

Association of Genomic Features with Integration

October 16, 2006

Contents

1	Introduction	3
2	Preference for Genes	4
2.1	Acembly Genes	4
2.2	refGenes	6
2.3	ensGenes	8
2.4	genScan Genes	10
2.5	uniGenes	12
3	CpG Island Neighborhoods	14
3.1	1 kilobase neighborhoods	14
3.2	5 kilobase neighborhoods	15
3.3	10 kilobase neighborhoods	16
3.4	25 kilobase neighborhoods	17
3.5	50 kilobase neighborhoods	18
4	Gene Density, Expression 'Density', and CpG Island Density	19
4.1	25 kiloBase Window	20
4.2	50 kiloBase Window	26
4.3	100 kiloBase Window	31
4.4	250 kiloBase Window	40
4.5	500 kiloBase Window	45
4.6	1 megaBase Window	54
4.7	2 megaBase Window	63
4.8	4 megaBase Window	72
4.9	8 megaBase Window	77
4.10	16 megaBase Window	82
4.11	32 megaBase Window	87

5	Juxtaposition with Gene Start and End Positions	92
5.1	Acembly Annotations	92
5.2	RefSeq Annotations	96
5.3	genScan Annotations	100
5.4	uniGene Annotations	104
6	Cytobands	108

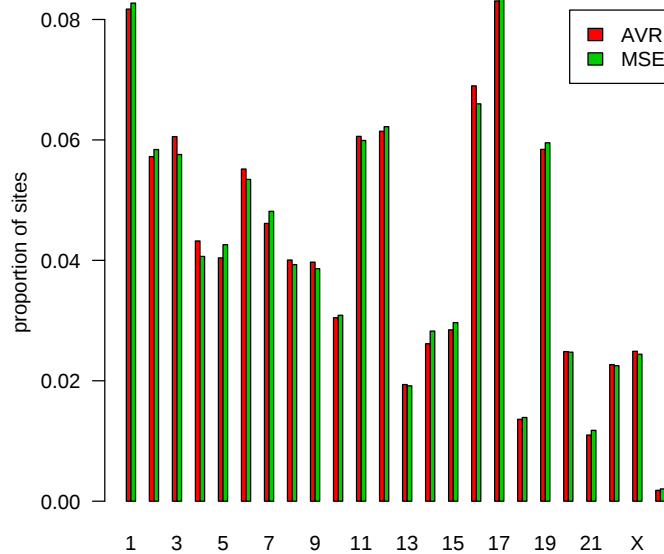
1 Introduction

In this document, I examine the association of integration siting with various genomic features.

The numbers are shown below:

```
Origin.of.data.set
  AVR   MSE
19962 20607
```

The distribution of relative frequency of insertions across the chromosomes is given in this barplot:

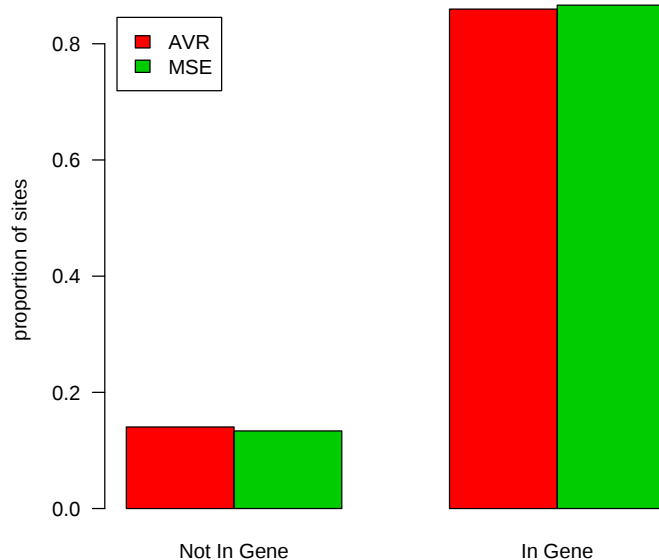


Are there chromosomes that are particularly favored for integration by one group over the other? This was tested for statistical significance. The test performed used the likelihood ratio statistic for the logistic regression model (reviewed in [1]) as implemented by the `glm` function of R using the `binomial` family. The null hypothesis tested is the ratio of true integration events in the two groups is constant across all chromosomes. This test attains a p-value of 0.97773.

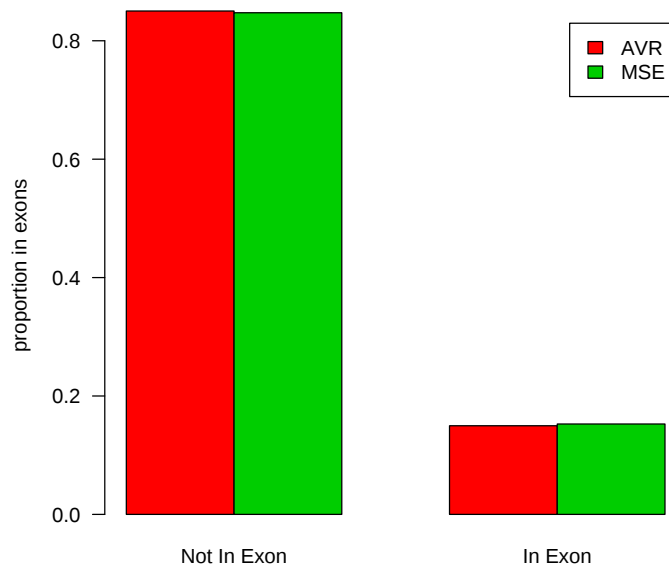
2 Preference for Genes

2.1 Acembly Genes

Here we examine the relative preference that integration events in the two groups have for genes. In the following plot we show the relative frequency of integrations in genes according to the 'Acembly' anotation. The bars grouped over the label "In Gene" give the relative frequency of integration events between bases located within Acembly gene annotations, while the label "Not in Gene" give the relative frequency of integration events between bases not located within Acembly gene annotations.



Is there is a difference in the tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.045753. In the following plot we show the relative frequency of insertions in exons according to the 'Acembly' anotation. The bars grouped over the label "In Exon" give the relative frequency of integration events between bases located in exons according to the Acembly annotation, while the label "Not in Exon" give the relative frequency of integration events between bases not located in exons according to the Acembly gene annotation.



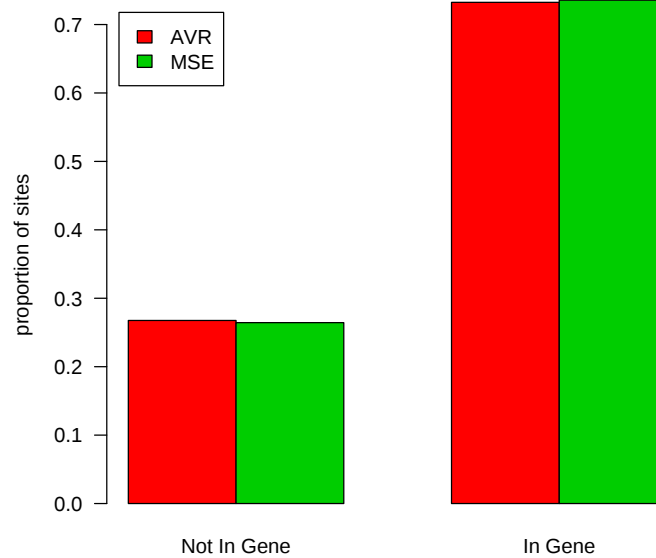
Here is the table of coefficients of the log ratio of intensities along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	-0.0180	0.0268	-0.671	0.5020
in.gene	0.0553	0.0293	1.890	0.0593
in.exon	0.0139	0.0281	0.495	0.6200

The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as 'in.exon' is net of that due to being in a gene. Note that in the barplot above the 'Not in Exon' bars include the both introns and intergenic regions, so the impression given by the table may differ from that for the barplot.

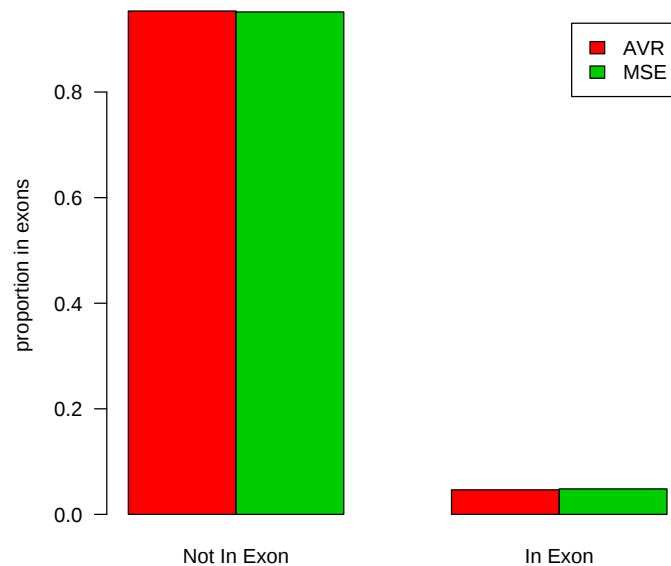
2.2 refGenes

Here we examine the relative preference that insertions of the two types have for genes. In the following plot we show the relative frequency of insertions in genes according to the 'refGene' annotation.



Is there is a tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.46832.

In the following plot we show the relative frequency of insertions in exons according to the 'refGene' anotation.



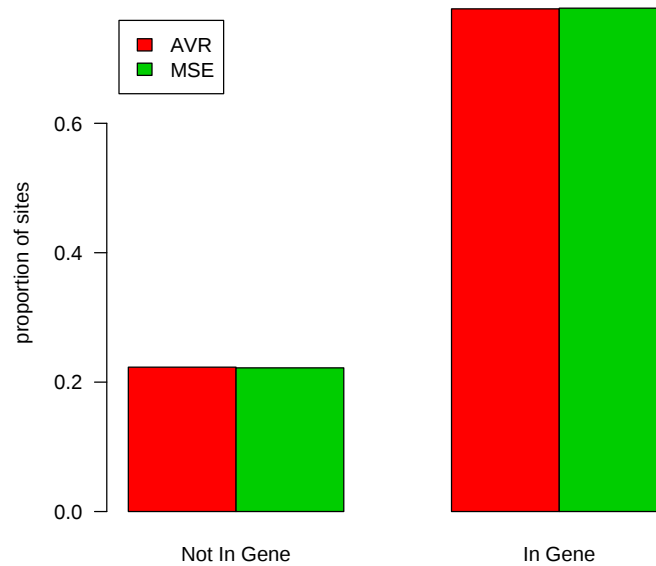
Here is the table of coefficients of the log ratio of intensities for along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	0.0198	0.0193	1.030	0.303
in.gene	0.0140	0.0227	0.617	0.538
in.exon	0.0359	0.0472	0.761	0.447

The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as 'in.exon' is net of that due to being in a gene. Note that in the barplot above the 'Not in Exon' bars include the both introns and intergenic regions, so the impression given by the table may differ from that for the barplot.

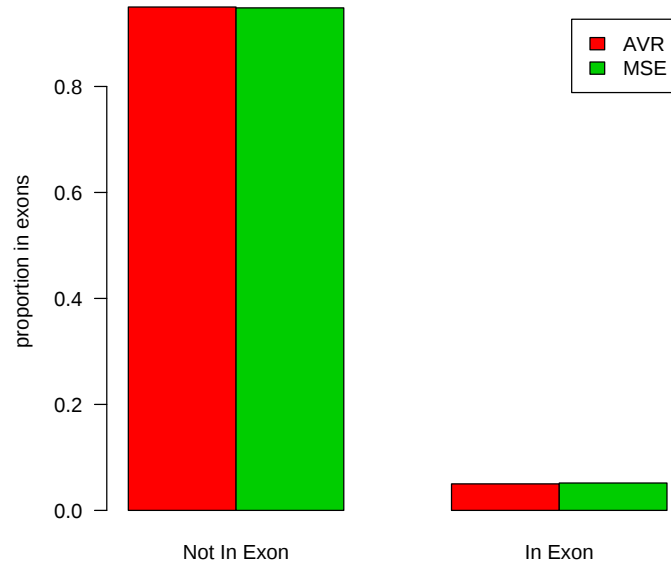
2.3 ensGenes

Here we examine the relative preference that insertions of the two types have for genes. In the following plot we show the relative frequency of insertions in genes according to the 'ensGene' annotation.



Is there is a tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.79685.

In the following plot we show the relative frequency of insertions in exons according to the 'ensGene' anotation.



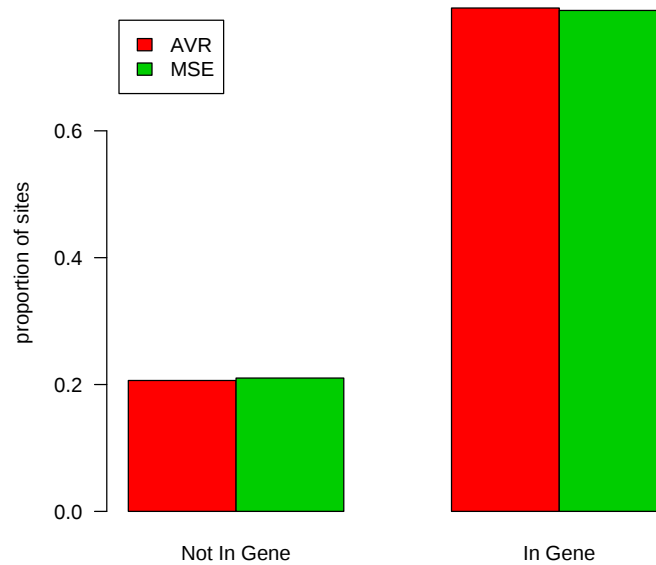
Here is the table of coefficients of the log ratio of intensities for along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	0.02700	0.0210	1.280	0.199
in.gene	0.00399	0.0241	0.166	0.868
in.exon	0.03300	0.0456	0.724	0.469

The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as 'in.exon' is net of that due to being in a gene. Note that in the barplot above the 'Not in Exon' bars include the both introns and intergenic regions, so the impression given by the table may differ from that for the barplot.

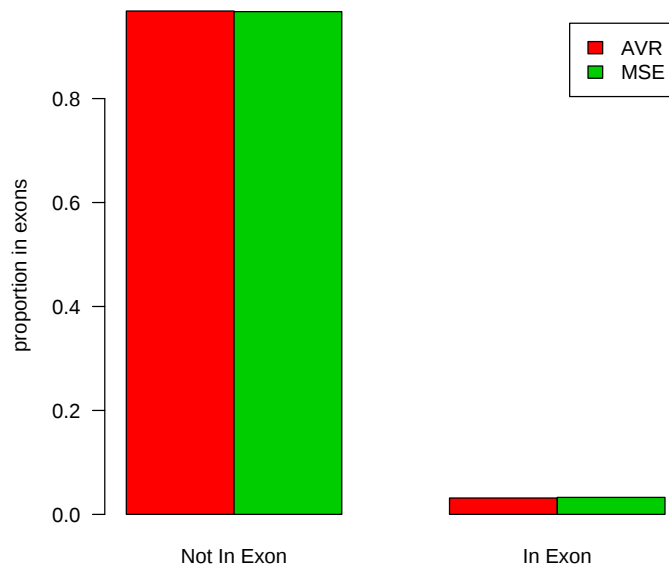
2.4 genScan Genes

Here we examine the preference that insertions have for genes. In the following plot we show the relative frequency of insertions in genes according to the 'genScan' annotation.



Is there is a tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.34257.

In the following plot we show the relative frequency of insertions in exons according to the 'genScan' annotation.



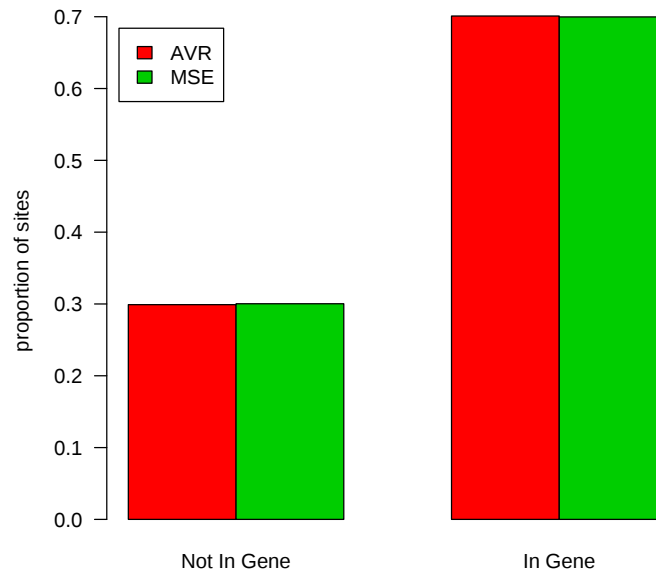
Here is the table of coefficients of the log ratio of intensities along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	0.0502	0.0218	2.310	0.0211
in.gene	-0.0250	0.0246	-1.020	0.3090
in.exon	0.0439	0.0566	0.776	0.4380

The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as 'in.exon' is net of that due to being in a gene. Note that in the barplot above the 'Not in Exon' bars include the both introns and intergenic regions, so the impression given by the table may differ from that for the barplot.

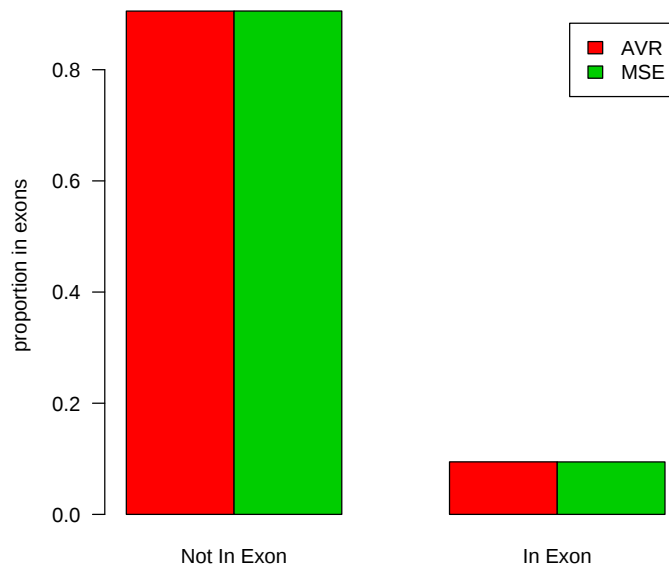
2.5 uniGenes

Here we examine the preference that insertions have for genes. In the following plot we show the relative frequency of insertions in genes according to the 'uniGene' anotation.



Is there is a tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.78043.

In the following plot we show the relative frequency of insertions in exons according to the 'uniGene' anotation.



Here is the table of coefficients of the log ratio of intensities along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	0.036000	0.0181	1.99000	0.047
in.gene	-0.006090	0.0222	-0.27400	0.784
in.exon	0.000328	0.0347	0.00945	0.992

The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as 'in.exon' is net of that due to being in a gene. Note that in the barplot above the 'Not in Exon' bars include the both introns and intergenic regions, so the impression given by the table may differ from that for the barplot.

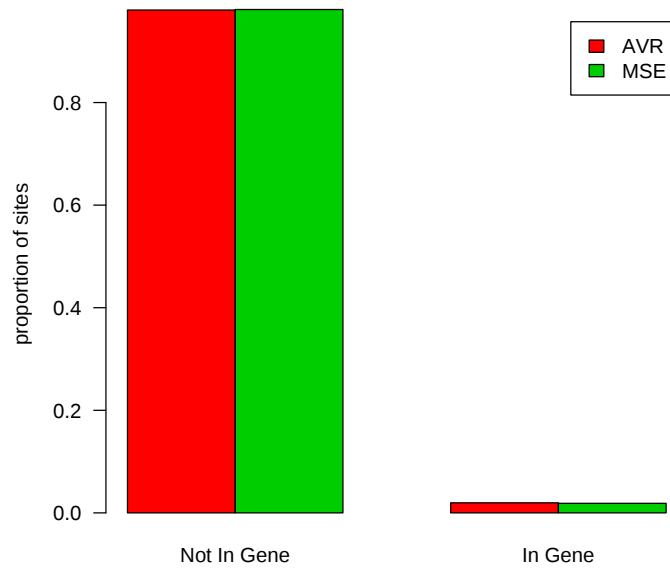
3 CpG Island Neighborhoods

Here we study the effect of being in the neighborhood of CpG Islands. Following Wu et al [2], who found that the neighborhoods within $\pm 1\text{kb}$ of CpG islands are enriched for MLV insertions, we study such neighborhoods.

3.1 1 kilobase neighborhoods

The following plot shows the effect of being in or within $\pm 1\text{kb}$ of a CpG island:

NULL

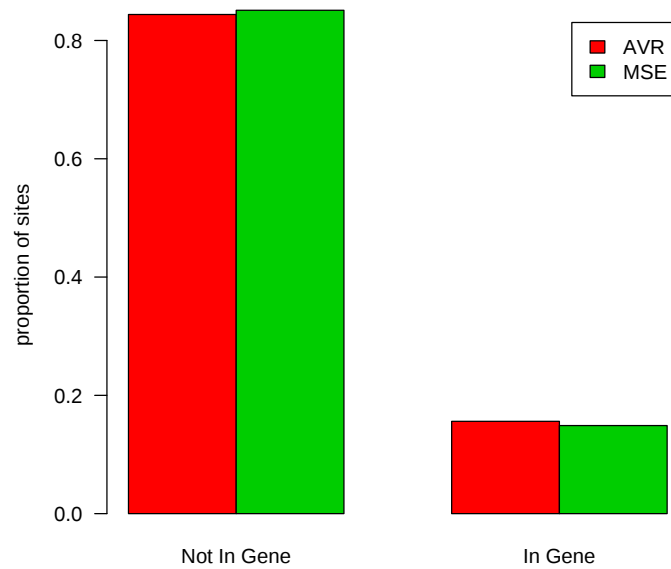


A formal test of significance comparing the difference attains a p-value of 0.46101.

3.2 5 kilobase neighborhoods

The following plot shows the effect of being in or within $\pm 5\text{kb}$ of a CpG island:

NULL

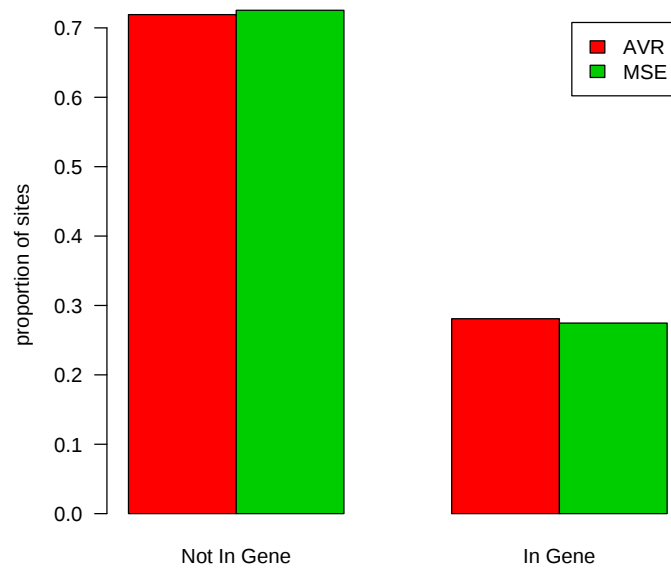


A formal test of significance comparing the difference attains a p-value of 0.044702.

3.3 10 kilobase neighborhoods

The following plot shows the effect of being in or within ± 10 kb of a CpG island:

NULL

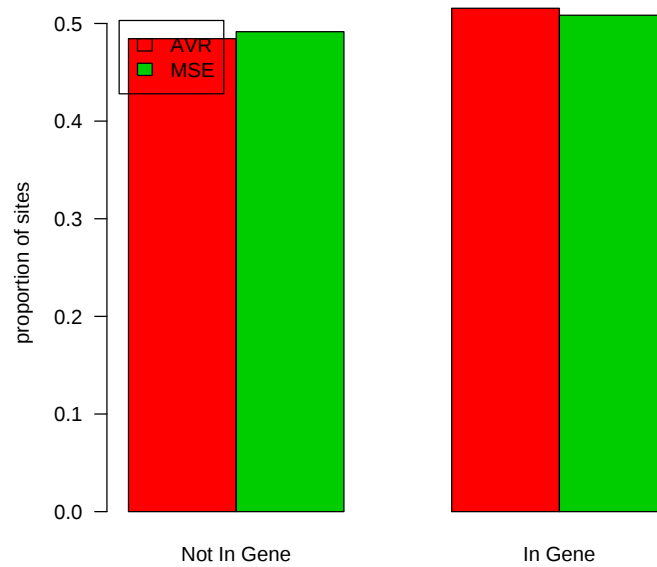


A formal test of significance comparing the difference attains a p-value of 0.15875.

3.4 25 kilobase neighborhoods

The following plot shows the effect of being in or within $\pm 25\text{kb}$ of a CpG island:

NULL

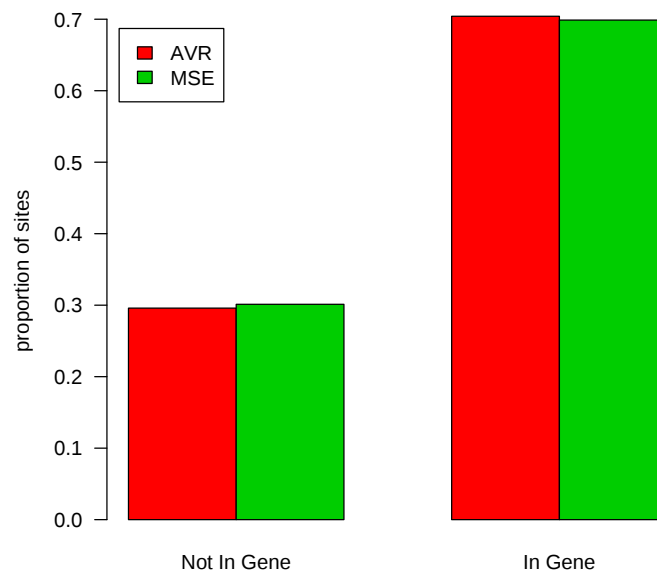


A formal test of significance comparing the difference attains a p-value of 0.14919.

3.5 50 kilobase neighborhoods

The following plot shows the effect of being in or within $\pm 50\text{kb}$ of a CpG island:

NULL



A formal test of significance comparing the difference attains a p-value of 0.23944.

4 Gene Density, Expression 'Density', and CpG Island Density

In this section the association with gene density is examined. We examine gene densities using the various gene calls - Acembly, RefSeq, Ensembl, and GenScan. For expression analysis, the 'genes' that are counted are the genes represented on the microarray. In addition, we the number of such genes expressed at various levels. The levels are

low.ex Count genes whose expression is in the upper half and divide by number of bases

med.ex Count genes whose expression is in the upper half $1/8^{th}$ and divide by number of bases

high.ex Count genes whose expression is in the upper half $1/16^{th}$ and divide by number of bases

The bolded terms are used as abbreviations in what follows. The abbreviation **dens** is used to indicate gene density as number of genes per base.

4.1 25 kiloBase Window

In the barplot that follows we examine the association of insertion sites with gene density in a 25 kilobase window surrounding each locus. More such plots will follow and the method of their construction is always to try to divide the data according to the deciles of density. However, it often happens that there is a very skewed distribution of density and often even the 90th percentile is zero. In that case, the barplots simply show the sites for which the density is zero and those for which it is non-zero. If there are fewer than ten groups of bars, then the groupings contain ten percent of the sites each except for the leftmost grouping which will contain all of the remaining sites.

Also note that the title of the plot contains clues as to its content; the prefix indicates the type of variable studied while the suffix indicates the window width in the number of bases. The p-value given is the result of fitting a cubic polynomial to the gene density values.

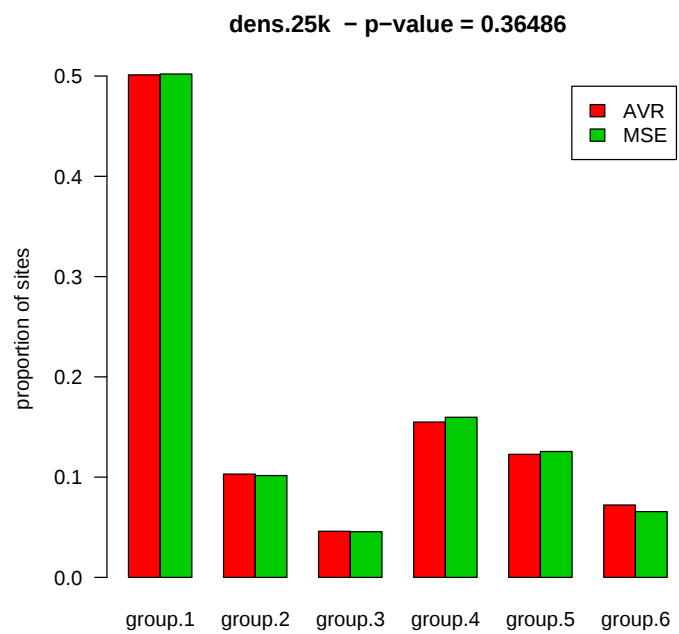
The following expression data and probe set were used for this report:

```
[1] "Jurkat-HU133Plus2"
```

```
[1] "HG-U133"
```

Category limits

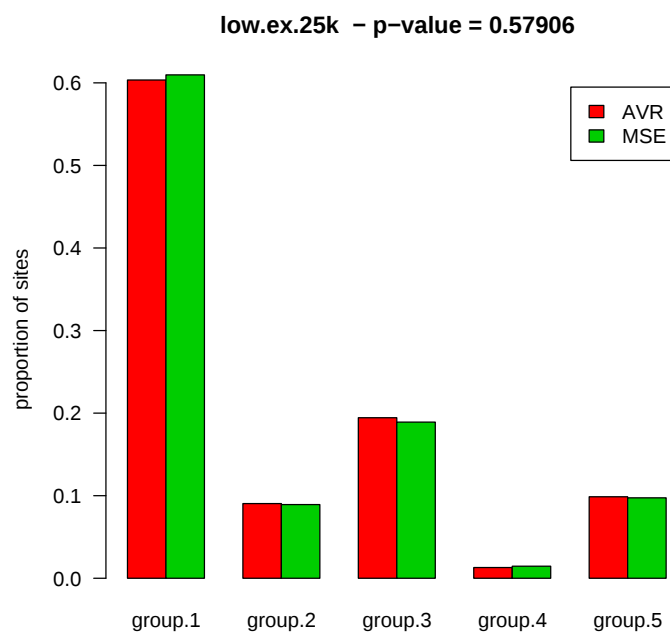
	lower	category	upper
1	0.000000e+00	group.1	1.333333e-05
2	1.333333e-05	group.2	2.666667e-05
3	2.666667e-05	group.3	4.000000e-05
4	4.000000e-05	group.4	4.000000e-05
5	4.000000e-05	group.5	8.000000e-05
6	8.000000e-05	group.6	4.000000e-04



Here are the results for expression density. First, we count just genes that are in the upper half.

Category limits

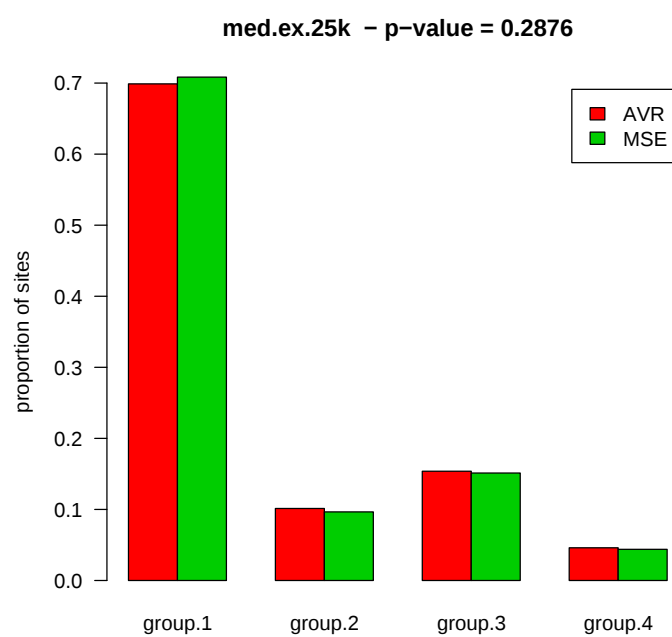
	lower	category	upper
1	0.000000e+00	group.1	1.333333e-05
2	1.333333e-05	group.2	2.666667e-05
3	2.666667e-05	group.3	4.000000e-05
4	4.000000e-05	group.4	5.000000e-05
5	5.000000e-05	group.5	2.400000e-04



Now we count genes in the upper 1/8th:

Category limits

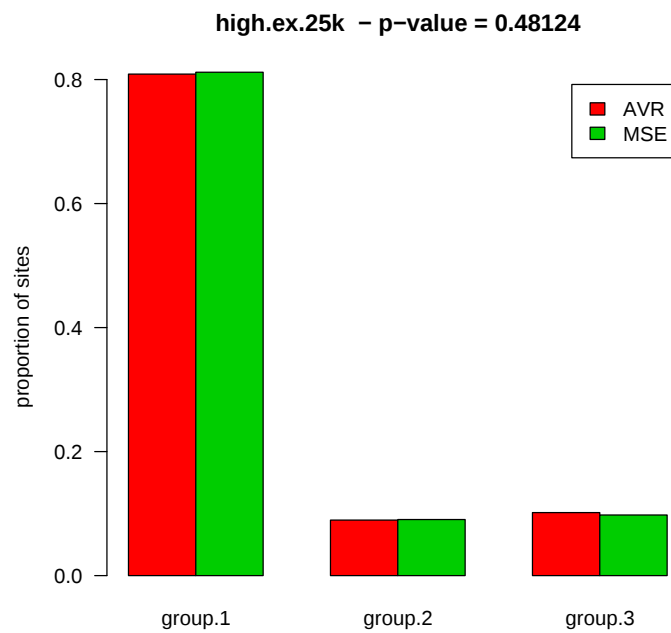
	lower	category	upper
1	0.000000e+00	group.1	1.333333e-05
2	1.333333e-05	group.2	2.000000e-05
3	2.000000e-05	group.3	4.000000e-05
4	4.000000e-05	group.4	1.733333e-04



And here we count genes in the upper 1/16th:

Category limits

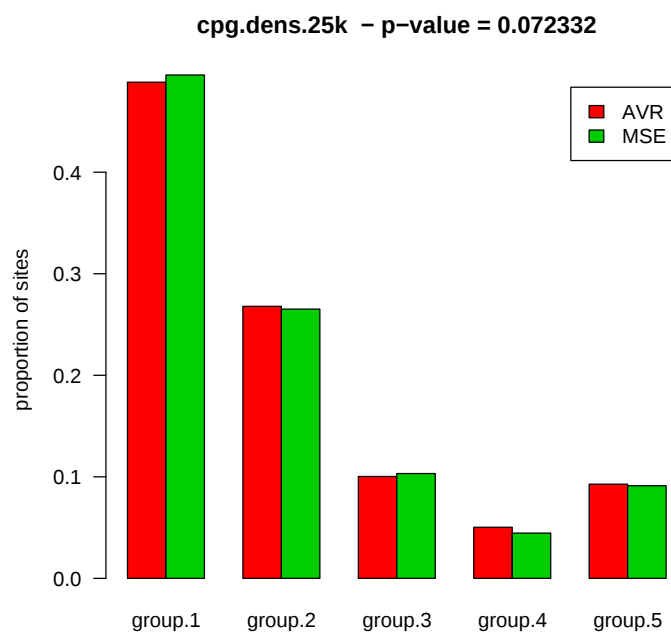
	lower	category	upper
1	0.000000e+00	group.1	1.000000e-05
2	1.000000e-05	group.2	2.666667e-05
3	2.666667e-05	group.3	1.400000e-04



Here the effect of density of CpG islands is studied:

Category limits

	lower	category	upper
1	0e+00	group.1	0.00002
2	2e-05	group.2	0.00002
3	2e-05	group.3	0.00004
4	4e-05	group.4	0.00006
5	6e-05	group.5	0.00036

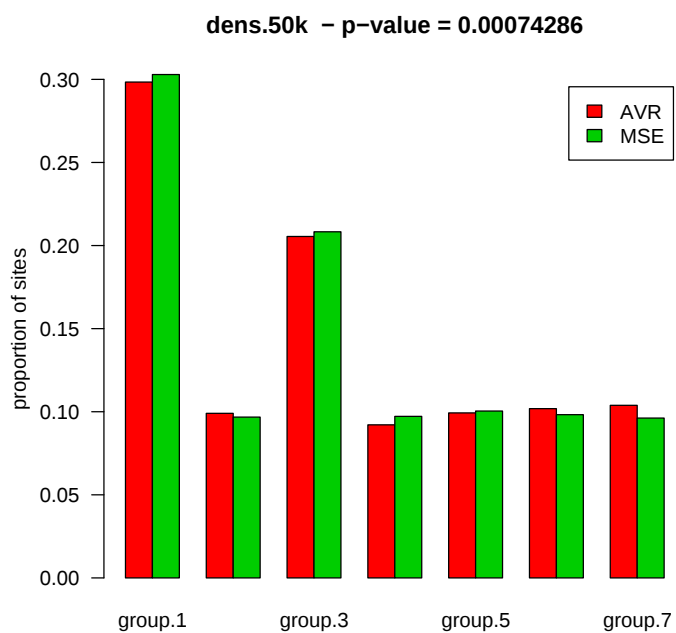


4.2 50 kiloBase Window

In the barplot that follows we examine the association of insertion sites with expression density in a 50 kilobase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category limits

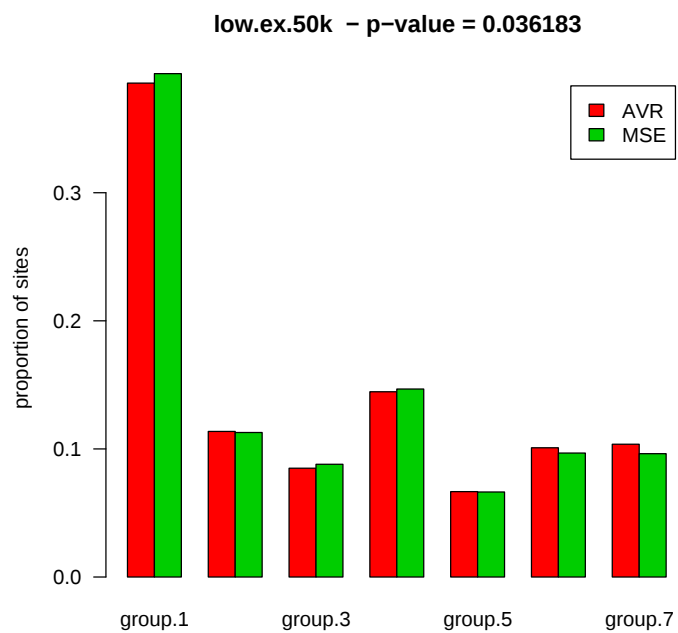
	lower	category	upper
1	0.000000e+00	group.1	6.666667e-06
2	6.666667e-06	group.2	1.333333e-05
3	1.333333e-05	group.3	2.000000e-05
4	2.000000e-05	group.4	3.500000e-05
5	3.500000e-05	group.5	4.142857e-05
6	4.142857e-05	group.6	6.333333e-05
7	6.333333e-05	group.7	3.000000e-04



Here are the results for expression density. First, we count just genes that are in the upper half.

Category limits

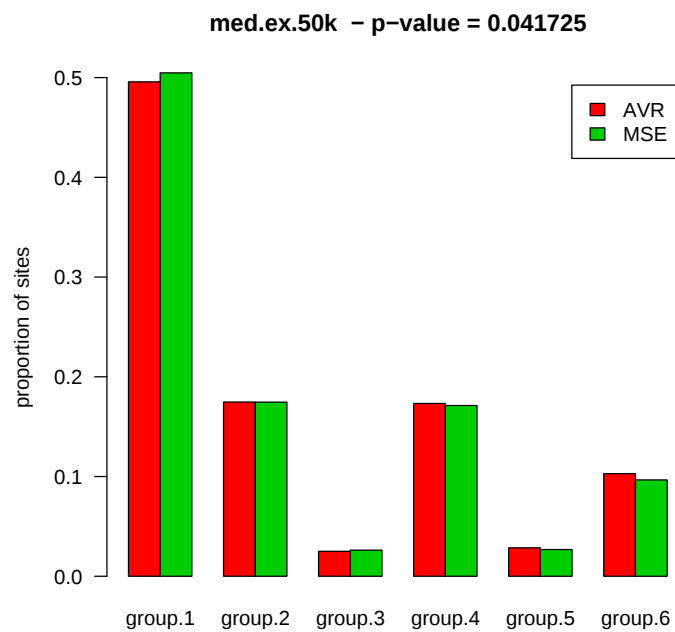
	lower	category	upper
1	0.000000e+00	group.1	6.666667e-06
2	6.666667e-06	group.2	1.000000e-05
3	1.000000e-05	group.3	2.000000e-05
4	2.000000e-05	group.4	2.000000e-05
5	2.000000e-05	group.5	3.000000e-05
6	3.000000e-05	group.6	4.666667e-05
7	4.666667e-05	group.7	2.000000e-04



Now we count genes in the upper 1/8th:

Category limits

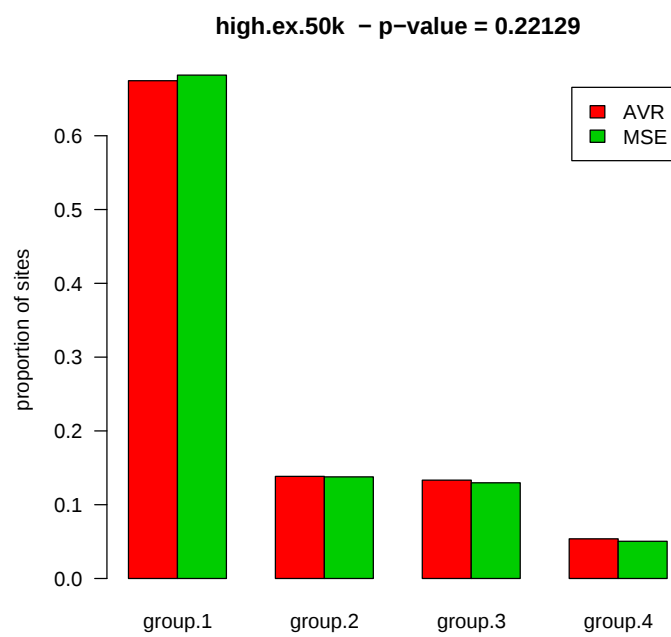
	lower	category	upper
1	0.000000e+00	group.1	4.000000e-06
2	4.000000e-06	group.2	1.000000e-05
3	1.000000e-05	group.3	1.333333e-05
4	1.333333e-05	group.4	2.000000e-05
5	2.000000e-05	group.5	2.857143e-05
6	2.857143e-05	group.6	1.200000e-04



And here we count genes in the upper 1/16th:

Category limits

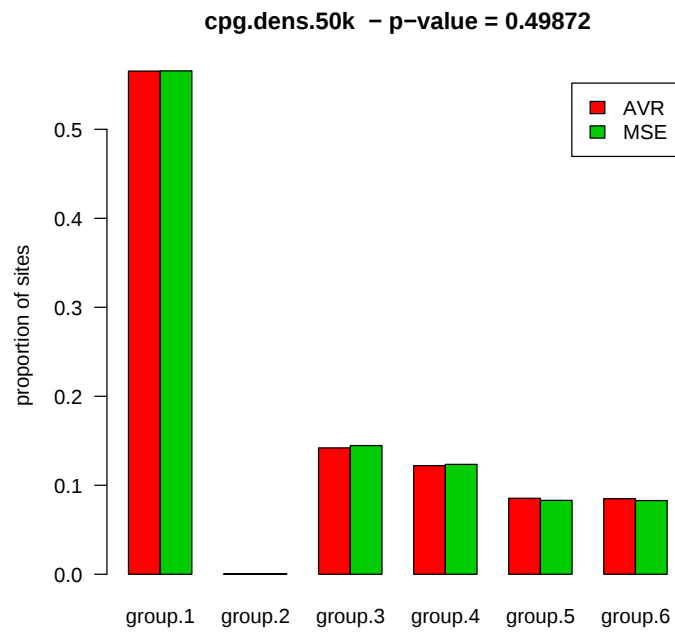
	lower	category	upper
1	0e+00	group.1	5e-06
2	5e-06	group.2	1e-05
3	1e-05	group.3	2e-05
4	2e-05	group.4	1e-04



Here the effect of density of CpG islands is studied:

Category limits

	lower	category	upper
1	0e+00	group.1	0.00001
2	1e-05	group.2	0.00002
3	2e-05	group.3	0.00002
4	2e-05	group.4	0.00004
5	4e-05	group.5	0.00007
6	7e-05	group.6	0.00027

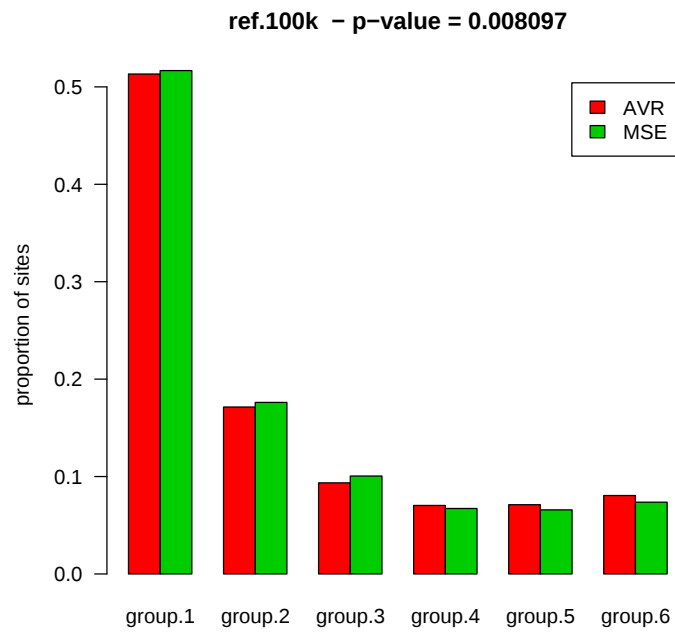


4.3 100 kiloBase Window

Here are gene densities for the various gene calls

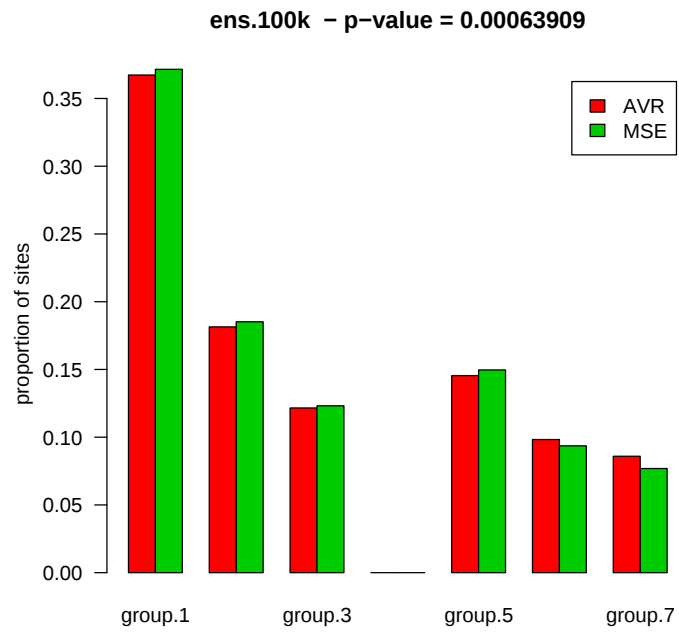
Category limits

	lower	category	upper
1	0e+00	group.1	0.00001
2	1e-05	group.2	0.00002
3	2e-05	group.3	0.00003
4	3e-05	group.4	0.00004
5	4e-05	group.5	0.00006
6	6e-05	group.6	0.00033



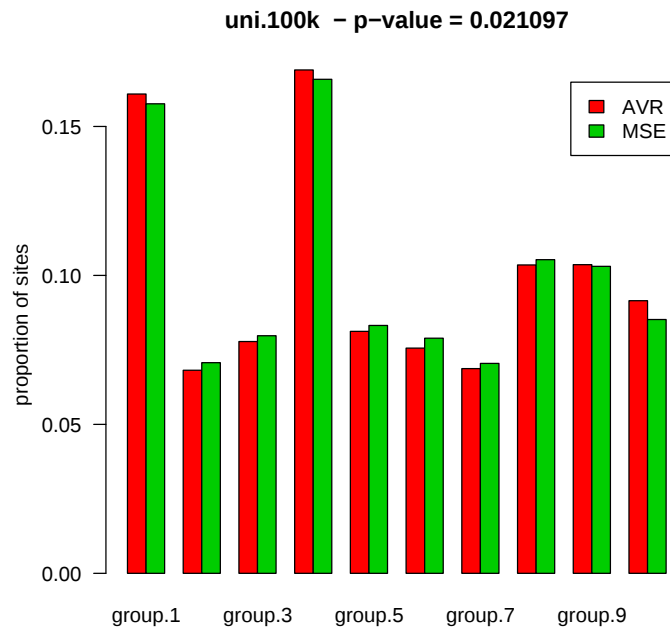
Category limits

	lower	category	upper
1	0e+00	group.1	0.00001
2	1e-05	group.2	0.00002
3	2e-05	group.3	0.00003
4	3e-05	group.4	0.00004
5	4e-05	group.5	0.00005
6	5e-05	group.6	0.00008
7	8e-05	group.7	0.00028



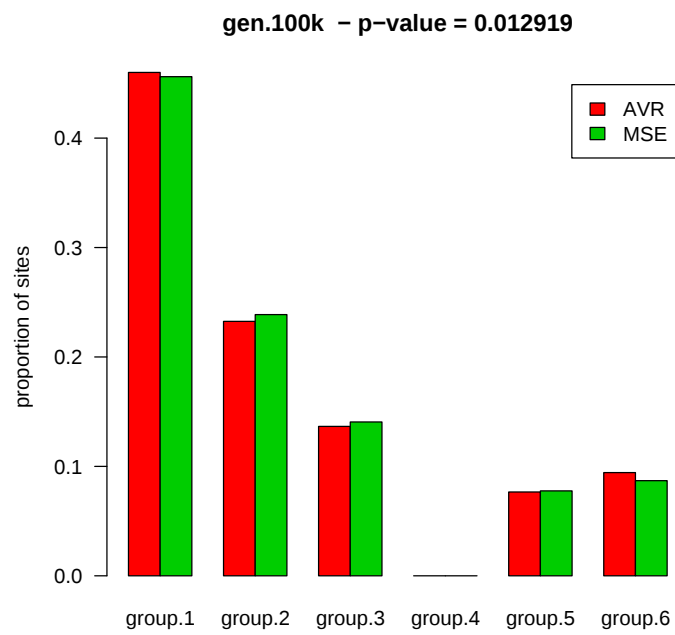
Category limits

	lower	category	upper
1	0.00000	group.1	0.00002
2	0.00002	group.2	0.00003
3	0.00003	group.3	0.00004
4	0.00004	group.4	0.00006
5	0.00006	group.5	0.00007
6	0.00007	group.6	0.00008
7	0.00008	group.7	0.00009
8	0.00009	group.8	0.00011
9	0.00011	group.9	0.00014
10	0.00014	group.10	0.00052



Category limits

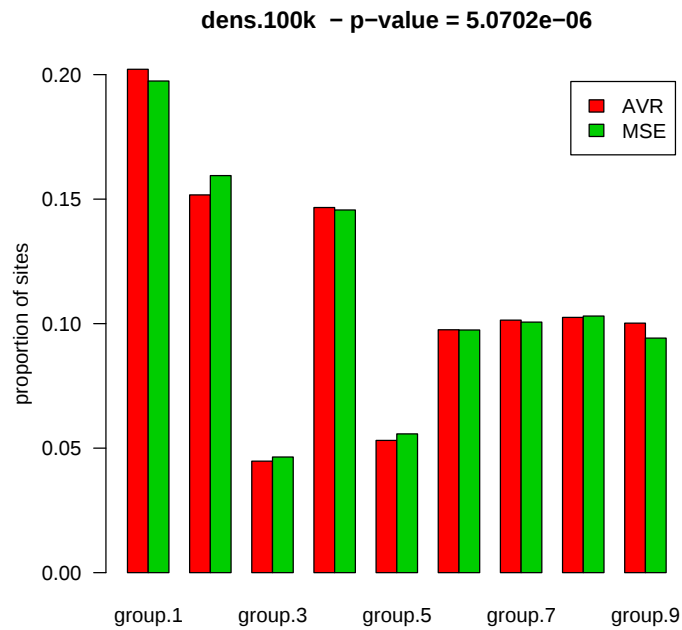
	lower	category	upper
1	0e+00	group.1	0.00001
2	1e-05	group.2	0.00002
3	2e-05	group.3	0.00003
4	3e-05	group.4	0.00003
5	3e-05	group.5	0.00004
6	4e-05	group.6	0.00018



In the barplot that follows we examine the association of insertion sites with expression density in a 100 kilobase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category limits

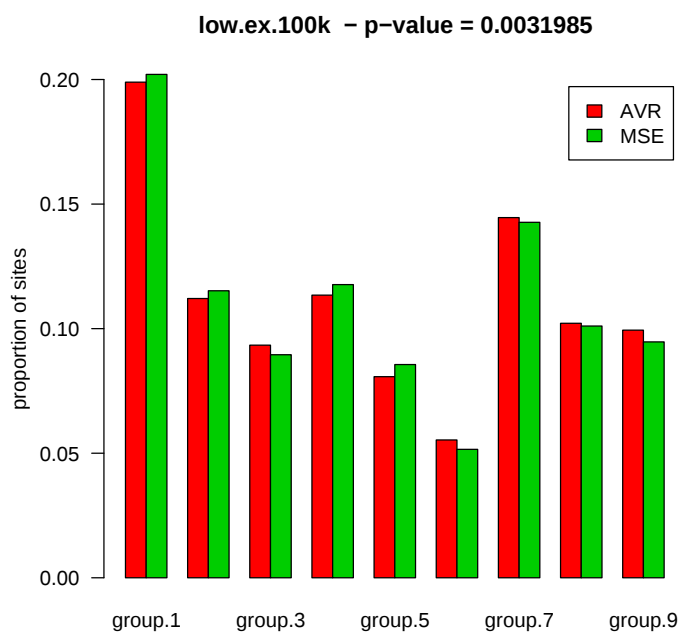
	lower	category	upper
1	0.0e+00	group.1	0.0000060000
2	6.0e-06	group.2	0.0000100000
3	1.0e-05	group.3	0.0000140000
4	1.4e-05	group.4	0.0000200000
5	2.0e-05	group.5	0.0000250000
6	2.5e-05	group.6	0.0000300000
7	3.0e-05	group.7	0.0000420000
8	4.2e-05	group.8	0.0000600000
9	6.0e-05	group.9	0.0002316667



Here are the results for expression density. First, we count just genes that are in the upper half.

Category limits

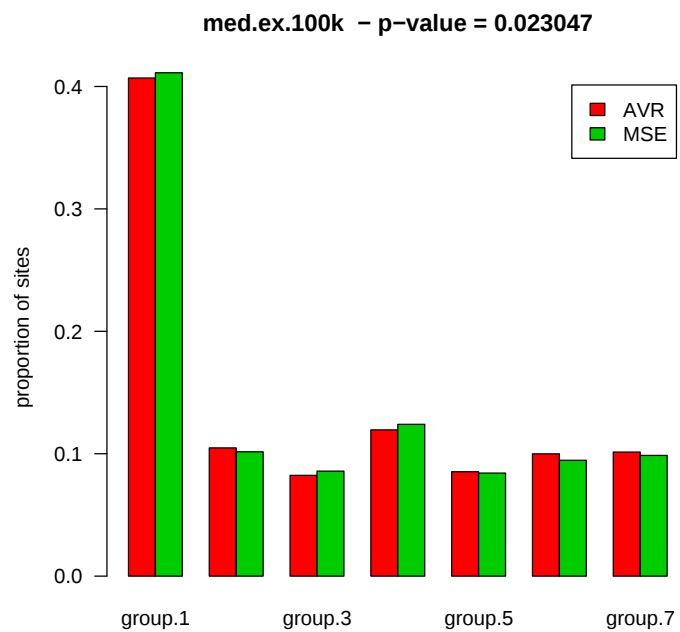
	lower	category	upper
1	0.000000e+00	group.1	1.666667e-06
2	1.666667e-06	group.2	5.000000e-06
3	5.000000e-06	group.3	1.000000e-05
4	1.000000e-05	group.4	1.000000e-05
5	1.000000e-05	group.5	1.500000e-05
6	1.500000e-05	group.6	2.000000e-05
7	2.000000e-05	group.7	3.000000e-05
8	3.000000e-05	group.8	4.000000e-05
9	4.000000e-05	group.9	1.316667e-04



Now we count genes in the upper 1/8th:

Category limits

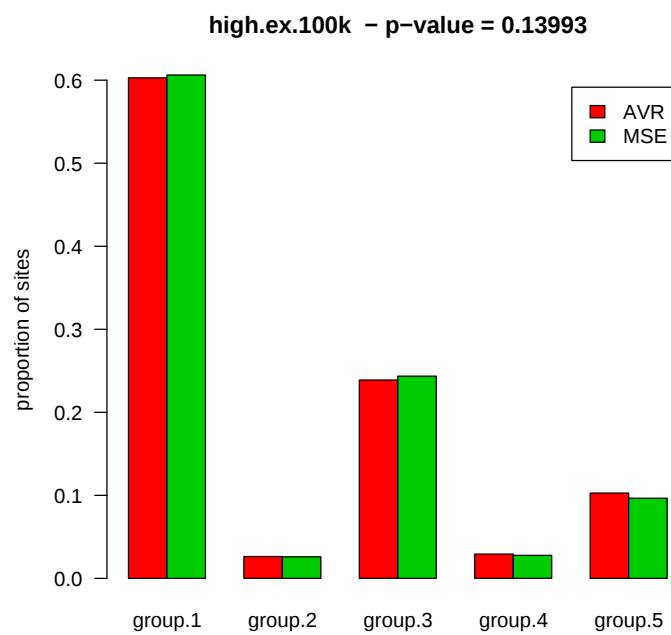
	lower	category	upper
1	0.000000e+00	group.1	3.333333e-06
2	3.333333e-06	group.2	5.000000e-06
3	5.000000e-06	group.3	1.000000e-05
4	1.000000e-05	group.4	1.000000e-05
5	1.000000e-05	group.5	1.500000e-05
6	1.500000e-05	group.6	2.333333e-05
7	2.333333e-05	group.7	9.000000e-05



And here we count genes in the upper 1/16th:

Category limits

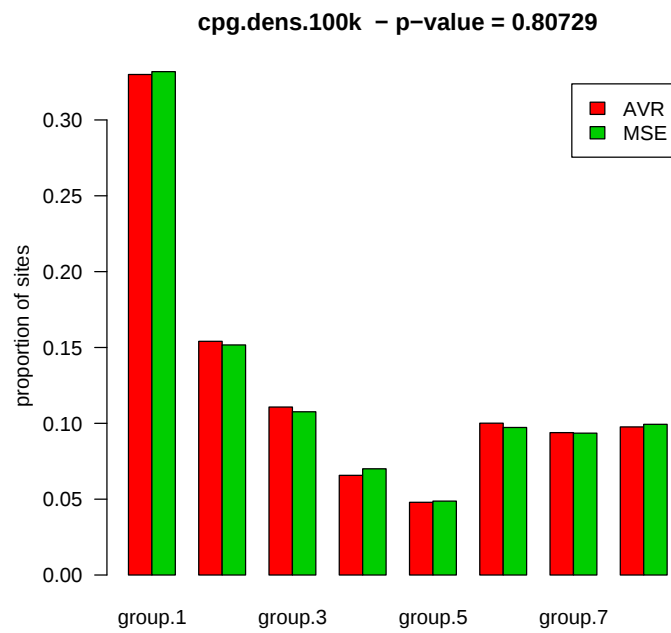
	lower	category	upper
1	0.000000e+00	group.1	3.333333e-06
2	3.333333e-06	group.2	5.000000e-06
3	5.000000e-06	group.3	1.000000e-05
4	1.000000e-05	group.4	1.333333e-05
5	1.333333e-05	group.5	5.666667e-05



Here the effect of density of CpG islands is studied:

Category limits

	lower	category	upper
1	0.0e+00	group.1	5.0e-06
2	5.0e-06	group.2	1.0e-05
3	1.0e-05	group.3	1.5e-05
4	1.5e-05	group.4	2.0e-05
5	2.0e-05	group.5	2.5e-05
6	2.5e-05	group.6	4.0e-05
7	4.0e-05	group.7	6.5e-05
8	6.5e-05	group.8	2.2e-04

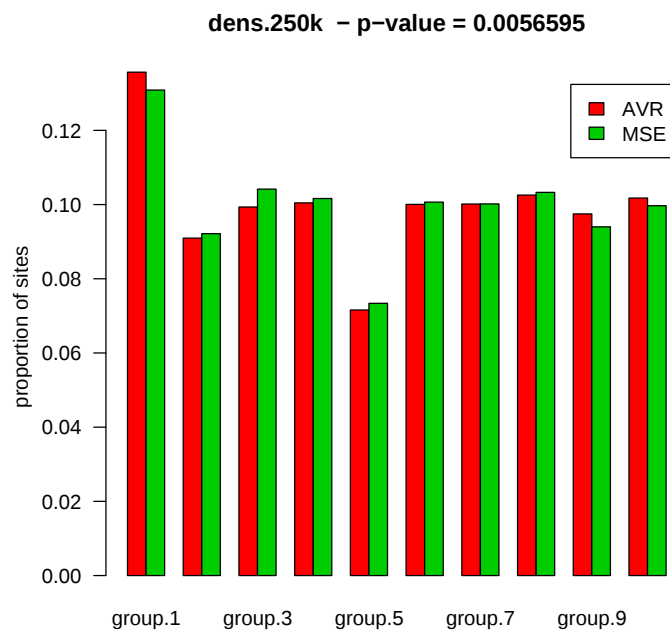


4.4 250 kiloBase Window

In the barplot that follows we examine the association of insertion sites with expression density in a 250 kilobase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category limits

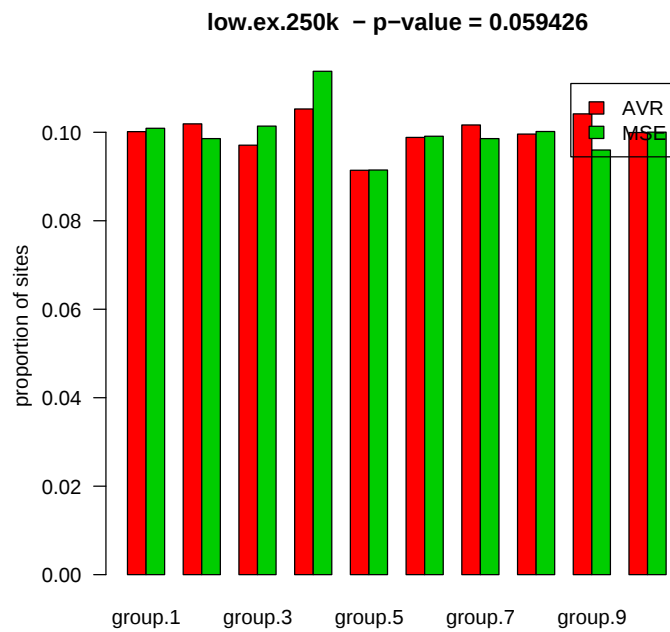
	lower	category	upper
1	0.0e+00	group.1	4.0e-06
2	4.0e-06	group.2	8.0e-06
3	8.0e-06	group.3	1.2e-05
4	1.2e-05	group.4	1.6e-05
5	1.6e-05	group.5	2.0e-05
6	2.0e-05	group.6	2.4e-05
7	2.4e-05	group.7	3.2e-05
8	3.2e-05	group.8	4.0e-05
9	4.0e-05	group.9	5.4e-05
10	5.4e-05	group.10	1.4e-04



Here are the results for expression density. First, we count just genes that are in the upper half.

Category limits

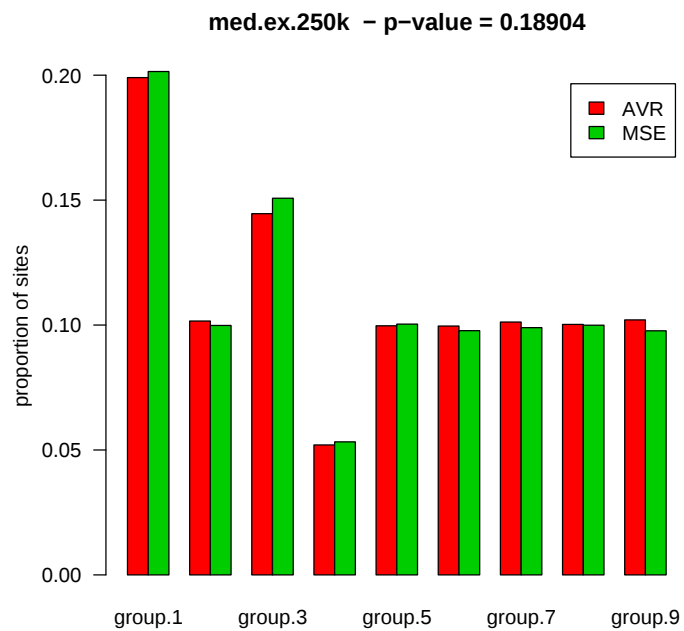
	lower	category	upper
1	0.000000e+00	group.1	6.666667e-07
2	6.666667e-07	group.2	4.000000e-06
3	4.000000e-06	group.3	5.777778e-06
4	5.777778e-06	group.4	8.000000e-06
5	8.000000e-06	group.5	1.100000e-05
6	1.100000e-05	group.6	1.400000e-05
7	1.400000e-05	group.7	1.866667e-05
8	1.866667e-05	group.8	2.401333e-05
9	2.401333e-05	group.9	3.400000e-05
10	3.400000e-05	group.10	8.800000e-05



Now we count genes in the upper $1/8^{th}$:

Category limits

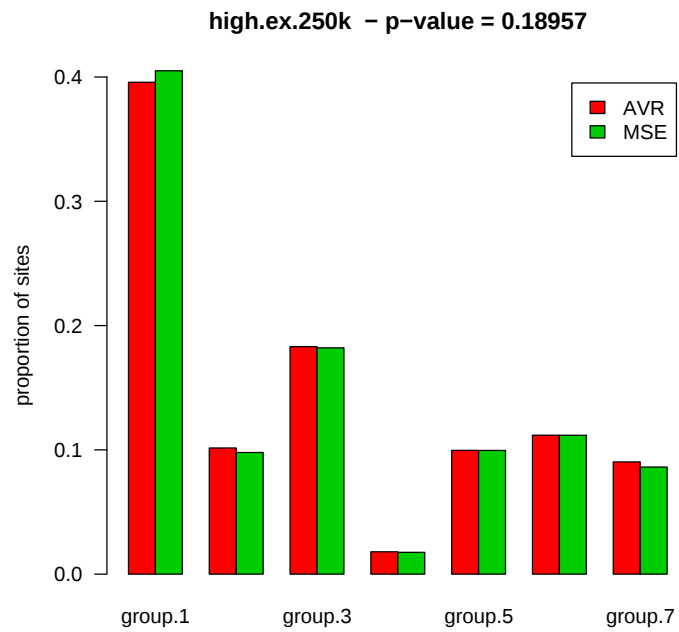
	lower	category	upper
1	0.000000e+00	group.1	1.090909e-06
2	1.090909e-06	group.2	2.400000e-06
3	2.400000e-06	group.3	4.000000e-06
4	4.000000e-06	group.4	5.333333e-06
5	5.333333e-06	group.5	7.333333e-06
6	7.333333e-06	group.6	9.466667e-06
7	9.466667e-06	group.7	1.266667e-05
8	1.266667e-05	group.8	1.800000e-05
9	1.800000e-05	group.9	5.400000e-05



And here we count genes in the upper 1/16th:

Category limits

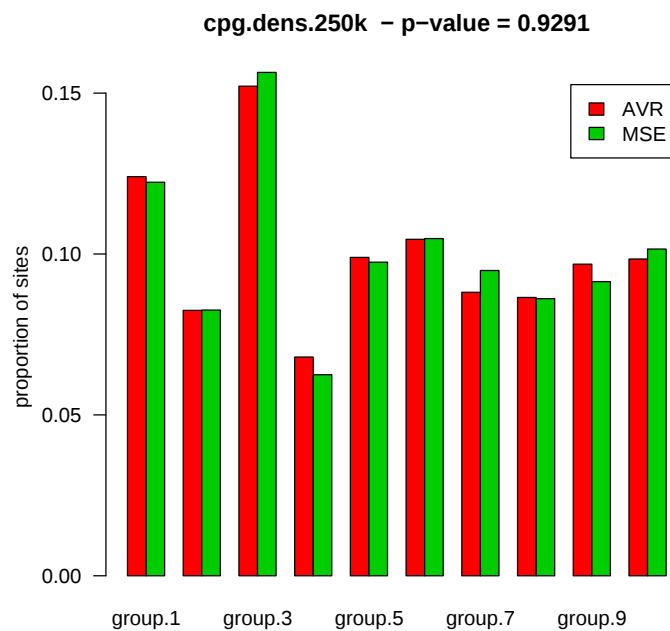
	lower	category	upper
1	0.000000e+00	group.1	1.333333e-06
2	1.333333e-06	group.2	2.000000e-06
3	2.000000e-06	group.3	4.000000e-06
4	4.000000e-06	group.4	4.666667e-06
5	4.666667e-06	group.5	6.666667e-06
6	6.666667e-06	group.6	1.000000e-05
7	1.000000e-05	group.7	3.066667e-05



Here the effect of density of CpG islands is studied:

Category limits

	lower	category	upper
1	0.000000e+00	group.1	2.000000e-06
2	2.000000e-06	group.2	4.000000e-06
3	4.000000e-06	group.3	8.000000e-06
4	8.000000e-06	group.4	1.000000e-05
5	1.000000e-05	group.5	1.400000e-05
6	1.400000e-05	group.6	2.000000e-05
7	2.000000e-05	group.7	2.800000e-05
8	2.800000e-05	group.8	4.000000e-05
9	4.000000e-05	group.9	6.778294e-05
10	6.778294e-05	group.10	2.020000e-04

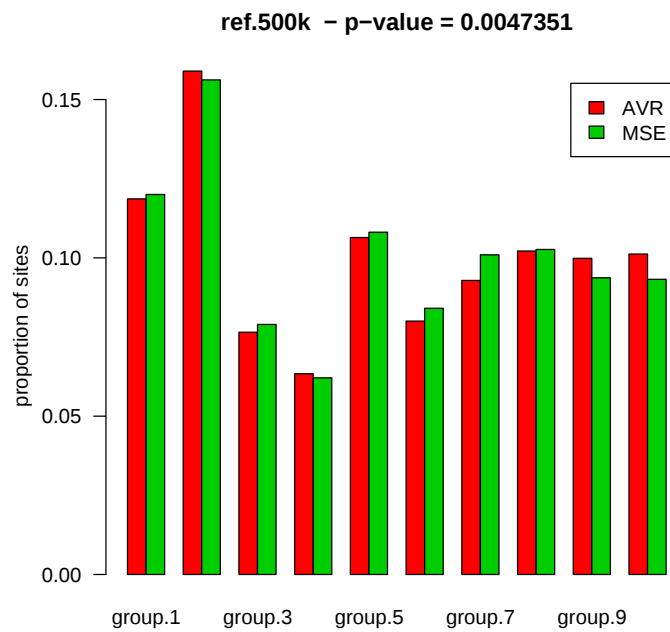


4.5 500 kiloBase Window

Here are gene densities for the various gene calls

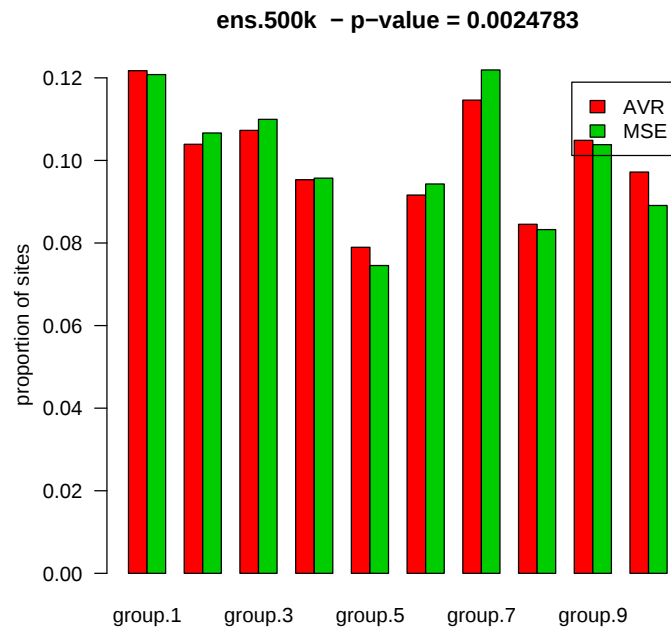
Category limits

	lower	category	upper
1	0.0e+00	group.1	2.0e-06
2	2.0e-06	group.2	6.0e-06
3	6.0e-06	group.3	8.0e-06
4	8.0e-06	group.4	1.0e-05
5	1.0e-05	group.5	1.4e-05
6	1.4e-05	group.6	1.8e-05
7	1.8e-05	group.7	2.4e-05
8	2.4e-05	group.8	3.4e-05
9	3.4e-05	group.9	4.8e-05
10	4.8e-05	group.10	1.5e-04



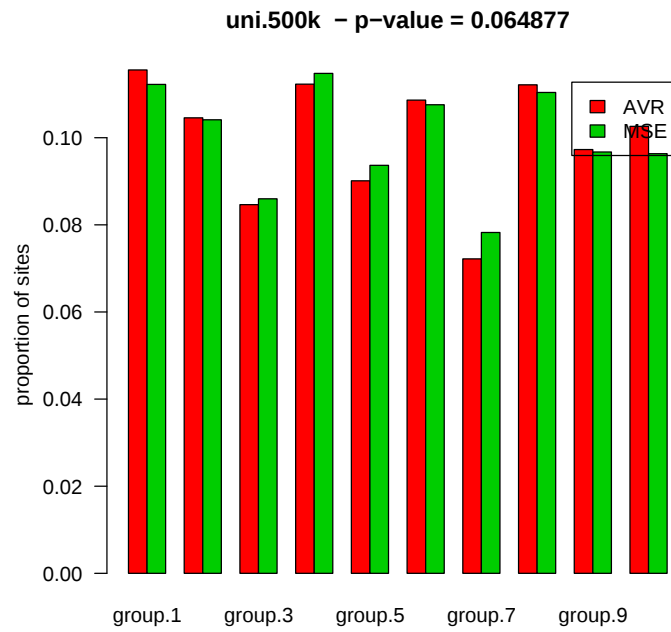
Category limits

	lower	category	upper
1	0.0e+00	group.1	0.000004
2	4.0e-06	group.2	0.000008
3	8.0e-06	group.3	0.000012
4	1.2e-05	group.4	0.000016
5	1.6e-05	group.5	0.000020
6	2.0e-05	group.6	0.000026
7	2.6e-05	group.7	0.000036
8	3.6e-05	group.8	0.000046
9	4.6e-05	group.9	0.000066
10	6.6e-05	group.10	0.000144



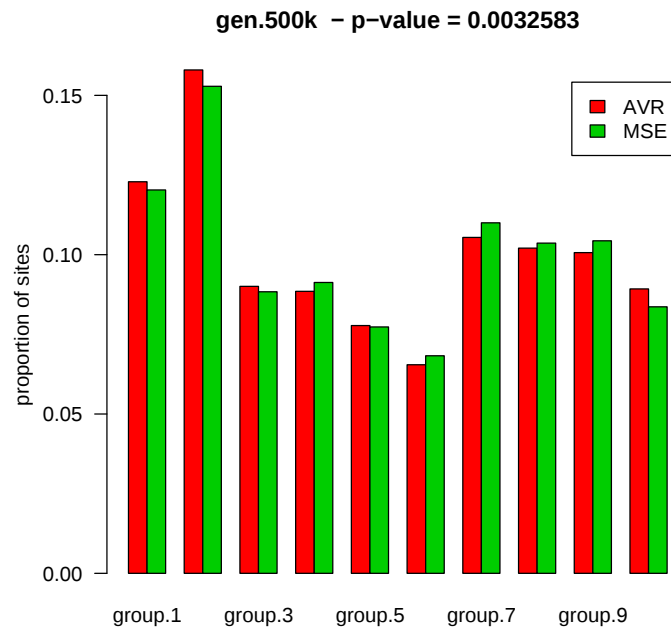
Category limits

	lower	category	upper
1	0.0e+00	group.1	0.000022
2	2.2e-05	group.2	0.000034
3	3.4e-05	group.3	0.000042
4	4.2e-05	group.4	0.000052
5	5.2e-05	group.5	0.000060
6	6.0e-05	group.6	0.000070
7	7.0e-05	group.7	0.000080
8	8.0e-05	group.8	0.000092
9	9.2e-05	group.9	0.000110
10	1.1e-04	group.10	0.000236



Category limits

	lower	category	upper
1	0.0e+00	group.1	8.0e-06
2	8.0e-06	group.2	1.2e-05
3	1.2e-05	group.3	1.4e-05
4	1.4e-05	group.4	1.6e-05
5	1.6e-05	group.5	1.8e-05
6	1.8e-05	group.6	2.0e-05
7	2.0e-05	group.7	2.4e-05
8	2.4e-05	group.8	3.0e-05
9	3.0e-05	group.9	4.0e-05
10	4.0e-05	group.10	7.8e-05

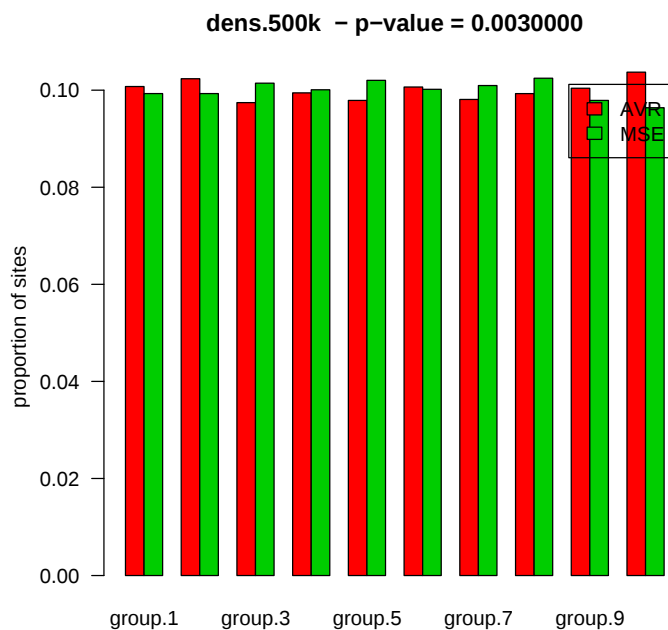


In the barplot

that follows we examine the association of insertion sites with expression density in a 500 kilobase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category limits

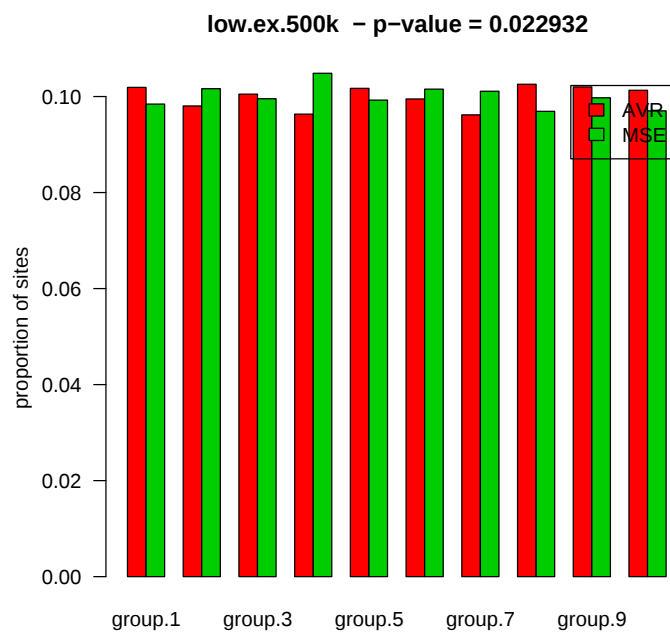
	lower	category	upper
1	0.000000e+00	group.1	4.333333e-06
2	4.333333e-06	group.2	8.000000e-06
3	8.000000e-06	group.3	1.200000e-05
4	1.200000e-05	group.4	1.480000e-05
5	1.480000e-05	group.5	1.800000e-05
6	1.800000e-05	group.6	2.300000e-05
7	2.300000e-05	group.7	2.982667e-05
8	2.982667e-05	group.8	3.800000e-05
9	3.800000e-05	group.9	5.000000e-05
10	5.000000e-05	group.10	1.050000e-04



Here are the results for expression density. First, we count just genes that are in the upper half.

Category limits

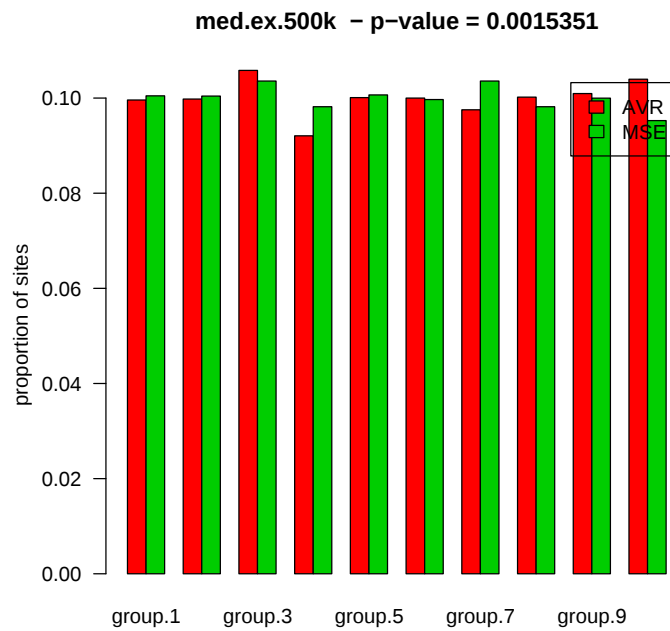
	lower	category	upper
1	0.000000e+00	group.1	1.600000e-06
2	1.600000e-06	group.2	3.638620e-06
3	3.638620e-06	group.3	5.518920e-06
4	5.518920e-06	group.4	7.666667e-06
5	7.666667e-06	group.5	1.000000e-05
6	1.000000e-05	group.6	1.300000e-05
7	1.300000e-05	group.7	1.700000e-05
8	1.700000e-05	group.8	2.220000e-05
9	2.220000e-05	group.9	3.000000e-05
10	3.000000e-05	group.10	6.522988e-05



Now we count genes in the upper 1/8th:

Category limits

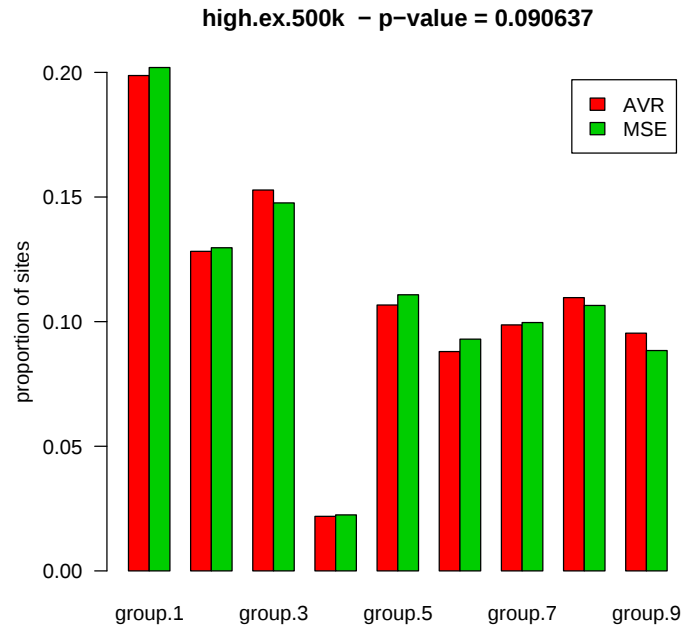
	lower	category	upper
1	0.000000e+00	group.1	1.111111e-07
2	1.111111e-07	group.2	1.466667e-06
3	1.466667e-06	group.3	2.500000e-06
4	2.500000e-06	group.4	3.728889e-06
5	3.728889e-06	group.5	5.000000e-06
6	5.000000e-06	group.6	6.333333e-06
7	6.333333e-06	group.7	8.200000e-06
8	8.200000e-06	group.8	1.116667e-05
9	1.116667e-05	group.9	1.566667e-05
10	1.566667e-05	group.10	3.536667e-05



And here we count genes in the upper 1/16th:

Category limits

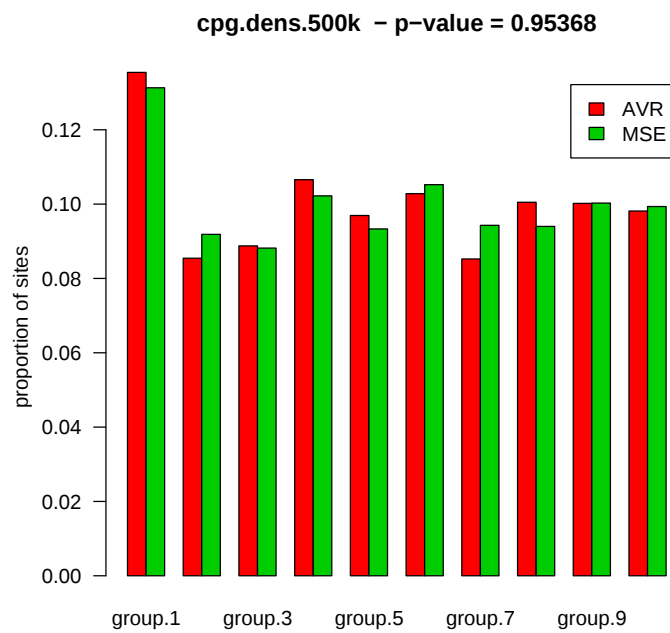
	lower	category	upper
1	0.000000e+00	group.1	2.222222e-07
2	2.222222e-07	group.2	1.000000e-06
3	1.000000e-06	group.3	2.000000e-06
4	2.000000e-06	group.4	2.333333e-06
5	2.333333e-06	group.5	3.000000e-06
6	3.000000e-06	group.6	4.000000e-06
7	4.000000e-06	group.7	5.533333e-06
8	5.533333e-06	group.8	8.000000e-06
9	8.000000e-06	group.9	2.100000e-05



Here the effect of density of CpG islands is studied:

Category limits

	lower	category	upper
1	0.0e+00	group.1	0.000003
2	3.0e-06	group.2	0.000005
3	5.0e-06	group.3	0.000007
4	7.0e-06	group.4	0.000010
5	1.0e-05	group.5	0.000014
6	1.4e-05	group.6	0.000020
7	2.0e-05	group.7	0.000027
8	2.7e-05	group.8	0.000038
9	3.8e-05	group.9	0.000064
10	6.4e-05	group.10	0.000193

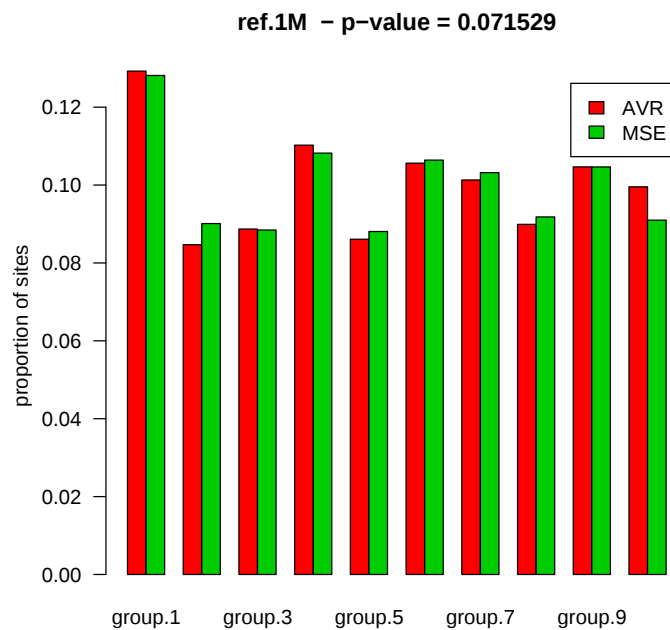


4.6 1 megaBase Window

Here are gene densities for the various gene calls

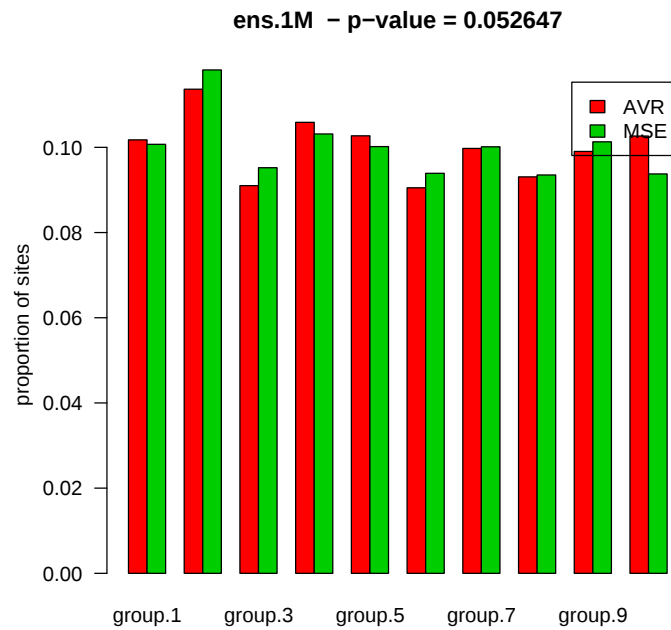
Category limits

	lower	category	upper
1	0.0e+00	group.1	3.0e-06
2	3.0e-06	group.2	5.0e-06
3	5.0e-06	group.3	7.0e-06
4	7.0e-06	group.4	1.0e-05
5	1.0e-05	group.5	1.3e-05
6	1.3e-05	group.6	1.8e-05
7	1.8e-05	group.7	2.5e-05
8	2.5e-05	group.8	3.3e-05
9	3.3e-05	group.9	4.5e-05
10	4.5e-05	group.10	9.8e-05



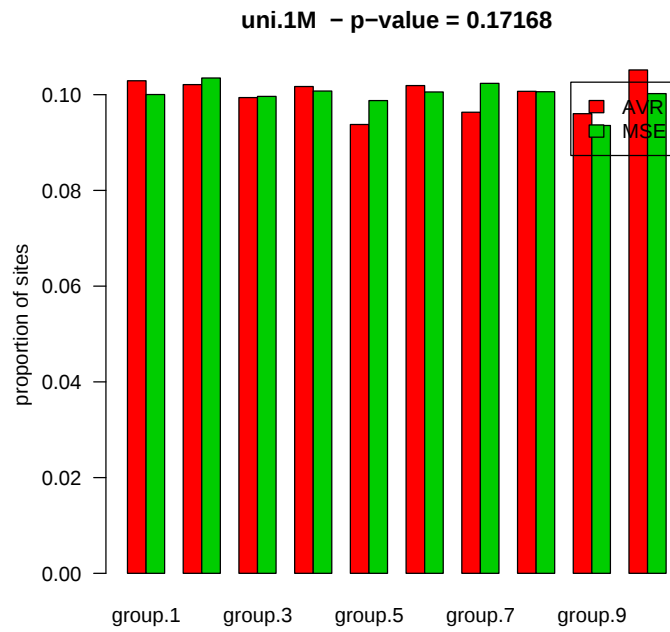
Category limits

	lower	category	upper
1	0.0e+00	group.1	0.0000040000
2	4.0e-06	group.2	0.0000080000
3	8.0e-06	group.3	0.0000110000
4	1.1e-05	group.4	0.0000150000
5	1.5e-05	group.5	0.0000200000
6	2.0e-05	group.6	0.0000260000
7	2.6e-05	group.7	0.0000350000
8	3.5e-05	group.8	0.0000460000
9	4.6e-05	group.9	0.0000610000
10	6.1e-05	group.10	0.0001164267



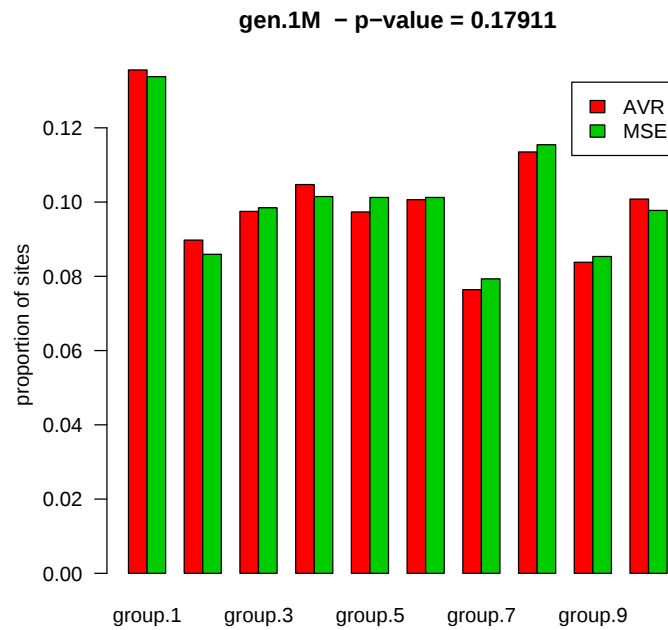
Category limits

	lower	category	upper
1	0.000000	group.1	0.000021
2	0.000021	group.2	0.000031
3	0.000031	group.3	0.000040
4	0.000040	group.4	0.000049
5	0.000049	group.5	0.000058
6	0.000058	group.6	0.000067
7	0.000067	group.7	0.000075
8	0.000075	group.8	0.000086
9	0.000086	group.9	0.000103
10	0.000103	group.10	0.000178



Category limits

	lower	category	upper
1	0.0e+00	group.1	1.000000e-05
2	1.0e-05	group.2	1.200000e-05
3	1.2e-05	group.3	1.400000e-05
4	1.4e-05	group.4	1.600000e-05
5	1.6e-05	group.5	1.800000e-05
6	1.8e-05	group.6	2.100000e-05
7	2.1e-05	group.7	2.400000e-05
8	2.4e-05	group.8	3.000000e-05
9	3.0e-05	group.9	3.600000e-05
10	3.6e-05	group.10	5.692527e-05

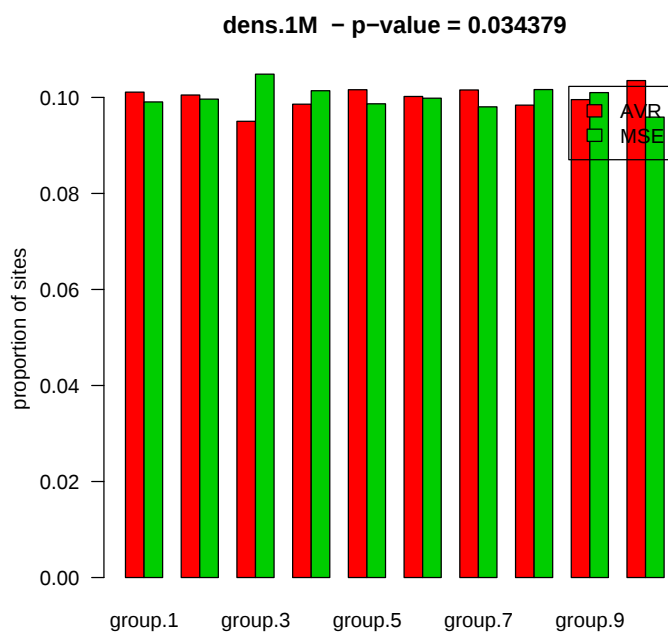


In the barplot

that follows we examine the association of insertion sites with expression density in a 1 megabase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category limits

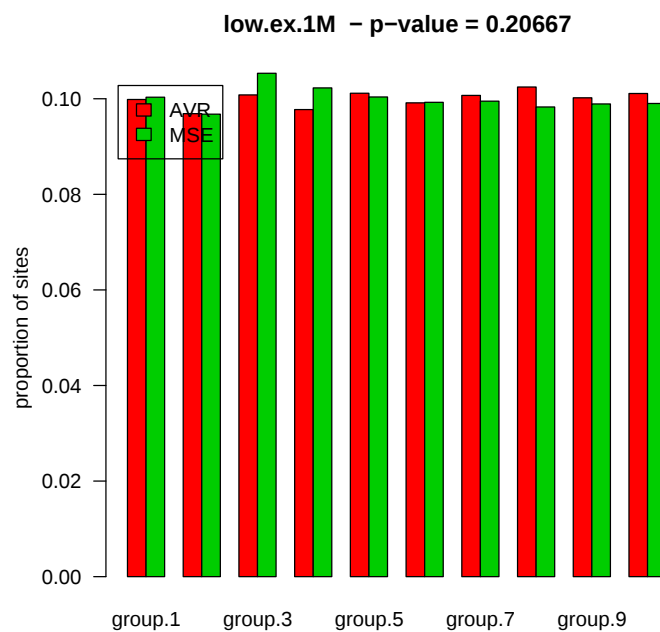
	lower	category	upper
1	0.000000e+00	group.1	5.333333e-06
2	5.333333e-06	group.2	8.333333e-06
3	8.333333e-06	group.3	1.120000e-05
4	1.120000e-05	group.4	1.400000e-05
5	1.400000e-05	group.5	1.733333e-05
6	1.733333e-05	group.6	2.200000e-05
7	2.200000e-05	group.7	2.825000e-05
8	2.825000e-05	group.8	3.650000e-05
9	3.650000e-05	group.9	4.600000e-05
10	4.600000e-05	group.10	7.938757e-05



Here are the results for expression density. First, we count just genes that are in the upper half.

Category limits

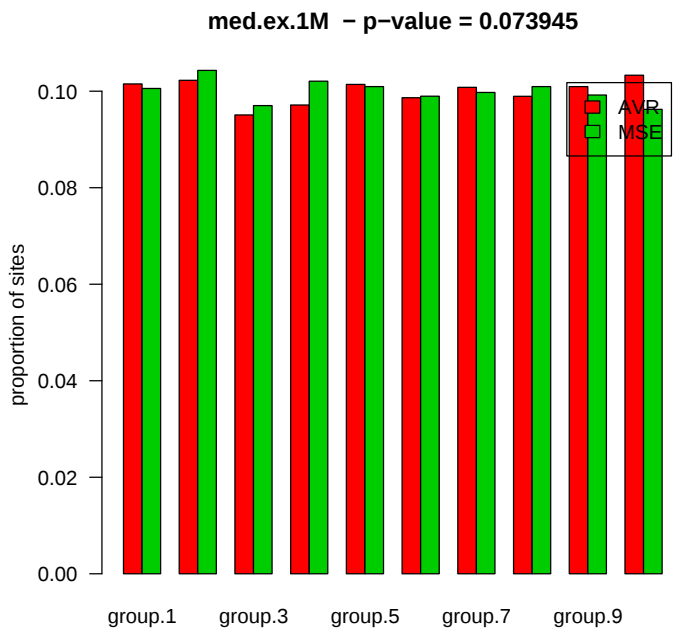
	lower	category	upper
1	0.000000e+00	group.1	1.750000e-06
2	1.750000e-06	group.2	3.500000e-06
3	3.500000e-06	group.3	5.166667e-06
4	5.166667e-06	group.4	6.900000e-06
5	6.900000e-06	group.5	9.133333e-06
6	9.133333e-06	group.6	1.186667e-05
7	1.186667e-05	group.7	1.541667e-05
8	1.541667e-05	group.8	2.066667e-05
9	2.066667e-05	group.9	2.733333e-05
10	2.733333e-05	group.10	5.200000e-05



Now we count genes in the upper 1/8th:

Category limits

	lower	category	upper
1	0.000000e+00	group.1	6.666667e-07
2	6.666667e-07	group.2	1.500000e-06
3	1.500000e-06	group.3	2.416667e-06
4	2.416667e-06	group.4	3.333333e-06
5	3.333333e-06	group.5	4.333333e-06
6	4.333333e-06	group.6	5.733333e-06
7	5.733333e-06	group.7	7.583333e-06
8	7.583333e-06	group.8	1.000000e-05
9	1.000000e-05	group.9	1.366667e-05
10	1.366667e-05	group.10	2.744346e-05

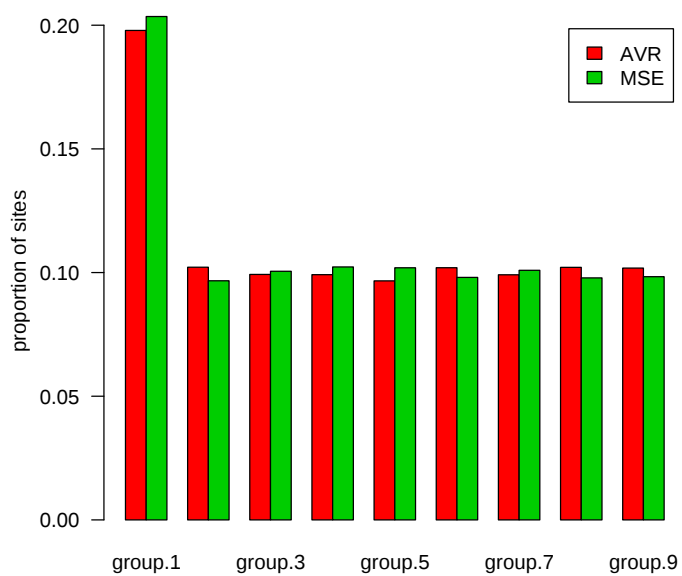


And here we count genes in the upper 1/16th:

Category limits

	lower	category	upper
1	0.000000e+00	group.1	6.000000e-07
2	6.000000e-07	group.2	1.100000e-06
3	1.100000e-06	group.3	1.566667e-06
4	1.566667e-06	group.4	2.166667e-06
5	2.166667e-06	group.5	2.833333e-06
6	2.833333e-06	group.6	3.616667e-06
7	3.616667e-06	group.7	4.833333e-06
8	4.833333e-06	group.8	6.666667e-06
9	6.666667e-06	group.9	1.507143e-05

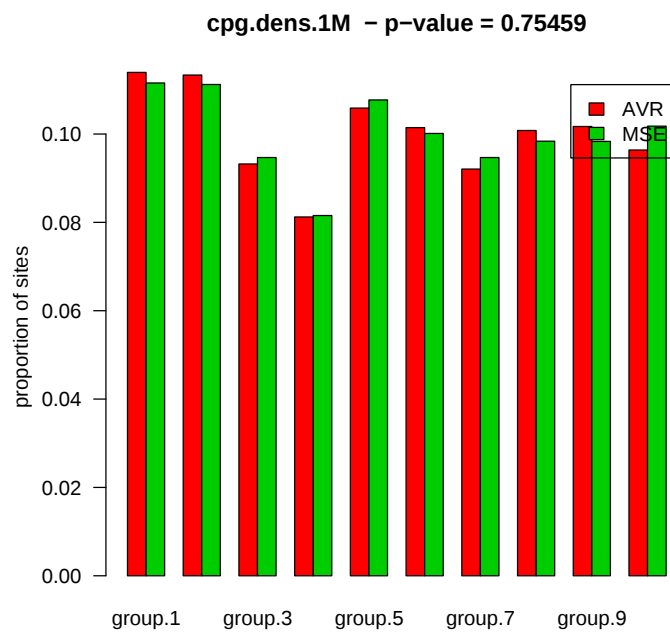
high.ex.1M - p-value = 0.27191



Here the effect of density of CpG islands is studied:

Category limits

	lower	category	upper
1	0.0e+00	group.1	0.0000030
2	3.0e-06	group.2	0.0000050
3	5.0e-06	group.3	0.0000070
4	7.0e-06	group.4	0.0000095
5	9.5e-06	group.5	0.0000140
6	1.4e-05	group.6	0.0000190
7	1.9e-05	group.7	0.0000250
8	2.5e-05	group.8	0.0000350
9	3.5e-05	group.9	0.0000590
10	5.9e-05	group.10	0.0001725

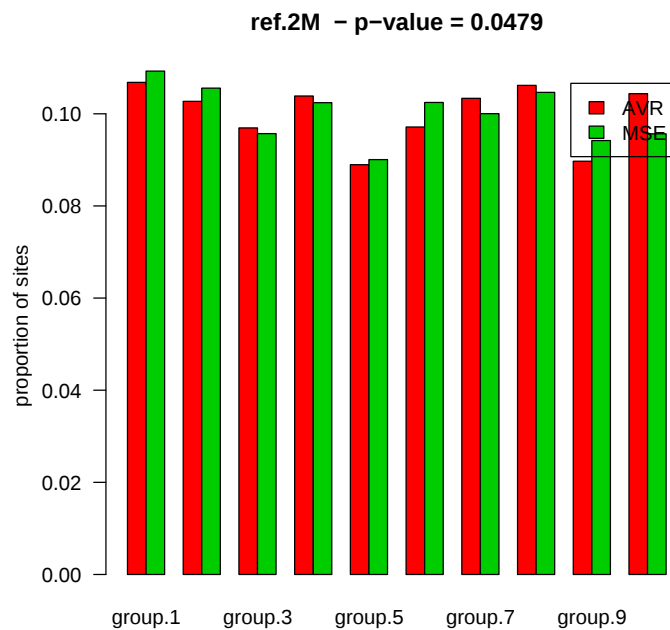


4.7 2 megaBase Window

Here are gene densities for the various gene calls

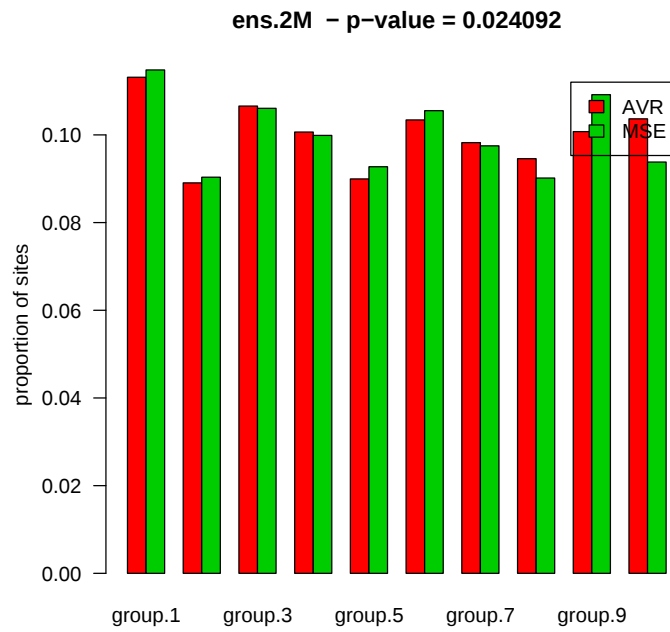
Category limits

	lower	category	upper
1	0.00e+00	group.1	3.00e-06
2	3.00e-06	group.2	5.00e-06
3	5.00e-06	group.3	7.00e-06
4	7.00e-06	group.4	9.50e-06
5	9.50e-06	group.5	1.25e-05
6	1.25e-05	group.6	1.80e-05
7	1.80e-05	group.7	2.40e-05
8	2.40e-05	group.8	3.10e-05
9	3.10e-05	group.9	3.95e-05
10	3.95e-05	group.10	6.85e-05



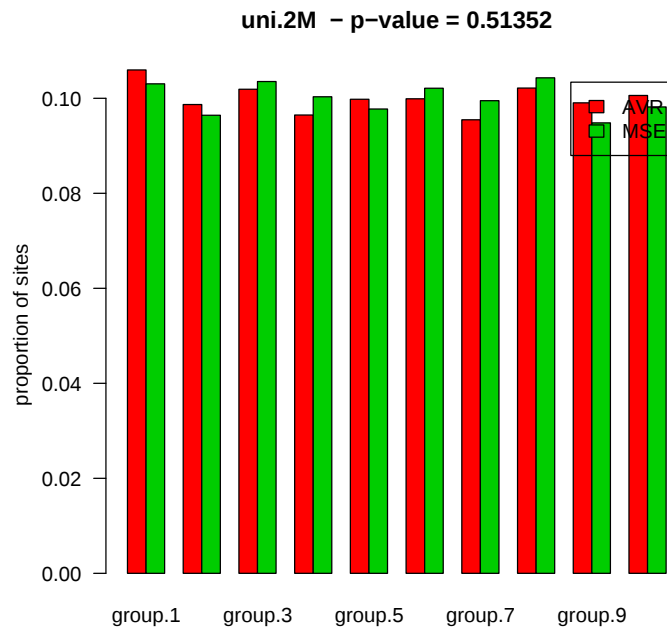
Category limits

	lower	category	upper
1	0.00e+00	group.1	5.000000e-06
2	5.00e-06	group.2	7.500000e-06
3	7.50e-06	group.3	1.050000e-05
4	1.05e-05	group.4	1.400000e-05
5	1.40e-05	group.5	1.850000e-05
6	1.85e-05	group.6	2.600000e-05
7	2.60e-05	group.7	3.350000e-05
8	3.35e-05	group.8	4.200000e-05
9	4.20e-05	group.9	5.400000e-05
10	5.40e-05	group.10	9.764105e-05



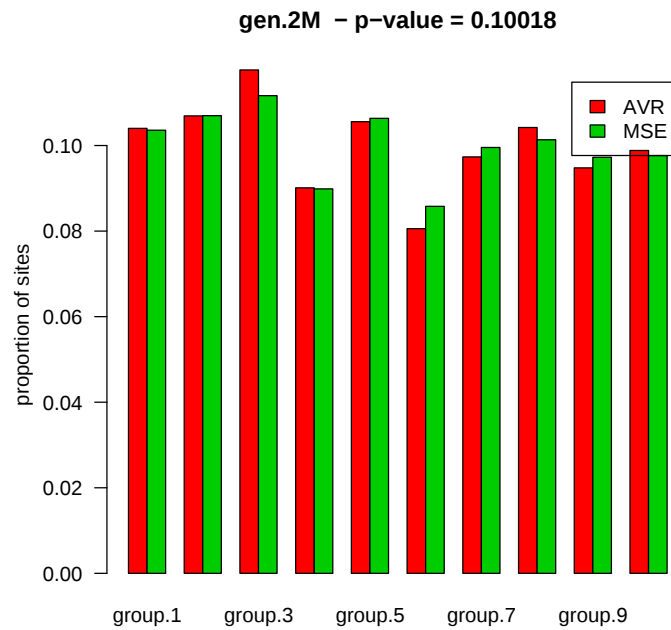
Category limits

	lower	category	upper
1	0.00e+00	group.1	0.0000215000
2	2.15e-05	group.2	0.0000295000
3	2.95e-05	group.3	0.0000375000
4	3.75e-05	group.4	0.0000465000
5	4.65e-05	group.5	0.0000555000
6	5.55e-05	group.6	0.0000635000
7	6.35e-05	group.7	0.0000710000
8	7.10e-05	group.8	0.0000815000
9	8.15e-05	group.9	0.0000950000
10	9.50e-05	group.10	0.0001355506



Category limits

	lower	category	upper
1	5.00e-07	group.1	1.000000e-05
2	1.00e-05	group.2	1.250000e-05
3	1.25e-05	group.3	1.450000e-05
4	1.45e-05	group.4	1.600000e-05
5	1.60e-05	group.5	1.800000e-05
6	1.80e-05	group.6	2.000000e-05
7	2.00e-05	group.7	2.350000e-05
8	2.35e-05	group.8	2.800000e-05
9	2.80e-05	group.9	3.550000e-05
10	3.55e-05	group.10	5.281451e-05

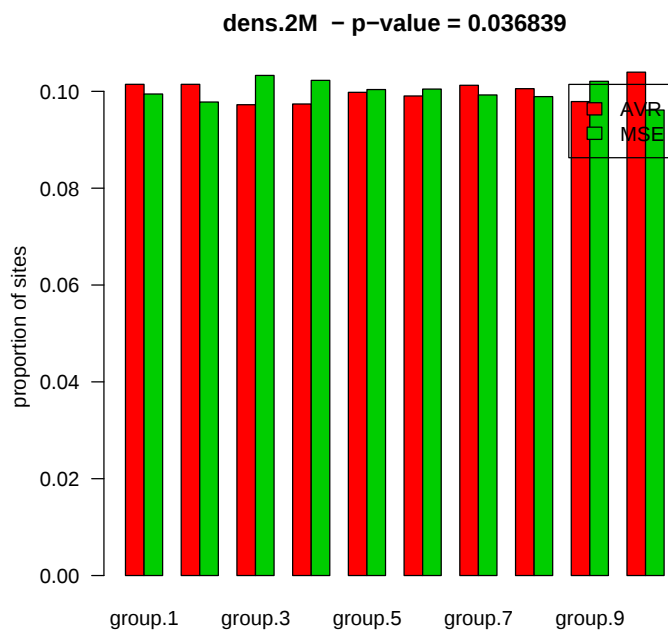


In the barplot

that follows we examine the association of insertion sites with expression density in a 2 megabase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category limits

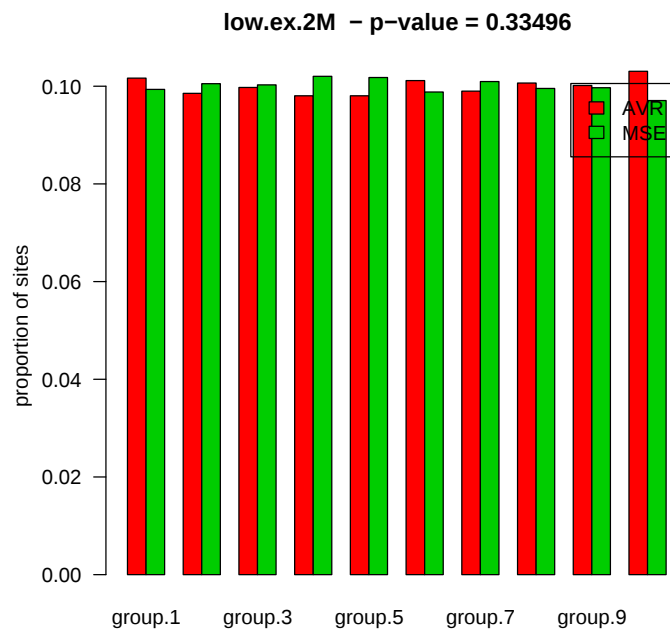
	lower	category	upper
1	0.000000e+00	group.1	5.500000e-06
2	5.500000e-06	group.2	8.000000e-06
3	8.000000e-06	group.3	1.050000e-05
4	1.050000e-05	group.4	1.325000e-05
5	1.325000e-05	group.5	1.655000e-05
6	1.655000e-05	group.6	2.093333e-05
7	2.093333e-05	group.7	2.688333e-05
8	2.688333e-05	group.8	3.383333e-05
9	3.383333e-05	group.9	4.200000e-05
10	4.200000e-05	group.10	6.726246e-05



Here are the results for expression density. First, we count just genes that are in the upper half.

Category limits

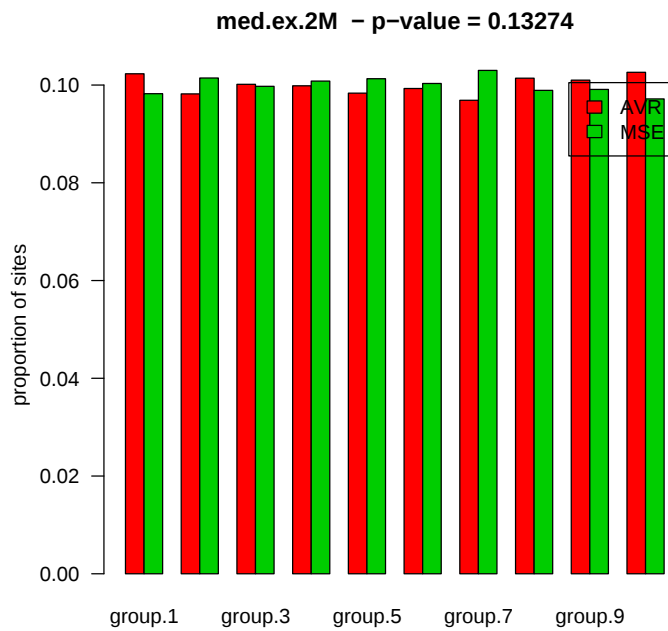
	lower	category	upper
1	0.000000e+00	group.1	1.916667e-06
2	1.916667e-06	group.2	3.250000e-06
3	3.250000e-06	group.3	4.583333e-06
4	4.583333e-06	group.4	6.100000e-06
5	6.100000e-06	group.5	8.433333e-06
6	8.433333e-06	group.6	1.115333e-05
7	1.115333e-05	group.7	1.414905e-05
8	1.414905e-05	group.8	1.843333e-05
9	1.843333e-05	group.9	2.357158e-05
10	2.357158e-05	group.10	4.175833e-05



Now we count genes in the upper 1/8th:

Category limits

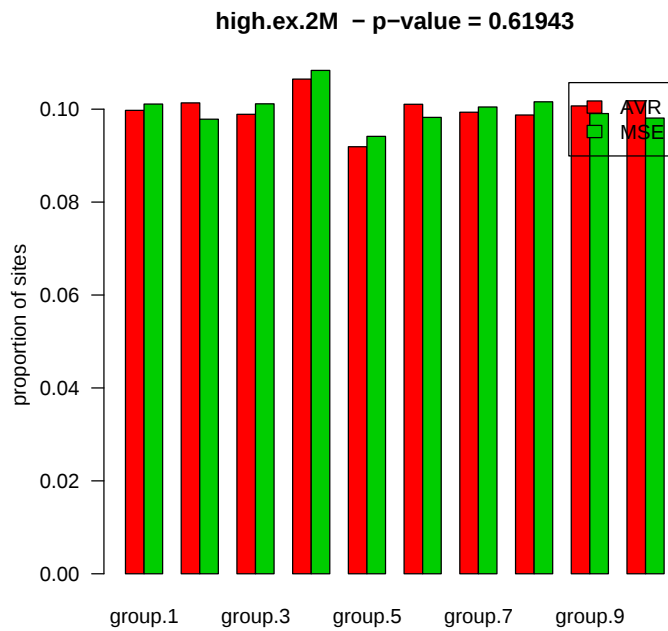
	lower	category	upper
1	0.000000e+00	group.1	7.805556e-07
2	7.805556e-07	group.2	1.500000e-06
3	1.500000e-06	group.3	2.208333e-06
4	2.208333e-06	group.4	2.908333e-06
5	2.908333e-06	group.5	3.800000e-06
6	3.800000e-06	group.6	5.279041e-06
7	5.279041e-06	group.7	6.958333e-06
8	6.958333e-06	group.8	8.825000e-06
9	8.825000e-06	group.9	1.142500e-05
10	1.142500e-05	group.10	2.183333e-05



And here we count genes in the upper $1/16^{th}$:

Category limits

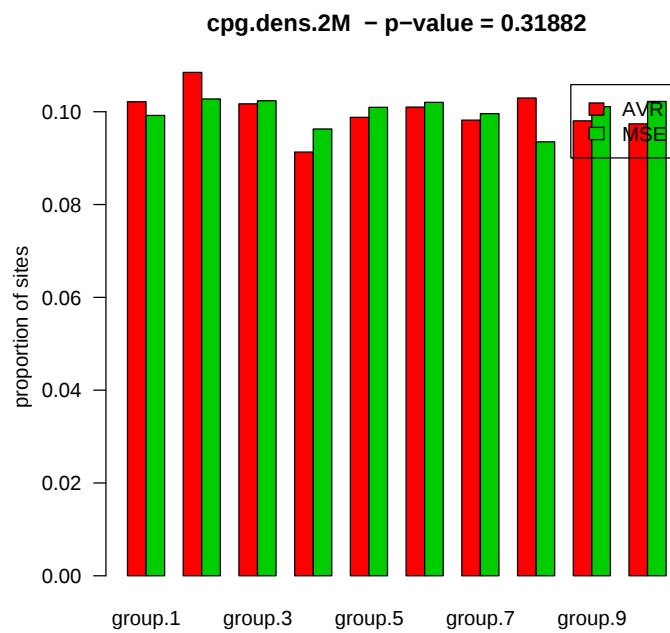
	lower	category	upper
1	0.000000e+00	group.1	3.000000e-07
2	3.000000e-07	group.2	7.320136e-07
3	7.320136e-07	group.3	1.112500e-06
4	1.112500e-06	group.4	1.500000e-06
5	1.500000e-06	group.5	1.958333e-06
6	1.958333e-06	group.6	2.583333e-06
7	2.583333e-06	group.7	3.266667e-06
8	3.266667e-06	group.8	4.250000e-06
9	4.250000e-06	group.9	5.500000e-06
10	5.500000e-06	group.10	1.275000e-05



Here the effect of density of CpG islands is studied:

Category limits

	lower	category	upper
1	0.000e+00	group.1	0.0000030000
2	3.000e-06	group.2	0.0000047500
3	4.750e-06	group.3	0.0000067500
4	6.750e-06	group.4	0.0000092500
5	9.250e-06	group.5	0.0000130000
6	1.300e-05	group.6	0.0000175000
7	1.750e-05	group.7	0.0000235000
8	2.350e-05	group.8	0.0000305000
9	3.050e-05	group.9	0.0000532500
10	5.325e-05	group.10	0.0001517623

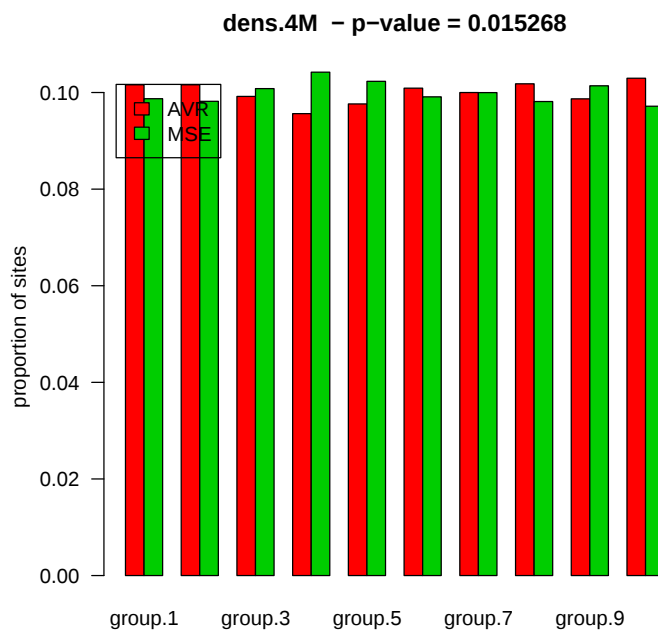


4.8 4 megaBase Window

In the barplot that follows we examine the association of insertion sites with expression density in a 4 megabase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category limits

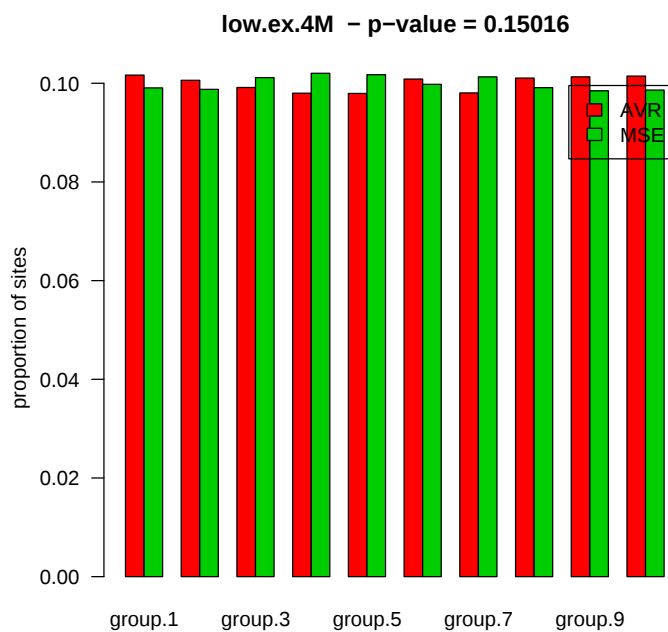
	lower	category	upper
1	1.875000e-07	group.1	5.750000e-06
2	5.750000e-06	group.2	7.886349e-06
3	7.886349e-06	group.3	9.827404e-06
4	9.827404e-06	group.4	1.233333e-05
5	1.233333e-05	group.5	1.580595e-05
6	1.580595e-05	group.6	1.965167e-05
7	1.965167e-05	group.7	2.424931e-05
8	2.424931e-05	group.8	2.962500e-05
9	2.962500e-05	group.9	3.710196e-05
10	3.710196e-05	group.10	6.624884e-05



Here are the results for expression density. First, we count just genes that are in the upper half.

Category limits

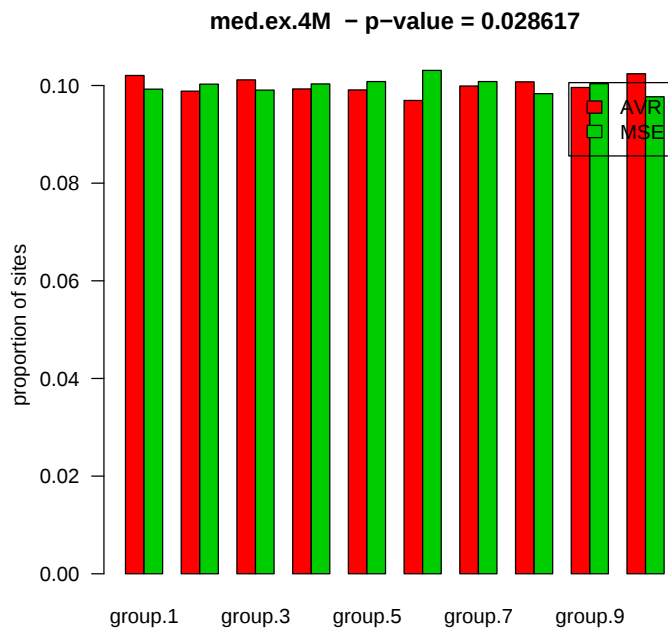
	lower	category	upper
1	0.000000e+00	group.1	2.041667e-06
2	2.041667e-06	group.2	3.078571e-06
3	3.078571e-06	group.3	4.159524e-06
4	4.159524e-06	group.4	5.566667e-06
5	5.566667e-06	group.5	7.654167e-06
6	7.654167e-06	group.6	9.970238e-06
7	9.970238e-06	group.7	1.272083e-05
8	1.272083e-05	group.8	1.613750e-05
9	1.613750e-05	group.9	1.973547e-05
10	1.973547e-05	group.10	3.458048e-05



Now we count genes in the upper 1/8th:

Category limits

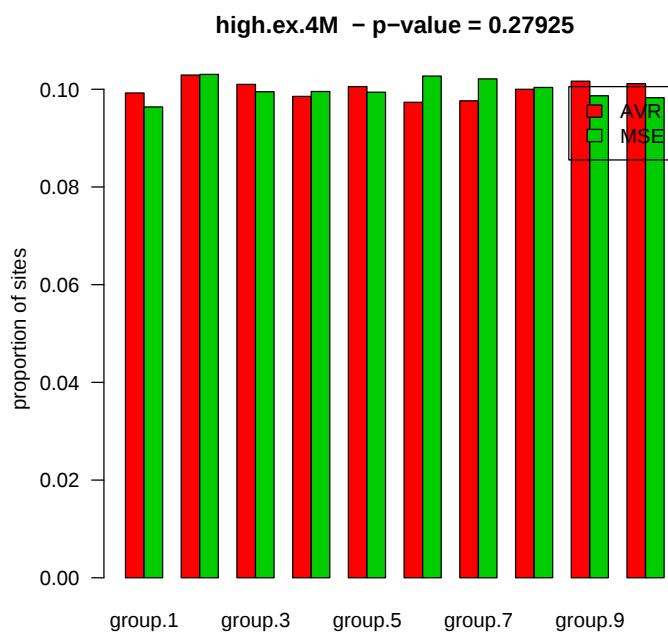
	lower	category	upper
1	0.000000e+00	group.1	8.750000e-07
2	8.750000e-07	group.2	1.487500e-06
3	1.487500e-06	group.3	2.015476e-06
4	2.015476e-06	group.4	2.687500e-06
5	2.687500e-06	group.5	3.600000e-06
6	3.600000e-06	group.6	4.829762e-06
7	4.829762e-06	group.7	6.183333e-06
8	6.183333e-06	group.8	7.344821e-06
9	7.344821e-06	group.9	9.335368e-06
10	9.335368e-06	group.10	1.852734e-05



And here we count genes in the upper 1/16th:

Category limits

	lower	category	upper
1	0.000000e+00	group.1	4.166667e-07
2	4.166667e-07	group.2	7.500000e-07
3	7.500000e-07	group.3	1.020833e-06
4	1.020833e-06	group.4	1.383333e-06
5	1.383333e-06	group.5	1.813104e-06
6	1.813104e-06	group.6	2.273810e-06
7	2.273810e-06	group.7	2.893714e-06
8	2.893714e-06	group.8	3.625000e-06
9	3.625000e-06	group.9	4.500000e-06
10	4.500000e-06	group.10	9.266667e-06



Here the effect of density of CpG islands is studied:

Category limits

	lower	category	upper
1	2.500000e-07	group.1	3.375000e-06
2	3.375000e-06	group.2	4.750000e-06
3	4.750000e-06	group.3	6.625000e-06
4	6.625000e-06	group.4	8.875000e-06
5	8.875000e-06	group.5	1.187500e-05
6	1.187500e-05	group.6	1.593864e-05
7	1.593864e-05	group.7	2.075000e-05
8	2.075000e-05	group.8	2.862500e-05
9	2.862500e-05	group.9	4.650000e-05
10	4.650000e-05	group.10	1.280920e-04

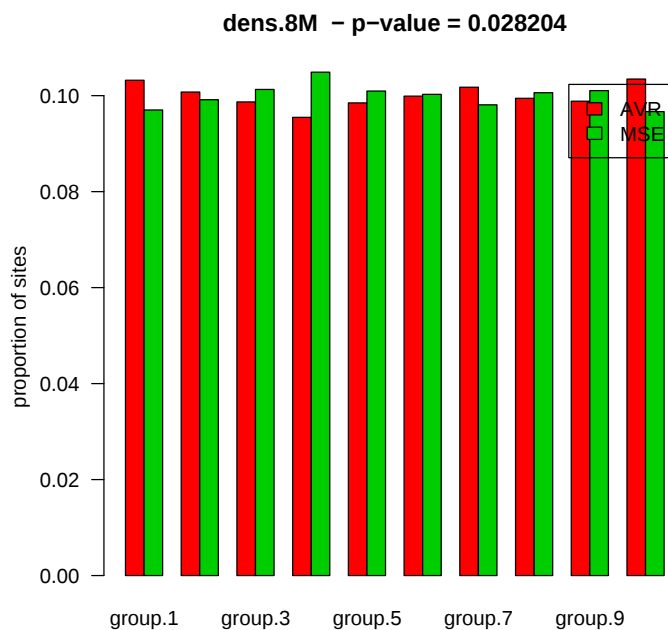


4.9 8 megaBase Window

In the barplot that follows we examine the association of insertion sites with expression density in a 8 megabase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category limits

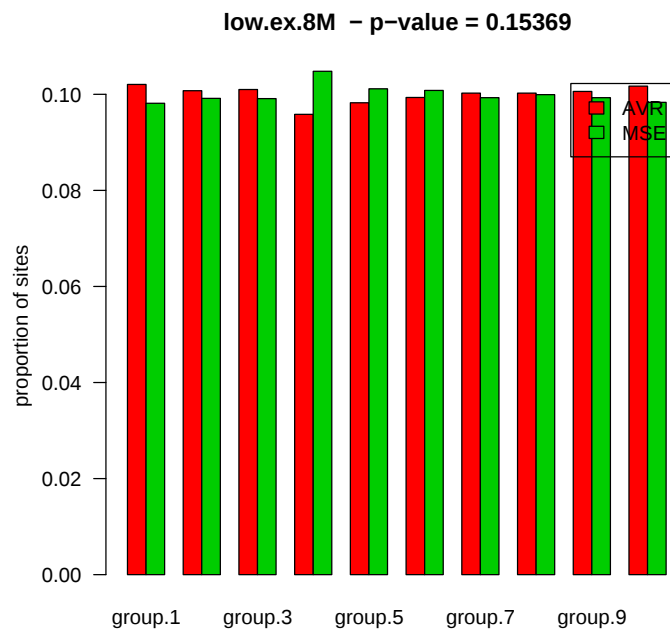
	lower	category	upper
1	9.375000e-08	group.1	5.922917e-06
2	5.922917e-06	group.2	7.870720e-06
3	7.870720e-06	group.3	9.640833e-06
4	9.640833e-06	group.4	1.172024e-05
5	1.172024e-05	group.5	1.460786e-05
6	1.460786e-05	group.6	1.796250e-05
7	1.796250e-05	group.7	2.123904e-05
8	2.123904e-05	group.8	2.610000e-05
9	2.610000e-05	group.9	3.350833e-05
10	3.350833e-05	group.10	5.707167e-05



Here are the results for expression density. First, we count just genes that are in the upper half.

Category limits

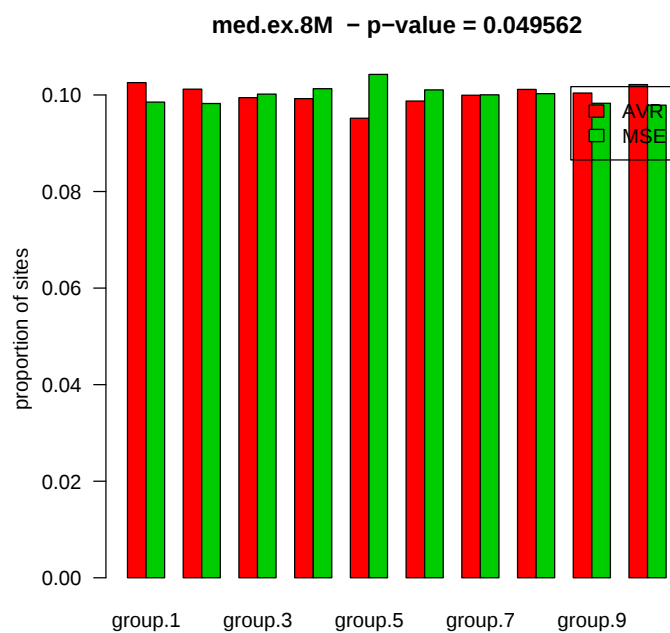
	lower	category	upper
1	0.000000e+00	group.1	2.175000e-06
2	2.175000e-06	group.2	3.065833e-06
3	3.065833e-06	group.3	4.085417e-06
4	4.085417e-06	group.4	5.194792e-06
5	5.194792e-06	group.5	6.748214e-06
6	6.748214e-06	group.6	8.781052e-06
7	8.781052e-06	group.7	1.114715e-05
8	1.114715e-05	group.8	1.387500e-05
9	1.387500e-05	group.9	1.740000e-05
10	1.740000e-05	group.10	2.995075e-05



Now we count genes in the upper 1/8th:

Category limits

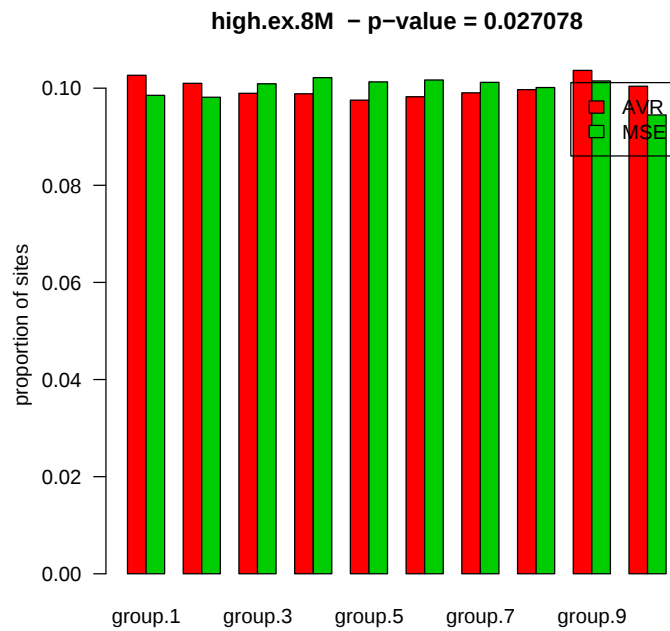
	lower	category	upper
1	0.000000e+00	group.1	1.034821e-06
2	1.034821e-06	group.2	1.470833e-06
3	1.470833e-06	group.3	1.943959e-06
4	1.943959e-06	group.4	2.531525e-06
5	2.531525e-06	group.5	3.283631e-06
6	3.283631e-06	group.6	4.203019e-06
7	4.203019e-06	group.7	5.233125e-06
8	5.233125e-06	group.8	6.204167e-06
9	6.204167e-06	group.9	7.895698e-06
10	7.895698e-06	group.10	1.398827e-05



And here we count genes in the upper 1/16th:

Category limits

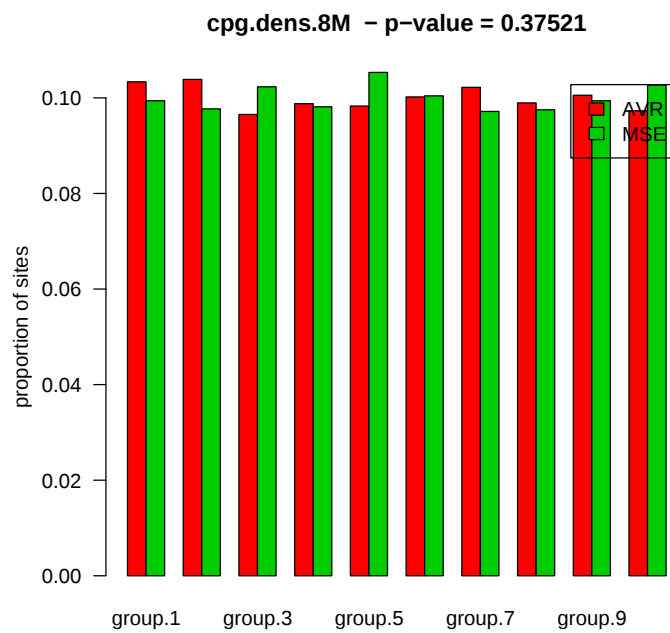
	lower	category	upper
1	0.000000e+00	group.1	5.104167e-07
2	5.104167e-07	group.2	7.708333e-07
3	7.708333e-07	group.3	1.011756e-06
4	1.011756e-06	group.4	1.271875e-06
5	1.271875e-06	group.5	1.635001e-06
6	1.635001e-06	group.6	1.972064e-06
7	1.972064e-06	group.7	2.450000e-06
8	2.450000e-06	group.8	2.968750e-06
9	2.968750e-06	group.9	3.924107e-06
10	3.924107e-06	group.10	6.725595e-06



Here the effect of density of CpG islands is studied:

Category limits

	lower	category	upper
1	2.387202e-07	group.1	3.750000e-06
2	3.750000e-06	group.2	5.125000e-06
3	5.125000e-06	group.3	6.625000e-06
4	6.625000e-06	group.4	8.422707e-06
5	8.422707e-06	group.5	1.050000e-05
6	1.050000e-05	group.6	1.431250e-05
7	1.431250e-05	group.7	1.837500e-05
8	1.837500e-05	group.8	2.515014e-05
9	2.515014e-05	group.9	3.563304e-05
10	3.563304e-05	group.10	9.320273e-05

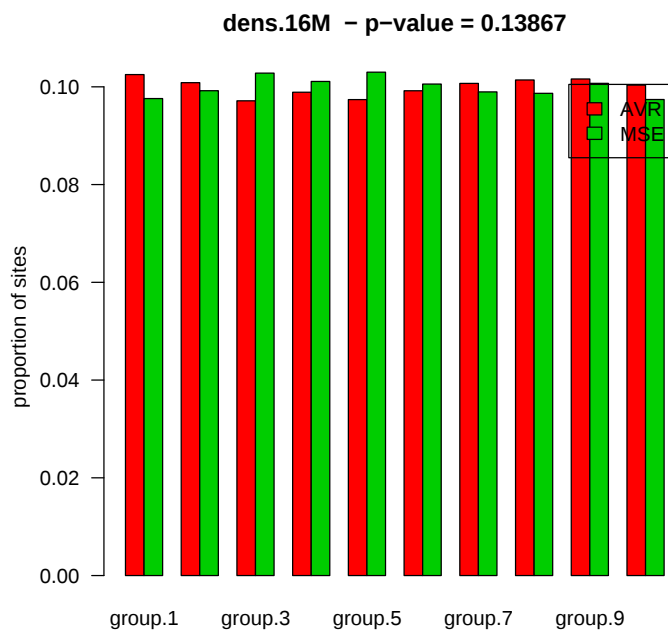


4.10 16 megaBase Window

In the barplot that follows we examine the association of insertion sites with expression density in a 16 megabase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category limits

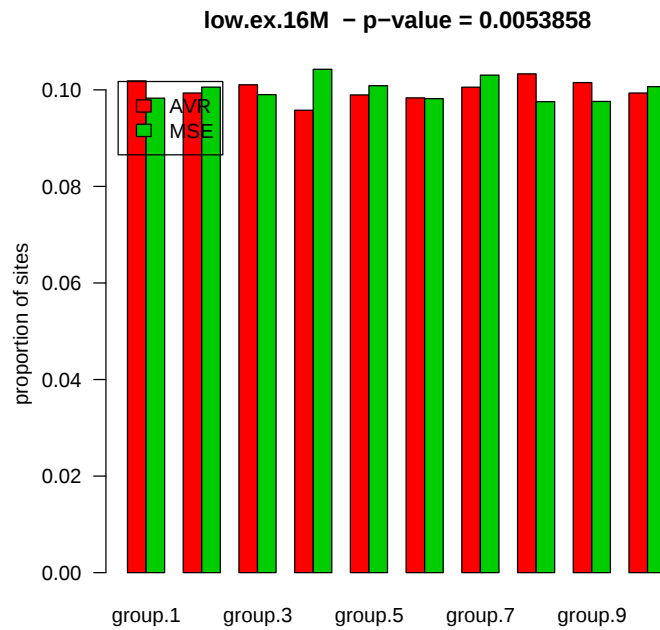
	lower	category	upper
1	5.968004e-08	group.1	6.252812e-06
2	6.252812e-06	group.2	7.813021e-06
3	7.813021e-06	group.3	9.422024e-06
4	9.422024e-06	group.4	1.110833e-05
5	1.110833e-05	group.5	1.315521e-05
6	1.315521e-05	group.6	1.571875e-05
7	1.571875e-05	group.7	1.841656e-05
8	1.841656e-05	group.8	2.204626e-05
9	2.204626e-05	group.9	2.805104e-05
10	2.805104e-05	group.10	4.420392e-05



Here are the results for expression density. First, we count just genes that are in the upper half.

Category limits

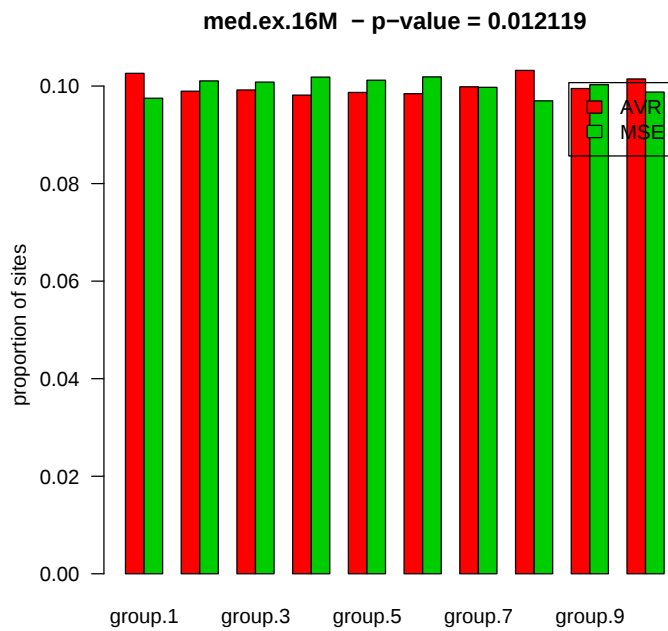
	lower	category	upper
1	0.000000e+00	group.1	2.342708e-06
2	2.342708e-06	group.2	3.102902e-06
3	3.102902e-06	group.3	3.924388e-06
4	3.924388e-06	group.4	4.997024e-06
5	4.997024e-06	group.5	6.129762e-06
6	6.129762e-06	group.6	7.548363e-06
7	7.548363e-06	group.7	9.070313e-06
8	9.070313e-06	group.8	1.137326e-05
9	1.137326e-05	group.9	1.482908e-05
10	1.482908e-05	group.10	2.194023e-05



Now we count genes in the upper 1/8th:

Category limits

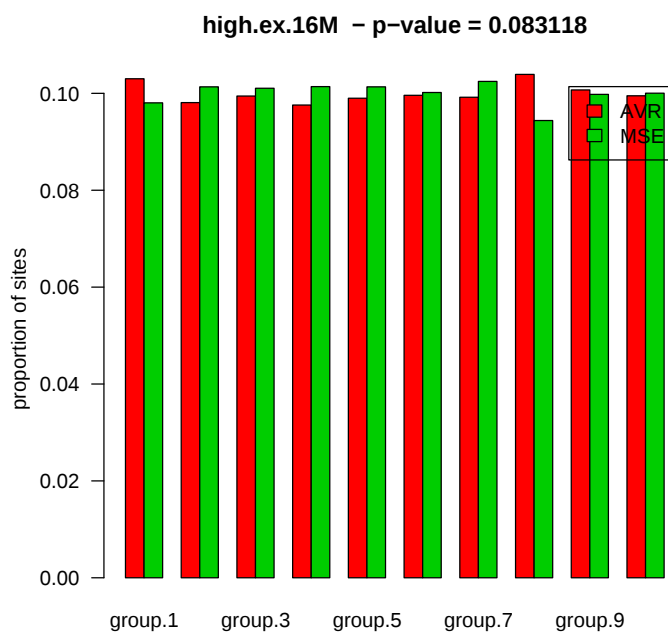
	lower	category	upper
1	0.000000e+00	group.1	1.144734e-06
2	1.144734e-06	group.2	1.489881e-06
3	1.489881e-06	group.3	1.939583e-06
4	1.939583e-06	group.4	2.432292e-06
5	2.432292e-06	group.5	2.982682e-06
6	2.982682e-06	group.6	3.477083e-06
7	3.477083e-06	group.7	4.189410e-06
8	4.189410e-06	group.8	5.445536e-06
9	5.445536e-06	group.9	6.514807e-06
10	6.514807e-06	group.10	1.003592e-05



And here we count genes in the upper 1/16th:

Category limits

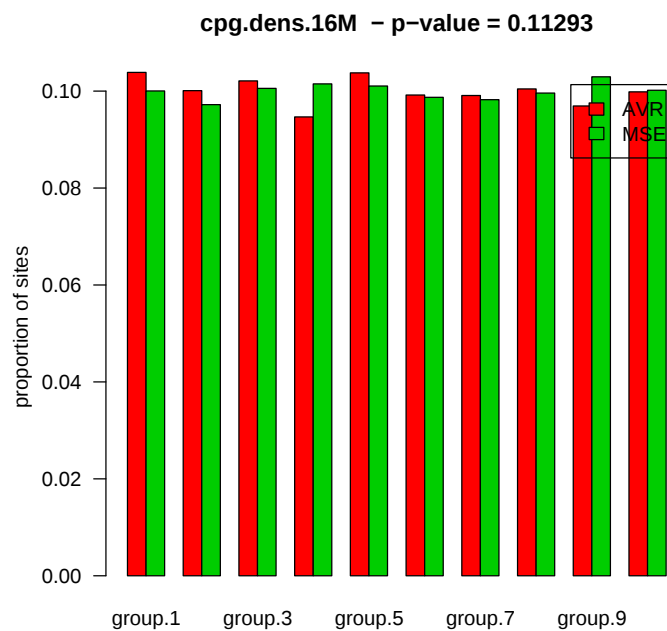
	lower	category	upper
1	0.000000e+00	group.1	5.949405e-07
2	5.949405e-07	group.2	7.822917e-07
3	7.822917e-07	group.3	9.895833e-07
4	9.895833e-07	group.4	1.218839e-06
5	1.218839e-06	group.5	1.490625e-06
6	1.490625e-06	group.6	1.708333e-06
7	1.708333e-06	group.7	1.989236e-06
8	1.989236e-06	group.8	2.471387e-06
9	2.471387e-06	group.9	3.213305e-06
10	3.213305e-06	group.10	4.589048e-06



Here the effect of density of CpG islands is studied:

Category limits

	lower	category	upper
1	1.221150e-07	group.1	4.125000e-06
2	4.125000e-06	group.2	5.531250e-06
3	5.531250e-06	group.3	6.875000e-06
4	6.875000e-06	group.4	8.230499e-06
5	8.230499e-06	group.5	1.006250e-05
6	1.006250e-05	group.6	1.212500e-05
7	1.212500e-05	group.7	1.616225e-05
8	1.616225e-05	group.8	1.933960e-05
9	1.933960e-05	group.9	2.596685e-05
10	2.596685e-05	group.10	6.774468e-05

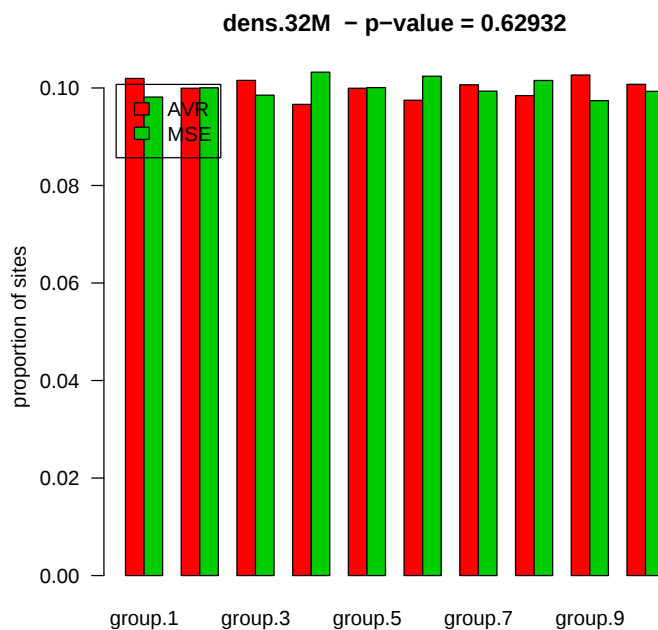


4.11 32 megaBase Window

In the barplot that follows we examine the association of insertion sites with expression density in a 32 megabase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category limits

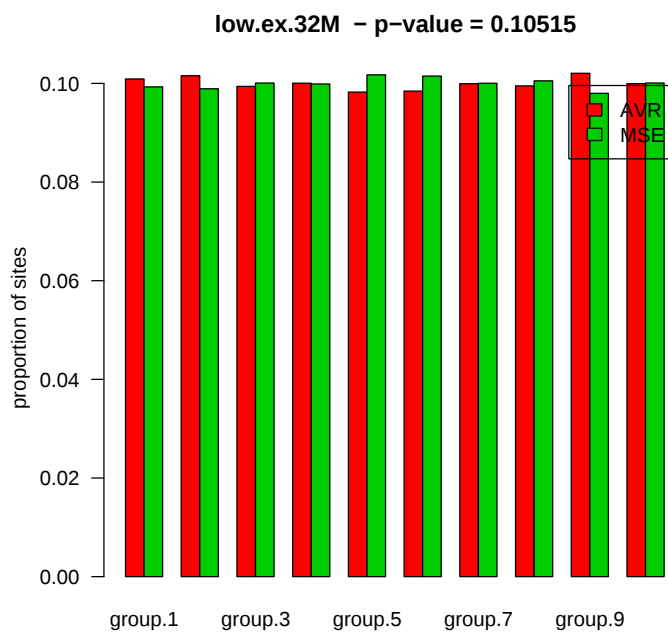
	lower	category	upper
1	3.052874e-08	group.1	6.630119e-06
2	6.630119e-06	group.2	7.930106e-06
3	7.930106e-06	group.3	9.129597e-06
4	9.129597e-06	group.4	1.089039e-05
5	1.089039e-05	group.5	1.219539e-05
6	1.219539e-05	group.6	1.378993e-05
7	1.378993e-05	group.7	1.670885e-05
8	1.670885e-05	group.8	1.899618e-05
9	1.899618e-05	group.9	2.124627e-05
10	2.124627e-05	group.10	4.136728e-05



Here are the results for expression density. First, we count just genes that are in the upper half.

Category limits

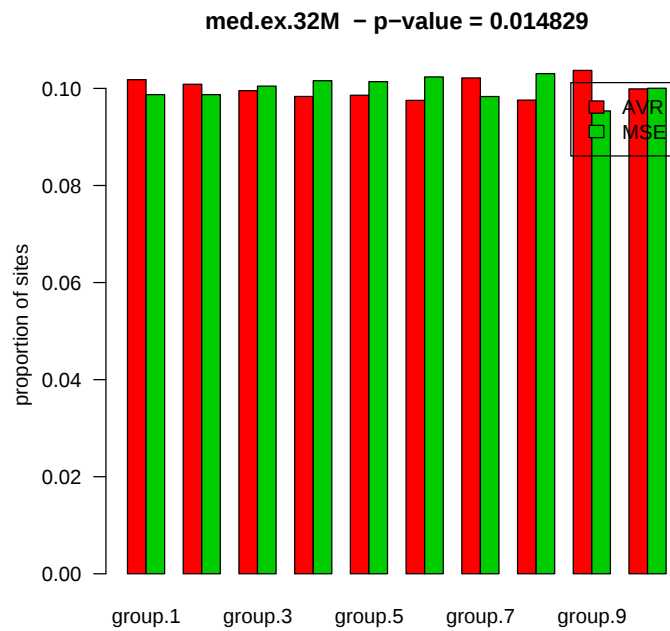
	lower	category	upper
1	0.000000e+00	group.1	2.538684e-06
2	2.538684e-06	group.2	3.287413e-06
3	3.287413e-06	group.3	3.918403e-06
4	3.918403e-06	group.4	4.734662e-06
5	4.734662e-06	group.5	5.549255e-06
6	5.549255e-06	group.6	6.460391e-06
7	6.460391e-06	group.7	7.833757e-06
8	7.833757e-06	group.8	9.649092e-06
9	9.649092e-06	group.9	1.058881e-05
10	1.058881e-05	group.10	2.003947e-05



Now we count genes in the upper 1/8th:

Category limits

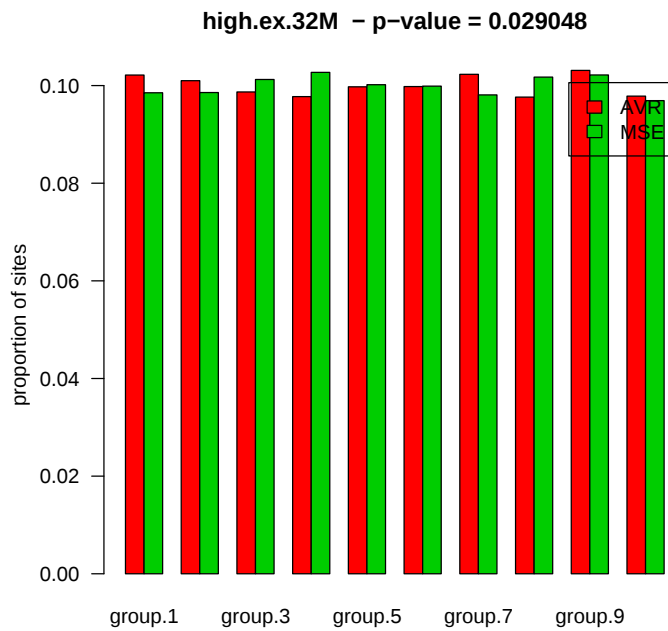
	lower	category	upper
1	0.000000e+00	group.1	1.253125e-06
2	1.253125e-06	group.2	1.570050e-06
3	1.570050e-06	group.3	1.858150e-06
4	1.858150e-06	group.4	2.242567e-06
5	2.242567e-06	group.5	2.598511e-06
6	2.598511e-06	group.6	3.213781e-06
7	3.213781e-06	group.7	3.843341e-06
8	3.843341e-06	group.8	4.344721e-06
9	4.344721e-06	group.9	4.885975e-06
10	4.885975e-06	group.10	9.497117e-06



And here we count genes in the upper 1/16th:

Category limits

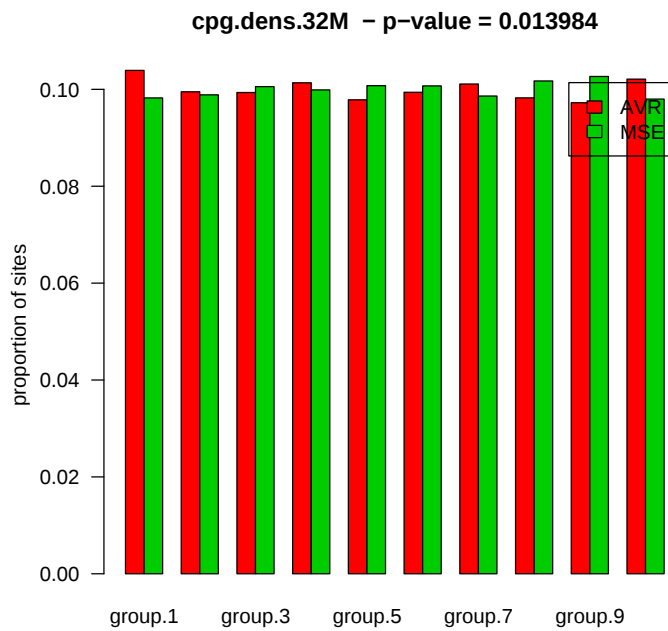
	lower	category	upper
1	0.000000e+00	group.1	6.385417e-07
2	6.385417e-07	group.2	8.202381e-07
3	8.202381e-07	group.3	9.542783e-07
4	9.542783e-07	group.4	1.142708e-06
5	1.142708e-06	group.5	1.339732e-06
6	1.339732e-06	group.6	1.522024e-06
7	1.522024e-06	group.7	1.830816e-06
8	1.830816e-06	group.8	2.055135e-06
9	2.055135e-06	group.9	2.451711e-06
10	2.451711e-06	group.10	4.213871e-06



Here the effect of density of CpG islands is studied:

Category limits

	lower	category	upper
1	1.544258e-07	group.1	4.562500e-06
2	4.562500e-06	group.2	5.953125e-06
3	5.953125e-06	group.3	6.828125e-06
4	6.828125e-06	group.4	7.562150e-06
5	7.562150e-06	group.5	9.156250e-06
6	9.156250e-06	group.6	1.065625e-05
7	1.065625e-05	group.7	1.252992e-05
8	1.252992e-05	group.8	1.703764e-05
9	1.703764e-05	group.9	2.096796e-05
10	2.096796e-05	group.10	4.082847e-05



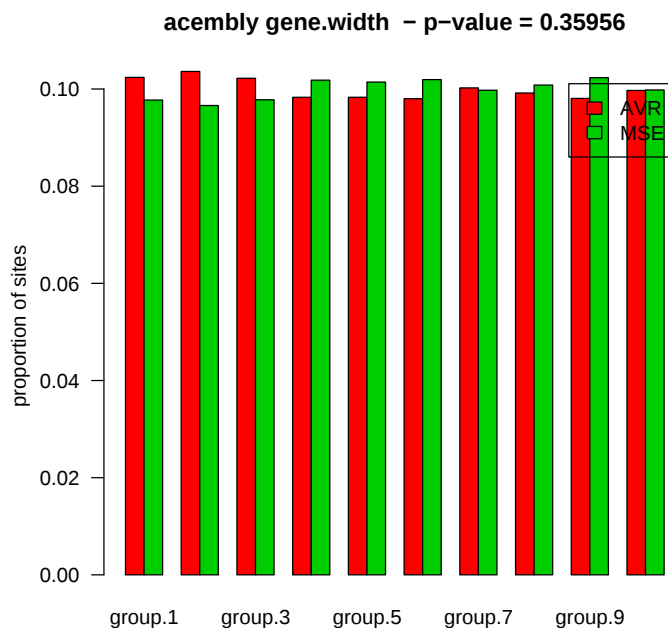
5 Juxtaposition with Gene Start and End Positions

5.1 Acembly Annotations

In this section we study the effect of juxtaposition in terms of gene start and end positions. The first barplot shows the effect of gene width for those insertions that are located within an Acembly gene.

Category limits

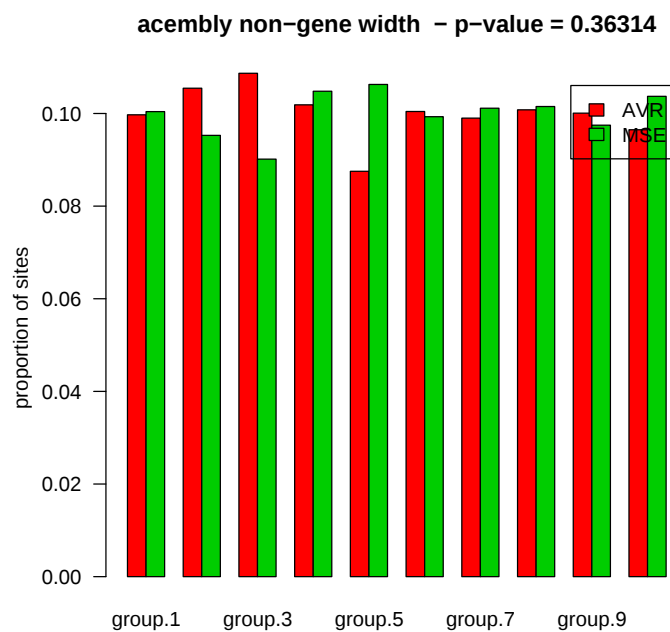
	lower	category	upper
1	141.0	group.1	11134.5
2	11134.5	group.2	20013.0
3	20013.0	group.3	29583.5
4	29583.5	group.4	40245.0
5	40245.0	group.5	53920.0
6	53920.0	group.6	73777.0
7	73777.0	group.7	96650.0
8	96650.0	group.8	134150.0
9	134150.0	group.9	210575.0
10	210575.0	group.10	1529446.0



The next plot uses the width of a non-gene region for insertions that fall into such regions.

Category limits

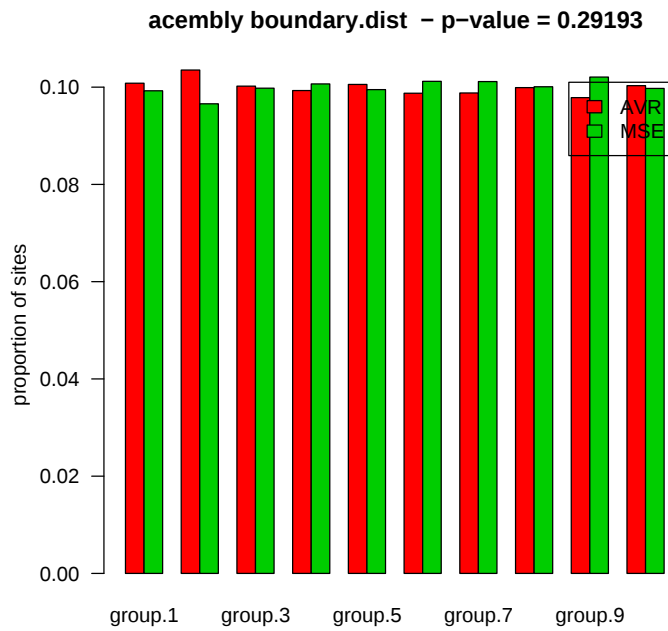
	lower	category	upper
1	69.0	group.1	7143.2
2	7143.2	group.2	13224.0
3	13224.0	group.3	21670.4
4	21670.4	group.4	32635.0
5	32635.0	group.5	49842.0
6	49842.0	group.6	81286.4
7	81286.4	group.7	122502.8
8	122502.8	group.8	189380.0
9	189380.0	group.9	307960.8
10	307960.8	group.10	30346008.0



The next plot studies the distance to the nearest boundary between a gene and a non-gene region. The distance is expressed as a fraction of the length of the region. Thus, '0.25' refers to one quarter of the distance from the site to nearest boundary divided by the total width of the region.

Category limits

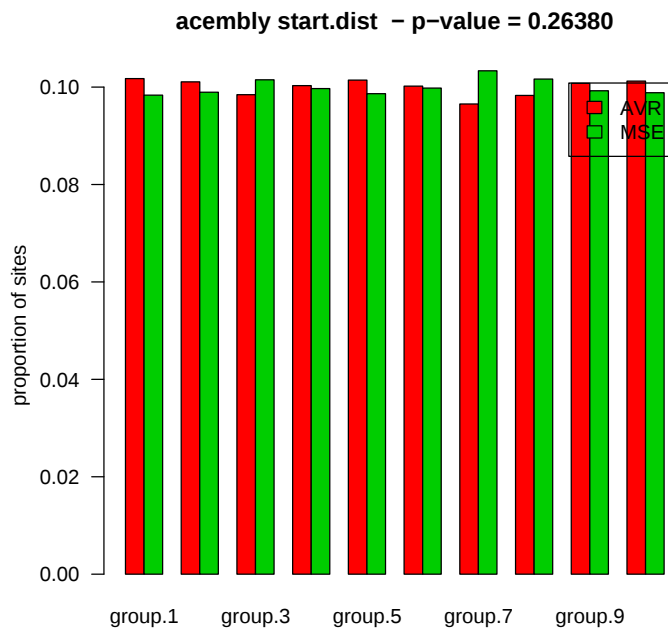
	lower	category	upper
1	9.852411e-06	group.1	0.05808594
2	5.808594e-02	group.2	0.10989332
3	1.098933e-01	group.3	0.15988619
4	1.598862e-01	group.4	0.21002717
5	2.100272e-01	group.5	0.26065890
6	2.606589e-01	group.6	0.30859731
7	3.085973e-01	group.7	0.35578540
8	3.557854e-01	group.8	0.40329985
9	4.032999e-01	group.9	0.45150765
10	4.515077e-01	group.10	0.49997407



This plot studies the effect of nearness to the beginning of a transcript. For sites in genes, it is the distance to the start of the gene divided by the width of the gene. For other sites it is the distance from the site to the nearer gene if that gene boundary is also a transcription starting point. Locations near '0' are relatively near the beginning of transcription, while those near '1' are near the termination of the transcript.

Category limits

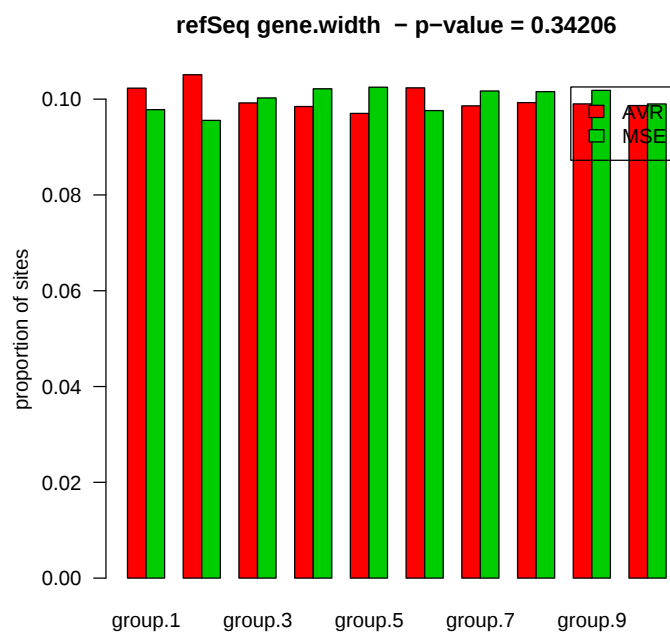
	lower	category	upper
1	7.000595e-05	group.1	0.1032365
2	1.032365e-01	group.2	0.1897125
3	1.897125e-01	group.3	0.2767975
4	2.767975e-01	group.4	0.3630682
5	3.630682e-01	group.5	0.4530168
6	4.530168e-01	group.6	0.5500514
7	5.500514e-01	group.7	0.6520067
8	6.520067e-01	group.8	0.7617847
9	7.617847e-01	group.9	0.8800670
10	8.800670e-01	group.10	0.9999901



5.2 RefSeq Annotations

Category limits

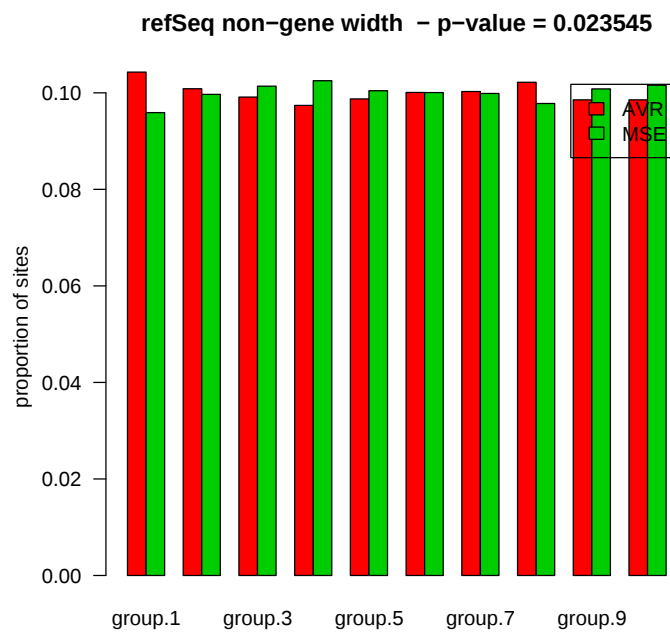
	lower	category	upper
1	990.0	group.1	18296.8
2	18296.8	group.2	29375.0
3	29375.0	group.3	40375.6
4	40375.6	group.4	54434.0
5	54434.0	group.5	74249.0
6	74249.0	group.6	92592.0
7	92592.0	group.7	115246.0
8	115246.0	group.8	155227.0
9	155227.0	group.9	245536.0
10	245536.0	group.10	2304258.0



Category limits

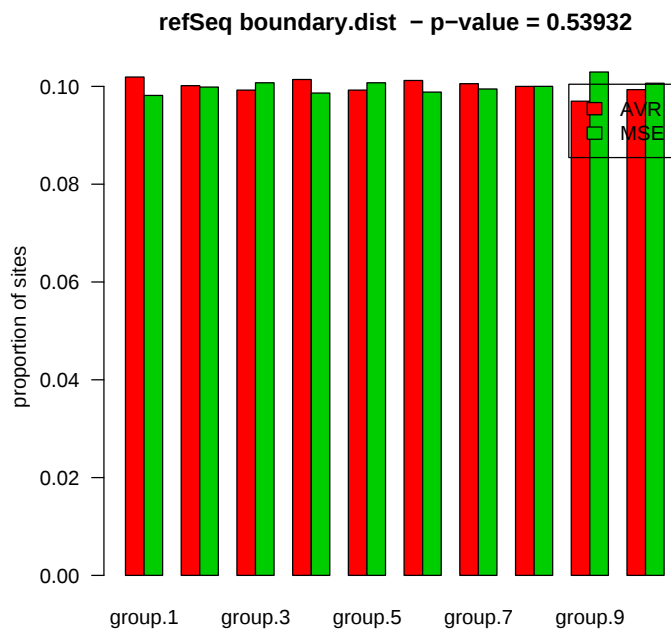
	lower	category	upper
1	70.0	group.1	14997.8
2	14997.8	group.2	30379.0
3	30379.0	group.3	58029.0
4	58029.0	group.4	99247.0

5	99247.0	group.5	155653.0
6	155653.0	group.6	270032.0
7	270032.0	group.7	468056.0
8	468056.0	group.8	881405.0
9	881405.0	group.9	1851242.0
10	1851242.0	group.10	21293005.0



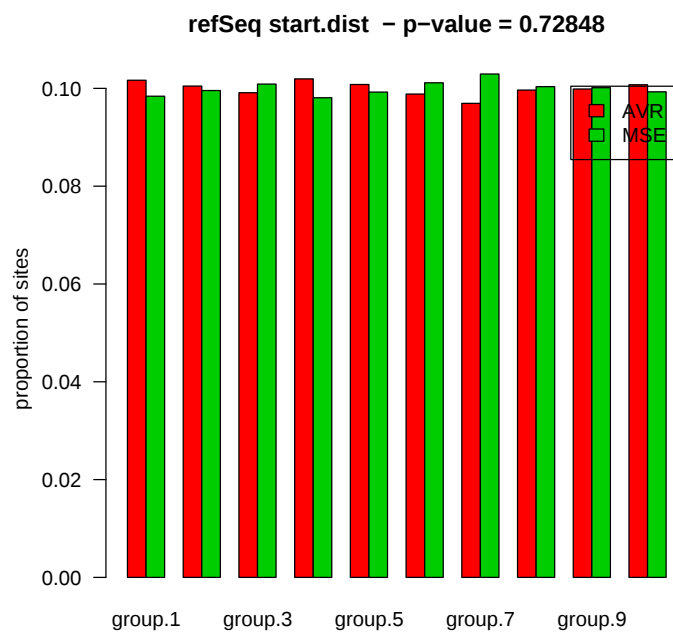
Category limits

	lower	category	upper
1	2.750918e-06	group.1	0.05945669
2	5.945669e-02	group.2	0.11268987
3	1.126899e-01	group.3	0.16269500
4	1.626950e-01	group.4	0.21338194
5	2.133819e-01	group.5	0.26310289
6	2.631029e-01	group.6	0.31135593
7	3.113559e-01	group.7	0.35925466
8	3.592547e-01	group.8	0.40615182
9	4.061518e-01	group.9	0.45424261
10	4.542426e-01	group.10	0.49999369



Category limits

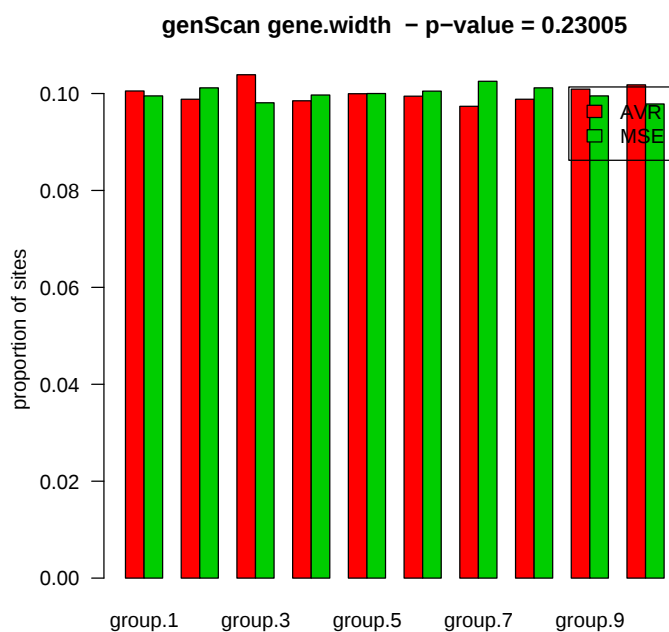
	lower	category	upper
1	2.750918e-06	group.1	0.09892256
2	9.892256e-02	group.2	0.18526898
3	1.852690e-01	group.3	0.27162230
4	2.716223e-01	group.4	0.35615411
5	3.561541e-01	group.5	0.44443861
6	4.444386e-01	group.6	0.53428061
7	5.342806e-01	group.7	0.63463944
8	6.346394e-01	group.8	0.74261160
9	7.426116e-01	group.9	0.86583189
10	8.658319e-01	group.10	0.99991841



5.3 genScan Annotations

Category limits

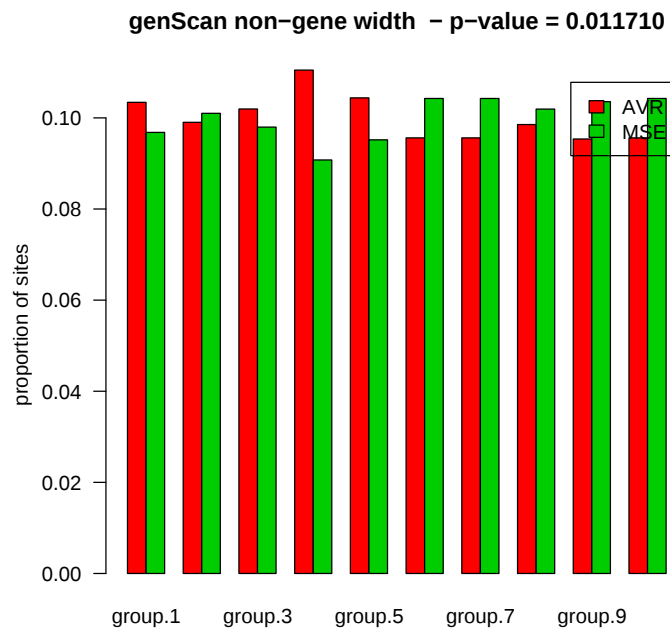
	lower	category	upper
1	207.0	group.1	21200.8
2	21200.8	group.2	34036.0
3	34036.0	group.3	46234.0
4	46234.0	group.4	59755.0
5	59755.0	group.5	75804.0
6	75804.0	group.6	94635.0
7	94635.0	group.7	120312.4
8	120312.4	group.8	155723.0
9	155723.0	group.9	222574.0
10	222574.0	group.10	872652.0



Category limits

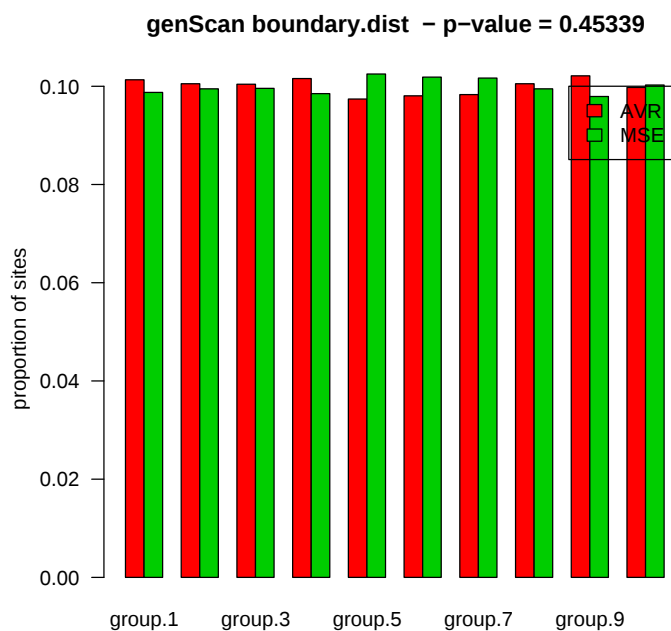
	lower	category	upper
1	168.0	group.1	6259.6
2	6259.6	group.2	9884.4
3	9884.4	group.3	14013.0
4	14013.0	group.4	18385.0

5	18385.0	group.5	22981.0
6	22981.0	group.6	29974.0
7	29974.0	group.7	39466.0
8	39466.0	group.8	51727.0
9	51727.0	group.9	72188.4
10	72188.4	group.10	30218312.0



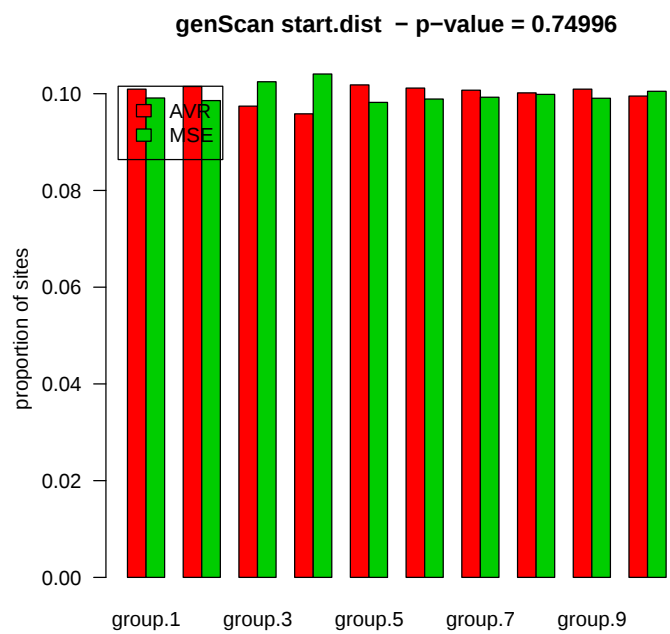
Category limits

	lower	category	upper
1	4.886153e-05	group.1	0.05955873
2	5.955873e-02	group.2	0.11233315
3	1.123332e-01	group.3	0.16250098
4	1.625010e-01	group.4	0.21284406
5	2.128441e-01	group.5	0.26064905
6	2.606490e-01	group.6	0.30903465
7	3.090346e-01	group.7	0.35627143
8	3.562714e-01	group.8	0.40434438
9	4.043444e-01	group.9	0.45242702
10	4.524270e-01	group.10	0.49999658



Category limits

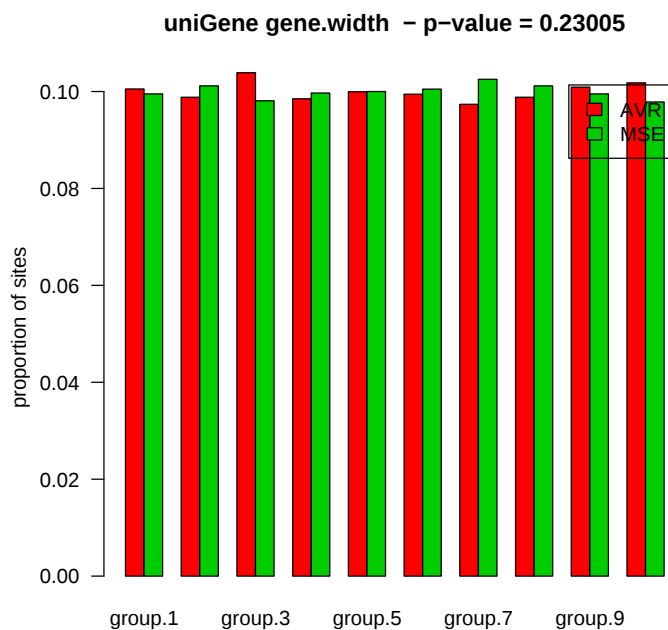
	lower	category	upper
1	4.886153e-05	group.1	0.1021515
2	1.021515e-01	group.2	0.1891642
3	1.891642e-01	group.3	0.2768891
4	2.768891e-01	group.4	0.3653818
5	3.653818e-01	group.5	0.4554740
6	4.554740e-01	group.6	0.5501795
7	5.501795e-01	group.7	0.6536011
8	6.536011e-01	group.8	0.7576950
9	7.576950e-01	group.9	0.8728831
10	8.728831e-01	group.10	0.9999507



5.4 uniGene Annotations

Category limits

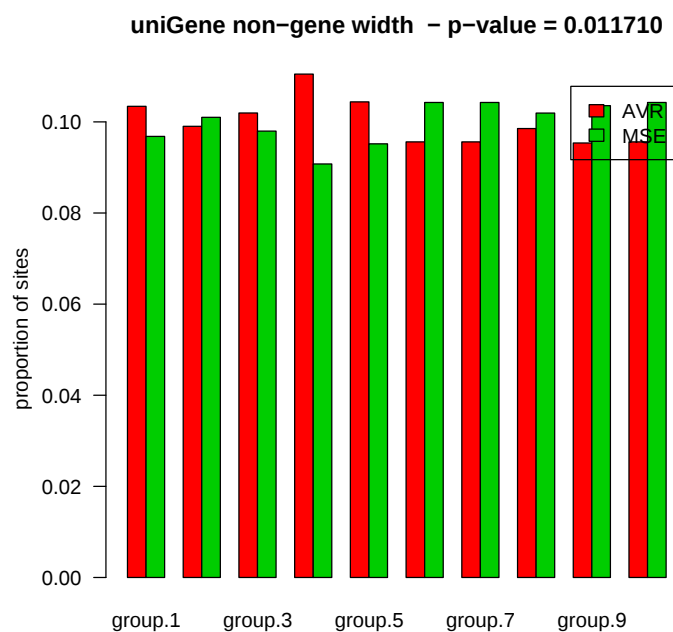
	lower	category	upper
1	207.0	group.1	21200.8
2	21200.8	group.2	34036.0
3	34036.0	group.3	46234.0
4	46234.0	group.4	59755.0
5	59755.0	group.5	75804.0
6	75804.0	group.6	94635.0
7	94635.0	group.7	120312.4
8	120312.4	group.8	155723.0
9	155723.0	group.9	222574.0
10	222574.0	group.10	872652.0



Category limits

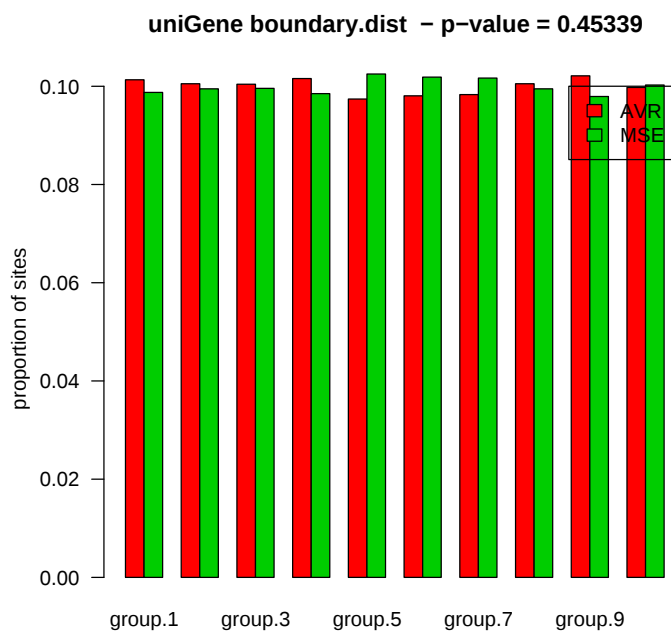
	lower	category	upper
1	168.0	group.1	6259.6
2	6259.6	group.2	9884.4
3	9884.4	group.3	14013.0
4	14013.0	group.4	18385.0

5	18385.0	group.5	22981.0
6	22981.0	group.6	29974.0
7	29974.0	group.7	39466.0
8	39466.0	group.8	51727.0
9	51727.0	group.9	72188.4
10	72188.4	group.10	30218312.0



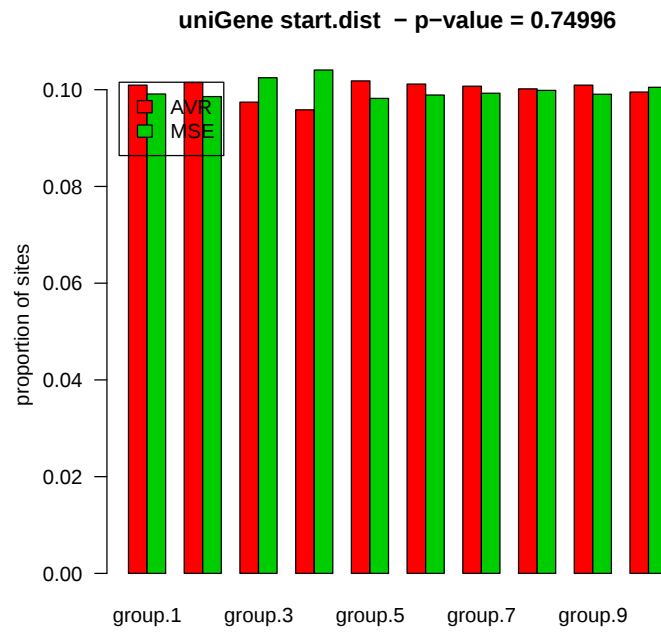
Category limits

	lower	category	upper
1	4.886153e-05	group.1	0.05955873
2	5.955873e-02	group.2	0.11233315
3	1.123332e-01	group.3	0.16250098
4	1.625010e-01	group.4	0.21284406
5	2.128441e-01	group.5	0.26064905
6	2.606490e-01	group.6	0.30903465
7	3.090346e-01	group.7	0.35627143
8	3.562714e-01	group.8	0.40434438
9	4.043444e-01	group.9	0.45242702
10	4.524270e-01	group.10	0.49999658



Category limits

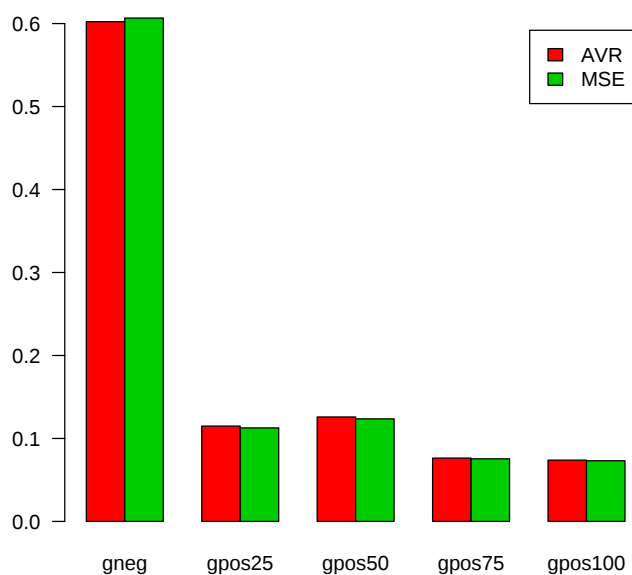
	lower	category	upper
1	4.886153e-05	group.1	0.1021515
2	1.021515e-01	group.2	0.1891642
3	1.891642e-01	group.3	0.2768891
4	2.768891e-01	group.4	0.3653818
5	3.653818e-01	group.5	0.4554740
6	4.554740e-01	group.6	0.5501795
7	5.501795e-01	group.7	0.6536011
8	6.536011e-01	group.8	0.7576950
9	7.576950e-01	group.9	0.8728831
10	8.728831e-01	group.10	0.9999507



6 Cytobands

Here we study the association of cytoBand with insertion intensity. The data are obtained from

<http://genome.ucsc.edu/goldenPath/hg17/database/cytoBand.txt.gz>.



A formal test of significance attains a p-value of 0.85718.

References

- [1] P. McCullagh and John A. Nelder. *Generalized linear models*. (Chapman & Hall ltd, 1999).
- [2] Xiaolin Wu, Yuan Li, Bruce Crise, Shawn M. Burgess “Transcription Start Regions in the Human Genome Are Favored Targets for MLV Integration,” *Science*, **300**(5626), (June 2003): 1749-1751.