG94

Lengthx2 - the branch length according to the double hit MG94

Lengthx3 - the branch length according to the triple hit MG94

```r
attach(petrov.dat)
```

```
## The following objects are masked from selectome.dat:
##
##     Branch Attributes - MG94 with double and triple instantaneous
##     substitutions, Branch Attributes - MG94 with double
##     instantaneous substitutions, Branch Attributes - Standard
##     MG94, distribution.double, distribution.single,
##     distribution.triple, distribution.triple_1, Double-hit vs
##     single-hit - LRT, Double-hit vs single-hit - p-value, File
##     name, GDD rate category 1.double, GDD rate category 1.single,
##     GDD rate category 1.triple, GDD rate category 1.triple_1, GDD
##     rate category 2.double, GDD rate category 2.single, GDD rate
##     category 2.triple, GDD rate category 2.triple_1, GDD rate
##     category 3.double, GDD rate category 3.single, GDD rate
##     category 3.triple, GDD rate category 3.triple_1, MG94 with
##     double and triple instantaneous substitutions - AIC-c, MG94
##     with double and triple instantaneous substitutions - Log
##     Likelihood, MG94 with double and triple instantaneous
##     substitutions [ISLANDS] - Log Likelihood, MG94 with double and
##     triple instantaneous substitutions [ISLANDS]- AIC-c, MG94 with
##     double instantaneous substitutions - AIC-c, MG94 with double
##     instantaneous substitutions - Log Likelihood, Mixture
##     auxiliary weight for GDD category 1.double, Mixture auxiliary
##     weight for GDD category 1.single, Mixture auxiliary weight for
##     GDD category 1.triple, Mixture auxiliary weight for GDD
##     category 1.triple_1, Mixture auxiliary weight for GDD category
##     2.double, Mixture auxiliary weight for GDD category 2.single,
##     Mixture auxiliary weight for GDD category 2.triple, Mixture
##     auxiliary weight for GDD category 2.triple_1,
##     non-synonymous/synonymous rate ratio,
##     non-synonymous/synonymous rate ratio_1,
##     non-synonymous/synonymous rate ratio_2,
##     non-synonymous/synonymous rate ratio_3, number of sequences,
##     number of sites, rate at which 2 nucleotides are changed
##     instantly within a single codon, rate at which 2 nucleotides
##     are changed instantly within a single codon_1, rate at which 2
##     nucleotides are changed instantly within a single codon_2,
##     rate at which 3 nucleotides are changed instantly within a
##     single codon, rate at which 3 nucleotides are changed
##     instantly within a single codon between synonymous codon
##     islands, rate at which 3 nucleotides are changed instantly
##     within a single codon between synonymous codon islands_1, rate
##     at which 3 nucleotides are changed instantly within a single
##     codon_1, Standard MG94 - AIC-c, Standard MG94 - Log
##     Likelihood, Substitution rate from nucleotide A to nucleotide
##     C, Substitution rate from nucleotide A to nucleotide C_1,
##     Substitution rate from nucleotide A to nucleotide C_2,
##     Substitution rate from nucleotide A to nucleotide C_3,
##     Substitution rate from nucleotide A to nucleotide G,
##     Substitution rate from nucleotide A to nucleotide G_1,
##     Substitution rate from nucleotide A to nucleotide G_2,
##     Substitution rate from nucleotide A to nucleotide G_3,
##     Substitution rate from nucleotide A to nucleotide T,
##     Substitution rate from nucleotide A to nucleotide T_1,
##     Substitution rate from nucleotide A to nucleotide T_2,
##     Substitution rate from nucleotide A to nucleotide T_3,
##     Substitution rate from nucleotide C to nucleotide G,
##     Substitution rate from nucleotide C to nucleotide G_1,
##     Substitution rate from nucleotide C to nucleotide G_2,
##     Substitution rate from nucleotide C to nucleotide G_3,
##     Substitution rate from nucleotide C to nucleotide T,
##     Substitution rate from nucleotide C to nucleotide T_1,
##     Substitution rate from nucleotide C to nucleotide T_2,
##     Substitution rate from nucleotide C to nucleotide T_3,
##     Substitution rate from nucleotide G to nucleotide T,
##     Substitution rate from nucleotide G to nucleotide T_1,
##     Substitution rate from nucleotide G to nucleotide T_2,
##     Substitution rate from nucleotide G to nucleotide T_3, Tree
##     Length - MG94 with double and triple instantaneous
##     substitutions, Tree Length - MG94 with double instantaneous
##     substitutions, Tree Length - Standard MG94, Triple-hit vs
##     double-hit - LRT, Triple-hit vs double-hit - p-value,
##     Triple-hit vs single-hit - LRT, Triple-hit vs single-hit -
##     p-value, Triple-hit vs Triple-hit-island - LRT, Triple-hit vs
##     Triple-hit-island - p-value, Triple-hit-island vs double-hit -
##     LRT, Triple-hit-island vs double-hit - p-value
```

```
temp.2 <- petrov.dat %>% separate_rows(., `Branch Attributes - MG94 with double and triple instantaneou

temp.2 <- temp.2 %>% separate(`Branch Attributes - MG94 with double and triple instantaneous substituti

temp.2 <- temp.2 %>% separate(`Branch Attributes - MG94 with double instantaneous substitutions`, c("Br

temp.2 <- temp.2 %>% separate(`Branch Attributes - Standard MG94`, c("Branchx1", "Lengthx1"), sep = ":"


temp.2$Branch <- temp.2$Branchx3 %>% str_remove_all("'") %>% str_remove("\\{") %>% str_remove("\\}") %>


temp.2$Lengthx1 %>% str_remove_all("'") %>% str_remove("\\{") %>% str_remove("\\}") %>% str_trim() -> t

temp.2$Lengthx2 %>% str_remove_all("'") %>% str_remove("\\{") %>% str_remove("\\}") %>% str_trim() -> t

temp.2$Lengthx3 %>% str_remove_all("'") %>% str_remove("\\{") %>% str_remove("\\}") %>% str_trim() -> t

pet.dat.bl <- temp.2
```

put branch length estimates for each model in own column for each branch for IMMUNE/Schultz

Lengthx1 - the branch length according to Standard MG94

Lengthx2 - the branch length according to the double hit MG94

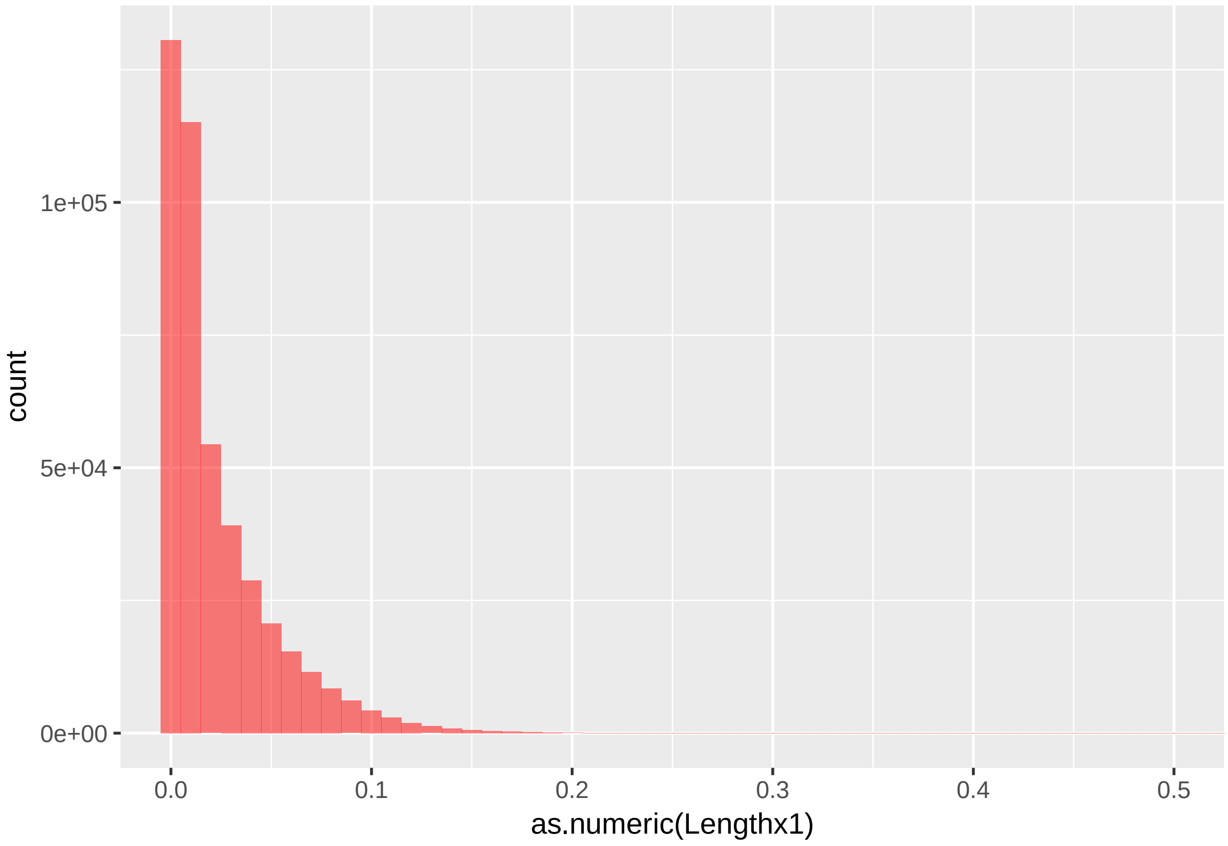
Lengthx3 - the branch length according to the triple hit MG94

```r
attach(schultz.dat)
```

```
## The following objects are masked from petrov.dat:
##
##     Branch Attributes - MG94 with double and triple instantaneous
##     substitutions, Branch Attributes - MG94 with double
##     instantaneous substitutions, Branch Attributes - Standard
##     MG94, distribution.double, distribution.single,
##     distribution.triple, distribution.triple_1, Double-hit vs
##     single-hit - LRT, Double-hit vs single-hit - p-value, File
##     name, GDD rate category 1.double, GDD rate category 1.single,
##     GDD rate category 1.triple, GDD rate category 1.triple_1, GDD
##     rate category 2.double, GDD rate category 2.single, GDD rate
##     category 2.triple, GDD rate category 2.triple_1, GDD rate
##     category 3.double, GDD rate category 3.single, GDD rate
##     category 3.triple, GDD rate category 3.triple_1, MG94 with
##     double and triple instantaneous substitutions - AIC-c, MG94
##     with double and triple instantaneous substitutions - Log
##     Likelihood, MG94 with double and triple instantaneous
##     substitutions [ISLANDS] - Log Likelihood, MG94 with double and
##     triple instantaneous substitutions [ISLANDS]- AIC-c, MG94 with
##     double instantaneous substitutions - AIC-c, MG94 with double
##     instantaneous substitutions - Log Likelihood, Mixture
##     auxiliary weight for GDD category 1.double, Mixture auxiliary
##     weight for GDD category 1.single, Mixture auxiliary weight for
##     GDD category 1.triple, Mixture auxiliary weight for GDD
##     category 1.triple_1, Mixture auxiliary weight for GDD category
##     2.double, Mixture auxiliary weight for GDD category 2.single,
##     Mixture auxiliary weight for GDD category 2.triple, Mixture
##     auxiliary weight for GDD category 2.triple_1,
##     non-synonymous/synonymous rate ratio,
##     non-synonymous/synonymous rate ratio_1,
##     non-synonymous/synonymous rate ratio_2,
##     non-synonymous/synonymous rate ratio_3, number of sequences,
##     number of sites, rate at which 2 nucleotides are changed
##     instantly within a single codon, rate at which 2 nucleotides
##     are changed instantly within a single codon_1, rate at which 2
##     nucleotides are changed instantly within a single codon_2,
##     rate at which 3 nucleotides are changed instantly within a
##     single codon, rate at which 3 nucleotides are changed
##     instantly within a single codon between synonymous codon
##     islands, rate at which 3 nucleotides are changed instantly
##     within a single codon between synonymous codon islands_1, rate
##     at which 3 nucleotides are changed instantly within a single
##     codon_1, Standard MG94 - AIC-c, Standard MG94 - Log
##     Likelihood, Substitution rate from nucleotide A to nucleotide
##     C, Substitution rate from nucleotide A to nucleotide C_1,
##     Substitution rate from nucleotide A to nucleotide C_2,
##     Substitution rate from nucleotide A to nucleotide C_3,
##     Substitution rate from nucleotide A to nucleotide G,
##     Substitution rate from nucleotide A to nucleotide G_1,
##     Substitution rate from nucleotide A to nucleotide G_2,
##     Substitution rate from nucleotide A to nucleotide G_3,
##     Substitution rate from nucleotide A to nucleotide T,
##     Substitution rate from nucleotide A to nucleotide T_1,
##     Substitution rate from nucleotide A to nucleotide T_2,
##     Substitution rate from nucleotide A to nucleotide T_3,
##     Substitution rate from nucleotide C to nucleotide G,
##     Substitution rate from nucleotide C to nucleotide G_1,
##     Substitution rate from nucleotide C to nucleotide G_2,
##     Substitution rate from nucleotide C to nucleotide G_3,
##     Substitution rate from nucleotide C to nucleotide T,
##     Substitution rate from nucleotide C to nucleotide T_1,
##     Substitution rate from nucleotide C to nucleotide T_2,
##     Substitution rate from nucleotide C to nucleotide T_3,
##     Substitution rate from nucleotide G to nucleotide T,
##     Substitution rate from nucleotide G to nucleotide T_1,
##     Substitution rate from nucleotide G to nucleotide T_2,
##     Substitution rate from nucleotide G to nucleotide T_3, Tree
##     Length - MG94 with double and triple instantaneous
##     substitutions, Tree Length - MG94 with double instantaneous
##     substitutions, Tree Length - Standard MG94, Triple-hit vs
##     double-hit - LRT, Triple-hit vs double-hit - p-value,
##     Triple-hit vs single-hit - LRT, Triple-hit vs single-hit -
##     p-value, Triple-hit vs Triple-hit-island - LRT, Triple-hit vs
##     Triple-hit-island - p-value, Triple-hit-island vs double-hit -
##     LRT, Triple-hit-island vs double-hit - p-value

## The following objects are masked from selectome.dat:
## 
##     Branch Attributes - MG94 with double and triple instantaneous
##     substitutions, Branch Attributes - MG94 with double
##     instantaneous substitutions, Branch Attributes - Standard
##     MG94, distribution.double, distribution.single,
##     distribution.triple, distribution.triple_1, Double-hit vs
##     single-hit - LRT, Double-hit vs single-hit - p-value, File
##     name, GDD rate category 1.double, GDD rate category 1.single,
##     GDD rate category 1.triple, GDD rate category 1.triple_1, GDD
##     rate category 2.double, GDD rate category 2.single, GDD rate
##     category 2.triple, GDD rate category 2.triple_1, GDD rate
##     category 3.double, GDD rate category 3.single, GDD rate
##     category 3.triple, GDD rate category 3.triple_1, MG94 with
##     double and triple instantaneous substitutions - AIC-c, MG94
##     with double and triple instantaneous substitutions - Log
##     Likelihood, MG94 with double and triple instantaneous
##     substitutions [ISLANDS] - Log Likelihood, MG94 with double and
##     triple instantaneous substitutions [ISLANDS]- AIC-c, MG94 with
##     double instantaneous substitutions - AIC-c, MG94 with double
##     instantaneous substitutions - Log Likelihood, Mixture
##     auxiliary weight for GDD category 1.double, Mixture auxiliary
##     weight for GDD category 1.single, Mixture auxiliary weight for
##     GDD category 1.triple, Mixture auxiliary weight for GDD
##     category 1.triple_1, Mixture auxiliary weight for GDD category
##     2.double, Mixture auxiliary weight for GDD category 2.single,
##     Mixture auxiliary weight for GDD category 2.triple, Mixture
##     auxiliary weight for GDD category 2.triple_1,
##     non-synonymous/synonymous rate ratio,
##     non-synonymous/synonymous rate ratio_1,
##     non-synonymous/synonymous rate ratio_2,
##     non-synonymous/synonymous rate ratio_3, number of sequences,
##     number of sites, rate at which 2 nucleotides are changed
##     instantly within a single codon, rate at which 2 nucleotides
##     are changed instantly within a single codon_1, rate at which 2
##     nucleotides are changed instantly within a single codon_2,
##     rate at which 3 nucleotides are changed instantly within a
##     single codon, rate at which 3 nucleotides are changed
##     instantly within a single codon between synonymous codon
##     islands, rate at which 3 nucleotides are changed instantly
##     within a single codon between synonymous codon islands_1, rate
##     at which 3 nucleotides are changed instantly within a single
##     codon_1, Standard MG94 - AIC-c, Standard MG94 - Log
##     Likelihood, Substitution rate from nucleotide A to nucleotide
##     C, Substitution rate from nucleotide A to nucleotide C_1,
##     Substitution rate from nucleotide A to nucleotide C_2,
##     Substitution rate from nucleotide A to nucleotide C_3,
##     Substitution rate from nucleotide A to nucleotide G,
##     Substitution rate from nucleotide A to nucleotide G_1,
##     Substitution rate from nucleotide A to nucleotide G_2,
##     Substitution rate from nucleotide A to nucleotide G_3,
##     Substitution rate from nucleotide A to nucleotide T,
##     Substitution rate from nucleotide A to nucleotide T_1,
##     Substitution rate from nucleotide A to nucleotide T_2,
##     Substitution rate from nucleotide A to nucleotide T_3,
##     Substitution rate from nucleotide C to nucleotide G,
##     Substitution rate from nucleotide C to nucleotide G_1,
##     Substitution rate from nucleotide C to nucleotide G_2,
##     Substitution rate from nucleotide C to nucleotide G_3,
##     Substitution rate from nucleotide C to nucleotide T,
##     Substitution rate from nucleotide C to nucleotide T_1,
##     Substitution rate from nucleotide C to nucleotide T_2,
##     Substitution rate from nucleotide C to nucleotide T_3,
##     Substitution rate from nucleotide G to nucleotide T,
##     Substitution rate from nucleotide G to nucleotide T_1,
##     Substitution rate from nucleotide G to nucleotide T_2,
##     Substitution rate from nucleotide G to nucleotide T_3, Tree
##     Length - MG94 with double and triple instantaneous
##     substitutions, Tree Length - MG94 with double instantaneous
##     substitutions, Tree Length - Standard MG94, Triple-hit vs
##     double-hit - LRT, Triple-hit vs double-hit - p-value,
##     Triple-hit vs single-hit - LRT, Triple-hit vs single-hit -
##     p-value, Triple-hit vs Triple-hit-island - LRT, Triple-hit vs
##     Triple-hit-island - p-value, Triple-hit-island vs double-hit -
##     LRT, Triple-hit-island vs double-hit - p-value
```

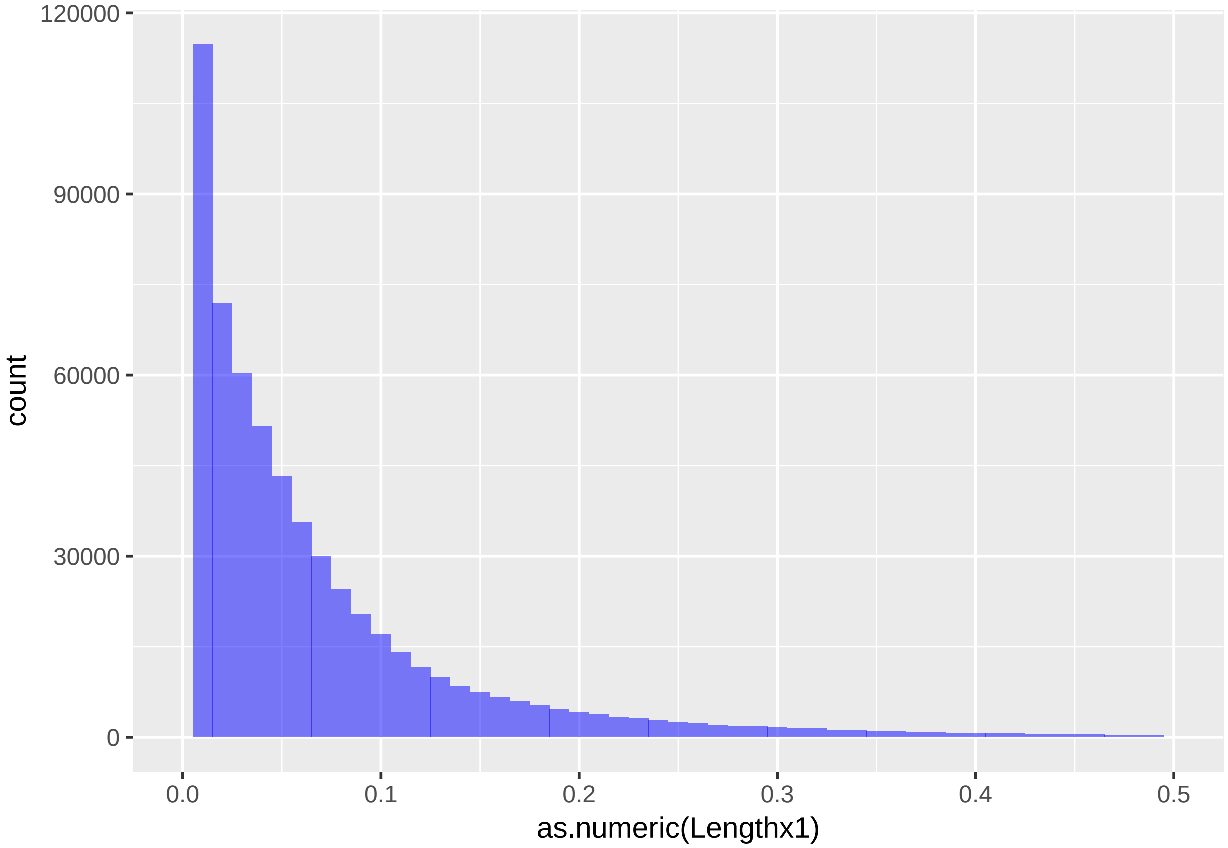
```
temp.2 <- schultz.dat %>% separate_rows(., `Branch Attributes - MG94 with double and triple instantaneou

temp.2 <- temp.2 %>% separate(`Branch Attributes - MG94 with double and triple instantaneous substituti

temp.2 <- temp.2 %>% separate(`Branch Attributes - MG94 with double instantaneous substitutions`, c("Bra

temp.2 <- temp.2 %>% separate(`Branch Attributes - Standard MG94`, c("Branchx1", "Lengthx1"), sep = ":"


temp.2$Branch <- temp.2$Branchx3 %>% str_remove_all("'") %>% str_remove("\\{") %>% str_remove("\\}") %>


temp.2$Lengthx1 %>% str_remove_all("'") %>% str_remove("\\{") %>% str_remove("\\}") %>% str_trim() -> te

temp.2$Lengthx2 %>% str_remove_all("'") %>% str_remove("\\{") %>% str_remove("\\}") %>% str_trim() -> te

temp.2$Lengthx3 %>% str_remove_all("'") %>% str_remove("\\{") %>% str_remove("\\}") %>% str_trim() -> te

schultz.dat.bl <- temp.2
```

```
pet.dat.bl %>% ggplot()+ geom_histogram(aes(as.numeric(Lengthx1)), binwidth = 0.01, fill = "red", alpha
```



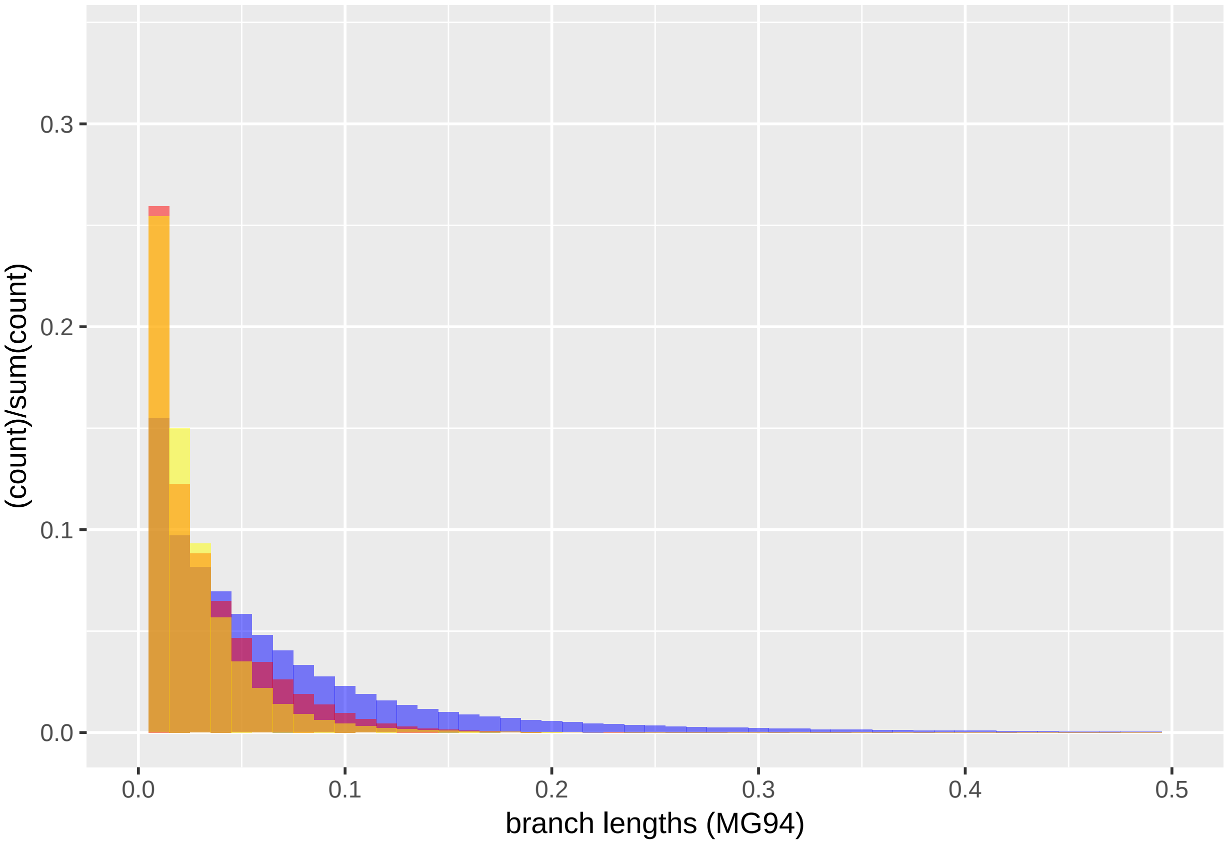
```
sel.dat.bl %>% filter(as.numeric(Lengthx1 )!= 0) %>% ggplot()+ geom_histogram(aes(as.numeric(Lengthx1))
```

```
## Warning: Removed 7802 rows containing non-finite values (stat_bin).

## Warning: Removed 2 rows containing missing values (geom_bar).
```



```
ggplot()+ geom_histogram(data = sel.dat.bl, aes(x =as.numeric(Lengthx1),  y = (..count..)/sum(..count..
```

```
## Warning: Removed 7802 rows containing non-finite values (stat_bin).

## Warning: Removed 13 rows containing non-finite values (stat_bin).

## Warning: Removed 1257 rows containing non-finite values (stat_bin).

## Warning: Removed 2 rows containing missing values (geom_bar).

## Warning: Removed 2 rows containing missing values (geom_bar).

## Warning: Removed 2 rows containing missing values (geom_bar).
```



```
summary(sel.dat.bl$Lengthx1 %>% as.numeric())
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##       0.0       0.0       0.0       5.5       0.1 2242368.4
```

```
summary(pet.dat.bl$Lengthx1 %>% as.numeric())
```

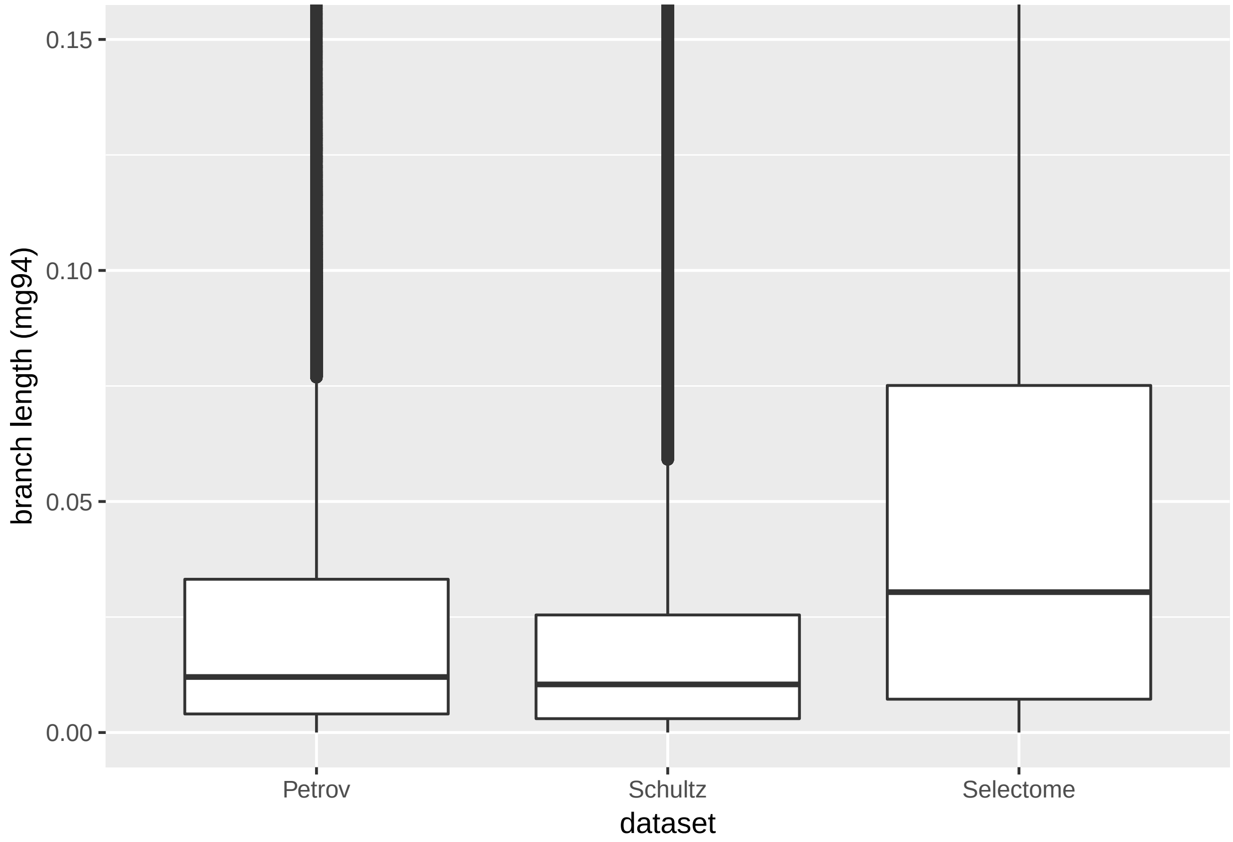
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00000   0.00402   0.01202   0.02367   0.03317 112.15786
```

```
summary(schultz.dat.bl$Lengthx1 %>% as.numeric())
```

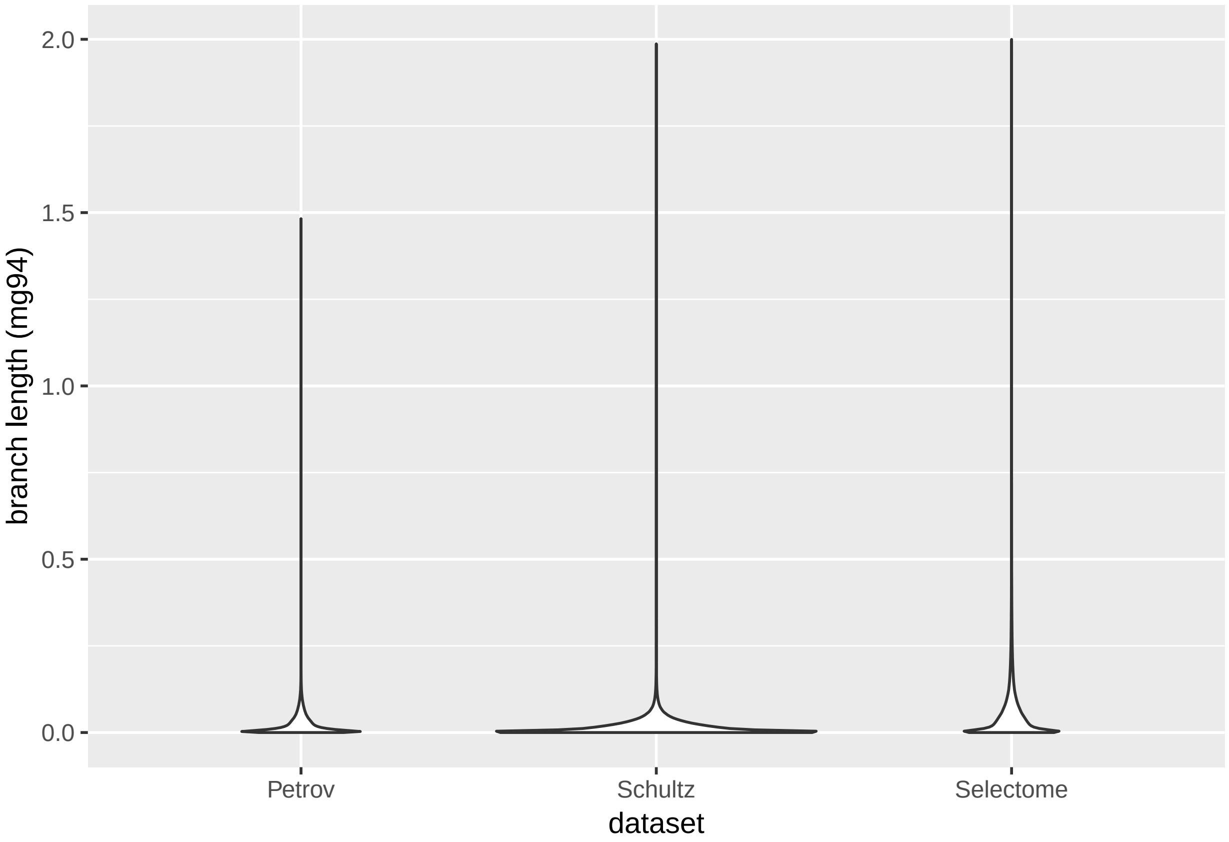
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0        0        0      104        0 80492031
```

```
comb.length <- bind_rows(
  sel.dat.bl %>% select(Lengthx1) %>% mutate(dataset = "Selectome"),
  pet.dat.bl %>% select(Lengthx1) %>% mutate(dataset = "Petrov"),
  schultz.dat.bl %>% select(Lengthx1) %>% mutate(dataset = "Schultz")
  )
```

```
comb.length %>% ggplot()+geom_boxplot(aes(y = as.numeric(Lengthx1), x = dataset)) + coord_cartesian(yli
```



```
comb.length %>% filter(Lengthx1 %>% as.numeric() <=2) %>% ggplot()+geom_violin(aes(y = as.numeric(Lengt
```

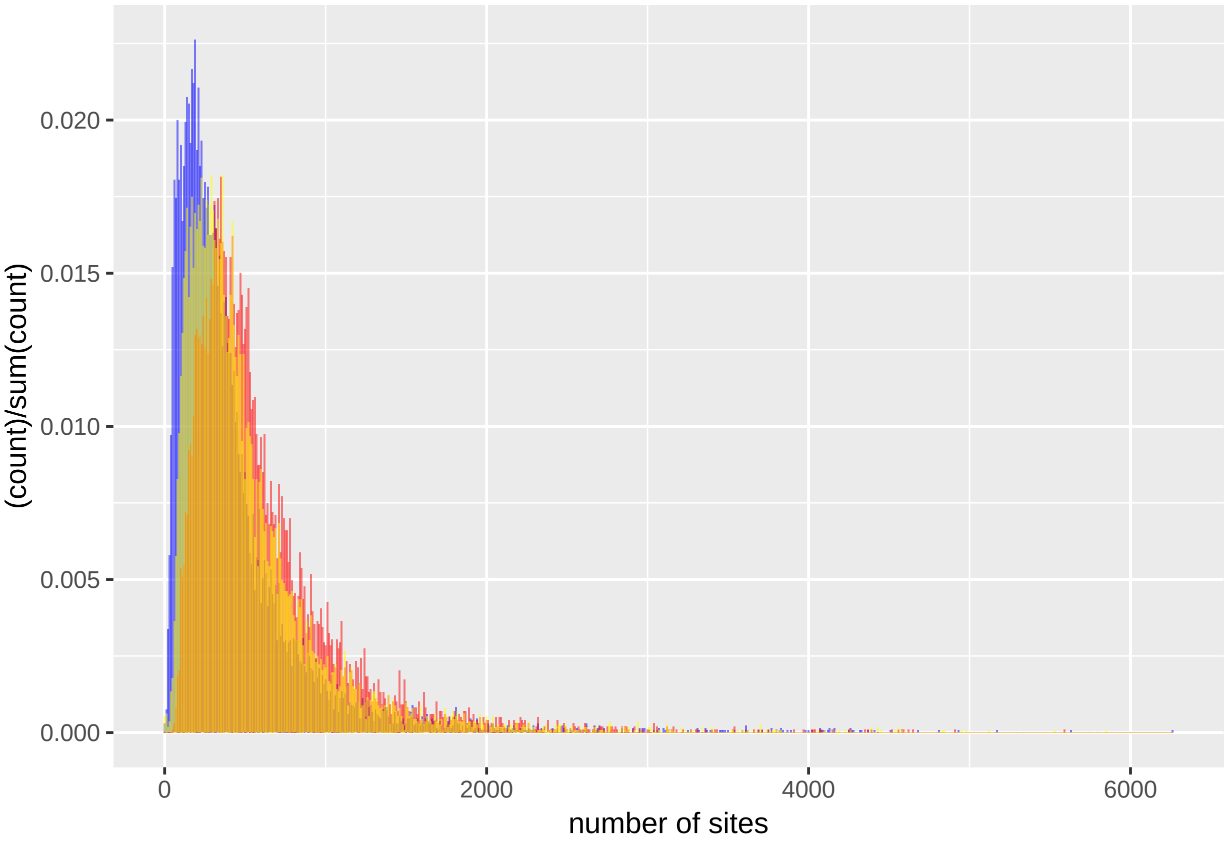


explore other variables

sequence length comparasion

histogram of number of sites

```
ggplot()+ geom_histogram(data = selectome.dat, aes(x = `number of sites`,  y = (..count..)/sum(..count.
```



not clear if there is any difference here. maybe selectome is a little shorter let's look at numbers

```
summary(selectome.dat$`number of sites`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   168.0   302.0   422.3   508.0  6265.0
```

```
summary(petrov.dat$`number of sites`)
```

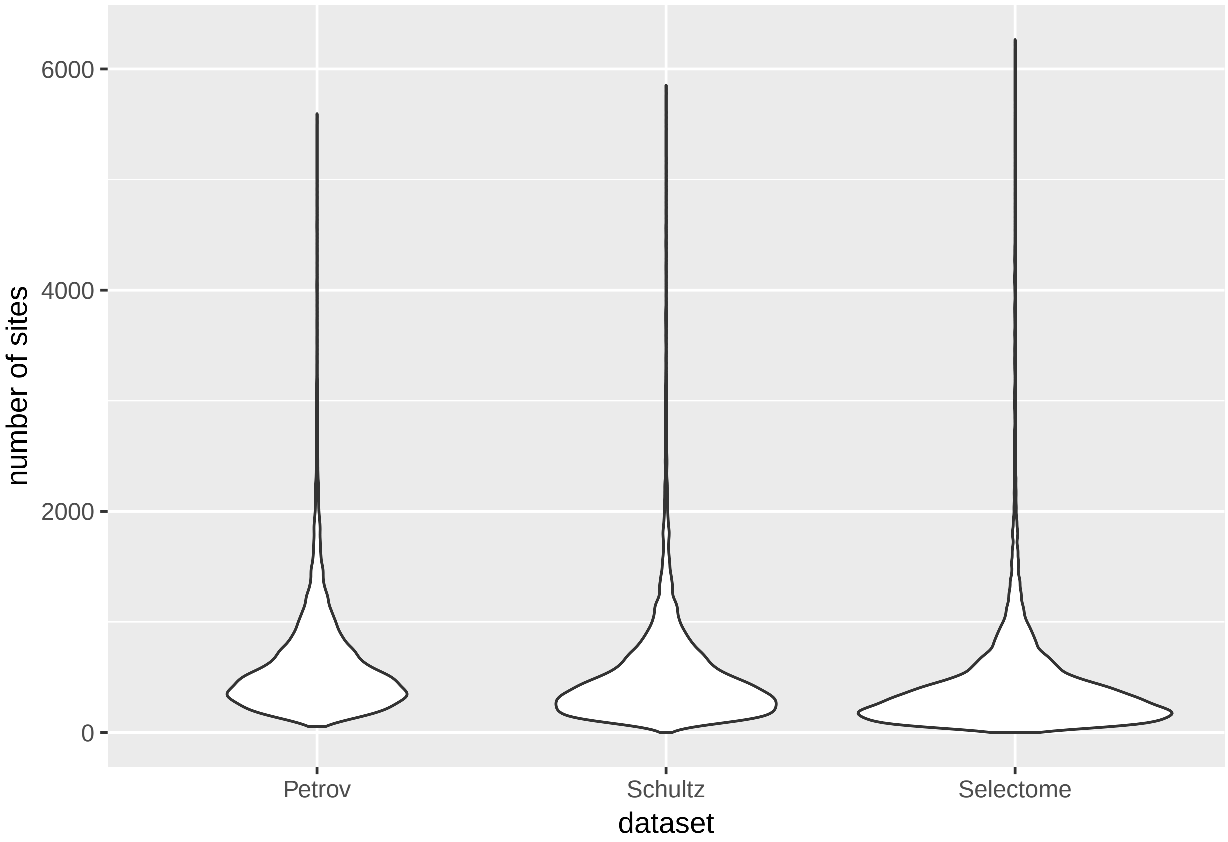
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    55.0   315.0   479.0   611.9   749.0  5595.0
```

```
summary(schultz.dat$`number of sites`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   234.0   388.0   517.1   634.0  5853.0
```

```
comb.length <- bind_rows(
  selectome.dat %>% select(`number of sites`) %>% mutate(dataset = "Selectome"),
  petrov.dat %>% select(`number of sites`) %>% mutate(dataset = "Petrov"),
  schultz.dat %>% select(`number of sites`) %>% mutate(dataset = "Schultz")
  )
```

```
comb.length  %>% ggplot()+geom_violin(aes(y = `number of sites`, x = dataset), scale = "count")  + labs
```


6000
4000
2000
0
number of sites
Petrov
Schultz
Selectome
dataset