Association of Genomic Features with Integration

Charles C. Berry

October 16, 2006

Contents

1 Introduction

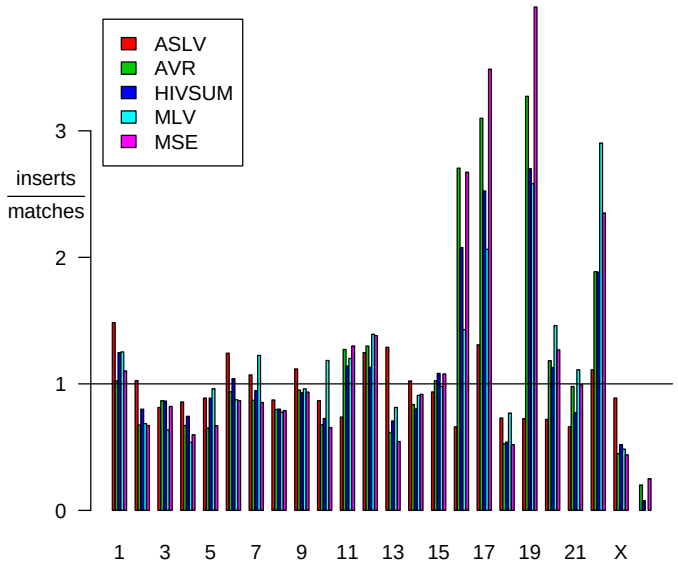
In this document, I examine the association of integration sites with various genomic features.

The data consist of both actual integration sites and sets of control sites, each set chosen to match the spacing (in bases) from the nearest restriction site (according to the direction in which the sequence was read) to an integration site. The numbers of insertion and matching sites for several data sets are shown below:

```
                  type
Origin.of.data.set insertion match
            ASLV         830  8290
            AVR        19962 59763
            HIVSUM      3710 37020
            MLV         1461 14550
            MSE        20607 61692
```

The advantage of choosing 'control' sites that match the spacing from the nearest restriction site is that biases due to location and density of restriction sites are eliminated by applying the classical multinomial logit model (reviewed in [2]). This model allows regression procedures to be applied to the study of integration intensity as a function of genomic features. The `clogit` function of the R `survival` library) implements estimation and fitting for such models along with the usual likelihood ratio and Wald tests.

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

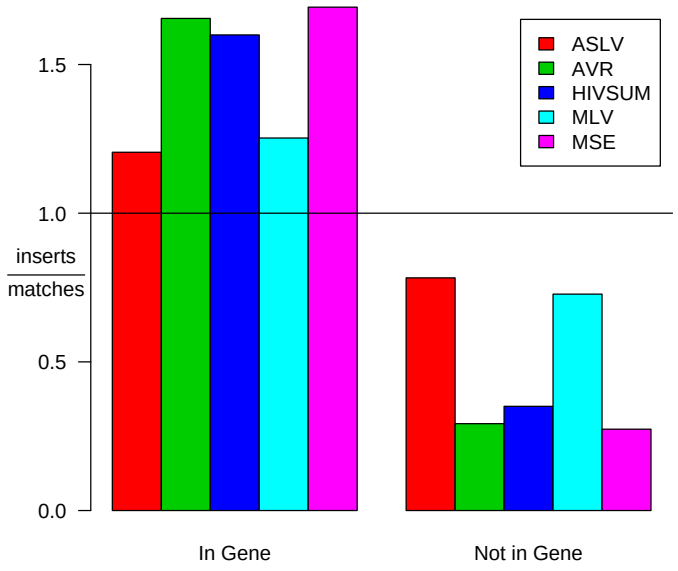


It seems evident that there are some chromosomes that are particularly favored for integration. This is reinforced by a test of statistical significance. The test performed used the likelihood ratio statistic for the multinomial logit model (reviewed in [2]) as implemented by the `clogit` function of the R `survival` library). The null hypothesis tested is that the ratio of true integration events to matched control sites is constant across all chromosomes. This test attains a p-value of $< 2.22e-16$.

2 Preference for Genes

2.1 Acembly Genes

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



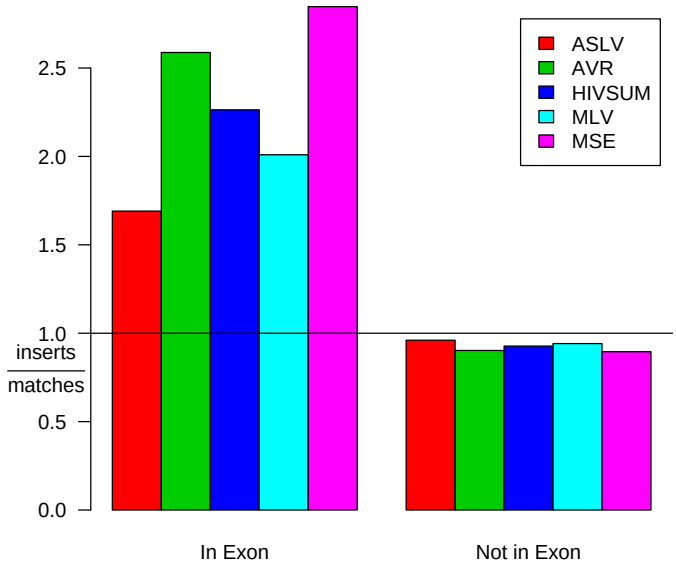
It seems evident that there is a strong tendency for insertions to occur in genes. A formal test of significance bears this out with a p-value of $< 2.22e-16$. Also, it appears that the tendency of different viruses to integrate into genes varies, and a test for this hypothesis attains $< 2.22e-16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
         coef     se     z         p
ASLV    0.434 0.0753  5.77  8.07e-09
AVR     1.740 0.0227 76.60  0.00e+00
HIVSUM  1.520 0.0453 33.60 2.41e-247
```

```
MLV     0.547 0.0575  9.52  1.81e-21
MSE     1.820 0.0227 80.20  0.00e+00
```

As is evident, there are some differences in the coefficients. The largest coefficient is seen in the MSE data set, while the smallest is seen in the ASLV data set.

In the following plot we show the relative frequency of insertions in exons according to the 'Acembly' annotation. The bars grouped over the label "In Exon" give the relative frequency of integration events (compared to control sites) between bases located in exons according to the Acembly annotation, while the label "Not in Exon" give the relative frequency of integration events (compared to control sites) between bases not located in exons according to the Acembly gene annotation.



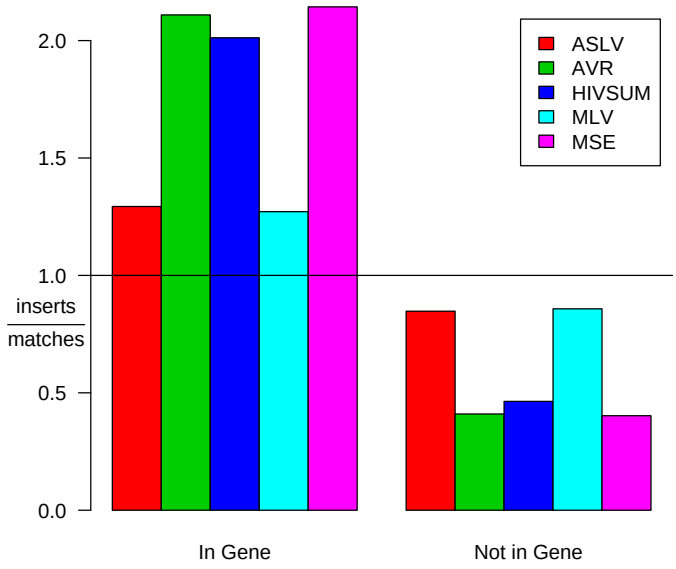
Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
         coef     se     z         p
ASLV    0.390 0.1340  2.91  3.59e-03
AVR     0.524 0.0278 18.80  4.69e-79
HIVSUM  0.394 0.0562  7.01  2.38e-12
MLV     0.546 0.0945  5.78  7.50e-09
MSE     0.600 0.0277 21.60 7.54e-104
```

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



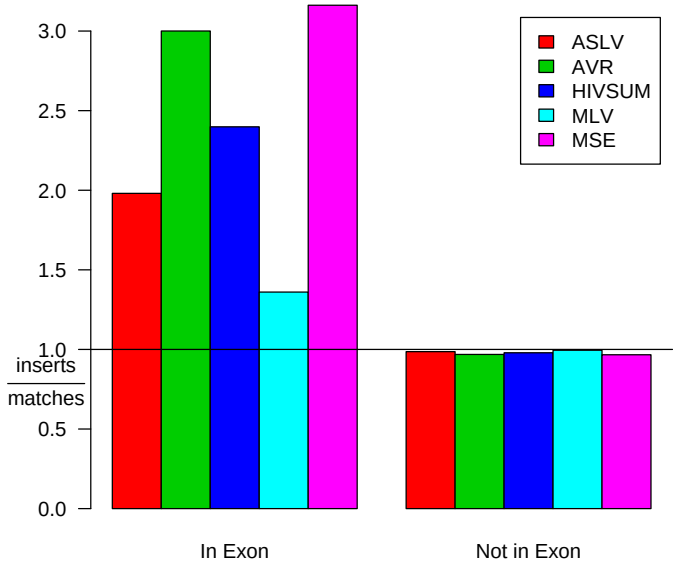
It seems evident that there is a strong tendency for insertions to occur in genes. A formal test of significance bears this out with a p-value of $< 2.22e-16$. Also, it appears that the tendency of different viruses to integrate into genes varies, and a test for this hypothesis attains $< 2.22e-16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
         coef     se     z        p
ASLV    0.429 0.0741  5.79 6.92e-09
AVR     1.640 0.0190 86.20 0.00e+00
HIVSUM  1.470 0.0376 39.20 0.00e+00
```

```
MLV    0.396 0.0557  7.11 1.15e-12
MSE    1.680 0.0189 89.00 0.00e+00
```

As is evident, there are some differences in the coefficients. The largest coefficient is seen in the MSE data set, while the smallest is seen in the MLV data set.

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



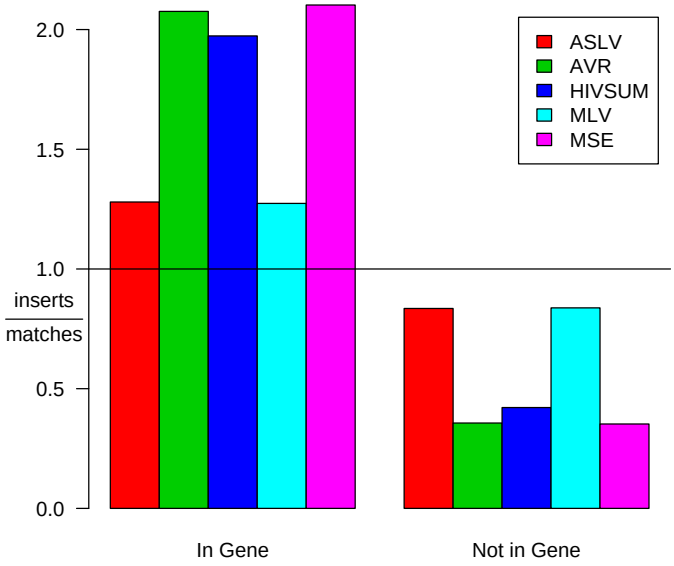
Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
         coef     se     z        p
ASLV   0.4410 0.2340 1.880 5.99e-02
AVR    0.3780 0.0493 7.680 1.64e-14
HIVSUM 0.1910 0.1000 1.910 5.63e-02
MLV    0.0798 0.2060 0.387 6.99e-01
MSE    0.4130 0.0485 8.530 1.46e-17
```

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.

2.3 ensGenes

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 'ensGene' annotation.

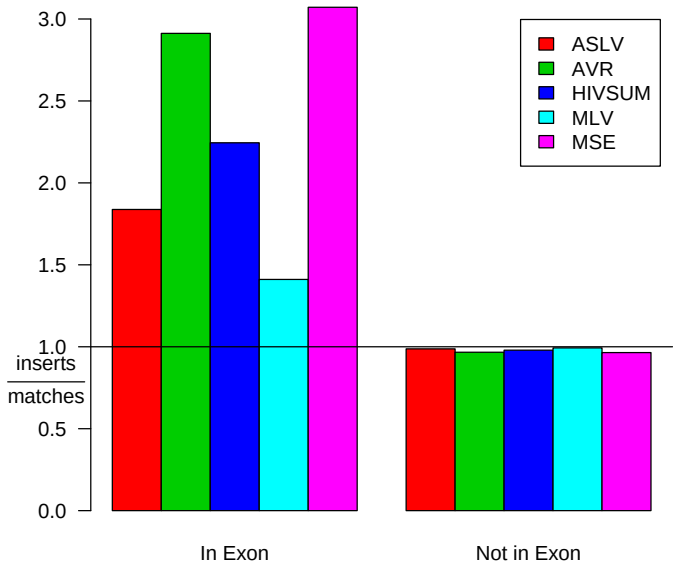


It seems evident that there is a strong tendency for insertions to occur in genes. A formal test of significance bears this out with a p-value of $< 2.22e-16$. Also, it appears that the tendency of different viruses to integrate into genes varies, and a test for this hypothesis attains $< 2.22e-16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
         coef     se     z        p
ASLV    0.435 0.0737  5.90 3.64e-09
AVR     1.760 0.0200 88.40 0.00e+00
HIVSUM 1.550 0.0390 39.80 0.00e+00
MLV     0.427 0.0552  7.73 1.09e-14
MSE     1.800 0.0198 90.80 0.00e+00
```

As is evident, there are some differences in the coefficients. The largest coefficient is seen in the MSE data set, while the smallest is seen in the MLV data set.

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



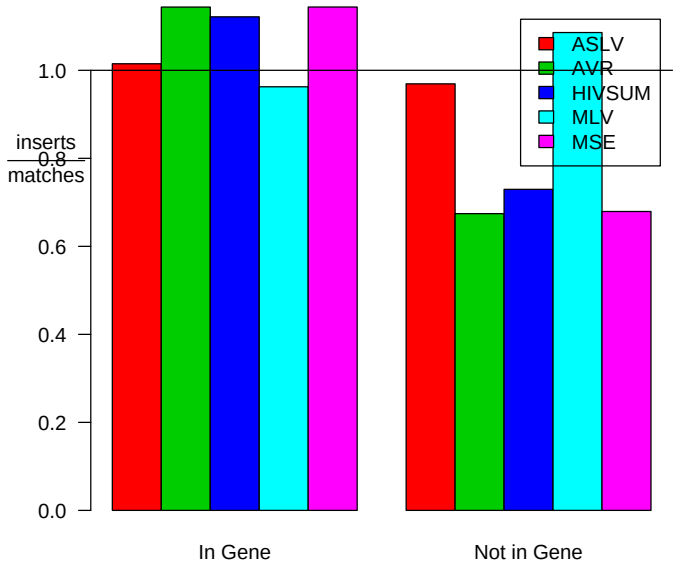
Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
         coef     se     z        p
ASLV    0.383 0.2330 1.640 1.00e-01
AVR     0.366 0.0472 7.750 9.22e-15
HIVSUM 0.138 0.0975 1.420 1.56e-01
MLV     0.117 0.1880 0.623 5.33e-01
MSE     0.406 0.0466 8.700 3.28e-18
```

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.

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.

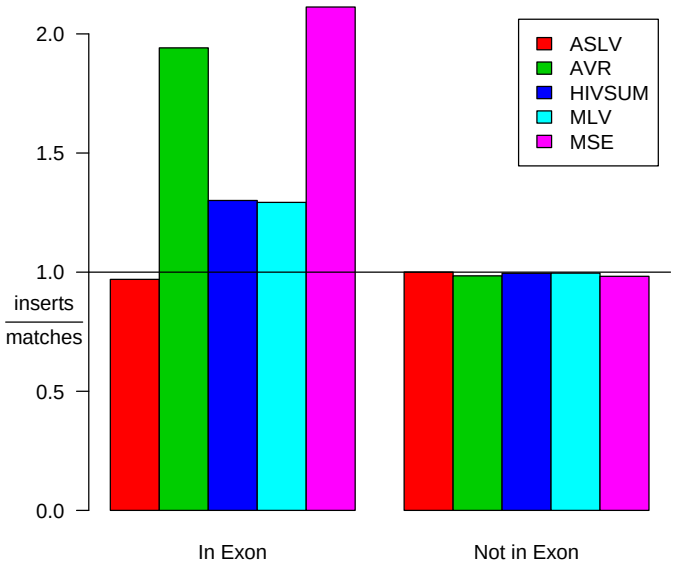


It seems evident that there is a strong tendency for insertions to occur in genes. A formal test of significance bears this out with a p-value of $< 2.22e-16$. Also, it appears that the tendency of different viruses to integrate into genes varies, and a test for this hypothesis attains $< 2.22e-16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
          coef     se      z         p
ASLV    0.0446 0.0786  0.568  5.70e-01
AVR     0.5310 0.0197 26.900 1.67e-159
HIVSUM  0.4350 0.0410 10.600  2.42e-26
MLV    -0.1180 0.0587 -2.000  4.52e-02
MSE     0.5200 0.0192 27.000 4.49e-161
```

As is evident, there are some differences in the coefficients. The largest coefficient is seen in the AVR data set, while the smallest is seen in the MLV data set.

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



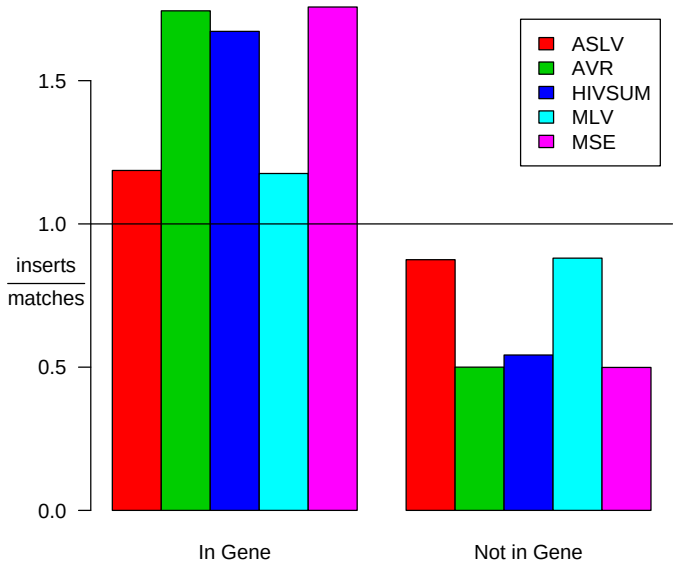
Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
          coef     se      z        p
ASLV   -0.0455 0.3330 -0.136 8.91e-01
AVR     0.5470 0.0524 10.400 1.65e-25
HIVSUM  0.1530 0.1240  1.240 2.14e-01
MLV     0.3080 0.2080  1.480 1.40e-01
MSE     0.6400 0.0517 12.400 2.78e-35
```

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.

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' annotation.

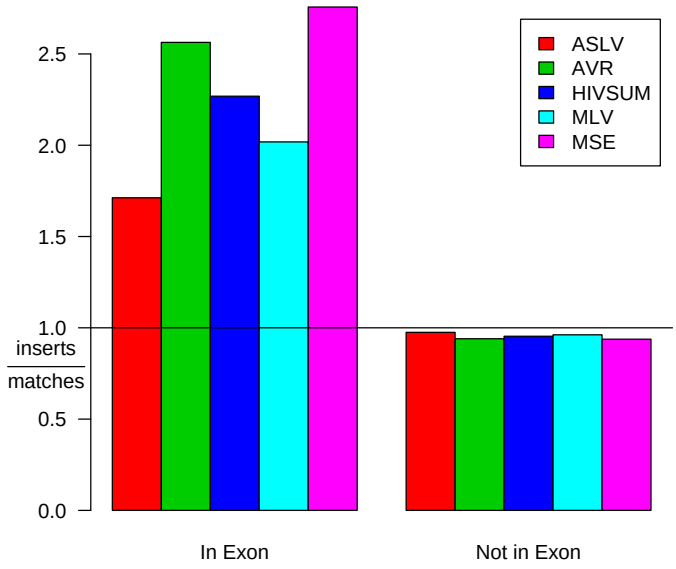


It seems evident that there is a strong tendency for insertions to occur in genes. A formal test of significance bears this out with a p-value of $< 2.22e-16$. Also, it appears that the tendency of different viruses to integrate into genes varies, and a test for this hypothesis attains $< 2.22e-16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
         coef     se     z         p
ASLV    0.309 0.0734  4.21  2.52e-05
AVR     1.250 0.0180 69.40  0.00e+00
HIVSUM  1.130 0.0369 30.70 9.38e-207
MLV     0.289 0.0552  5.24  1.58e-07
MSE     1.260 0.0178 71.00  0.00e+00
```

As is evident, there are some differences in the coefficients. The largest coefficient is seen in the MSE data set, while the smallest is seen in the MLV data set.

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



Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
ASLV	0.408	0.1650	2.47	1.35e-02
AVR	0.434	0.0339	12.80	1.57e-37
HIVSUM	0.347	0.0687	5.05	4.43e-07
MLV	0.613	0.1150	5.33	1.01e-07
MSE	0.492	0.0340	14.50	1.44e-47

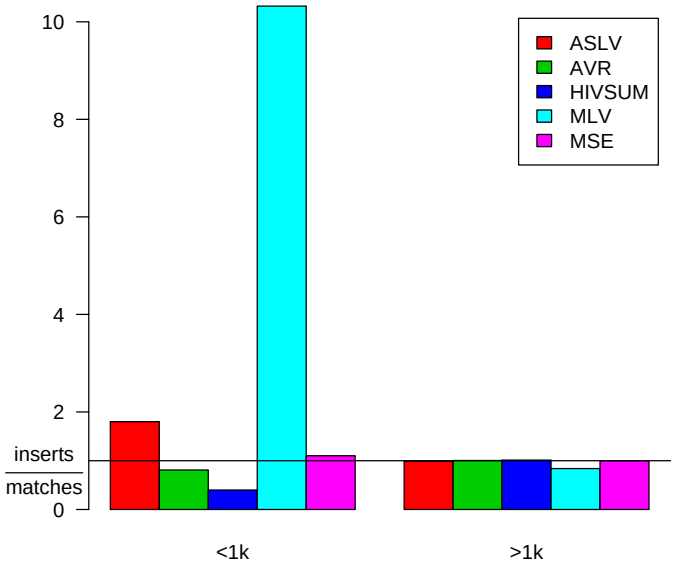
The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as "In Exon" is net of that due to being in a gene.

3 CpG Island Neighborhoods

Here we study the effect of being in the neighborhood of CpG Islands. Following Wu et al [3], who found that the neighborhoods within ±1kb 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 ±1kb of a CpG island:



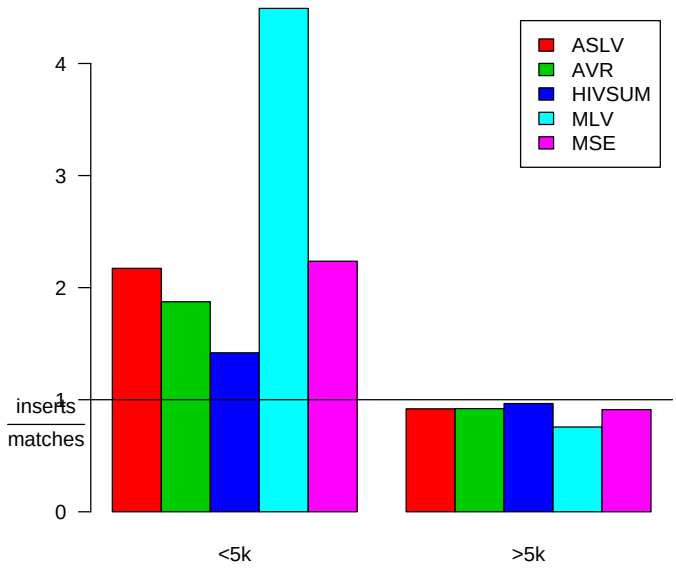
A formal test of significance comparing the difference attains a p-value of $1.382e-06$. A test for differences between viruses attains $< 2.22e-16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
          coef     se     z         p
ASLV     0.6030 0.2350  2.57  1.03e-02
AVR     -0.2160 0.0576 -3.75  1.74e-04
HIVSUM -0.9280 0.1840 -5.04  4.55e-07
MLV      2.5300 0.0979 25.80 5.68e-147
MSE      0.0964 0.0605  1.59  1.11e-01
```

The largest coefficient is seen in the MLV data set, while the smallest is seen in the HIVSUM data set.

3.2 5 kilobase neighborhoods

The following plot shows the effect of being in or within ±5kb of a CpG island:



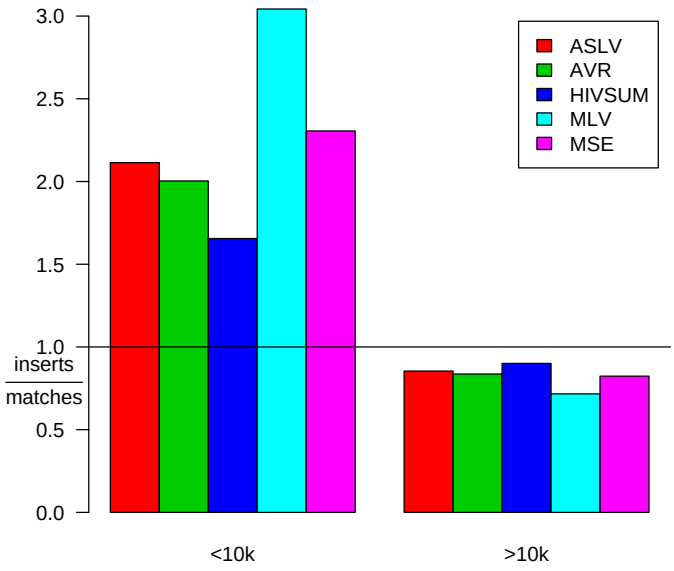
A formal test of significance comparing the difference attains a p-value of $< 2.22e-16$. A test for differences between viruses attains $< 2.22e-16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
         coef     se     z         p
ASLV    0.865 0.1100  7.88  3.20e-15
AVR     0.713 0.0247 28.90 9.49e-184
HIVSUM  0.383 0.0568  6.74  1.54e-11
MLV     1.780 0.0678 26.20 1.42e-151
MSE     0.907 0.0258 35.20 4.54e-271
```

The largest coefficient is seen in the MLV data set, while the smallest is seen in the HIVSUM data set.

3.3 10 kilobase neighborhoods

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



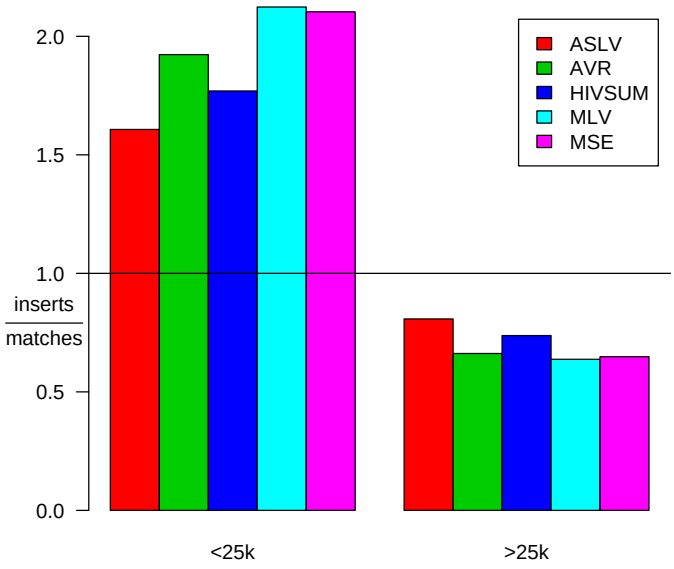
A formal test of significance comparing the difference attains a p-value of $< 2.22e-16$. A test for differences between viruses attains $< 2.22e-16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
        coef     se    z         p
ASLV   0.909 0.0881 10.3  5.66e-25
AVR    0.878 0.0200 44.0  0.00e+00
HIVSUM 0.610 0.0428 14.3  3.77e-46
MLV    1.440 0.0604 23.9 4.22e-126
MSE    1.030 0.0204 50.8  0.00e+00
```

The largest coefficient is seen in the MLV data set, while the smallest is seen in the HIVSUM data set.

3.4 25 kilobase neighborhoods

The following plot shows the effect of being in or within ±25kb of a CpG island:



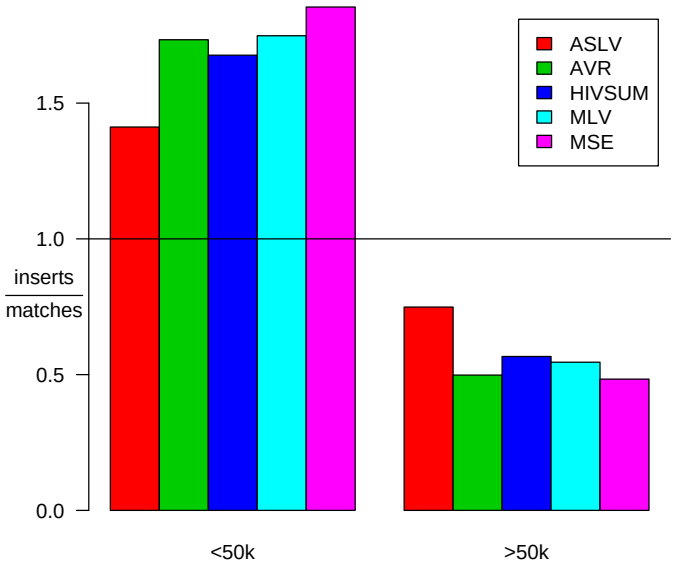
A formal test of significance comparing the difference attains a p-value of $< 2.22e-16$. A test for differences between viruses attains $< 2.22e-16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
        coef     se    z         p
ASLV   0.692 0.0760  9.1  8.90e-20
AVR    1.070 0.0173 61.8  0.00e+00
HIVSUM 0.875 0.0352 24.9 1.53e-136
MLV    1.210 0.0566 21.5 3.08e-102
MSE    1.190 0.0174 68.3  0.00e+00
```

The largest coefficient is seen in the MLV data set, while the smallest is seen in the ASLV data set.

3.5 50 kilobase neighborhoods

The following plot shows the effect of being in or within ±50kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of $< 2.22e-16$. A test for differences between viruses attains $< 2.22e-16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

```
         coef     se     z         p
ASLV    0.637 0.0734  8.69  3.72e-18
AVR     1.250 0.0181 69.10  0.00e+00
HIVSUM  1.090 0.0363 30.00 2.91e-197
MLV     1.170 0.0583 20.10  1.41e-89
MSE     1.350 0.0179 75.40  0.00e+00
```

The largest coefficient is seen in the MSE data set, while the smallest is seen in the ASLV data set.

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

In this section the association with gene density is examined. 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 $1/8^{th}$ and divide by number of bases

high.ex Count genes whose expression is in the upper $1/16^{th}$ 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

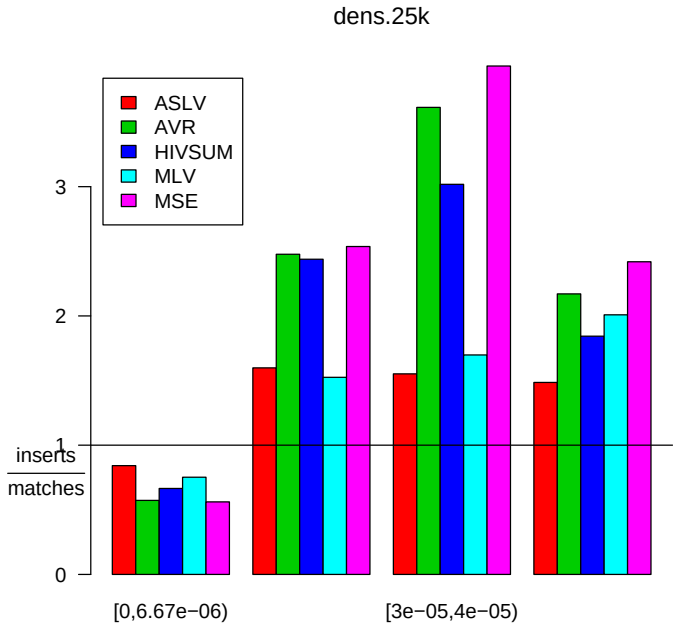
In the barplot that follows we examine the association of insertion sites with gene density in a 25 kilobase window surrounding each locus. More such plots will follow and the method of their construction is always to try to divide the data according to the deciles of density. However, it often happens that there is a very skewed distribution of density and even the 90^{th} 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, 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:

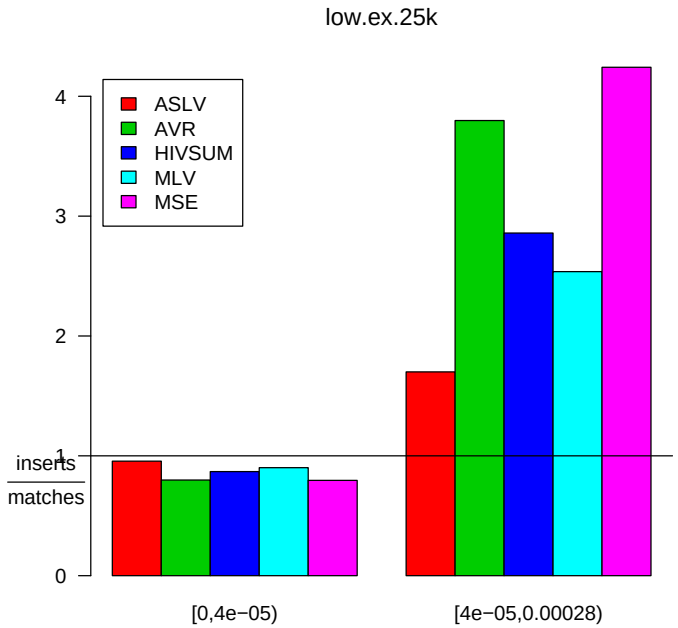
```
[1] "Jurkat-HU133Plus2"
```

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



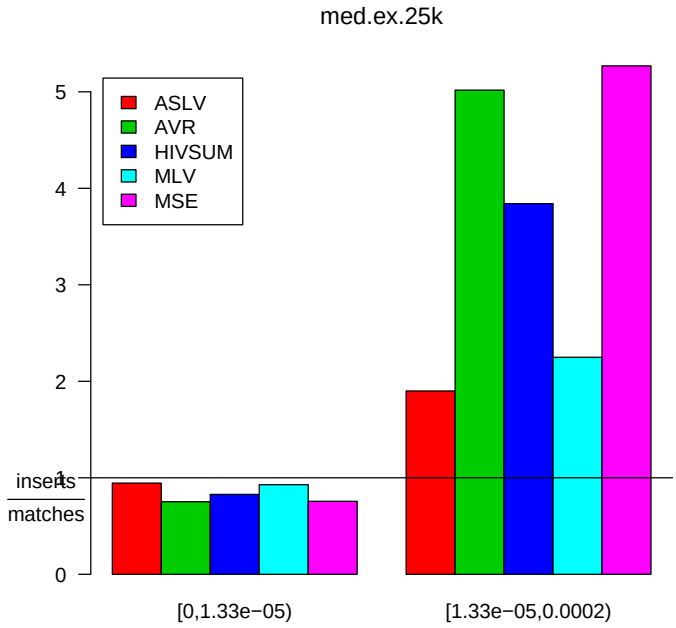
	coef	se	z	p
ASLV	0.595	0.0774	7.69	1.45e-14
AVR	1.410	0.0179	78.80	0.00e+00
HIVSUM	1.140	0.0353	32.30	5.31e-229
MLV	0.904	0.0570	15.90	1.27e-56
MSE	1.510	0.0179	84.30	0.00e+00

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



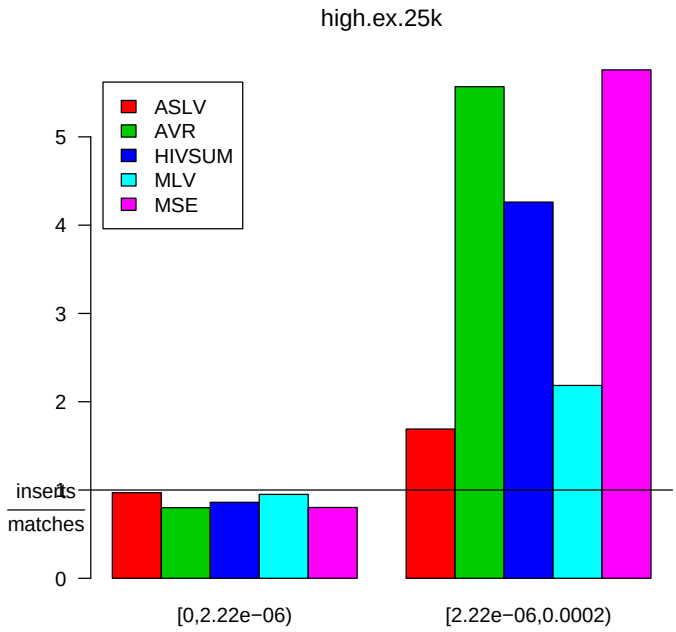
	coef	se	z	p
ASLV	0.655	0.0912	7.18	6.86e-13
AVR	1.820	0.0200	91.10	0.00e+00
HIVSUM	1.470	0.0374	39.30	0.00e+00
MLV	0.872	0.0662	13.20	1.24e-39
MSE	1.890	0.0201	94.20	0.00e+00

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



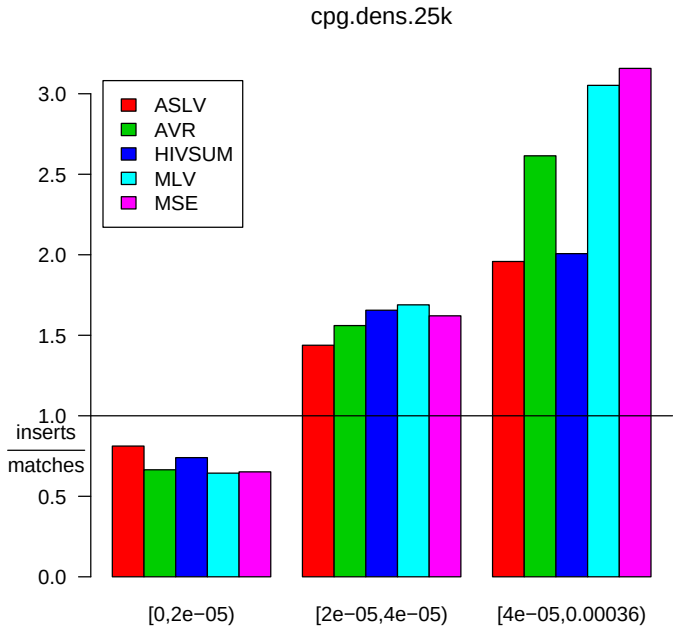
	coef	se	z	p
ASLV	0.681	0.1110	6.14	8.14e-10
AVR	1.960	0.0230	85.50	0.00e+00
HIVSUM	1.600	0.0419	38.10	0.00e+00
MLV	0.844	0.0807	10.50	1.36e-25
MSE	1.980	0.0229	86.50	0.00e+00

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



	coef	se	z	p
ASLV	0.557	0.1460	3.82	1.35e-04
AVR	1.960	0.0280	70.00	0.00e+00
HIVSUM	1.600	0.0511	31.40	4.20e-216
MLV	0.837	0.1020	8.20	2.47e-16
MSE	1.960	0.0276	71.00	0.00e+00

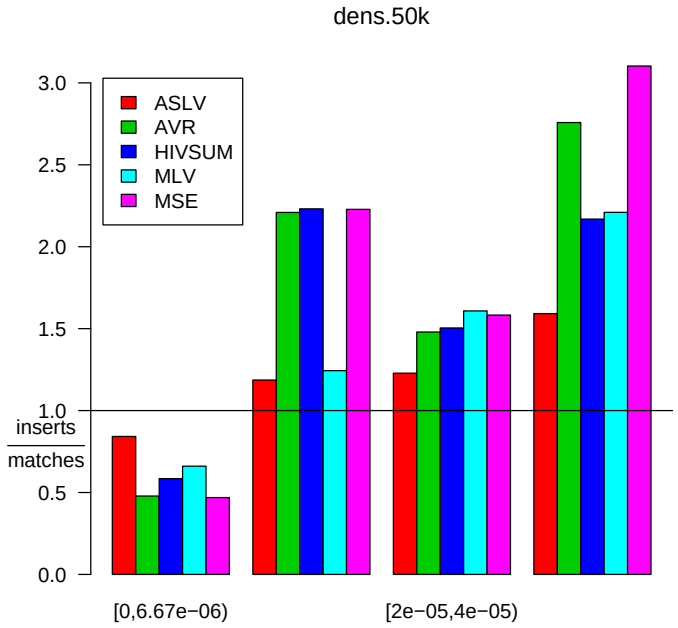
Here the effect of density of CpG islands is studied:



	coef	se	z	p
ASLV	0.681	0.0762	8.94	3.98e-19
AVR	1.070	0.0173	61.70	0.00e+00
HIVSUM	0.875	0.0353	24.80	9.08e-136
MLV	1.210	0.0567	21.40	2.54e-101
MSE	1.180	0.0174	68.00	0.00e+00

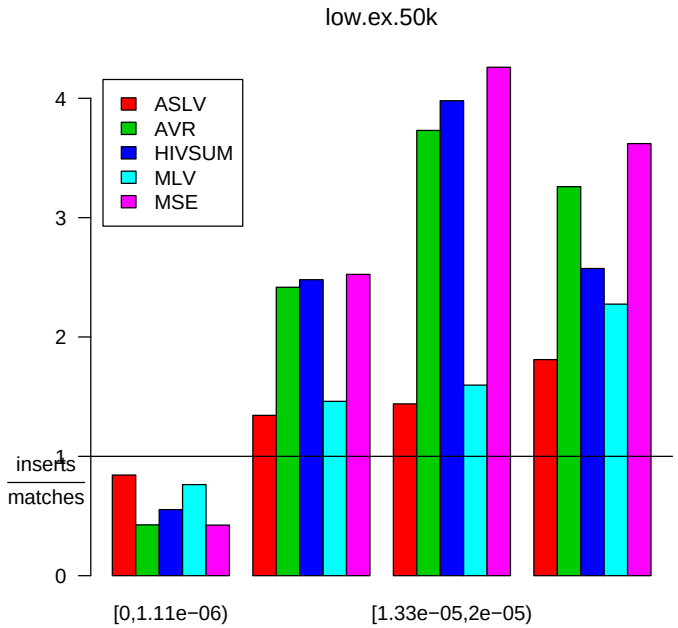
4.2 50 kilobase Window

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 represented on the chip.



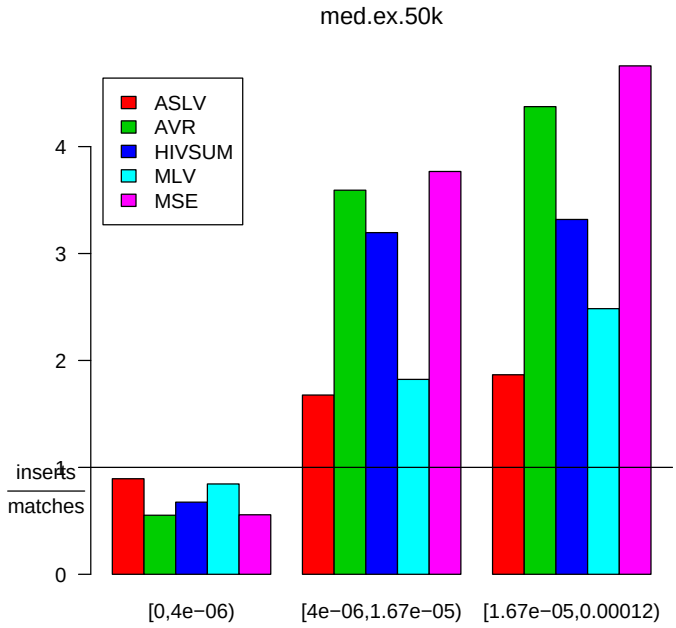
	coef	se	z	p
ASLV	0.513	0.0735	6.98	2.98e-12
AVR	1.590	0.0193	82.20	0.00e+00
HIVSUM	1.290	0.0372	34.60	1.27e-262
MLV	0.982	0.0567	17.30	4.42e-67
MSE	1.680	0.0192	87.40	0.00e+00

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



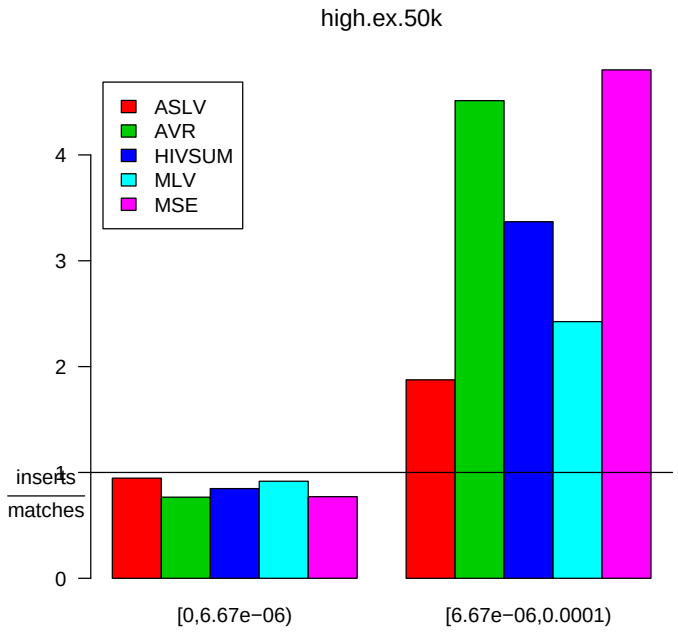
	coef	se	z	p
ASLV	0.654	0.0792	8.26	1.45e-16
AVR	1.950	0.0196	99.60	0.00e+00
HIVSUM	1.560	0.0358	43.60	0.00e+00
MLV	0.925	0.0577	16.00	8.33e-58
MSE	2.040	0.0197	104.00	0.00e+00

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



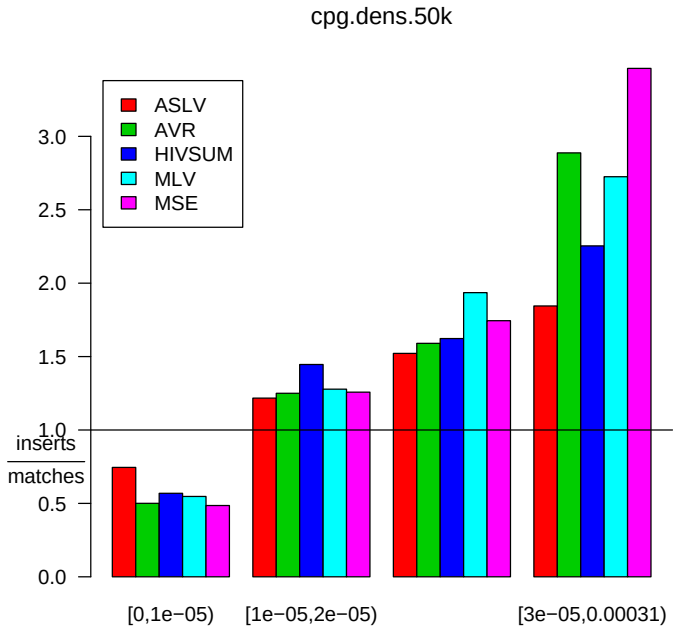
	coef	se	z	p
ASLV	0.663	0.0905	7.32	2.44e-13
AVR	1.990	0.0203	97.90	0.00e+00
HIVSUM	1.610	0.0372	43.30	0.00e+00
MLV	0.916	0.0646	14.20	1.14e-45
MSE	2.040	0.0203	100.00	0.00e+00

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



	coef	se	z	p
ASLV	0.611	0.1110	5.5	3.79e-08
AVR	1.870	0.0227	82.5	0.00e+00
HIVSUM	1.500	0.0423	35.5	3.86e-276
MLV	0.927	0.0774	12.0	5.10e-33
MSE	1.910	0.0227	84.1	0.00e+00

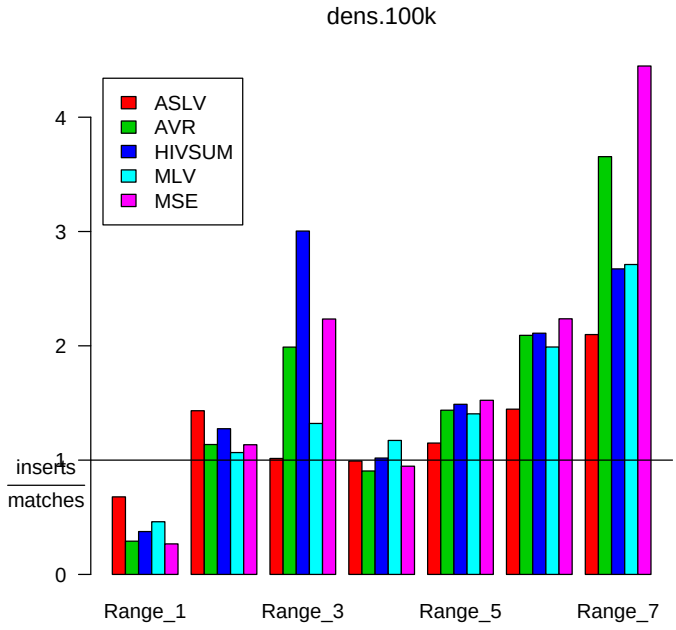
Here the effect of density of CpG islands is studied:



	coef	se	z	p
ASLV	0.646	0.0734	8.81	1.24e-18
AVR	1.240	0.0180	68.90	0.00e+00
HIVSUM	1.080	0.0362	29.90	3.60e-196
MLV	1.170	0.0582	20.10	4.32e-90
MSE	1.340	0.0179	75.30	0.00e+00

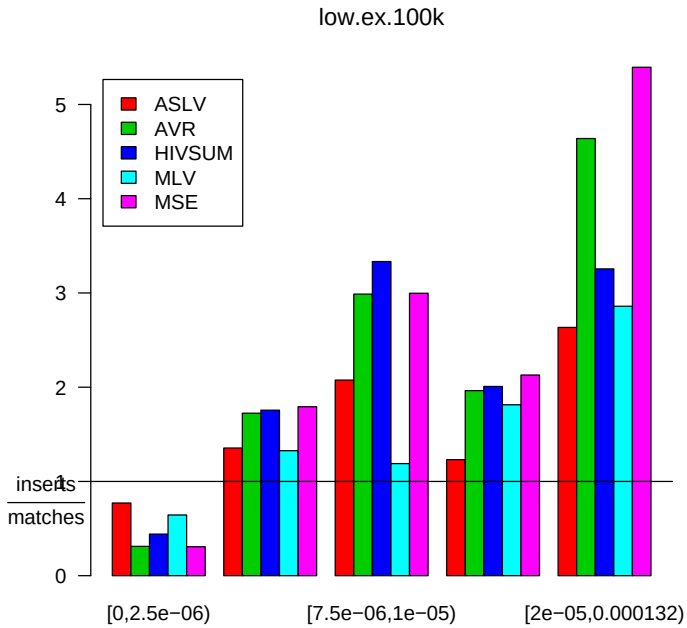
4.3 100 kilobase Window

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 represented on the chip.



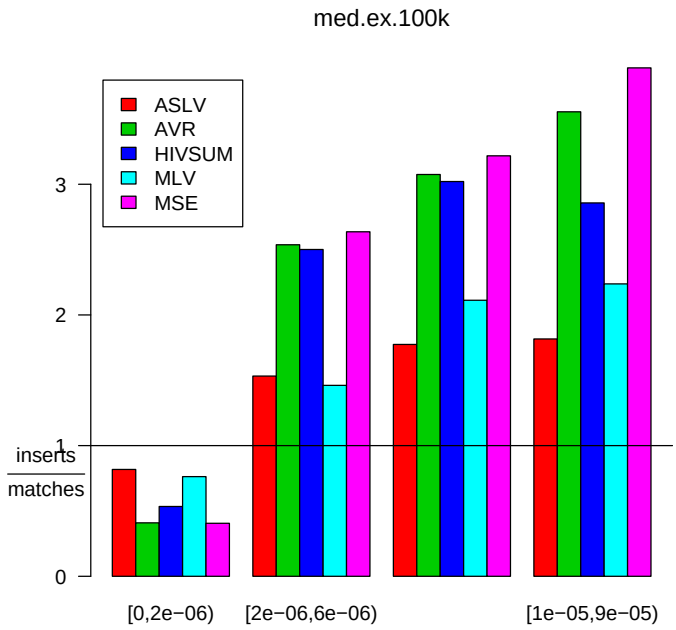
	coef	se	z	p
ASLV	0.432	0.0732	5.9	3.68e-09
AVR	1.330	0.0189	70.5	0.00e+00
HIVSUM	1.070	0.0374	28.7	1.98e-181
MLV	1.030	0.0584	17.7	9.46e-70
MSE	1.440	0.0187	77.0	0.00e+00

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



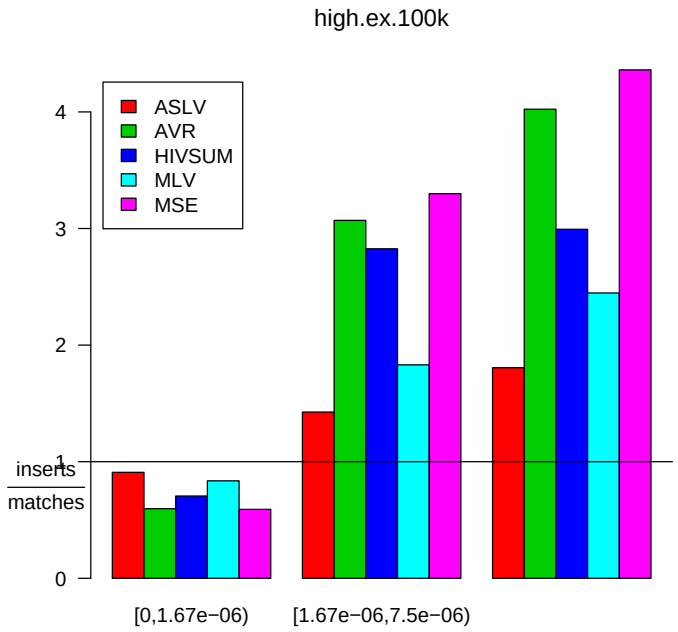
	coef	se	z	p
ASLV	0.718	0.0739	9.71	2.61e-22
AVR	2.120	0.0215	98.60	0.00e+00
HIVSUM	1.690	0.0387	43.70	0.00e+00
MLV	1.060	0.0562	18.90	9.79e-80
MSE	2.210	0.0214	103.00	0.00e+00

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



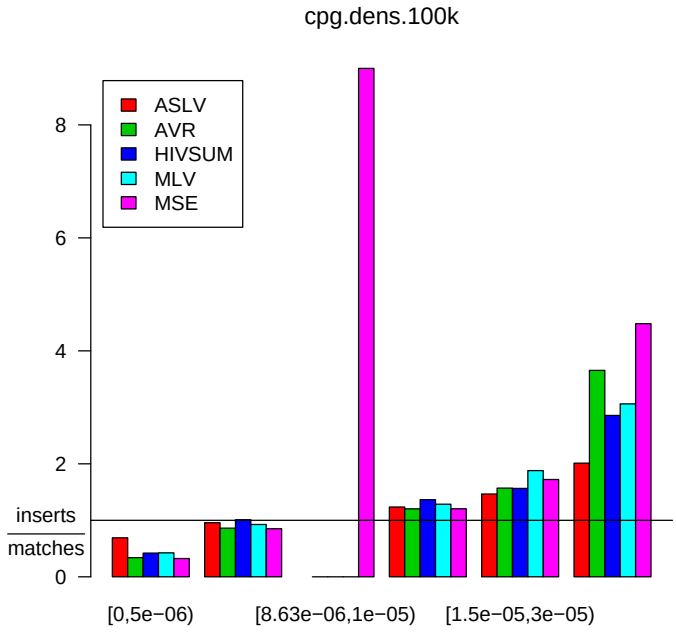
	coef	se	z	p
ASLV	0.722	0.0774	9.33	1.11e-20
AVR	2.080	0.0202	103.00	0.00e+00
HIVSUM	1.680	0.0362	46.30	0.00e+00
MLV	0.937	0.0573	16.30	5.33e-60
MSE	2.140	0.0201	107.00	0.00e+00

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



	coef	se	z	p
ASLV	0.576	0.0911	6.32	2.62e-10
AVR	1.810	0.0199	90.80	0.00e+00
HIVSUM	1.440	0.0371	38.80	0.00e+00
MLV	0.938	0.0636	14.70	3.29e-49
MSE	1.870	0.0199	94.40	0.00e+00

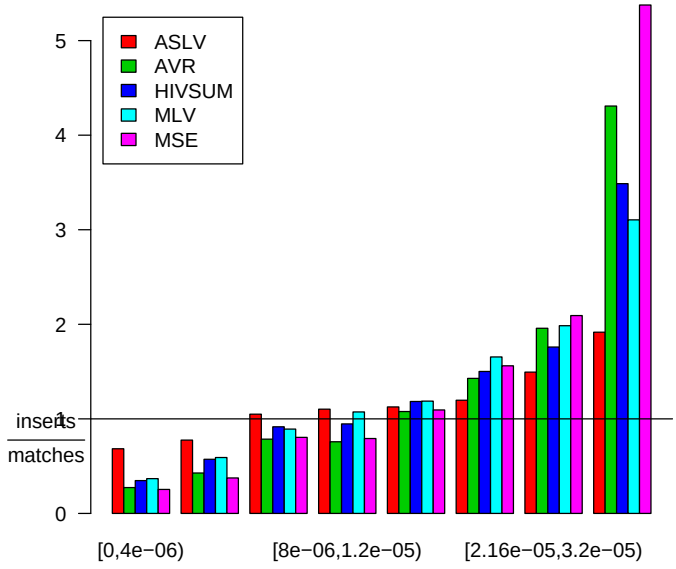
Here the effect of density of CpG islands is studied:



	coef	se	z	p
ASLV	0.646	0.0739	8.75	2.15e-18
AVR	1.330	0.0179	74.10	0.00e+00
HIVSUM	1.070	0.0354	30.20	8.16e-201
MLV	1.180	0.0569	20.80	8.02e-96
MSE	1.450	0.0179	81.30	0.00e+00

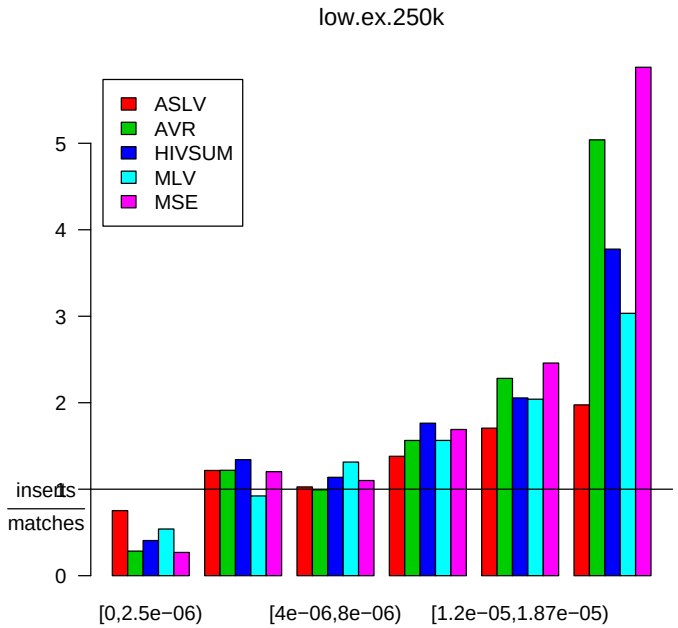
4.4 250 kilobase Window

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 represented on the chip.



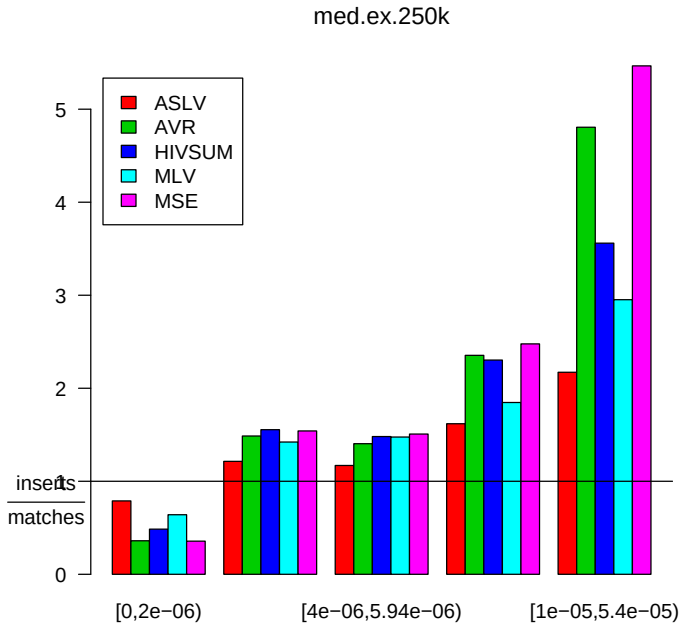
	coef	se	z	p
ASLV	0.533	0.0733	7.28	3.38e-13
AVR	1.560	0.0196	79.80	0.00e+00
HIVSUM	1.270	0.0379	33.50	1.67e-246
MLV	1.210	0.0591	20.40	1.68e-92
MSE	1.700	0.0196	86.80	0.00e+00

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



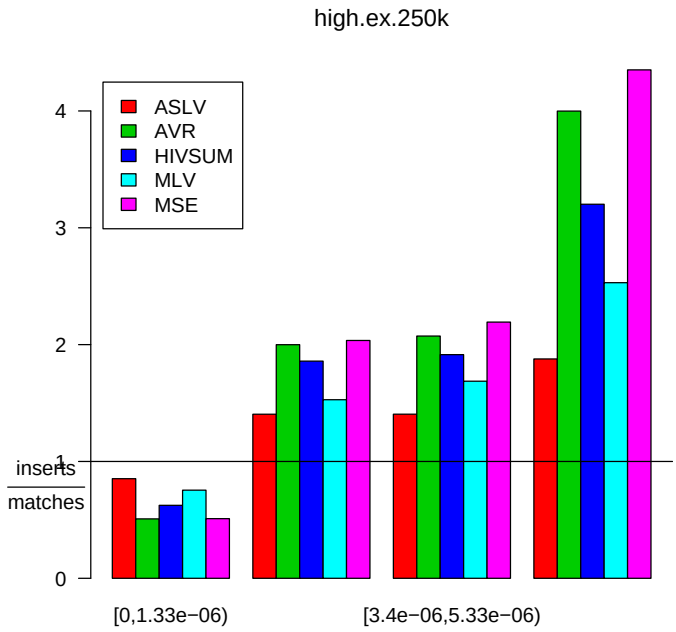
	coef	se	z	p
ASLV	0.599	0.0738	8.11	4.87e-16
AVR	1.930	0.0220	87.40	0.00e+00
HIVSUM	1.490	0.0403	37.00	4.97e-300
MLV	1.140	0.0588	19.30	2.35e-83
MSE	2.050	0.0220	93.40	0.00e+00

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



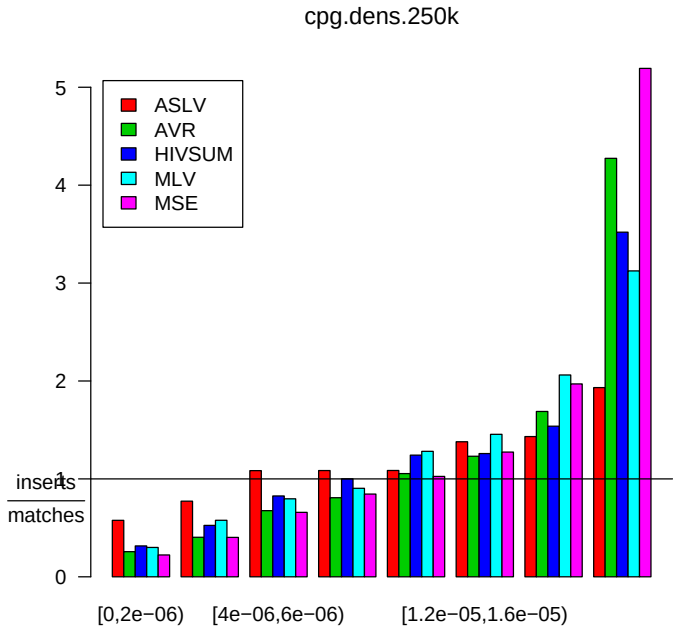
	coef	se	z	p
ASLV	0.63	0.0736	8.56	1.12e-17
AVR	2.09	0.0224	93.40	0.00e+00
HIVSUM	1.62	0.0401	40.40	0.00e+00
MLV	1.05	0.0571	18.40	1.67e-75
MSE	2.17	0.0222	98.00	0.00e+00

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



	coef	se	z	p
ASLV	0.525	0.0762	6.89	5.47e-12
AVR	1.780	0.0191	93.10	0.00e+00
HIVSUM	1.370	0.0356	38.50	0.00e+00
MLV	0.952	0.0560	17.00	8.89e-65
MSE	1.840	0.0189	97.00	0.00e+00

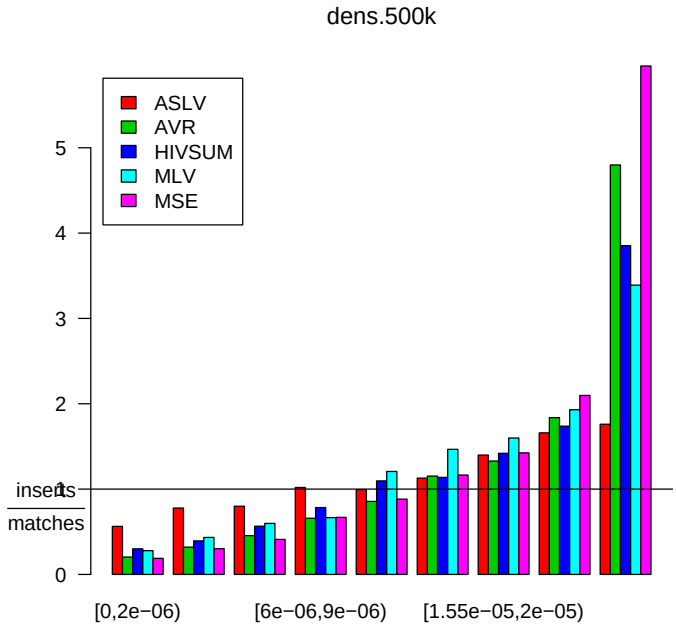
Here the effect of density of CpG islands is studied:



	coef	se	z	p
ASLV	0.538	0.0737	7.3	2.84e-13
AVR	1.380	0.0184	75.1	0.00e+00
HIVSUM	1.100	0.0360	30.5	6.72e-204
MLV	1.200	0.0579	20.7	4.25e-95
MSE	1.500	0.0183	81.9	0.00e+00

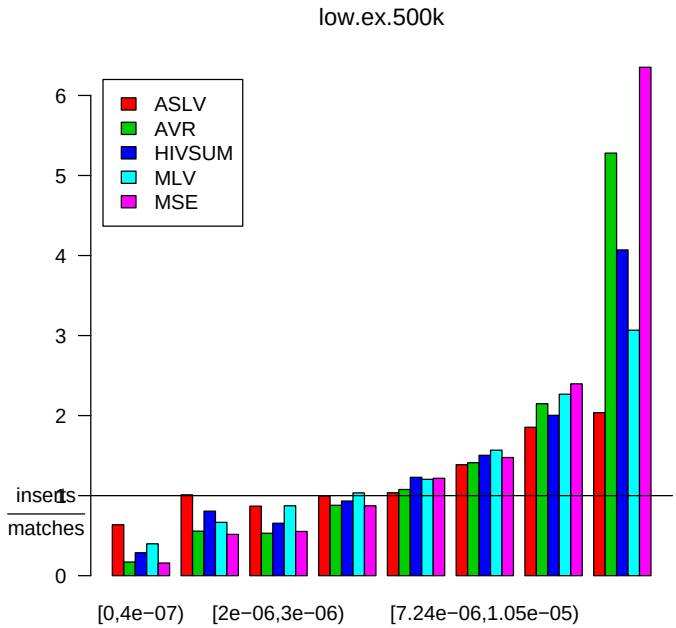
4.5 500 kilobase Window

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 represented on the chip.



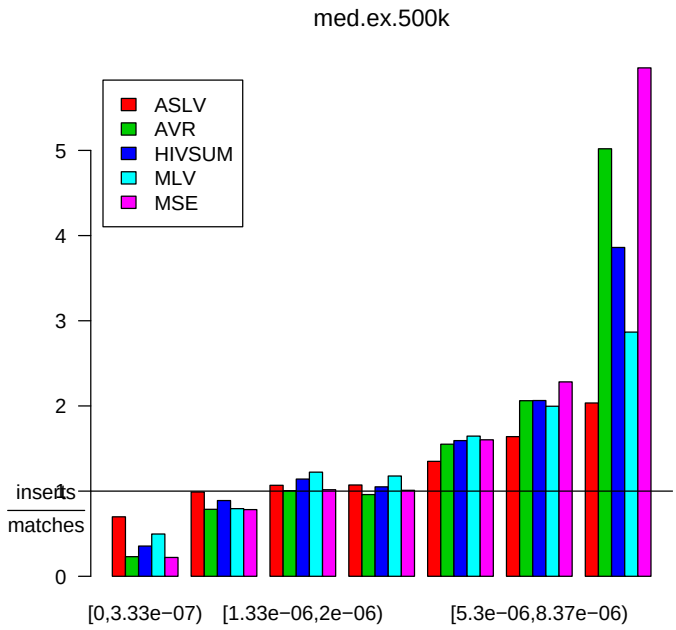
	coef	se	z	p
ASLV	0.558	0.0738	7.56	4.14e-14
AVR	1.520	0.0198	76.90	0.00e+00
HIVSUM	1.220	0.0382	31.90	2.23e-223
MLV	1.320	0.0615	21.50	2.49e-102
MSE	1.650	0.0197	83.70	0.00e+00

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



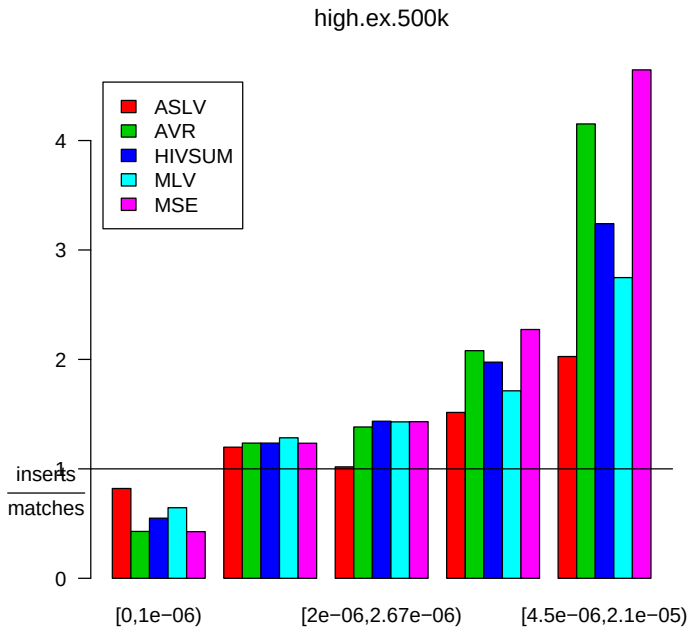
	coef	se	z	p
ASLV	0.598	0.0739	8.1	5.71e-16
AVR	1.780	0.0211	84.2	0.00e+00
HIVSUM	1.340	0.0390	34.4	3.31e-259
MLV	1.140	0.0591	19.3	2.77e-83
MSE	1.900	0.0210	90.4	0.00e+00

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



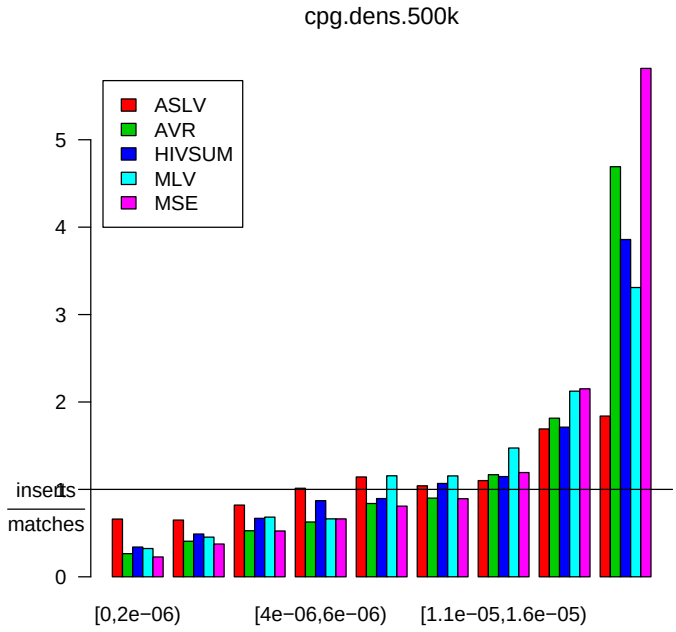
	coef	se	z	p
ASLV	0.582	0.0737	7.89	2.95e-15
AVR	1.720	0.0207	83.20	0.00e+00
HIVSUM	1.340	0.0388	34.50	2.62e-260
MLV	1.100	0.0586	18.80	1.57e-78
MSE	1.820	0.0205	88.40	0.00e+00

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



	coef	se	z	p
ASLV	0.529	0.0736	7.19	6.37e-13
AVR	1.690	0.0205	82.70	0.00e+00
HIVSUM	1.270	0.0384	33.20	1.79e-241
MLV	1.040	0.0582	17.90	1.20e-71
MSE	1.760	0.0202	87.40	0.00e+00

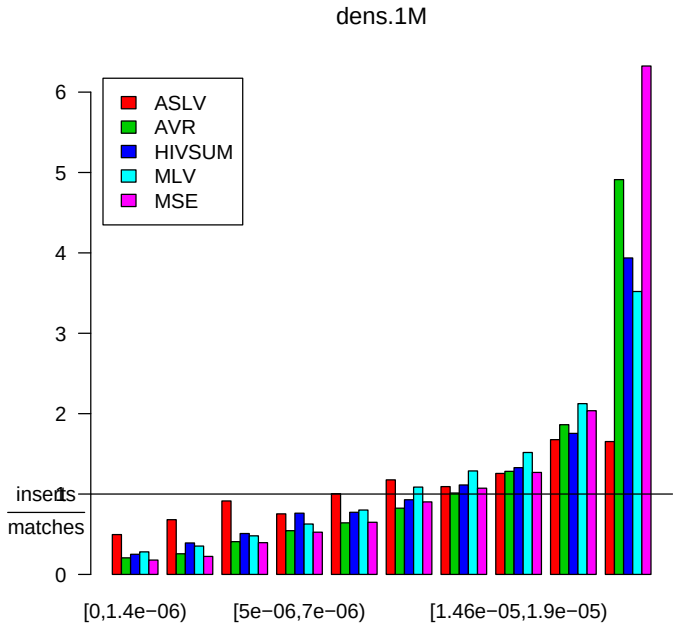
Here the effect of density of CpG islands is studied:



	coef	se	z	p
ASLV	0.481	0.0737	6.52	6.89e-11
AVR	1.300	0.0184	70.40	0.00e+00
HIVSUM	0.998	0.0362	27.60	1.11e-167
MLV	1.170	0.0589	19.90	9.43e-88
MSE	1.430	0.0184	78.00	0.00e+00

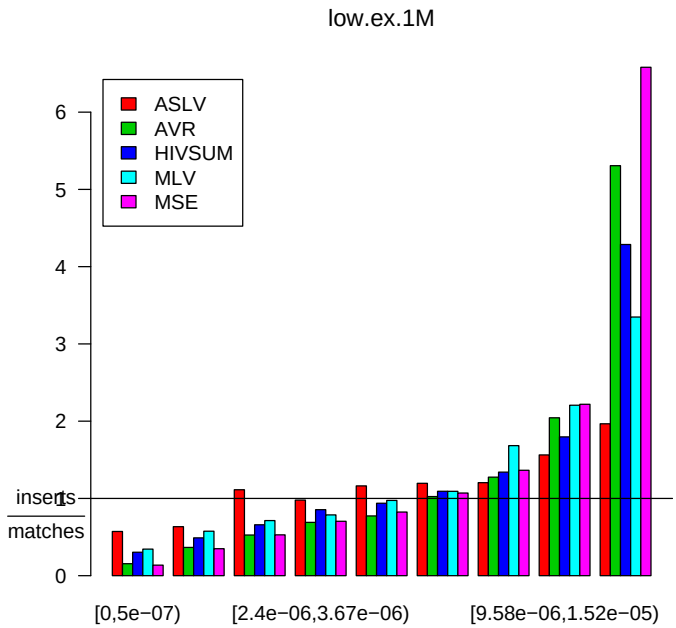
4.6 1 megabase Window

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 represented on the chip.



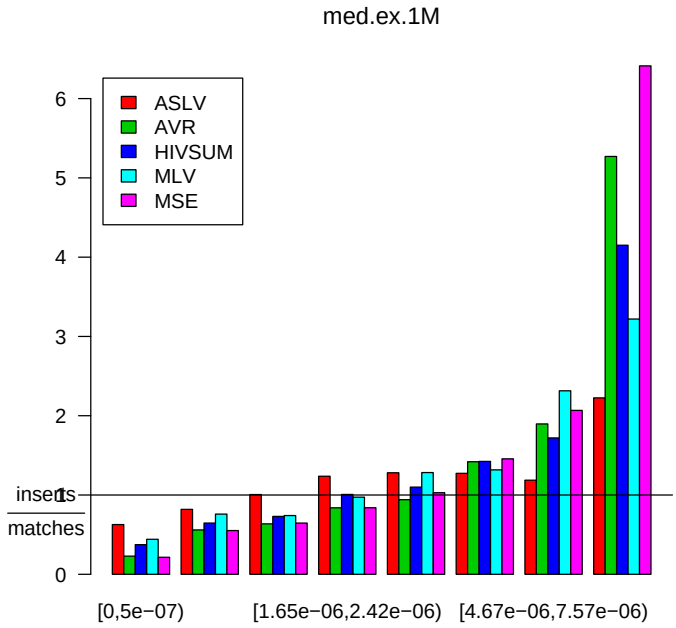
	coef	se	z	p
ASLV	0.55	0.0738	7.44	9.80e-14
AVR	1.44	0.0194	74.40	0.00e+00
HIVSUM	1.10	0.0374	29.40	1.78e-190
MLV	1.23	0.0604	20.30	4.82e-92
MSE	1.57	0.0194	81.10	0.00e+00

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



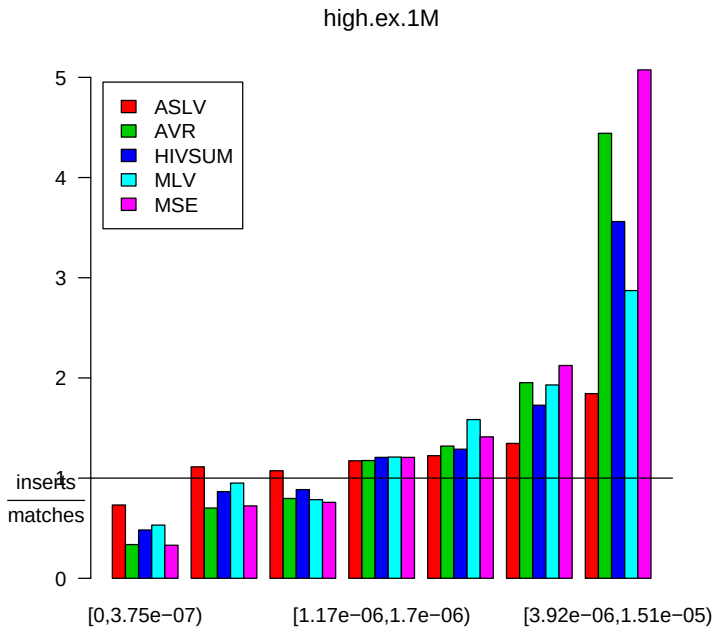
	coef	se	z	p
ASLV	0.577	0.0738	7.82	5.14e-15
AVR	1.560	0.0199	78.50	0.00e+00
HIVSUM	1.160	0.0377	30.70	2.14e-207
MLV	1.120	0.0589	19.00	1.45e-80
MSE	1.680	0.0199	84.40	0.00e+00

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



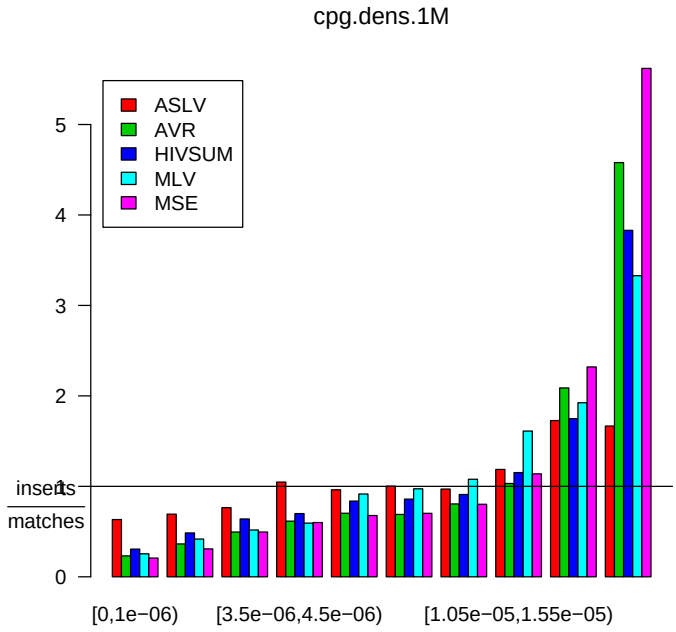
	coef	se	z	p
ASLV	0.598	0.0739	8.1	5.61e-16
AVR	1.540	0.0198	77.6	0.00e+00
HIVSUM	1.190	0.0379	31.3	2.19e-214
MLV	1.070	0.0584	18.3	1.65e-74
MSE	1.630	0.0197	83.0	0.00e+00

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



	coef	se	z	p
ASLV	0.468	0.0735	6.36	2.01e-10
AVR	1.450	0.0194	74.80	0.00e+00
HIVSUM	1.050	0.0371	28.20	4.90e-175
MLV	0.929	0.0576	16.10	2.07e-58
MSE	1.510	0.0191	79.00	0.00e+00

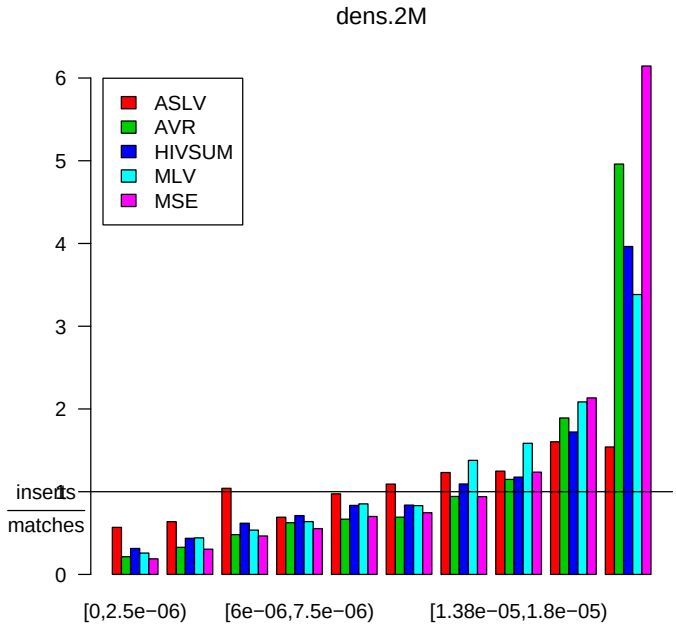
Here the effect of density of CpG islands is studied:



	coef	se	z	p
ASLV	0.444	0.0732	6.07	1.32e-09
AVR	1.180	0.0182	64.70	0.00e+00
HIVSUM	0.865	0.0359	24.10	1.35e-128
MLV	1.080	0.0585	18.40	1.27e-75
MSE	1.320	0.0182	72.40	0.00e+00

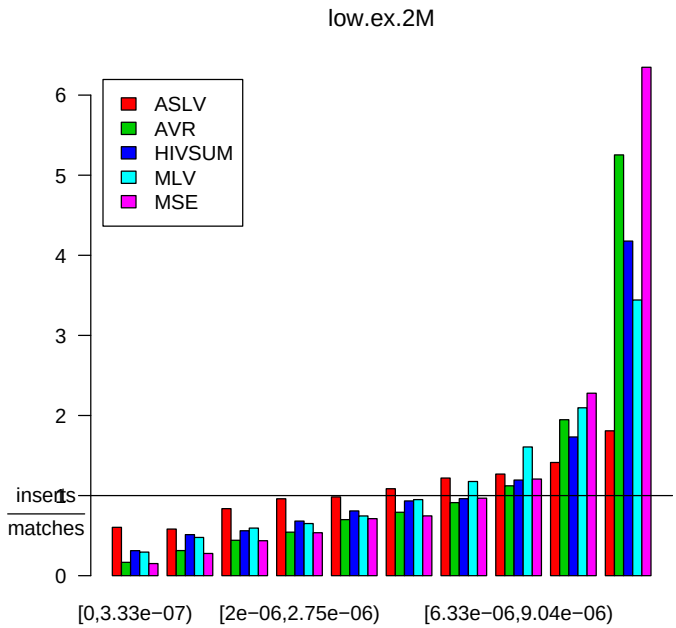
4.7 2 megabase Window

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 represented on the chip.



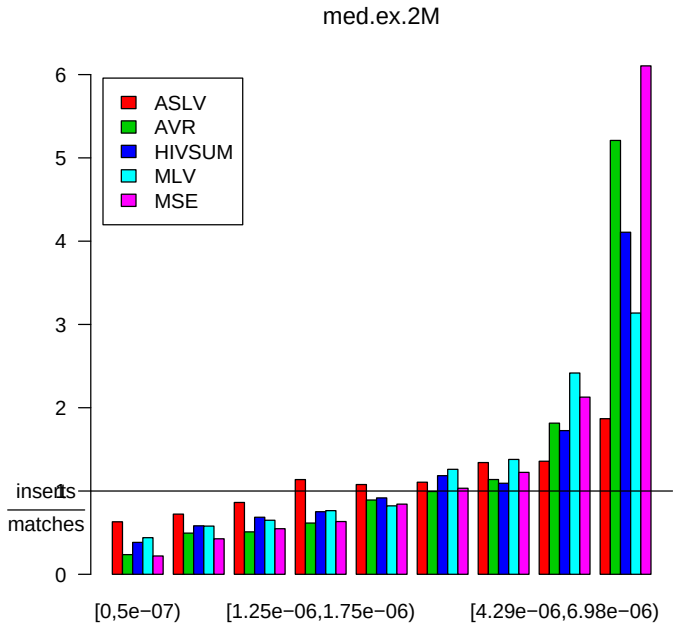
	coef	se	z	p
ASLV	0.529	0.0737	7.17	7.63e-13
AVR	1.280	0.0187	68.30	0.00e+00
HIVSUM	0.973	0.0367	26.50	9.88e-155
MLV	1.130	0.0596	18.90	8.97e-80
MSE	1.410	0.0187	75.60	0.00e+00

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



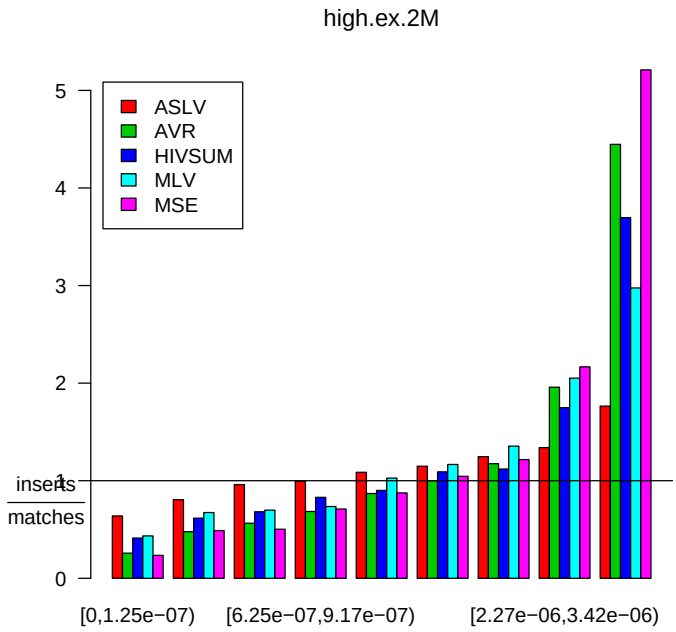
	coef	se	z	p
ASLV	0.506	0.0733	6.9	5.08e-12
AVR	1.370	0.0190	72.1	0.00e+00
HIVSUM	0.996	0.0367	27.1	6.71e-162
MLV	1.110	0.0594	18.6	1.30e-77
MSE	1.470	0.0189	77.8	0.00e+00

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



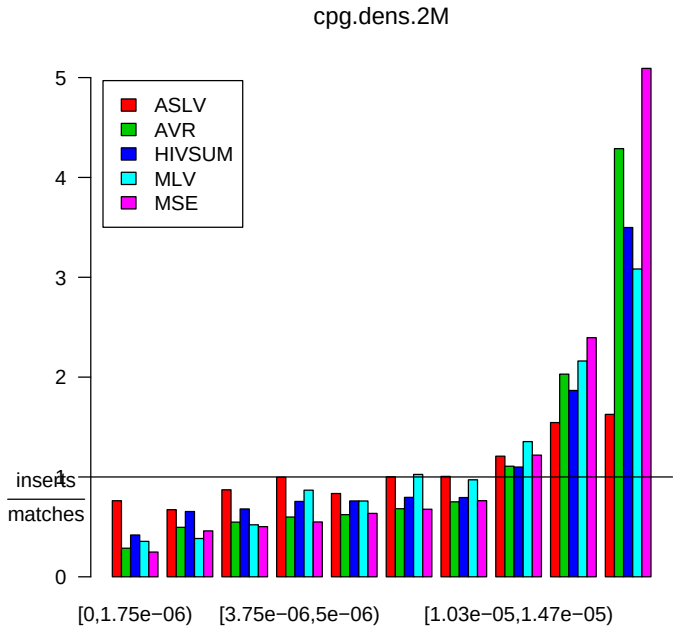
	coef	se	z	p
ASLV	0.497	0.0734	6.77	1.31e-11
AVR	1.410	0.0191	73.80	0.00e+00
HIVSUM	1.050	0.0370	28.40	6.02e-178
MLV	1.040	0.0582	17.90	1.01e-71
MSE	1.510	0.0190	79.40	0.00e+00

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



	coef	se	z	p
ASLV	0.442	0.0732	6.05	1.48e-09
AVR	1.330	0.0189	70.50	0.00e+00
HIVSUM	0.955	0.0366	26.10	1.40e-150
MLV	0.983	0.0581	16.90	2.75e-64
MSE	1.420	0.0187	76.10	0.00e+00

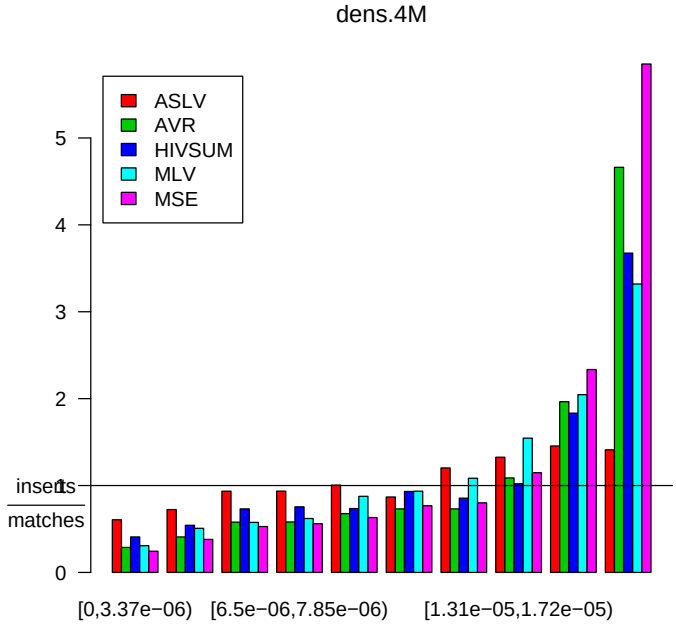
Here the effect of density of CpG islands is studied:



	coef	se	z	p
ASLV	0.373	0.0732	5.1	3.45e-07
AVR	1.120	0.0180	62.1	0.00e+00
HIVSUM	0.779	0.0356	21.9	1.81e-106
MLV	0.963	0.0577	16.7	1.36e-62
MSE	1.270	0.0181	70.6	0.00e+00

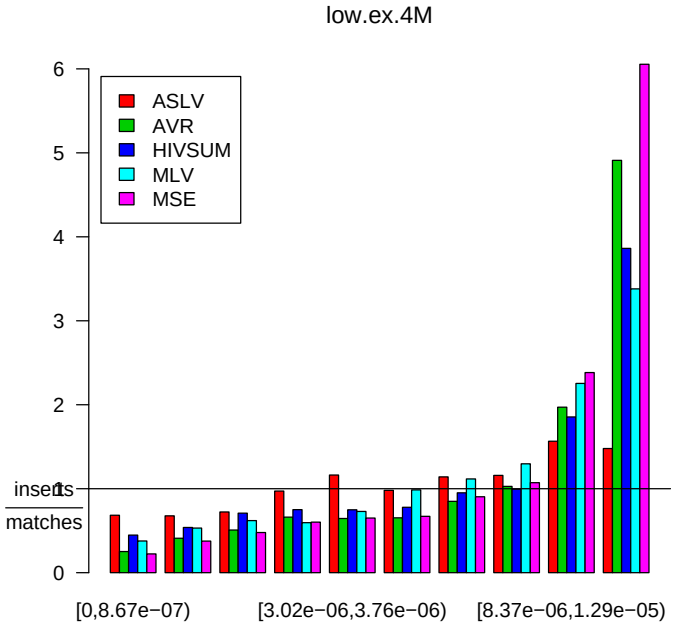
4.8 4 megabase Window

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 represented on the chip.



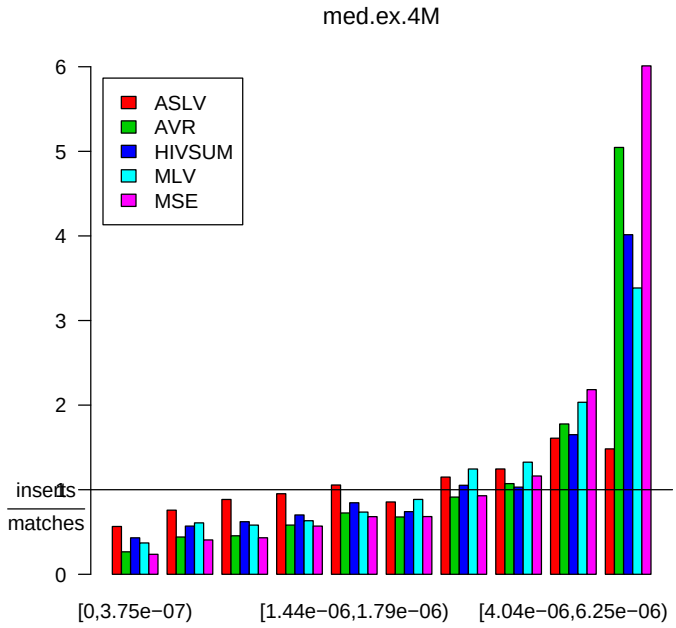
	coef	se	z	p
ASLV	0.374	0.0729	5.12	3.00e-07
AVR	1.140	0.0183	62.50	0.00e+00
HIVSUM	0.839	0.0361	23.30	9.26e-120
MLV	1.020	0.0587	17.40	3.60e-68
MSE	1.320	0.0183	71.80	0.00e+00

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



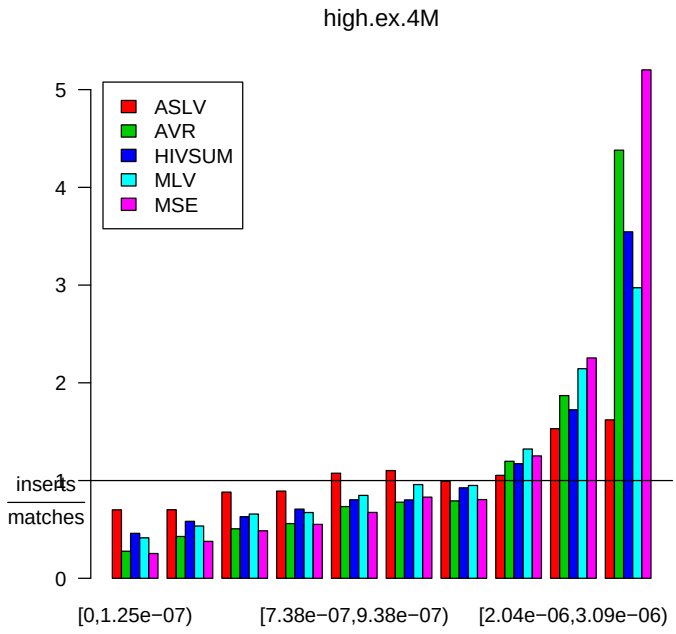
	coef	se	z	p
ASLV	0.378	0.0731	5.17	2.36e-07
AVR	1.180	0.0184	64.10	0.00e+00
HIVSUM	0.827	0.0360	23.00	1.34e-116
MLV	1.040	0.0588	17.70	7.29e-70
MSE	1.330	0.0184	72.30	0.00e+00

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



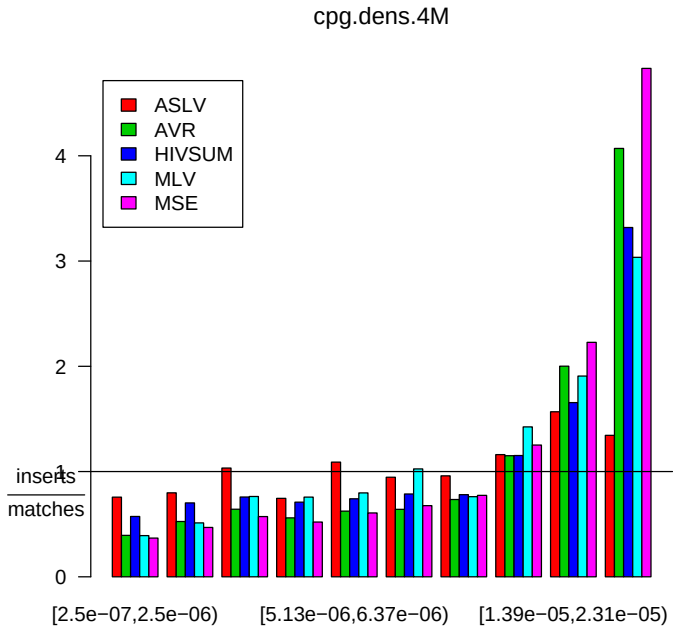
	coef	se	z	p
ASLV	0.369	0.0730	5.05	4.36e-07
AVR	1.180	0.0184	64.20	0.00e+00
HIVSUM	0.839	0.0360	23.30	6.75e-120
MLV	0.993	0.0583	17.00	5.24e-65
MSE	1.330	0.0184	72.50	0.00e+00

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



	coef	se	z	p
ASLV	0.354	0.0729	4.86	1.17e-06
AVR	1.150	0.0183	63.10	0.00e+00
HIVSUM	0.821	0.0360	22.80	3.84e-115
MLV	0.878	0.0574	15.30	8.28e-53
MSE	1.310	0.0183	71.50	0.00e+00

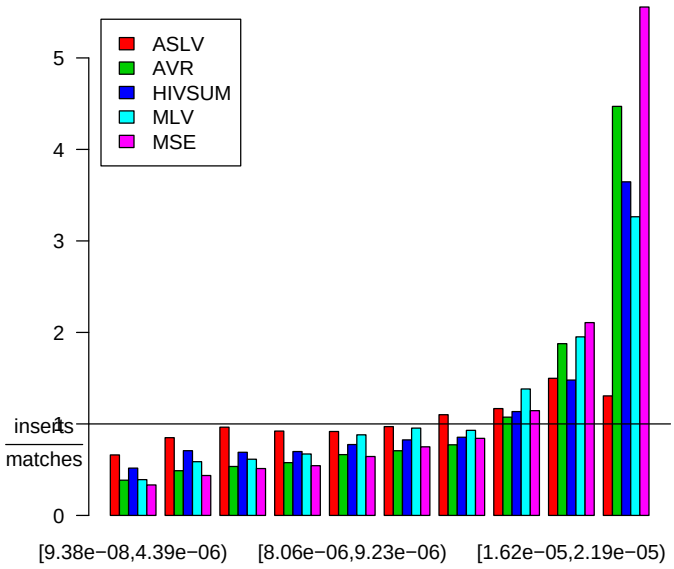
Here the effect of density of CpG islands is studied:




```
         coef     se    z        p
ASLV    0.262 0.0729  3.6 3.16e-04
AVR     1.010 0.0178 56.7 0.00e+00
HIVSUM  0.676 0.0354 19.1 1.82e-81
MLV     0.816 0.0571 14.3 2.16e-46
MSE     1.170 0.0179 65.7 0.00e+00
```

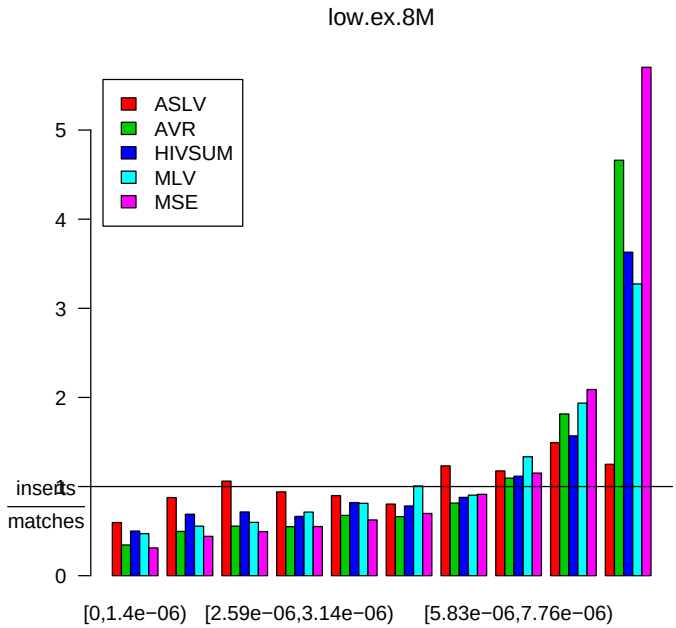
4.9 8 megabase Window

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 represented on the chip.



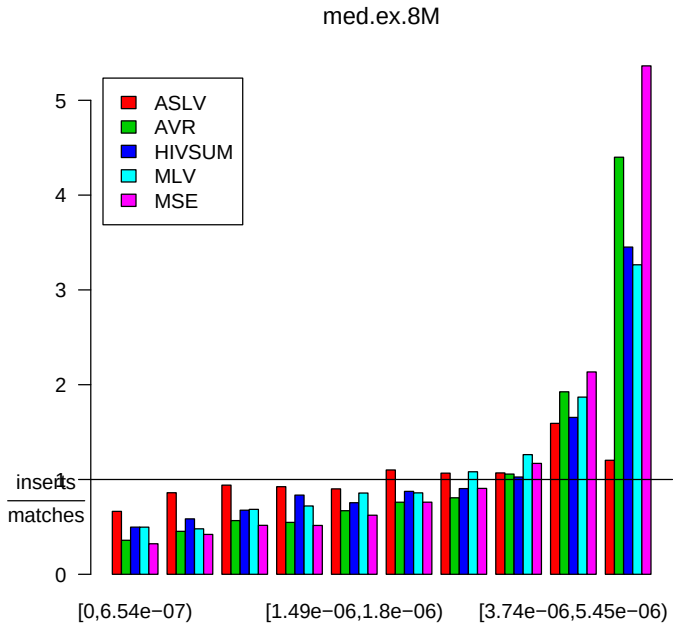
	coef	se	z	p
ASLV	0.316	0.0729	4.33	1.49e−05
AVR	1.070	0.0180	59.50	0.00e+00
HIVSUM	0.726	0.0356	20.40	1.53e−92
MLV	0.876	0.0575	15.20	1.87e−52
MSE	1.230	0.0181	68.00	0.00e+00

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



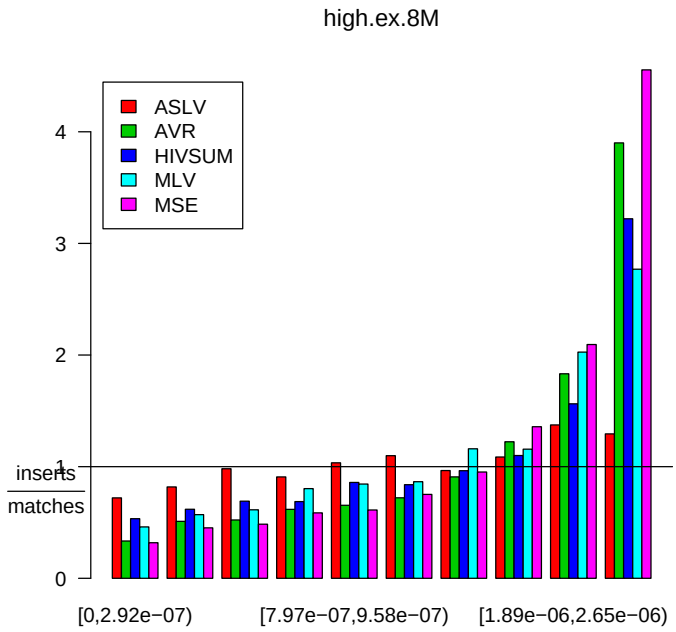
	coef	se	z	p
ASLV	0.294	0.0730	4.03	5.58e-05
AVR	1.100	0.0182	60.30	0.00e+00
HIVSUM	0.725	0.0356	20.40	2.73e-92
MLV	0.867	0.0573	15.10	9.42e-52
MSE	1.260	0.0182	69.20	0.00e+00

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



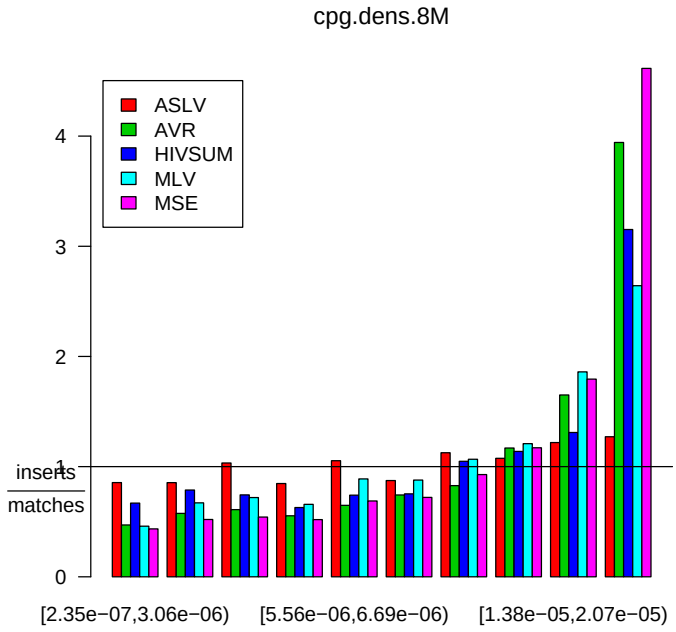
	coef	se	z	p
ASLV	0.325	0.0730	4.44	8.83e-06
AVR	1.110	0.0181	61.00	0.00e+00
HIVSUM	0.747	0.0356	21.00	1.40e-97
MLV	0.831	0.0571	14.50	6.14e-48
MSE	1.270	0.0182	69.90	0.00e+00

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



	coef	se	z	p
ASLV	0.250	0.0729	3.43	6.12e-04
AVR	1.080	0.0181	60.00	0.00e+00
HIVSUM	0.722	0.0356	20.30	1.51e-91
MLV	0.795	0.0567	14.00	1.05e-44
MSE	1.240	0.0181	68.40	0.00e+00

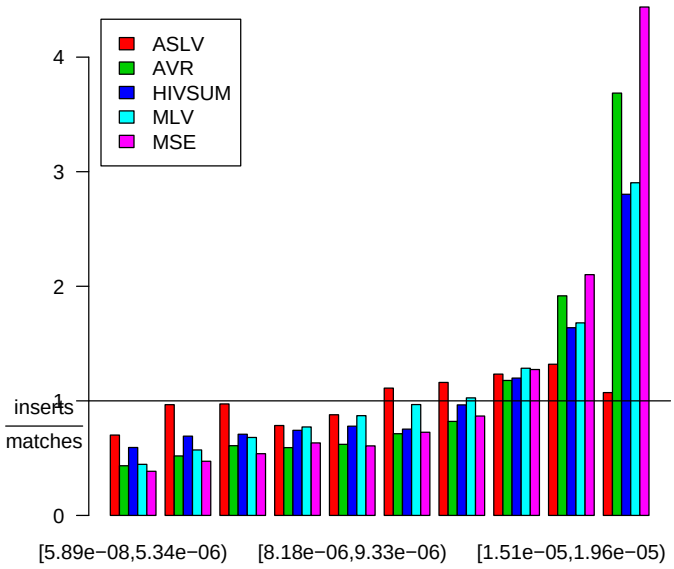
Here the effect of density of CpG islands is studied:



	coef	se	z	p
ASLV	0.172	0.0732	2.35	1.89e-02
AVR	0.953	0.0177	54.00	0.00e+00
HIVSUM	0.632	0.0353	17.90	7.11e-72
MLV	0.748	0.0566	13.20	7.17e-40
MSE	1.080	0.0176	61.00	0.00e+00

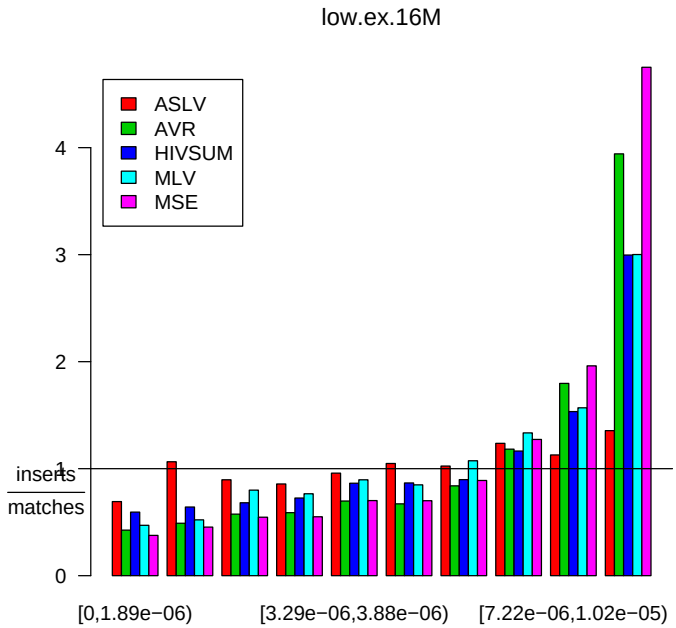
4.10 16 megabase Window

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 represented on the chip.



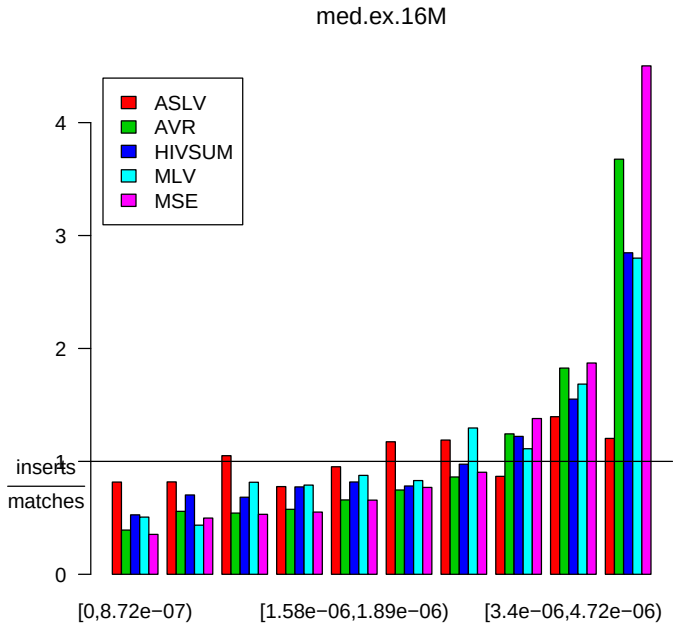
	coef	se	z	p
ASLV	0.318	0.0729	4.36	1.32e-05
AVR	1.010	0.0179	56.40	0.00e+00
HIVSUM	0.661	0.0354	18.70	5.70e-78
MLV	0.768	0.0567	13.50	1.00e-41
MSE	1.120	0.0178	63.30	0.00e+00

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



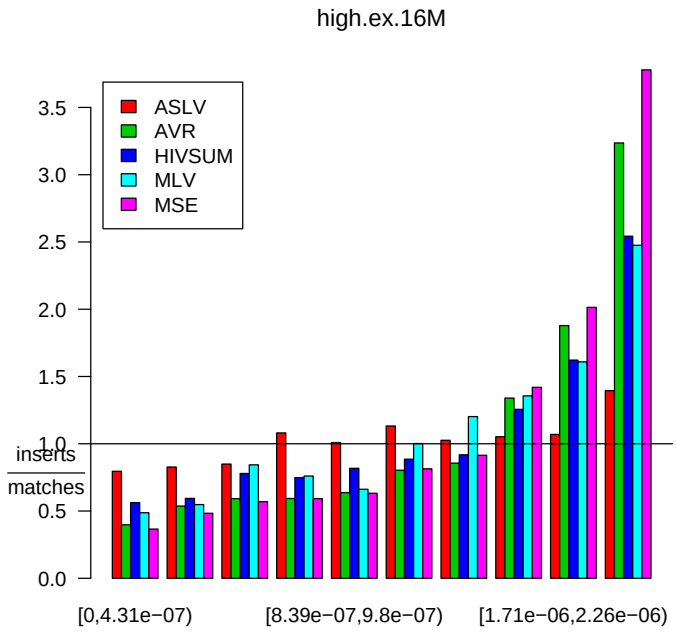
	coef	se	z	p
ASLV	0.230	0.0728	3.16	1.56e-03
AVR	1.010	0.0179	56.30	0.00e+00
HIVSUM	0.671	0.0354	18.90	5.24e-80
MLV	0.724	0.0565	12.80	1.16e-37
MSE	1.130	0.0177	63.60	0.00e+00

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



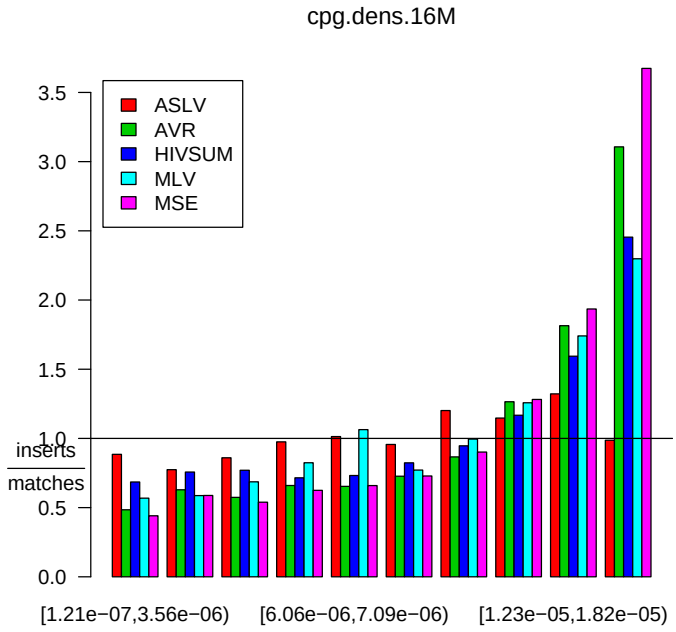
	coef	se	z	p
ASLV	0.272	0.0728	3.74	1.83e-04
AVR	1.030	0.0179	57.50	0.00e+00
HIVSUM	0.672	0.0354	19.00	4.37e-80
MLV	0.739	0.0567	13.00	7.33e-39
MSE	1.150	0.0178	64.60	0.00e+00

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



	coef	se	z	p
ASLV	0.200	0.0728	2.75	5.91e-03
AVR	1.010	0.0179	56.70	0.00e+00
HIVSUM	0.673	0.0354	19.00	1.69e-80
MLV	0.773	0.0570	13.60	5.94e-42
MSE	1.120	0.0177	63.10	0.00e+00

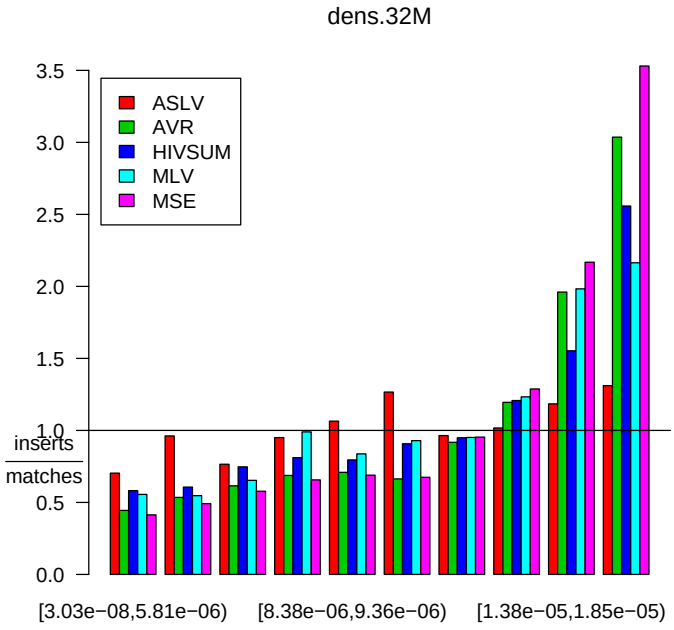
Here the effect of density of CpG islands is studied:




```
         coef     se     z        p
ASLV    0.229 0.0730  3.14 1.69e-03
AVR     0.893 0.0175 51.00 0.00e+00
HIVSUM  0.591 0.0352 16.80 2.52e-63
MLV     0.586 0.0560 10.50 1.38e-25
MSE     1.010 0.0175 57.60 0.00e+00
```

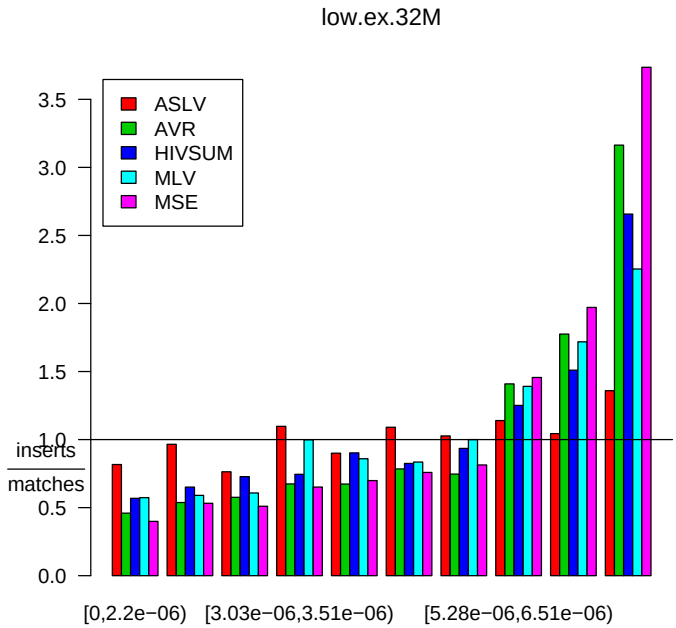
4.11 32 megabase Window

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 represented on the chip.



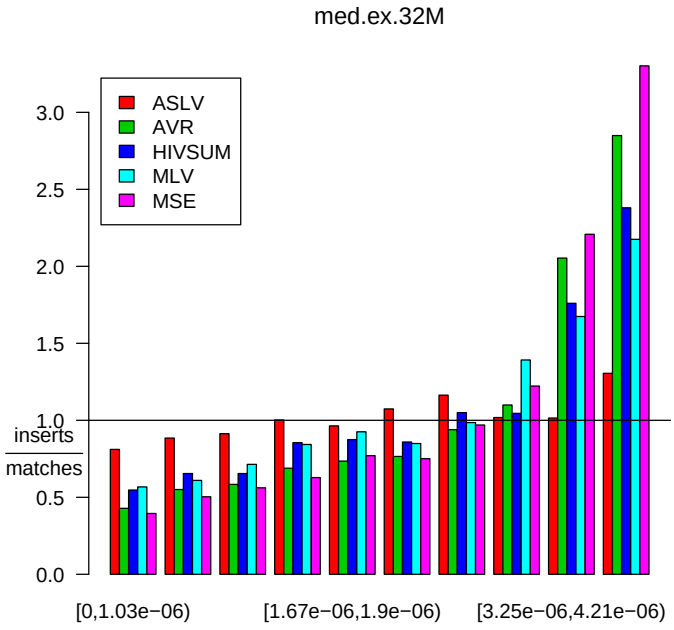
	coef	se	z	p
ASLV	0.251	0.0731	3.44	5.92e-04
AVR	0.893	0.0175	51.00	0.00e+00
HIVSUM	0.650	0.0354	18.40	3.02e-75
MLV	0.655	0.0563	11.60	2.61e-31
MSE	1.010	0.0174	58.20	0.00e+00

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



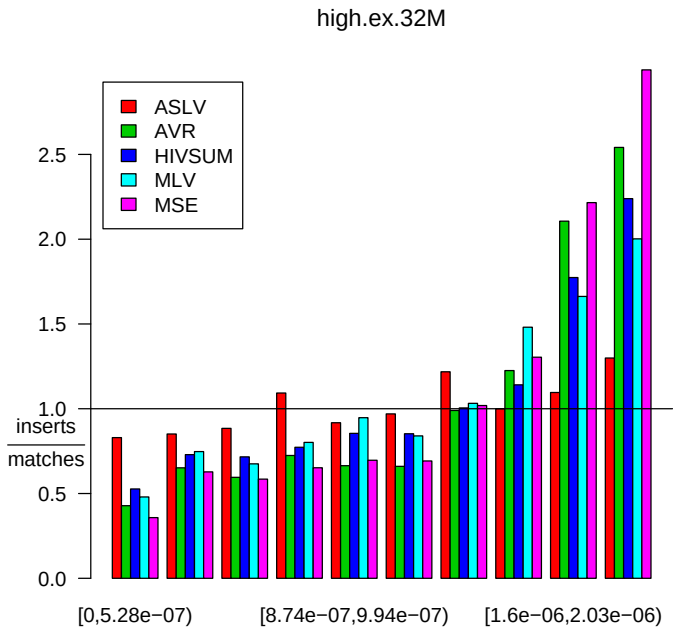
	coef	se	z	p
ASLV	0.206	0.0732	2.81	4.97e-03
AVR	0.929	0.0176	52.70	0.00e+00
HIVSUM	0.626	0.0353	17.70	3.17e-70
MLV	0.640	0.0562	11.40	5.58e-30
MSE	1.040	0.0175	59.30	0.00e+00

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



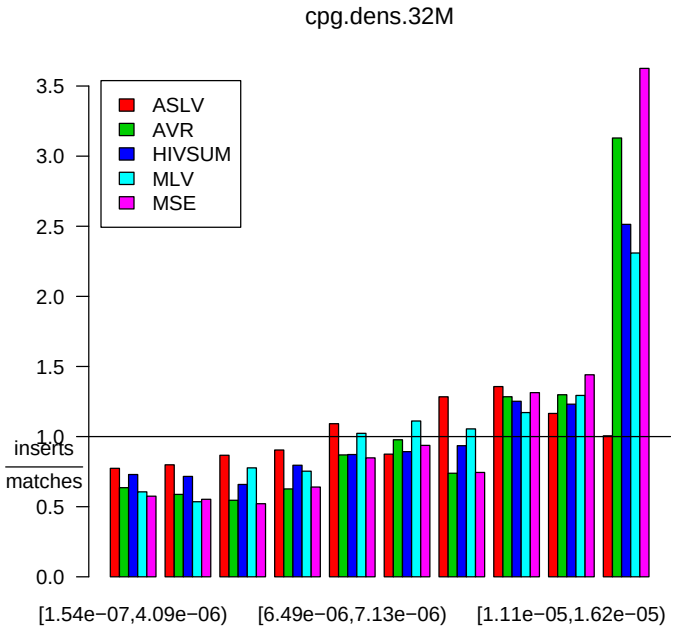
	coef	se	z	p
ASLV	0.192	0.0732	2.62	8.78e-03
AVR	0.893	0.0175	51.10	0.00e+00
HIVSUM	0.632	0.0354	17.90	1.38e-71
MLV	0.618	0.0561	11.00	2.88e-28
MSE	1.000	0.0174	57.60	0.00e+00

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



	coef	se	z	p
ASLV	0.189	0.0729	2.59	9.57e-03
AVR	0.852	0.0174	49.00	0.00e+00
HIVSUM	0.624	0.0353	17.70	8.46e-70
MLV	0.622	0.0562	11.10	1.86e-28
MSE	0.969	0.0173	56.00	0.00e+00

Here the effect of density of CpG islands is studied:

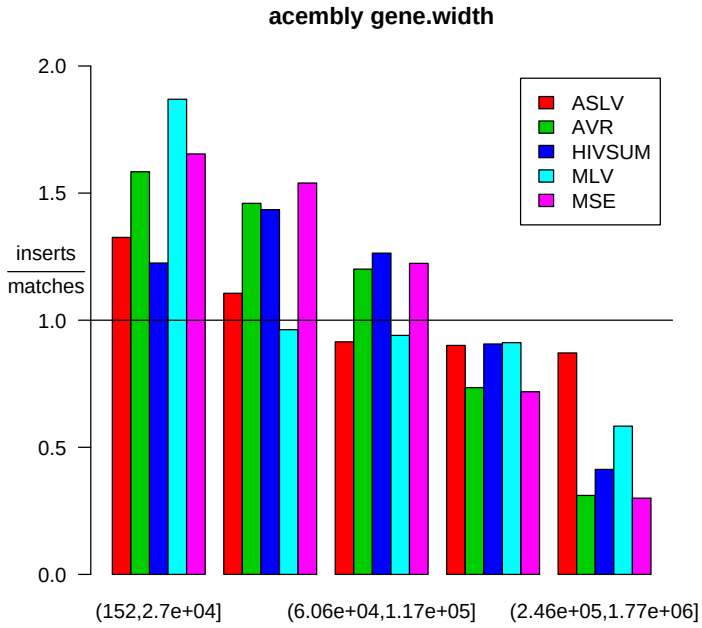


	coef	se	z	p
ASLV	0.259	0.0734	3.53	4.16e-04
AVR	0.767	0.0172	44.50	0.00e+00
HIVSUM	0.533	0.0350	15.20	2.84e-52
MLV	0.594	0.0559	10.60	2.51e-26
MSE	0.843	0.0170	49.50	0.00e+00

5 Juxtaposition with Gene Start and End Positions

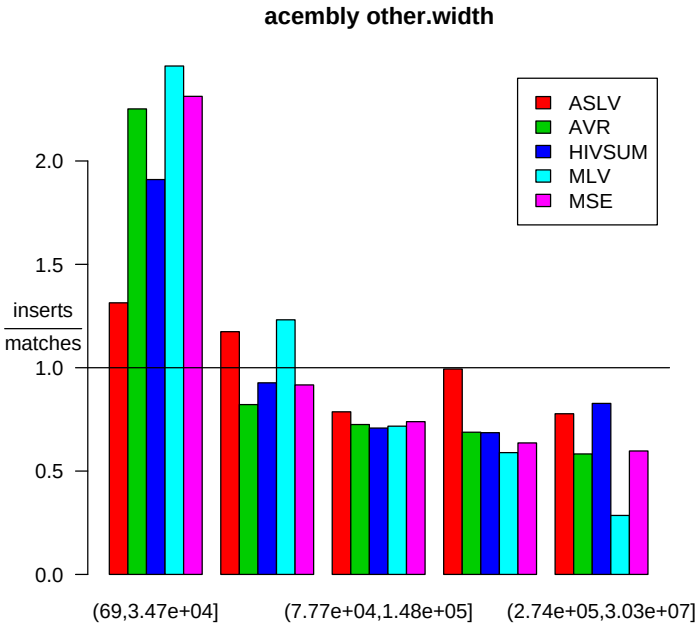
5.1 Acembly Annotations

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. The table following the barplot shows the p-values for a test of the hypothesis that the proportions in each of the categories that define the bars are equal in the insertions and their matches. This p-value is obtained from the $5 \times 2 \times k$ table of counts defined by gene width category, insertion/match status, and stratum (consisting of an insertion and its matched sites) using a likelihood ratio test for the hypothesis of no association between gene width category and insertion/match status. The test used compared the log-linear model [1] with all two-way configurations to that with no gene width category and insertion/match status configuration.



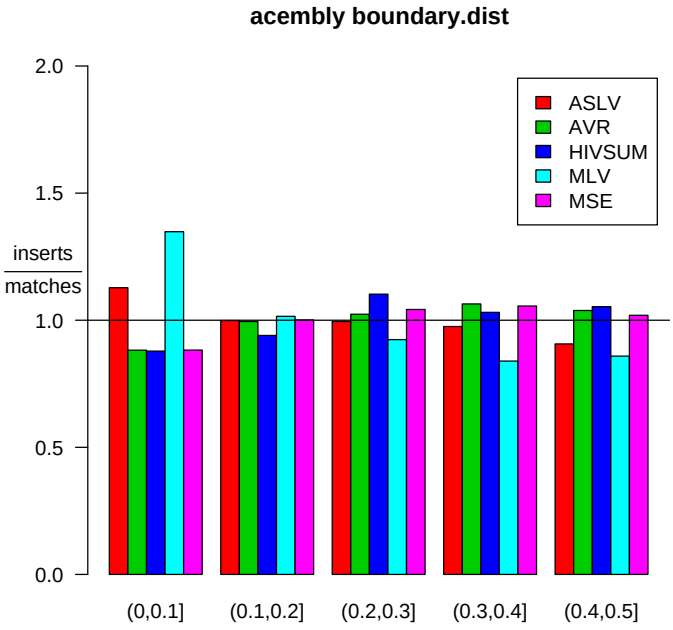
ASLV	AVR	HIVSUM	MLV	MSE
5.72e-03	0.00e+00	2.73e-105	1.92e-32	0.00e+00

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



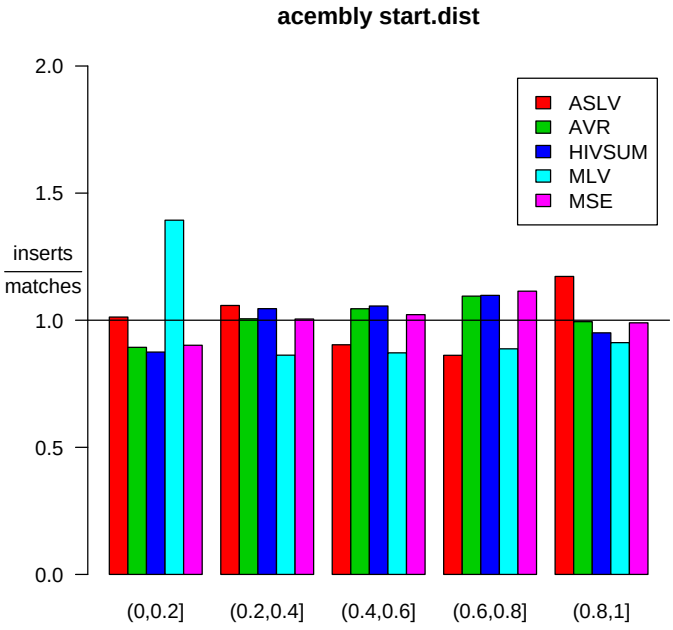

```
    ASLV       AVR    HIVSUM       MLV       MSE
5.51e-03 1.72e-138  5.50e-21  2.69e-49 2.45e-136
```

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.



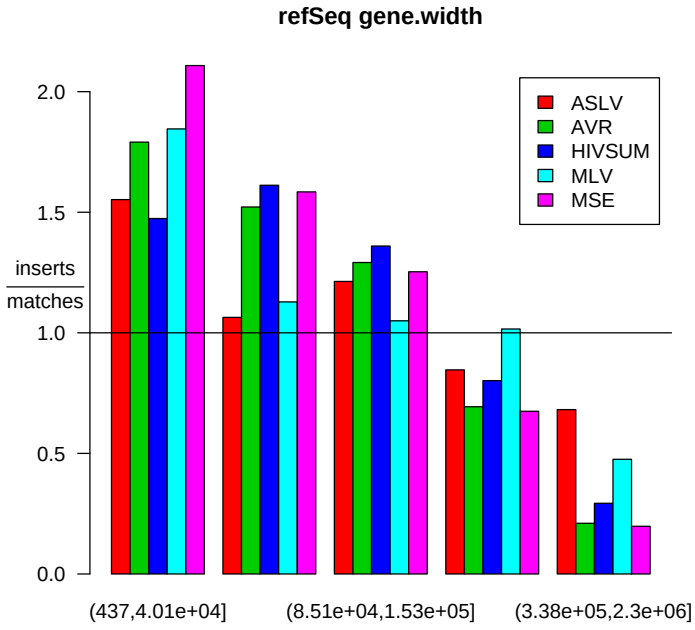

```
    ASLV      AVR   HIVSUM      MLV      MSE
3.84e-01 3.84e-17 6.07e-05 6.43e-10 7.13e-17
```

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.

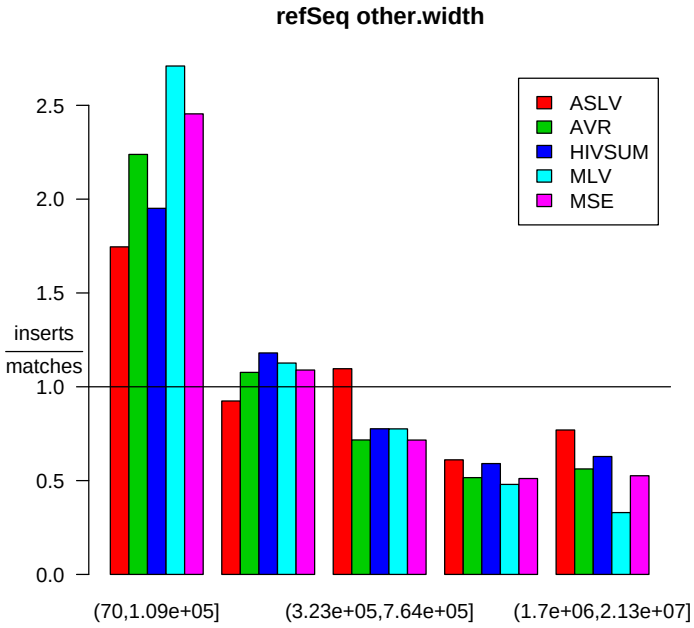



```
    ASLV      AVR   HIVSUM      MLV      MSE
6.49e-02 2.69e-20 4.65e-05 4.15e-11 1.08e-18
```

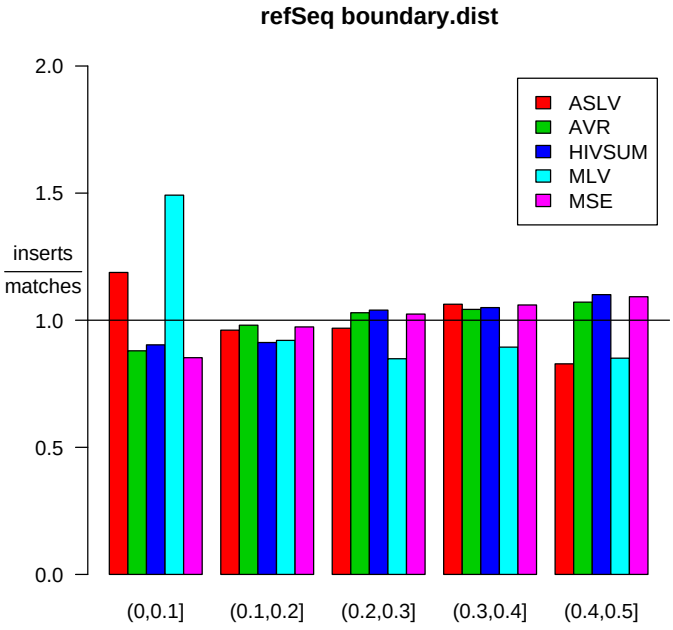
5.2 RefSeq Annotations



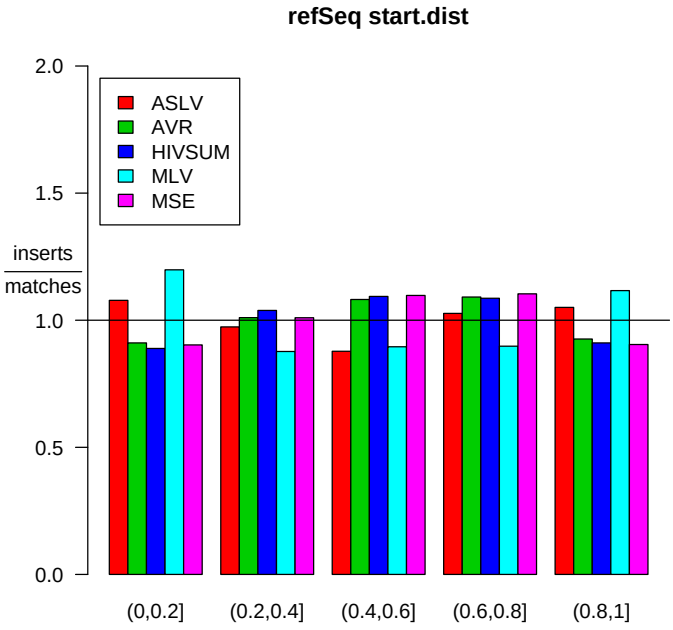

```
    ASLV       AVR    HIVSUM       MLV       MSE
1.51e-05  0.00e+00 2.85e-169  5.91e-26  0.00e+00
```




```
    ASLV      AVR   HIVSUM      MLV      MSE
1.76e-13 0.00e+00 2.68e-46 3.74e-96 0.00e+00
```

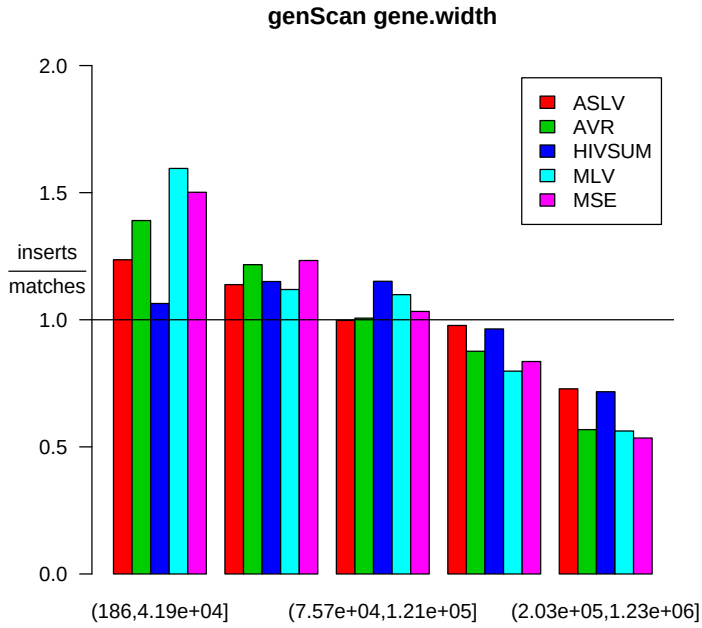



```
    ASLV      AVR   HIVSUM      MLV      MSE
1.90e-02 5.56e-19 1.01e-04 2.31e-16 1.94e-31
```

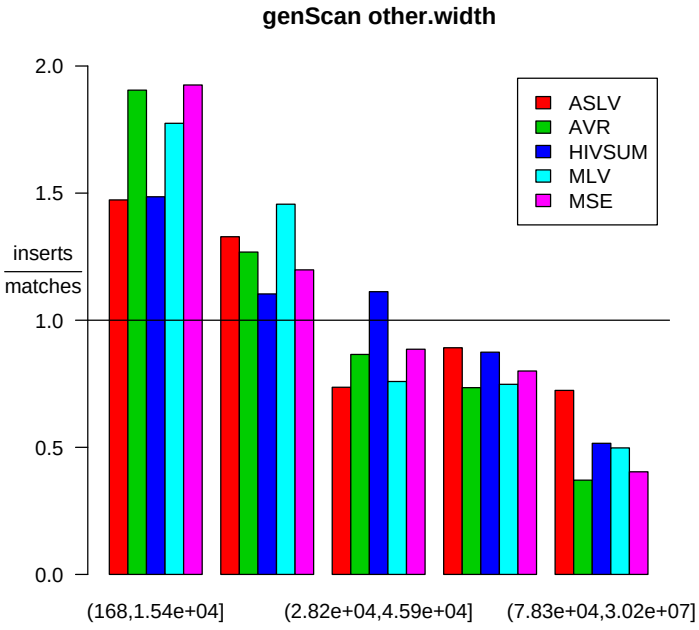



```
    ASLV      AVR   HIVSUM      MLV      MSE
3.85e-01 2.63e-21 1.24e-05 3.73e-04 2.11e-27
```

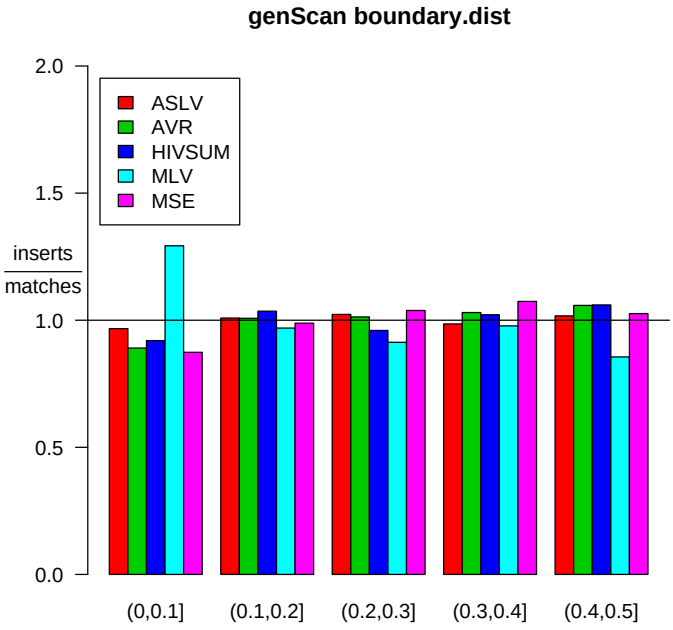
5.3 genScan Annotations



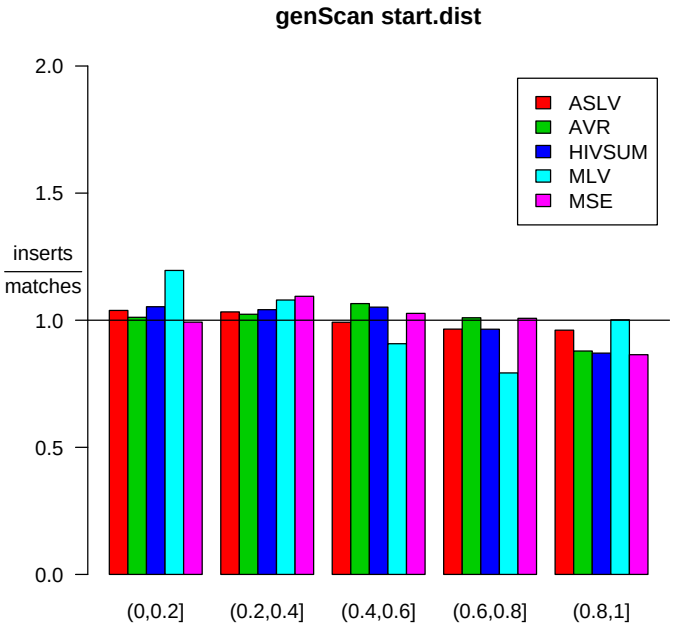

```
    ASLV       AVR    HIVSUM       MLV       MSE
8.72e-04 1.33e-288  1.06e-16  3.85e-24  0.00e+00
```




```
    ASLV        AVR     HIVSUM        MLV        MSE
5.68e-04 1.19e-151  4.77e-20  1.33e-21 2.13e-141
```

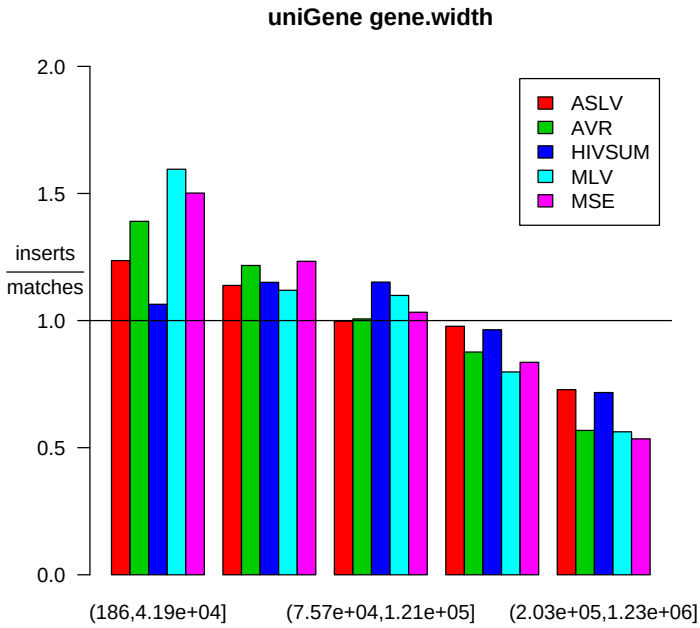



```
    ASLV      AVR   HIVSUM      MLV      MSE
9.86e-01 1.05e-13 4.14e-02 2.50e-06 1.58e-20
```

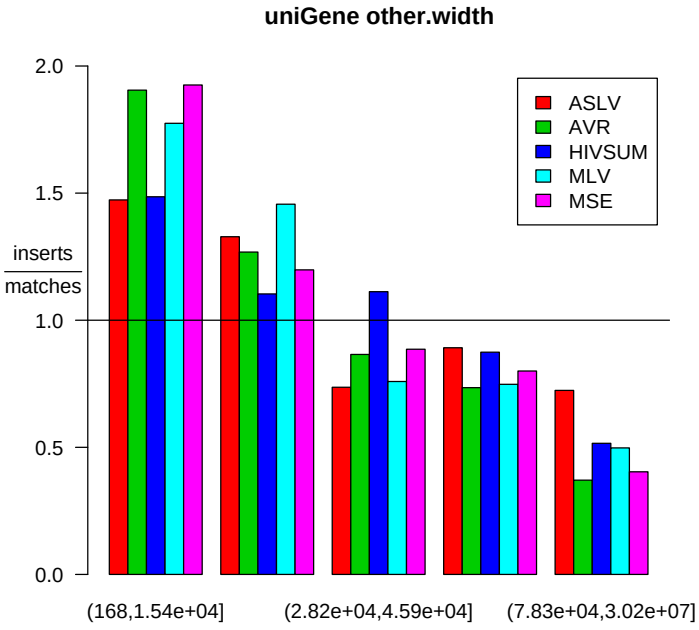



```
    ASLV      AVR   HIVSUM      MLV      MSE
9.78e-01 3.13e-13 4.27e-03 5.68e-05 4.07e-21
```

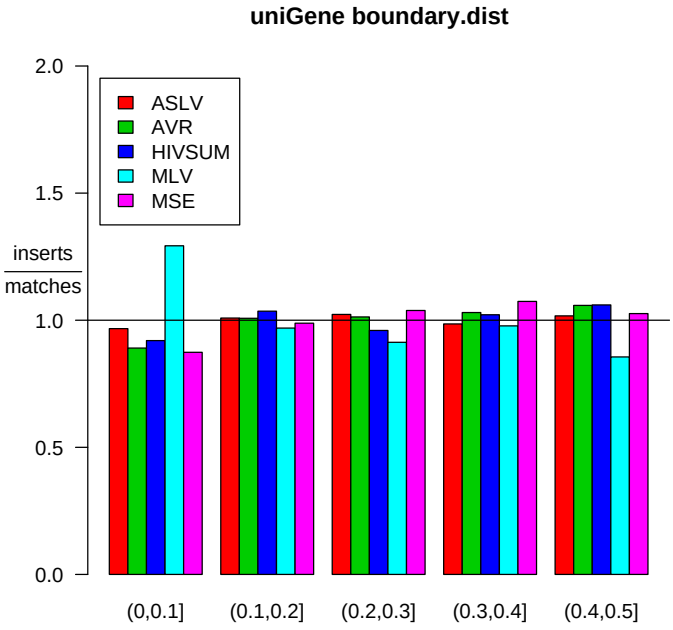
5.4 uniGene Annotations



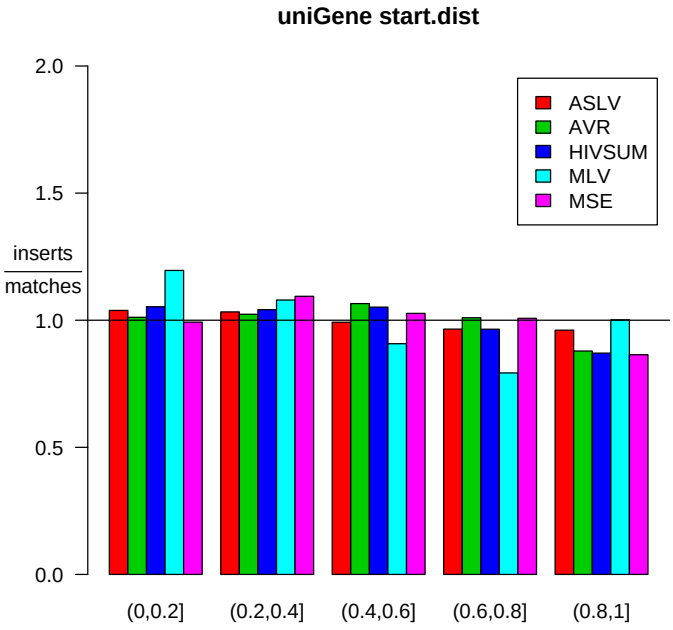

```
    ASLV       AVR    HIVSUM       MLV       MSE
8.72e-04 1.33e-288  1.06e-16  3.85e-24  0.00e+00
```




```
    ASLV        AVR     HIVSUM        MLV        MSE
5.68e-04 1.19e-151  4.77e-20  1.33e-21 2.13e-141
```



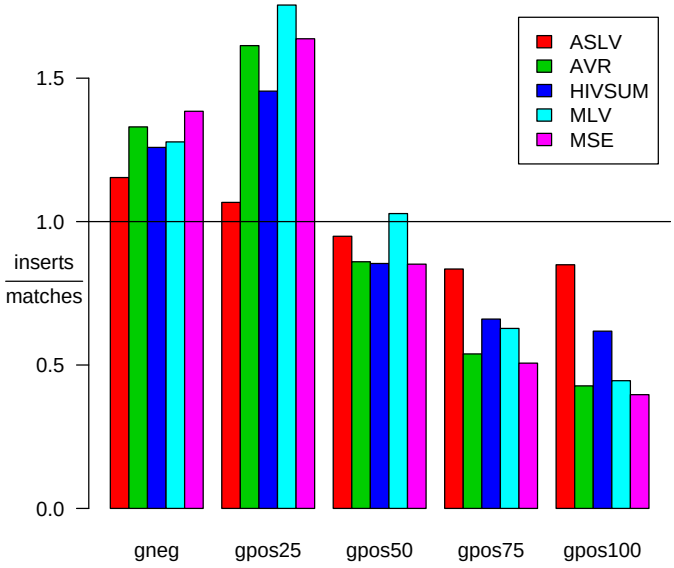

```
    ASLV      AVR   HIVSUM      MLV      MSE
9.86e-01 1.05e-13 4.14e-02 2.50e-06 1.58e-20
```




```
    ASLV      AVR   HIVSUM      MLV      MSE
9.78e-01 3.13e-13 4.27e-03 5.68e-05 4.07e-21
```

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 $< 2.22e-16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites (comparing each category of Giemsa staining to 'gneg') along with their standard errors, z statistics, and p-values:

```
                   coef     se      z         p
cyto.typegpos100 -1.120 0.0191 -58.50  0.00e+00
cyto.typegpos25   0.179 0.0184   9.76  1.74e-22
cyto.typegpos50  -0.438 0.0164 -26.70 8.94e-157
cyto.typegpos75  -0.904 0.0193 -46.80  0.00e+00
```

References

[1] Yvonne M.M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete multivariate analyses: Theory and practice* (MIT Press, 1975).

[2] P. McCullagh and John A. Nelder. *Generalized linear models.* (Chapman & Hall ltd, 1999).

[3] 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.