

Association of Genomic Features with Integration

Charles C. Berry

July 13, 2007

Contents

1	Introduction	2
2	Preference for Genes	4
2.1	refGene Genes	4
2.2	ensGenes	6
2.3	genScan Genes	8
2.4	oncogenes	11
3	CpG Island Neighborhoods	13
3.1	1 kilobase neighborhoods	13
3.2	5 kilobase neighborhoods	14
3.3	10 kilobase neighborhoods	15
3.4	25 kilobase neighborhoods	16
3.5	50 kilobase neighborhoods	17
4	Gene Density, Expression 'Density', and CpG Island Density	19
4.1	25 kilobase Window	19
4.2	50 kilobase Window	24
4.3	100 kilobase Window	29
4.4	250 kilobase Window	34
4.5	500 kilobase Window	39
4.6	1 megabase Window	44
4.7	2 megabase Window	49
4.8	4 megabase Window	54
4.9	8 megabase Window	59
4.10	16 megabase Window	64
4.11	32 megabase Window	69
5	Juxtaposition with Gene Start and End Positions	75
5.1	Refseq Annotations	75
5.2	genScan Annotations	79
6	Oncogenes	83

7 GC content	85
8 Cytobands	102

1 Introduction

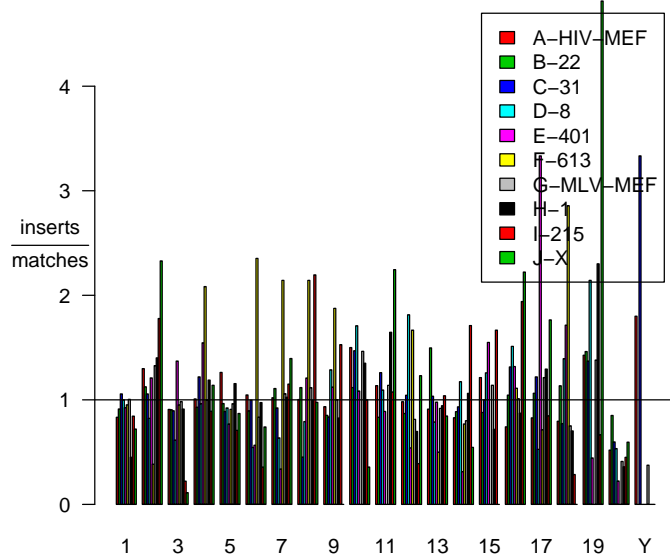
In this document, I examine the association of integration sites with various genomic features.

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:

Origin.of.data.set	type	
	insertion	match
A-HIV-MEF	2441	7308
B-22	716	7140
C-31	558	5550
D-8	418	4170
E-401	199	1990
F-613	37	330
G-MLV-MEF	4828	14475
H-1	478	4780
I-215	184	1840
J-X	136	1360

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:

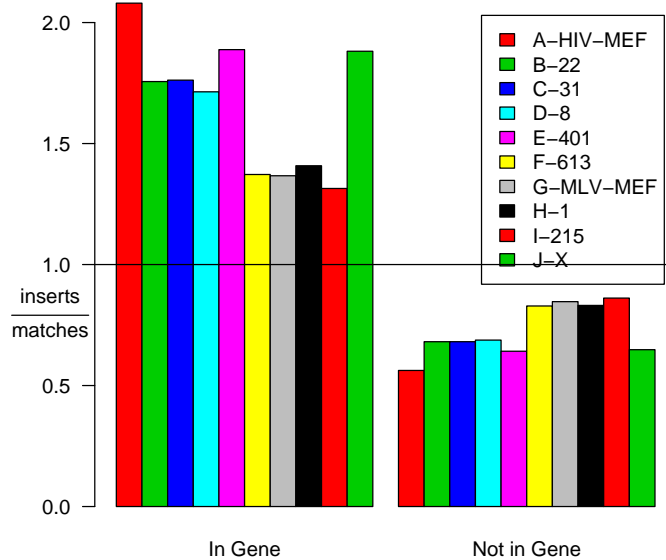


It seems evident that there are some chromosomes that are particularly favored for integration. This is reinforced by a test of statistical significance. The test performed used the likelihood ratio statistic for the multinomial logit model (reviewed in [2]) as implemented by the `clogit` function of 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 refGene 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 'refGene' annotation. The bars grouped over the label "In Gene" give the relative frequency of integration events (compared to control sites) between bases located within refGene gene annotations, while the label "Not in Gene" give the relative frequency of integration events (compared to control sites) between bases not located within refGene gene annotations.



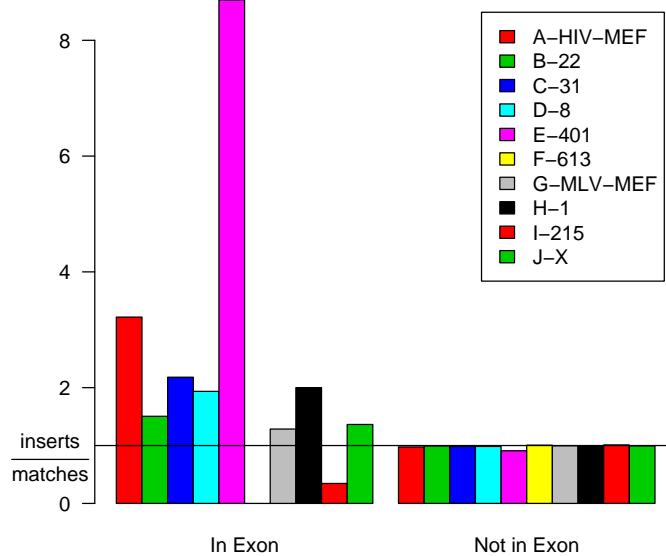
It seems evident that there is a strong tendency for insertions to occur in genes. A formal test of significance bears this out with a p-value of $< 2.22e-16$. Also, it appears that the tendency of different viruses to integrate into genes varies, and a test for this hypothesis attains $< 2.22e-16$. 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
A-HIV-MEF	1.320	0.0506	26.10	5.52e-150
B-22	0.950	0.0793	12.00	4.38e-33
C-31	0.959	0.0900	10.70	1.64e-26

D-8	0.902	0.1030	8.72	2.68e-18
E-401	1.080	0.1510	7.16	7.92e-13
F-613	0.597	0.3700	1.61	1.07e-01
G-MLV-MEF	0.475	0.0345	13.80	4.11e-43
H-1	0.527	0.0983	5.37	8.06e-08
I-215	0.423	0.1590	2.67	7.70e-03
J-X	1.080	0.1840	5.86	4.73e-09

As is evident, there are some differences in the coefficients. The largest coefficient is seen in the A-HIV-MEF data set, while the smallest is seen in the I-215 data set.

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

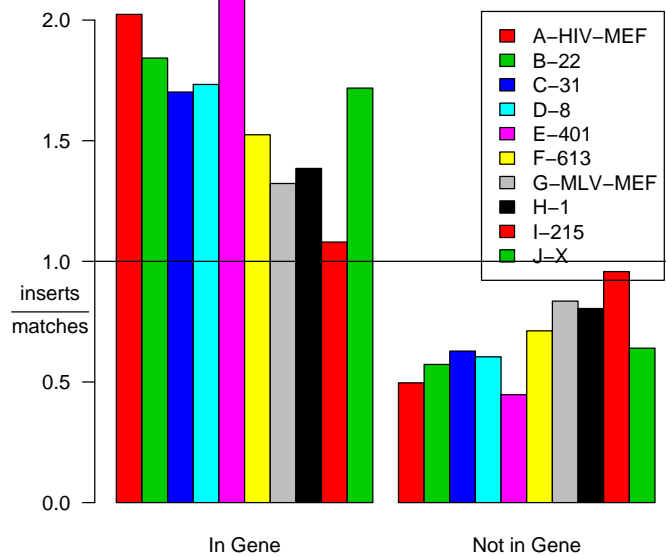
coef	se	z	p
------	----	---	---

A-HIV-MEF	0.5660	0.163	3.4700	5.22e-04
B-22	-0.1670	0.274	-0.6080	5.43e-01
C-31	0.2340	0.283	0.8250	4.09e-01
D-8	0.1510	0.311	0.4850	6.28e-01
E-401	1.7300	0.340	5.0800	3.76e-07
F-613	-11.3000	461.000	-0.0244	9.81e-01
G-MLV-MEF	-0.0655	0.131	-0.4990	6.18e-01
H-1	0.3850	0.315	1.2200	2.22e-01
I-215	-1.4200	1.030	-1.3800	1.68e-01
J-X	-0.3370	0.627	-0.5370	5.91e-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 both the introns and intergenic regions, so the impression given by the table may differ from that for the barplot.

2.2 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.



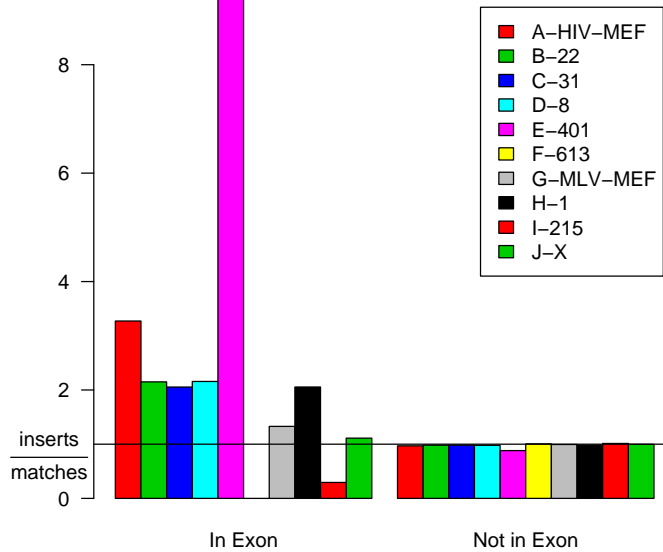
It seems evident that there is a strong tendency for insertions to occur in

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
A-HIV-MEF	1.420	0.0517	27.500	3.28e-166
B-22	1.170	0.0813	14.400	3.20e-47
C-31	1.020	0.0917	11.100	8.33e-29
D-8	1.050	0.1060	9.970	2.11e-23
E-401	1.570	0.1650	9.540	1.47e-21
F-613	0.766	0.3640	2.100	3.56e-02
G-MLV-MEF	0.456	0.0338	13.500	1.90e-41
H-1	0.545	0.0968	5.630	1.78e-08
I-215	0.120	0.1600	0.749	4.54e-01
J-X	0.988	0.1830	5.390	7.12e-08

As is evident, there are some differences in the coefficients. The largest coefficient is seen in the E-401 data set, while the smallest is seen in the I-215 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:

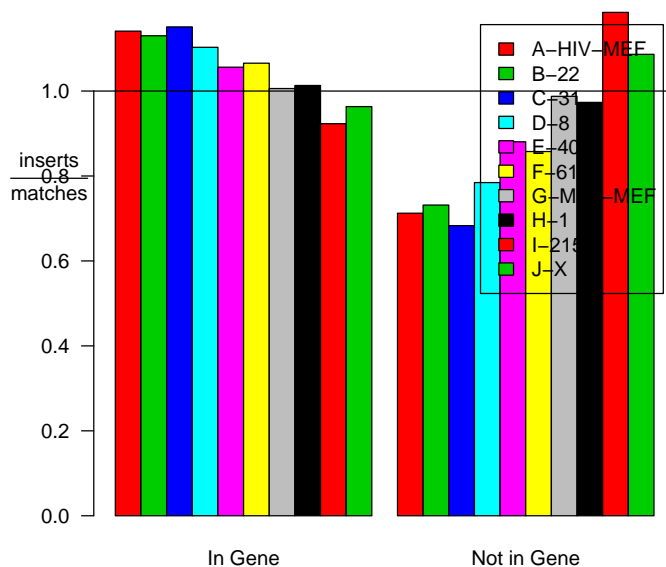
	coef	se	z	p
A-HIV-MEF	0.55400	0.147	3.7600	1.69e-04
B-22	0.15700	0.226	0.6940	4.88e-01
C-31	0.21100	0.260	0.8100	4.18e-01
D-8	0.26100	0.282	0.9230	3.56e-01
E-401	1.67000	0.300	5.5800	2.47e-08
F-613	-11.30000	464.000	-0.0243	9.81e-01
G-MLV-MEF	0.00932	0.120	0.0780	9.38e-01
H-1	0.42900	0.294	1.4600	1.45e-01
I-215	-1.35000	1.020	-1.3200	1.88e-01
J-X	-0.45500	0.624	-0.7300	4.65e-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.

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.



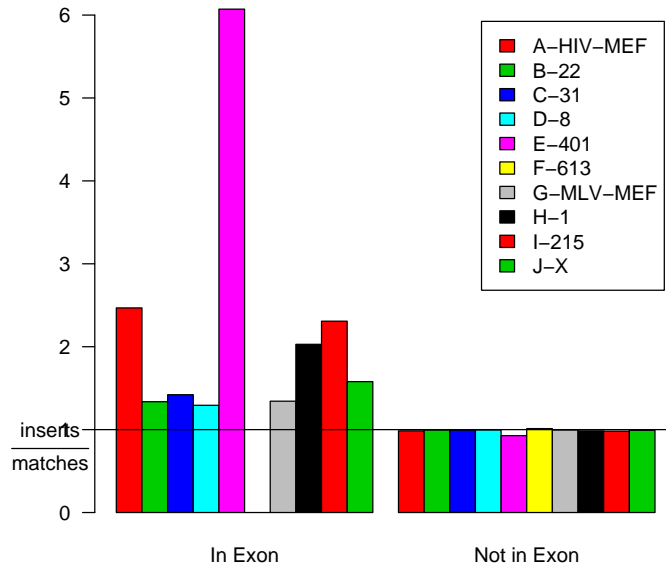
It seems evident that there is a strong tendency for insertions to occur in genes. A formal test of significance bears this out with a p-value of $1.3959e-15$. Also, it appears that the tendency of different viruses to integrate into genes varies, and a test for this hypothesis attains $3.3114e-14$. 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
A-HIV-MEF	0.4790	0.0542	8.840	$9.86e-19$
B-22	0.4450	0.0914	4.860	$1.15e-06$
C-31	0.5220	0.1060	4.920	$8.76e-07$
D-8	0.3370	0.1170	2.870	$4.05e-03$
E-401	0.1820	0.1650	1.100	$2.69e-01$
F-613	0.3590	0.4210	0.853	$3.94e-01$
G-MLV-MEF	0.0202	0.0356	0.569	$5.70e-01$
H-1	0.0404	0.1030	0.392	$6.95e-01$
I-215	-0.2500	0.1630	-1.530	$1.25e-01$
J-X	-0.1210	0.1940	-0.627	$5.30e-01$

As is evident, there are some differences in the coefficients. The largest coefficient is seen in the C-31 data set, while the smallest is seen in the I-215

data set.

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



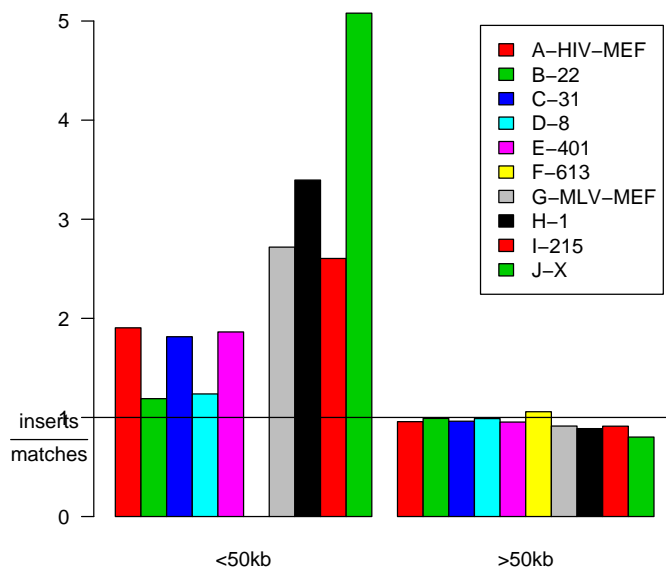
Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
A-HIV-MEF	0.828	0.179	4.6300	3.68e-06
B-22	0.164	0.299	0.5470	5.84e-01
C-31	0.222	0.325	0.6810	4.96e-01
D-8	0.169	0.403	0.4200	6.74e-01
E-401	1.860	0.324	5.7400	9.21e-09
F-613	-11.300	426.000	-0.0266	9.79e-01
G-MLV-MEF	0.294	0.131	2.2500	2.47e-02
H-1	0.733	0.303	2.4200	1.56e-02
I-215	0.939	0.460	2.0400	4.12e-02
J-X	0.502	0.629	0.7990	4.24e-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.

2.4 oncogenes

Here we examine the preference that insertions have for oncogenes. In the following plot we show the relative frequency of insertions with 50kb of an oncogene 5' end.



A formal test of oncogenic insertion returns p-value of $< 2.22e - 16$. The tendency of different viruses to integrate near oncogenes may vary, and a test for this hypothesis attains $5.7508e - 15$. 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
A-HIV-MEF	-0.678	0.0917	-7.4000	1.35e-13
B-22	-0.184	0.1660	-1.1100	2.67e-01
C-31	-0.631	0.1680	-3.7600	1.72e-04
D-8	-0.223	0.2220	-1.0100	3.15e-01
E-401	-0.657	0.2590	-2.5400	1.11e-02
F-613	12.200	329.0000	0.0372	9.70e-01
G-MLV-MEF	-1.090	0.0584	-18.7000	3.02e-78
H-1	-1.370	0.1480	-9.2600	2.14e-20
I-215	-1.060	0.2420	-4.3900	1.15e-05
J-X	-1.940	0.2540	-7.6600	1.93e-14
A-HIV-MEF	NA	0.0000	NA	NA

B-22	NA	0.0000	NA	NA
C-31	NA	0.0000	NA	NA
D-8	NA	0.0000	NA	NA
E-401	NA	0.0000	NA	NA
F-613	NA	0.0000	NA	NA
G-MLV-MEF	NA	0.0000	NA	NA
H-1	NA	0.0000	NA	NA
I-215	NA	0.0000	NA	NA
J-X	NA	0.0000	NA	NA

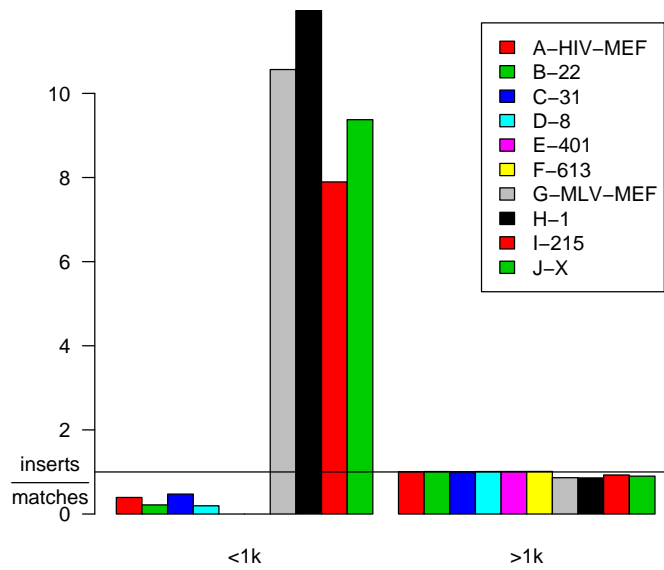
As is evident, there are some differences in the coefficients. The largest coefficient is seen in the F-613 data set, while the smallest is seen in the J-X data set.

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 $\pm 1\text{kb}$ of CpG islands are enriched for MLV insertions, we study such neighborhoods.

3.1 1 kilobase neighborhoods

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



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:

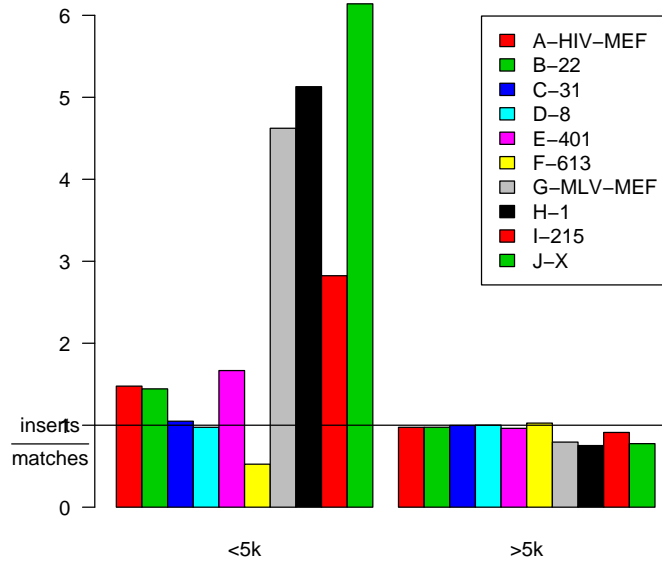
	coef	se	z	p
A-HIV-MEF	-0.947	3.39e-01	-2.8000	5.19e-03
B-22	-1.540	7.16e-01	-2.1400	3.20e-02
C-31	-0.749	5.93e-01	-1.2600	2.07e-01
D-8	-1.640	1.01e+00	-1.6200	1.05e-01
E-401	-12.900	4.27e+02	-0.0302	9.76e-01
F-613	-13.000	1.01e+03	-0.0128	9.90e-01
G-MLV-MEF	2.490	8.44e-02	29.5000	1.09e-190

H-1	2.640	1.87e-01	14.1000	1.96e-45
I-215	2.130	3.56e-01	5.9800	2.21e-09
J-X	2.320	3.73e-01	6.2100	5.21e-10

The largest coefficient is seen in the H-1 data set, while the smallest is seen in the F-613 data set.

3.2 5 kilobase neighborhoods

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

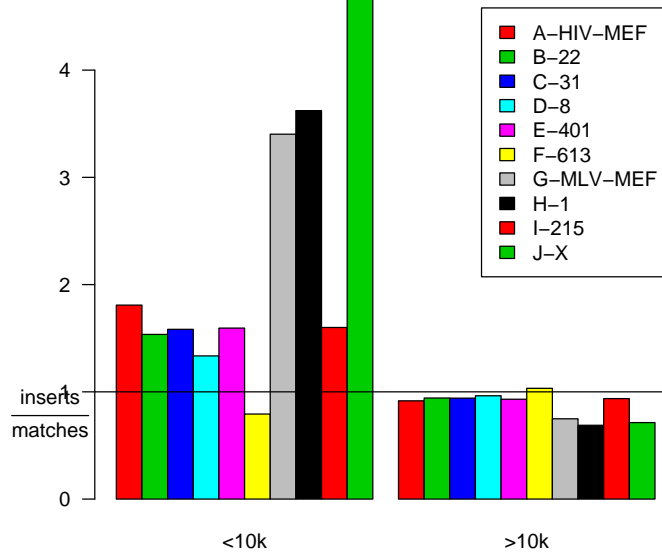
	coef	se	z	p
A-HIV-MEF	0.4120	0.0929	4.440	9.11e-06
B-22	0.3970	0.1500	2.640	8.26e-03
C-31	0.0559	0.1940	0.288	7.73e-01
D-8	-0.0249	0.2350	-0.106	9.16e-01
E-401	0.5540	0.2680	2.070	3.83e-02
F-613	-0.5560	1.0500	-0.530	5.96e-01

G-MLV-MEF	1.7800	0.0523	34.000	8.54e-254
H-1	1.9400	0.1220	15.900	3.34e-57
I-215	1.1300	0.2450	4.590	4.35e-06
J-X	2.0700	0.2430	8.480	2.23e-17

The largest coefficient is seen in the J-X data set, while the smallest is seen in the F-613 data set.

3.3 10 kilobase neighborhoods

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

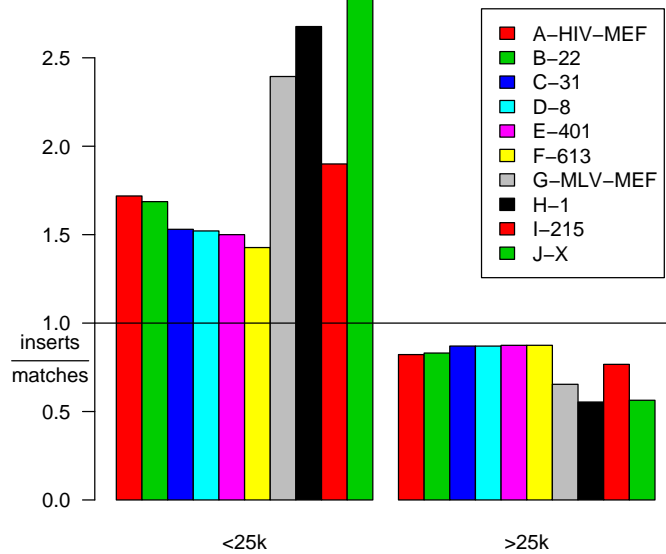
	coef	se	z	p
A-HIV-MEF	0.681	0.0676	10.100	7.28e-24
B-22	0.484	0.1130	4.280	1.87e-05
C-31	0.523	0.1290	4.070	4.67e-05
D-8	0.324	0.1520	2.120	3.36e-02
E-401	0.544	0.2060	2.640	8.19e-03

F-613	-0.448	0.6210	-0.721	4.71e-01
G-MLV-MEF	1.530	0.0439	34.800	8.27e-266
H-1	1.670	0.1060	15.700	2.04e-55
I-215	0.534	0.2200	2.430	1.52e-02
J-X	1.910	0.2160	8.830	1.06e-18

The largest coefficient is seen in the J-X data set, while the smallest is seen in the F-613 data set.

3.4 25 kilobase neighborhoods

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

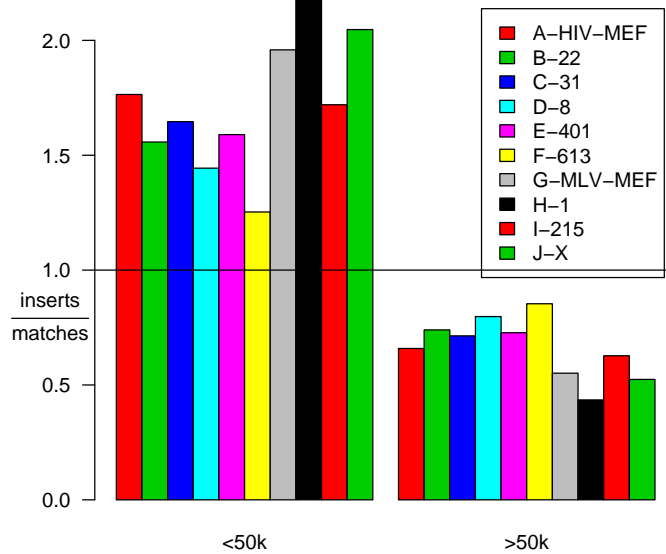
	coef	se	z	p
A-HIV-MEF	0.734	0.0521	14.10	4.33e-45
B-22	0.706	0.0848	8.32	8.63e-17
C-31	0.573	0.0986	5.81	6.08e-09
D-8	0.556	0.1130	4.94	7.73e-07

E-401	0.544	0.1650	3.29	9.93e-04
F-613	0.514	0.3850	1.33	1.82e-01
G-MLV-MEF	1.310	0.0370	35.30	1.04e-272
H-1	1.580	0.1000	15.80	3.48e-56
I-215	0.929	0.1640	5.67	1.45e-08
J-X	1.620	0.1880	8.59	8.40e-18

The largest coefficient is seen in the J-X data set, while the smallest is seen in the F-613 data set.

3.5 50 kilobase neighborhoods

The following plot shows the effect of being in or within ± 50 kb 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
A-HIV-MEF	0.985	0.0488	20.20	8.86e-91
B-22	0.746	0.0792	9.42	4.42e-21
C-31	0.850	0.0902	9.42	4.46e-21

D-8	0.598	0.1040	5.75	8.68e-09
E-401	0.782	0.1500	5.22	1.81e-07
F-613	0.488	0.3680	1.33	1.85e-01
G-MLV-MEF	1.280	0.0359	35.50	1.74e-276
H-1	1.620	0.1060	15.30	5.57e-53
I-215	1.030	0.1590	6.44	1.19e-10
J-X	1.360	0.1890	7.19	6.39e-13

The largest coefficient is seen in the H-1 data set, while the smallest is seen in the F-613 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] "MEF-GSE3400-MGU74Av2"
```

```
[1] "MG_U74"
```

Density data too sparse for barplot

	coef	se	z	p
A-HIV-MEF	1.010	0.0694	14.500	7.21e-48
B-22	0.717	0.1120	6.390	1.67e-10
C-31	0.743	0.1300	5.700	1.22e-08

D-8	0.448	0.1600	2.800	5.11e-03
E-401	1.070	0.2000	5.320	1.03e-07
F-613	0.212	0.4690	0.452	6.51e-01
G-MLV-MEF	0.943	0.0483	19.500	1.03e-84
H-1	1.270	0.1250	10.100	5.32e-24
I-215	0.424	0.2500	1.700	8.99e-02
J-X	1.190	0.2320	5.120	2.98e-07

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

Density data too sparse for barplot

	coef	se	z	p
A-HIV-MEF	1.2200	0.0856	14.200	1.02e-45
B-22	0.8940	0.1340	6.660	2.82e-11
C-31	0.8760	0.1620	5.390	6.98e-08
D-8	0.6740	0.1910	3.530	4.18e-04
E-401	1.1800	0.2480	4.750	2.04e-06
F-613	-0.0704	0.6300	-0.112	9.11e-01
G-MLV-MEF	1.1300	0.0593	19.000	2.02e-80
H-1	1.2000	0.1530	7.840	4.36e-15
I-215	0.1240	0.3450	0.360	7.18e-01
J-X	1.2700	0.2910	4.360	1.31e-05

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

Density data too sparse for barplot

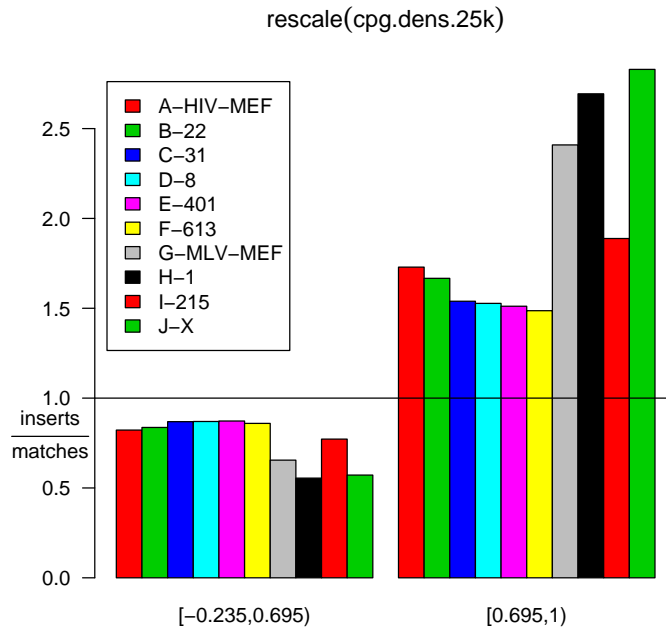
	coef	se	z	p
A-HIV-MEF	1.270	0.1120	11.4000	7.35e-30
B-22	0.679	0.1890	3.5900	3.30e-04
C-31	0.851	0.2160	3.9500	7.89e-05
D-8	0.534	0.2590	2.0600	3.92e-02
E-401	1.190	0.3180	3.7600	1.72e-04
F-613	-12.200	335.0000	-0.0363	9.71e-01
G-MLV-MEF	1.350	0.0768	17.6000	1.80e-69
H-1	1.190	0.1980	6.0100	1.87e-09
I-215	0.434	0.3910	1.1100	2.67e-01
J-X	0.743	0.4680	1.5900	1.12e-01

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

Density data too sparse for barplot

	coef	se	z	p
A-HIV-MEF	1.120	0.149	7.520	5.61e-14
B-22	0.489	0.264	1.850	6.38e-02
C-31	1.030	0.258	3.980	6.84e-05
D-8	0.659	0.367	1.790	7.27e-02
E-401	1.220	0.412	2.960	3.11e-03
F-613	-12.200	487.000	-0.025	9.80e-01
G-MLV-MEF	1.530	0.102	14.900	2.37e-50
H-1	1.430	0.248	5.770	7.80e-09
I-215	1.070	0.440	2.430	1.50e-02
J-X	-0.410	1.040	-0.395	6.93e-01

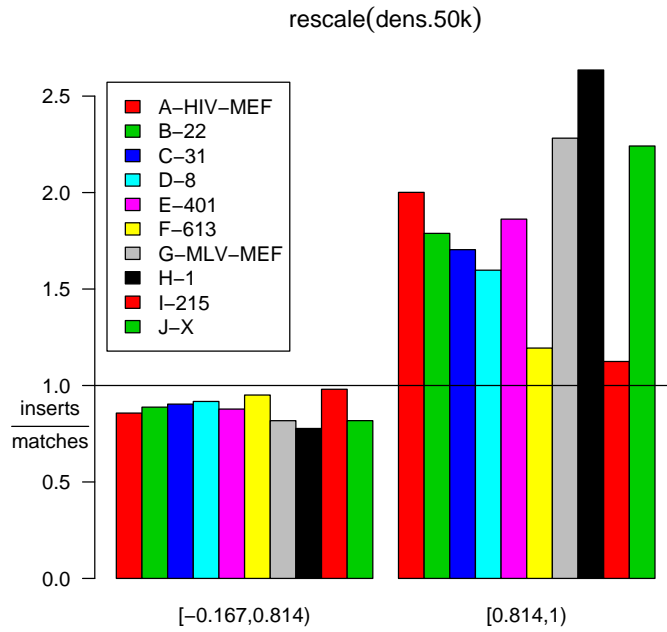
Here the effect of density of CpG islands is studied:



	coef	se	z	p
A-HIV-MEF	0.740	0.0523	14.10	2.05e-45
B-22	0.689	0.0852	8.09	5.96e-16
C-31	0.572	0.0987	5.80	6.77e-09
D-8	0.557	0.1130	4.94	7.97e-07
E-401	0.554	0.1650	3.35	8.05e-04
F-613	0.531	0.3850	1.38	1.68e-01
G-MLV-MEF	1.310	0.0371	35.30	1.89e-273
H-1	1.580	0.0999	15.80	1.72e-56
I-215	0.919	0.1650	5.58	2.40e-08
J-X	1.600	0.1880	8.50	1.83e-17

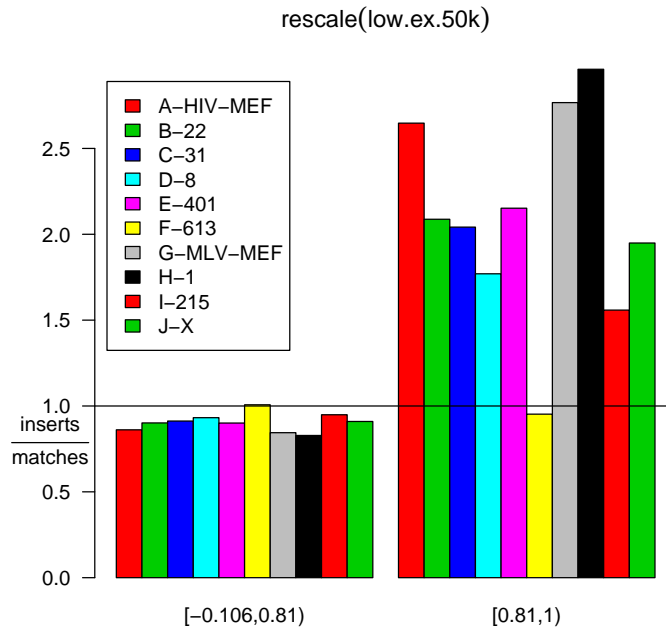
4.2 50 kilobase Window

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
A-HIV-MEF	0.924	0.0573	16.10	1.68e-58
B-22	0.736	0.0915	8.04	8.73e-16
C-31	0.663	0.1070	6.22	4.96e-10
D-8	0.530	0.1270	4.17	3.06e-05
E-401	0.847	0.1720	4.91	8.96e-07
F-613	0.275	0.4100	0.67	5.03e-01
G-MLV-MEF	1.040	0.0400	26.00	1.52e-148
H-1	1.190	0.1060	11.20	4.63e-29
I-215	0.353	0.1980	1.79	7.41e-02
J-X	0.964	0.2010	4.80	1.57e-06

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



	coef	se	z	p
A-HIV-MEF	1.1300	0.0674	16.800	2.55e-63
B-22	0.8340	0.1050	7.920	2.45e-15
C-31	0.7740	0.1250	6.200	5.75e-10
D-8	0.6790	0.1460	4.650	3.36e-06
E-401	0.8590	0.2070	4.150	3.29e-05
F-613	-0.0825	0.5570	-0.148	8.82e-01
G-MLV-MEF	1.1700	0.0462	25.300	8.00e-141
H-1	1.2900	0.1190	10.900	1.29e-27
I-215	0.4710	0.2350	2.000	4.52e-02
J-X	0.7760	0.2440	3.180	1.48e-03

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

Density data too sparse for barplot

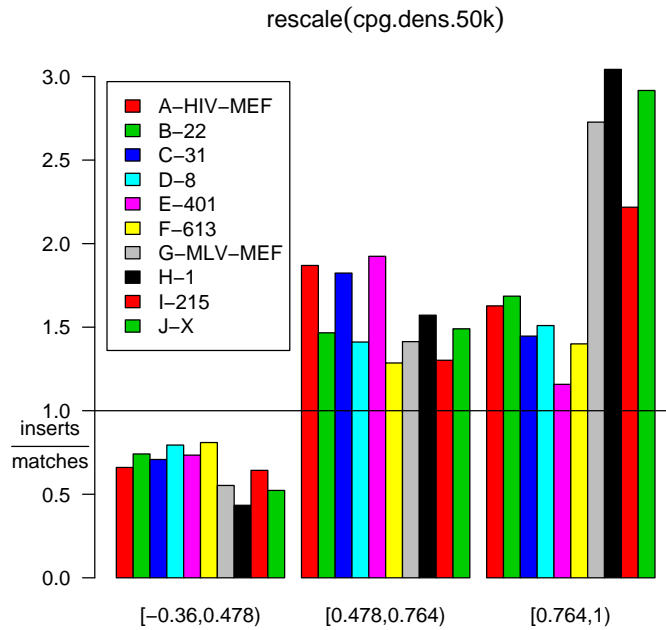
	coef	se	z	p
A-HIV-MEF	1.130	0.0841	13.500	3.02e-41
B-22	0.710	0.1410	5.050	4.35e-07
C-31	0.739	0.1580	4.660	3.11e-06
D-8	0.383	0.2020	1.900	5.79e-02
E-401	0.557	0.2810	1.980	4.77e-02
F-613	-0.906	1.0300	-0.878	3.80e-01
G-MLV-MEF	1.270	0.0574	22.100	2.02e-108
H-1	1.300	0.1440	9.050	1.39e-19
I-215	0.725	0.2740	2.650	8.08e-03
J-X	0.830	0.3010	2.750	5.89e-03

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

Density data too sparse for barplot

	coef	se	z	p
A-HIV-MEF	1.0300	0.1120	9.1800	4.29e-20
B-22	0.6600	0.1880	3.5000	4.63e-04
C-31	0.8570	0.1980	4.3200	1.57e-05
D-8	0.0671	0.3200	0.2090	8.34e-01
E-401	0.8540	0.3410	2.5000	1.23e-02
F-613	-12.2000	399.0000	-0.0305	9.76e-01
G-MLV-MEF	1.4000	0.0748	18.7000	3.05e-78
H-1	1.3400	0.1800	7.4400	1.01e-13
I-215	1.1400	0.3280	3.4800	5.04e-04
J-X	1.0800	0.3660	2.9500	3.18e-03

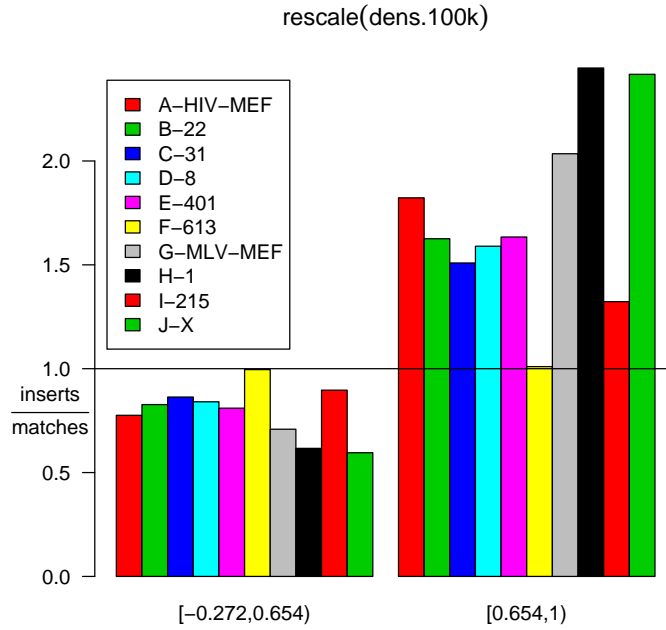
Here the effect of density of CpG islands is studied:



	coef	se	z	p
A-HIV-MEF	0.981	0.0487	20.10	4.00e-90
B-22	0.746	0.0792	9.42	4.62e-21
C-31	0.856	0.0902	9.50	2.16e-21
D-8	0.604	0.1040	5.81	6.16e-09
E-401	0.765	0.1500	5.10	3.38e-07
F-613	0.500	0.3680	1.36	1.73e-01
G-MLV-MEF	1.270	0.0359	35.40	3.45e-275
H-1	1.620	0.1060	15.40	2.59e-53
I-215	0.982	0.1590	6.19	6.13e-10
J-X	1.370	0.1890	7.24	4.63e-13

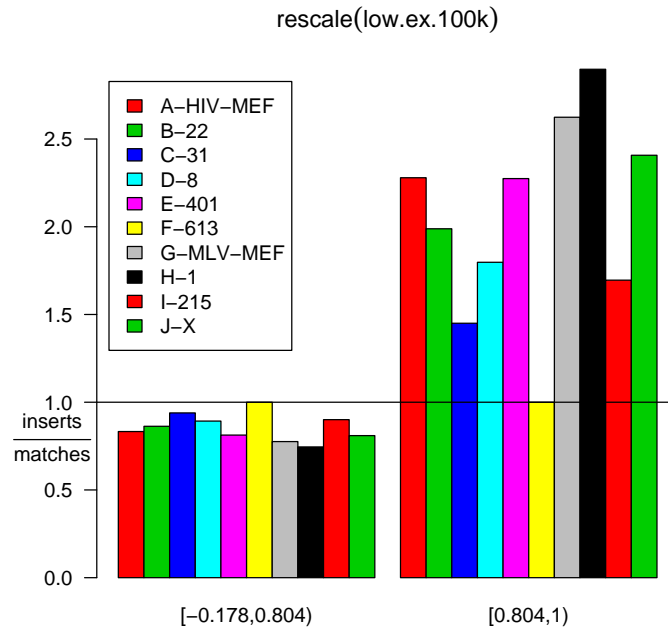
4.3 100 kilobase Window

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.



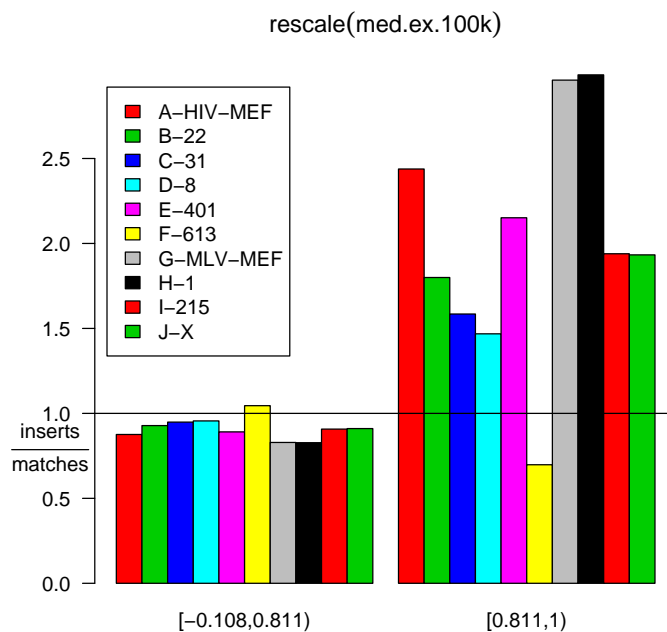
	coef	se	z	p
A-HIV-MEF	0.9280	0.0504	18.400	1.15e-75
B-22	0.7100	0.0820	8.660	4.74e-18
C-31	0.5730	0.0949	6.040	1.51e-09
D-8	0.6250	0.1080	5.780	7.34e-09
E-401	0.6890	0.1540	4.480	7.39e-06
F-613	-0.0707	0.3970	-0.178	8.58e-01
G-MLV-MEF	1.0700	0.0358	29.900	7.97e-196
H-1	1.3700	0.0989	13.900	6.66e-44
I-215	0.5460	0.1630	3.350	8.22e-04
J-X	1.6400	0.1890	8.700	3.22e-18

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



	coef	se	z	p
A-HIV-MEF	1.080	0.0561	19.30	1.29e-82
B-22	0.884	0.0886	9.97	2.05e-23
C-31	0.619	0.1060	5.83	5.62e-09
D-8	0.709	0.1210	5.88	4.06e-09
E-401	0.927	0.1660	5.60	2.12e-08
F-613	-0.195	0.4740	-0.41	6.81e-01
G-MLV-MEF	1.210	0.0389	31.20	1.39e-213
H-1	1.320	0.1040	12.70	4.99e-37
I-215	0.913	0.1750	5.22	1.81e-07
J-X	1.250	0.1950	6.42	1.33e-10

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



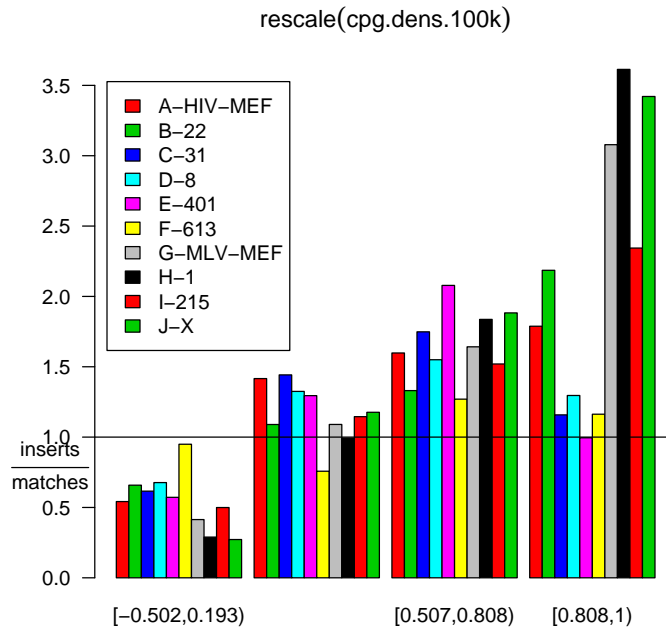
	coef	se	z	p
A-HIV-MEF	1.040	0.0662	15.800	5.24e-56
B-22	0.628	0.1130	5.570	2.57e-08
C-31	0.526	0.1320	3.980	7.02e-05
D-8	0.463	0.1530	3.020	2.49e-03
E-401	0.896	0.1990	4.500	6.88e-06
F-613	-0.135	0.5560	-0.242	8.09e-01
G-MLV-MEF	1.250	0.0455	27.500	4.47e-167
H-1	1.260	0.1200	10.500	1.20e-25
I-215	0.764	0.2150	3.550	3.83e-04
J-X	0.852	0.2370	3.590	3.27e-04

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

Density data too sparse for barplot

	coef	se	z	p
A-HIV-MEF	0.922	0.0874	10.5000	5.29e-26
B-22	0.684	0.1450	4.7400	2.19e-06
C-31	0.573	0.1640	3.5000	4.70e-04
D-8	0.143	0.2310	0.6190	5.36e-01
E-401	0.834	0.2670	3.1300	1.78e-03
F-613	-13.300	463.0000	-0.0286	9.77e-01
G-MLV-MEF	1.300	0.0569	22.9000	2.71e-116
H-1	1.160	0.1460	7.9400	2.03e-15
I-215	1.050	0.2620	4.0000	6.34e-05
J-X	1.010	0.2960	3.4100	6.61e-04

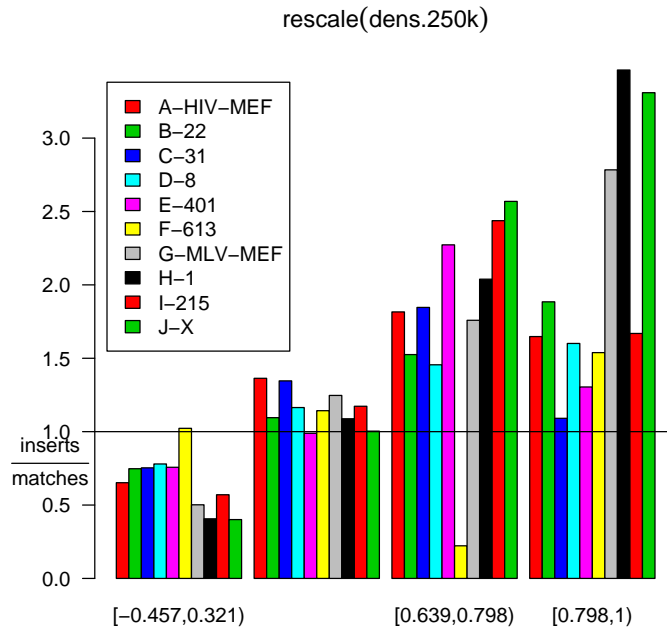
Here the effect of density of CpG islands is studied:



	coef	se	z	p
A-HIV-MEF	0.779	0.0491	15.90	1.24e-56
B-22	0.746	0.0802	9.30	1.35e-20
C-31	0.613	0.0925	6.63	3.27e-11
D-8	0.554	0.1070	5.19	2.07e-07
E-401	0.778	0.1500	5.17	2.28e-07
F-613	0.429	0.3660	1.17	2.40e-01
G-MLV-MEF	1.290	0.0358	36.00	3.27e-283
H-1	1.640	0.1020	16.10	2.70e-58
I-215	1.030	0.1600	6.44	1.22e-10
J-X	1.540	0.1900	8.09	6.05e-16

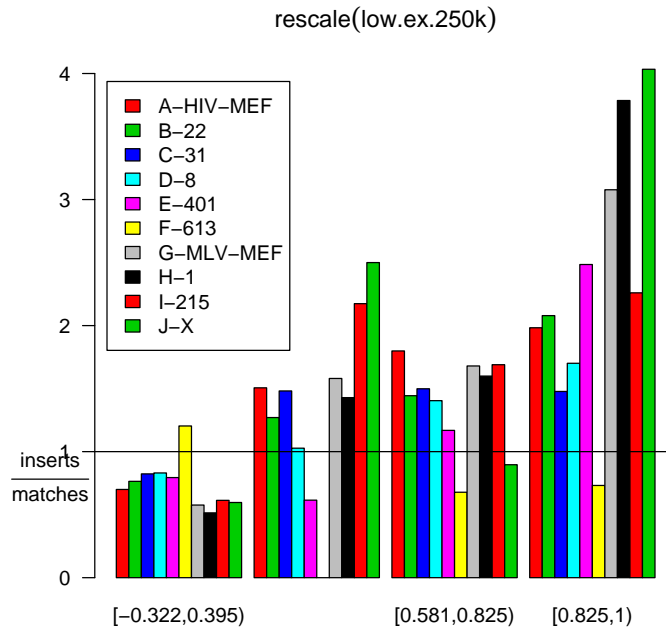
4.4 250 kilobase Window

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.



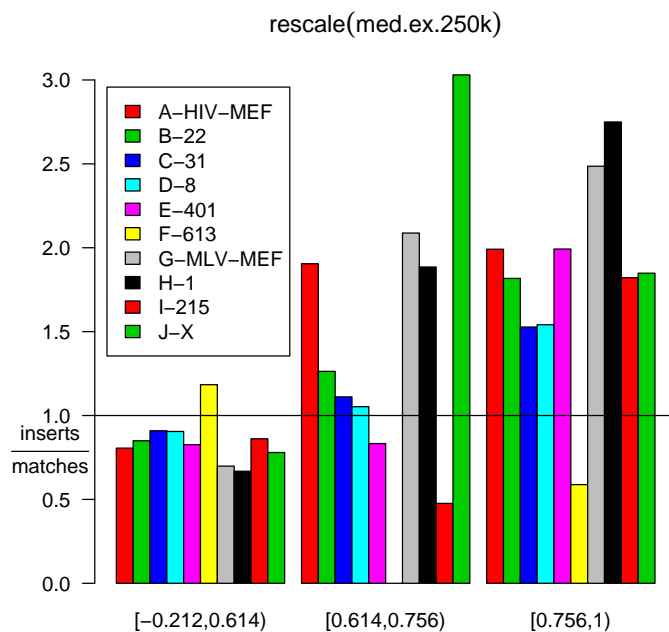
	coef	se	z	p
A-HIV-MEF	0.8650	0.0490	17.7000	8.22e-70
B-22	0.6380	0.0793	8.0400	8.69e-16
C-31	0.6140	0.0899	6.8300	8.54e-12
D-8	0.5110	0.1030	4.9700	6.65e-07
E-401	0.6410	0.1510	4.2400	2.21e-05
F-613	0.0122	0.3660	0.0333	9.73e-01
G-MLV-MEF	1.2800	0.0371	34.4000	1.31e-259
H-1	1.5500	0.1130	13.7000	6.61e-43
I-215	1.0800	0.1690	6.4100	1.48e-10
J-X	1.5100	0.2080	7.2300	4.89e-13

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



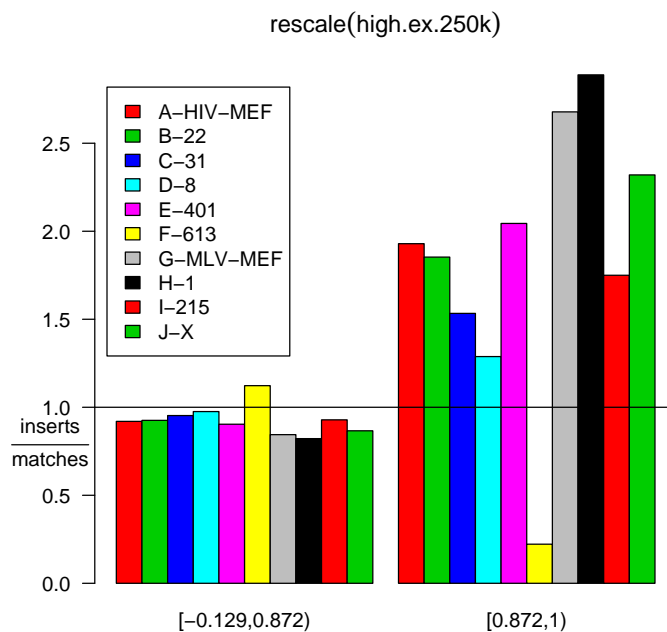
	coef	se	z	p
A-HIV-MEF	0.978	0.0494	19.80	2.87e-87
B-22	0.752	0.0796	9.46	3.22e-21
C-31	0.603	0.0927	6.51	7.46e-11
D-8	0.561	0.1050	5.32	1.02e-07
E-401	0.646	0.1520	4.26	2.08e-05
F-613	-0.499	0.4220	-1.18	2.37e-01
G-MLV-MEF	1.300	0.0358	36.40	1.32e-289
H-1	1.490	0.1010	14.80	1.60e-49
I-215	1.120	0.1580	7.07	1.58e-12
J-X	1.230	0.1850	6.64	3.14e-11

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



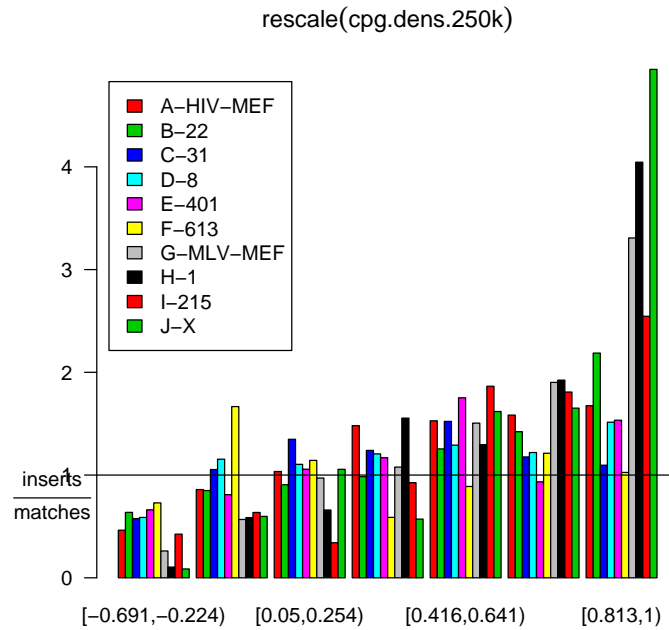
	coef	se	z	p
A-HIV-MEF	0.907	0.0536	16.90	2.82e-64
B-22	0.692	0.0876	7.90	2.74e-15
C-31	0.500	0.1050	4.78	1.78e-06
D-8	0.501	0.1180	4.26	2.07e-05
E-401	0.774	0.1620	4.77	1.84e-06
F-613	-0.659	0.4950	-1.33	1.83e-01
G-MLV-MEF	1.250	0.0371	33.70	1.76e-248
H-1	1.350	0.1010	13.40	9.80e-41
I-215	0.569	0.1720	3.31	9.19e-04
J-X	0.922	0.1900	4.86	1.18e-06

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



	coef	se	z	p
A-HIV-MEF	0.817	0.0649	12.60	2.41e-36
B-22	0.621	0.1060	5.86	4.75e-09
C-31	0.466	0.1260	3.71	2.10e-04
D-8	0.282	0.1520	1.86	6.30e-02
E-401	0.806	0.1890	4.27	1.98e-05
F-613	-1.920	1.0300	-1.87	6.12e-02
G-MLV-MEF	1.220	0.0427	28.50	3.61e-178
H-1	1.230	0.1130	10.90	1.04e-27
I-215	0.722	0.1960	3.68	2.35e-04
J-X	1.020	0.2080	4.90	9.70e-07

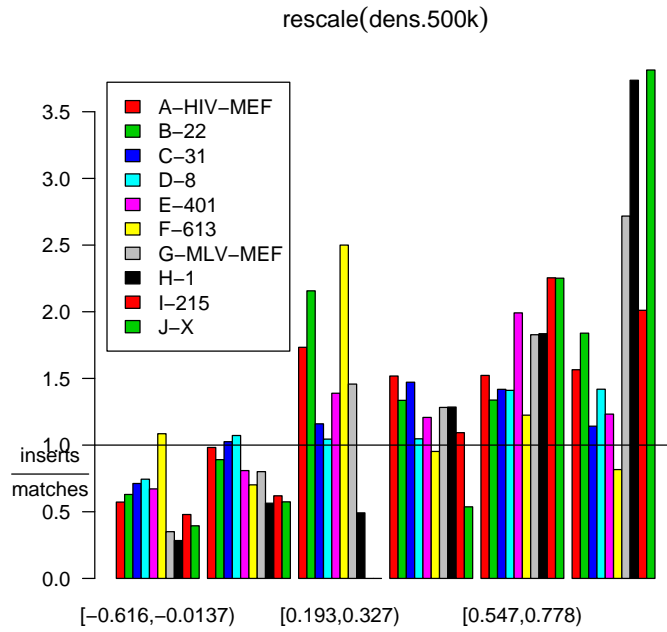
Here the effect of density of CpG islands is studied:



	coef	se	z	p
A-HIV-MEF	0.840	0.0481	17.500	2.06e-68
B-22	0.656	0.0791	8.290	1.11e-16
C-31	0.431	0.0896	4.810	1.49e-06
D-8	0.453	0.1030	4.410	1.03e-05
E-401	0.565	0.1490	3.790	1.48e-04
F-613	-0.138	0.3730	-0.371	7.11e-01
G-MLV-MEF	1.380	0.0372	37.100	5.83e-302
H-1	1.840	0.1170	15.800	3.23e-56
I-215	1.370	0.1730	7.940	2.04e-15
J-X	1.590	0.2020	7.880	3.22e-15

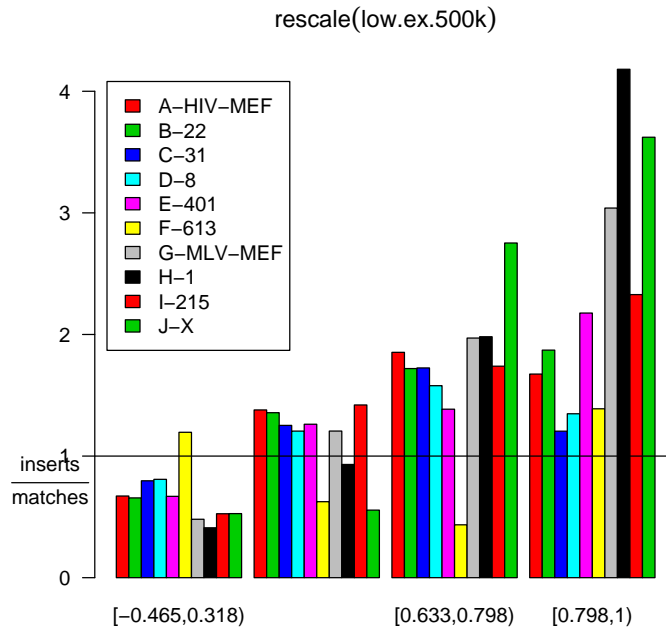
4.5 500 kilobase Window

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.



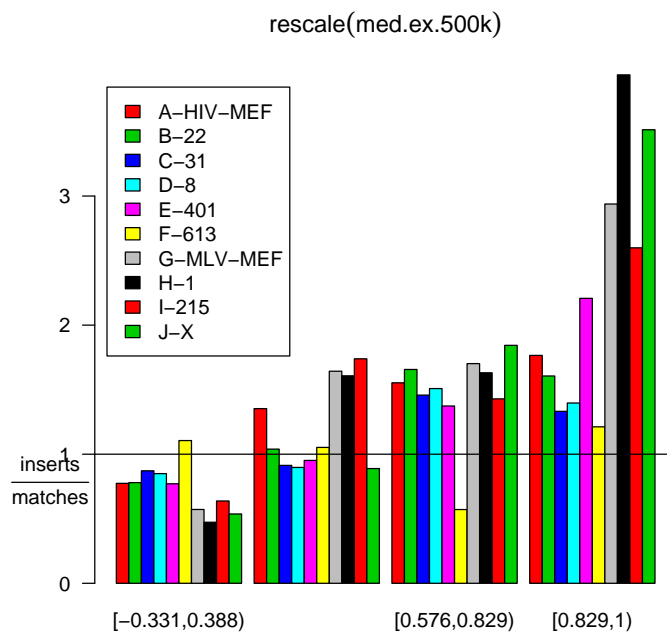
	coef	se	z	p
A-HIV-MEF	0.801	0.0484	16.600	1.24e-61
B-22	0.724	0.0791	9.160	5.37e-20
C-31	0.534	0.0896	5.970	2.44e-09
D-8	0.425	0.1040	4.100	4.10e-05
E-401	0.710	0.1500	4.740	2.11e-06
F-613	0.181	0.3630	0.498	6.19e-01
G-MLV-MEF	1.350	0.0367	36.700	5.83e-295
H-1	1.730	0.1120	15.400	1.83e-53
I-215	1.170	0.1660	7.050	1.80e-12
J-X	1.480	0.1960	7.580	3.40e-14

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



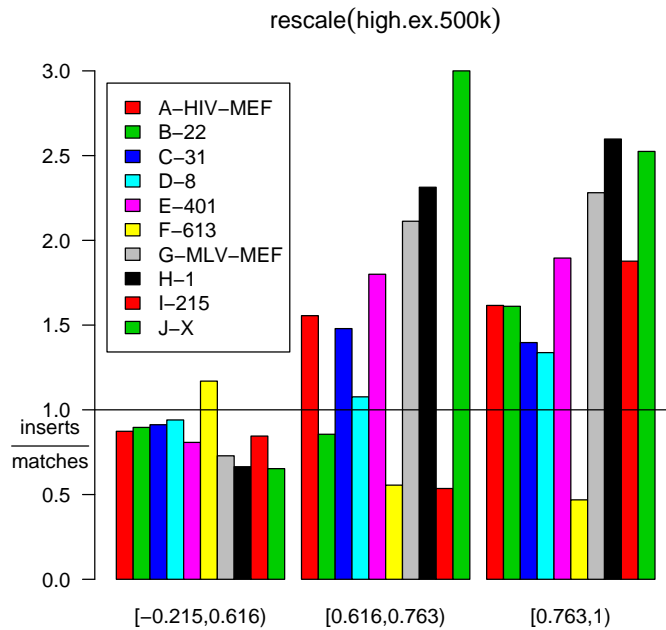
	coef	se	z	p
A-HIV-MEF	0.904	0.0489	18.500	3.23e-76
B-22	0.877	0.0816	10.800	5.92e-27
C-31	0.573	0.0897	6.400	1.60e-10
D-8	0.530	0.1030	5.120	3.04e-07
E-401	0.723	0.1530	4.720	2.32e-06
F-613	-0.216	0.3650	-0.591	5.55e-01
G-MLV-MEF	1.450	0.0387	37.300	3.74e-305
H-1	1.720	0.1190	14.400	4.68e-47
I-215	1.090	0.1670	6.510	7.70e-11
J-X	1.550	0.2120	7.330	2.26e-13

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



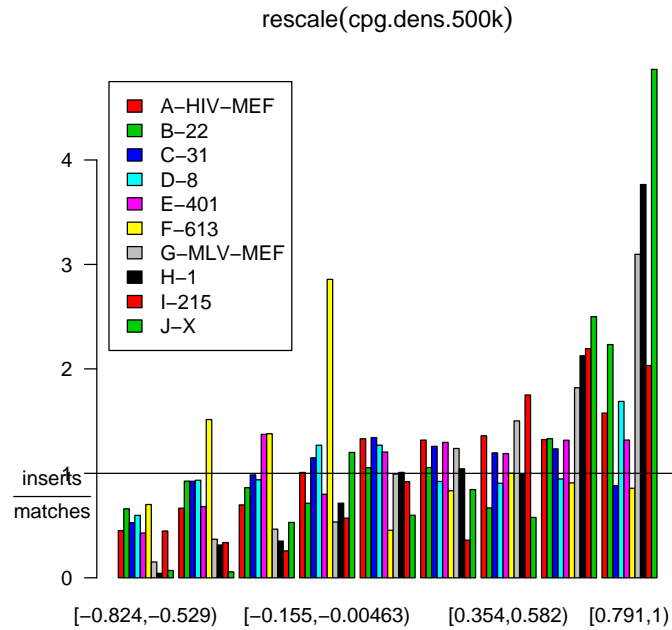
	coef	se	z	p
A-HIV-MEF	0.738	0.0487	15.200	5.43e-52
B-22	0.656	0.0800	8.200	2.39e-16
C-31	0.458	0.0927	4.950	7.55e-07
D-8	0.505	0.1060	4.790	1.69e-06
E-401	0.725	0.1510	4.810	1.53e-06
F-613	-0.165	0.3780	-0.436	6.63e-01
G-MLV-MEF	1.300	0.0357	36.300	1.13e-288
H-1	1.610	0.1030	15.600	1.13e-54
I-215	1.010	0.1570	6.450	1.14e-10
J-X	1.480	0.1900	7.790	6.75e-15

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



	coef	se	z	p
A-HIV-MEF	0.623	0.0542	11.50	1.30e-30
B-22	0.530	0.0894	5.93	3.05e-09
C-31	0.446	0.1040	4.30	1.73e-05
D-8	0.397	0.1200	3.31	9.42e-04
E-401	0.846	0.1610	5.26	1.46e-07
F-613	-0.659	0.5000	-1.32	1.87e-01
G-MLV-MEF	1.150	0.0371	31.00	2.66e-211
H-1	1.310	0.0990	13.20	6.16e-40
I-215	0.719	0.1670	4.30	1.73e-05
J-X	1.390	0.1860	7.49	6.78e-14

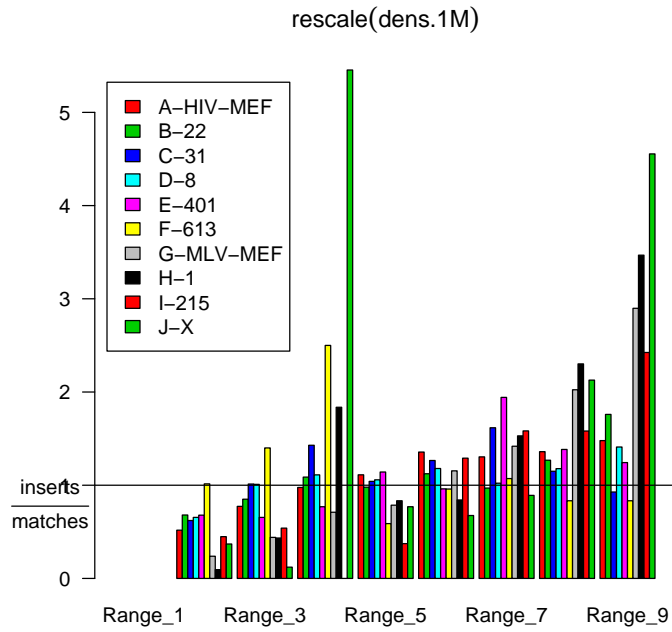
Here the effect of density of CpG islands is studied:



	coef	se	z	p
A-HIV-MEF	0.706	0.0479	14.700	3.49e-49
B-22	0.523	0.0790	6.630	3.44e-11
C-31	0.259	0.0893	2.900	3.76e-03
D-8	0.306	0.1030	2.970	2.94e-03
E-401	0.390	0.1490	2.610	9.01e-03
F-613	-0.175	0.3720	-0.472	6.37e-01
G-MLV-MEF	1.520	0.0393	38.800	0.00e+00
H-1	1.570	0.1160	13.500	1.04e-41
I-215	1.200	0.1740	6.850	7.30e-12
J-X	1.600	0.2130	7.500	6.45e-14

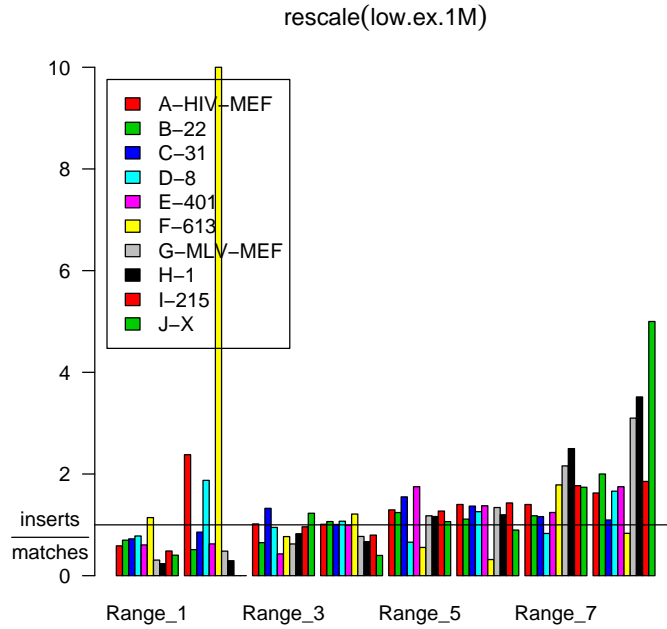
4.6 1 megabase Window

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.



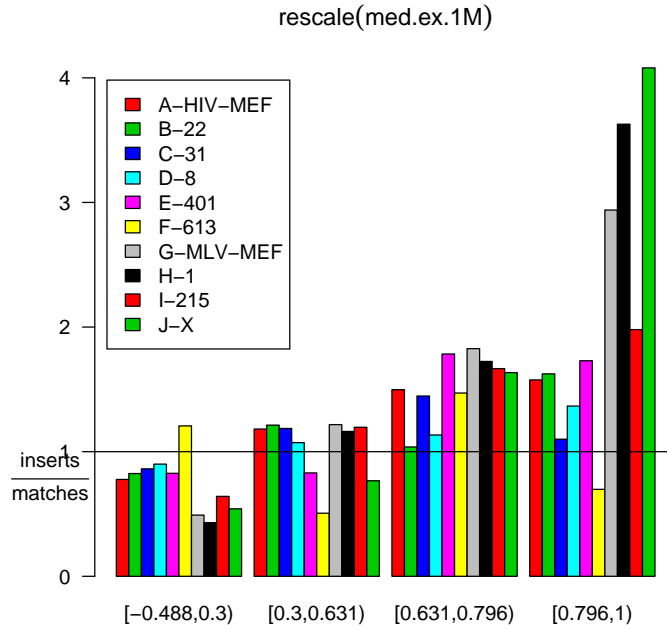
	coef	se	z	p
A-HIV-MEF	0.671	0.0481	13.900	3.86e-44
B-22	0.445	0.0788	5.650	1.64e-08
C-31	0.393	0.0895	4.390	1.12e-05
D-8	0.329	0.1030	3.190	1.40e-03
E-401	0.456	0.1510	3.010	2.60e-03
F-613	-0.119	0.3610	-0.329	7.42e-01
G-MLV-MEF	1.450	0.0396	36.600	3.76e-293
H-1	1.550	0.1170	13.200	4.81e-40
I-215	1.240	0.1820	6.830	8.78e-12
J-X	1.290	0.2060	6.260	3.87e-10

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



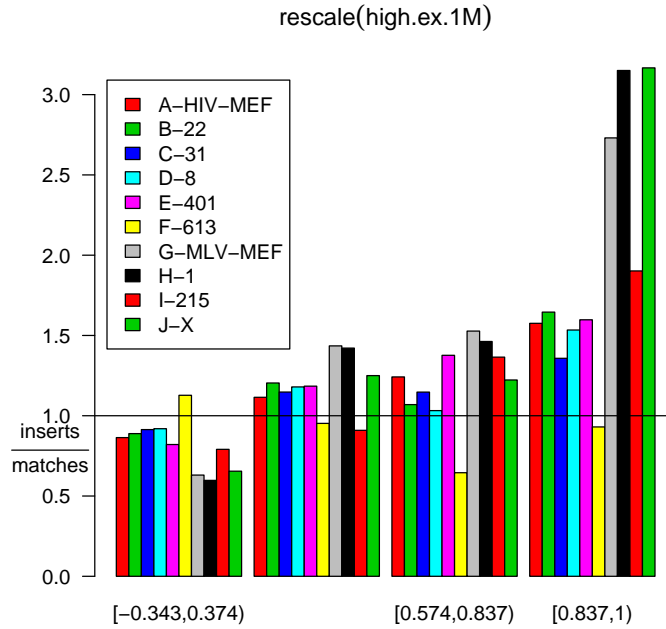
	coef	se	z	p
A-HIV-MEF	0.667	0.0476	14.00	1.61e-44
B-22	0.528	0.0787	6.72	1.85e-11
C-31	0.400	0.0893	4.48	7.45e-06
D-8	0.308	0.1030	2.98	2.87e-03
E-401	0.700	0.1520	4.61	4.05e-06
F-613	-0.214	0.3760	-0.57	5.68e-01
G-MLV-MEF	1.410	0.0374	37.70	0.00e+00
H-1	1.580	0.1100	14.30	3.67e-46
I-215	1.020	0.1660	6.17	6.63e-10
J-X	1.410	0.1980	7.12	1.06e-12

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



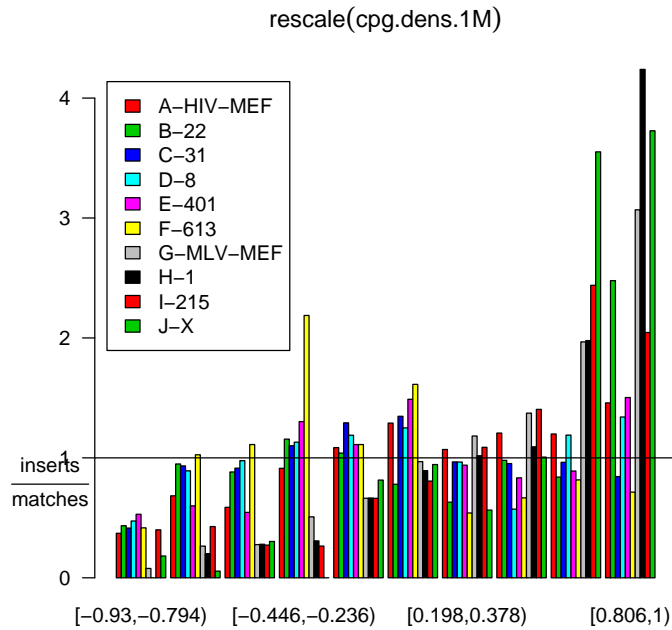
	coef	se	z	p
A-HIV-MEF	0.592	0.0476	12.40	1.52e-35
B-22	0.392	0.0789	4.98	6.51e-07
C-31	0.385	0.0894	4.31	1.65e-05
D-8	0.270	0.1030	2.62	8.70e-03
E-401	0.419	0.1500	2.79	5.32e-03
F-613	-0.498	0.3690	-1.35	1.78e-01
G-MLV-MEF	1.350	0.0386	35.00	4.97e-269
H-1	1.600	0.1190	13.50	1.91e-41
I-215	1.050	0.1720	6.14	8.30e-10
J-X	1.250	0.2040	6.13	9.03e-10

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



	coef	se	z	p
A-HIV-MEF	0.454	0.0486	9.340	9.85e-21
B-22	0.357	0.0809	4.410	1.04e-05
C-31	0.276	0.0934	2.960	3.09e-03
D-8	0.283	0.1060	2.650	7.93e-03
E-401	0.514	0.1510	3.410	6.38e-04
F-613	-0.384	0.3930	-0.977	3.28e-01
G-MLV-MEF	1.120	0.0353	31.800	1.35e-221
H-1	1.150	0.0977	11.800	5.18e-32
I-215	0.557	0.1560	3.580	3.49e-04
J-X	1.140	0.1850	6.160	7.07e-10

Here the effect of density of CpG islands is studied:

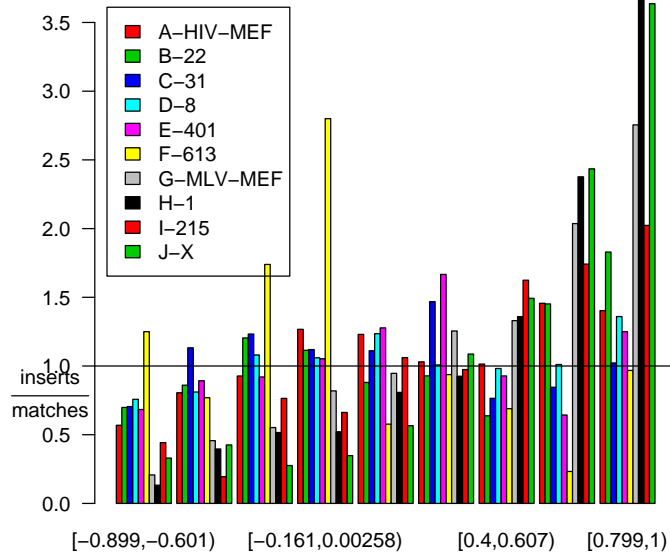


	coef	se	z	p
A-HIV-MEF	0.4490	0.0473	9.490	2.42e-21
B-22	0.1460	0.0783	1.860	6.28e-02
C-31	-0.0362	0.0893	-0.405	6.85e-01
D-8	0.1640	0.1030	1.600	1.10e-01
E-401	0.1430	0.1490	0.959	3.37e-01
F-613	-0.3820	0.3760	-1.020	3.09e-01
G-MLV-MEF	1.4300	0.0395	36.200	2.96e-287
H-1	1.7400	0.1260	13.800	3.00e-43
I-215	1.2300	0.1820	6.740	1.55e-11
J-X	1.6900	0.2310	7.340	2.13e-13

4.7 2 megabase Window

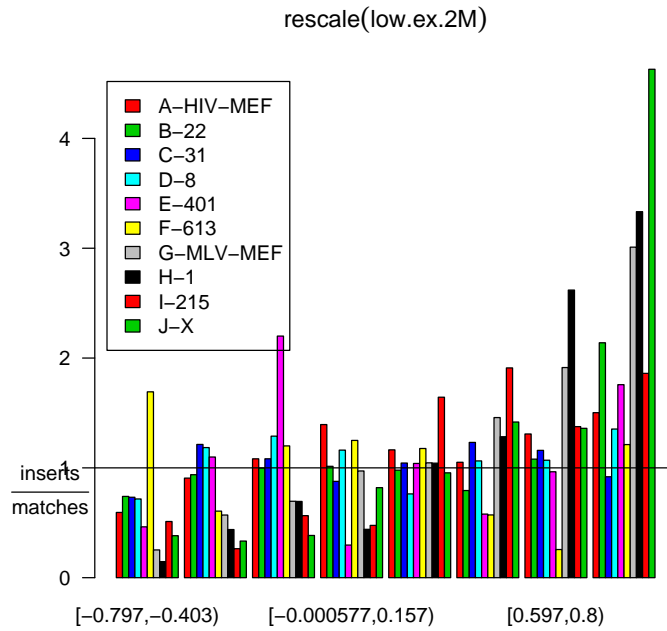
In the barplot that follows we examine the association of insertion sites with expression density in a 2 megabase window surrounding each locus. First, we count just the number of genes represented on the chip.

rescale(dens.2M)



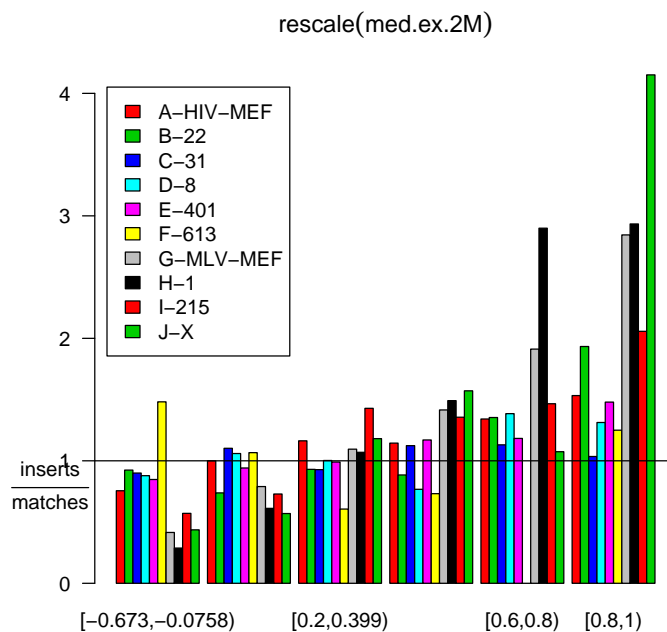
	coef	se	z	p
A-HIV-MEF	0.341	0.0468	7.280	3.45e-13
B-22	0.277	0.0783	3.540	4.02e-04
C-31	0.070	0.0899	0.779	4.36e-01
D-8	0.188	0.1030	1.830	6.70e-02
E-401	0.410	0.1510	2.720	6.57e-03
F-613	-0.857	0.3950	-2.170	2.99e-02
G-MLV-MEF	1.260	0.0375	33.700	7.30e-249
H-1	1.610	0.1180	13.600	2.14e-42
I-215	1.100	0.1740	6.330	2.44e-10
J-X	1.700	0.2260	7.530	5.15e-14

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



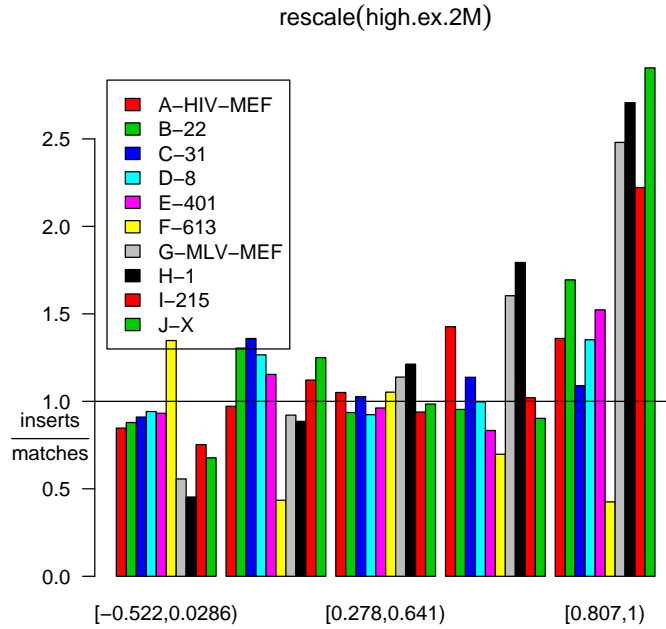
	coef	se	z	p
A-HIV-MEF	0.4730	0.0474	9.970	2.03e-23
B-22	0.3140	0.0787	3.990	6.59e-05
C-31	0.1130	0.0889	1.270	2.04e-01
D-8	0.0665	0.1030	0.646	5.18e-01
E-401	-0.0406	0.1500	-0.272	7.86e-01
F-613	-0.2970	0.3700	-0.802	4.22e-01
G-MLV-MEF	1.3100	0.0385	34.100	3.66e-255
H-1	1.6200	0.1220	13.300	3.33e-40
I-215	1.2800	0.1860	6.900	5.29e-12
J-X	1.5700	0.2220	7.090	1.39e-12

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



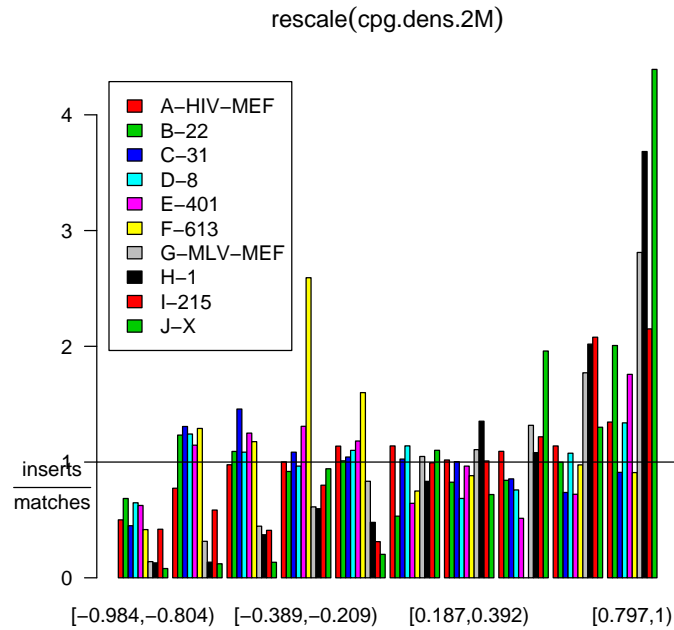
	coef	se	z	p
A-HIV-MEF	0.466	0.0471	9.89	4.71e-23
B-22	0.280	0.0783	3.58	3.46e-04
C-31	0.133	0.0894	1.49	1.36e-01
D-8	0.131	0.1030	1.27	2.04e-01
E-401	0.262	0.1490	1.76	7.91e-02
F-613	-0.744	0.3830	-1.94	5.19e-02
G-MLV-MEF	1.210	0.0366	33.20	4.30e-241
H-1	1.620	0.1170	13.90	1.12e-43
I-215	0.755	0.1620	4.65	3.35e-06
J-X	1.510	0.2090	7.21	5.62e-13

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



	coef	se	z	p
A-HIV-MEF	0.3270	0.0470	6.960	3.52e-12
B-22	0.2260	0.0785	2.880	4.04e-03
C-31	0.0980	0.0892	1.100	2.72e-01
D-8	0.0565	0.1030	0.547	5.85e-01
E-401	0.0785	0.1490	0.528	5.98e-01
F-613	-0.4700	0.3690	-1.270	2.03e-01
G-MLV-MEF	0.9780	0.0356	27.400	1.41e-165
H-1	1.2300	0.1070	11.500	1.16e-30
I-215	0.4450	0.1580	2.820	4.87e-03
J-X	0.5690	0.1820	3.120	1.80e-03

Here the effect of density of CpG islands is studied:

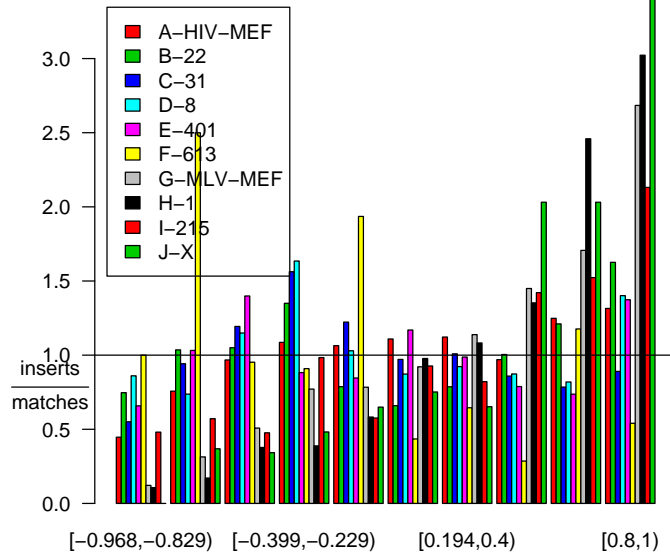


	coef	se	z	p
A-HIV-MEF	0.2500	0.0469	5.330	9.63e-08
B-22	0.0750	0.0784	0.957	3.39e-01
C-31	-0.1460	0.0903	-1.610	1.07e-01
D-8	0.0357	0.1030	0.347	7.29e-01
E-401	-0.1690	0.1510	-1.120	2.64e-01
F-613	-0.7690	0.3800	-2.020	4.33e-02
G-MLV-MEF	1.1400	0.0370	30.900	1.71e-209
H-1	1.6300	0.1210	13.500	1.80e-41
I-215	1.1500	0.1760	6.530	6.40e-11
J-X	1.7100	0.2270	7.510	5.80e-14

4.8 4 megabase Window

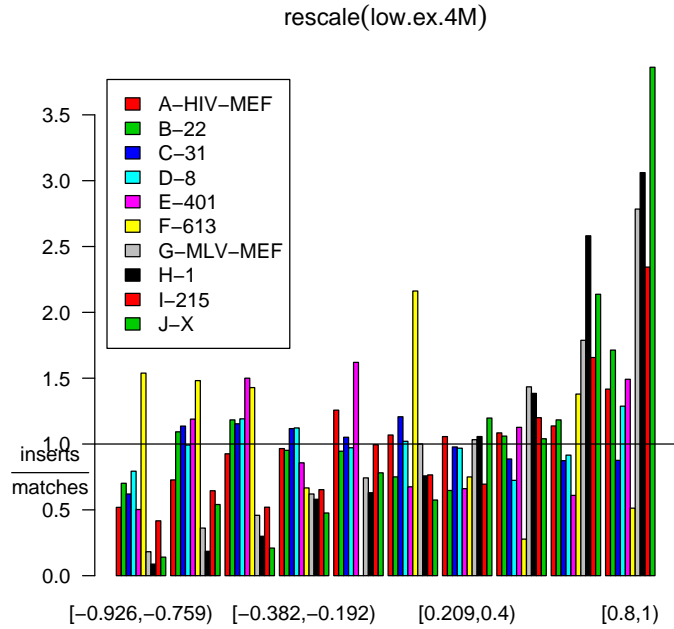
In the barplot that follows we examine the association of insertion sites with expression density in a 4 megabase window surrounding each locus. First, we count just the number of genes represented on the chip.

rescale(dens.4M)



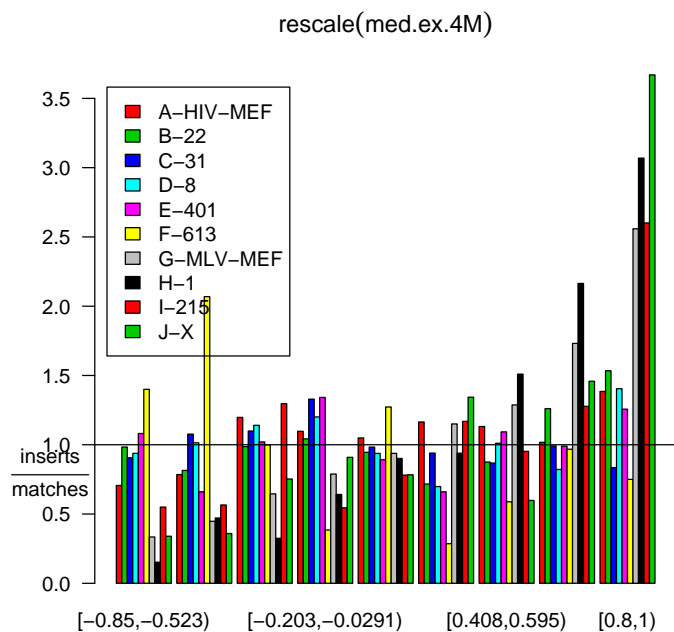
	coef	se	z	p
A-HIV-MEF	0.2800	0.0473	5.930	3.01e-09
B-22	0.0928	0.0781	1.190	2.35e-01
C-31	-0.1730	0.0901	-1.910	5.56e-02
D-8	-0.0702	0.1030	-0.682	4.95e-01
E-401	0.1410	0.1500	0.941	3.47e-01
F-613	-0.9530	0.4060	-2.350	1.88e-02
G-MLV-MEF	1.1300	0.0373	30.200	8.06e-201
H-1	1.5600	0.1200	13.000	8.45e-39
I-215	0.9500	0.1720	5.510	3.50e-08
J-X	1.5800	0.2250	7.010	2.39e-12

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



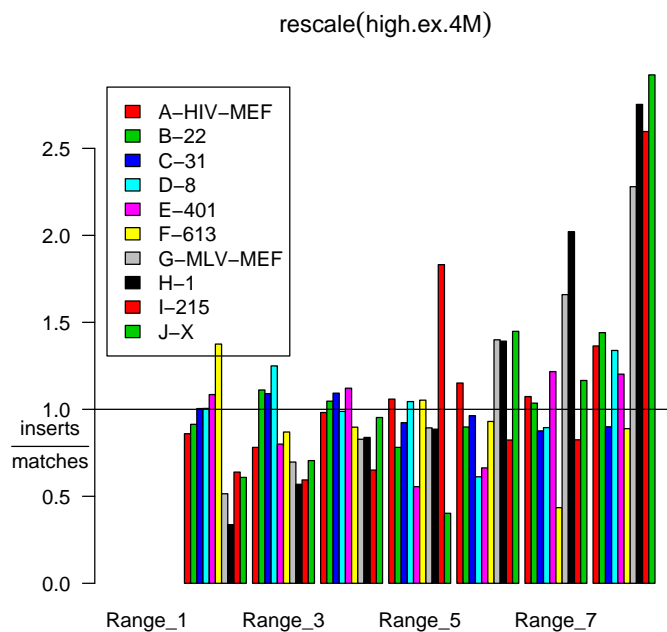
	coef	se	z	p
A-HIV-MEF	0.2740	0.0471	5.830	5.67e-09
B-22	0.0611	0.0784	0.779	4.36e-01
C-31	-0.0835	0.0895	-0.932	3.51e-01
D-8	-0.0444	0.1030	-0.431	6.67e-01
E-401	-0.1480	0.1510	-0.982	3.26e-01
F-613	0.0736	0.3680	0.200	8.41e-01
G-MLV-MEF	1.1400	0.0370	30.800	1.36e-208
H-1	1.5500	0.1180	13.100	3.23e-39
I-215	0.6790	0.1620	4.180	2.89e-05
J-X	1.4500	0.2170	6.700	2.04e-11

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



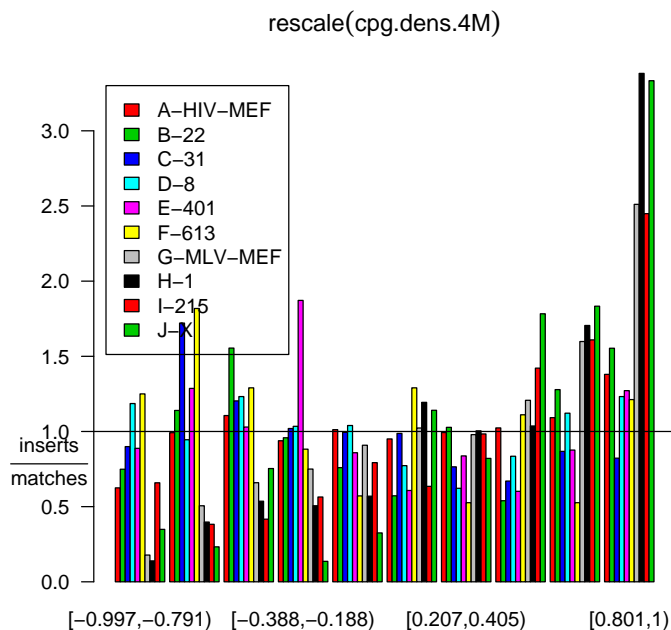
	coef	se	z	p
A-HIV-MEF	0.2760	0.0471	5.870	4.35e-09
B-22	0.1020	0.0783	1.300	1.95e-01
C-31	-0.1680	0.0901	-1.870	6.19e-02
D-8	-0.0413	0.1030	-0.402	6.88e-01
E-401	-0.1480	0.1500	-0.988	3.23e-01
F-613	-0.3780	0.3680	-1.030	3.05e-01
G-MLV-MEF	1.0700	0.0369	29.000	8.82e-185
H-1	1.5100	0.1170	12.900	4.93e-38
I-215	0.8640	0.1680	5.150	2.54e-07
J-X	1.0500	0.1950	5.380	7.49e-08

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



	coef	se	z	p
A-HIV-MEF	0.2680	0.0471	5.680	1.33e-08
B-22	0.0835	0.0785	1.060	2.87e-01
C-31	-0.0137	0.0892	-0.154	8.78e-01
D-8	-0.0434	0.1030	-0.421	6.74e-01
E-401	-0.3420	0.1520	-2.250	2.46e-02
F-613	-0.3950	0.3700	-1.070	2.86e-01
G-MLV-MEF	0.8910	0.0357	25.000	1.75e-137
H-1	1.1900	0.1090	11.000	6.55e-28
I-215	0.7810	0.1640	4.760	1.98e-06
J-X	0.5680	0.1830	3.100	1.92e-03

Here the effect of density of CpG islands is studied:

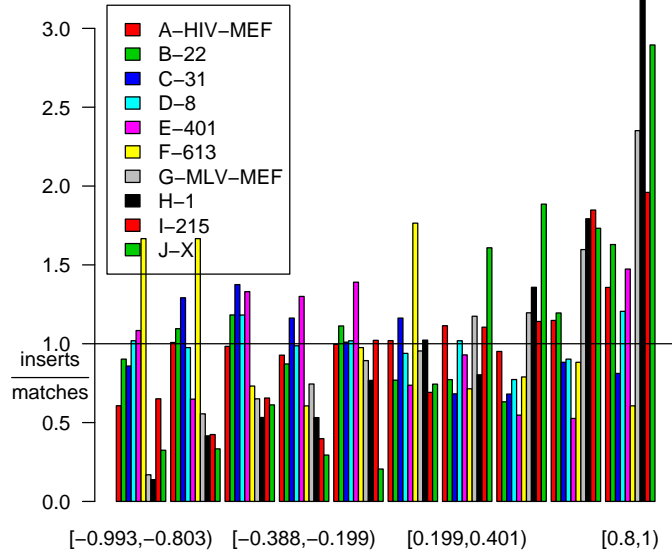


	coef	se	z	p
A-HIV-MEF	0.1400	0.0470	2.980	2.93e-03
B-22	-0.0269	0.0786	-0.342	7.32e-01
C-31	-0.3710	0.0922	-4.030	5.63e-05
D-8	-0.1680	0.1030	-1.620	1.04e-01
E-401	-0.3840	0.1520	-2.520	1.18e-02
F-613	-0.1460	0.3670	-0.399	6.90e-01
G-MLV-MEF	0.8630	0.0355	24.300	8.83e-131
H-1	1.2800	0.1110	11.500	1.56e-30
I-215	0.9900	0.1720	5.760	8.31e-09
J-X	1.4000	0.2120	6.610	3.74e-11

4.9 8 megabase Window

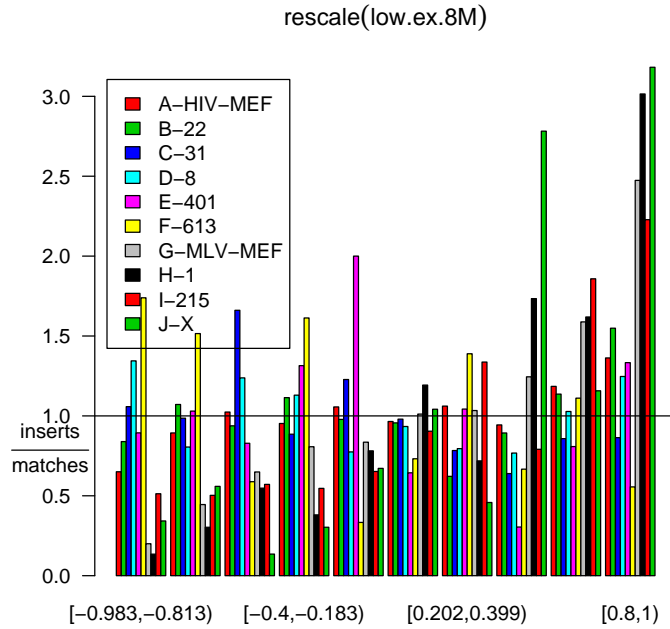
In the barplot that follows we examine the association of insertion sites with expression density in a 8 megabase window surrounding each locus. First, we count just the number of genes represented on the chip.

rescale(dens.8M)



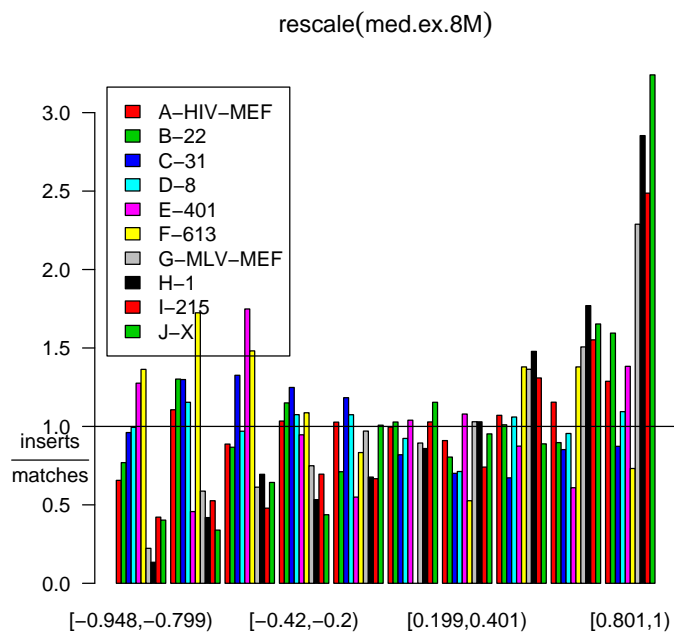
	coef	se	z	p
A-HIV-MEF	0.2060	0.0469	4.390	1.14e-05
B-22	-0.0600	0.0786	-0.764	4.45e-01
C-31	-0.2920	0.0909	-3.210	1.31e-03
D-8	-0.0736	0.1030	-0.712	4.76e-01
E-401	-0.3200	0.1520	-2.110	3.50e-02
F-613	-0.0607	0.3660	-0.166	8.68e-01
G-MLV-MEF	0.8650	0.0357	24.200	1.04e-129
H-1	1.1600	0.1090	10.600	2.80e-26
I-215	0.7880	0.1670	4.720	2.40e-06
J-X	1.5900	0.2250	7.080	1.42e-12

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



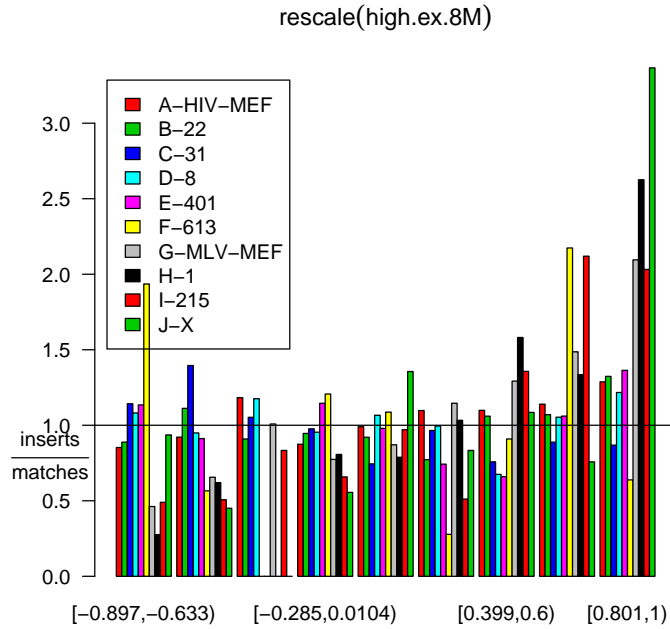
	coef	se	z	p
A-HIV-MEF	0.1650	0.0470	3.510	4.52e-04
B-22	0.0326	0.0786	0.415	6.78e-01
C-31	-0.3350	0.0916	-3.660	2.57e-04
D-8	-0.1280	0.1040	-1.240	2.16e-01
E-401	-0.3610	0.1520	-2.370	1.76e-02
F-613	-0.2470	0.3610	-0.685	4.93e-01
G-MLV-MEF	0.8830	0.0356	24.800	1.47e-135
H-1	1.2400	0.1110	11.200	4.31e-29
I-215	1.0100	0.1720	5.890	3.88e-09
J-X	1.3400	0.2090	6.400	1.57e-10

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



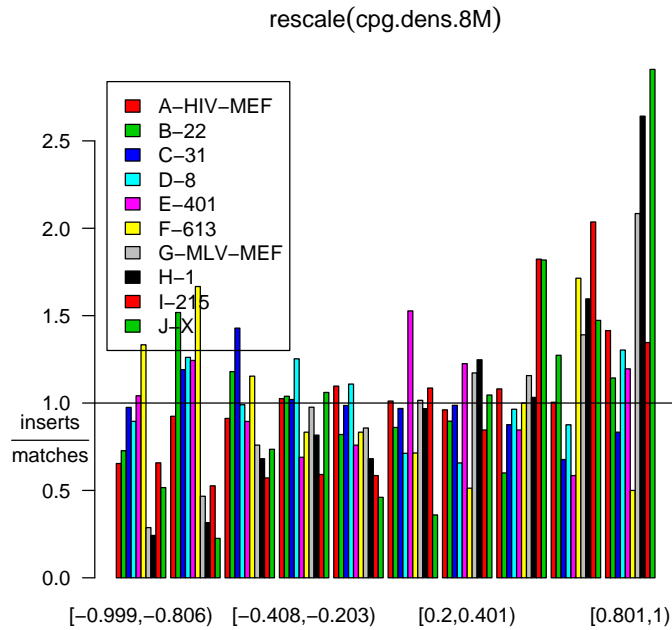
	coef	se	z	p
A-HIV-MEF	0.14200	0.0469	3.0200	2.53e-03
B-22	0.08980	0.0783	1.1500	2.51e-01
C-31	-0.42700	0.0920	-4.6400	3.42e-06
D-8	-0.09980	0.1030	-0.9700	3.32e-01
E-401	-0.00204	0.1500	-0.0136	9.89e-01
F-613	-0.47900	0.3680	-1.3000	1.94e-01
G-MLV-MEF	0.80000	0.0354	22.6000	7.90e-113
H-1	1.15000	0.1090	10.6000	4.04e-26
I-215	0.91800	0.1720	5.3300	9.80e-08
J-X	0.98000	0.1950	5.0200	5.21e-07

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



	coef	se	z	p
A-HIV-MEF	0.2210	0.0471	4.690	2.79e-06
B-22	0.0714	0.0788	0.905	3.65e-01
C-31	-0.2630	0.0916	-2.870	4.05e-03
D-8	0.0376	0.1030	0.365	7.15e-01
E-401	-0.1790	0.1510	-1.180	2.36e-01
F-613	-0.1200	0.3640	-0.330	7.42e-01
G-MLV-MEF	0.7420	0.0348	21.300	9.04e-101
H-1	0.9300	0.1020	9.150	5.63e-20
I-215	0.6580	0.1610	4.090	4.33e-05
J-X	0.7020	0.1830	3.840	1.24e-04

Here the effect of density of CpG islands is studied:

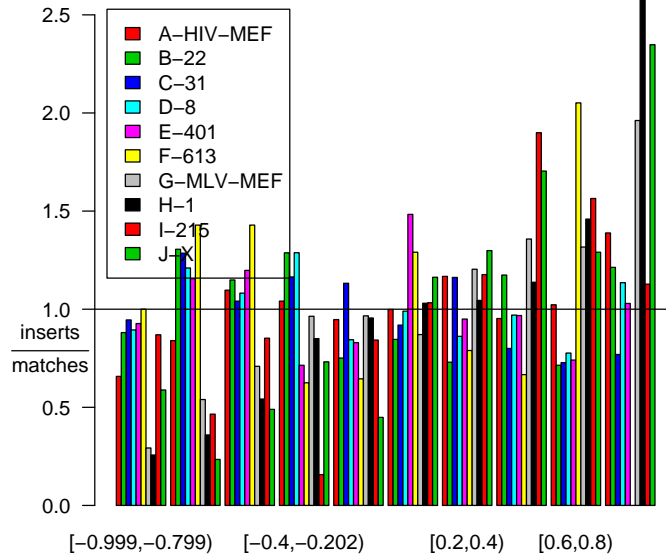


	coef	se	z	p
A-HIV-MEF	0.162	0.0470	3.440	5.83e-04
B-22	-0.113	0.0789	-1.440	1.51e-01
C-31	-0.258	0.0905	-2.850	4.34e-03
D-8	-0.186	0.1030	-1.800	7.13e-02
E-401	0.037	0.1500	0.246	8.06e-01
F-613	-0.245	0.3680	-0.665	5.06e-01
G-MLV-MEF	0.701	0.0346	20.200	5.20e-91
H-1	0.940	0.1040	9.020	1.94e-19
I-215	0.902	0.1690	5.320	1.02e-07
J-X	0.896	0.1920	4.680	2.93e-06

4.10 16 megabase Window

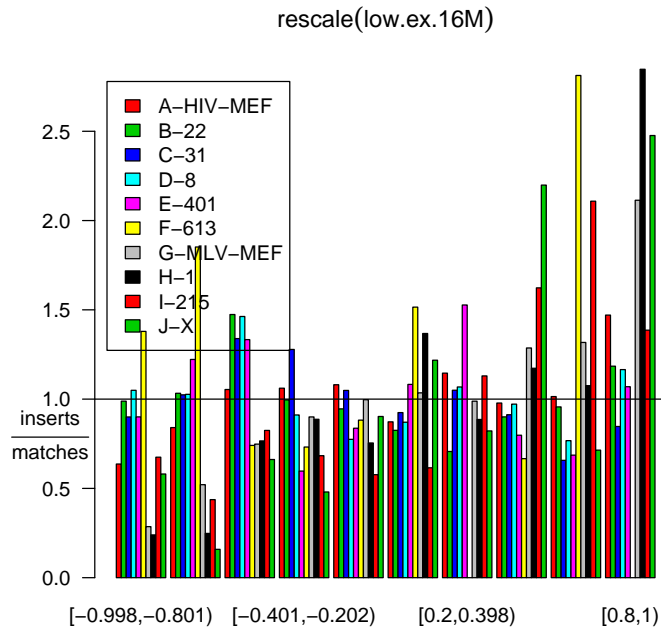
In the barplot that follows we examine the association of insertion sites with expression density in a 16 megabase window surrounding each locus. First, we count just the number of genes represented on the chip.

rescale(dens.16M)



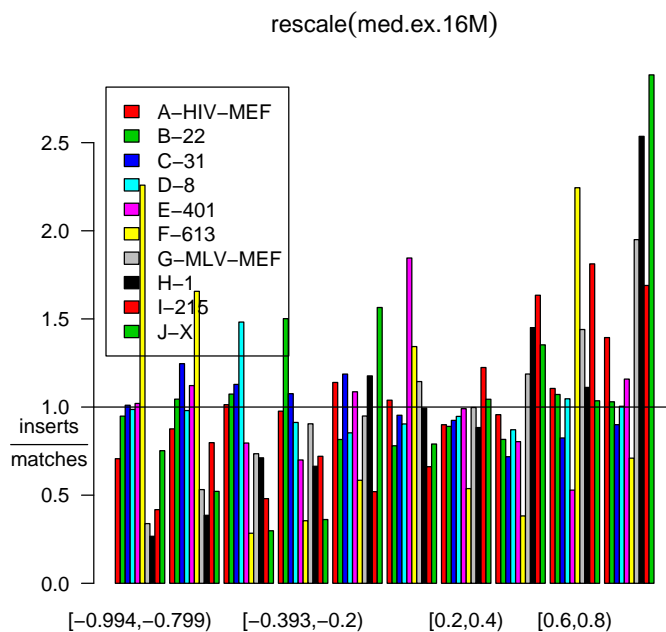
	coef	se	z	p
A-HIV-MEF	0.1880	0.0469	4.020	5.92e-05
B-22	-0.1570	0.0790	-1.990	4.65e-02
C-31	-0.2320	0.0901	-2.580	9.93e-03
D-8	-0.0994	0.1030	-0.968	3.33e-01
E-401	0.0917	0.1510	0.605	5.45e-01
F-613	-0.0470	0.3590	-0.131	8.96e-01
G-MLV-MEF	0.6560	0.0345	19.000	2.21e-80
H-1	0.8950	0.1040	8.600	8.01e-18
I-215	0.7750	0.1690	4.600	4.31e-06
J-X	1.1100	0.2020	5.490	4.07e-08

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



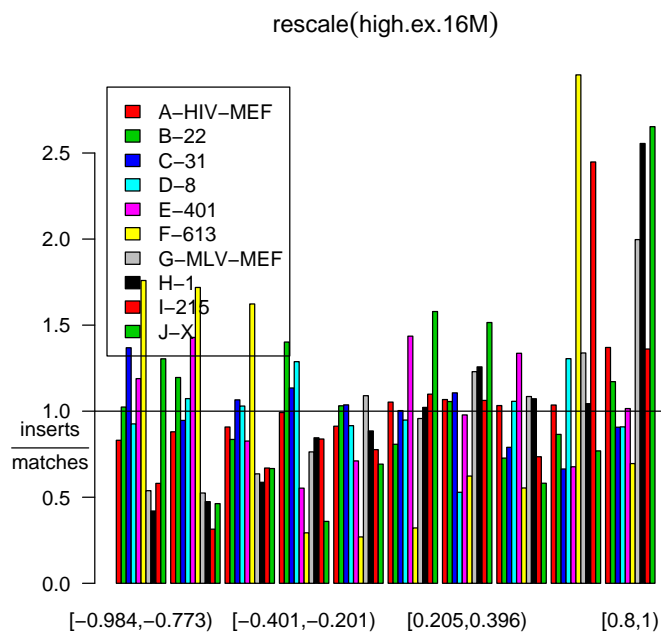
	coef	se	z	p
A-HIV-MEF	0.1660	0.0470	3.540	4.07e-04
B-22	-0.1850	0.0791	-2.340	1.93e-02
C-31	-0.2490	0.0901	-2.760	5.75e-03
D-8	-0.0563	0.1020	-0.549	5.83e-01
E-401	0.0770	0.1510	0.509	6.11e-01
F-613	-0.2650	0.3650	-0.726	4.68e-01
G-MLV-MEF	0.6770	0.0347	19.500	7.07e-85
H-1	0.9000	0.1040	8.680	4.05e-18
I-215	0.7780	0.1670	4.670	3.03e-06
J-X	0.9220	0.1950	4.720	2.36e-06

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



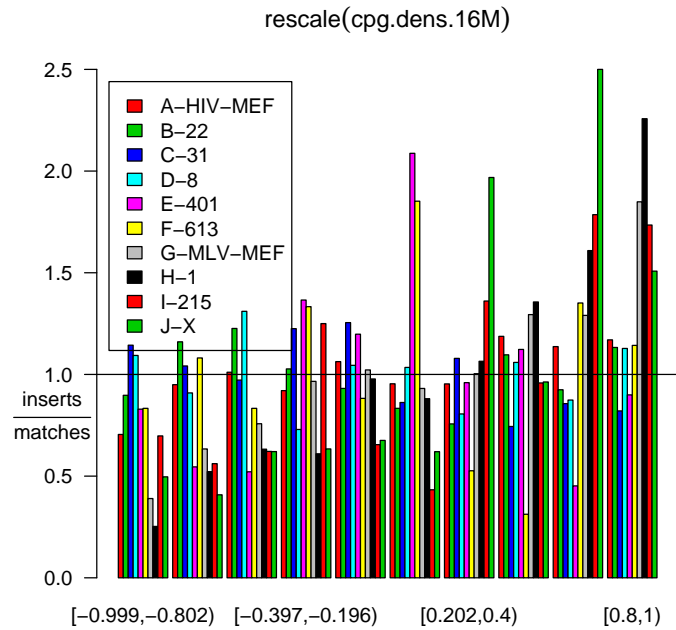
	coef	se	z	p
A-HIV-MEF	0.1320	0.0470	2.810	5.01e-03
B-22	-0.1680	0.0790	-2.130	3.34e-02
C-31	-0.2760	0.0905	-3.040	2.34e-03
D-8	-0.0939	0.1030	-0.914	3.61e-01
E-401	0.1170	0.1500	0.778	4.37e-01
F-613	0.0593	0.3610	0.164	8.70e-01
G-MLV-MEF	0.6600	0.0347	19.000	8.39e-81
H-1	0.7730	0.1010	7.610	2.65e-14
I-215	0.8600	0.1700	5.070	3.92e-07
J-X	0.5300	0.1840	2.890	3.91e-03

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



	coef	se	z	p
A-HIV-MEF	0.2010	0.0469	4.280	1.85e-05
B-22	-0.1620	0.0789	-2.050	4.00e-02
C-31	-0.2190	0.0896	-2.440	1.45e-02
D-8	-0.0996	0.1030	-0.964	3.35e-01
E-401	0.1960	0.1510	1.300	1.94e-01
F-613	-0.1200	0.3640	-0.331	7.41e-01
G-MLV-MEF	0.6140	0.0346	17.700	2.36e-70
H-1	0.7520	0.1020	7.390	1.42e-13
I-215	0.7380	0.1650	4.460	8.11e-06
J-X	0.6090	0.1850	3.290	1.02e-03

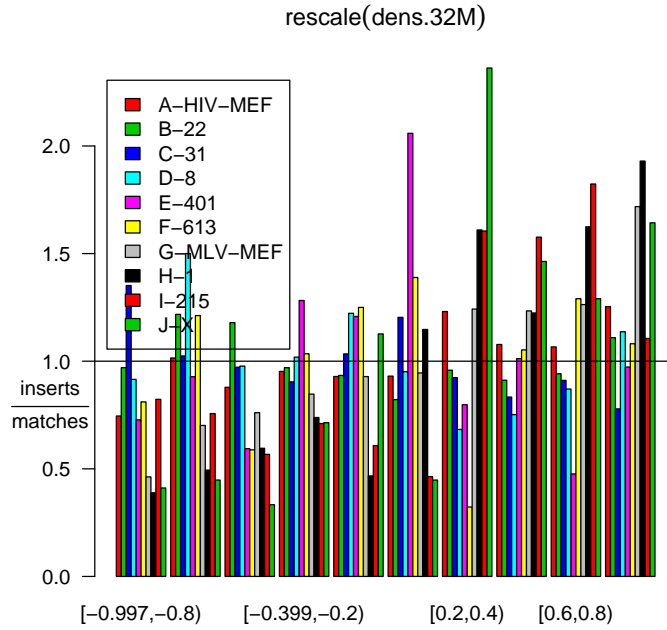
Here the effect of density of CpG islands is studied:



	coef	se	z	p
A-HIV-MEF	0.1540	0.0469	3.2800	1.05e-03
B-22	-0.0871	0.0787	-1.1100	2.69e-01
C-31	-0.2490	0.0904	-2.7600	5.85e-03
D-8	-0.0191	0.1020	-0.1860	8.52e-01
E-401	0.2690	0.1490	1.8000	7.12e-02
F-613	0.0364	0.3660	0.0997	9.21e-01
G-MLV-MEF	0.5180	0.0340	15.2000	2.52e-52
H-1	0.8800	0.1040	8.4800	2.30e-17
I-215	0.4930	0.1610	3.0600	2.19e-03
J-X	0.9740	0.1940	5.0300	5.03e-07

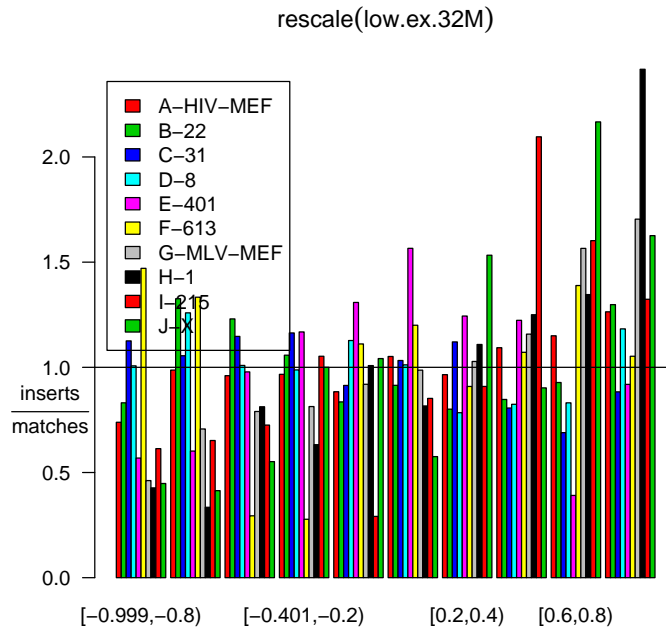
4.11 32 megabase Window

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.



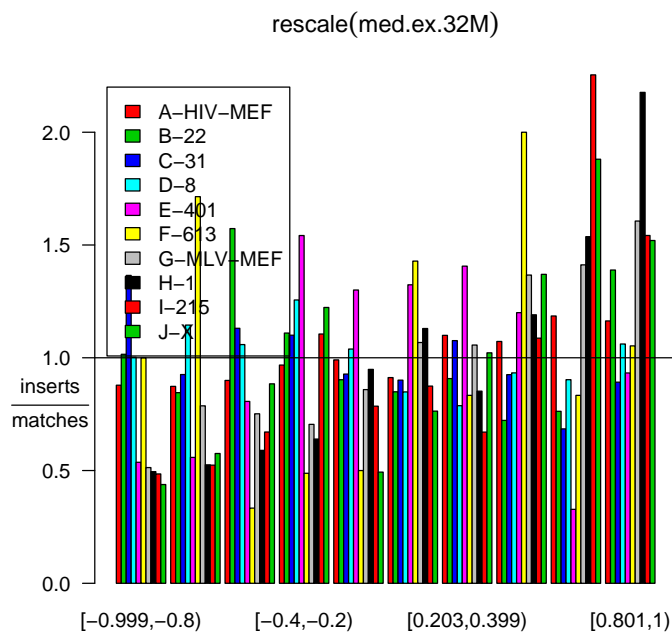
	coef	se	z	p
A-HIV-MEF	0.2040	0.0467	4.380	1.21e-05
B-22	-0.1040	0.0786	-1.330	1.85e-01
C-31	-0.1290	0.0898	-1.440	1.50e-01
D-8	-0.2390	0.1030	-2.320	2.06e-02
E-401	0.1390	0.1500	0.926	3.54e-01
F-613	0.0837	0.3630	0.230	8.18e-01
G-MLV-MEF	0.5470	0.0341	16.100	5.69e-58
H-1	1.0300	0.1070	9.690	3.25e-22
I-215	0.6380	0.1630	3.920	8.80e-05
J-X	0.8570	0.1910	4.490	7.15e-06

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



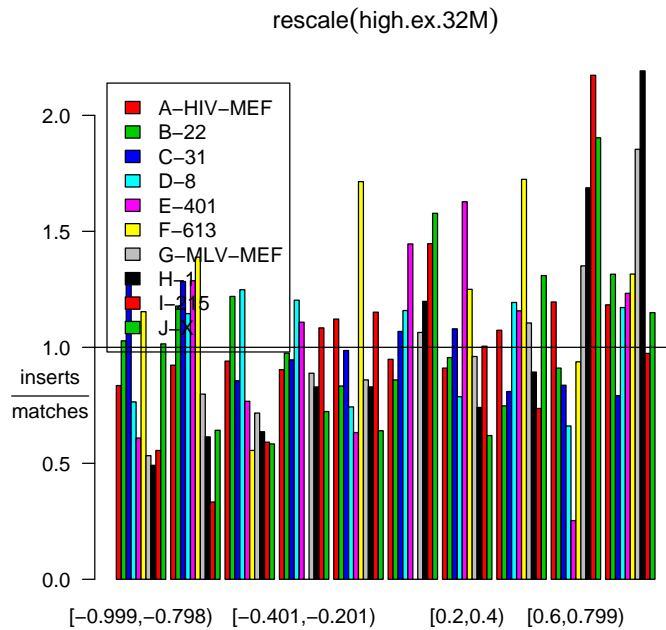
	coef	se	z	p
A-HIV-MEF	0.1980	0.0470	4.210	2.52e-05
B-22	-0.0998	0.0787	-1.270	2.04e-01
C-31	-0.1720	0.0899	-1.910	5.64e-02
D-8	-0.1540	0.1030	-1.500	1.34e-01
E-401	0.1590	0.1490	1.070	2.85e-01
F-613	0.2450	0.3680	0.665	5.06e-01
G-MLV-MEF	0.5530	0.0342	16.200	6.40e-59
H-1	0.7600	0.1010	7.510	6.14e-14
I-215	0.7170	0.1630	4.390	1.15e-05
J-X	0.6410	0.1850	3.470	5.21e-04

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



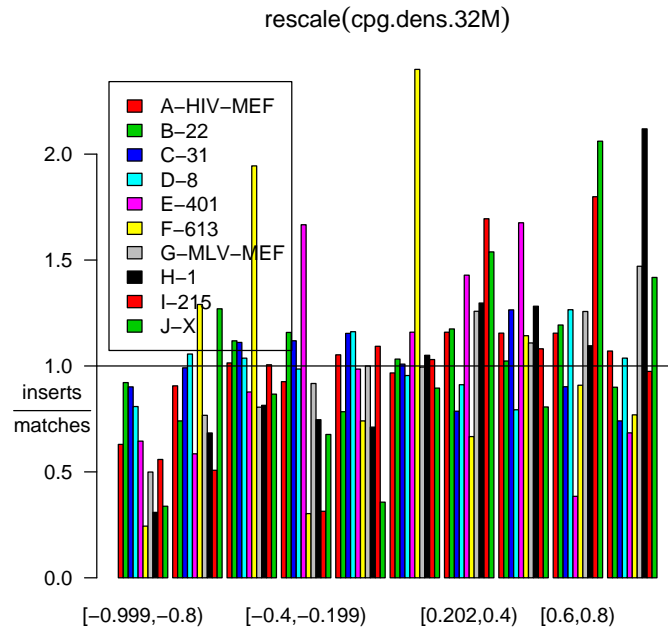
	coef	se	z	p
A-HIV-MEF	0.1620	0.0468	3.470	5.27e-04
B-22	-0.1650	0.0788	-2.090	3.64e-02
C-31	-0.1940	0.0902	-2.150	3.18e-02
D-8	-0.1940	0.1030	-1.880	5.99e-02
E-401	0.0863	0.1480	0.581	5.61e-01
F-613	0.4310	0.3660	1.180	2.39e-01
G-MLV-MEF	0.5850	0.0342	17.100	2.11e-65
H-1	0.7610	0.1020	7.490	6.90e-14
I-215	0.5940	0.1620	3.660	2.56e-04
J-X	0.5710	0.1840	3.110	1.90e-03

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



	coef	se	z	p
A-HIV-MEF	0.1110	0.0468	2.370	1.77e-02
B-22	-0.1070	0.0789	-1.350	1.77e-01
C-31	-0.1550	0.0901	-1.730	8.45e-02
D-8	-0.0239	0.1020	-0.233	8.16e-01
E-401	0.2730	0.1510	1.810	7.03e-02
F-613	0.1560	0.3630	0.429	6.68e-01
G-MLV-MEF	0.5030	0.0341	14.800	2.44e-49
H-1	0.6660	0.0999	6.670	2.58e-11
I-215	0.5190	0.1610	3.230	1.24e-03
J-X	0.5960	0.1830	3.250	1.14e-03

Here the effect of density of CpG islands is studied:

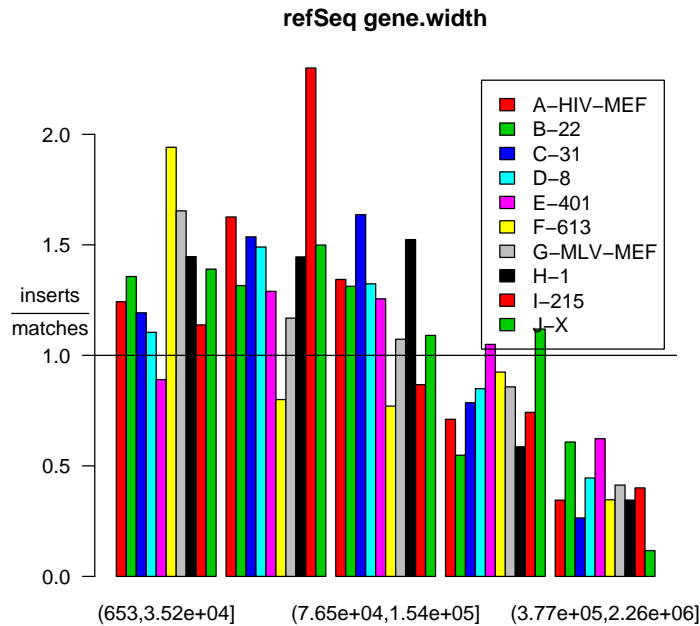


	coef	se	z	p
A-HIV-MEF	0.203000	0.0473	4.29000	1.77e-05
B-22	0.125000	0.0783	1.60000	1.09e-01
C-31	-0.132000	0.0901	-1.47000	1.43e-01
D-8	-0.000966	0.1030	-0.00937	9.93e-01
E-401	0.104000	0.1490	0.70300	4.82e-01
F-613	0.228000	0.3640	0.62700	5.31e-01
G-MLV-MEF	0.421000	0.0338	12.50000	1.40e-35
H-1	0.751000	0.1020	7.38000	1.56e-13
I-215	0.651000	0.1620	4.02000	5.78e-05
J-X	0.650000	0.1840	3.53000	4.18e-04

5 Juxtaposition with Gene Start and End Positions

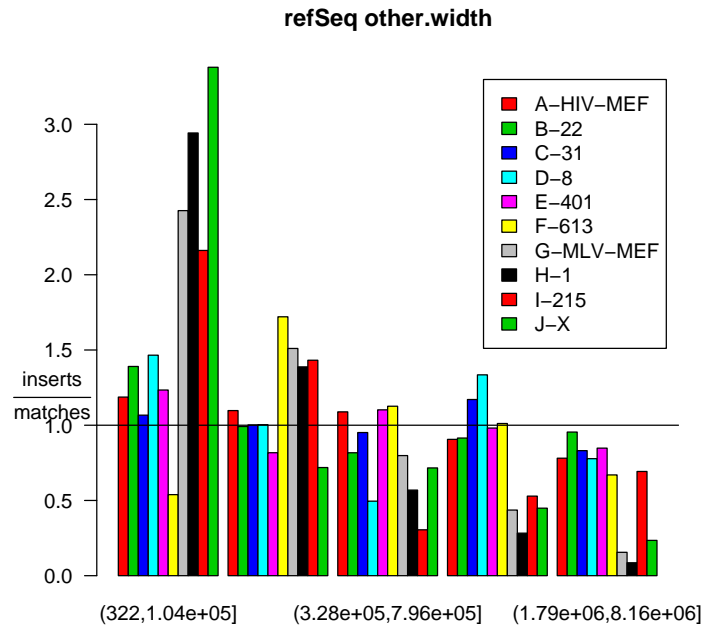
5.1 Refseq Annotations

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 RefSeq 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.



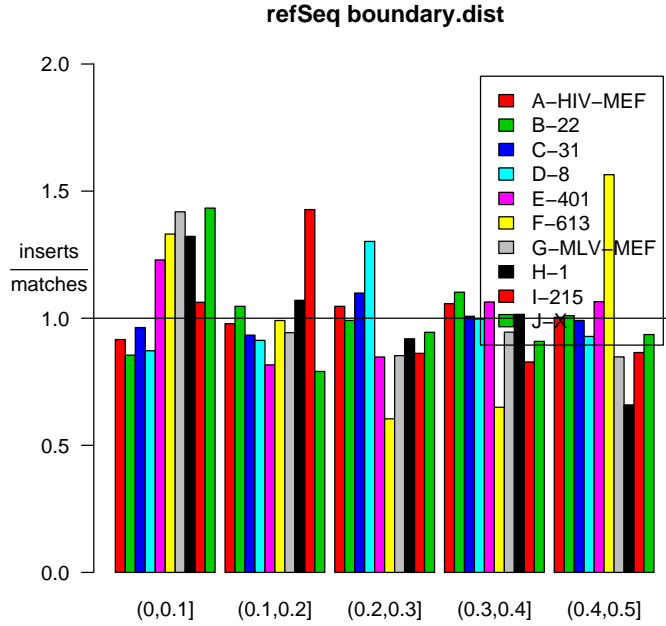
A-HIV-MEF	B-22	C-31	D-8	E-401	F-613	G-MLV-MEF	H-1
2.69e-60	2.45e-10	2.14e-16	4.19e-06	2.62e-01	6.89e-01	2.87e-39	1.72e-12
I-215	J-X						
1.27e-04	7.41e-06						

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



A-HIV-MEF	B-22	C-31	D-8	E-401	F-613	G-MLV-MEF	H-1
2.61e-05	2.02e-02	6.92e-01	3.88e-05	8.01e-01	4.48e-01	0.00e+00	1.76e-54
I-215	J-X						
1.94e-10	4.59e-12						

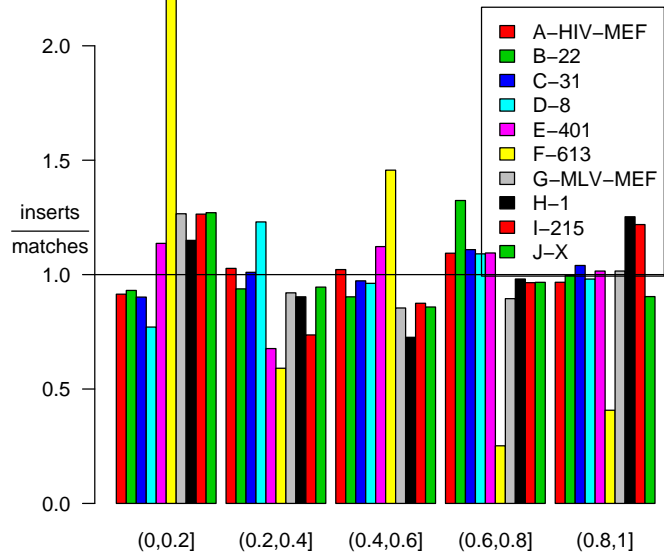
The next plot studies the distance to the nearest boundary between a gene and a non-gene region. The distance is expressed as a fraction of the length of the region. Thus, '0.25' refers to one quarter of the distance from the site to nearest boundary divided by the total width of the region.



A-HIV-MEF	B-22	C-31	D-8	E-401	F-613	G-MLV-MEF	H-1
1.46e-01	2.92e-01	7.97e-01	6.55e-02	3.46e-01	3.31e-01	6.58e-41	1.45e-04
I-215	J-X						
8.42e-02	1.81e-01						

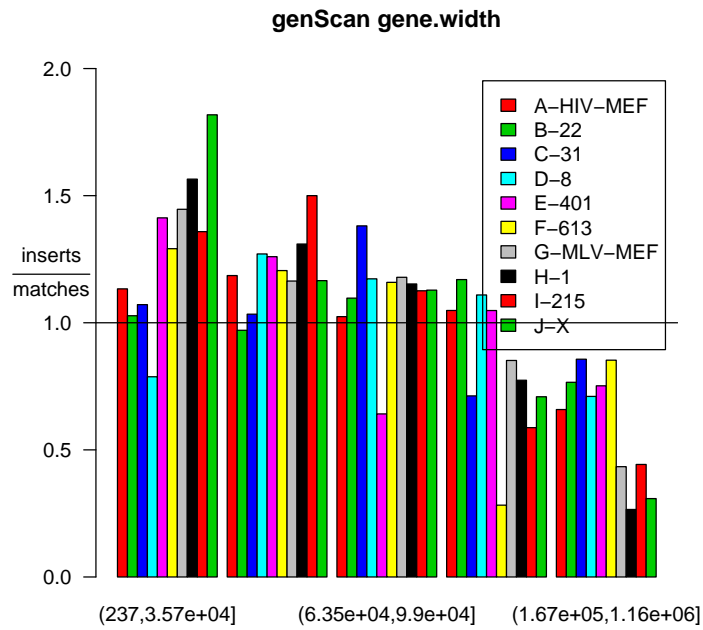
This plot studies the effect of nearness to the beginning of a transcript. For sites in genes, it is the distance to the start of the gene divided by the width of the gene. For other sites it is the distance from the site to the nearer gene if that gene boundary is also a transcription starting point.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

refSeq start.dist



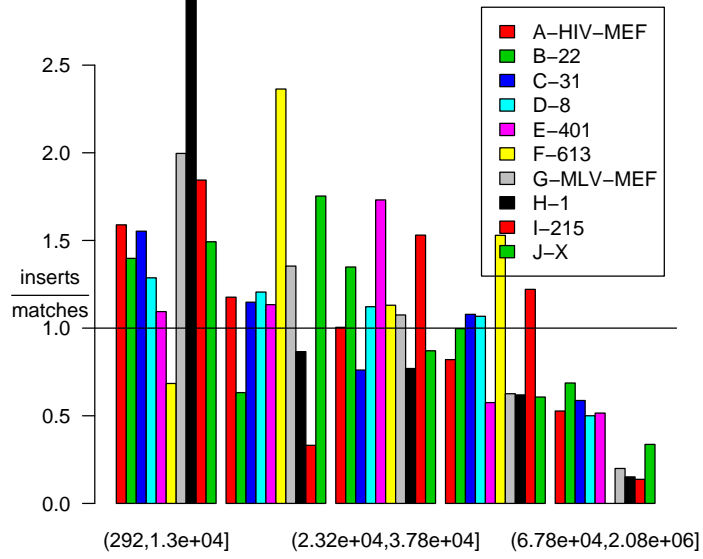
A-HIV-MEF	B-22	C-31	D-8	E-401	F-613	G-MLV-MEF	H-1
1.17e-01	3.15e-02	6.51e-01	5.56e-02	2.26e-01	1.03e-02	5.59e-20	6.94e-03
I-215	J-X						
1.59e-01	6.21e-01						

5.2 genScan Annotations



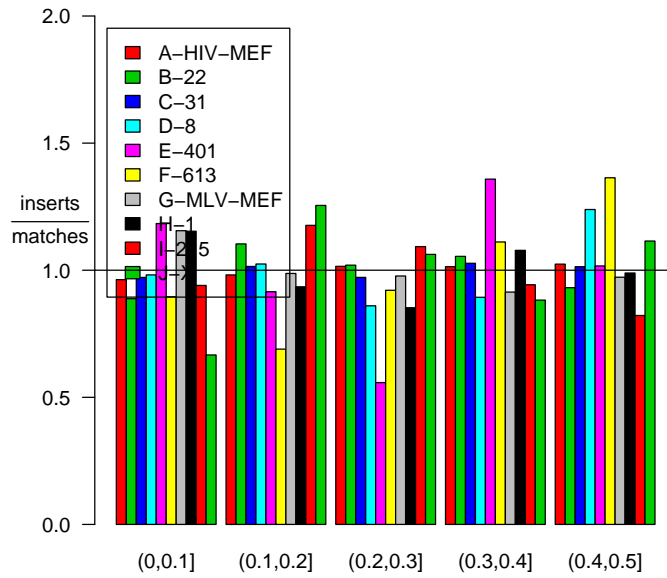
A-HIV-MEF	B-22	C-31	D-8	E-401	F-613	G-MLV-MEF	H-1
9.71e-14	1.63e-02	3.49e-04	4.76e-03	1.46e-02	3.37e-01	2.17e-87	3.16e-16
I-215	J-X						
2.58e-04	2.04e-04						

genScan other.width



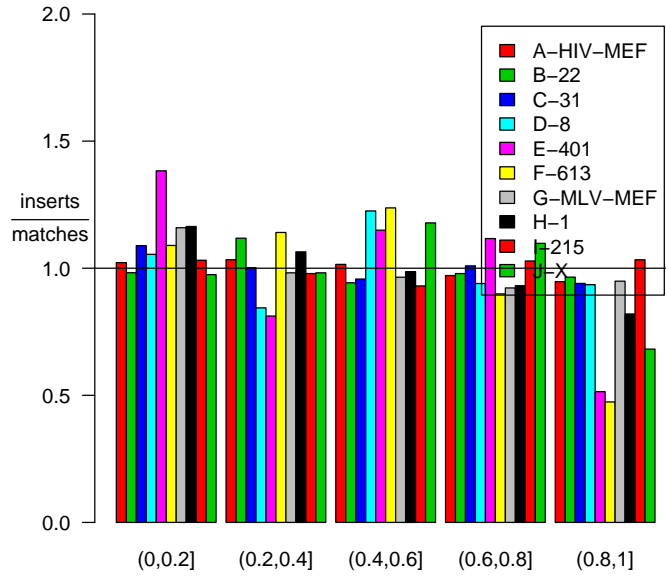
A-HIV-MEF	B-22	C-31	D-8	E-401	F-613	G-MLV-MEF	H-1
1.17e-05	4.40e-03	2.45e-03	1.98e-03	4.10e-03	3.41e-02	1.02e-95	1.94e-22
I-215	J-X						
1.54e-05	8.93e-03						

genScan boundary.dist



A-HIV-MEF	B-22	C-31	D-8	E-401	F-613	G-MLV-MEF	H-1
8.63e-01	3.46e-01	9.92e-01	1.11e-01	3.51e-03	7.99e-01	4.64e-06	2.54e-01
I-215	J-X						
5.60e-01	2.40e-01						

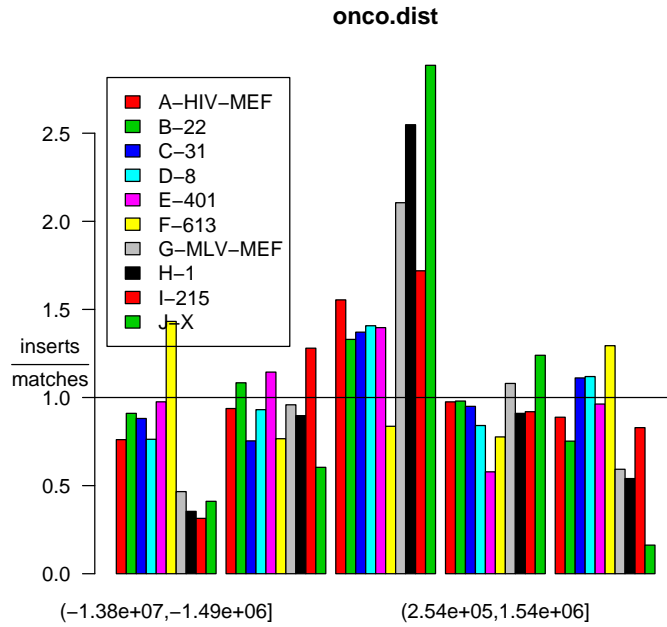
genScan start.dist



A-HIV-MEF	B-22	C-31	D-8	E-401	F-613	G-MLV-MEF	H-1
7.09e-01	6.79e-01	8.03e-01	1.31e-01	1.01e-03	7.17e-01	1.16e-06	1.82e-01
I-215	J-X						
9.91e-01	4.34e-01						

6 Oncogenes

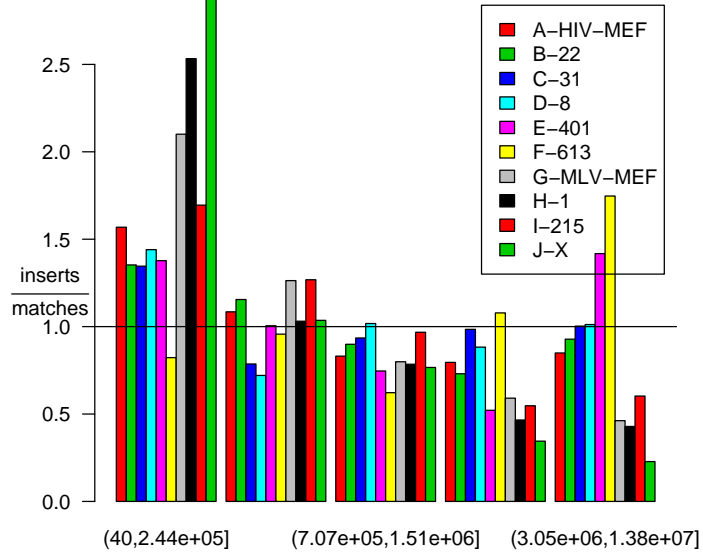
This plot studies the effect of nearness to the 5' end of an oncogene transcript. Positive values represent distances in which the integration site is upstream of the nearest oncogene 5' end, negative downstream.



A-HIV-MEF	B-22	C-31	D-8	E-401	F-613	G-MLV-MEF	H-1
5.75e-31	4.93e-05	1.47e-04	5.43e-04	5.80e-03	5.42e-01	7.42e-290	1.69e-47
I-215	J-X						
7.98e-09	3.64e-23						

Here is the same plot using absolute distance

abs abs.onco.dist

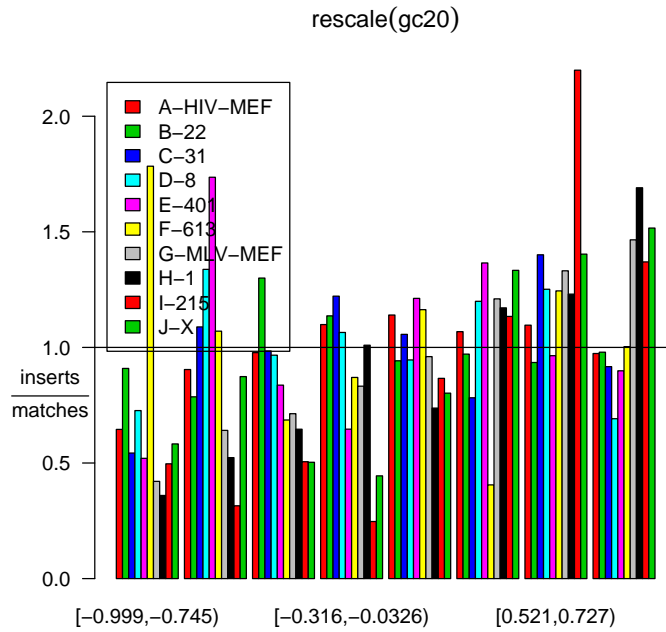


A-HIV-MEF	B-22	C-31	D-8	E-401	F-613	G-MLV-MEF	H-1
6.80e-35	2.45e-06	3.22e-03	2.96e-04	4.82e-05	3.64e-01	7.74e-309	7.18e-47
I-215	J-X						
1.00e-06	1.03e-21						

7 GC content

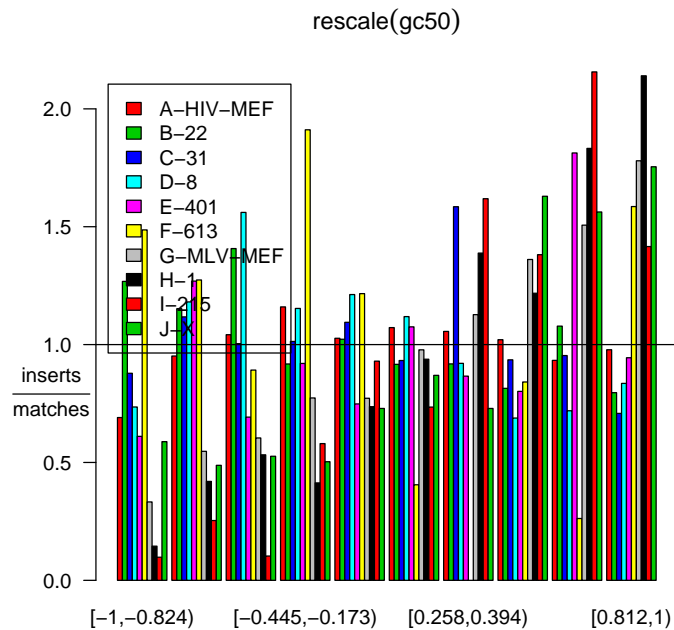
Here we study the effect of GC content on insertion. The GC content is taken from the Mouse Genome Draft at GoldenPath from the table

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

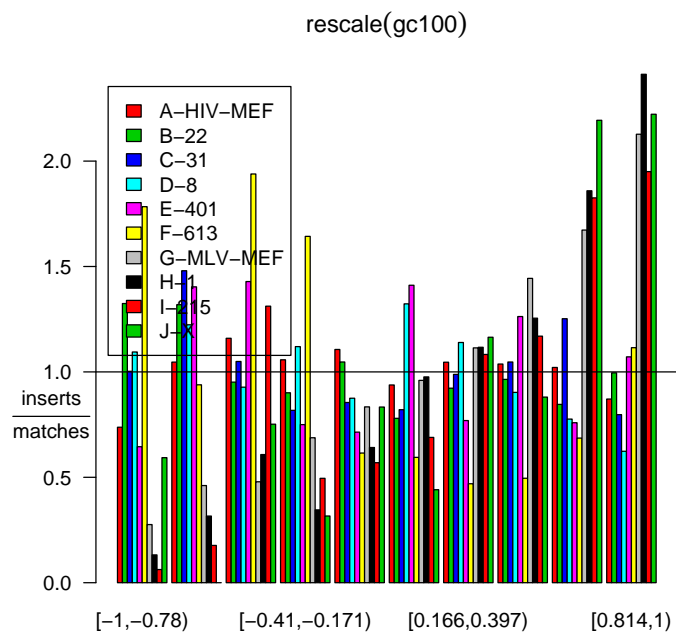


	coef	se	z	p
A-HIV-MEF	0.0651	0.0479	1.3600	1.75e-01
B-22	-0.0648	0.0801	-0.8090	4.19e-01
C-31	0.0015	0.0909	0.0165	9.87e-01
D-8	0.0245	0.1050	0.2340	8.15e-01
E-401	0.1070	0.1490	0.7200	4.72e-01
F-613	-0.0844	0.3640	-0.2320	8.17e-01
G-MLV-MEF	0.5950	0.0340	17.5000	1.45e-68
H-1	0.7080	0.0992	7.1400	9.28e-13
I-215	1.1100	0.1750	6.3200	2.69e-10
J-X	0.8030	0.1870	4.2900	1.78e-05

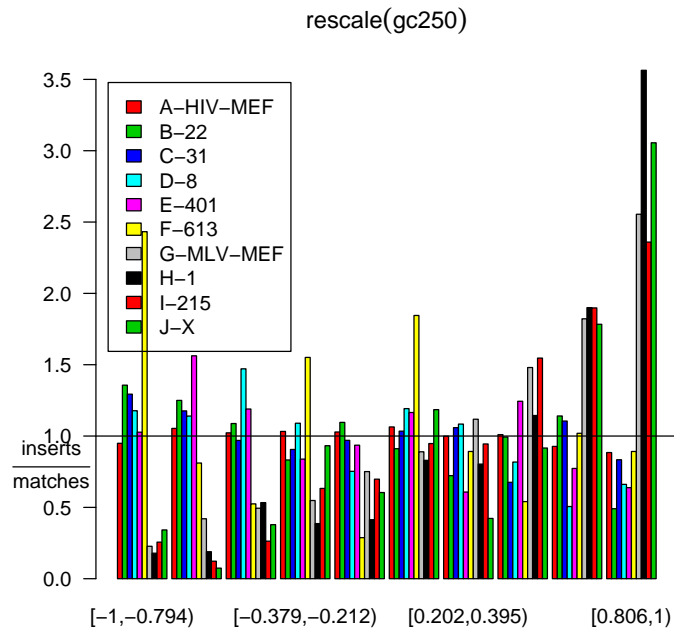
	coef	se	z	p
A-HIV-MEF	9.06e-03	0.0476	1.90e-01	8.49e-01
B-22	-1.73e-01	0.0799	-2.16e+00	3.05e-02
C-31	-1.03e-02	0.0901	-1.15e-01	9.09e-01
D-8	-3.24e-01	0.1060	-3.07e+00	2.17e-03
E-401	-3.98e-17	0.1500	-2.66e-16	1.00e+00
F-613	-1.92e-01	0.3640	-5.29e-01	5.97e-01
G-MLV-MEF	7.67e-01	0.0349	2.20e+01	8.50e-107
H-1	1.01e+00	0.1050	9.57e+00	1.10e-21
I-215	1.03e+00	0.1780	5.77e+00	7.84e-09
J-X	8.71e-01	0.1940	4.48e+00	7.39e-06



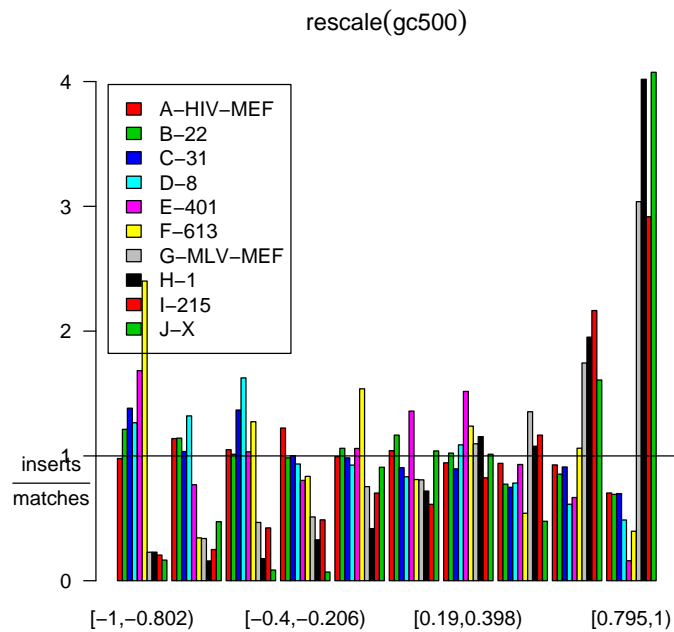
	coef	se	z	p
A-HIV-MEF	-1.72e-02	0.0479	-3.59e-01	7.20e-01
B-22	-1.82e-01	0.0805	-2.26e+00	2.38e-02
C-31	7.52e-04	0.0909	8.27e-03	9.93e-01
D-8	-1.80e-01	0.1050	-1.71e+00	8.73e-02
E-401	-2.13e-16	0.1500	-1.42e-15	1.00e+00
F-613	-4.65e-01	0.3690	-1.26e+00	2.08e-01
G-MLV-MEF	1.01e+00	0.0365	2.76e+01	7.04e-168
H-1	1.24e+00	0.1120	1.11e+01	1.58e-28
I-215	8.97e-01	0.1720	5.22e+00	1.83e-07
J-X	1.17e+00	0.2100	5.58e+00	2.35e-08



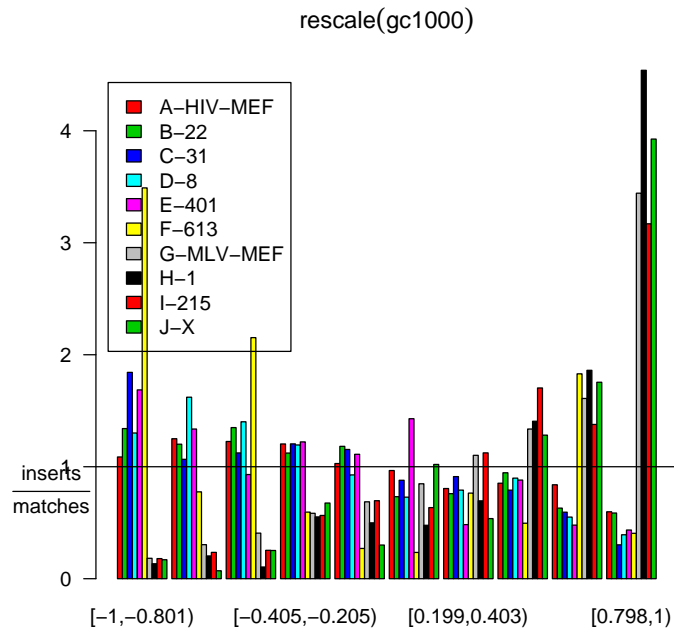
	coef	se	z	p
A-HIV-MEF	-0.0322	0.0472	-0.683	4.94e-01
B-22	-0.2470	0.0807	-3.060	2.21e-03
C-31	-0.1070	0.0903	-1.180	2.36e-01
D-8	-0.3540	0.1060	-3.330	8.66e-04
E-401	-0.1550	0.1500	-1.030	3.02e-01
F-613	0.2030	0.3770	0.539	5.90e-01
G-MLV-MEF	1.1400	0.0371	30.700	4.87e-207
H-1	1.5200	0.1200	12.600	2.08e-36
I-215	1.3600	0.1900	7.170	7.70e-13
J-X	1.2500	0.2140	5.810	6.16e-09



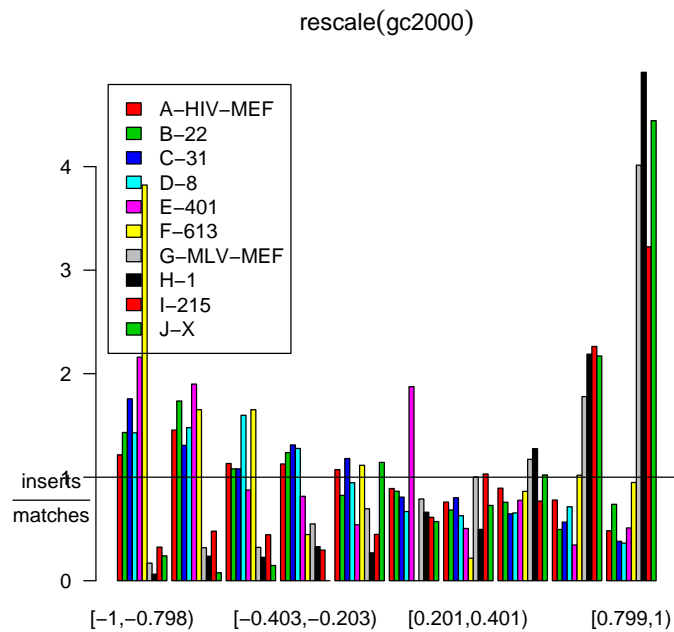
	coef	se	z	p
A-HIV-MEF	-0.163	0.0475	-3.430	6.03e-04
B-22	-0.169	0.0797	-2.120	3.37e-02
C-31	-0.338	0.0918	-3.690	2.27e-04
D-8	-0.500	0.1070	-4.670	3.00e-06
E-401	-0.150	0.1510	-0.994	3.20e-01
F-613	-0.297	0.3690	-0.805	4.21e-01
G-MLV-MEF	1.250	0.0384	32.600	4.34e-233
H-1	1.840	0.1350	13.600	2.57e-42
I-215	1.370	0.1910	7.160	7.81e-13
J-X	1.490	0.2290	6.520	7.10e-11



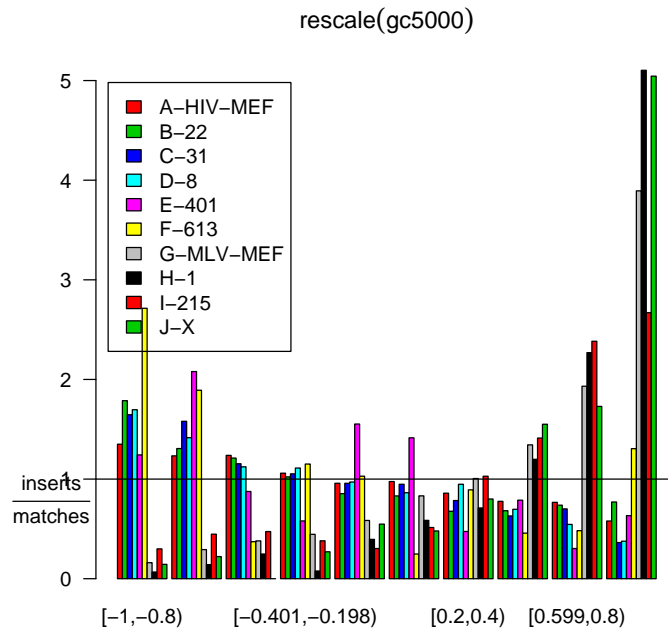
	coef	se	z	p
A-HIV-MEF	-0.325	0.0478	-6.80	1.07e-11
B-22	-0.512	0.0831	-6.17	7.04e-10
C-31	-0.576	0.0936	-6.15	7.64e-10
D-8	-0.637	0.1100	-5.81	6.29e-09
E-401	-0.514	0.1550	-3.31	9.18e-04
F-613	-0.468	0.3640	-1.28	1.99e-01
G-MLV-MEF	1.310	0.0390	33.60	6.32e-247
H-1	1.760	0.1320	13.40	7.47e-41
I-215	1.410	0.1920	7.36	1.89e-13
J-X	1.740	0.2490	6.98	3.03e-12



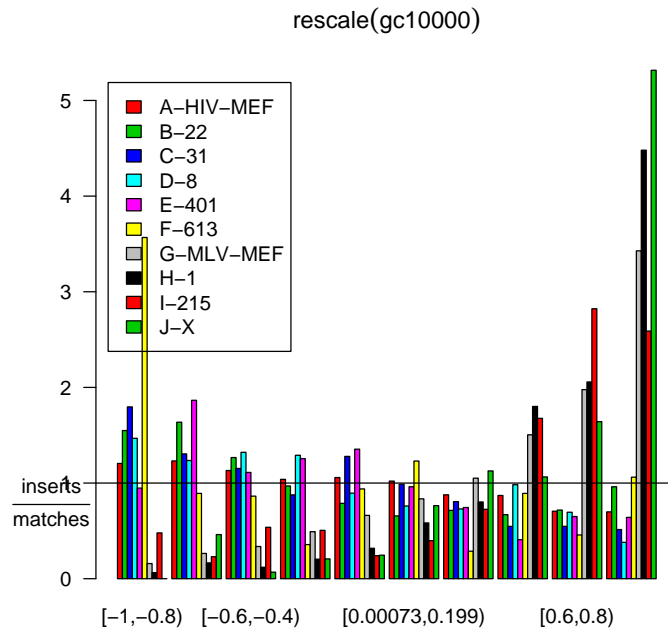
	coef	se	z	p
A-HIV-MEF	-0.435	0.0484	-9.00	2.33e-19
B-22	-0.568	0.0831	-6.83	8.35e-12
C-31	-0.741	0.0972	-7.62	2.47e-14
D-8	-0.771	0.1130	-6.85	7.33e-12
E-401	-0.387	0.1530	-2.54	1.11e-02
F-613	-0.986	0.3770	-2.62	8.89e-03
G-MLV-MEF	1.400	0.0399	35.00	3.78e-268
H-1	2.110	0.1480	14.20	7.36e-46
I-215	1.420	0.1930	7.36	1.88e-13
J-X	1.660	0.2410	6.87	6.34e-12



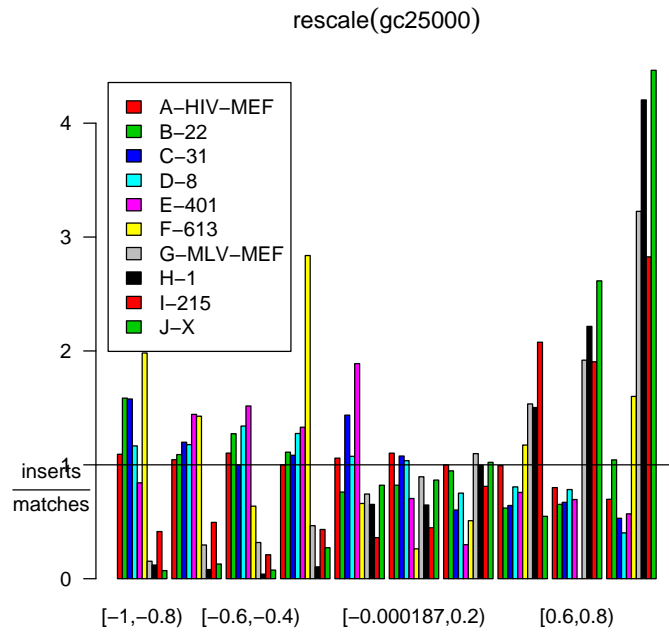
	coef	se	z	p
A-HIV-MEF	-0.366	0.0480	-7.63	2.33e-14
B-22	-0.516	0.0825	-6.25	4.17e-10
C-31	-0.593	0.0951	-6.24	4.51e-10
D-8	-0.583	0.1090	-5.34	9.33e-08
E-401	-0.571	0.1570	-3.64	2.76e-04
F-613	-0.644	0.3720	-1.73	8.34e-02
G-MLV-MEF	1.510	0.0411	36.80	7.34e-297
H-1	2.310	0.1600	14.50	1.91e-47
I-215	1.470	0.1980	7.45	9.52e-14
J-X	1.960	0.2650	7.40	1.34e-13



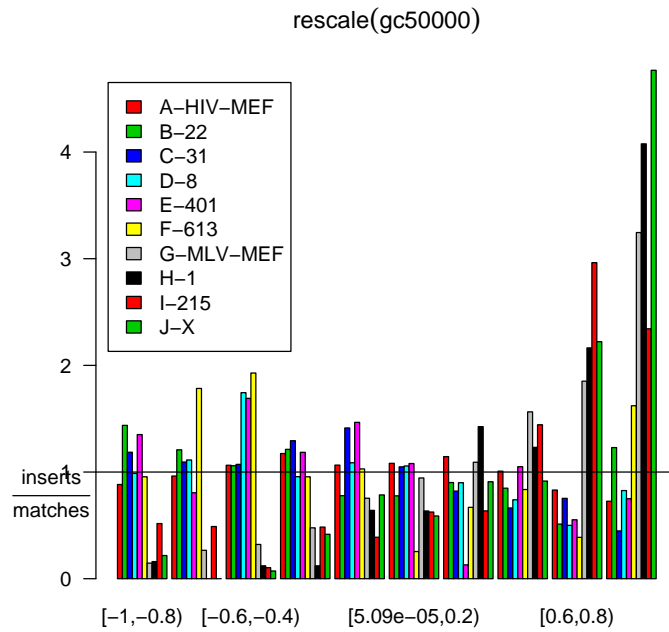
	coef	se	z	p
A-HIV-MEF	-0.290	0.0476	-6.10	1.09e-09
B-22	-0.525	0.0824	-6.37	1.92e-10
C-31	-0.601	0.0952	-6.32	2.70e-10
D-8	-0.550	0.1080	-5.08	3.74e-07
E-401	-0.646	0.1580	-4.09	4.41e-05
F-613	-0.380	0.3680	-1.03	3.03e-01
G-MLV-MEF	1.470	0.0407	36.20	2.18e-287
H-1	2.400	0.1630	14.70	9.99e-49
I-215	1.420	0.1930	7.34	2.07e-13
J-X	2.280	0.2990	7.63	2.37e-14



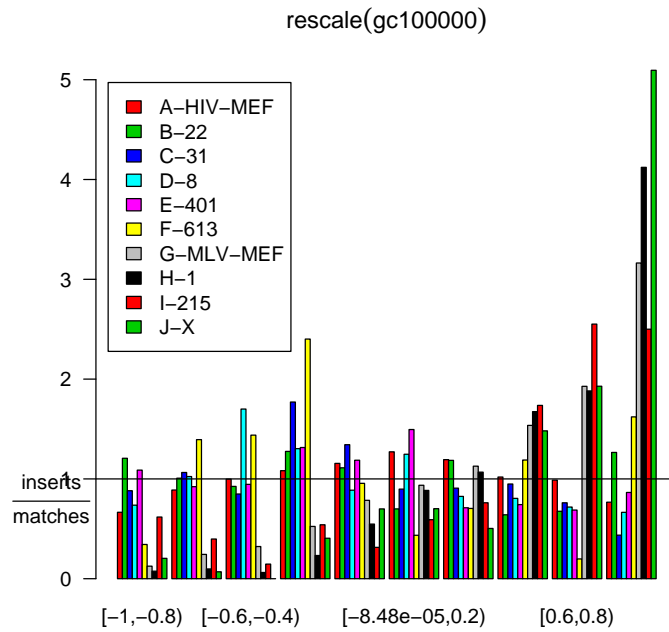
	coef	se	z	p
A-HIV-MEF	-0.131	0.0473	-2.78	5.49e-03
B-22	-0.361	0.0807	-4.47	7.86e-06
C-31	-0.554	0.0940	-5.90	3.73e-09
D-8	-0.447	0.1060	-4.20	2.64e-05
E-401	-0.859	0.1620	-5.30	1.15e-07
F-613	-0.583	0.3740	-1.56	1.19e-01
G-MLV-MEF	1.450	0.0403	35.90	7.04e-283
H-1	2.260	0.1550	14.60	5.71e-48
I-215	1.460	0.1940	7.52	5.44e-14
J-X	1.890	0.2540	7.47	8.30e-14



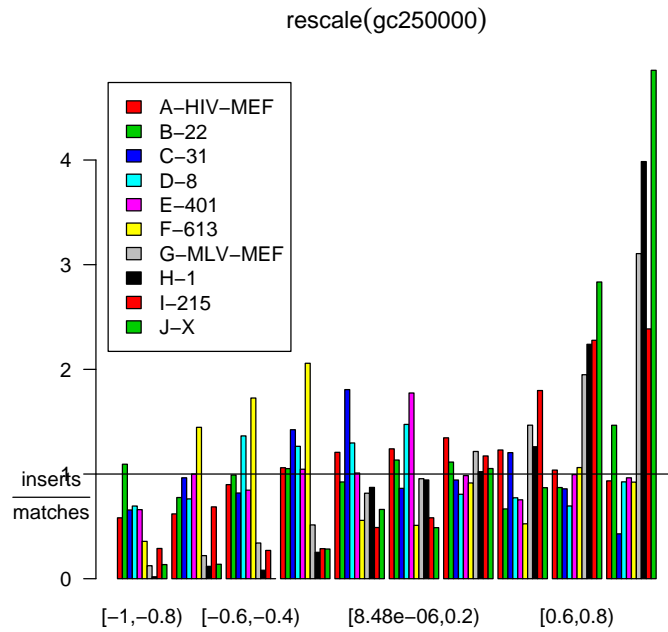
	coef	se	z	p
A-HIV-MEF	-0.0581	0.0471	-1.23	2.17e-01
B-22	-0.2990	0.0802	-3.72	1.97e-04
C-31	-0.4530	0.0929	-4.88	1.06e-06
D-8	-0.3720	0.1060	-3.52	4.36e-04
E-401	-0.6160	0.1560	-3.95	7.79e-05
F-613	-0.5770	0.3740	-1.54	1.24e-01
G-MLV-MEF	1.4600	0.0405	36.10	1.53e-285
H-1	2.2500	0.1560	14.50	1.66e-47
I-215	1.4200	0.1930	7.39	1.46e-13
J-X	1.7900	0.2430	7.38	1.63e-13



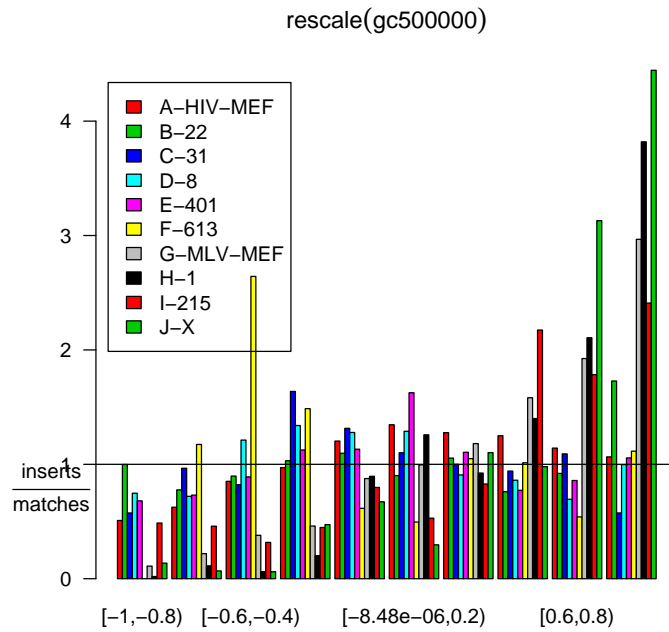
	coef	se	z	p
A-HIV-MEF	0.103	0.0469	2.19	2.89e-02
B-22	-0.218	0.0796	-2.74	6.20e-03
C-31	-0.363	0.0915	-3.96	7.41e-05
D-8	-0.268	0.1050	-2.56	1.04e-02
E-401	-0.202	0.1510	-1.34	1.82e-01
F-613	-0.448	0.3700	-1.21	2.27e-01
G-MLV-MEF	1.460	0.0405	36.00	2.90e-284
H-1	2.250	0.1540	14.60	2.87e-48
I-215	1.420	0.1910	7.47	7.99e-14
J-X	1.870	0.2470	7.55	4.21e-14



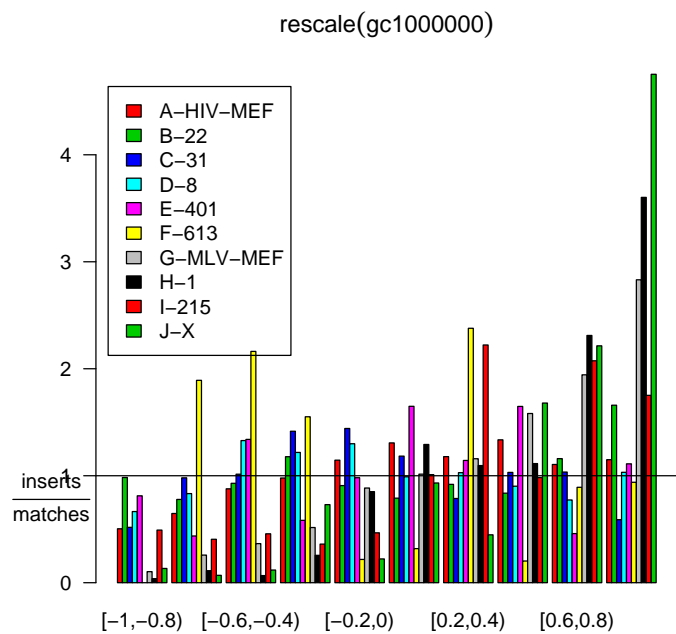
	coef	se	z	p
A-HIV-MEF	0.2990	0.0467	6.410	1.45e-10
B-22	0.0728	0.0784	0.928	3.53e-01
C-31	-0.2280	0.0907	-2.520	1.19e-02
D-8	-0.1250	0.1030	-1.210	2.26e-01
E-401	0.1720	0.1490	1.150	2.49e-01
F-613	-0.3580	0.3690	-0.969	3.32e-01
G-MLV-MEF	1.4500	0.0402	36.000	1.55e-284
H-1	1.9600	0.1380	14.200	6.94e-46
I-215	1.4000	0.1870	7.470	7.72e-14
J-X	2.0800	0.2640	7.870	3.60e-15



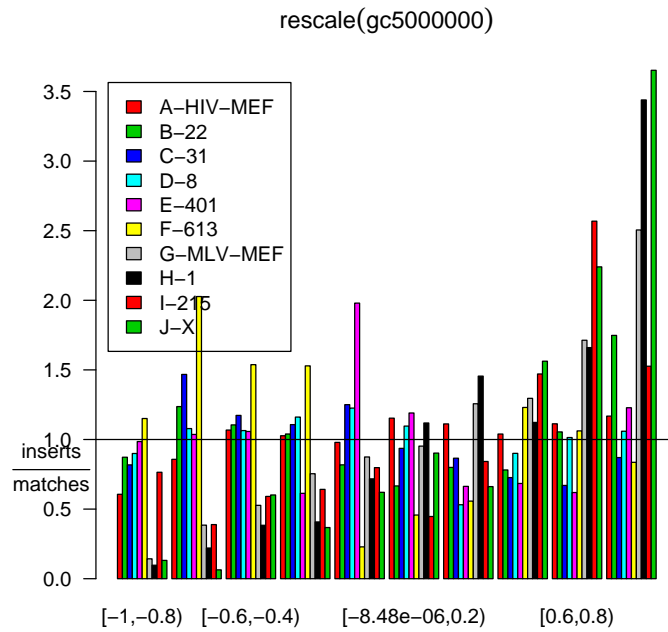
	coef	se	z	p
A-HIV-MEF	0.3980	0.0469	8.490	2.09e-17
B-22	0.1010	0.0784	1.280	1.99e-01
C-31	-0.0888	0.0900	-0.986	3.24e-01
D-8	-0.0902	0.1030	-0.876	3.81e-01
E-401	0.1710	0.1500	1.140	2.54e-01
F-613	-0.2320	0.3670	-0.633	5.27e-01
G-MLV-MEF	1.4500	0.0403	36.000	1.32e-283
H-1	2.0200	0.1410	14.400	8.39e-47
I-215	1.1300	0.1750	6.470	9.91e-11
J-X	1.9300	0.2530	7.630	2.30e-14



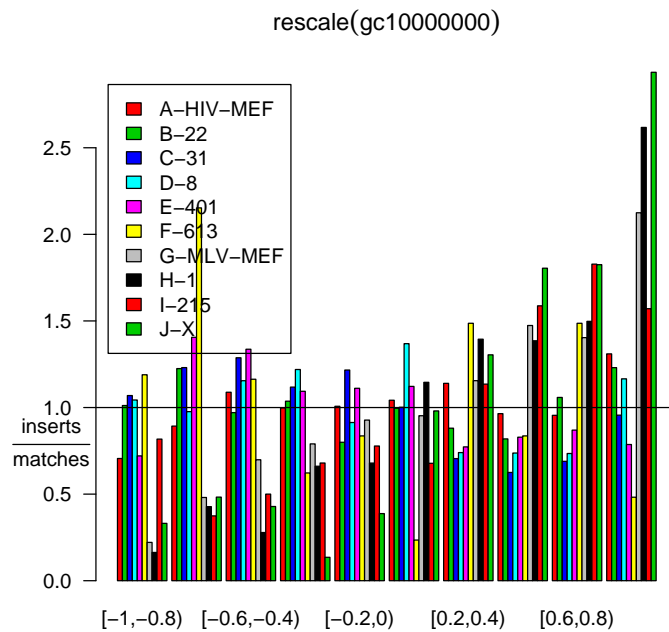
	coef	se	z	p
A-HIV-MEF	0.3940	0.0470	8.390	4.90e-17
B-22	0.0981	0.0785	1.250	2.11e-01
C-31	-0.1190	0.0902	-1.320	1.87e-01
D-8	-0.0967	0.1030	-0.940	3.47e-01
E-401	0.3660	0.1510	2.430	1.52e-02
F-613	-0.1220	0.3670	-0.333	7.39e-01
G-MLV-MEF	1.4000	0.0397	35.200	1.07e-270
H-1	1.9600	0.1380	14.200	8.78e-46
I-215	1.3100	0.1840	7.120	1.07e-12
J-X	2.0200	0.2580	7.830	4.79e-15



	coef	se	z	p
A-HIV-MEF	0.2200	0.0469	4.690	2.74e-06
B-22	-0.0302	0.0785	-0.385	7.00e-01
C-31	-0.3440	0.0914	-3.770	1.64e-04
D-8	-0.1640	0.1030	-1.580	1.13e-01
E-401	-0.2640	0.1520	-1.740	8.26e-02
F-613	-0.2940	0.3680	-0.799	4.24e-01
G-MLV-MEF	1.0400	0.0366	28.500	7.66e-179
H-1	1.5400	0.1210	12.800	2.54e-37
I-215	0.7770	0.1660	4.690	2.72e-06
J-X	1.6300	0.2310	7.050	1.85e-12



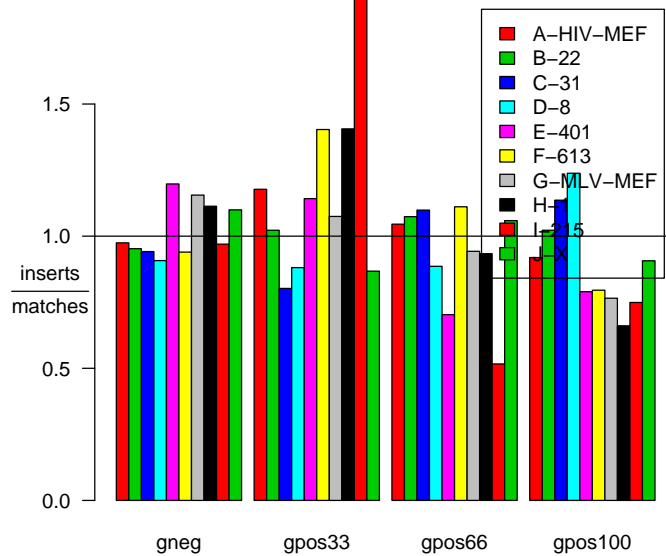
	coef	se	z	p
A-HIV-MEF	0.141	0.0467	3.030	2.49e-03
B-22	-0.019	0.0784	-0.242	8.09e-01
C-31	-0.388	0.0919	-4.220	2.44e-05
D-8	-0.117	0.1030	-1.140	2.56e-01
E-401	-0.246	0.1500	-1.640	1.01e-01
F-613	-0.165	0.3740	-0.441	6.59e-01
G-MLV-MEF	0.817	0.0353	23.200	1.03e-118
H-1	1.300	0.1140	11.300	7.62e-30
I-215	0.787	0.1670	4.730	2.26e-06
J-X	1.560	0.2270	6.870	6.55e-12



8 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	-0.2080	0.0276	-7.52	5.43e-14
cyto.typegpos33	0.0352	0.0319	1.10	2.69e-01
cyto.typegpos66	-0.0905	0.0372	-2.43	1.50e-02

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.