Location of MPRA variants in repeats, Vijay's data

Grace Hansen

06/29/2019

Location of MPRA variants in repeats

Variant type	Number of variants	Proportion in repeats	Proportion in Alus
EMVar	68	0.4853	0.2059
Enhancer	339	0.5015	0.2301
Nonsignificant	2223	0.5888	0.2519

Retrovirus types in MPRA

Types of repeats in EMVars:

RV	count
SINE	15
LINE	9
LTR	8
Simple_repeat	1
DNA	0
DNA?	0
snRNA	0
Retroposon	0

Sub-types of repeats in EMVars:

RV subtype	count
AluY	4
AluSp	2
AluYk3	2
AluJo	1
AluSg	1
AluSx	1
AluYf1	1
AluYk2	1
AluYm1	1
FLAM_C	0
HAL1	0
HERV9	0
L1M5_orf2	0
L1MA4_3end	0

RV subtype	count
L1MC1_3end	0
L1MC4_3end	0
L1ME1_3end	0
L1ME4b_3end	0
L1MEd_5end	0
L1P1_orf2	0
L1PA4_3end	0
LTR12C	0
LTR12E	0
LTR53	0
MLT1E2	0
THE1B	0
THE1-int	0

Plot results with chi-square p values

Are there more repeats and Alus in repeats than expected by chance?

```
## [1] "Repeat Chi square observed vs expected tables:"
```

```
##           repeats no repeats
## EMVars         33         35
## enhancers     170        169
## nonsig       1309        914
```

```
##              repeats no repeats
## EMVars      39.09354   28.90646
## enhancers  194.89278  144.10722
## nonsig    1278.01369  944.98631
```

```
## [1] "Alu Chi square observed vs expected tables:"
```

```
##           repeats no repeats
## EMVars         14         54
## enhancers      78        261
## nonsig        560       1663
```

```
##              repeats no repeats
## EMVars     16.85779   51.14221
## enhancers  84.04106  254.95894
## nonsig    551.10114 1671.89886
```

```
## pdf
##   2
```

```
## pdf
##   2
```

Expression of Alu elements

Are sequences containing Alu elements more highly expressed than sequences not containing Alu elements?

```
## [1] "qnorm from nonsignificant sequences without Alu elements"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.2955 -0.4690 -0.2969 -0.2991 -0.1289  1.6376
```

```
## [1] "qnorm from nonsignificant sequences with Alu elements"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.1930 -0.4693 -0.2916 -0.3110 -0.1582  0.3839
```

```
## [1] "qnorm from enhancer sequences without Alu elements"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.1584  0.1520  0.2983  0.5848  0.6244  5.7360
```

```
## [1] "qnorm from enhancer sequences with Alu elements"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.1665  0.1147  0.2770  0.3801  0.4592  1.8944
```

```
## [1] "qnorm from EMVar sequences without Alu elements"
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.23262  0.01322  0.19763  0.51134  0.87895  3.76455
```

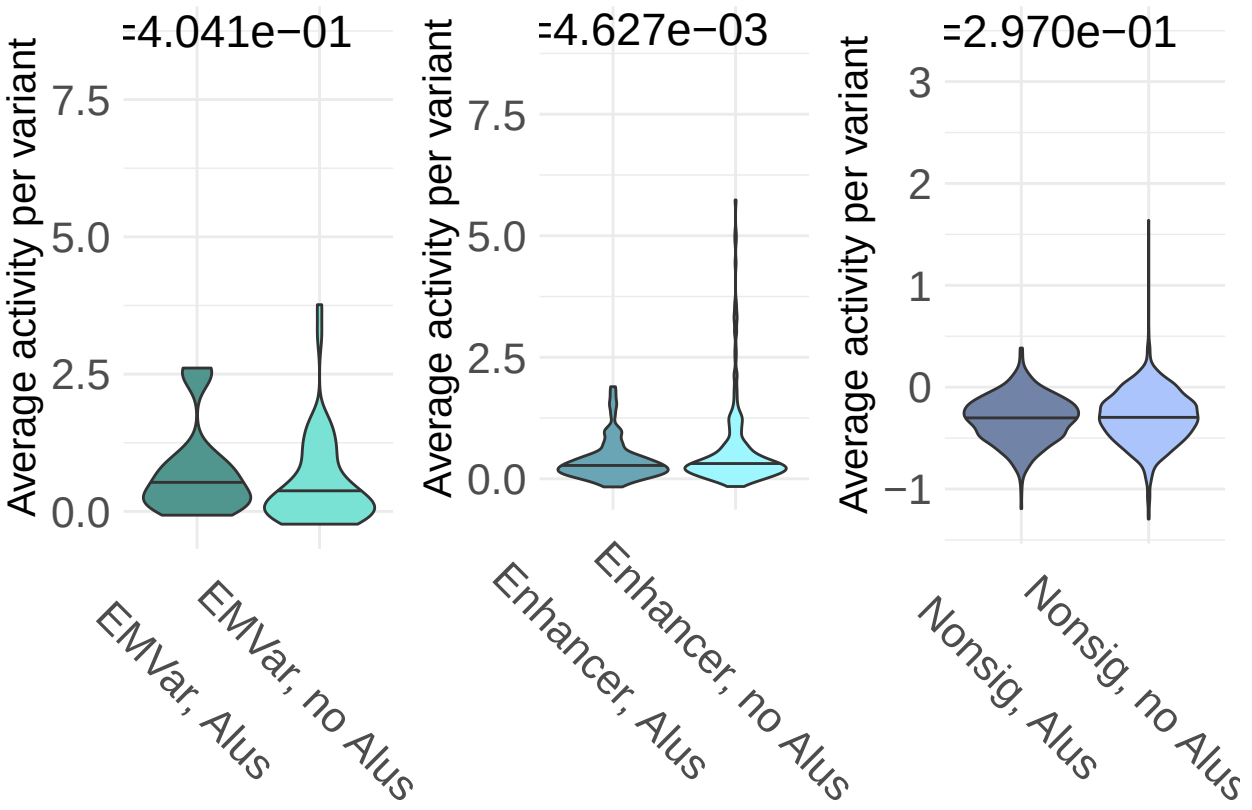
```
## [1] "qnorm from EMVar sequences with Alu elements"
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.06914  0.17000  0.48447  0.72253  0.82545  2.61034
```

Significance and visualization

In the plots below, you can see that the Alu-containing sequences have higher expression in enhancers and EMVars, but this isn't true for nonsignificant sequences.

```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
## 'x' values
```




```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
## 'x' values
```

```
## pdf
##   2
```