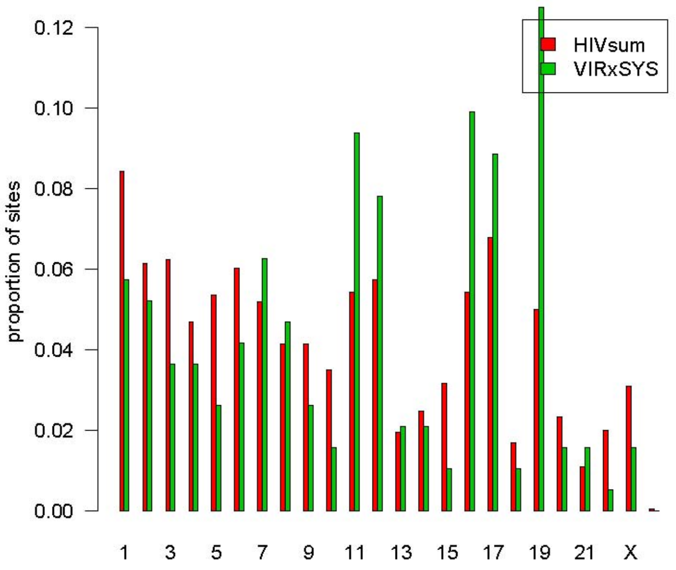
Association of Various Genomic Features with VRX496 Integration Sites

1. Introduction

Here we examine the association of integration sites with various genomic features. The numbers of integration events analyzed are shown below:

Origin.of.data.set	# integrations
HIVsum	2969
VIRxSYS	192

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

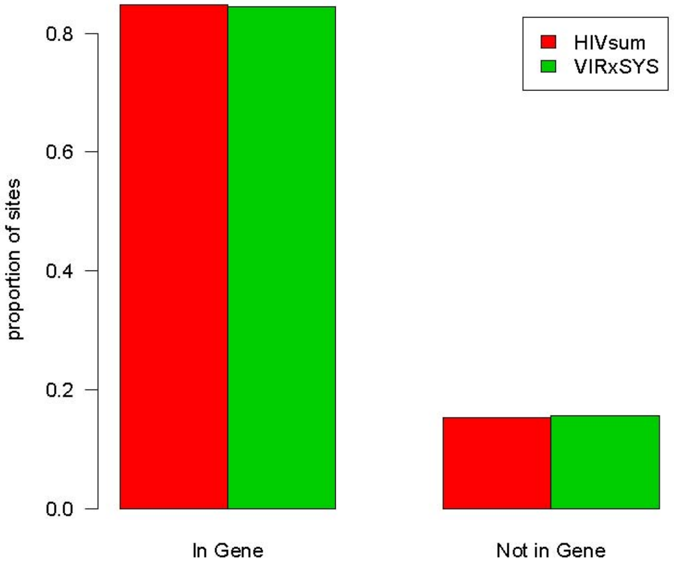


Are there chromosomes that are particularly favored for integration by one group over the other? This was tested for statistical significance. The test performed used the likelihood ratio statistic for the logistic regression model [reviewed in McCulllagh et al (17)] as implemented by the glm function 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 0.00136.

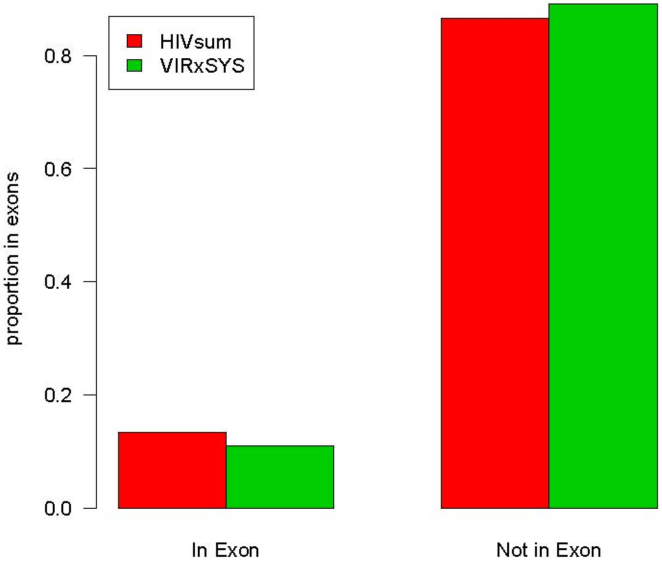
2. Preference for Genes

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' 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.



Is there a difference in the tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.901. 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.



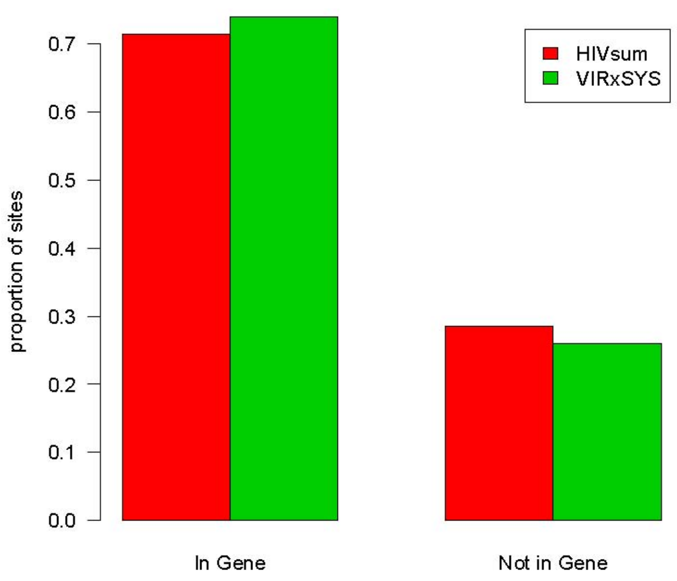
Here is the table of coefficients of the log ratio of intensities along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	-2.72000	0.189	-14.4000	4.31e-47
in.gene	0.00838	0.208	0.0404	9.68e-01
in.exon -	0.23600	0.240	-0.9820	3.26e-01

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.

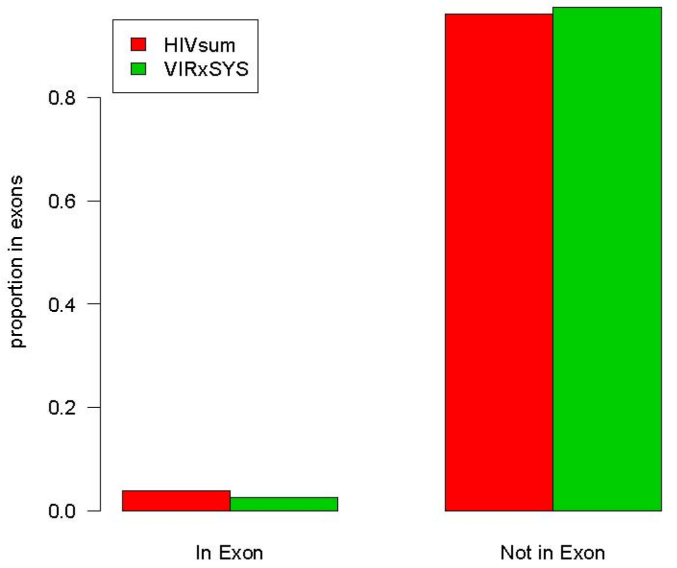
2.2 refGenes

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 0.443.

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



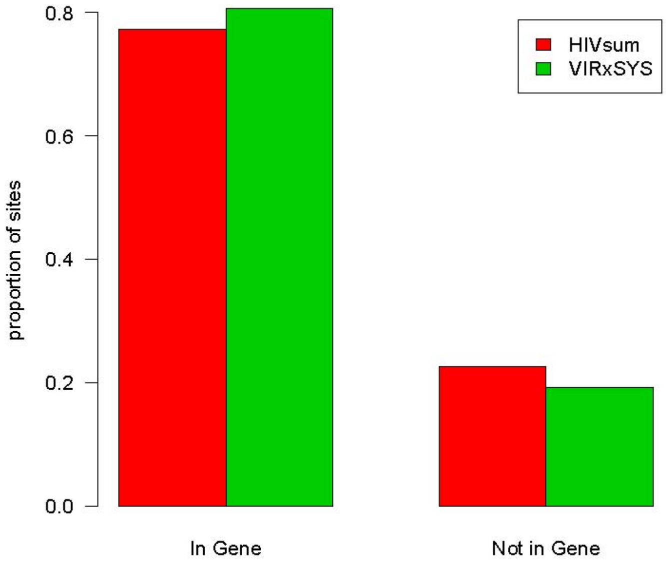
Here is the table of coefficients of the log ratio of intensities for along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	-2.830	0.146	-19.50	2.35e-84
in.gene	0.149	0.170	0.873	3.83e-01
in.exon	-0.452	0.465	-0.972	3.31e-01

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.

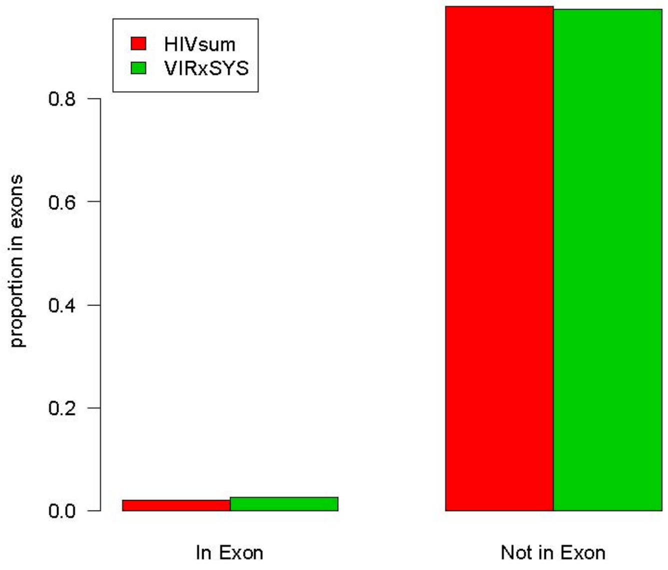
2.3 genScan Genes

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 0.257.

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



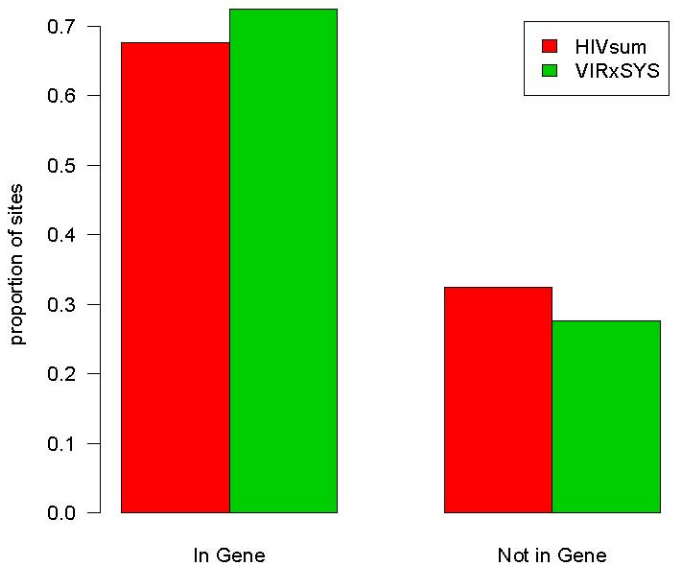
Here is the table of coefficients of the log ratio of intensities along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	-2.900	0.169	-17.200	2.72e-66
in.gene	0.204	0.189	1.080	2.80e-01
in.exon	0.182	0.472	0.386	7.00e-01

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.

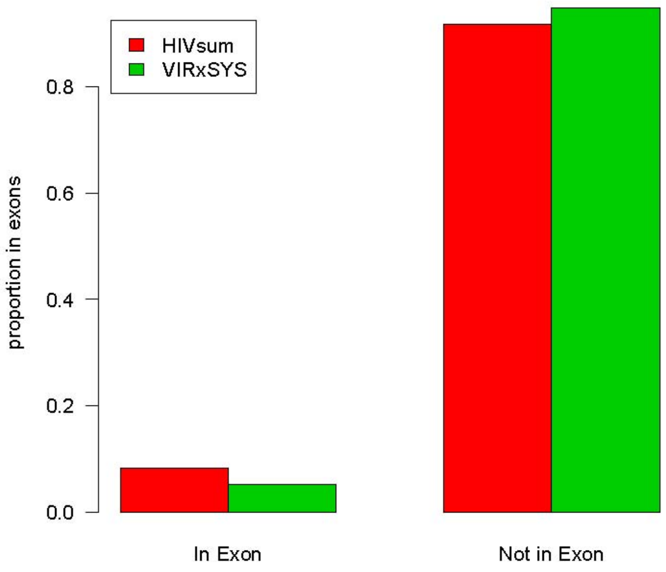
2.4 uniGenes

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.



Is there is a tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.162.

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 along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	-2.900	0.141	-20.50	8.46e-94
in.gene	0.285	0.168	1.70	8.99e-02
in.exon	-0.589	0.335	-1.76	7.89e-02

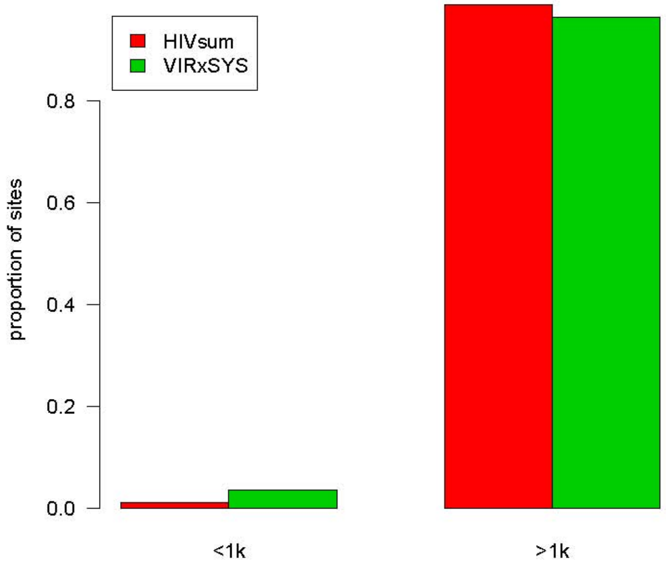
The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as 'in.exon' is net of that due to being in a gene. Note that in the barplot above the 'Not in Exon' bars include the both introns and intergenic regions, so the impression given 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 (11), who found that the neighborhoods within ±1kb of CpG islands are enriched for MLV insertions, we studied such neighborhoods for VRX496.

3.1 1 kilobase neighborhoods

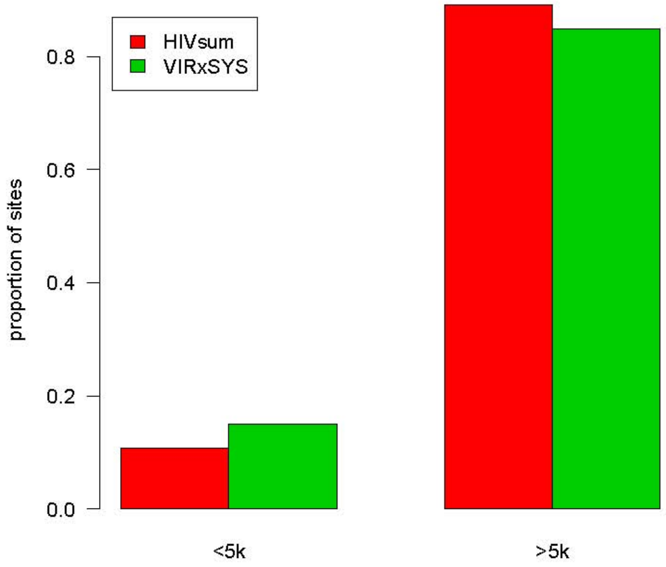
The following plot shows the effect of being in or within ± 1kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.0116.

3.2 5 kilobase neighborhoods

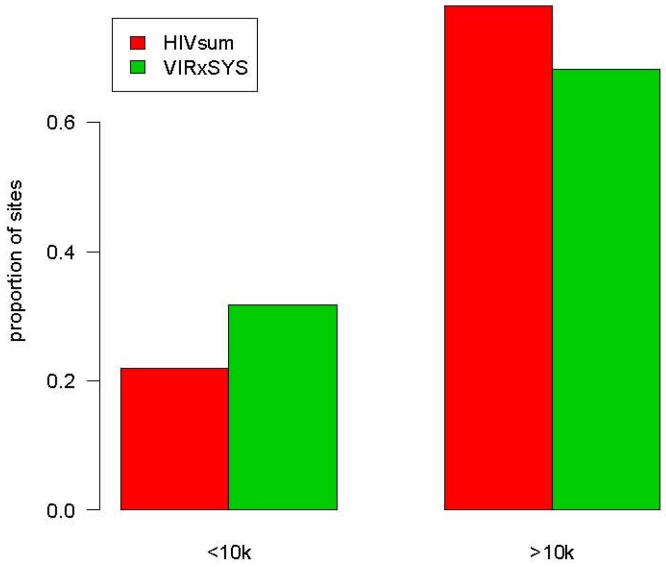
The following plot shows the effect of being in or within ± 5kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.0814.

3.3 10 kilobase neighborhoods

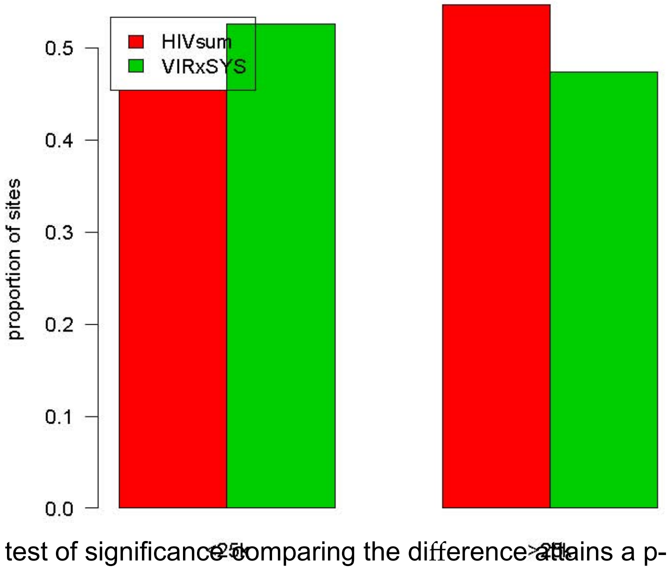
The following plot shows the effect of being in or within ± 10kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.00241.

3.4 25 kilobase neighborhoods

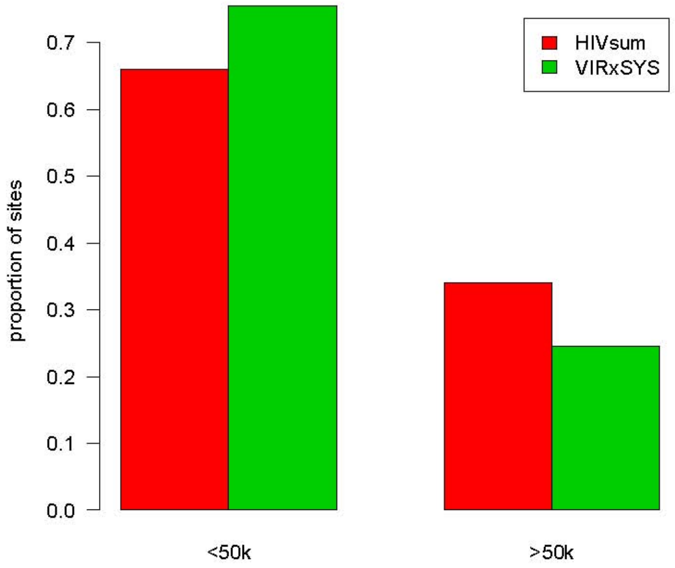
The following plot shows the effect of being in or within ± 25kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.051.

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 0.0055.

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

In this section the association with gene density was examined. The 'genes' that are counted are the genes represented on the Affymetrix Hu133A microarray. In addition, we categorized 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 1/8th and divide by number of bases
high.ex	Count genes whose expression is in the upper half 1/16th and divide by number of bases

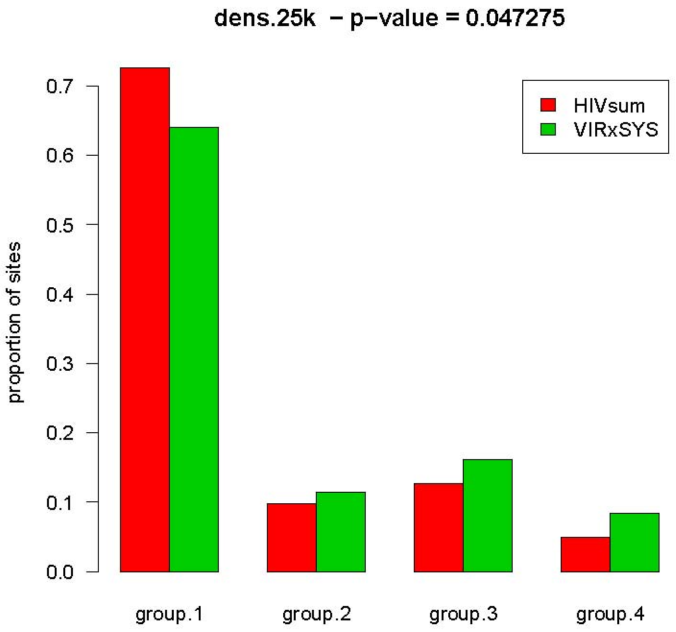
The bolded terms are used as abbreviations in what follows. The abbreviation 'dens' is used to indicate gene density as number of genes per base.

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.

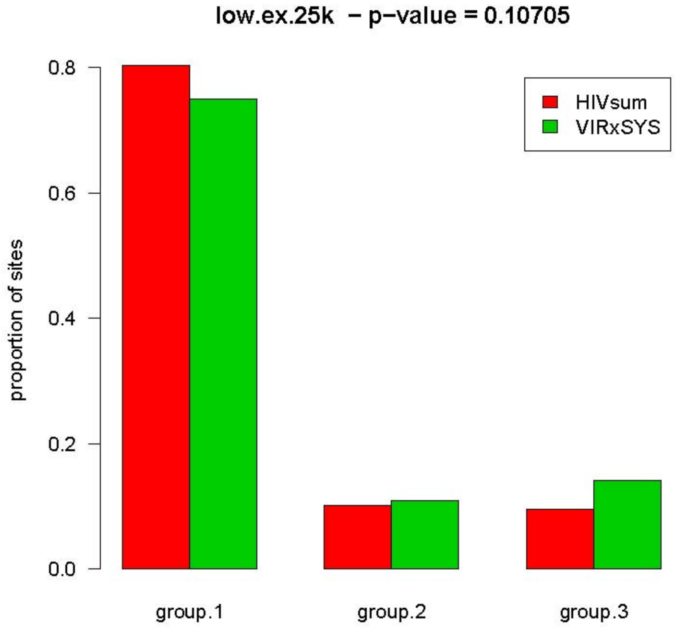
Also note that the title of the plot contains clues as to its content; the prefix indicates the type of variable studied while the suffix indicates the window width in the number of bases. The p-value given is the result of fitting a cubic polynomial to the gene density values.

Category	limits	
	lower category	upper
1	0e+00 group.1	0.00001
2	1e-05 group.2	0.00002
3	2e-05 group.3	0.00004
4	4e-05 group.4	0.00018



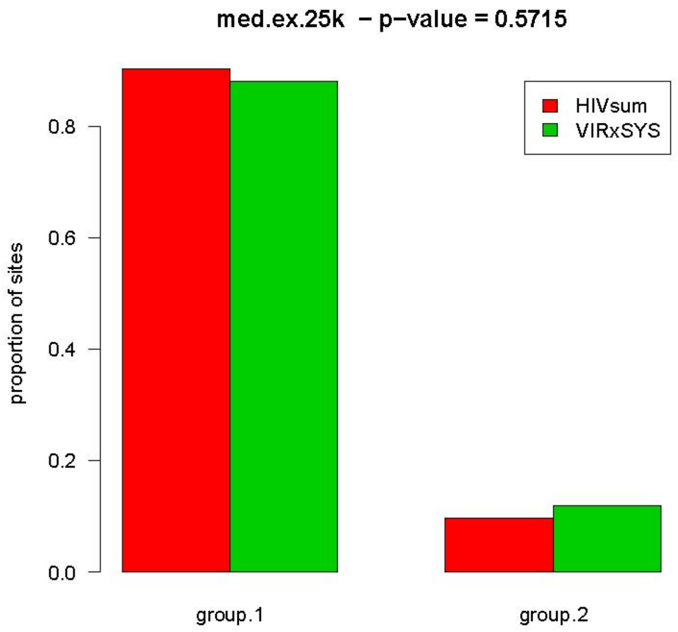
Here are the results for expression density. First, we count just genes that are in the upper half.

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	1.400000e-04



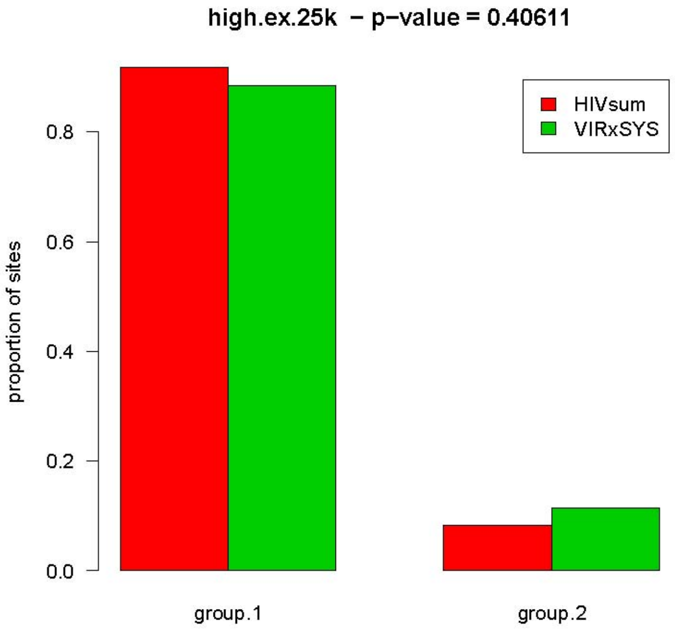
Now we count genes in the upper 1/8th:

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	1.333333e-05
2	1.333333e-05	group.2	1.200000e-04



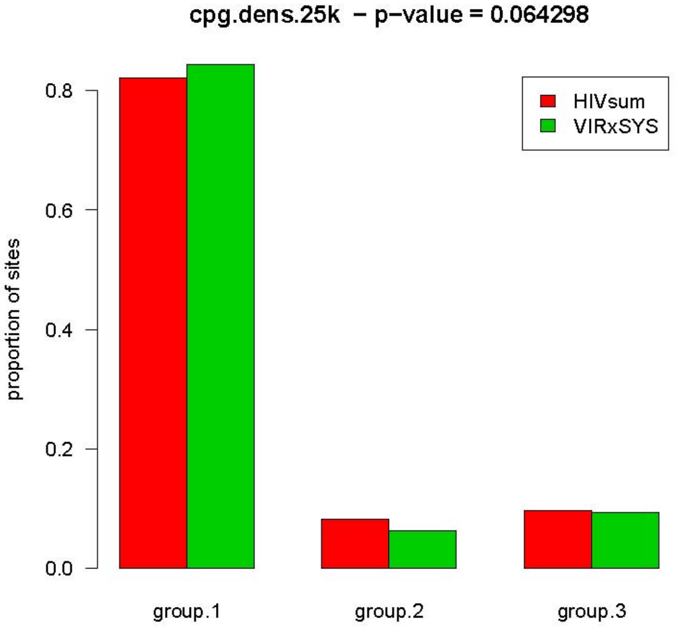
And here we count genes in the upper 1/16th:

Category	limits	
	lower category	upper
0%	0.000000e+00 group.	1 3.636364e-06
100%	3.636364e-06 group.2	1.200000e-04



Here the effect of density of CpG islands is studied:

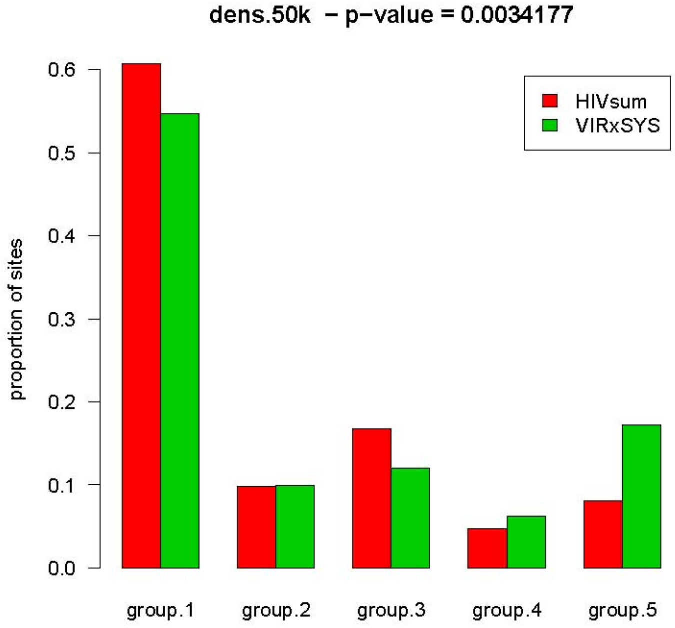
Category		limits	
	lower	category	upper
1	0e+00	group.1	0.00002
2	2e-05	group.2	0.00004
3	4e-05	group.3	0.00024



4.2 50 kiloBase Window

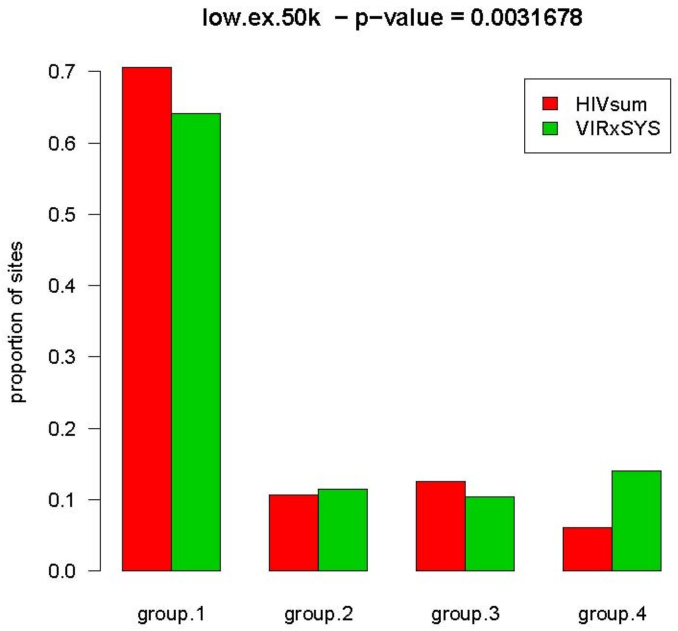
In the barplot that follows we examine the association of insertion sites with expression density in a 50 kilobase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category	limits		
	lower	category	upper
1	0.000000e+00	group.	1 6.666667e-06
2	6.666667e-06	group.2	1.000000e-05
3	1.000000e-05	group.3	2.000000e-05
4	2.000000e-05	group.4	3.000000e-05
5	3.000000e-05	group.5	1.400000e-04



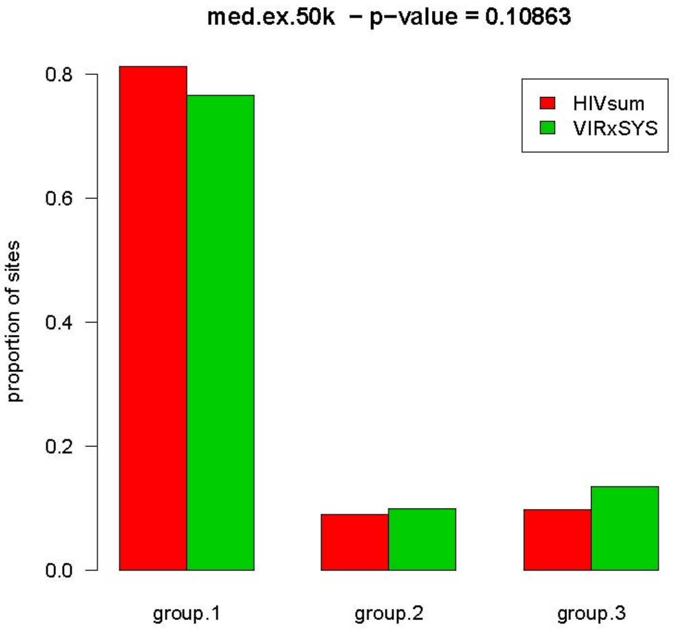
Below are the results for expression density. First, we count just genes that are in the upper half.

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	9.000000e-05



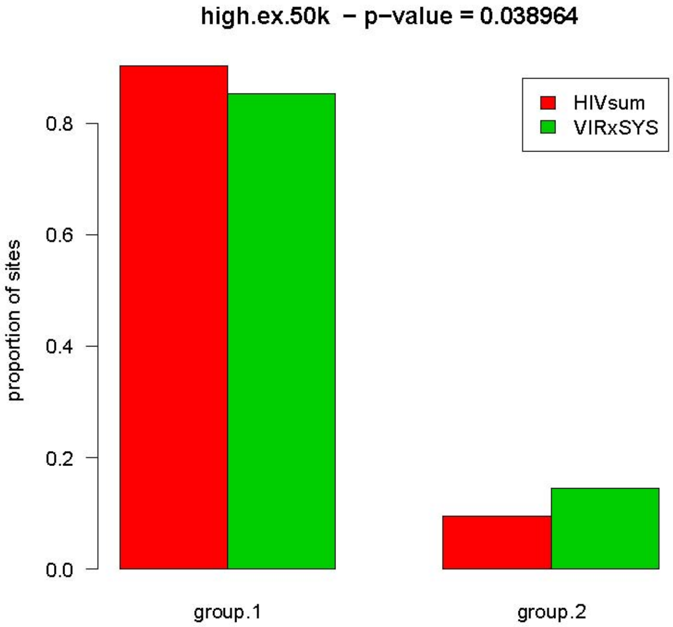
Now we count genes in the upper 1/8th:

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	5.000000e-06
2	5.000000e-06	group.2	1.166667e-05
3	1.166667e-05	group.3	8.000000e-05



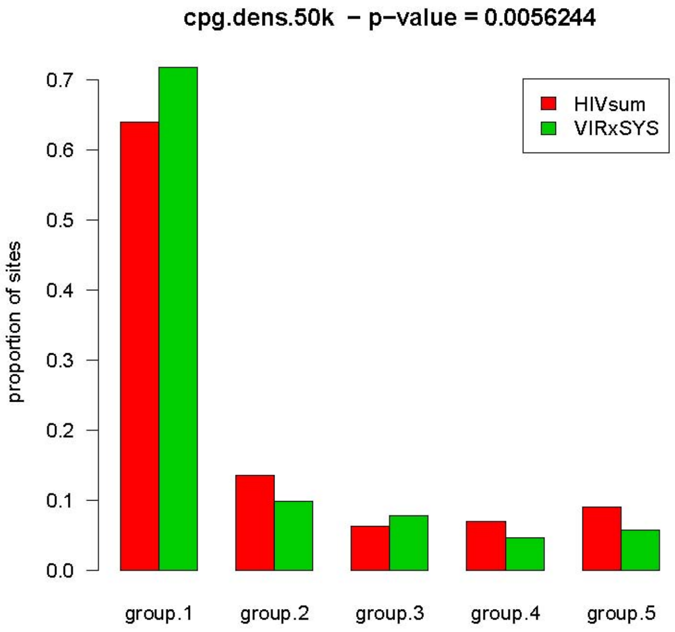
And here we count genes in the upper 1/16th:

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	6.666667e-06
2	6.666667e-06	group.2	8.000000e-05



Below the effect of density of CpG islands is studied:

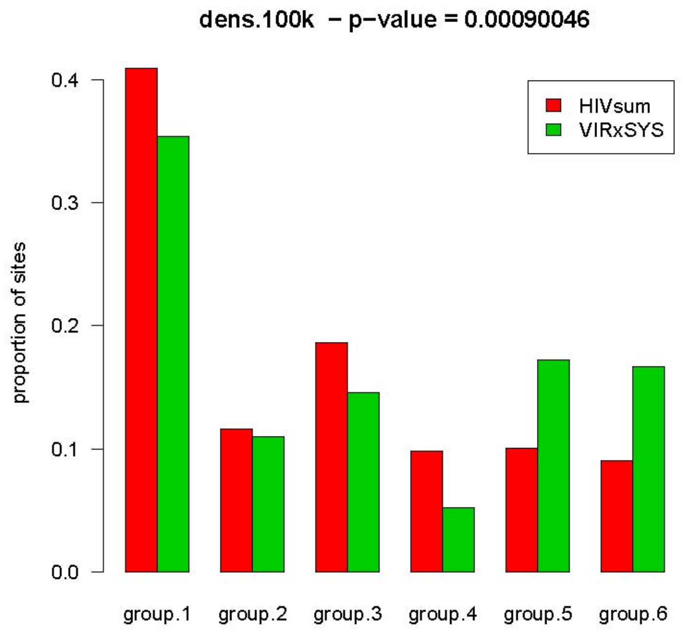
Category	limits	
	lower category	upper
1	0e+00 group.1	0.00001
2	1e-05 group.2	0.00002
3	2e-05 group.3	0.00003
4	3e-05 group.4	0.00005
5	5e-05 group.5	0.00025



4.3 100 kiloBase Window

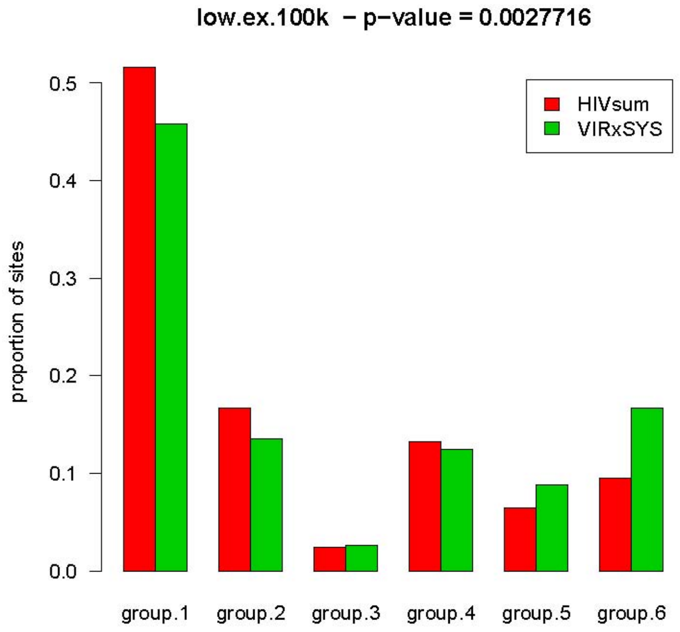
In the barplot that follows we examine the association of insertion sites with expression density in a 100 kilobase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

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	1.600000e-05
5	1.600000e-05 group.5	2.500000e-05
6	2.500000e-05 group.6	1.050000e-04



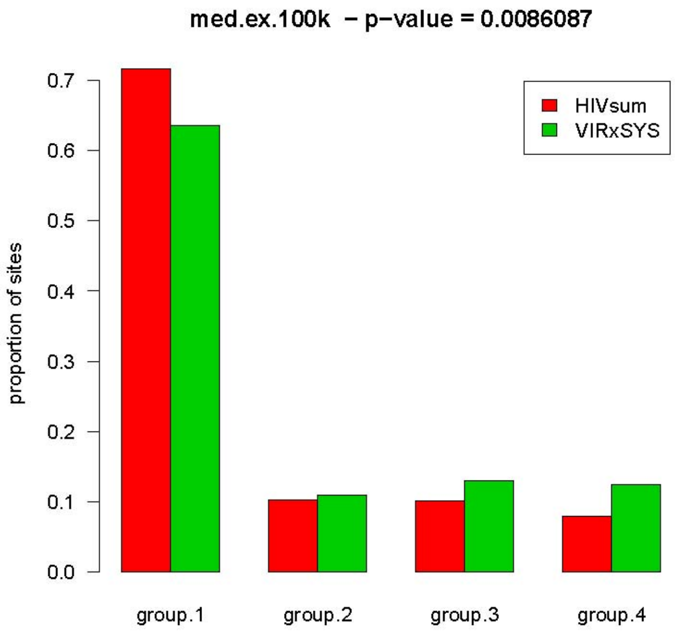
Here are the results for expression density. First, we count just genes that are in the upper half.

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	2.500000e-06
2	2.500000e-06	group.2	5.000000e-06
3	5.000000e-06	group.3	6.666667e-06
4	6.666667e-06	group.4	1.000000e-05
5	1.000000e-05	group.5	1.600000e-05
6	1.600000e-05	group.6	7.000000e-05



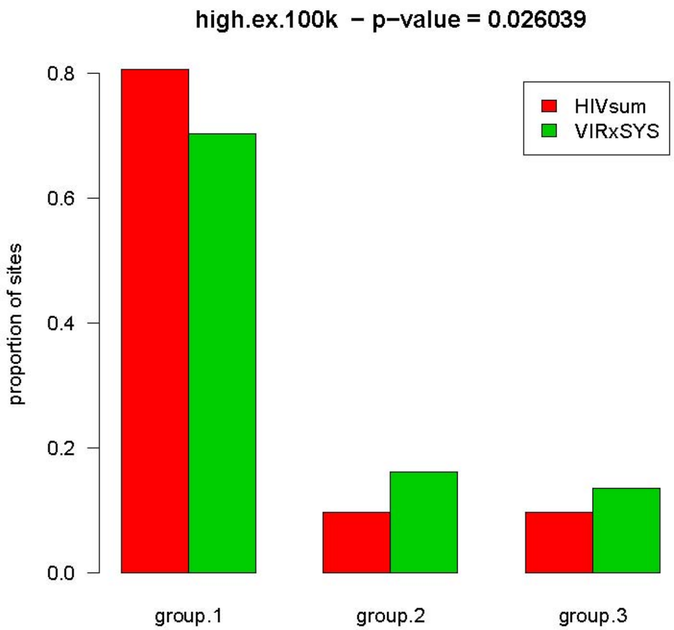
Now we count genes in the upper 1/8th:

```
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  5.500000e-05
```



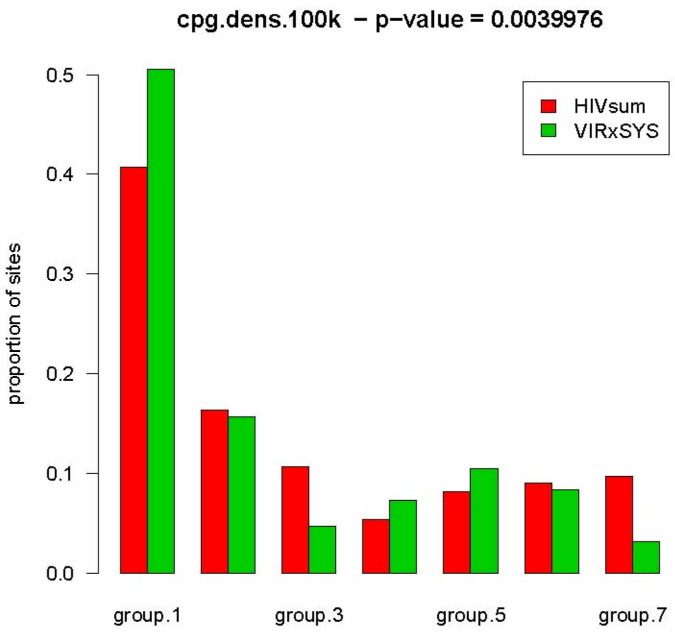
Below we count genes in the upper 1/16th:

```
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  4.000000e-05
```



Here the effect of density of CpG islands is studied:

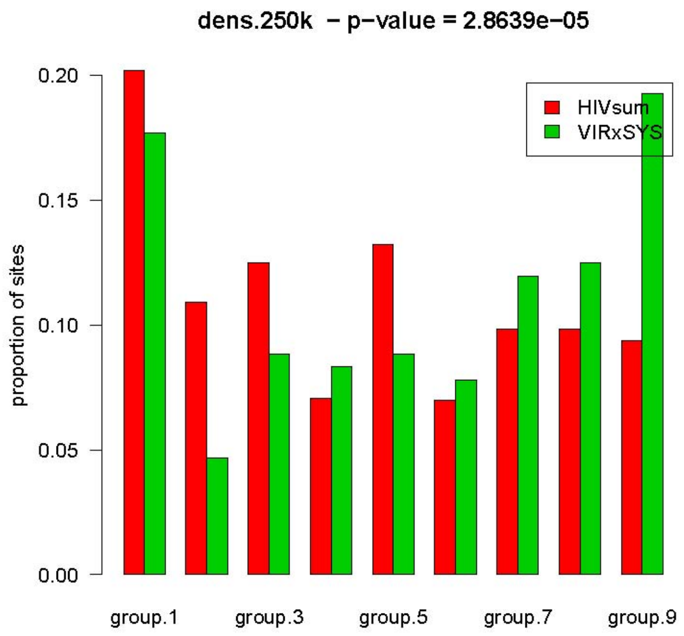
Category	limits		
	lower	category	upper
1	0.0e+00	group.1	0.000005
2	5.0e-06	group.2	0.000010
3	1.0e-05	group.3	0.000015
4	1.5e-05	group.4	0.000020
5	2.0e-05	group.5	0.000030
6	3.0e-05	group.6	0.000055
7	5.5e-05	group.7	0.000215



4.4 250 kiloBase Window

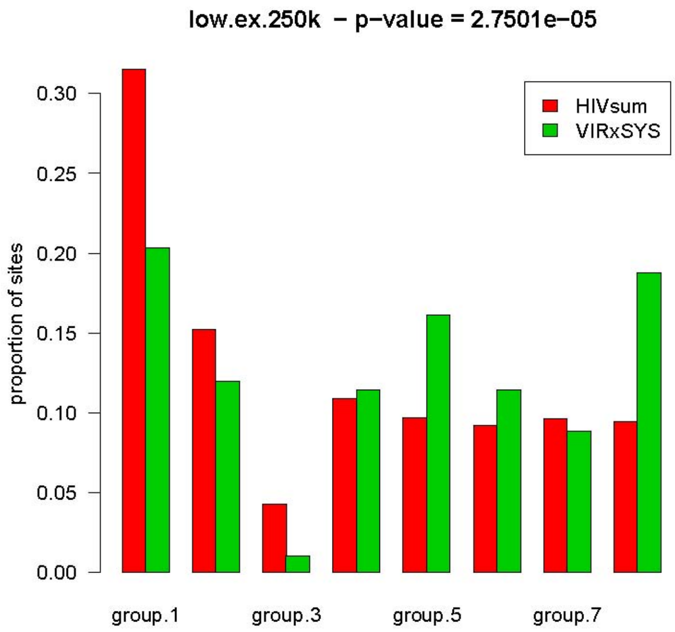
In the barplot that follows we examine the association of insertion sites with expression density in a 250 kilobase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category	limits	
	lower category	upper
1	0.000000e+00 group.1	1.142857e-06
2	1.142857e-06 group.2	2.666667e-06
3	2.666667e-06 group.3	4.000000e-06
4	4.000000e-06 group.4	5.674768e-06
5	5.674768e-06 group.5	8.000000e-06
6	8.000000e-06 group.6	1.033333e-05
7	1.033333e-05 group.7	1.520000e-05
8	1.520000e-05 group.8	2.466667e-05
9	2.466667e-05 group.9	9.546667e-05



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

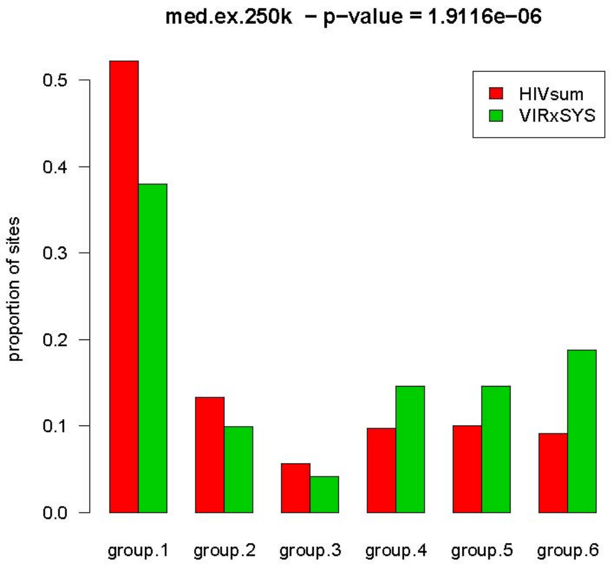
Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	8.000000e-07
2	8.000000e-07	group.2	2.000000e-06
3	2.000000e-06	group.3	2.666667e-06
4	2.666667e-06	group.4	4.000000e-06
5	4.000000e-06	group.5	6.000000e-06
6	6.000000e-06	group.6	9.000000e-06
7	9.000000e-06	group.7	1.440000e-05
8	1.440000e-05	group.8	6.000000e-05



low.ex.250k – p–value = 2.7501e–05

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

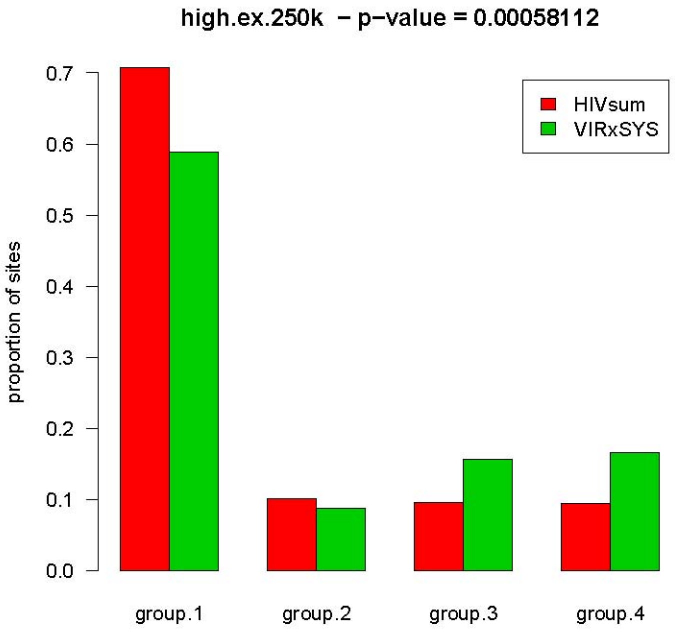
Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	1.000000e-06
2	1.000000e-06	group.2	2.000000e-06
3	2.000000e-06	group.3	3.333333e-06
4	3.333333e-06	group.4	4.800000e-06
5	4.800000e-06	group.5	8.000000e-06
6	8.000000e-06	group.6	3.033333e-05



med.ex.250k – p–value = 1.9116e–06

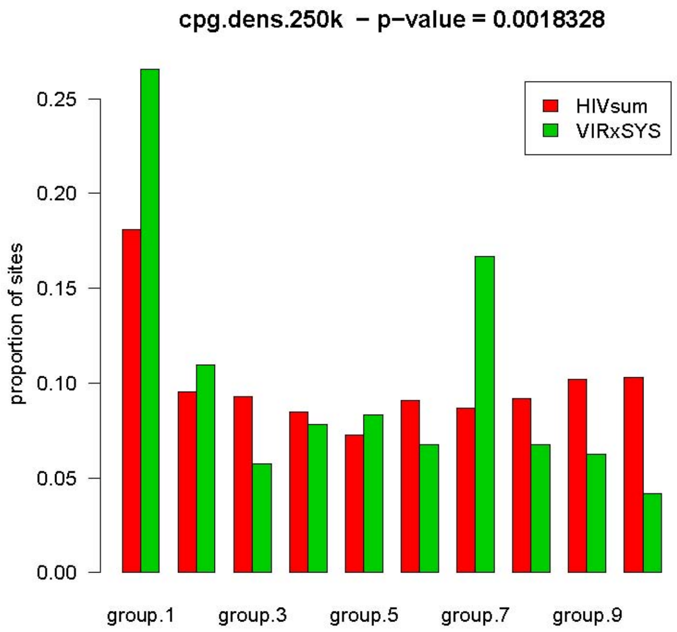
And here we count genes in the upper 1/16th:

Category	limits	
	lower category	upper
1	0.000000e+00 group.1	1.333333e-06
2	1.333333e-06 group.2	2.666667e-06
3	2.666667e-06 group.3	4.000000e-06
4	4.000000e-06 group.4	2.400000e-05



Here the effect of density of CpG islands is studied:

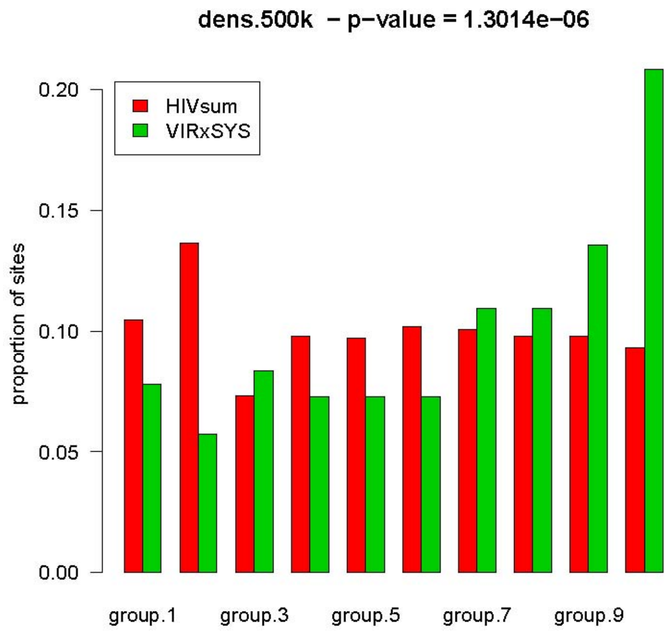
Category	limits	
	lower category	upper
1	0.0e+00 group.1	0.000002
2	2.0e-06 group.2	0.000004
3	4.0e-06 group.3	0.000006
4	6.0e-06 group.4	0.000008
5	8.0e-06 group.5	0.000010
6	1.0e-05 group.6	0.000014
7	1.4e-05 group.7	0.000020
8	2.0e-05 group.8	0.000030
9	3.0e-05 group.9	0.000054
10	5.4e-05 group.10	0.000202



4.5 500 kiloBase Window

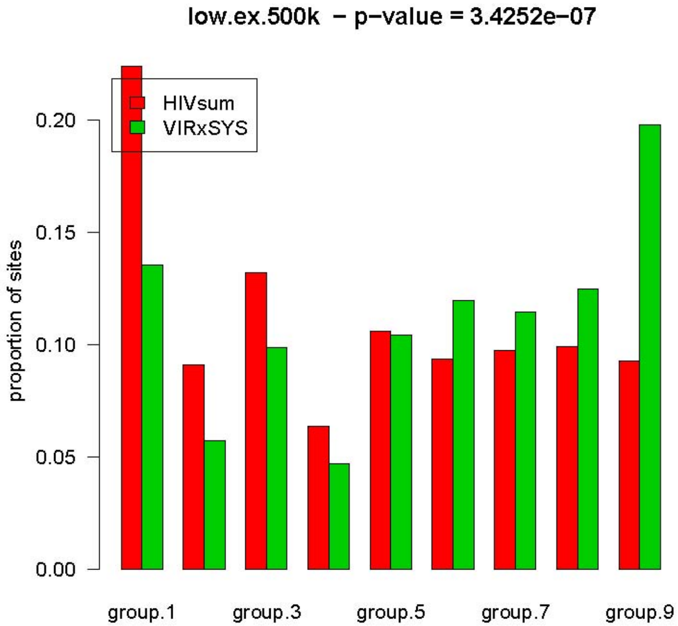
In the barplot that follows we examine the association of insertion sites with expression density in a 500 kilobase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	5.000000e-07
2	5.000000e-07	group.2	2.000000e-06
3	2.000000e-06	group.3	3.000000e-06
4	3.000000e-06	group.4	4.000000e-06
5	4.000000e-06	group.5	5.400000e-06
6	5.400000e-06	group.6	7.333333e-06
7	7.333333e-06	group.7	1.000000e-05
8	1.000000e-05	group.8	1.500000e-05
9	1.500000e-05	group.9	2.300000e-05
10	2.300000e-05	group.10	7.633333e-05



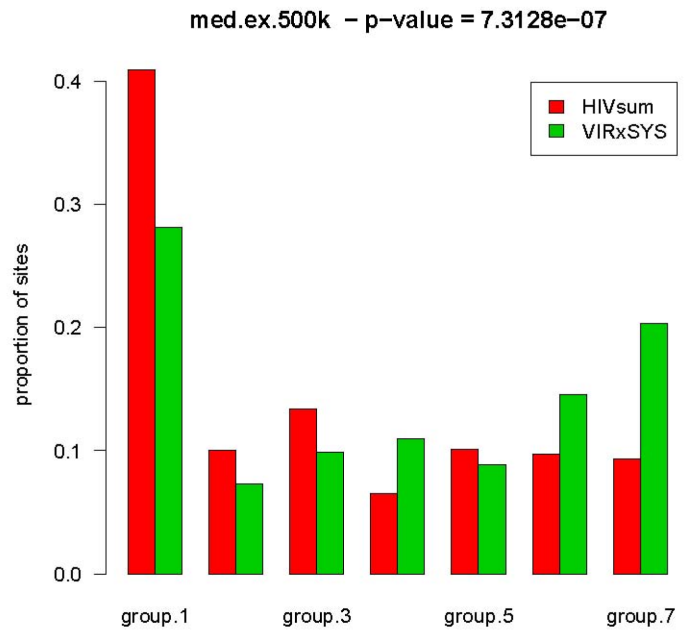
Below are the results for expression density. First, we count just genes that are in the upper half.

Category		limits	
	lower	category	upper
1	0.000000e+00	group.1	5.000000e-07
2	5.000000e-07	group.2	1.000000e-06
3	1.000000e-06	group.3	2.000000e-06
4	2.000000e-06	group.4	2.800000e-06
5	2.800000e-06	group.5	4.000000e-06
6	4.000000e-06	group.6	5.666667e-06
7	5.666667e-06	group.7	8.400000e-06
8	8.400000e-06	group.8	1.300000e-05
9	1.300000e-05	group.9	5.366667e-05



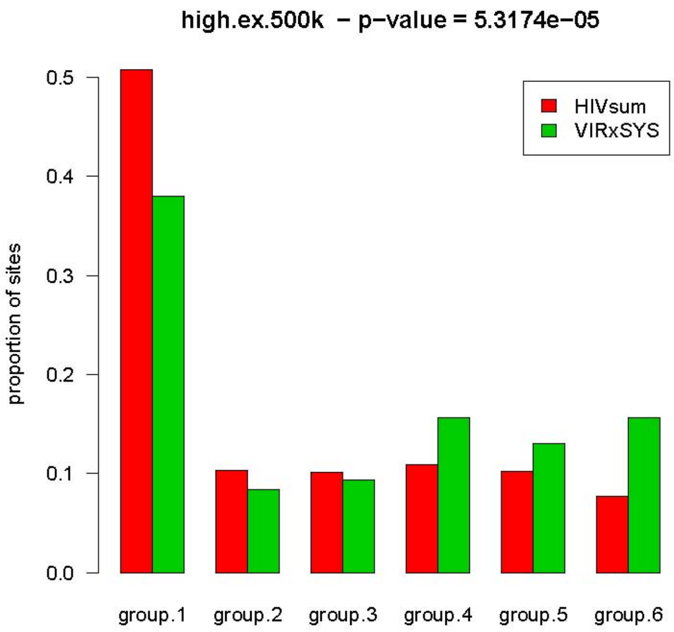
Now we count genes in the upper 1/8th:

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	6.666667e-07
2	6.666667e-07	group.2	1.222222e-06
3	1.222222e-06	group.3	2.000000e-06
4	2.000000e-06	group.4	3.000000e-06
5	3.000000e-06	group.5	4.285714e-06
6	4.285714e-06	group.6	7.166667e-06
7	7.166667e-06	group.7	2.383333e-05



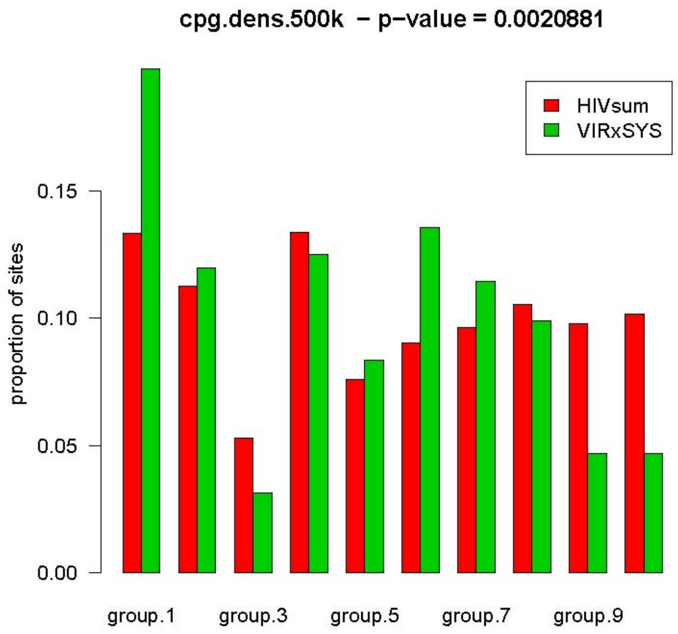
Here we count genes in the upper 1/16th:

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	3.333333e-07
2	3.333333e-07	group.2	8.000000e-07
3	8.000000e-07	group.3	1.333333e-06
4	1.333333e-06	group.4	2.000000e-06
5	2.000000e-06	group.5	4.000000e-06
6	4.000000e-06	group.6	1.500000e-05



Here the eﬀect of density of CpG islands is studied:

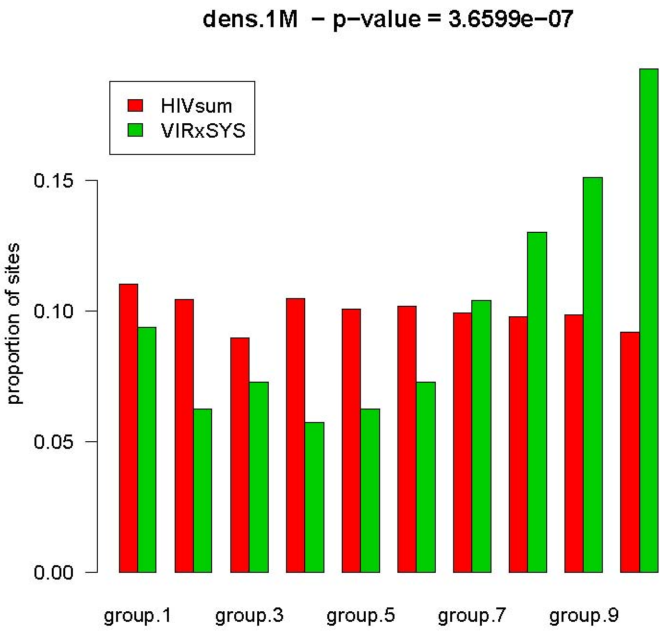
Category	limits		
	lower	category	upper
1	0.0e+00	group.1	0.000002
2	2.0e-06	group.2	0.000004
3	4.0e-06	group.3	0.000005
4	5.0e-06	group.4	0.000008
5	8.0e-06	group.5	0.000010
6	1.0e-05	group.6	0.000014
7	1.4e-05	group.7	0.000020
8	2.0e-05	group.8	0.000031
9	3.1e-05	group.9	0.000052
10	5.2e-05	group.10	0.000192



4.6 1 megaBase Window

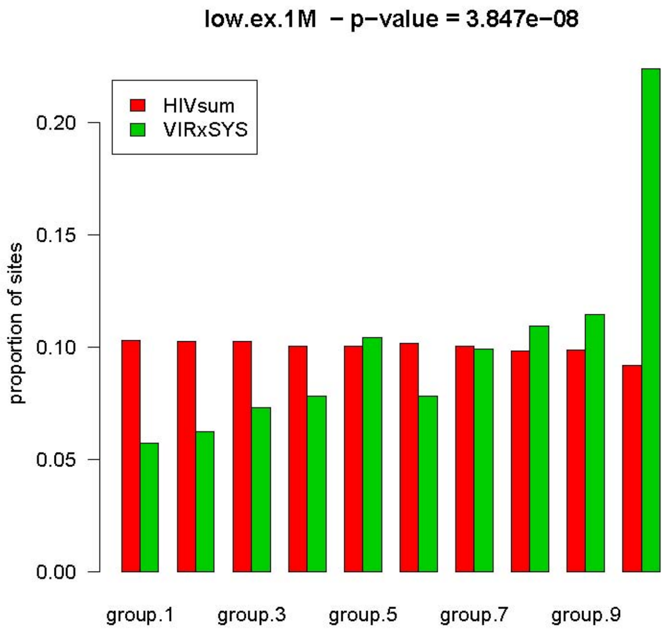
In the barplot that follows we examine the association of insertion sites with expression density in a 1 megabase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	1.000000e-06
2	1.000000e-06	group.2	2.000000e-06
3	2.000000e-06	group.3	3.000000e-06
4	3.000000e-06	group.4	4.000000e-06
5	4.000000e-06	group.5	5.319048e-06
6	5.319048e-06	group.6	6.916667e-06
7	6.916667e-06	group.7	9.750000e-06
8	9.750000e-06	group.8	1.425000e-05
9	1.425000e-05	group.9	2.186667e-05
10	2.186667e-05	group.10	5.866667e-05



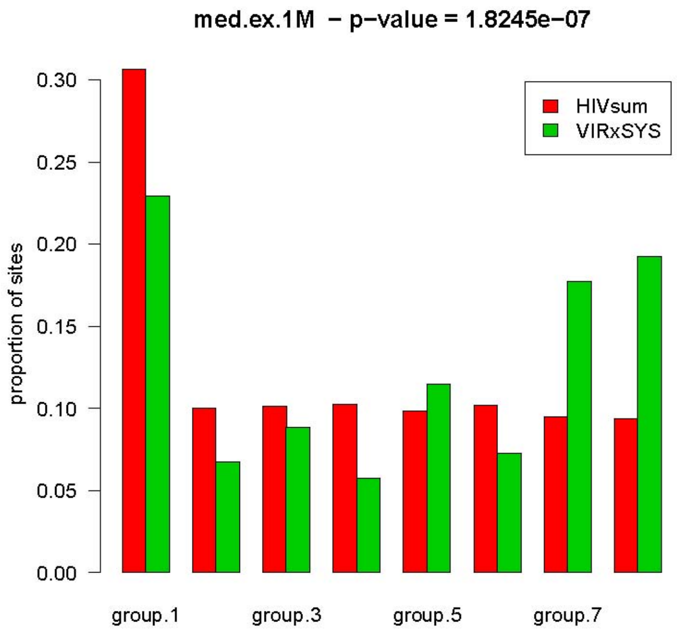
Here are the results for expression density. First, we count just genes that are in the upper half.

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	1.428571e-07
2	1.428571e-07	group.2	6.666667e-07
3	6.666667e-07	group.3	1.166667e-06
4	1.166667e-06	group.4	1.750000e-06
5	1.750000e-06	group.5	2.500000e-06
6	2.500000e-06	group.6	3.583333e-06
7	3.583333e-06	group.7	5.000000e-06
8	5.000000e-06	group.8	7.642857e-06
9	7.642857e-06	group.9	1.217691e-05
10	1.217691e-05	group.10	4.058333e-05



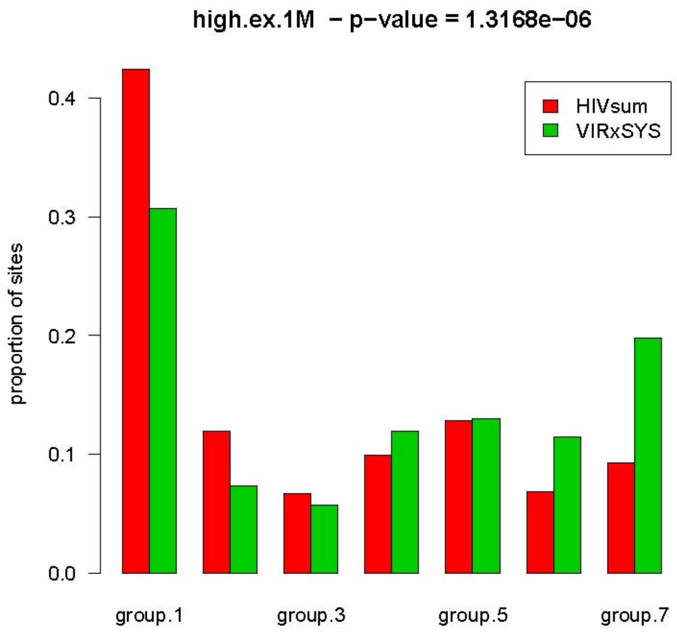
Now we count genes in the upper 1/8th:

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	4.000000e-07
2	4.000000e-07	group.2	7.857143e-07
3	7.857143e-07	group.3	1.150000e-06
4	1.150000e-06	group.4	1.700000e-06
5	1.700000e-06	group.5	2.450000e-06
6	2.450000e-06	group.6	3.750000e-06
7	3.750000e-06	group.7	6.500000e-06
8	6.500000e-06	group.8	2.041667e-05



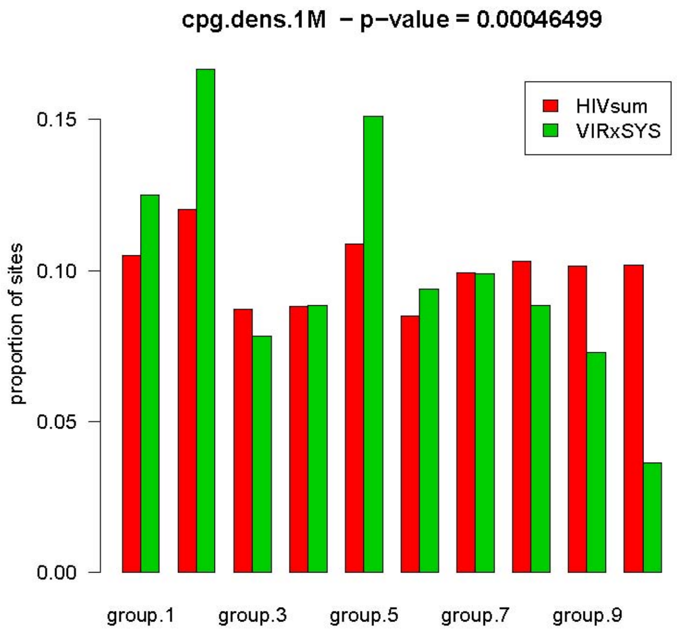
And here we count genes in the upper 1/16th:

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	2.500000e-07
2	2.500000e-07	group.2	5.000000e-07
3	5.000000e-07	group.3	8.666667e-07
4	8.666667e-07	group.4	1.000000e-06
5	1.000000e-06	group.5	2.000000e-06
6	2.000000e-06	group.6	3.000000e-06
7	3.000000e-06	group.7	1.239996e-05



Below the effect of density of CpG islands is studied:

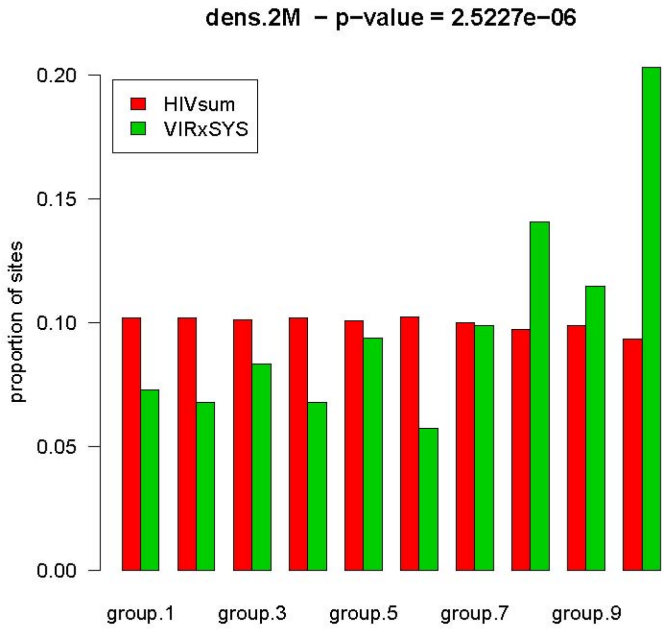
Category	limits		
	lower	category	upper
1	0.00e+00	group.1	2.00e-06
2	2.00e-06	group.2	4.00e-06
3	4.00e-06	group.3	5.50e-06
4	5.50e-06	group.4	7.00e-06
5	7.00e-06	group.5	1.00e-05
6	1.00e-05	group.6	1.35e-05
7	1.35e-05	group.7	1.95e-05
8	1.95e-05	group.8	2.85e-05
9	2.85e-05	group.9	5.00e-05
10	5.00e-05	group.10	1.72e-04



4.7 2 megaBase Window

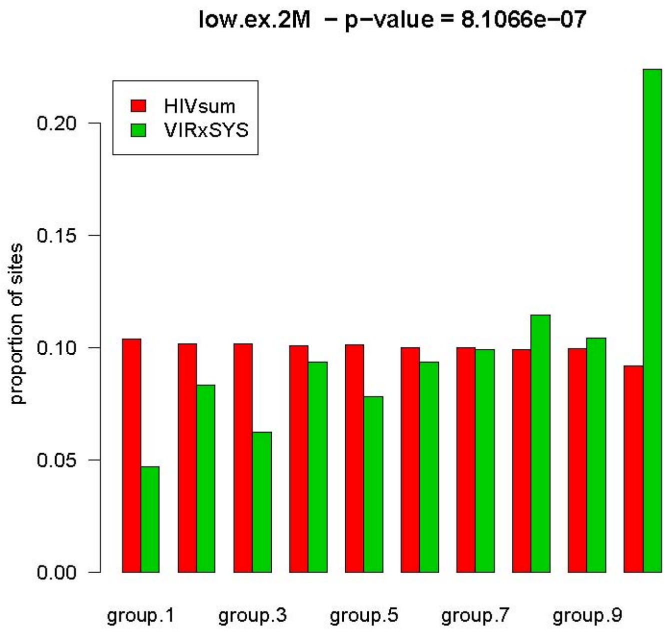
In the barplot that follows we examine the association of insertion sites with expression density in a 2 megabase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	1.283333e-06
2	1.283333e-06	group.2	2.083333e-06
3	2.083333e-06	group.3	2.833333e-06
4	2.833333e-06	group.4	3.858333e-06
5	3.858333e-06	group.5	4.966667e-06
6	4.966667e-06	group.6	6.595455e-06
7	6.595455e-06	group.7	9.441667e-06
8	9.441667e-06	group.8	1.375833e-05
9	1.375833e-05	group.9	1.926667e-05
10	1.926667e-05	group.10	4.108333e-05



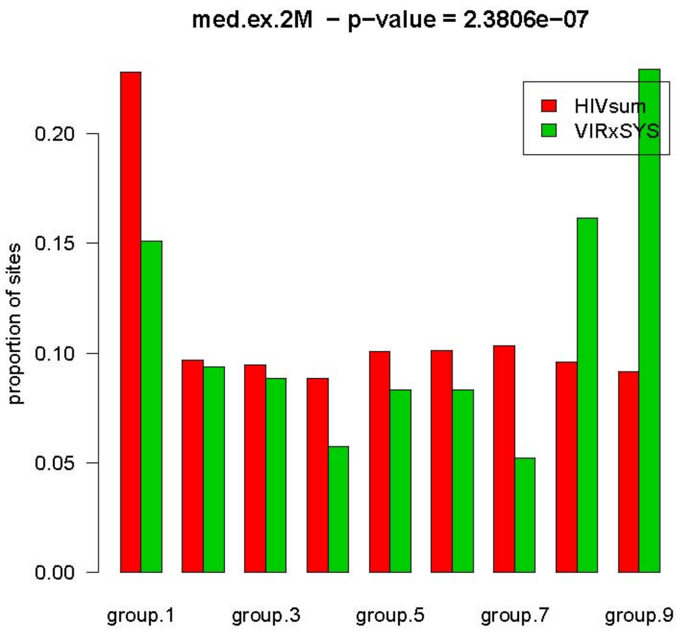
Below are the results for expression density. First, we count just genes that are in the upper half.

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	4.047619e-07
2	4.047619e-07	group.2	7.666667e-07
3	7.666667e-07	group.3	1.166667e-06
4	1.166667e-06	group.4	1.666667e-06
5	1.666667e-06	group.5	2.366667e-06
6	2.366667e-06	group.6	3.308333e-06
7	3.308333e-06	group.7	5.083333e-06
8	5.083333e-06	group.8	7.258333e-06
9	7.258333e-06	group.9	1.046667e-05
10	1.046667e-05	group.10	2.487500e-05



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

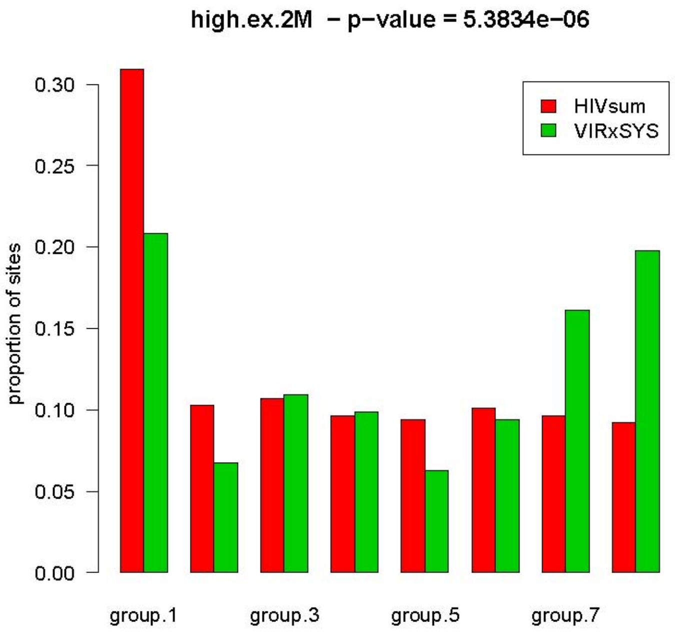
Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	2.500000e-07
2	2.500000e-07	group.2	5.000000e-07
3	5.000000e-07	group.3	7.500000e-07
4	7.500000e-07	group.4	1.125000e-06
5	1.125000e-06	group.5	1.616667e-06
6	1.616667e-06	group.6	2.408333e-06
7	2.408333e-06	group.7	3.571429e-06
8	3.571429e-06	group.8	5.489139e-06
9	5.489139e-06	group.9	1.397500e-05



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

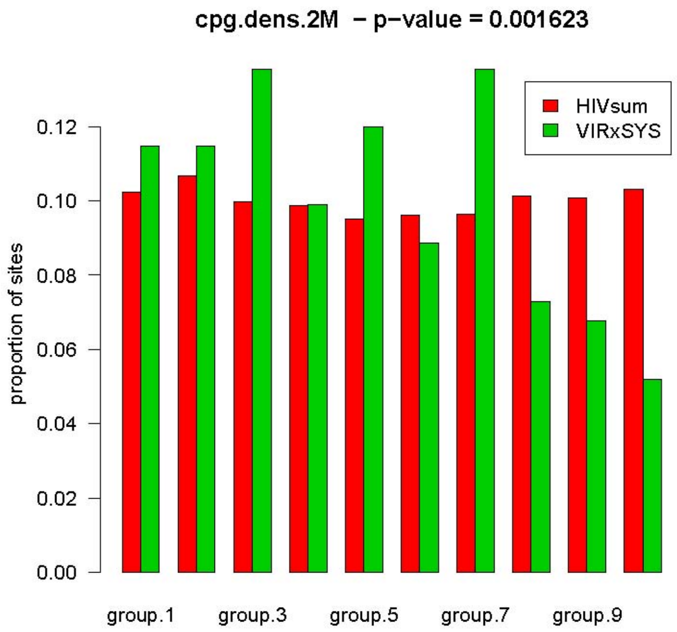
Category limits

lower category		upper
1	0.000000e+00 group.1	1.666667e-07
2	1.666667e-07 group.2	3.333333e-07
3	3.333333e-07 group.3	5.000000e-07 4
5.	000000e-07 group.4	7.500000e-07
5	7.500000e-07 group.5	1.053571e-06
6	1.053571e-06 group.6	1.666667e-06
7	1.666667e-06 group.7	2.708333e-06
8	2.708333e-06 group.8	8.959995e-06



Here the effect of density of CpG islands is studied:

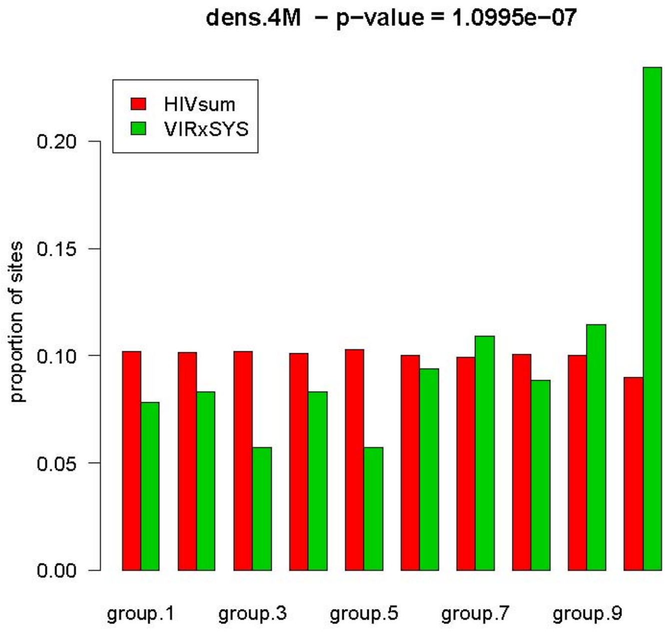
Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	2.500000e-06
2	2.500000e-06	group.2	4.000000e-06
3	4.000000e-06	group.3	5.250000e-06
4	5.250000e-06	group.4	7.000000e-06
5	7.000000e-06	group.5	9.250000e-06
6	9.250000e-06	group.6	1.325000e-05
7	1.325000e-05	group.7	1.875000e-05
8	1.875000e-05	group.8	2.625000e-05
9	2.625000e-05	group.9	4.701325e-05
10	4.701325e-05	group.10	1.515144e-04



4.8 4 megaBase Window

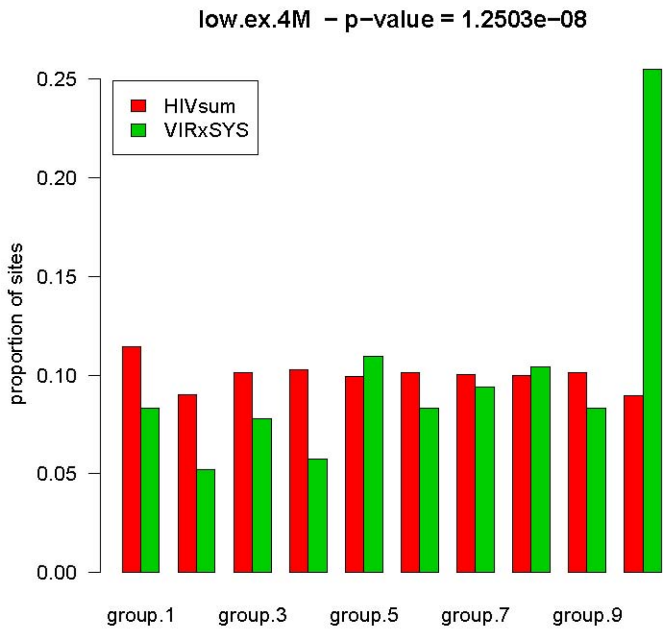
In the barplot that follows we examine the association of insertion sites with expression density in a 4 megabase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	1.458333e-06
2	1.458333e-06	group.2	2.091667e-06
3	2.091667e-06	group.3	2.731250e-06
4	2.731250e-06	group.4	3.463095e-06
5	3.463095e-06	group.5	4.487500e-06
6	4.487500e-06	group.6	6.358333e-06
7	6.358333e-06	group.7	8.779167e-06
8	8.779167e-06	group.8	1.233333e-05
9	1.233333e-05	group.9	1.643958e-05
10	1.643958e-05	group.10	3.604167e-05



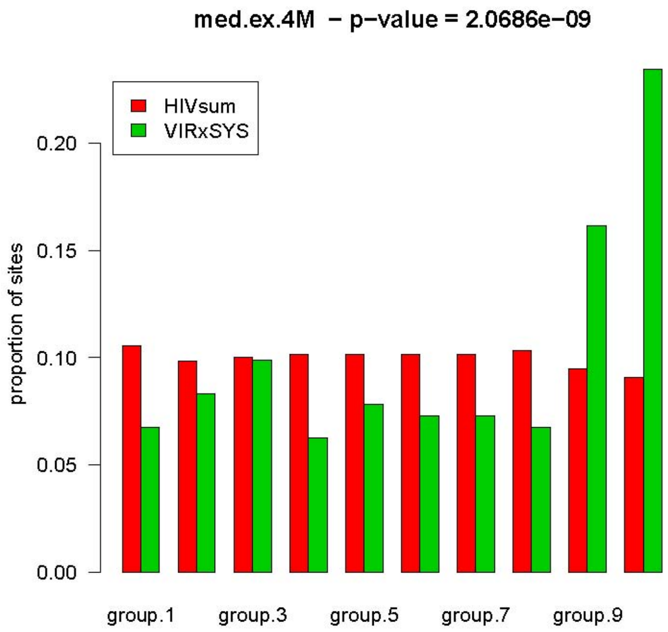
Below are the results for expression density. First, we count just genes that are in the upper half.

Category		limits	
	lower	category	upper
1	0.000000e+00	group.1	5.000000e-07
2	5.000000e-07	group.2	8.000000e-07
3	8.000000e-07	group.3	1.133333e-06
4	1.133333e-06	group.4	1.508333e-06
5	1.508333e-06	group.5	2.041667e-06
6	2.041667e-06	group.6	3.147917e-06
7	3.147917e-06	group.7	4.815181e-06
8	4.815181e-06	group.8	6.350595e-06
9	6.350595e-06	group.9	8.741667e-06
10	8.741667e-06	group.10	2.255000e-05



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

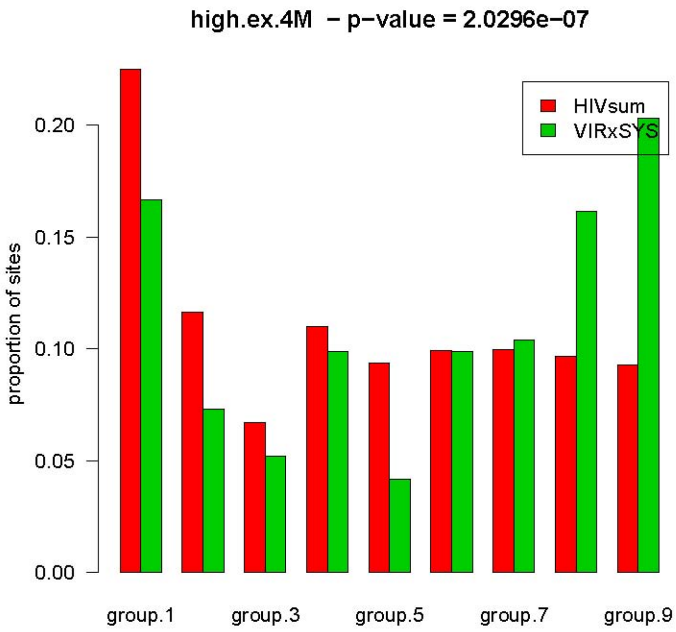
Category	limits	
	lower category	upper
1	0.000000e+00 group.1	1.250000e-07
2	1.250000e-07 group.2	3.238095e-07
3	3.238095e-07 group.3	5.000000e-07
4	5.000000e-07 group.4	7.000000e-07
5	7.000000e-07 group.5	1.020833e-06
6	1.020833e-06 group.6	1.566667e-06
7	1.566667e-06 group.7	2.343148e-06
8	2.343148e-06 group.8	3.187500e-06
9	3.187500e-06 group.9	4.350000e-06
10	4.350000e-06 group.10	1.333750e-05



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

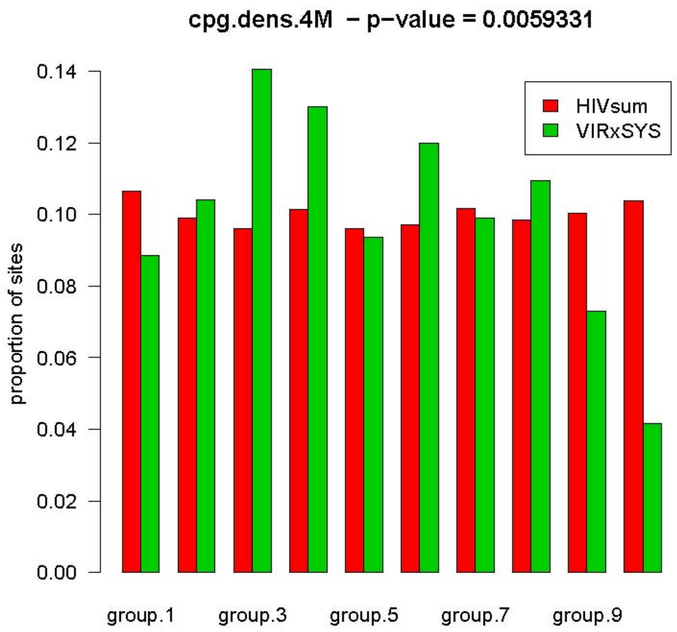
Category limits

	lower	category	upper
1	0.000000e+00	group.1	1.250000e-07
2	1.250000e-07	group.2	2.500000e-07
3	2.500000e-07	group.3	3.333333e-07
4	3.333333e-07	group.4	5.000000e-07
5	5.000000e-07	group.5	7.291667e-07
6	7.291667e-07	group.6	1.000000e-06
7	1.000000e-06	group.7	1.495833e-06
8	1.495833e-06	group.8	2.154167e-06
9	2.154167e-06	group.9	8.258333e-06



Here the effect of density of CpG islands is studied:

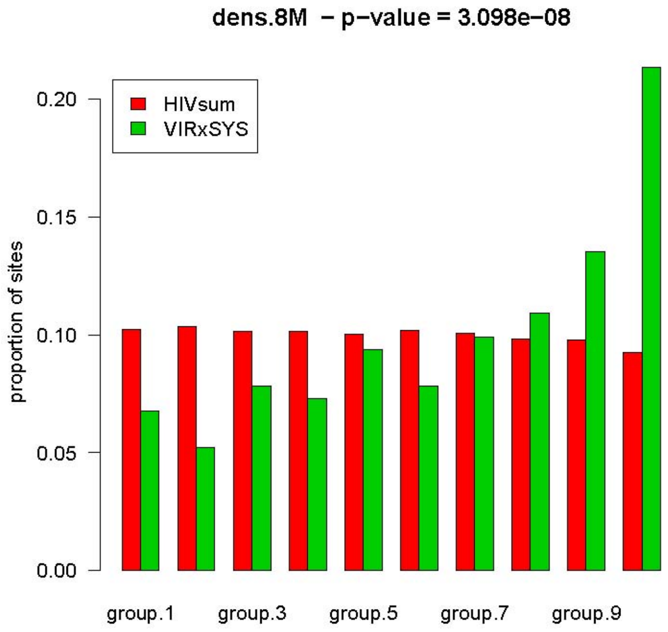
Category	limits		
	lower	category	upper
1	5.000000e-07	group.1	3.000000e-06
2	3.000000e-06	group.2	4.000000e-06
3	4.000000e-06	group.3	5.250000e-06
4	5.250000e-06	group.4	6.875000e-06
5	6.875000e-06	group.5	9.250000e-06
6	9.250000e-06	group.6	1.225000e-05
7	1.225000e-05	group.7	1.687500e-05
8	1.687500e-05	group.8	2.387500e-05
9	2.387500e-05	group.9 4	.246248e-05
10	4.246248e-05	group.10	1.268003e-04



4.9 8 megaBase Window

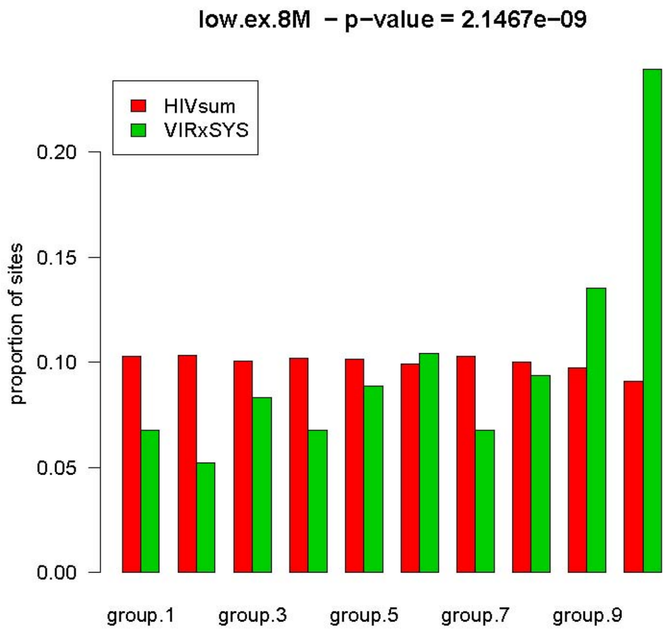
In the barplot that follows we examine the association of insertion sites with expression density in an 8 megabase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category	limits	
	lower category	upper
1	4.166667e-8 group 1	1.5875e-06
2	1.5875e-06 group.2	2.150000e-06 2
3	2.150000e-06 group.3	2.658631e-06
4	2.658631e-06 group.4	3.381250e-06
5	3.381250e-06 group.5	4.397917e-06
6	4.397917e-06 group.6	6.000962e-06
7	6.000962e-06 group.7	7.764583e-06
8	7.764583e-06 group.8	1.051634e-05
9	1.051634e-05 group.9	1.530215e-05
10	1.530215e-05 group.10	2.956875e-05



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

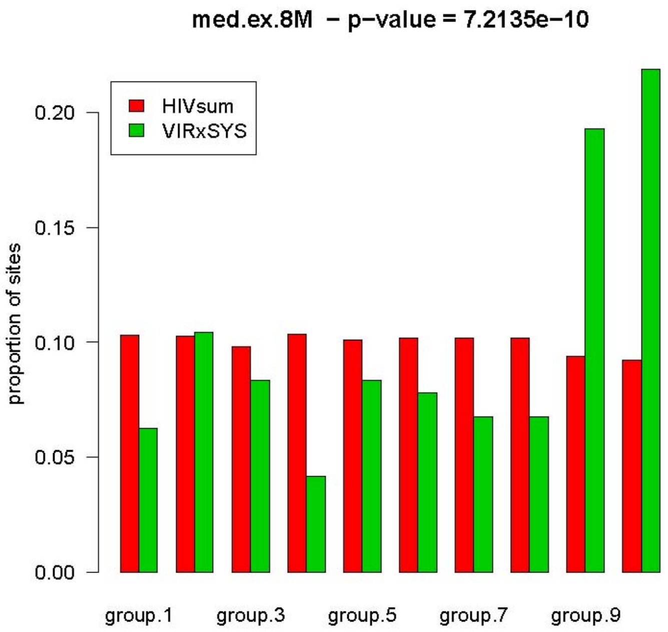
Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	5.708333e-07
2	5.708333e-07	group.2	8.385417e-07
3	8.385417e-07	group.3	1.140972e-06
4	1.140972e-06	group.4	1.433333e-06
5	1.433333e-06	group.5	2.064205e-06
6	2.064205e-06	group.6	2.976190e-06
7	2.976190e-06	group.7	4.018899e-06
8	4.018899e-06	group.8	5.522917e-06
9	5.522917e-06	group.9	7.635913e-06
10	7.635913e-06	group.10	1.508542e-05



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

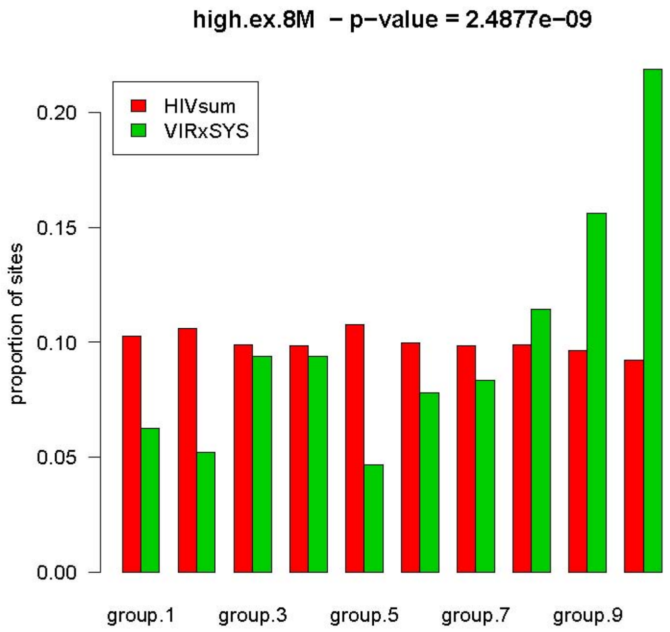
Category limits

	lower	category	upper
1	0.000000e+00	group.1	2.291667e-07
2	2.291667e-07	group.2	3.645833e-07
3	3.645833e-07	group.3	5.000000e-07
4	5.000000e-07	group.4	7.083333e-07
5	7.083333e-07	group.5	9.812500e-07
6	9.812500e-07	group.6	1.435417e-06
7	1.435417e-06	group.7	1.947917e-06
8	1.947917e-06	group.8	2.870833e-06
9	2.870833e-06	group.9	3.580745e-06
10	3.580745e-06	group.10	8.177083e-06



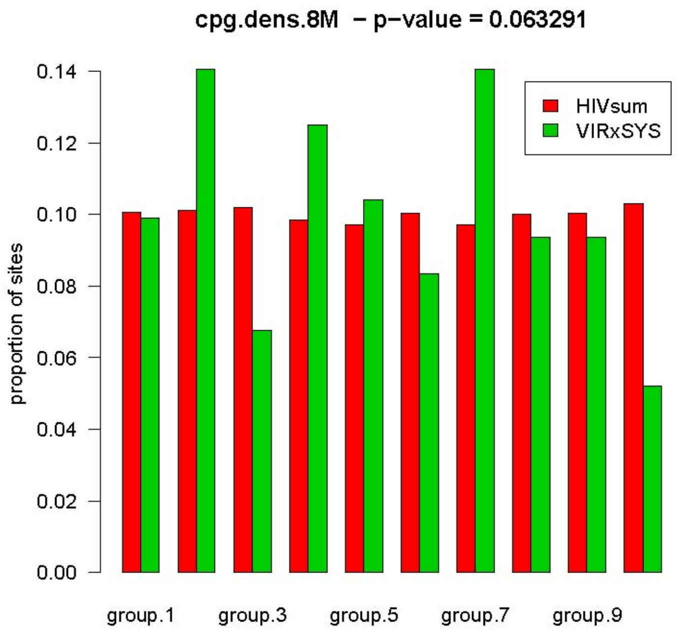
And here we count genes in the upper 1/16th:

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	8.035714e-08
2	8.035714e-08	group.2	1.666667e-07
3	1.666667e-07	group.3	2.395833e-07
4	2.395833e-07	group.4	3.520833e-07
5	3.520833e-07	group.5	4.791667e-07
6	4.791667e-07	group.6	6.562500e-07
7	6.562500e-07	group.7	9.134815e-07
8	9.134815e-07	group.8	1.278769e-06
9	1.278769e-06	group.9	1.760417e-06
10	1.760417e-06	group.10	5.283333e-06



Here the eﬀect of density of CpG islands is studied:

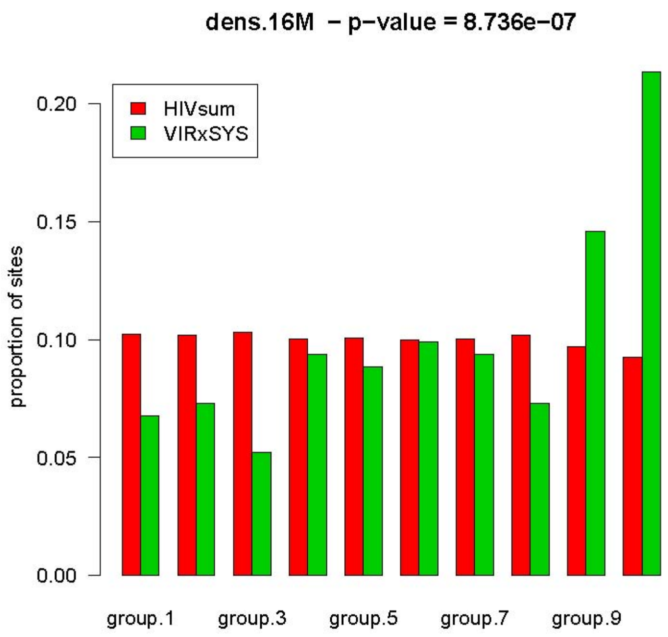
Category	limits		
	lower	category	upper
1	7.500000e-07	group.1	3.312500e-06
2	3.312500e-06	group.2	4.250000e-06
3	4.250000e-06	group.3	5.437500e-06
4	5.437500e-06	group.4	7.187500e-06
5	7.187500e-06	group.5	9.125000e-06
6	9.125000e-06	group.6	1.125000e-05
7	1.125000e-05	group.7	1.518750e-05
8	1.518750e-05	group.8	2.126807e-05
9	2.126807e-05	group.9	3.226199e-05
10	3.226199e-05	group.10	9.244836e-05



4.10 16 megaBase Window

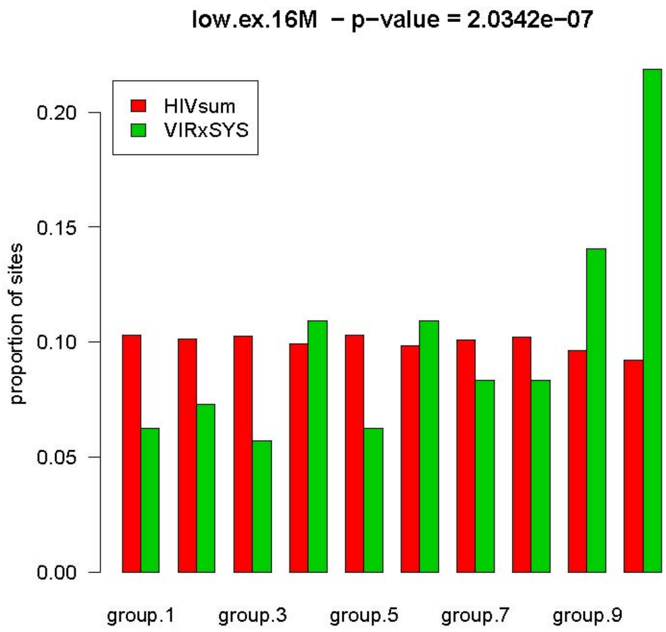
In the barplot that follows we examine the association of insertion sites with expression density in a 16 megabase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category	limits		
	lower	category	upper
1	5.234375e-07	group.1	1.785937e-06
2	1.785937e-06	group.2	2.255357e-06
3	2.255357e-06	group.2.	732366e-06
4	2.732366e-06	group.4	3.310640e-06
5	3.310640e-06	group.5	4.253125e-06
6	4.253125e-06	group.6	5.313889e-06
7	5.313889e-06	group.7	6.636458e-06
8	6.636458e-06	group.8	8.991695e-06
9	8.991695e-06	group.9	1.253222e-05
10	1.253222e-05	group.10	1.919554e-05



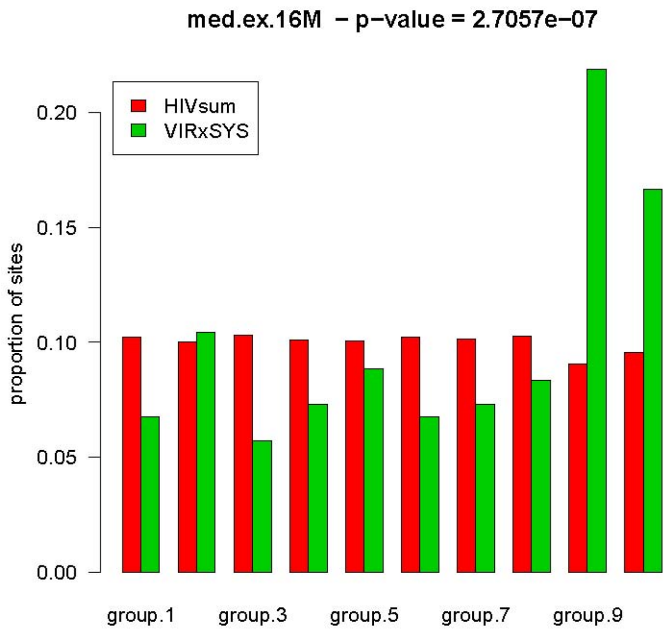
Below are the results for expression density. First, we count just genes that are in the upper half.

Category	limits		
	lower	category	upper
1	9.375000e-08	group.1	6.728175e-07
2	6.728175e-07	group.2	8.936508e-07
3	8.936508e-07	group.3	1.187500e-06
4	1.187500e-06	group.4	1.487054e-06
5	1.487054e-06	group.5	2.017113e-06
6	2.017113e-06	group.6	2.636161e-06
7	2.636161e-06	group.7	3.344345e-06
8	3.344345e-06	group.8	4.626190e-06
9	4.626190e-06	group.9	6.207258e-06
10	6.207258e-06	group.10	9.449033e-06



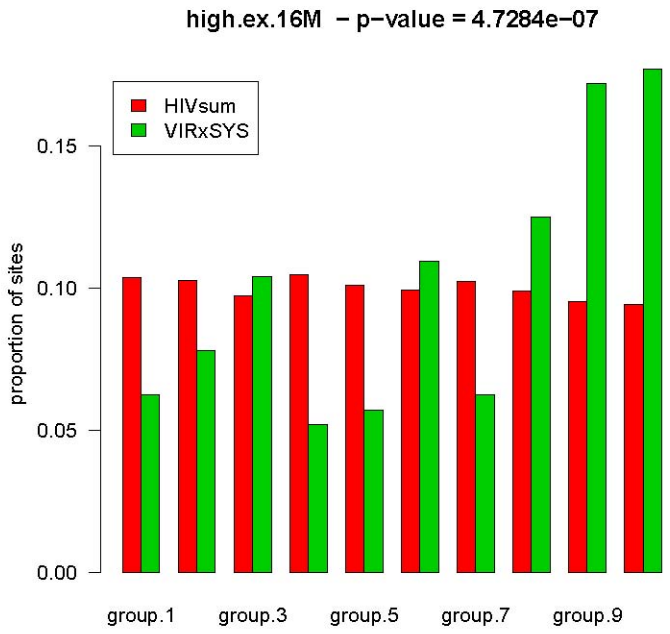
Now we count genes in the upper 1/8th:

Category	limits lower	category	upper
1	6.944444e-09	group.1	2.520833e-07
2	2.520833e-07	group.2	3.866071e-07
3	3.866071e-07	group.3	5.302083e-07
4	5.302083e-07	group.4	7.235119e-07
5	7.235119e-07	group.5	1.016121e-06
6	1.016121e-06	group.6	1.269494e-06
7	1.269494e-06	group.7	1.676562e-06
8	1.676562e-06	group.8	2.278720e-06
9	2.278720e-06	group.9	2.956250e-06
10	2.956250e-06	group.10	4.935491e-06



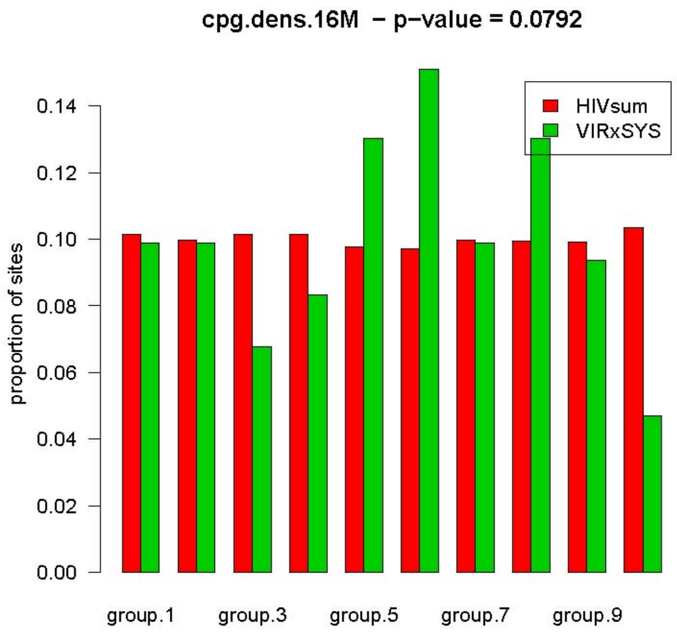
And here we count genes in the upper 1/16th:

Category	limits		
	lower	category	upper
1	0.000000e+00	group.1	1.145833e-07
2	1.145833e-07	group.2	1.875000e-07
3	1.875000e-07	group.3	2.635417e-07
4	2.635417e-07	group.4	3.593750e-07
5	3.593750e-07	group.5	4.699405e-07
6	4.699405e-07	group.6	5.875000e-07
7	5.875000e-07	group.7	7.602310e-07
8	7.602310e-07	group.8	1.063616e-06
9	1.063616e-06	group.9	1.427083e-06
10	1.427083e-06	group.10	3.197396e-06



Below the effect of density of CpG islands is studied:

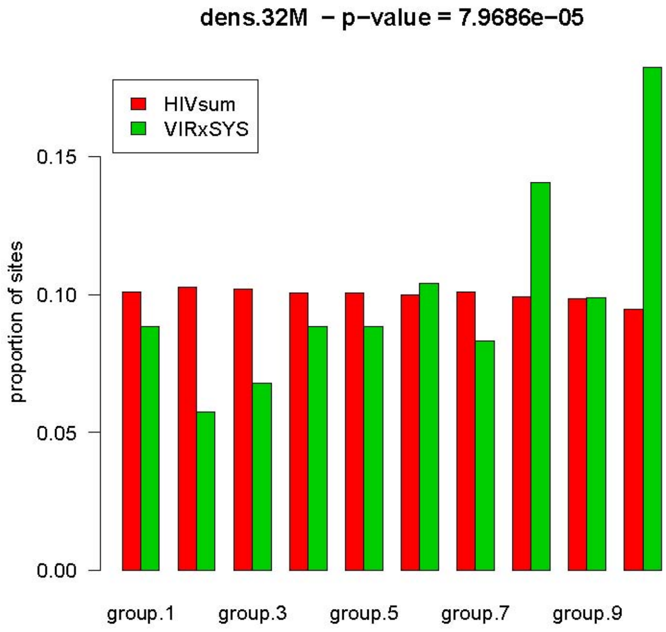
Category	limits		
	lower	category	upper
1	1.281250e-06	group.1	3.812500e-06
2	3.812500e-06	group.2	4.656250e-06
3	4.656250e-06	group.3	5.875000e-06
4	5.875000e-06	group.4	7.375000e-06
5	7.375000e-06	group.5	8.708920e-06
6	8.708920e-06	group.6	1.081250e-05
7	1.081250e-05	group.7	1.405475e-05
8	1.405475e-05	group.8	1.821875e-05
9	1.821875e-05	group.9	2.471450e-05
10	2.471450e-05	group.10	6.746383e-05



4.11 32 megaBase Window

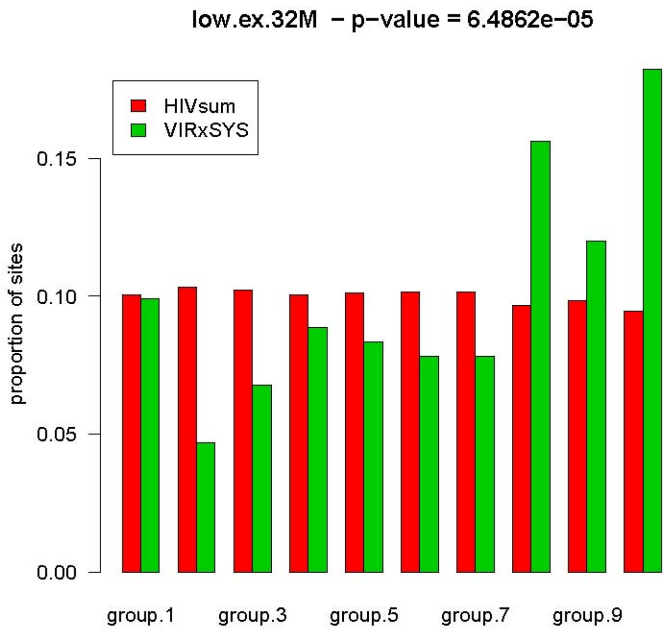
In the barplot that follows we examine the association of insertion sites with expression density in a 32 megabase window surrounding each locus. First, we count just the number of genes on the represented on the chip.

Category	limits		
	lower	category	upper
1	5.751546e-07	group.1	2.012364e-06
2	2.012364e-06	group.2	2.42194 e-06
3	2.421941e-06	group.3	2.852431e-06
4	2.852431e-06	group.4	3.51461 e-06
5	3.514612e-06	group.5	4.175781e-06
6	4.175781e-06	group.6	4.89815 e-06
7	4.898153e-06	group.7	5.907006e-06
8	5.907006e-06	group.8	7.8539 3e-06
9	7.853943e-06	group.9	9.208854e-06
10	9.208854e-06	group.10	1.85336 e-05



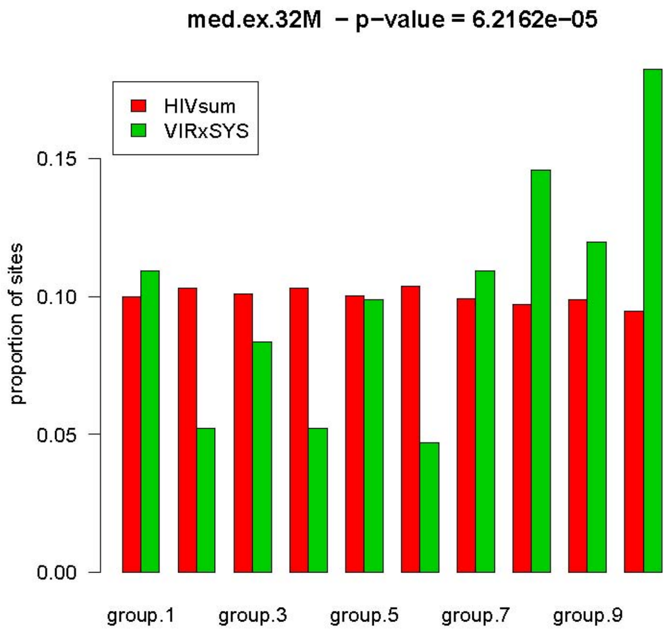
Here are the results for expression density. First, we count just genes that are in the upper half.

Category	limits	
	lower category	upper
1	2.440476e-07 group.1	8.057292e-07
2	8.057292e-07 group.2	1.025521e-06
3	1.025521e-06 group.3	1.261235e-06
4	1.261235e-06 group.4	1.615201e-06
5	1.615201e-06 group.5	1.920337e-06
6	1.920337e-06 group.6	2.323990e-06
7	2.323990e-06 group.7	3.006076e-06
8	3.006076e-06 group.8	3.946391e-06
9	3.946391e-06 group.9	4.559812e-06
10	4.559812e-06 group.10	9.250907e-06



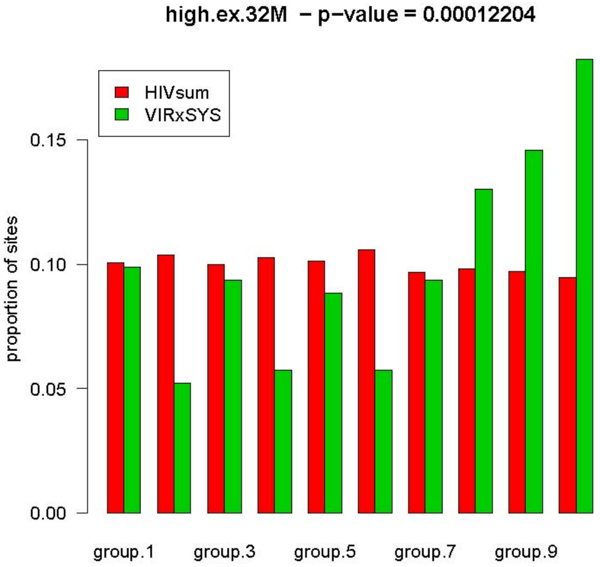
Now we count genes in the upper 1/8th:

Category	limits		
	lower	category	upper
1	9.173908e-08	group.1	3.326761e-07
2	3.326761e-07	group.2	4.633185e-07
3	4.633185e-07	group.3	5.906250e-07
4	5.906250e-07	group.4	7.760087e-07
5	7.760087e-07	group.5	9.290179e-07
6	9.290179e-07	group.6	1.147917e-06
7	1.147917e-06	group.7	1.535038e-06
8	1.535038e-06	group.8	1.859747e-06
9	1.859747e-06	group.9	2.364062e-06
10	2.364062e-06	group.10 4.	710958e-06



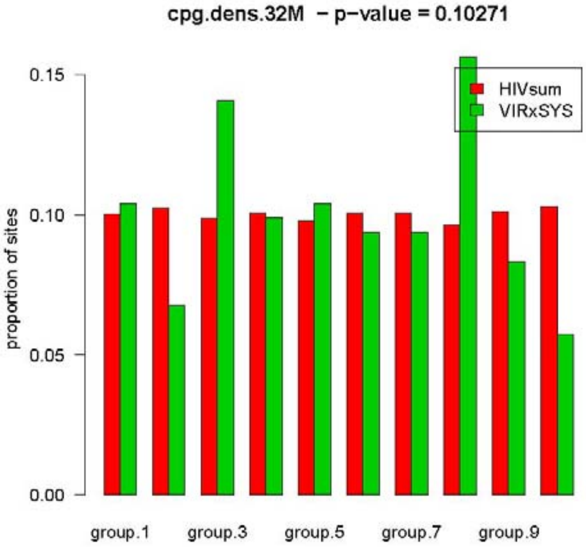
And here we count genes in the upper 1/16th:

Category	limits		
	lower	category	upper
1	1.896911e-08	group.1	1.598958e-07
2	1.598958e-07	group.2	2.213542e-07
3	2.213542e-07	group.3	2.904211e-07
4	2.904211e-07	group.4	3.878653e-07
5	3.878653e-07	group.5	4.654018e-07
6	4.654018e-07	group.6	5.791667e-07
7	5.791667e-07	group.7	7.002796e-07
8	7.002796e-07	group.8	8.462893e-07
9	8.462893e-07	group.9	1.235322e-06
10	1.235322e-06	group.10	2.499212e-06



Here the effect of density of CpG islands is studied:

Category	limits		
	lower	category	upper
1	2.671875e-06	group.1	4.468516e-06
2	4.468516e-06	group.2	5.609375e-06
3	5.609375e-06	group.3	6.468750e-06
4	6.468750e-06	group.4	7.296875e-06
5	7.296875e-06	group.5	8.266942e-06
6	8.266942e-06	group.6	1.000252e-05
7	1.000252e-05	group.7	1.227835e-05
8	1.227835e-05	group.8	1.594645e-05
9	1.594645e-05	group.9	2.066955e-05
10	2.066955e-05	group.10	4.072442e-05



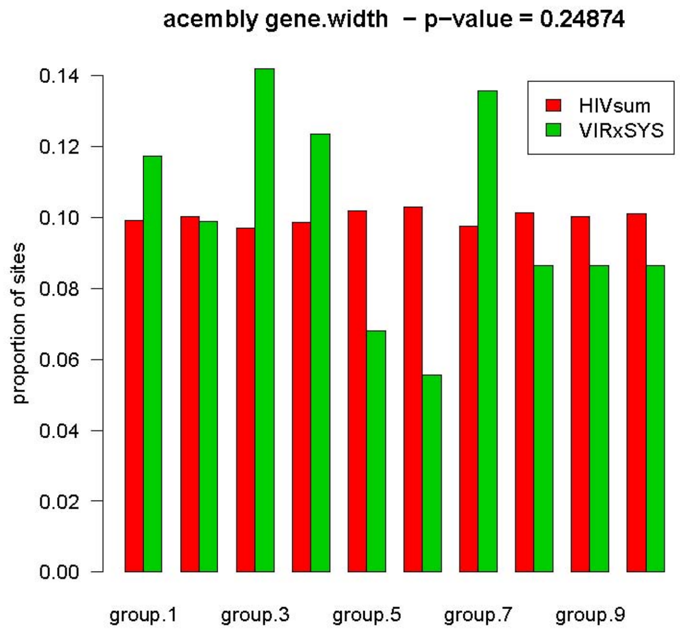
5 Juxtaposition with Gene Start and End Positions

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.

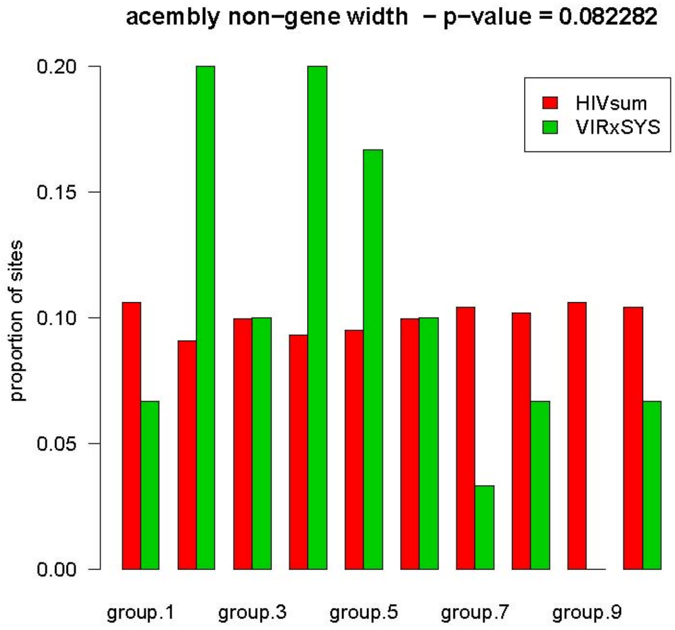
Category limits

	lower	category	upper
1	357.0	group.1	14301.5
2	14301.5	group.2	24460.0
3	24460.0	group.3	35436.0
4	35436.0	group.4	48532.0
5	48532.0	group.5	64500.0
6	64500.0	group.6	84305.0
7	84305.0	group.7	109498.0
8	109498.0	group.8	149730.0
9	149730.0	group.9	230286.5
10	230286.5	group.10	1529446.0



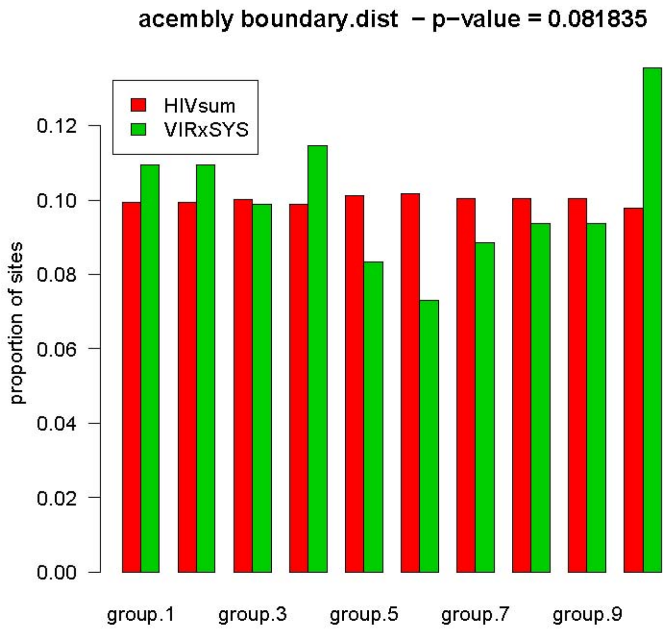
The next plot uses the width of a non-gene region for insertions that fall into such regions.

Category	limits		
	lower	category	upper
1	288.0	group.1	8670.0
2	8670.0	group.2	15495.0
3	15495.0	group.3	24554.0
4	24554.0	group.4	38135.0
5	38135.0	group.5	58338.5
6	58338.5	group.6	79834.2
7	79834.2	group.7	119887.2
8	119887.2	group.8	176510.4
9	176510.4	group.9	333186.5
10	333186.5	group.10	833936.0



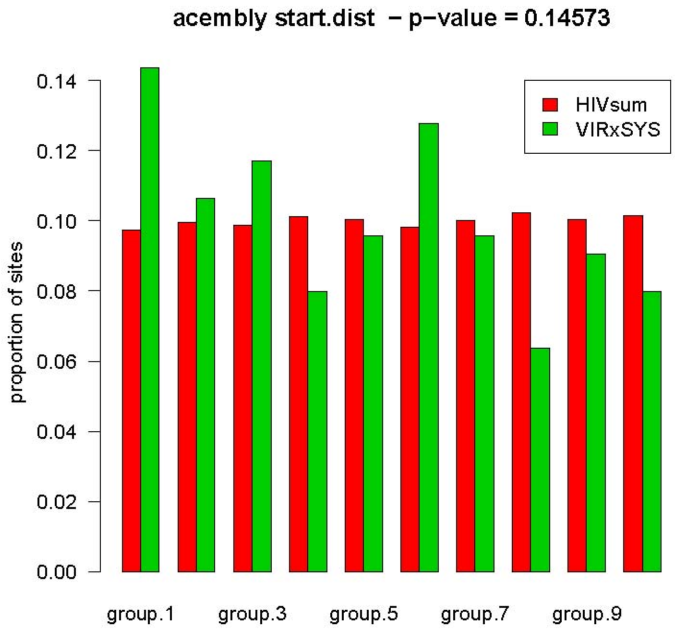
The next plot studies the distance to the nearest boundary between a gene and a non-gene region. The distance is expressed as a fraction of the length of the region. Thus, ’0.25’ refers to one quarter of the distance from the site to nearest boundary divided by the total width of the region.

Category	limits		
	lower	category	upper
1	5.665043e-05	group.1	0.05446415
2	5.446415e-02	group.2	0.10687480
3	1.068748e-01	group.3	0.16139174
4	1.613917e-01	group.4	0.21251964
5	2.125196e-01	group.5	0.25635738
6	2.563574e-01	group.6	0.30151305
7	3.015131e-01	group.7	0.35040906
8	3.504091e-01	group.8	0.40200305
9	4.020031e-01	group.9	0.45061212
10	4.506121e-01	group.10	0.49967076

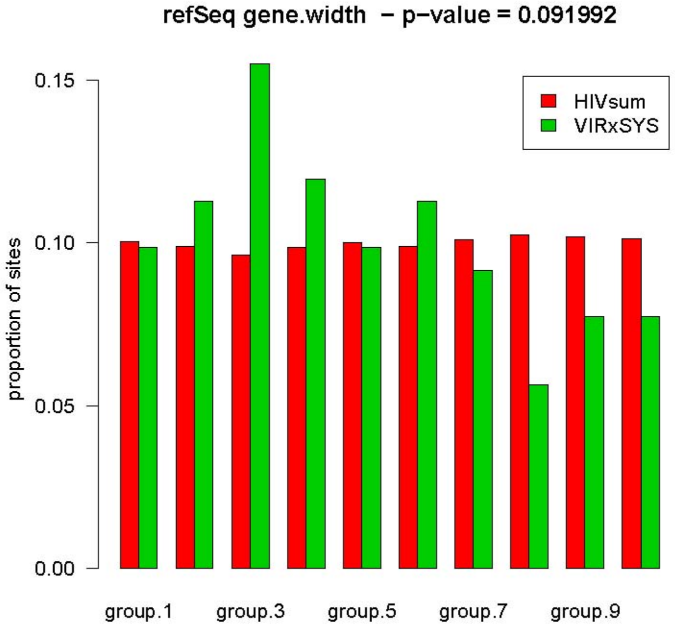


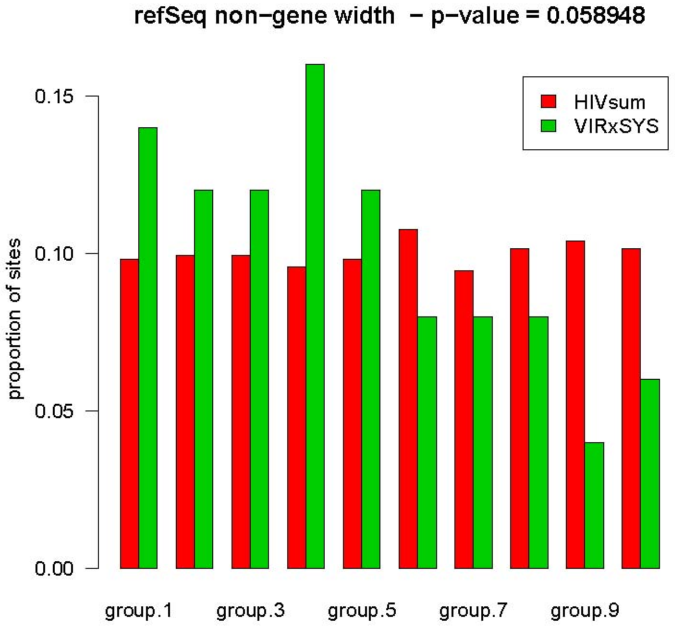
This plot studies the effect of nearness to the beginning of a transcript. For sites in genes, it is the distance to the start of the gene divided by the width of the gene. For other sites it is the distance from the site to the nearer gene if that gene boundary is also a transcription starting point.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

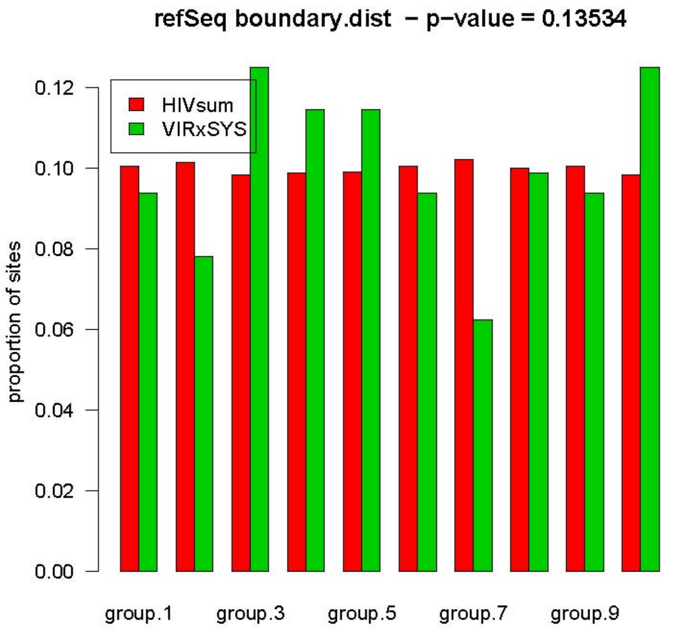
Category	limits		
	lower	category	upper
1	0.0009344262	group.1	0.09771473
2	0.0977147279	group.2	0.18896134
3	0.1889613398	group.3	0.27139459
4	0.2713945891	group.4	0.35206577
5	0.3520657743	group.5	0.44295463
6	0.4429546271	group.6	0.54284215
7	0.5428421477	group.7	0.64988290
8	0.6498828987	group.8	0.75860630
9	0.7586062998	group.9	0.87600479
10	0.8760047932	group.10	0.99994335

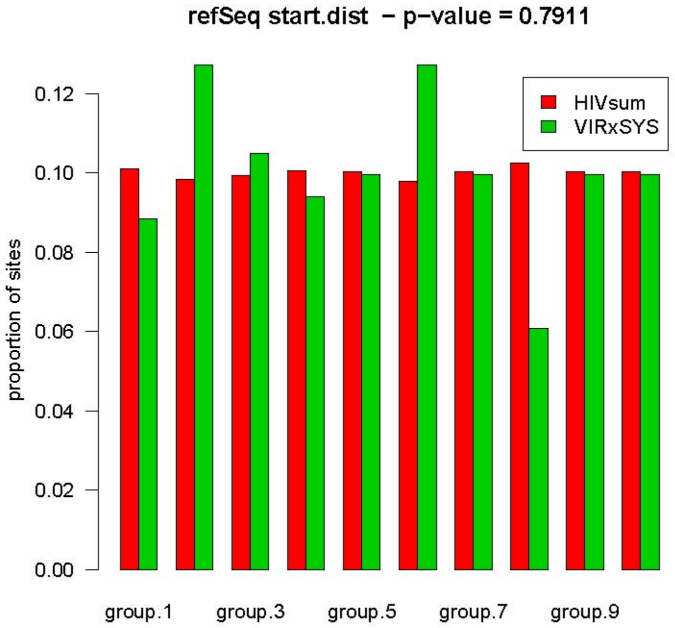

acembly start.dist – p–value = 0.14573

5.2 RefSeq Annotations

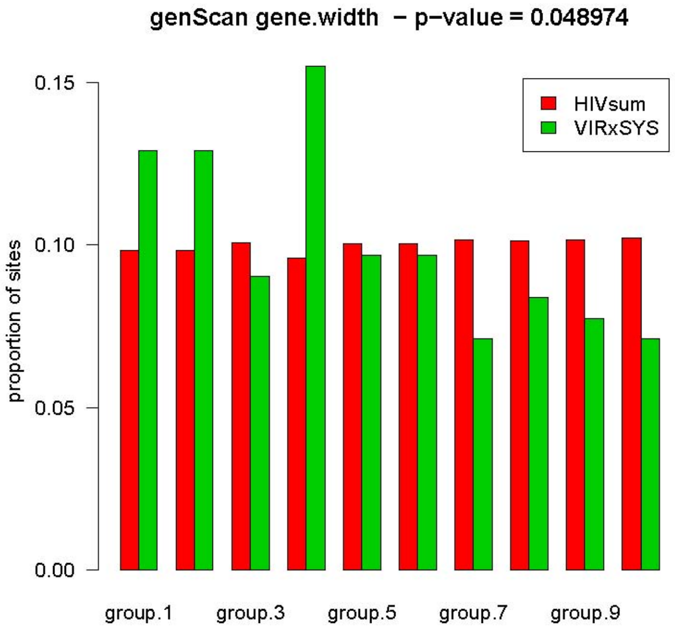


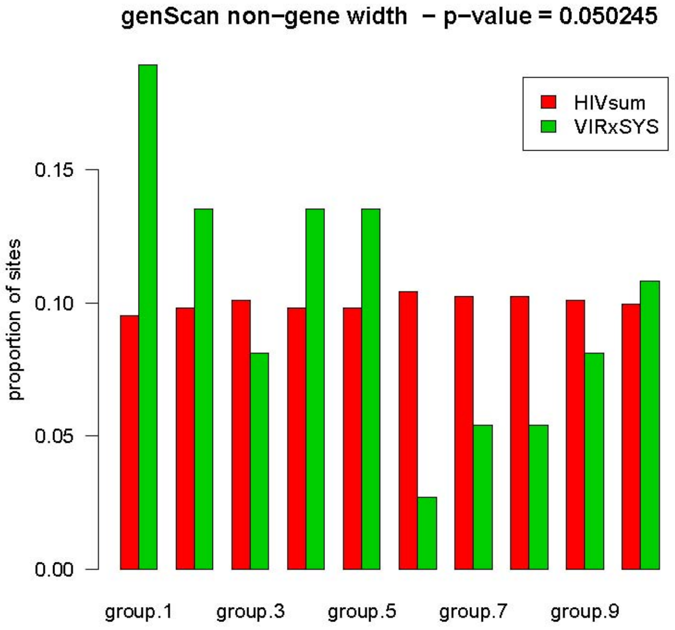


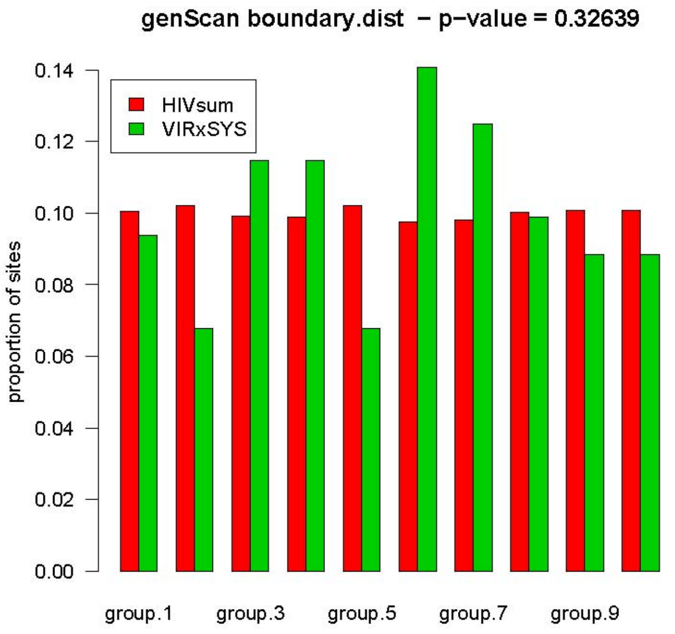

refSeq boundary.dist − p−value = 0.13534
HIVsum
VIRxSYS
proportion of sites
0.00
0.02
0.04
0.06
0.08
0.10
0.12
group.1
group.3
group.5
group.7
group.9

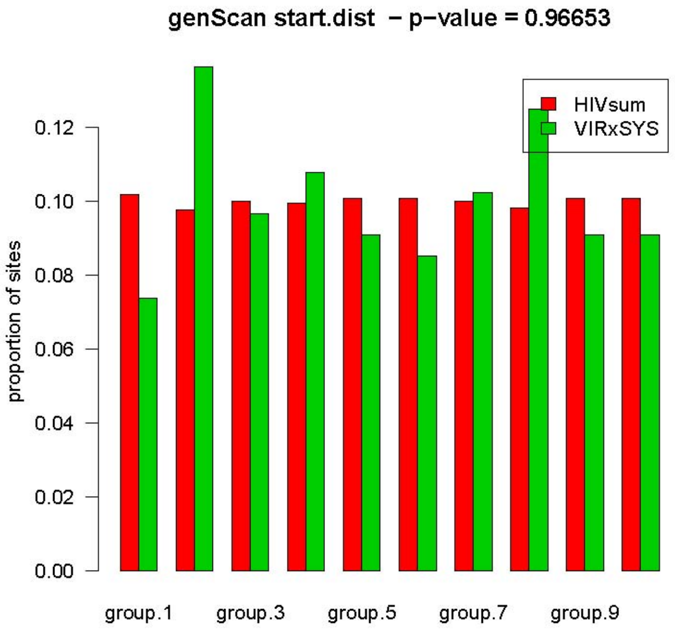

refSeq start.dist − p−value = 0.7911
HIVsum
VIRxSYS
proportion of sites
0.00
0.02
0.04
0.06
0.08
0.10
0.12
group.1
group.3
group.5
group.7
group.9

5.3 genScan Annotations

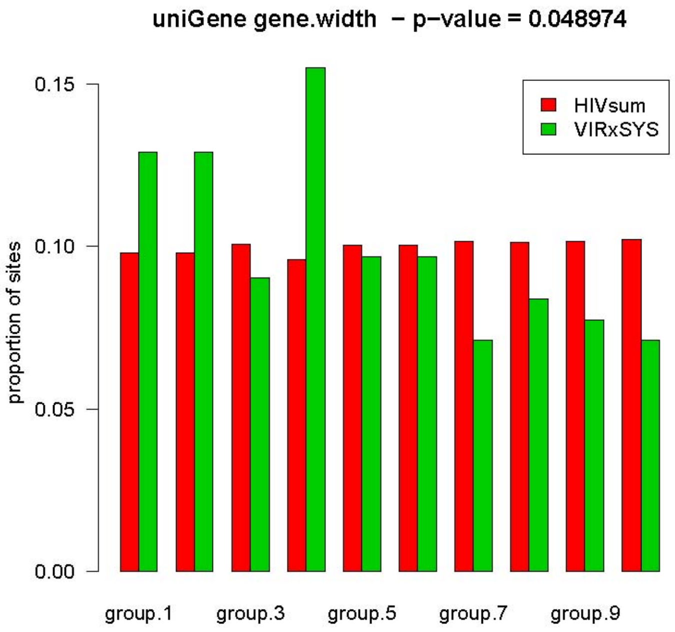


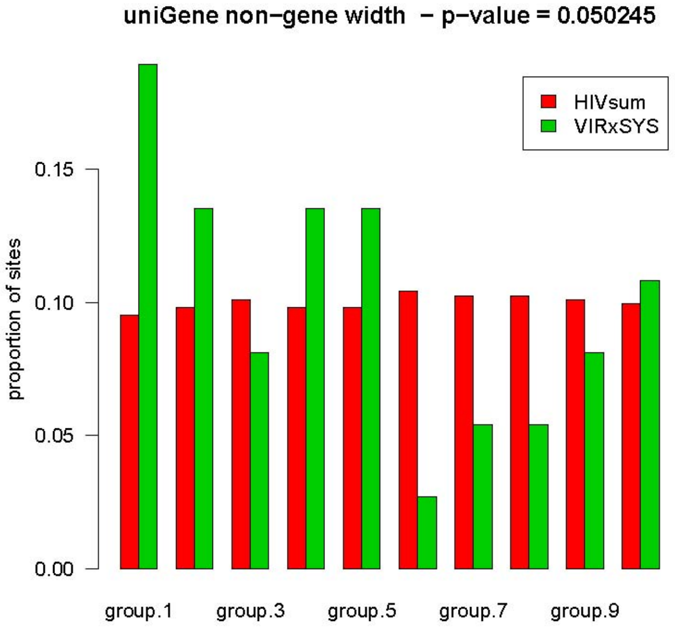


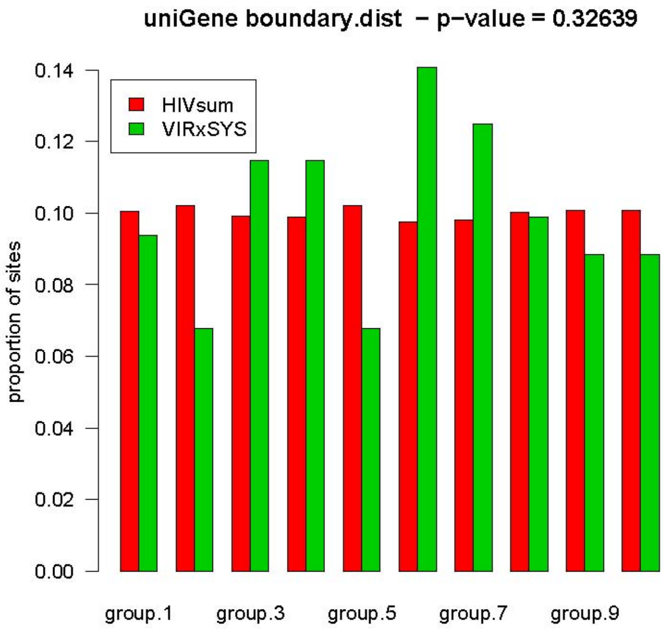

genScan boundary.dist − p−value = 0.32639
HIVsum
VIRxSYS
proportion of sites
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
group.1
group.3
group.5
group.7
group.9

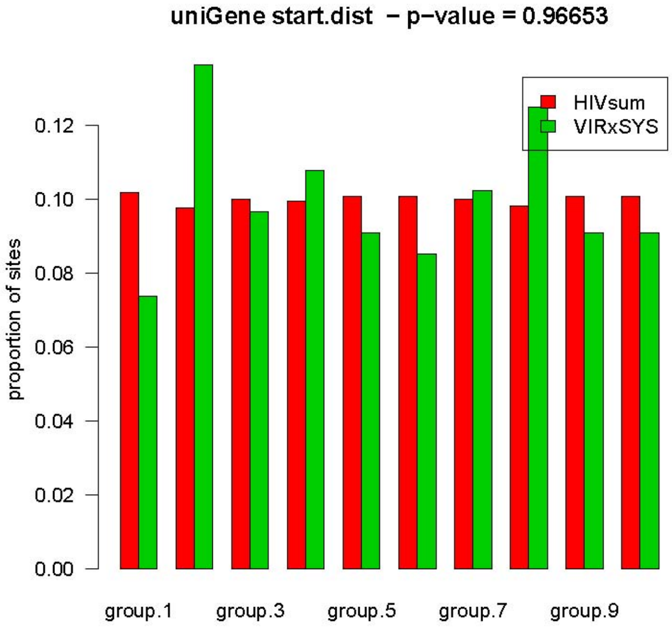

genScan start.dist − p−value = 0.96653
HIVsum
VIRxSYS
proportion of sites
0.12
0.10
0.08
0.06
0.04
0.02
0.00
group.1
group.3
group.5
group.7
group.9

5.4 uniGene Annotations






uniGene boundary.dist − p−value = 0.32639
HIVsum
VIRxSYS
proportion of sites
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
group.1
group.3
group.5
group.7
group.9

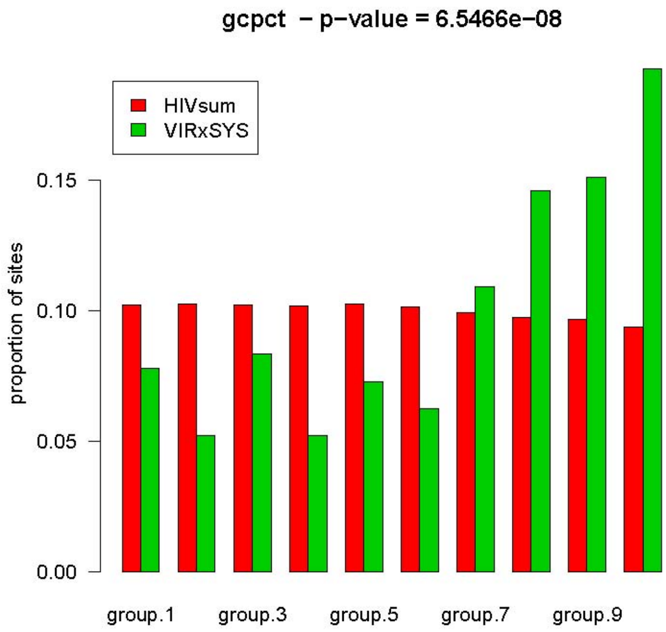

uniGene start.dist − p−value = 0.96653
HIVsum
VIRxSYS
proportion of sites
0.12
0.10
0.08
0.06
0.04
0.02
0.00
group.1
group.3
group.5
group.7
group.9

6 GC content

Here we study the effect of GC content on insertion. The GC content is taken from the Human Genome Draft at GoldenPath from the table http://genome.ucsc.edu/goldenPath/hg17/database/gc5Base.txt.gz.

Following the plot is a table of fitted coefficients based on splitting the GC percent data at the median.

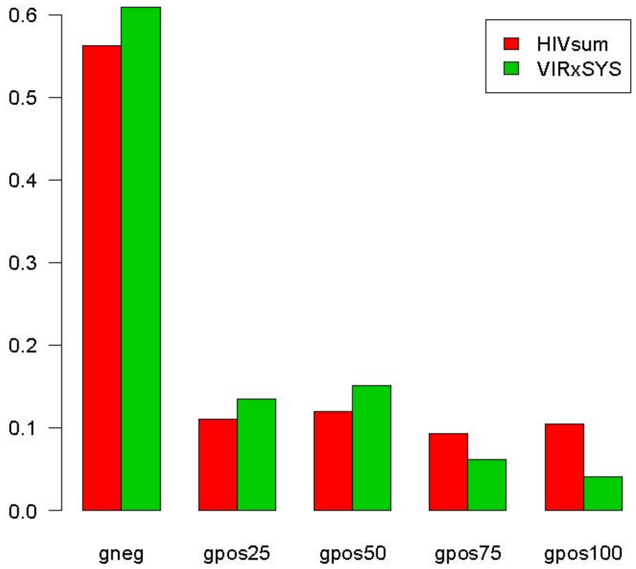
Category	limits		
	lower	category	upper
1	26.26953	group.1	33.86719
2	33.86719	group.2	35.44922
3	35.44922	group.3	36.73828
4	36.73828	group.4	38.06641
5	38.06641	group.5	39.33594
6	39.33594	group.6	40.76172
7	40.76172	group.7	42.50000
8	42.50000	group.8	44.72656
9	44.72656	group.9	48.69141
10	48.69141	group.10	66.44531



7 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 0.007.

Supporting References

1. MacGregor, R. R. (2001) *Hum. Gene Ther* **12**, 2028-2029.

2. Levine, B. L., Mosca, J., Riley, J. L., Carroll, R. G., Vahey, M. T., Jagodizinski, L., Wagner, K. F., Mayers, D. L., Burke, D. S., Weislow, O. S. *et al.* (1996) *Science* **272**, 1939-1943.

3. Levine, B. L., Bernstein, W., Craighead, N., Lindsten, T., Thompson, C. B. & June, C. H. (1997) *J Immunol* **159**, 5921-5930.

4. Hanenberg, H., Xiao, X. L., Dilloo, D., Hashino, K., Kato, I. & Williams, D. A. (1996) *Nat. Med* **2**, 876-882.

5. Pollok, K. E., Hanenberg, H., Noblitt, T. W., Schroeder, W. L., Kato, I., Emanuel, D. & Williams, D. A. (1998) *J Virol.* **72**, 4882-4892.

6. Levine, B. L., Cotte, J., Small, C. C., Carroll, R. G., Riley, J. L., Bernstein, W. B., Van Epps, D., Hardwick, R. A. & June, C. H. (1998) *J. Hematotherapy* **7**, 437-448.

7. Manilla, P., Rebello, T., Afable, C., Lu, X., Slepushkin, V., Humeau, L. M., Schonely, K., Ni, Y., Binder, G. K., Levine, B. L. *et al.* (2005) *Hum. Gene Ther* **16**, 17-25.

8. Schonely, K., Afable C, Slepushkin V, Lu X, Andre K, Boehmer J, Bengston K, Doub M, Cohen R, Berlinger D *et al.* (2003) *Bioprocessing Journal* **2**, 39-47.

9. Akatsuka, Y., Martin, E. G., Madonik, A., Barsoukov, A. A. & Hansen, J. A. (1999) *Tissue Antigens.* **53**, 122-134.

10. Schroder, A. R., Shinn, P., Chen, H., Berry, C., Ecker, J. R. & Bushman, F. (2002) *Cell* **110**, 521-529.

11. Wu, X., Li, Y., Crise, B. & Burgess, S. M. (2003) *Science.* **300**, 1749-1751.

12. Lewinski, M. K., Bisgrove, D., Shinn, P., Chen, H., Hoffmann, C., Hannenhalli, S., Verdin, E., Berry, C. C., Ecker, J. R. & Bushman, F. D. (2005) *J Virol.* **79**, 6610-6619.

13. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G. *et al.* (2004) *Proc. Natl. Acad. Sci. U. S. A.* **101**, 6062-6067.

14. McCullagh P & Nelder JA (1999) in *Generalized linear models* (Chapman & Hall, Ltd.