logFC error Normalization method comparison

Log fold-change correlations between normalization method

```
prepost_logFC <- readRDS("~/Desktop/norm_logFC_prepost.RDS")
library(GGally)
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
prepost_logFC %>%
    select(norm_method, biosample_id, feature_id, logFC) %>%
    spread(norm_method, logFC) %>%
    ggpairs(aes(color = biosample_id), columns = 3:8) + theme_bw()
```

```
## Warning: Removed 339 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 1 rows containing missing values (geom_text).
```

```
## Warning: Removed 5 rows containing missing values (geom_text).
```

```
## Warning: Removed 1 rows containing missing values (geom_text).
```

```
## Warning: Removed 5 rows containing missing values (geom_text).
```

```
## Warning: Removed 1 rows containing missing values (geom_text).
```

```
## Warning: Removed 5 rows containing missing values (geom_text).
```

```
## Warning: Removed 1 rows containing missing values (geom_text).
```

```
## Warning: Removed 5 rows containing missing values (geom_text).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 498 rows containing missing values
```

```
## Warning: Removed 339 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 1 rows containing missing values (geom_text).
```

```
## Warning: Removed 5 rows containing missing values (geom_text).
```

```
## Warning: Removed 1 rows containing missing values (geom_text).
```

```
## Warning: Removed 5 rows containing missing values (geom_text).
```

```
## Warning: Removed 1 rows containing missing values (geom_text).
```

```
## Warning: Removed 5 rows containing missing values (geom_text).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 498 rows containing missing values
```

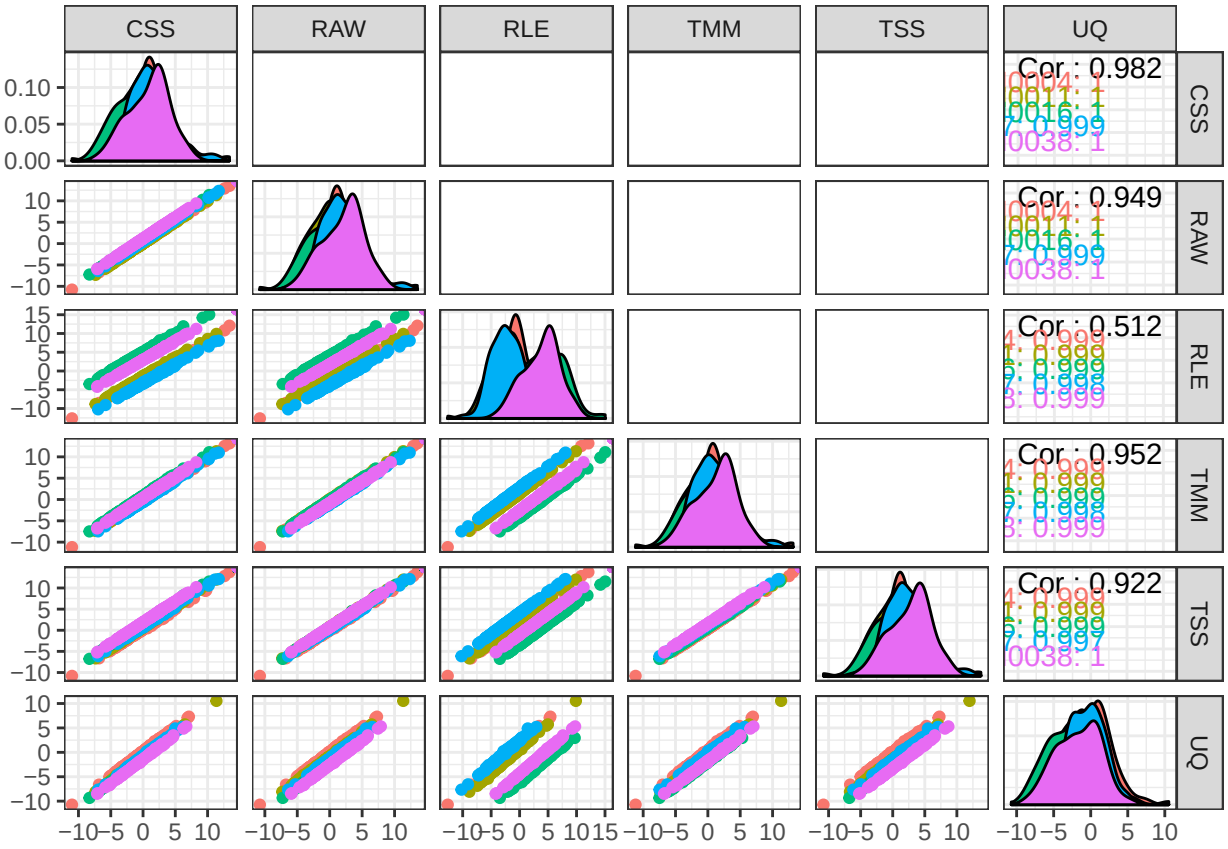
```
## Warning: Removed 339 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 1 rows containing missing values (geom_text).
```

```
## Warning: Removed 5 rows containing missing values (geom_text).
```

```
## Warning: Removed 1 rows containing missing values (geom_text).
```

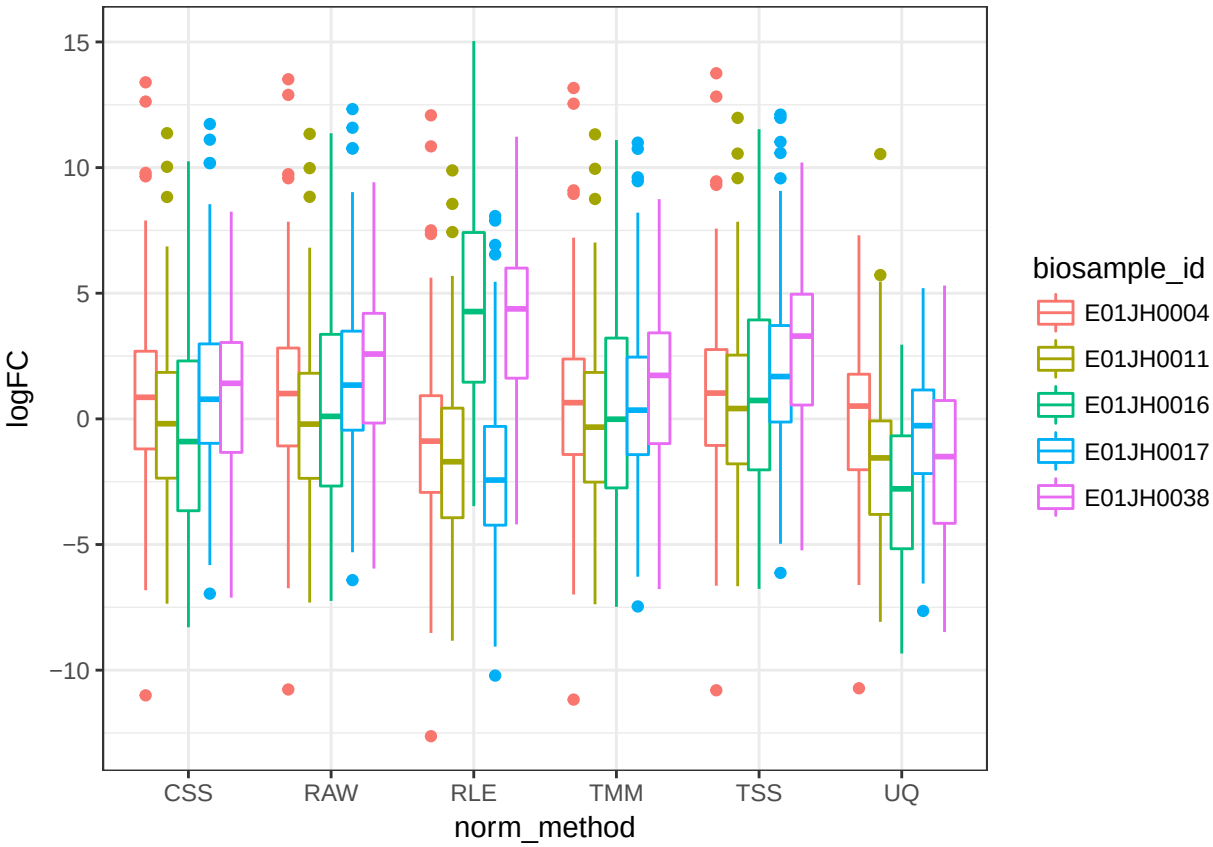
```
## Warning: Removed 5 rows containing missing values (geom_text).
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 498 rows containing missing values
## Warning: Removed 339 rows containing non-finite values (stat_density).
## Warning: Removed 1 rows containing missing values (geom_text).
## Warning: Removed 5 rows containing missing values (geom_text).
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 498 rows containing missing values
## Warning: Removed 339 rows containing non-finite values (stat_density).
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 498 rows containing missing values
## Warning: Removed 498 rows containing missing values (geom_point).
## Warning: Removed 498 rows containing missing values (geom_point).
## Warning: Removed 498 rows containing missing values (geom_point).
## Warning: Removed 498 rows containing missing values (geom_point).
## Warning: Removed 498 rows containing missing values (geom_point).
## Warning: Removed 498 rows containing non-finite values (stat_density).
```



Comparison of log fold-change distributions across normalization methods. UQ has consistently lower log fold-change estimates compared to the other normalization methods. In general the distribution of logFC estimates is consistent across individuals for the different normalization methods excluding RLE.

```
prepost_logFC %>% ggplot() +
    geom_boxplot(aes(x = norm_method, y = logFC, color = biosample_id)) +
    theme_bw()
```

```
## Warning: Removed 2193 rows containing non-finite values (stat_boxplot).
```



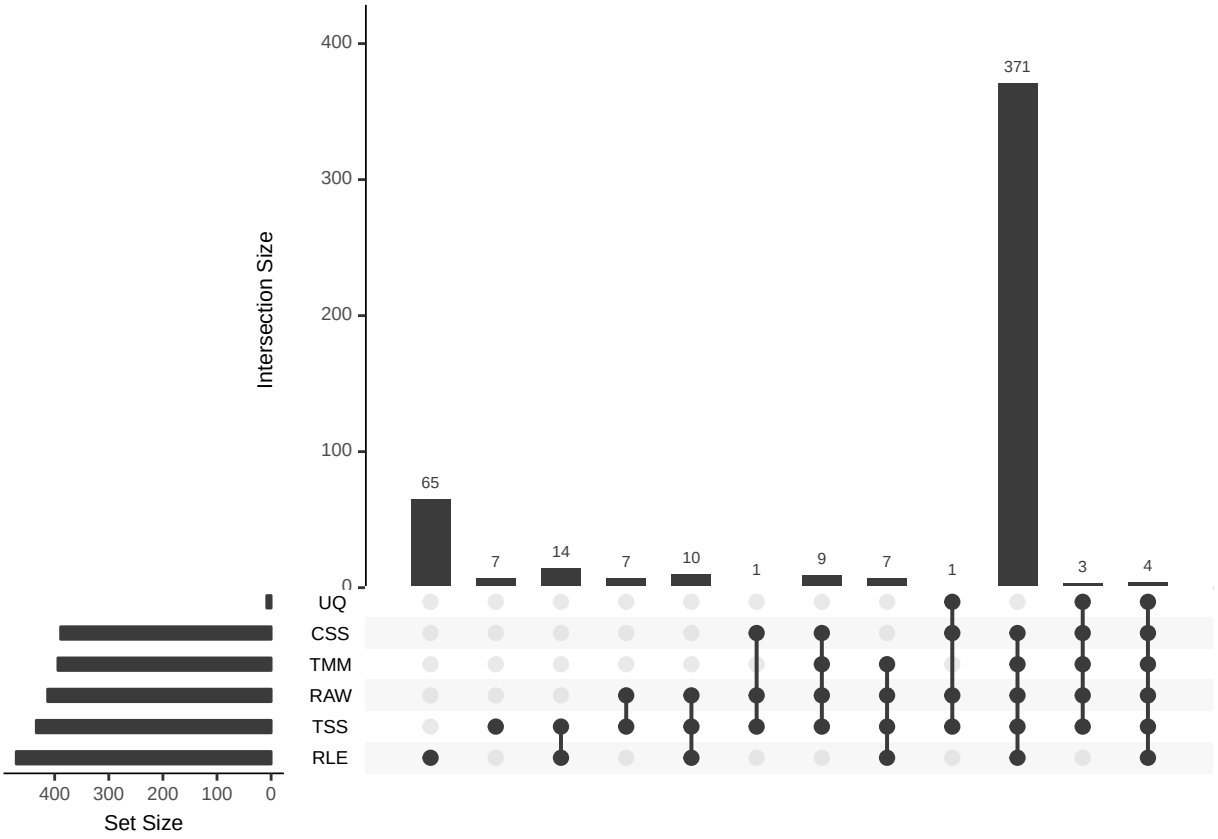
Few features categorized a pre-specific and dominant for UQ normalized data due to the lower distribution of log fold-change estimates compared to the other methods. One option is to choose a common sent of features to compare across normalization methods.

Overlap in pre specific and pre-dominant features between normalization mehtods. Might want to consider using the features that are classified as pre-specific or pre-dominant in 5 or the 6 method.

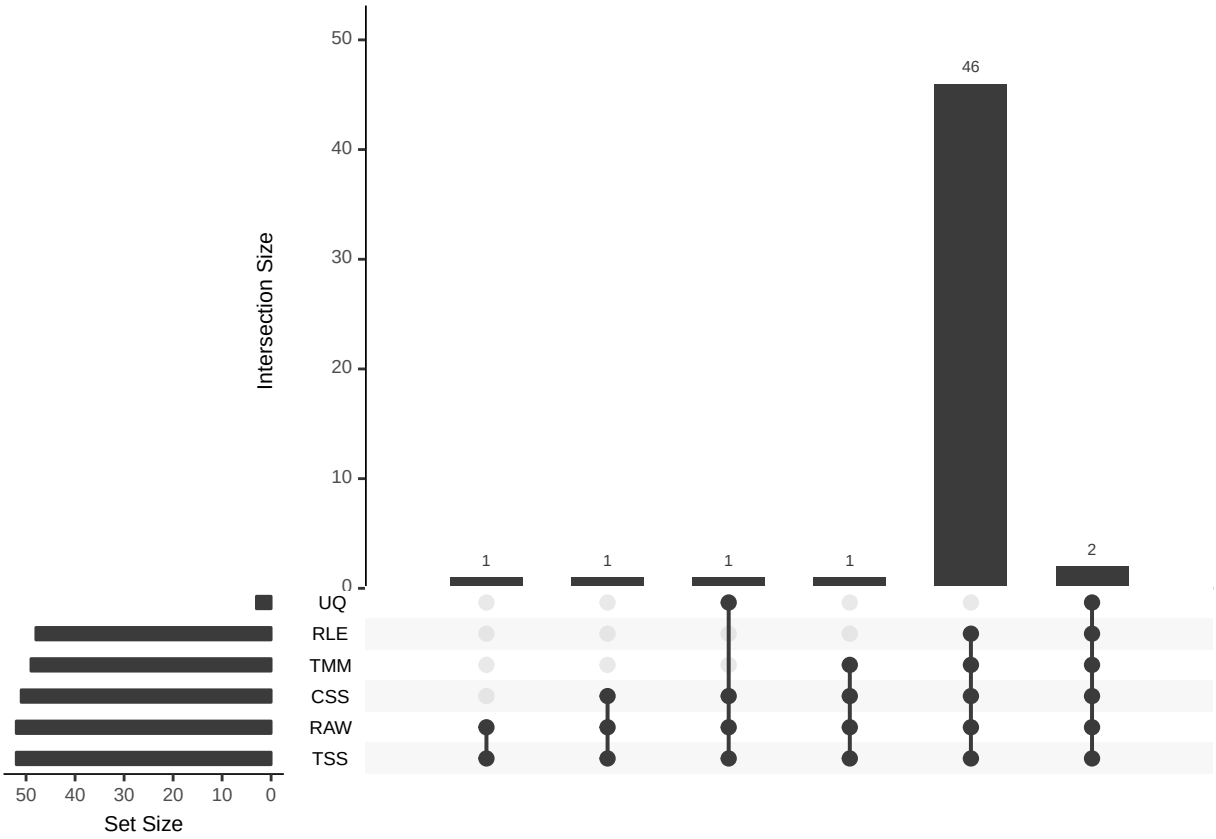
```
library(UpSetR)
logFC_pre_set <- logFC_pre %>% select(biosample_id, norm_method, feature_id) %>% unique() %>%
    add_column(x = 1) %>%
    spread(norm_method, x, fill = 0)
upset(as.data.frame(logFC_pre_set), nset = 7)
```

Table 1: Number of pre-specific and pre-dominant features by individual and normalization method for Mothur

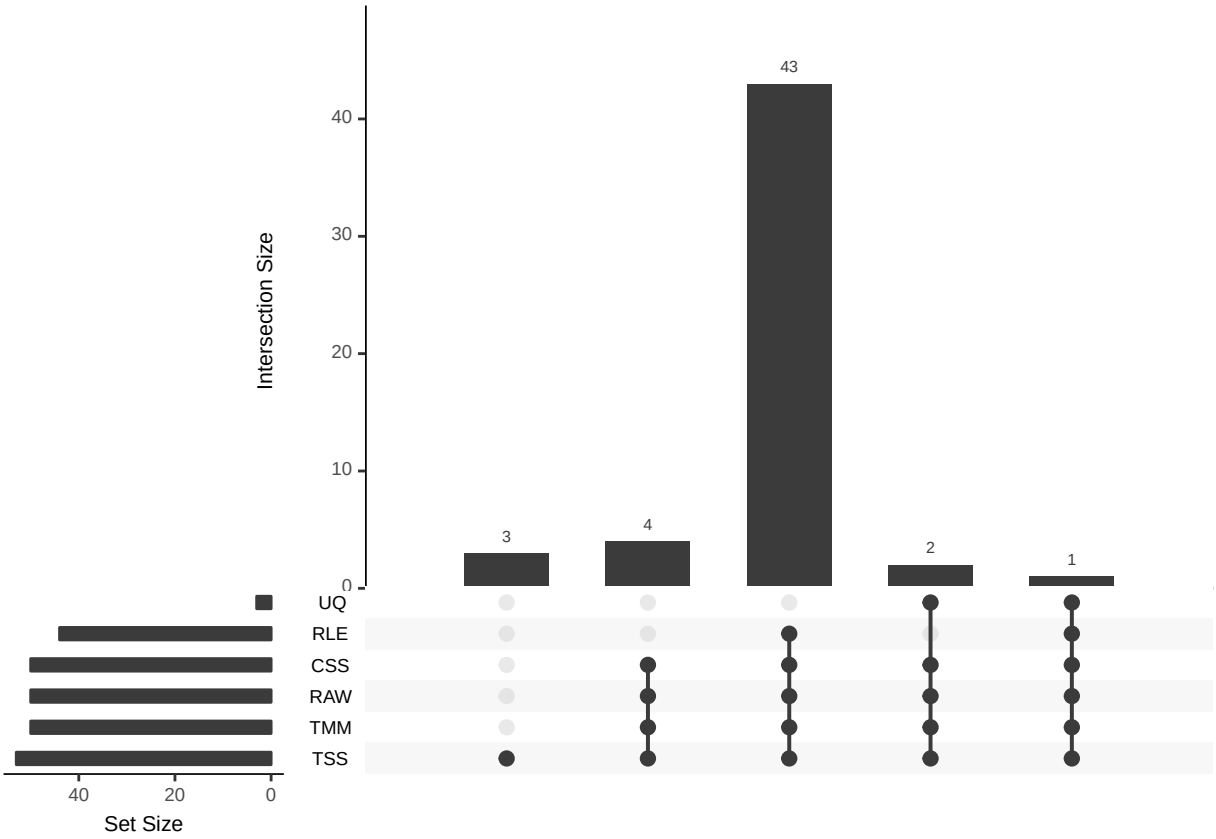
Individual	Type	CSS	RAW	RLE	TMM	TSS	UQ
E01JH0004	dominant	13	14	10	11	14	3
E01JH0004	specific	38	38	38	38	38	0
E01JH0011	dominant	10	10	4	10	13	3
E01JH0011	specific	40	40	40	40	40	0
E01JH0016	dominant	5	10	50	10	12	0
E01JH0016	specific	107	107	107	107	107	0
E01JH0017	dominant	10	16	5	10	20	1
E01JH0017	specific	84	84	84	84	84	0
E01JH0038	dominant	12	24	63	14	36	1
E01JH0038	specific	70	70	70	70	70	0



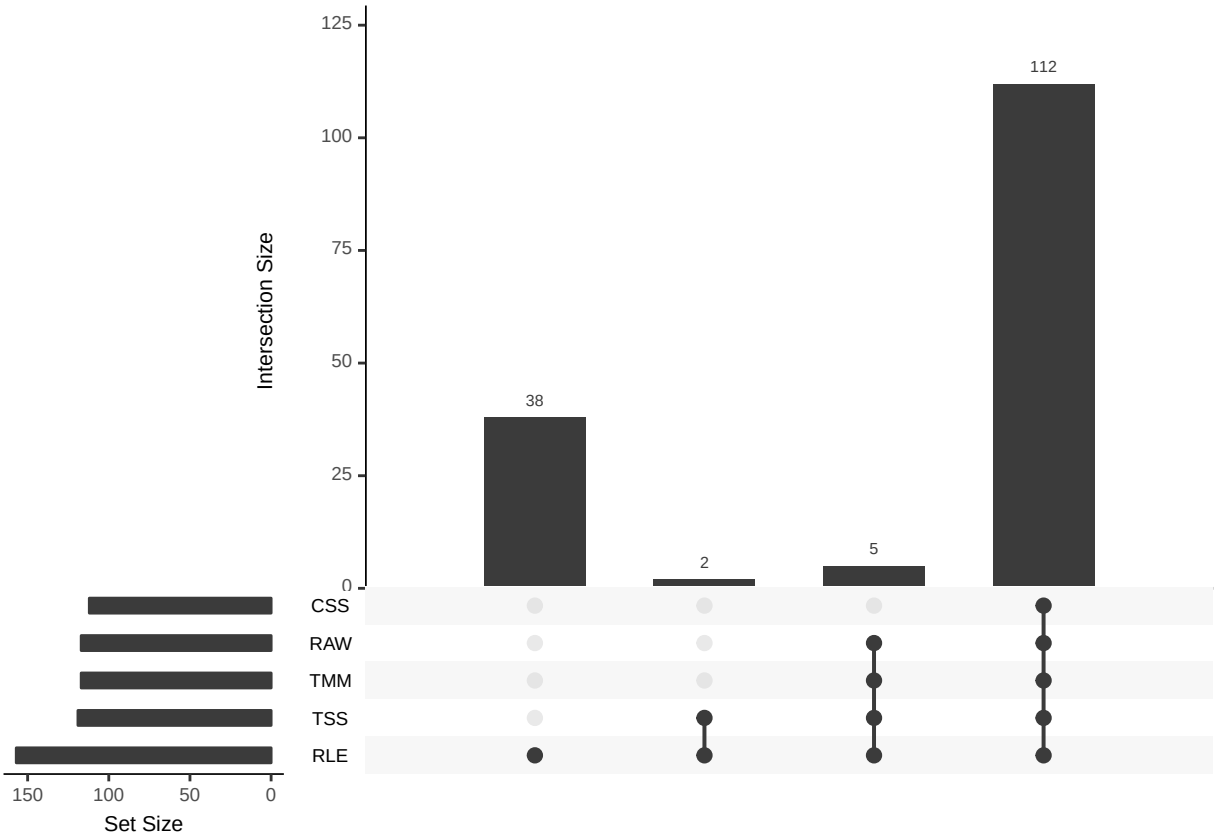

```
logFC_pre_set %>% filter(biosample_id == "E01JH0004") %>% as.data.frame() %>%
                                upset(nset = 6)
```



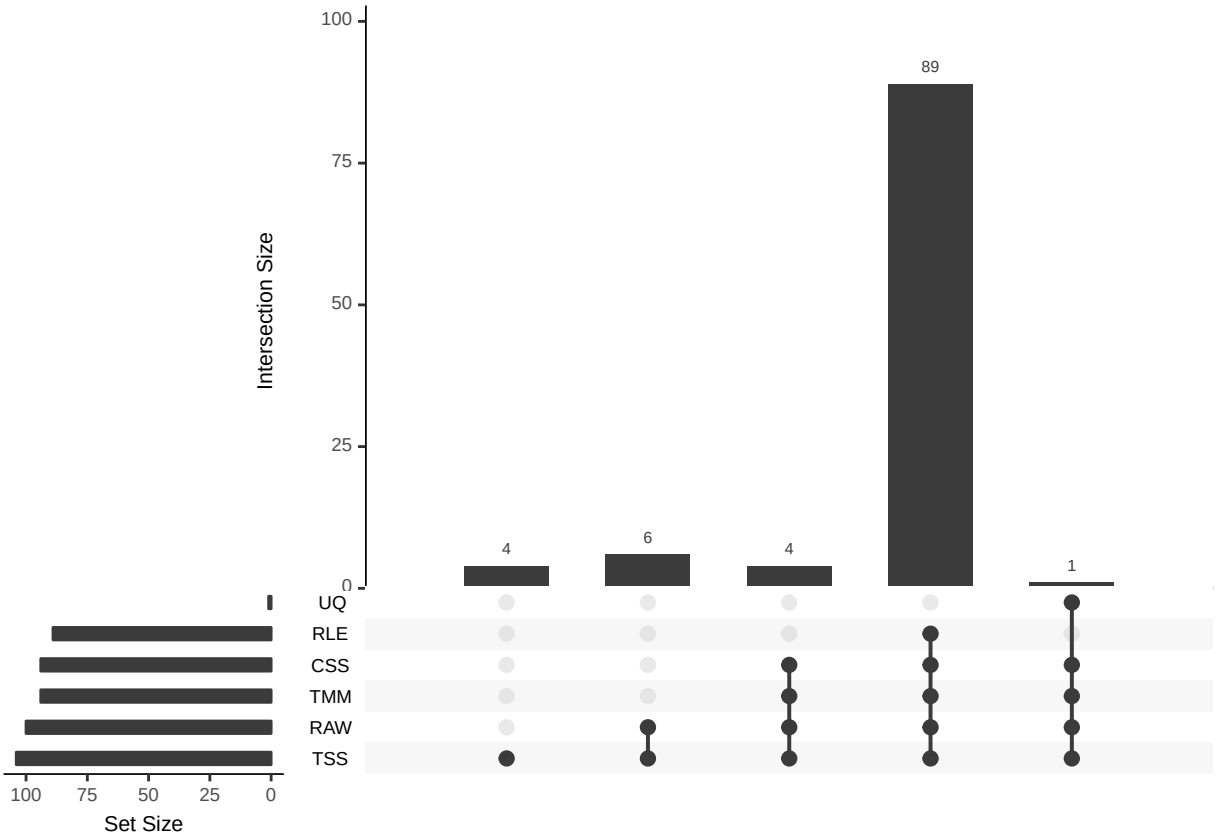

```
logFC_pre_set %>% filter(biosample_id == "E01JH0011") %>% as.data.frame() %>%
                            upset(nset = 6)
```



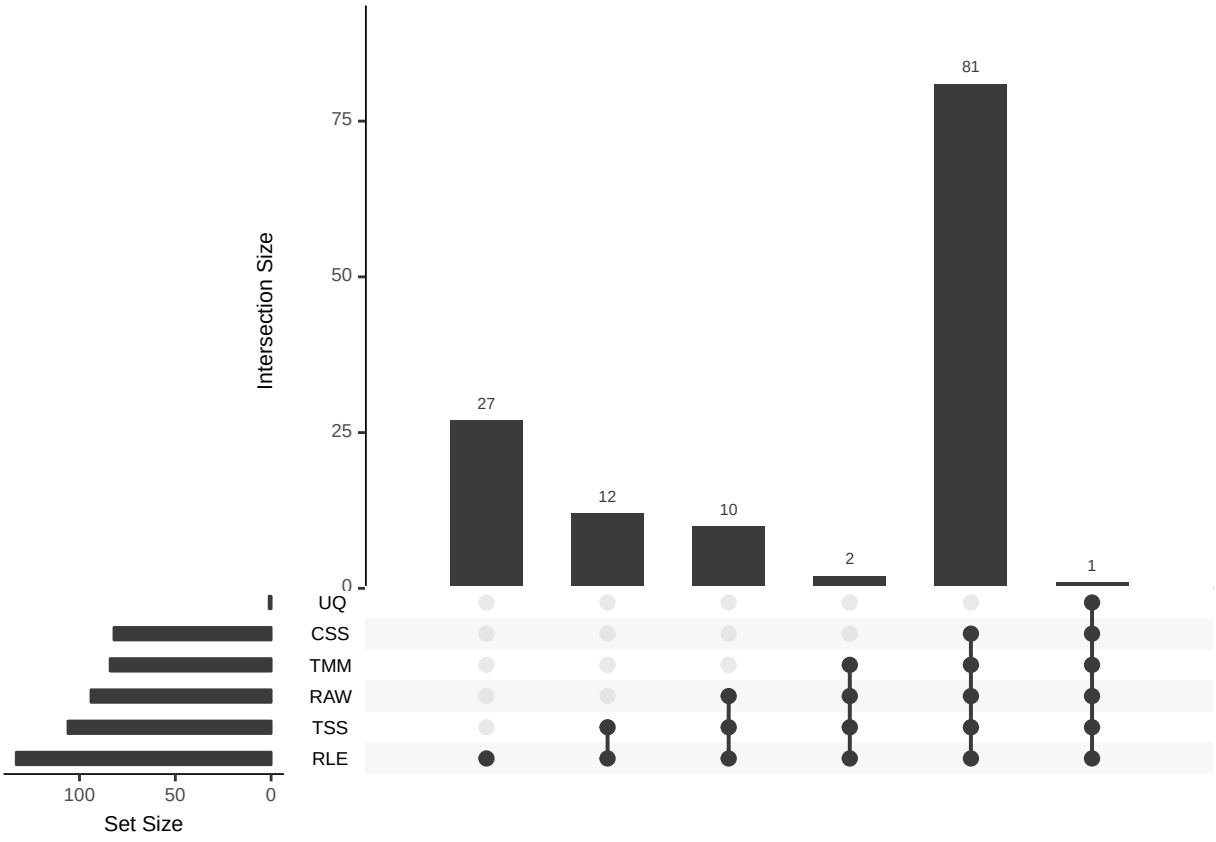

```
logFC_pre_set %>% filter(biosample_id == "E01JH0016") %>% as.data.frame() %>%
                            upset(nset = 6)
```




```
logFC_pre_set %>% filter(biosample_id == "E01JH0017") %>% as.data.frame() %>%
                            upset(nset = 6)
```




```
logFC_pre_set %>% filter(biosample_id == "E01JH0038") %>% as.data.frame() %>%
                                upset(nset = 6)
```




```
## Warning: Removed 1328 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1328 rows containing non-finite values (stat_smooth).
## Warning: Removed 1328 rows containing non-finite values (stat_binline).
## Warning: Removed 20580 rows containing missing values (geom_text).
## Warning: Removed 1328 rows containing non-finite values (stat_binline).
## Warning: Removed 20580 rows containing missing values (geom_text).
```

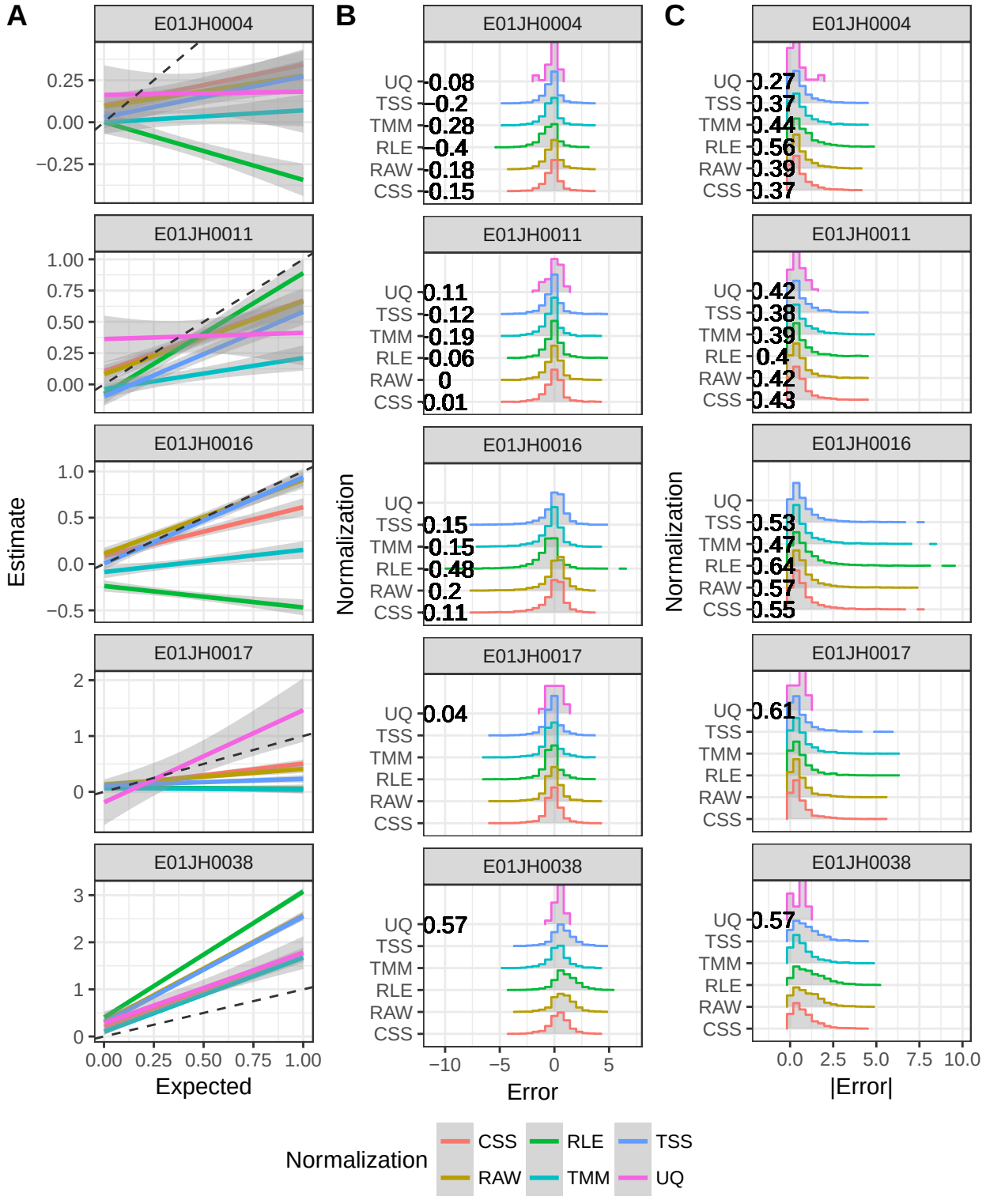


Figure 1: Impact of normalization methods on the agreement between log fold-change estimates and expected values for pre-specific and pre-dominant features. (A) Linear model relating the log fold-change estimates with the expected values by individual and normalization method. (B) Distribution of log fold-change (B) absolute error and (C) error by normalization method and individual.