Location of MPRA variants in repeats

Grace Hansen

06/29/2019

Location of MPRA variants in repeats

Variant type	Number of variants	Proportion in repeats	Proportion in Alus
EMVar	59	0.9153	0.7627
Enhancer	367	0.6649	0.4414
Nonsignificant	1429	0.5843	0.2407

Retrovirus types in MPRA

Types of repeats in EMVars:

RV	count
SINE	45
LINE	5
LTR	3
DNA	1
Retroposon	0
Satellite	0
Simple_repeat	0
snRNA	0
tRNA	0

Sub-types of repeats in EMVars:

RV subtype	count
AluSx	7
AluY	7
AluSq2	4
AluJb	3
AluSg	3
AluSp	3
AluSx3	3
AluSz	3
AluJr	2
AluSc8	2
AluSc	1
AluSc5	1
AluSg4	1
AluSq10	1
AluSz6	1
AluYe6	1
AluYk2	1
AluYk3	1

RV subtype	count
HAL1	1
L1M4_orf2	1
L1MA8_3end	1
L1MC4a_3end	1
L2c_3end	1
LTR12C	1
LTR15	1
MER5B	1
MLT1L	1

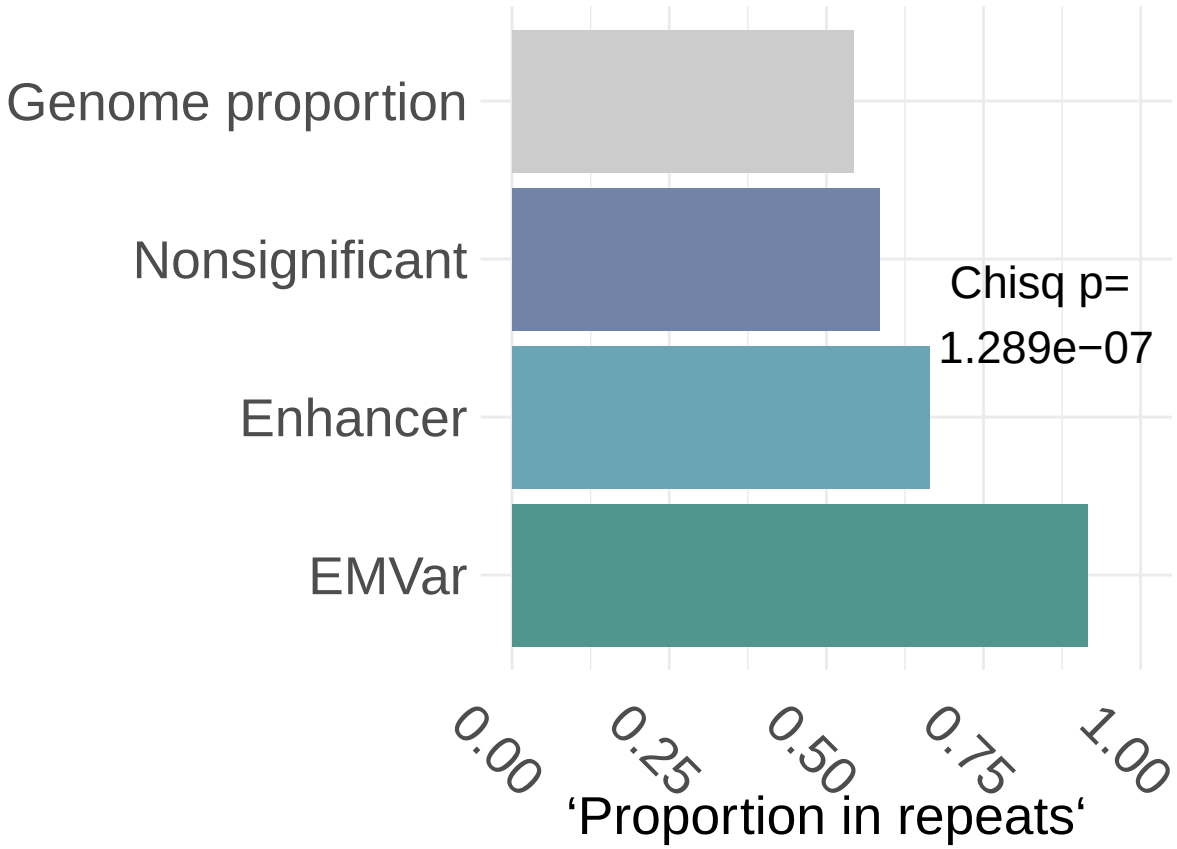
Strandedness of enhancer elements

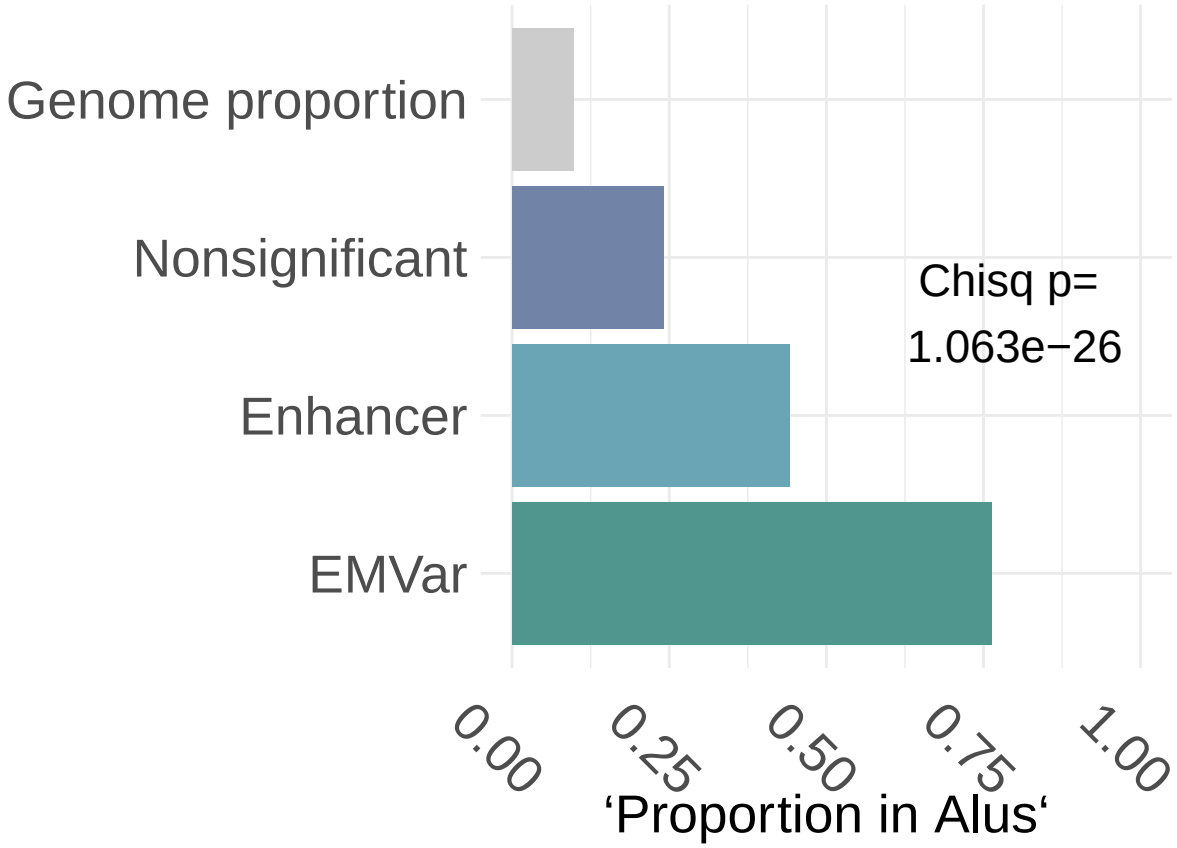
Are the Alus that drive expression in the same orientation as the barcode (i.e. on the + strand?)

	-	+
EMVar	5	40
enhancer	44	118
nonsig	204	140

Plot results with chi-square p values

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


Genome proportion
Nonsignificant
Enhancer
EMVar
Chisq p=
1.289e−07
0.00
0.25
0.50
0.75
1.00
‘Proportion in repeats‘


Genome proportion
Nonsignificant
Enhancer
EMVar
Chisq p=
1.063e−26
0.00
0.25
0.50
0.75
1.00
‘Proportion in Alus‘

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, Rep 1"
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.32001 -0.40233 -0.13417 -0.07431  0.14177  8.41046
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.0289 -0.4245 -0.1493 -0.1023  0.1228  6.3667
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.2693  0.3218  0.6367  0.8580  1.1103  8.6051
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.1010  0.5719  1.2580  2.2045  3.1462  9.3478
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.9793  0.6912  1.1989  1.5775  1.7758  5.4650
```

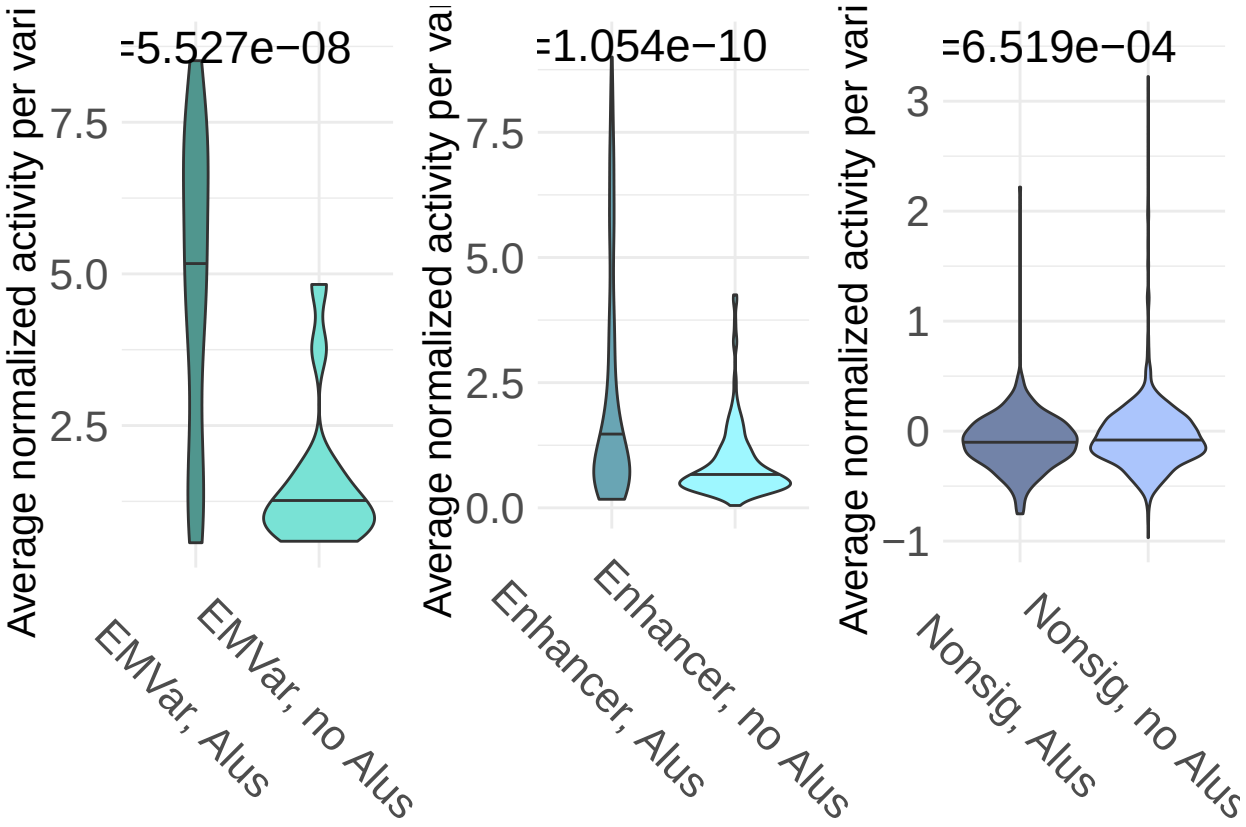
```
## [1] "qnorm from EMVar sequences with Alu elements, Rep 1"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -0.437   2.608   5.263   4.853   7.019   8.822
```

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
```



Motifs in expressed Alus

Do these Alus with increased expression contain the GAGGTCA motif?

```
## [1] 0.6962025
```

```
## [1] "GAGGCCGA"
```