Association of Genomic Features with Integration in Stably Expressed or Inducible Cell Lines

Charles C. Berry

August 31, 2004

Contents

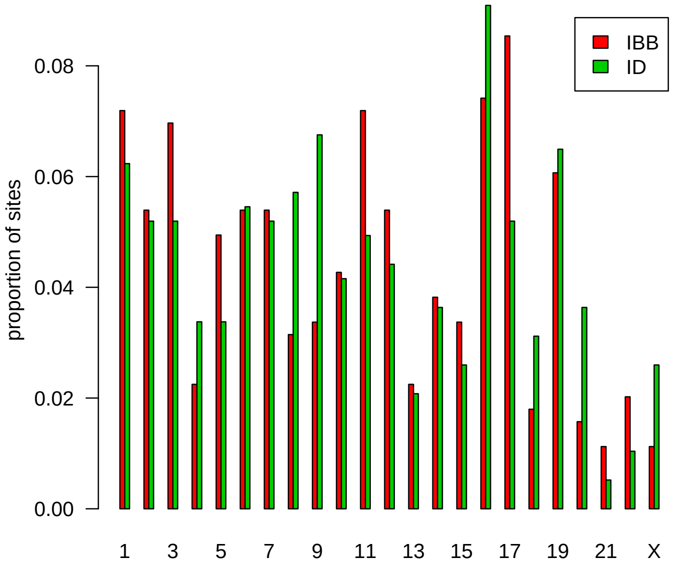
1 Introduction

In this document, I examine the association of integration siting in cells selected as **stably expressed** (labelled 'IBB' hereafter) or **inducible** (labelled 'ID') with various genomic features.

The numbers are shown below:

```
exp.group
IBB  ID
447 388
```

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

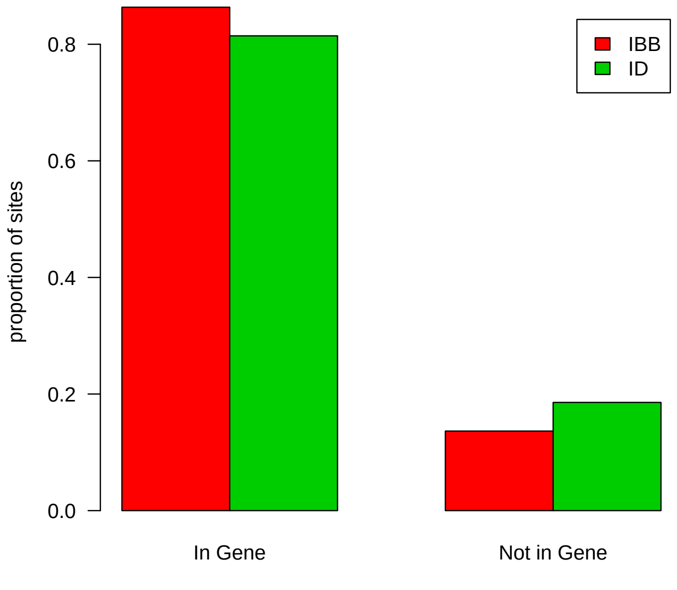


Are there chromosomes that are particularly favored for integration by one group over the other? This was tested for statistical significance. The test performed used the likelihood ratio statistic for the logistic regression model (reviewed in [2]) as implemented by the `glm` function of R using the `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 0.17674.

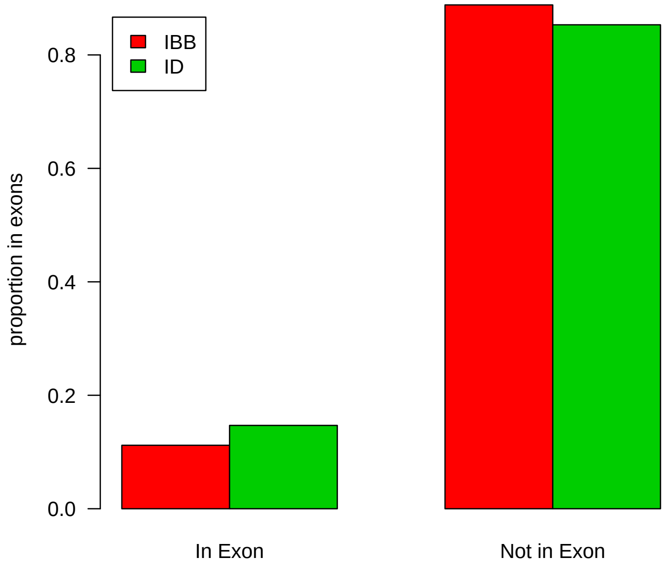
2 Preference for Genes

2.1 Acembly Genes

Here we examine the relative preference that integration events in the two groups have for genes. In the following plot we show the relative frequency of integrations in genes according to the 'Acembly' anotation. The bars grouped over the label "In Gene" give the relative frequency of integration events (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 is a difference in the tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.053439. In the following plot we show the relative frequency of insertions in exons according to the 'Acembly' anotation The bars grouped over the label “In Exon” give the relative frequency of integration events (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.



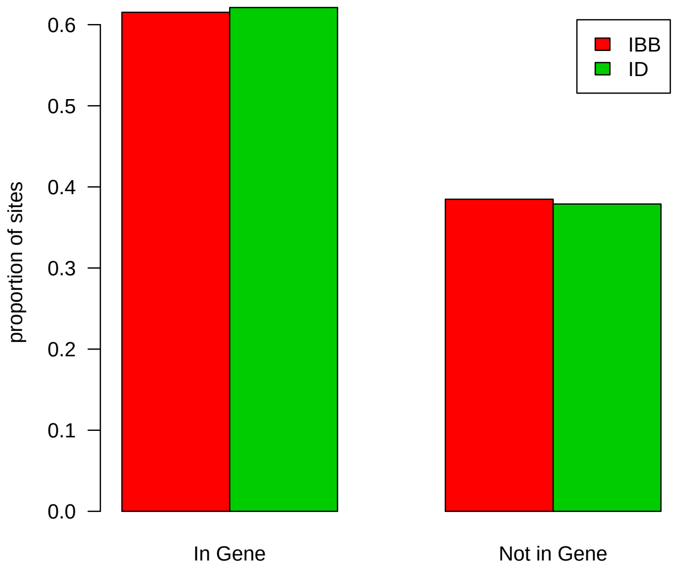
Here is the table of coefficients of the log ratio of intensities for ID sites versus IBB sites along with their standard errors, z statistics, and p-values:

```
               coef    se      z      p
(Intercept)  0.166 0.174  0.953 0.3410
in.gene     -0.426 0.193 -2.210 0.0270
in.exon      0.391 0.211  1.860 0.0632
```

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.

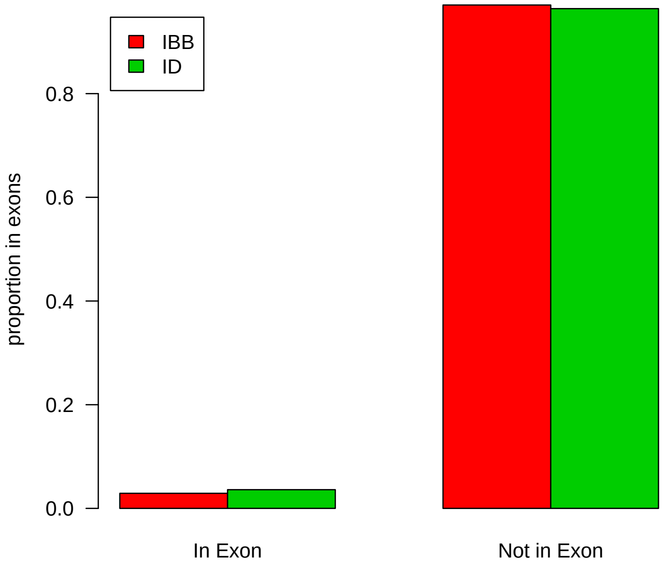
2.2 refGenes

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’ anotation.



Is there is a tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.86057.

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



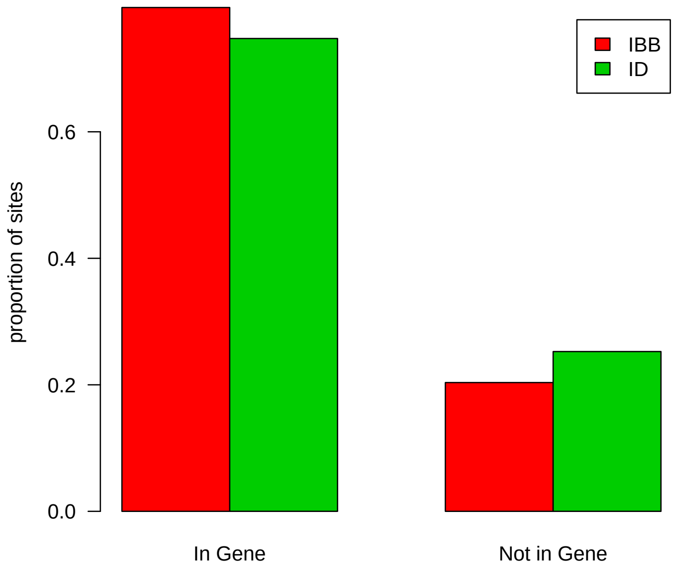
Here is the table of coefficients of the log ratio of intensities for ID sites versus IBB sites along with their standard errors, z statistics, and p-values:

```
               coef    se       z     p
(Intercept) -0.1570 0.112 -1.4000 0.162
in.gene      0.0137 0.144  0.0947 0.925
in.exon      0.2180 0.396  0.5500 0.583
```

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.

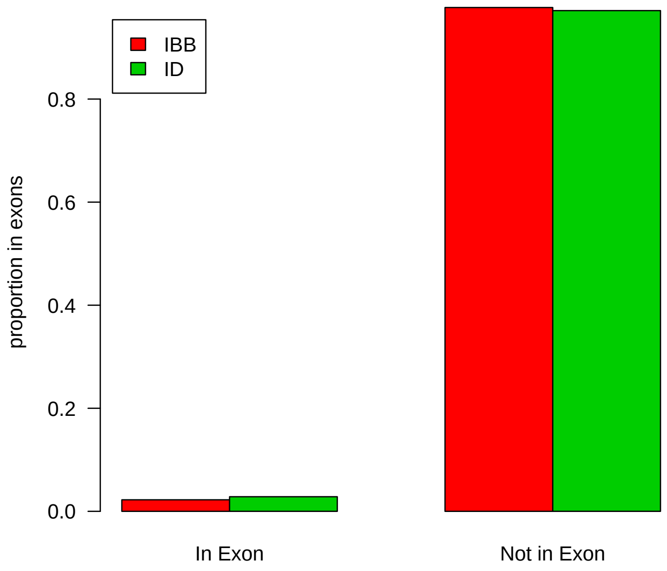
2.3 genScan Genes

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' anotation.



Is there is a tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.091842.

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



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

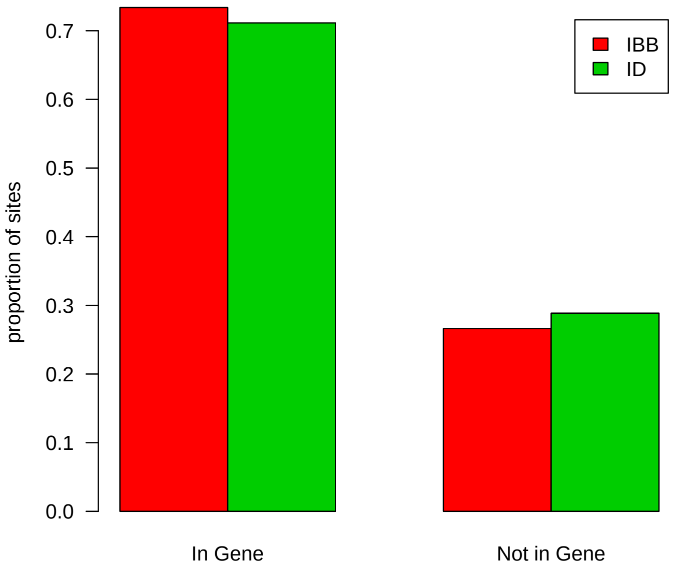
```
                coef    se      z     p
(Intercept)  0.0741 0.146  0.509 0.611
in.gene     -0.2890 0.166 -1.740 0.082
in.exon      0.3110 0.444  0.699 0.485
```

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.

data set.

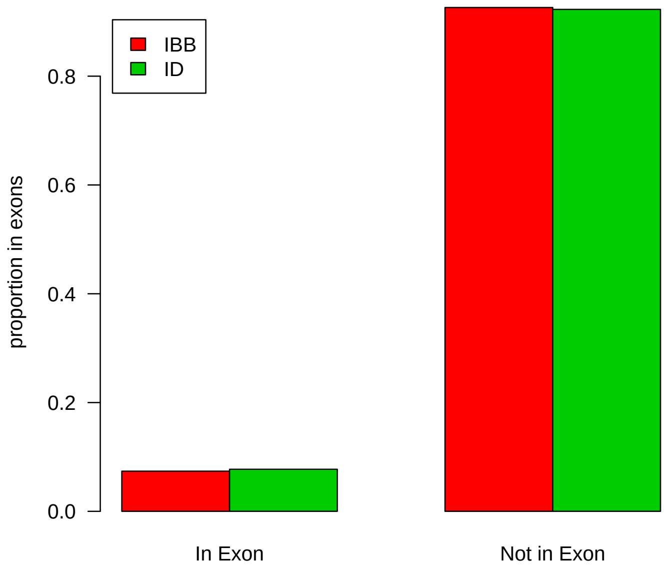
2.4 uniGenes

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' anotation.



Is there is a tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.46991.

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



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

```
               coef    se      z     p
(Intercept) -0.0606 0.132 -0.460 0.645
in.gene     -0.1210 0.157 -0.769 0.442
in.exon      0.0863 0.267  0.324 0.746
```

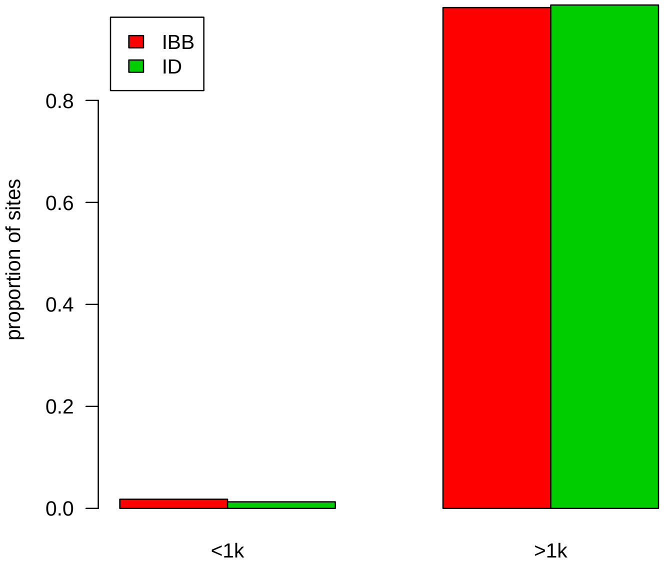
The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as 'in.exon' is net of that due to being in a gene. Note that in the barplot above the 'Not in Exon' bars include the both introns and intergenic regions, so the impression given 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 [4], who found that the neighborhoods within ±1kb of CpG islands are enriched for MLV insertions, we study such neighborhoods.

3.1 1 kilobase neighborhoods

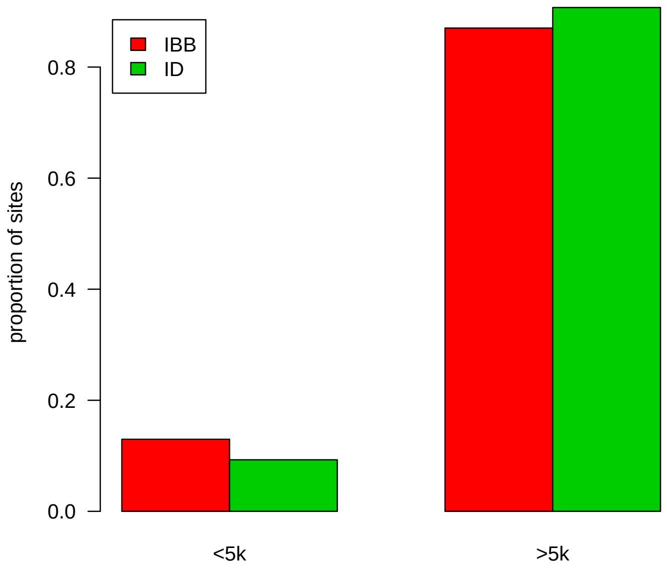
The following plot shows the effect of being in or within ±1kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.55736.

3.2 5 kilobase neighborhoods

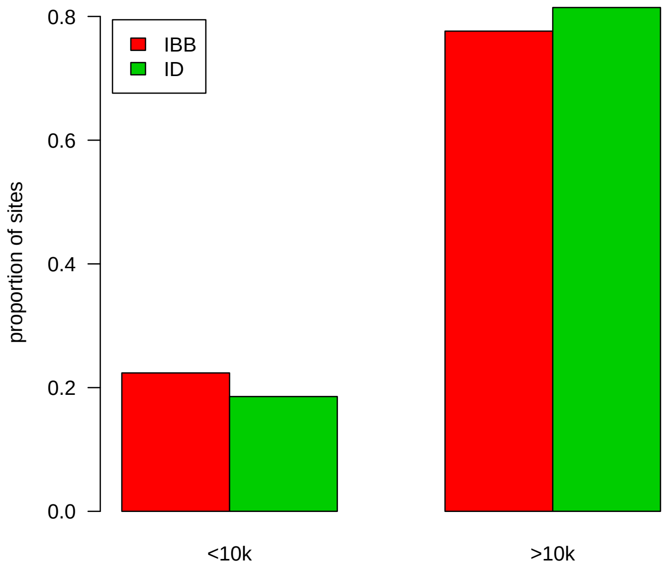
The following plot shows the effect of being in or within ±5kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.09008.

3.3 10 kilobase neighborhoods

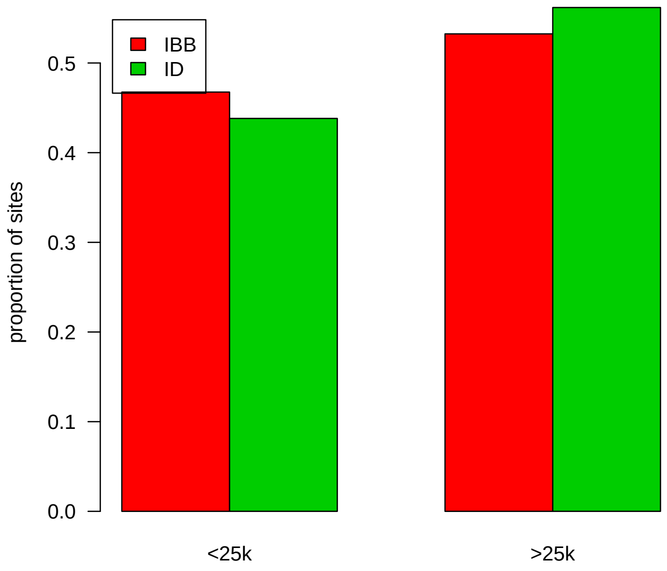
The following plot shows the effect of being in or within ±10kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.17307.

3.4 25 kilobase neighborhoods

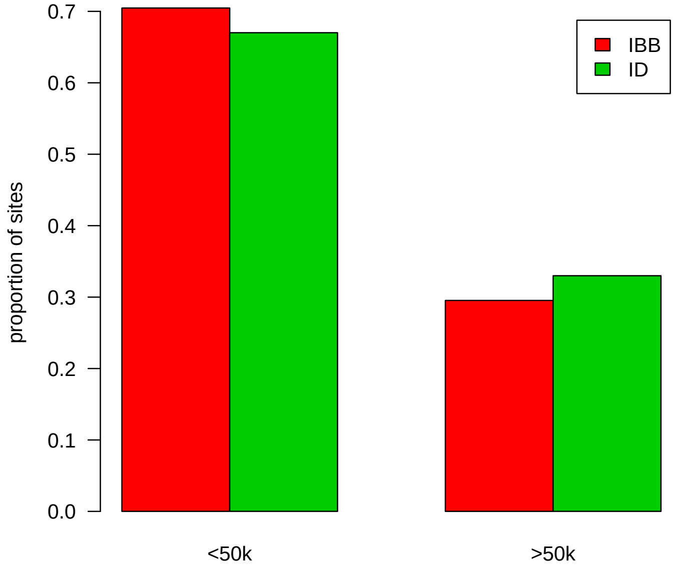
The following plot shows the effect of being in or within ±25kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.39436.

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 0.28185.

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

In this section the association with gene density is examined. The 'genes' that are counted are the Ensembl genes. In addition, we study various functions of the EST counts for the Ensembl genes using data described in Versteeg et al [3] and CpG Island density. Based on preliminary observations, it was decided to determine the density of ESTs found in a region in the following ways:

count.exprs Count only one EST per gene and divide by number of bases

exprs Count up to 200 ESTs per gene and divide by number of bases

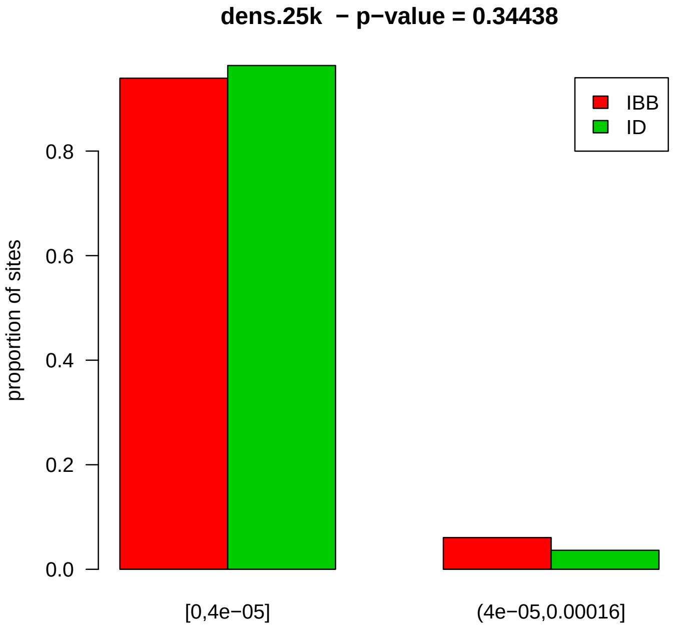
big.exprs Counting only the ESTs in excess of two hundred per gene and divide by number of bases

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.

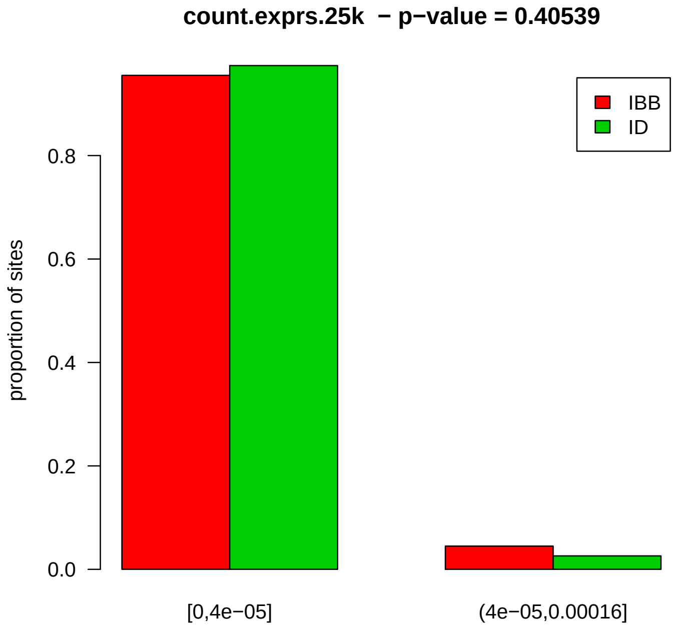
4.1 25 kiloBase Window

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 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, then the groupings contain ten percent of the sites each except for the leftmost grouping which will contain all of the remaining sites.

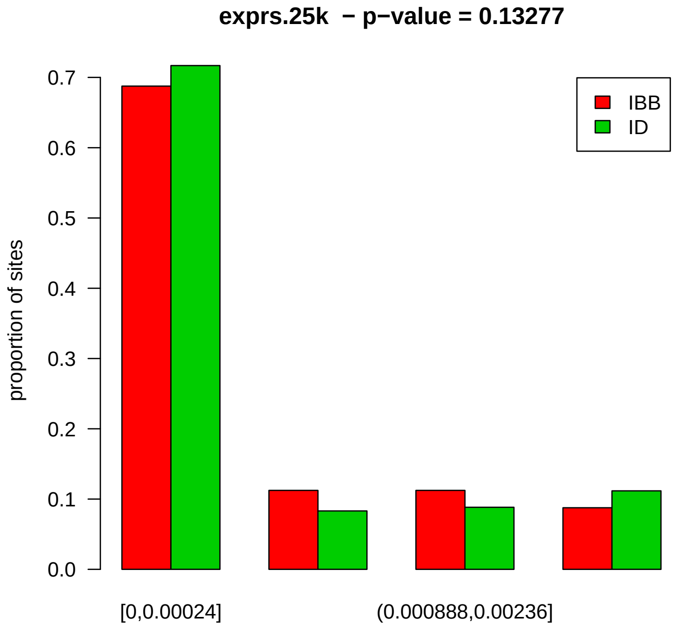
Also note that the title of the plot contains clues as to its content; the prefix indicates the type of variable studied while the suffix indicates the window width in the number of bases. The p-value given is the result of fitting a quadratic polynomial to the gene density values.



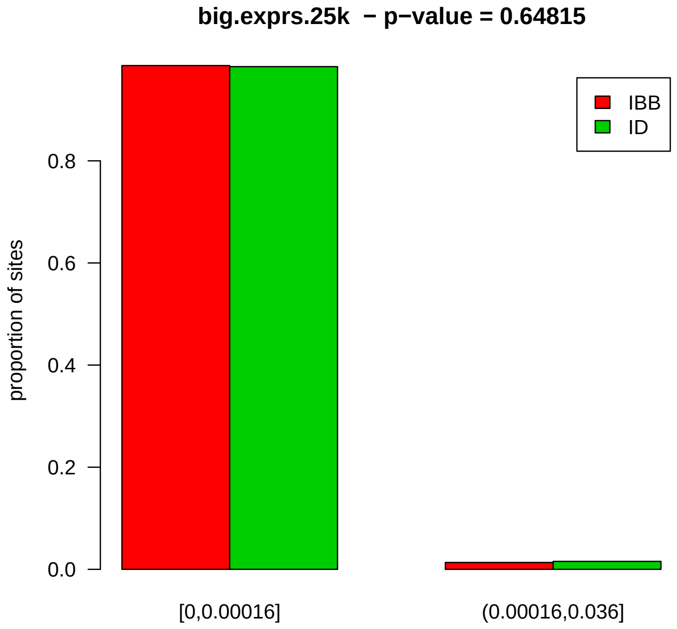
In the barplot that follows we examine the association of insertion sites with expression density in a 25 kilobase window surrounding each locus. First, we count just one EST per gene.



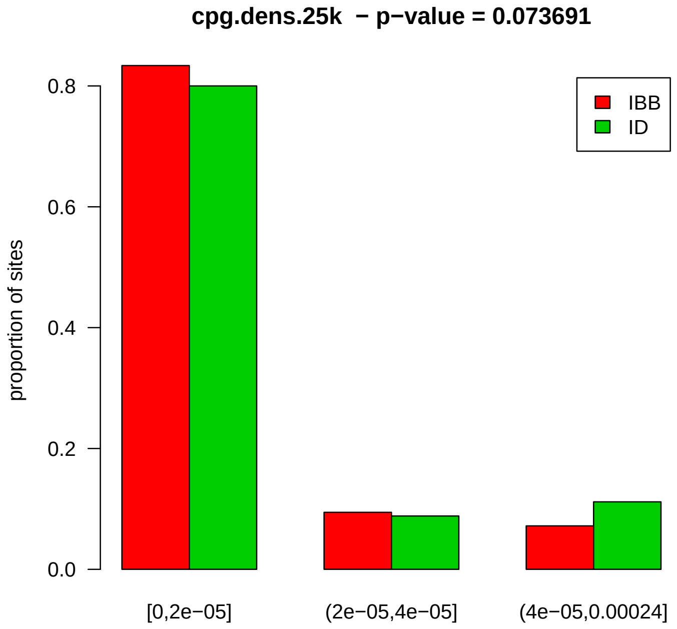
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene

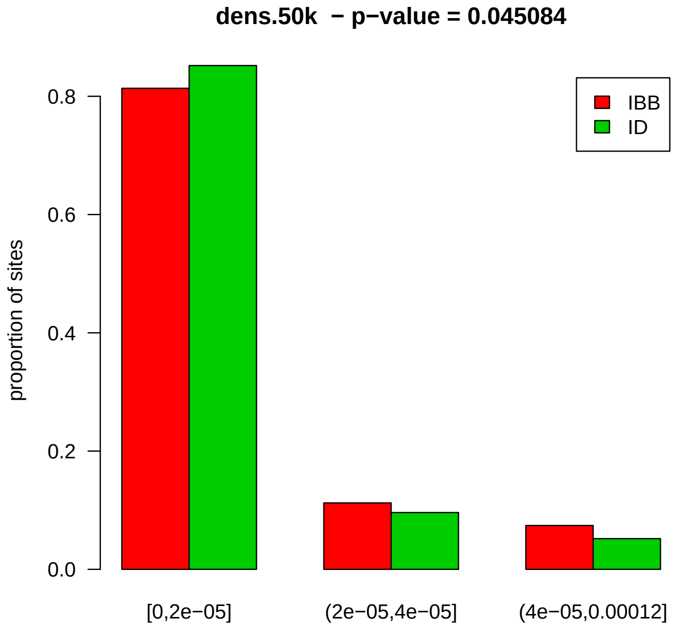


Here the effect of density of CpG islands is studied:

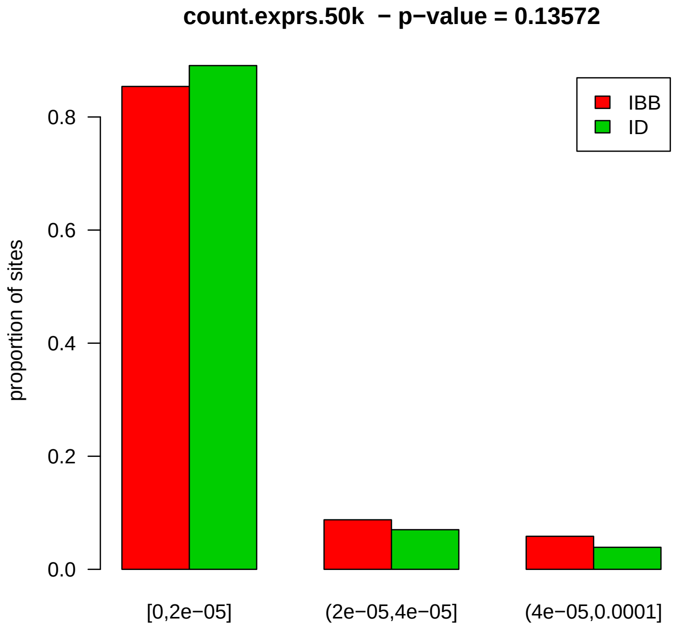


4.2 50 kiloBase Window

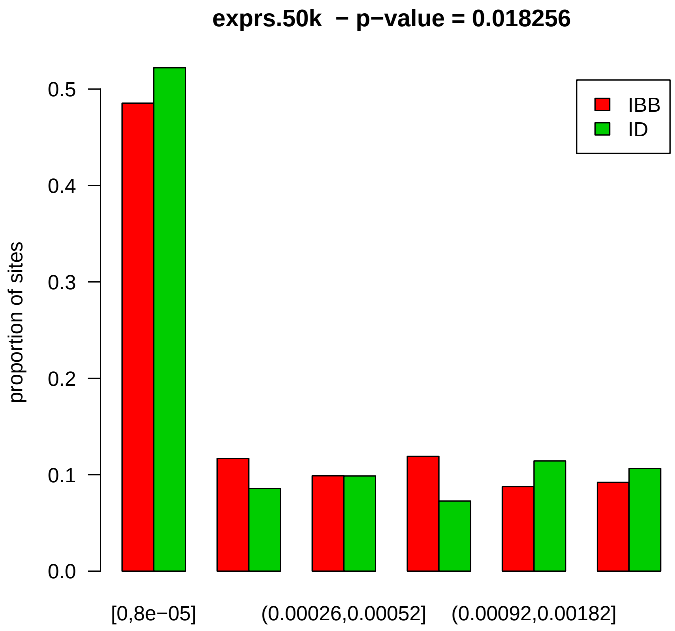
First, we see gene density:



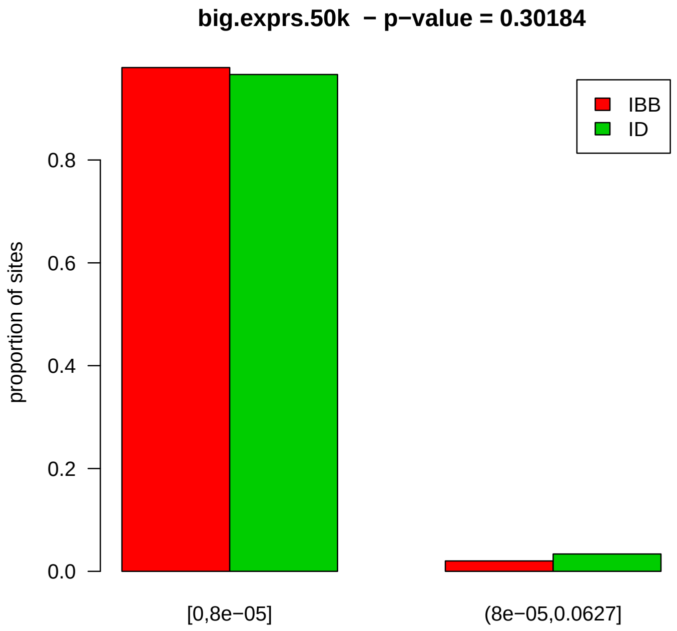
Here are the results for EST density. First, we count just one EST per gene.



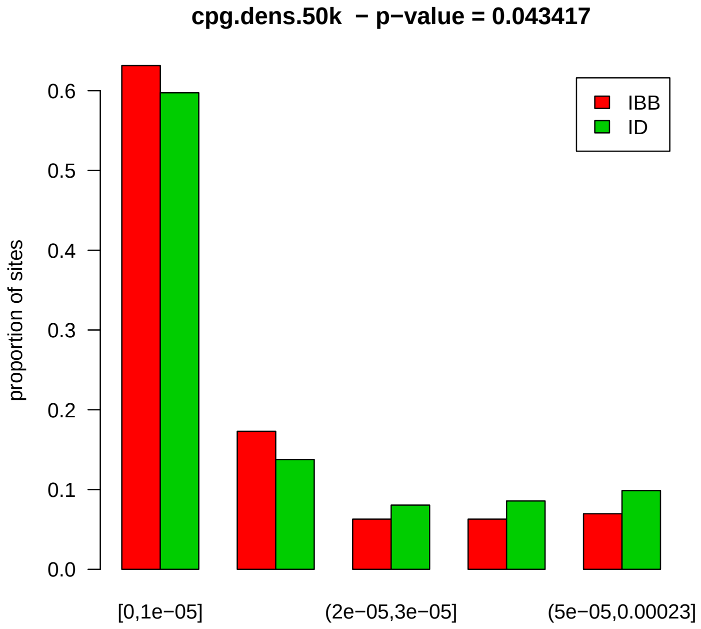
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene

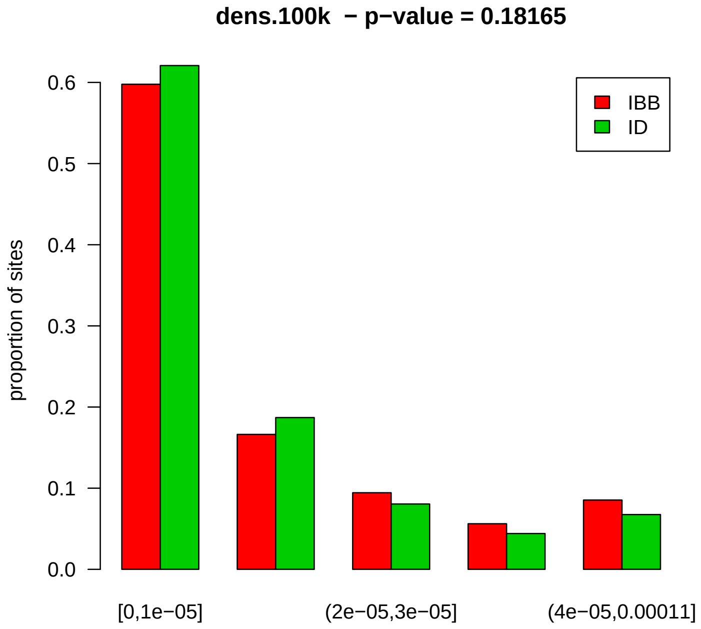


Here the effect of density of CpG islands is studied:

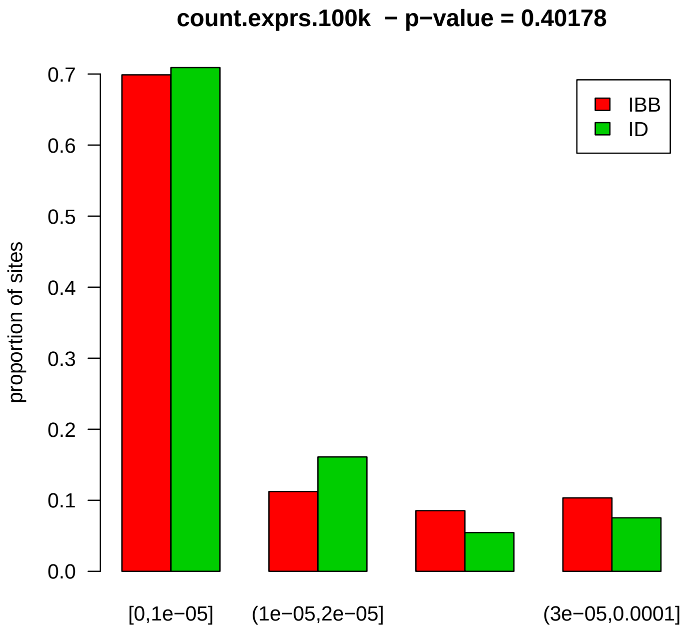


4.3 100 kiloBase Window

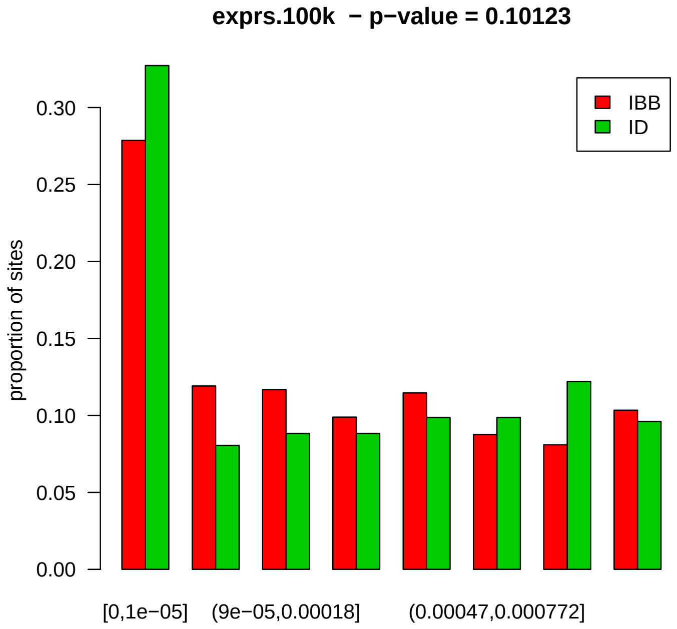
First, we see gene density:



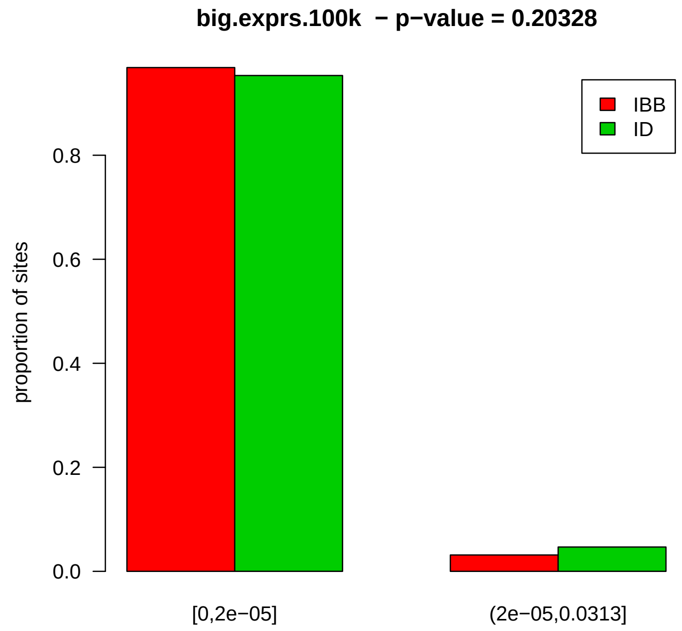
Here are the results for EST density. First, we count just one EST per gene.



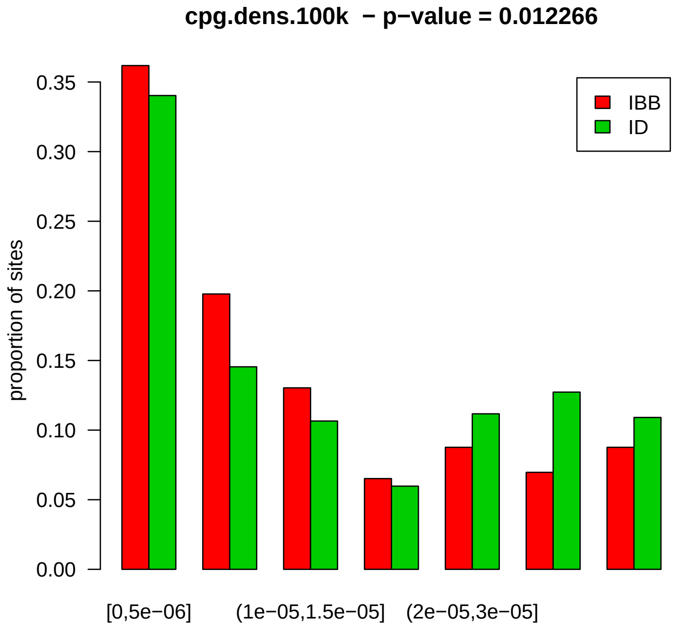
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene

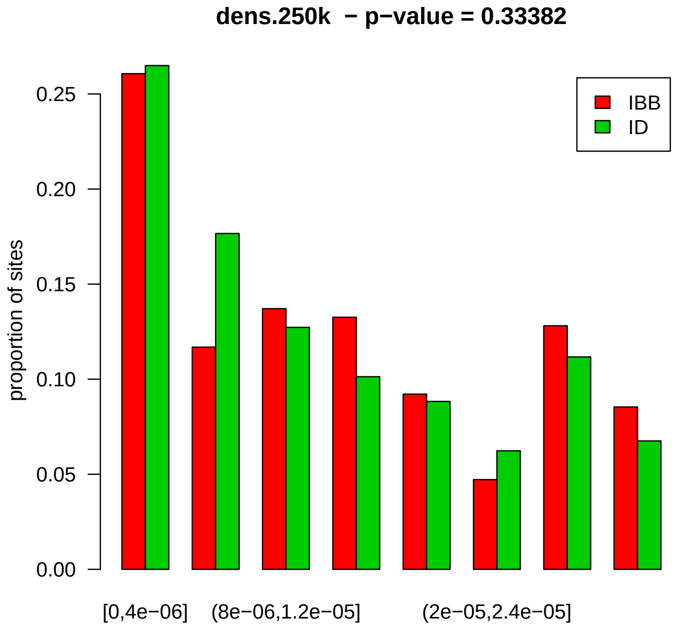


Here the effect of density of CpG islands is studied:

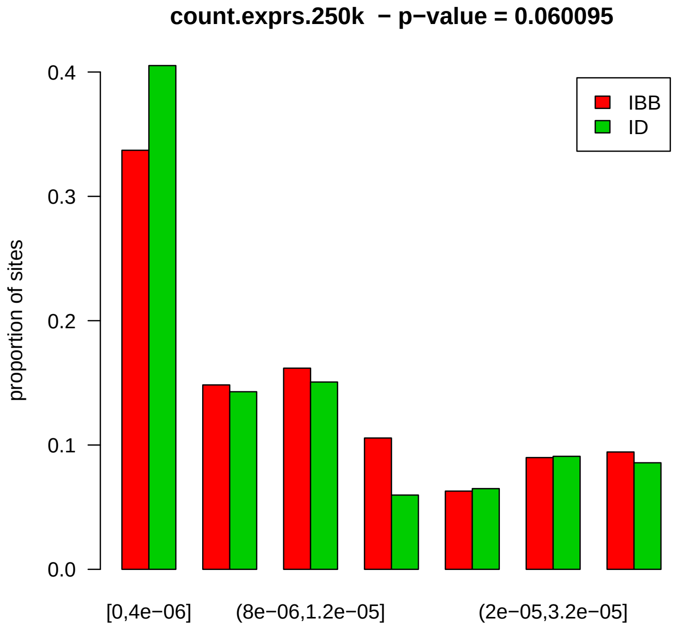


4.4 250 kiloBase Window

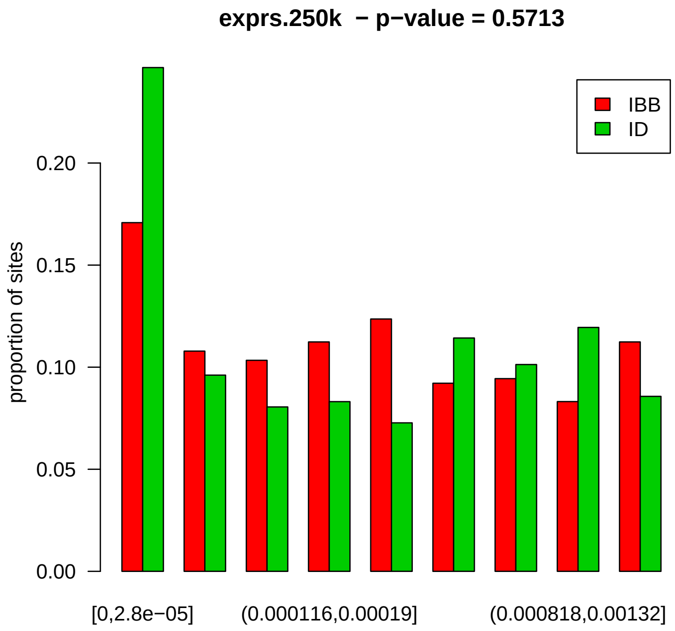
First, we see gene density:



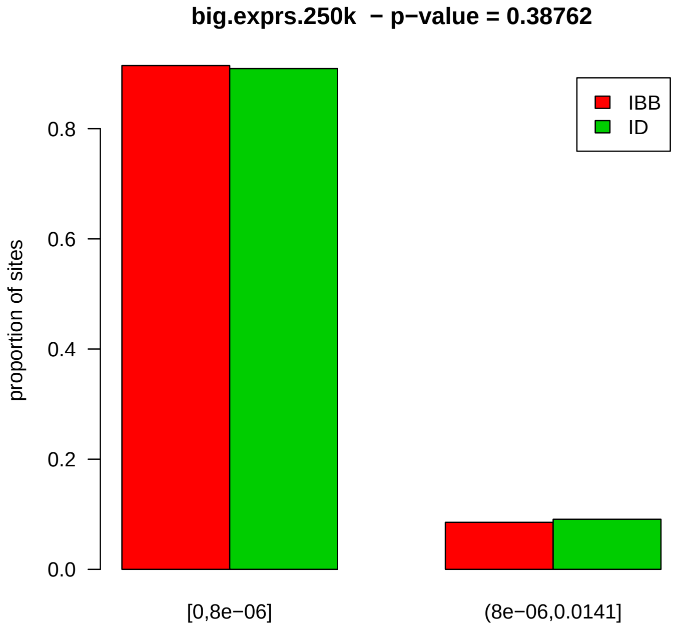
Here are the results for EST density. First, we count just one EST per gene.



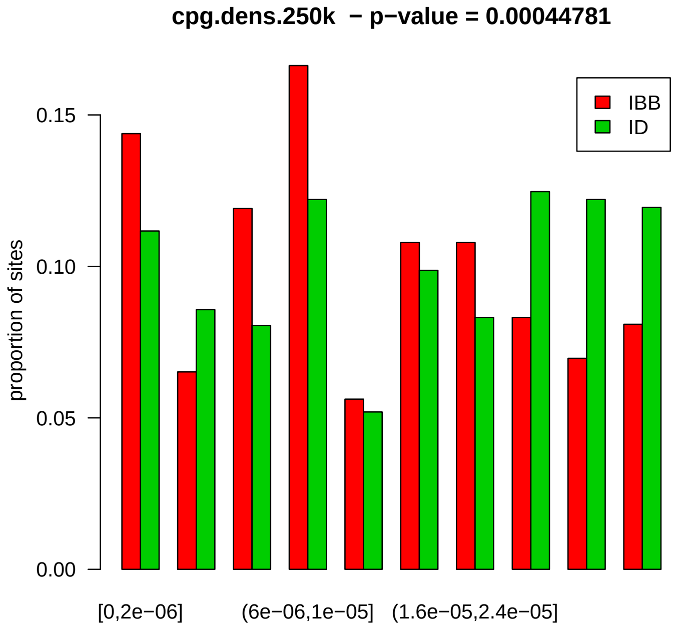
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene

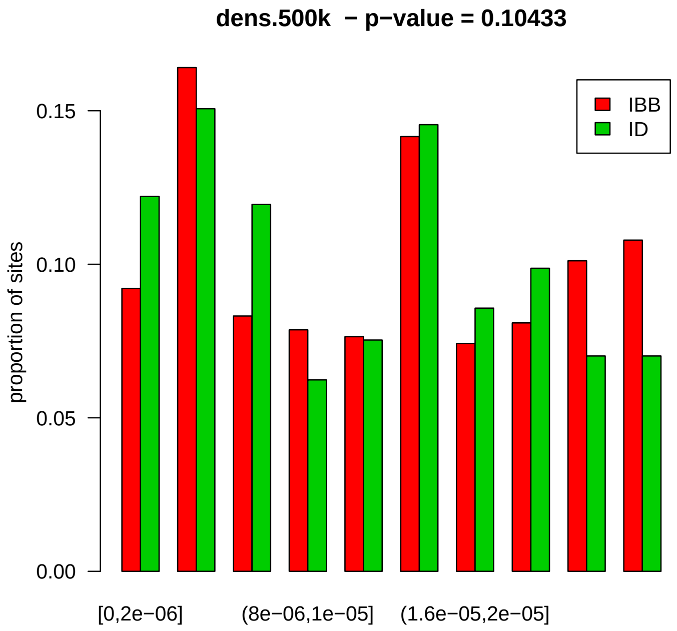


Here the effect of density of CpG islands is studied:

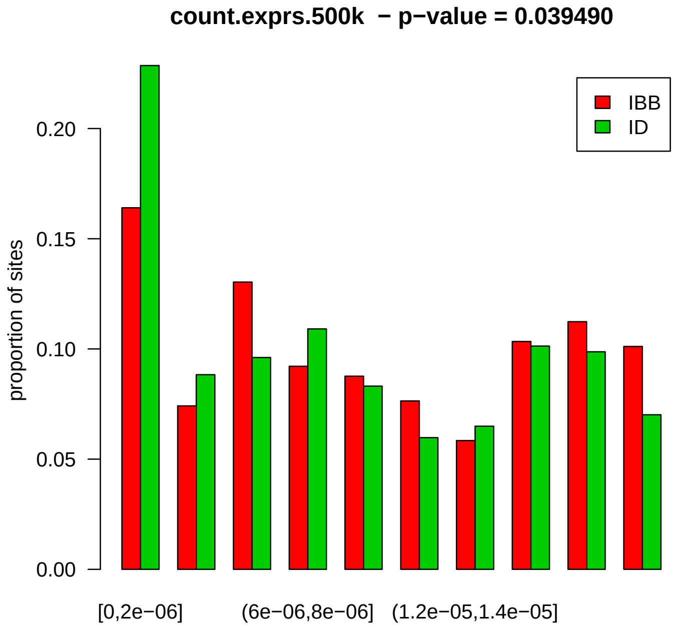


4.5 500 kiloBase Window

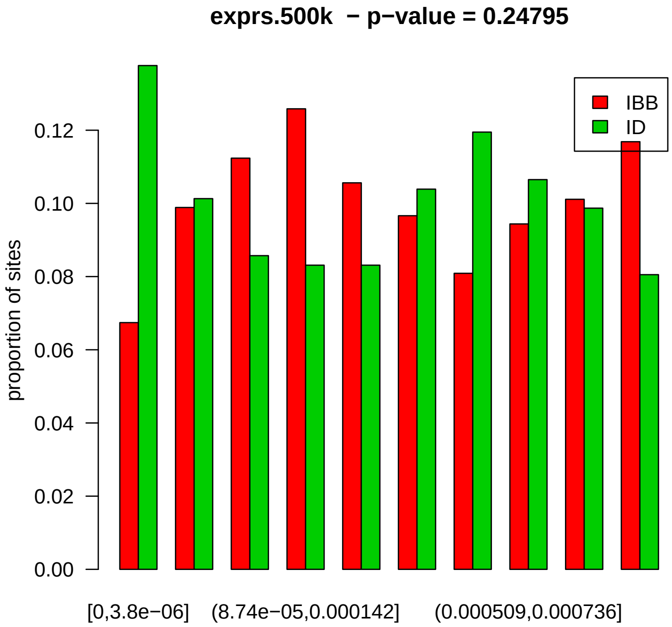
First, we see gene density:



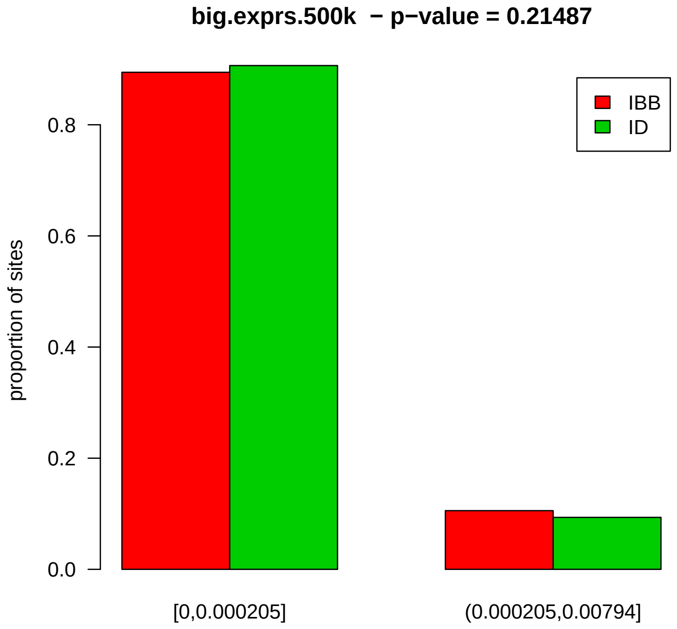
Here are the results for EST density. First, we count just one EST per gene.



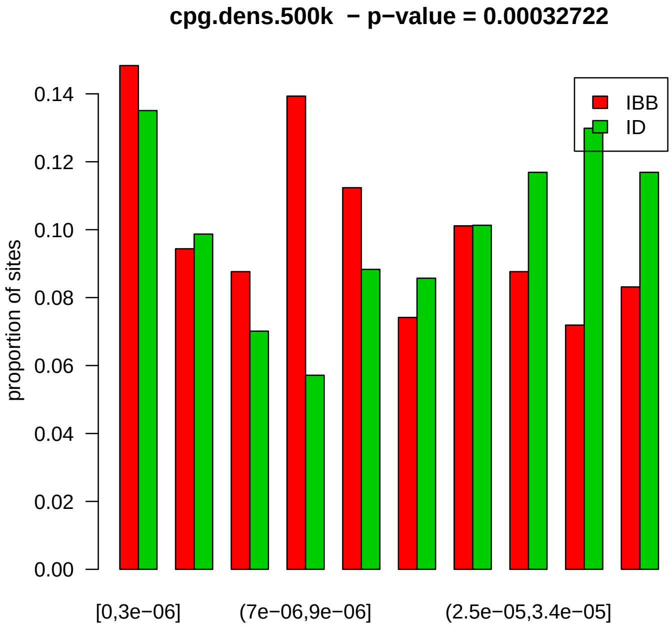
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene

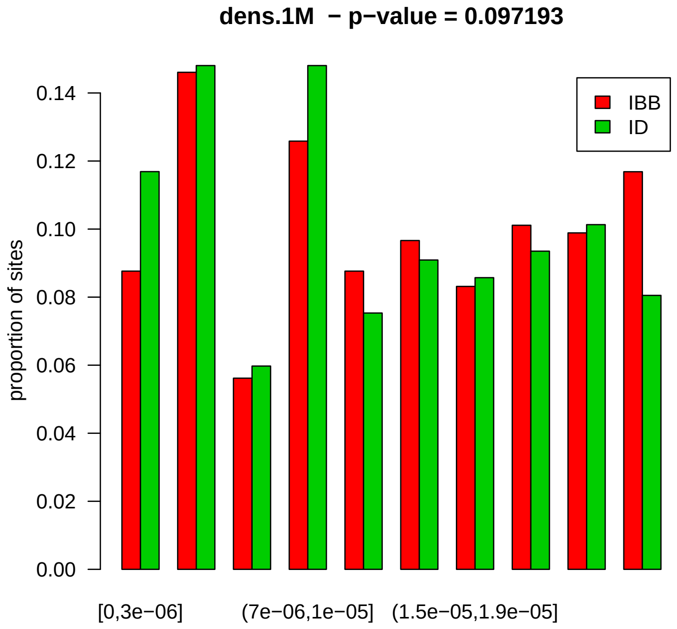


Here the effect of density of CpG islands is studied:

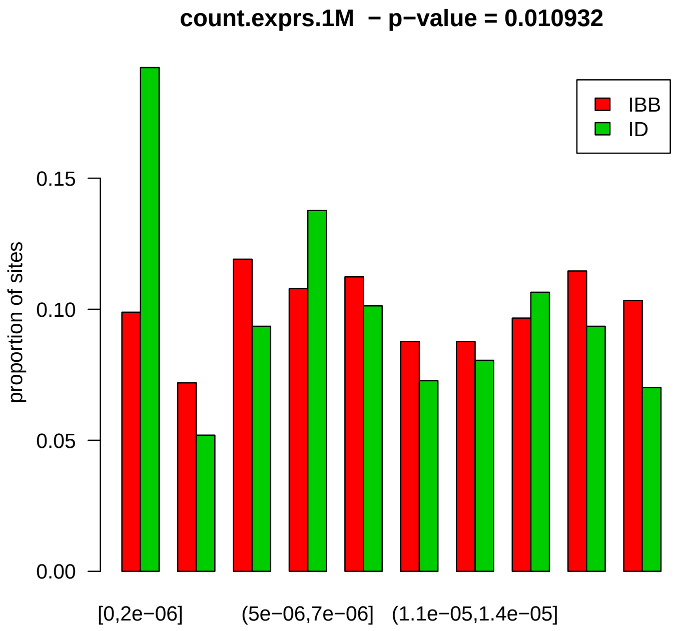


4.6 1 megaBase Window

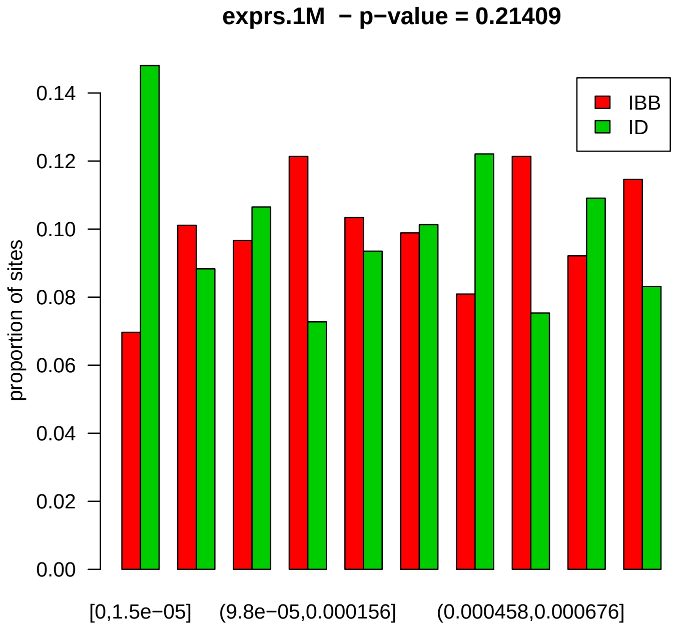
First, we see gene density:



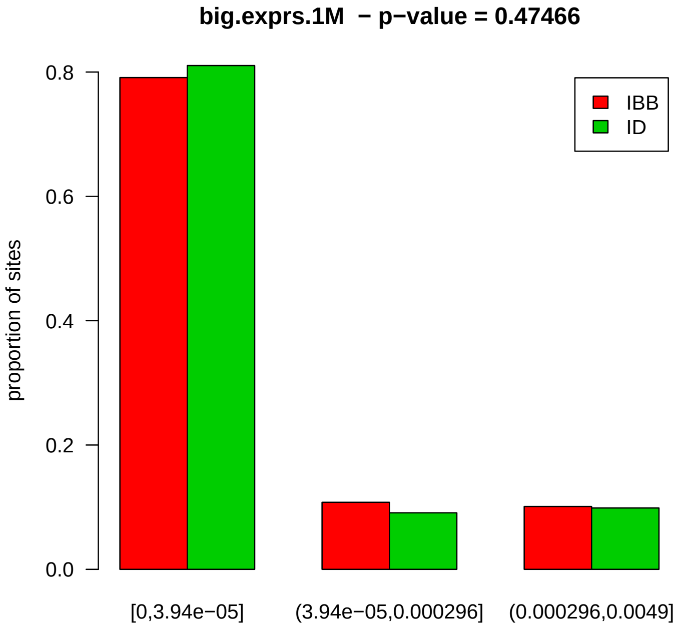
Here are the results for EST density. First, we count just one EST per gene.



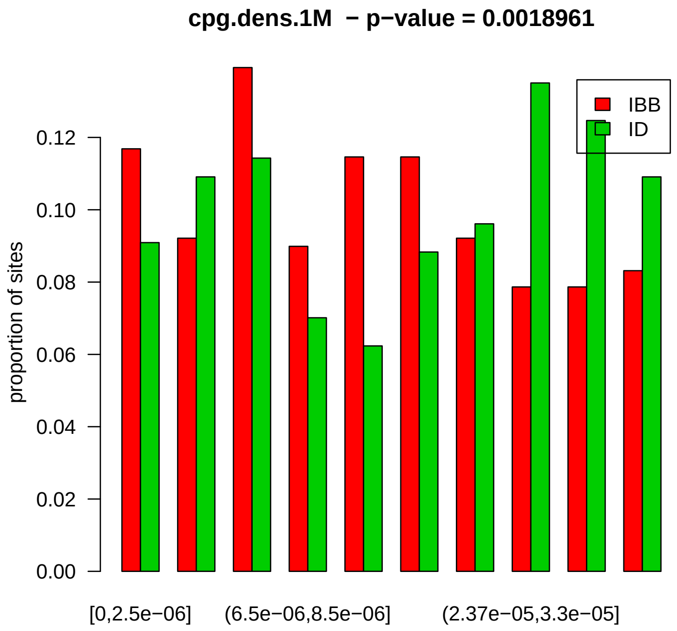
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene

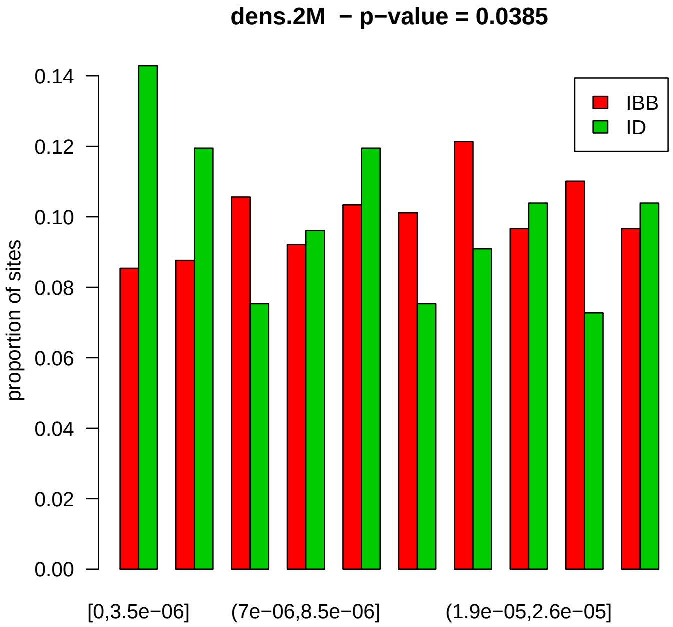


Here the effect of density of CpG islands is studied:

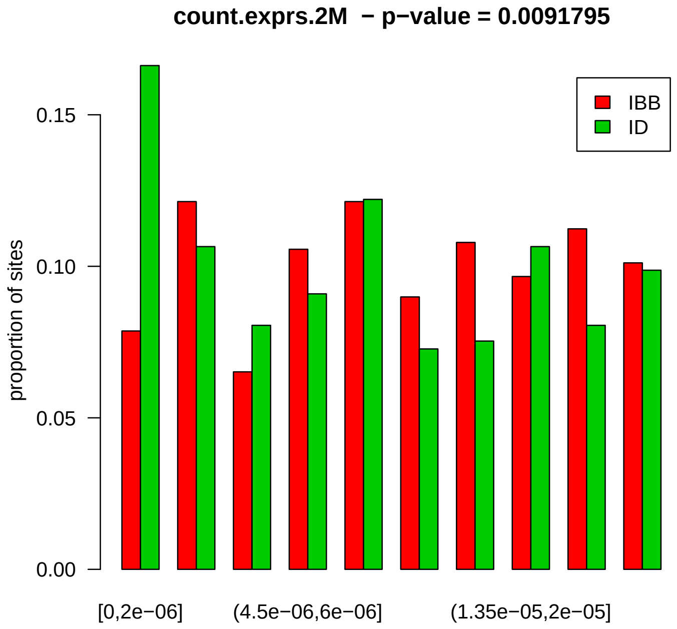


4.7 2 megaBase Window

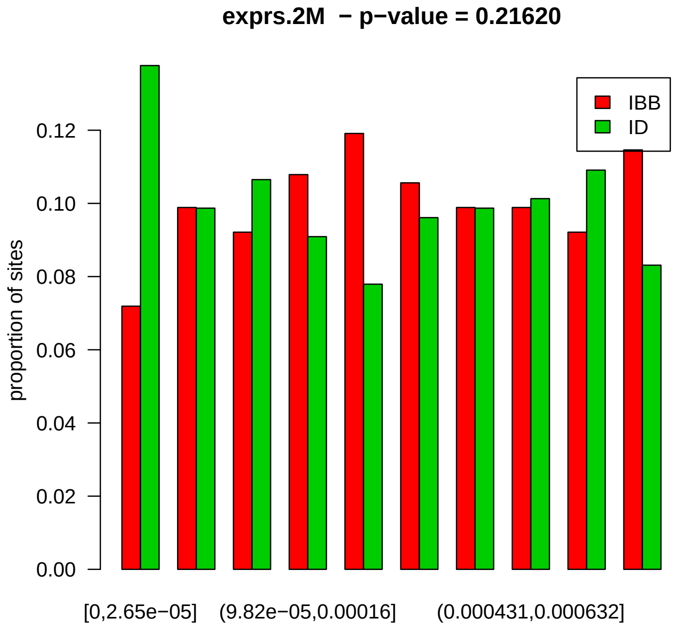
First, we see gene density:



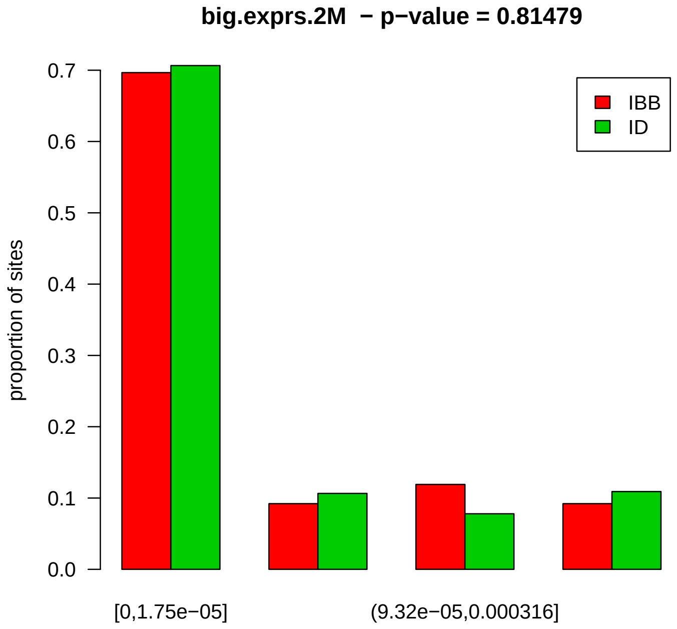
Here are the results for EST density. First, we count just one EST per gene.



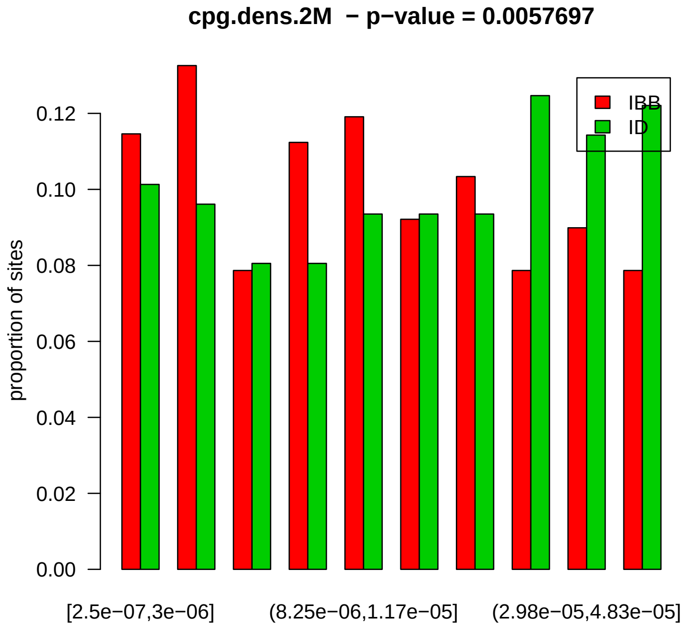
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene

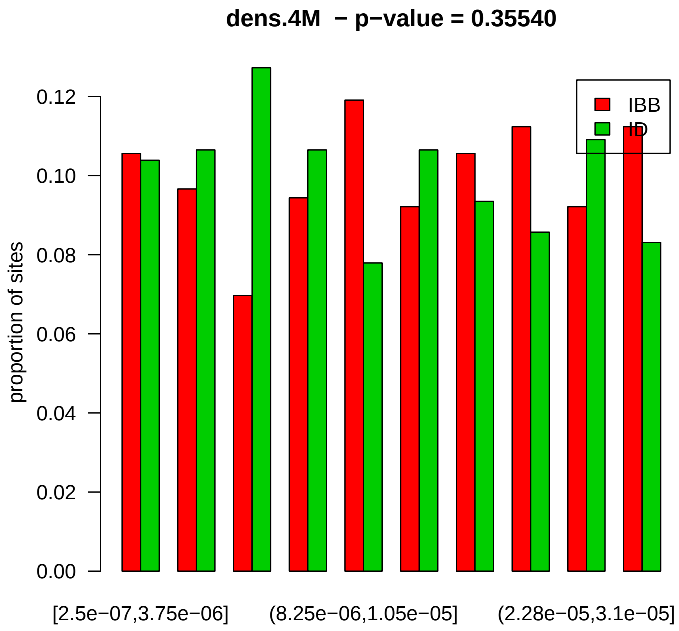


Here the effect of density of CpG islands is studied:

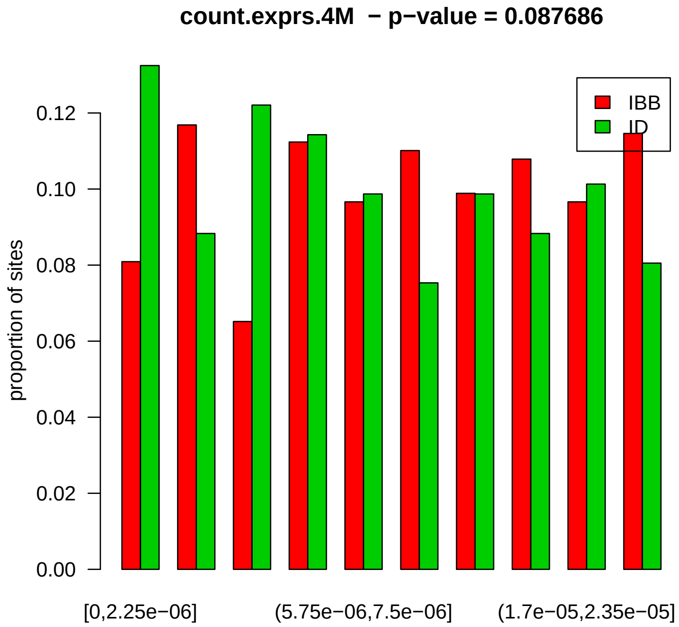


4.8 4 megaBase Window

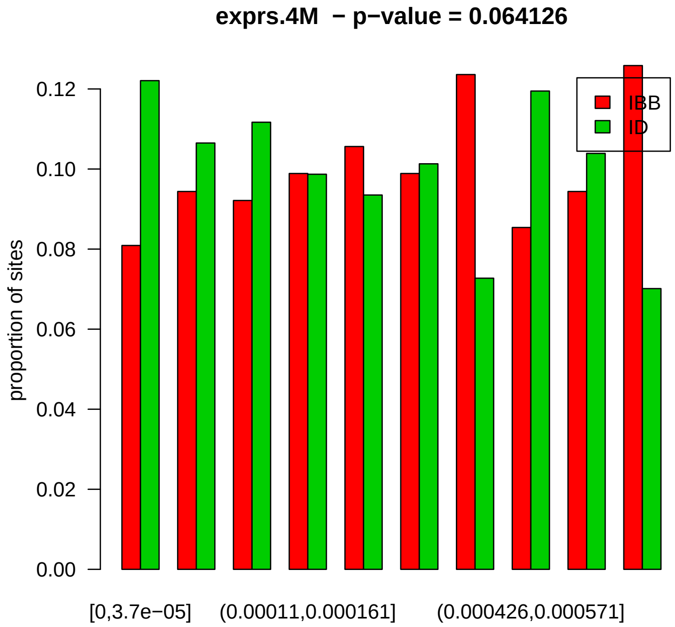
First, we see gene density:



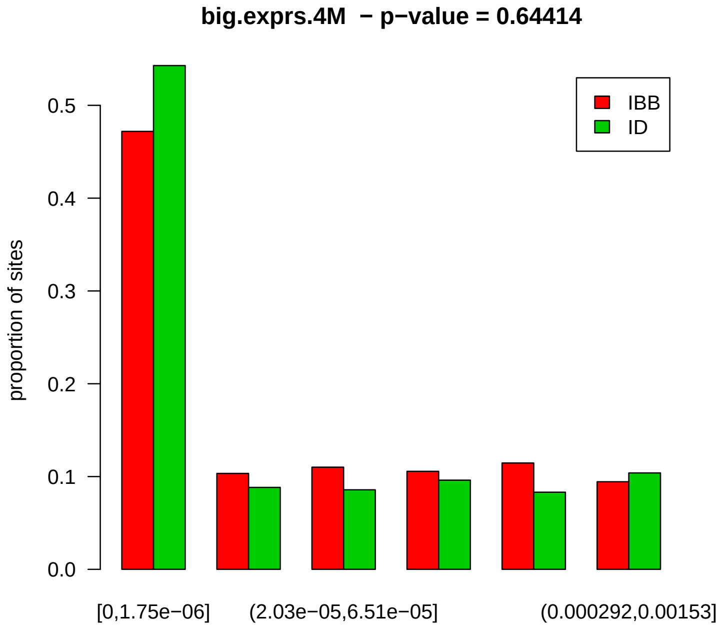
Here are the results for EST density. First, we count just one EST per gene.



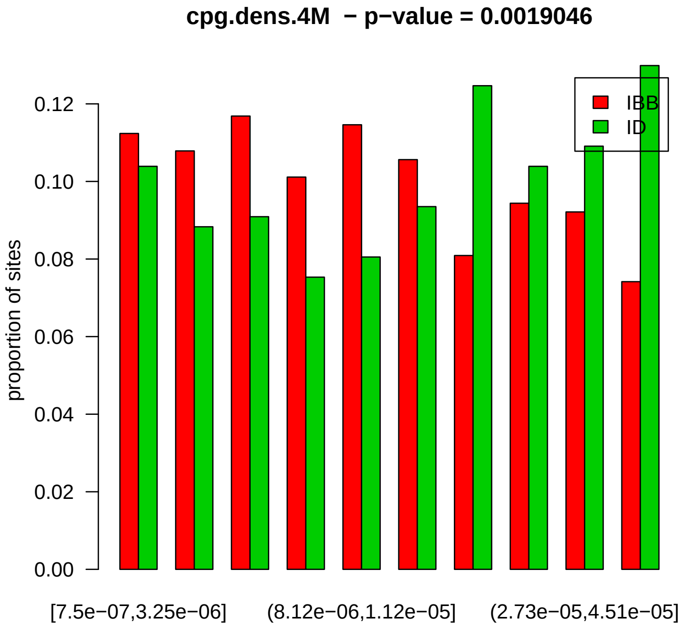
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene

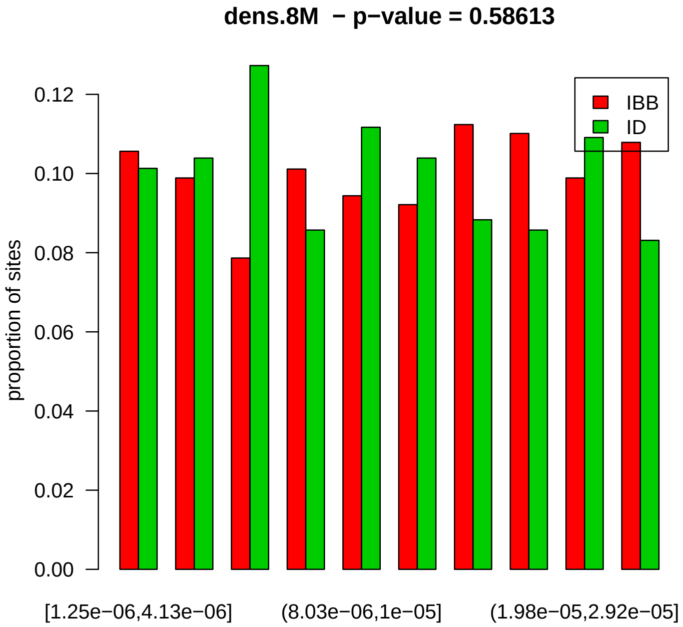


Here the effect of density of CpG islands is studied:

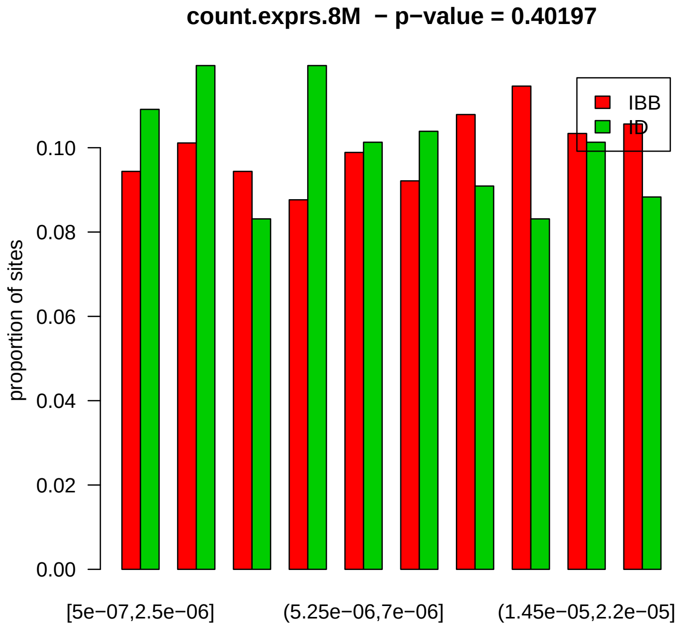


4.9 4 megaBase Window

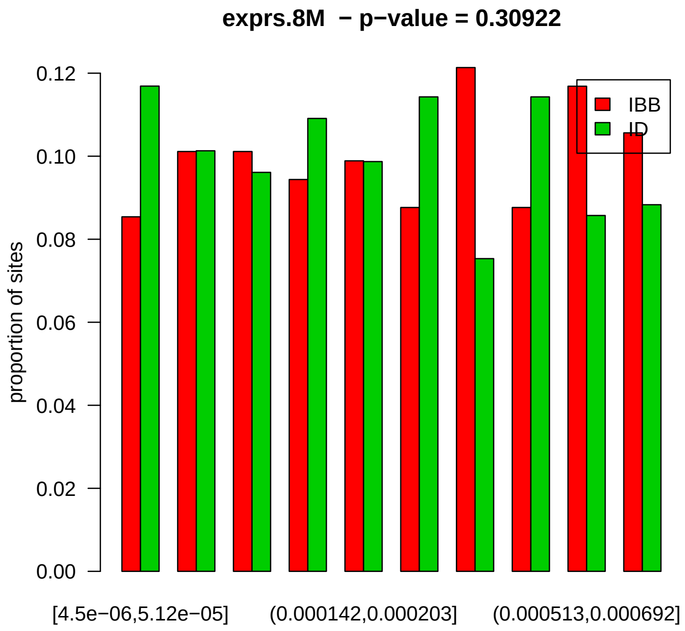
First, we see gene density:



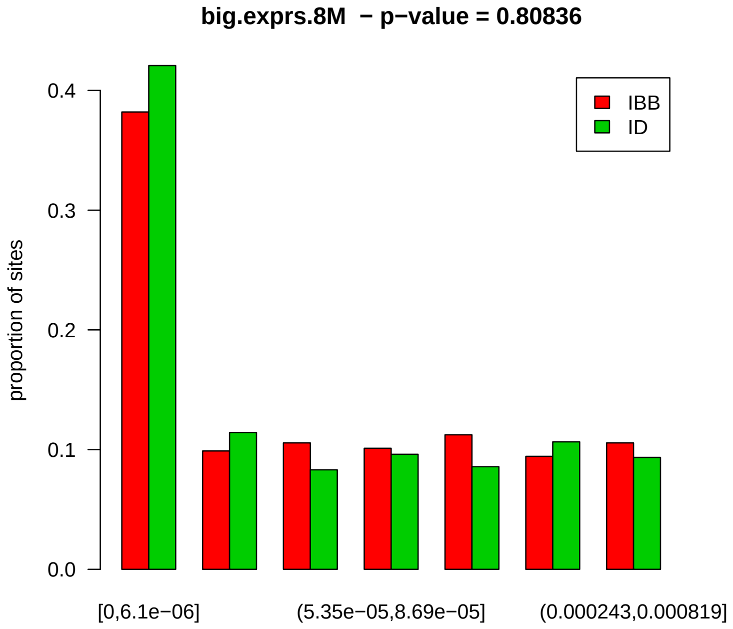
Here are the results for EST density. First, we count just one EST per gene.



Now we count up to 200 ESTs per gene:

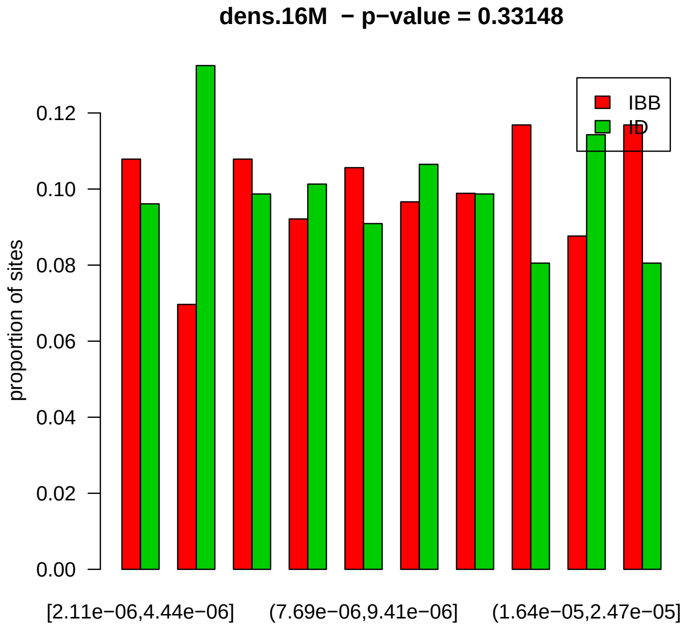


And here counting starts only after 200 ESTs per gene

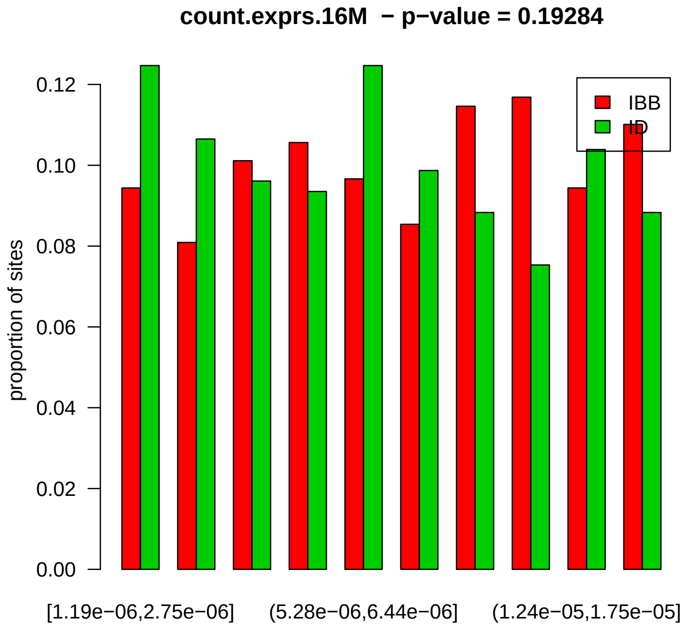


4.10 16 megaBase Window

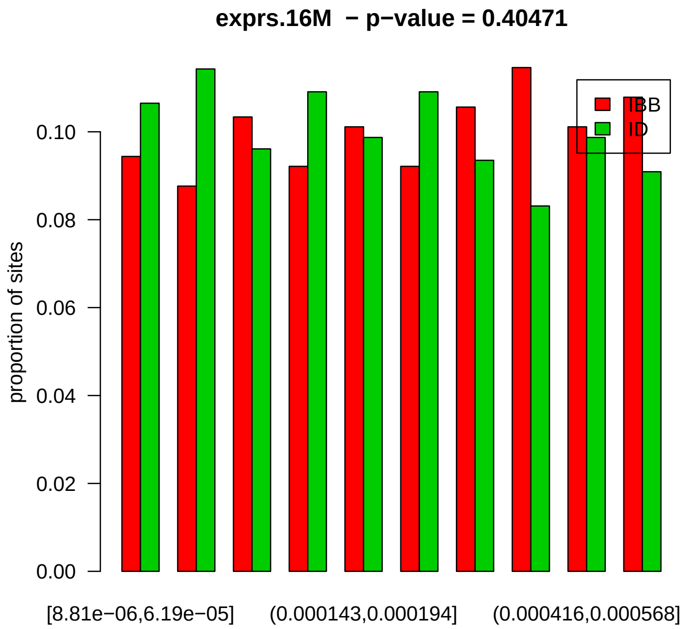
First, we see gene density:



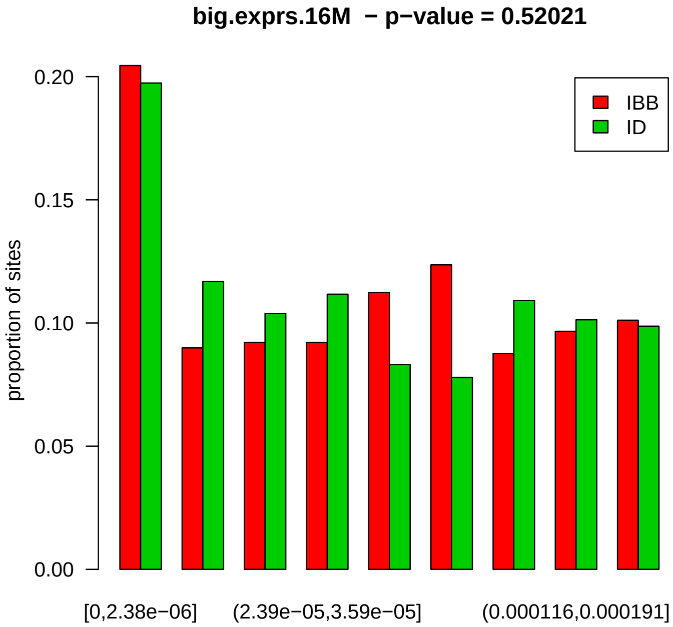
Here are the results for EST density. First, we count just one EST per gene.



Now we count up to 200 ESTs per gene:

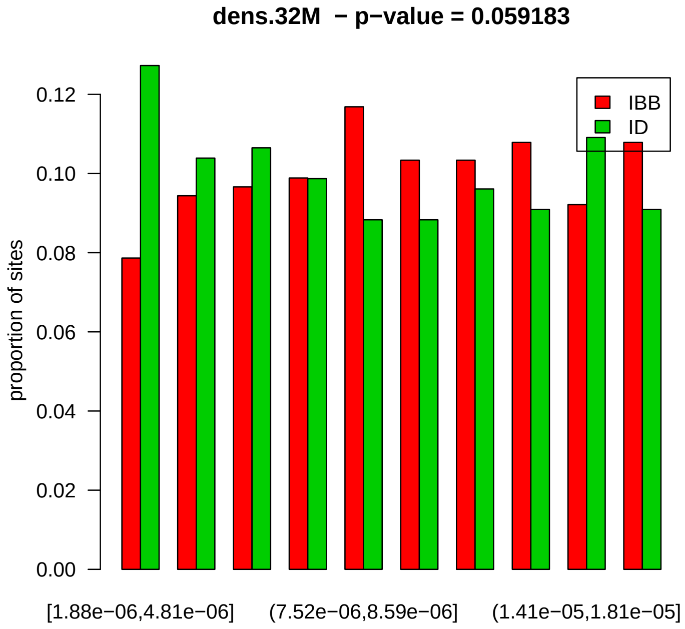


And here counting starts only after 200 ESTs per gene

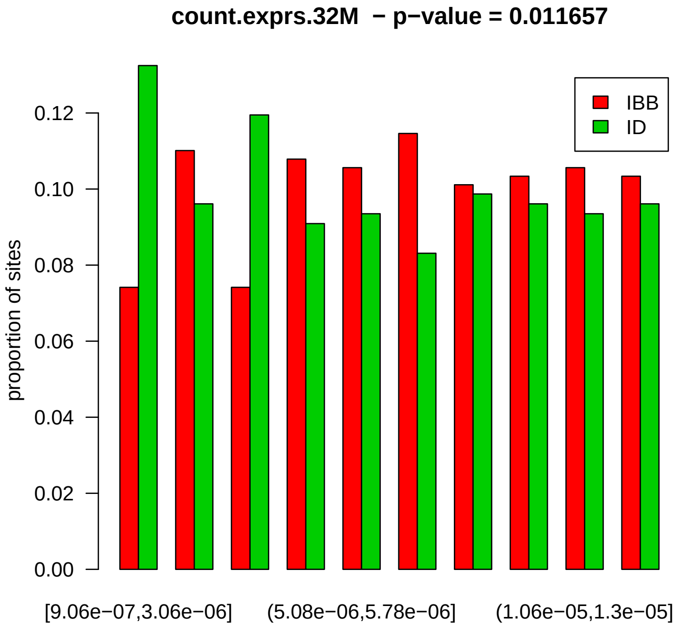


4.11 32 megaBase Window

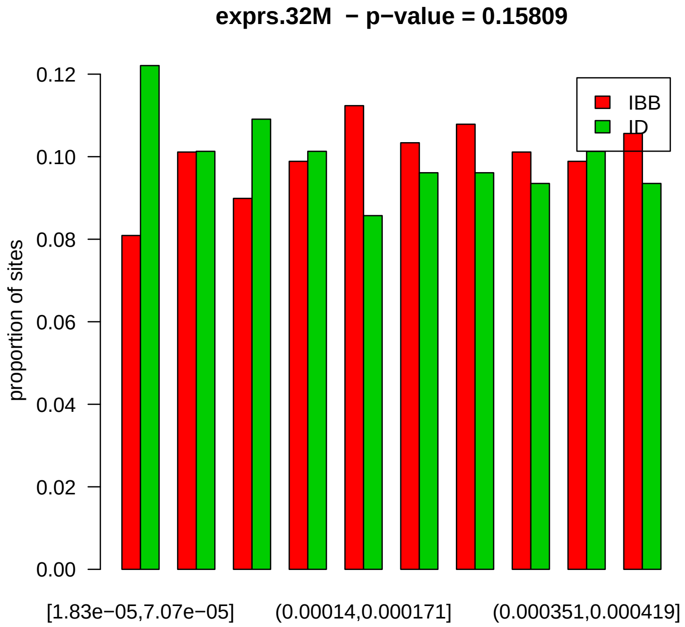
First, we see gene density:



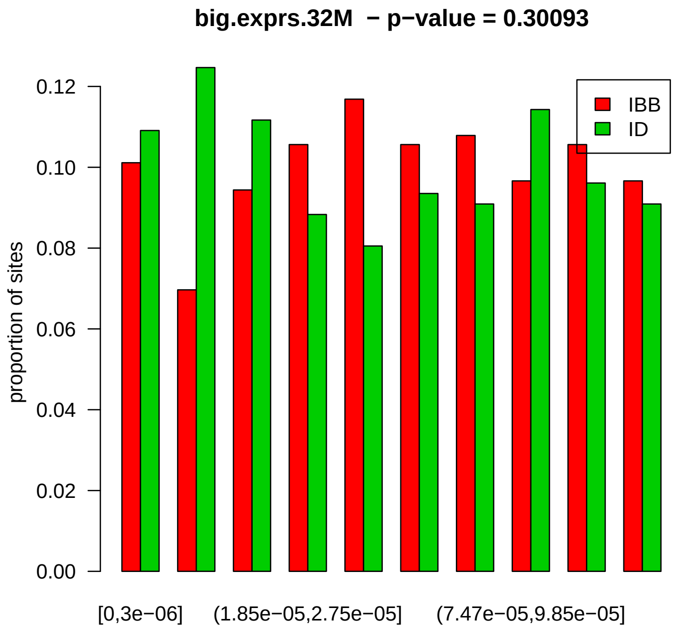
Here are the results for EST density. First, we count just one EST per gene.



Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene

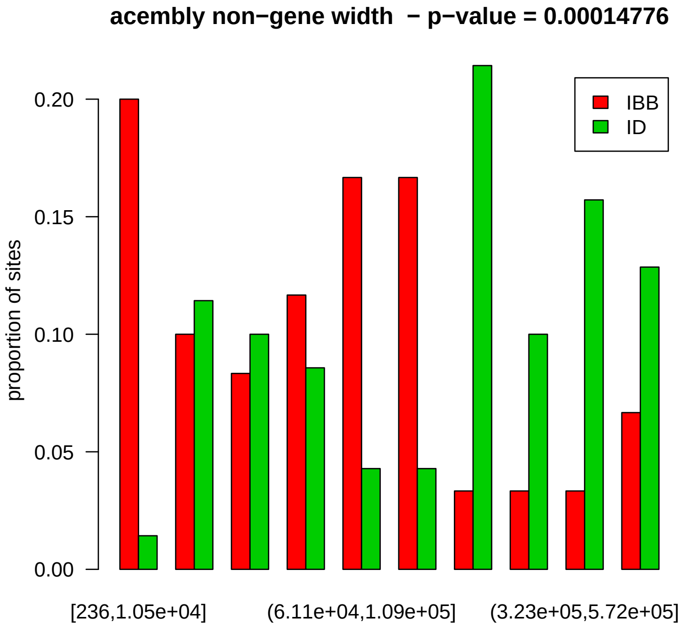


5 Juxtaposition with Gene Start and End Positions

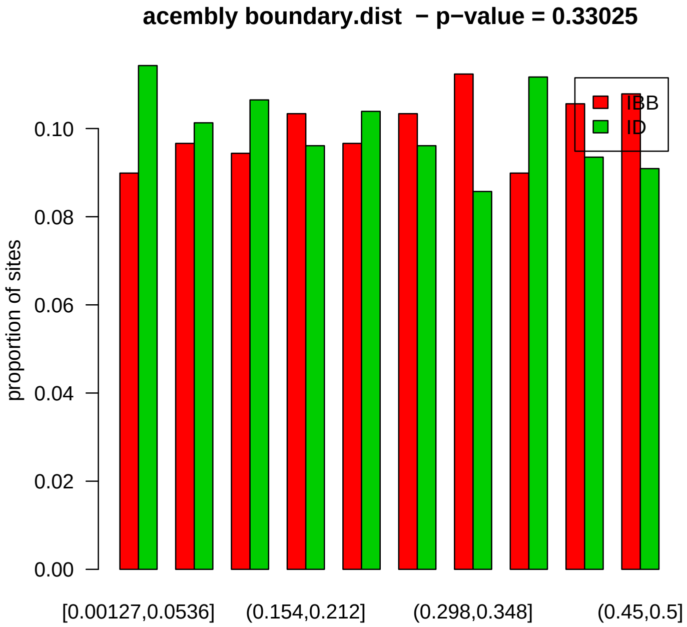
5.1 Acembly Annotations

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

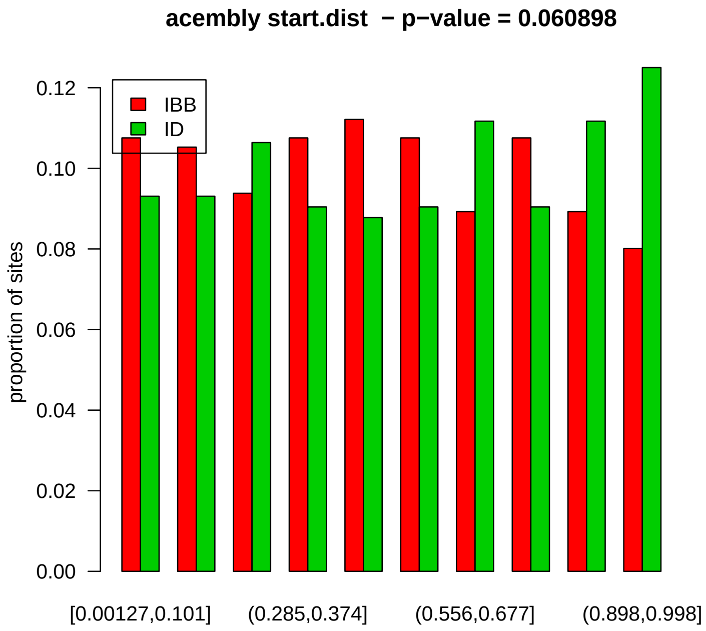
The next plot uses the width of a non-gene region for insertions that fall into such regions.



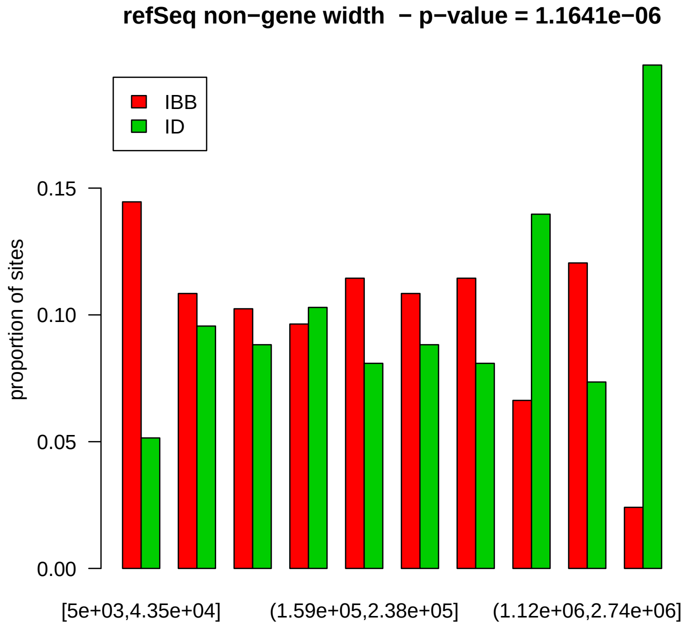
The next plot studies the distance to the nearest boundary between a gene and a non-gene region. The distance is expressed as a fraction of the length of the region. Thus, '0.25' refers to one quarter of the distance from the site to nearest boundary divided by the total width of the region.

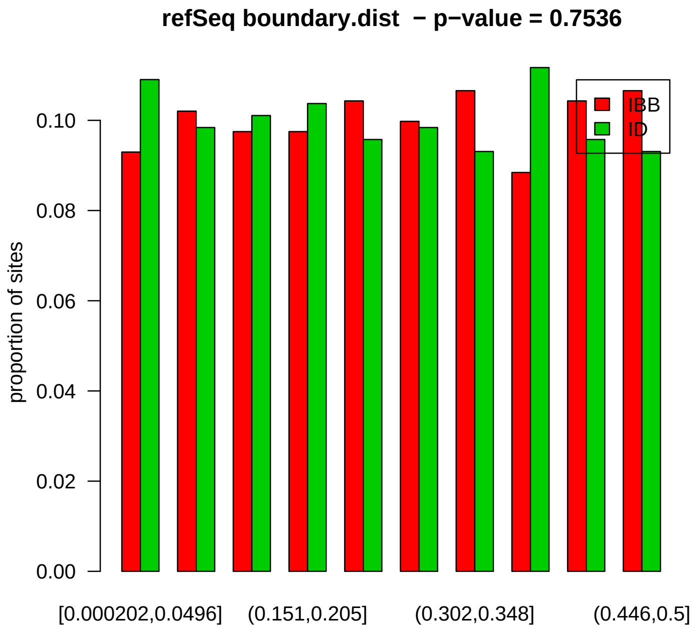


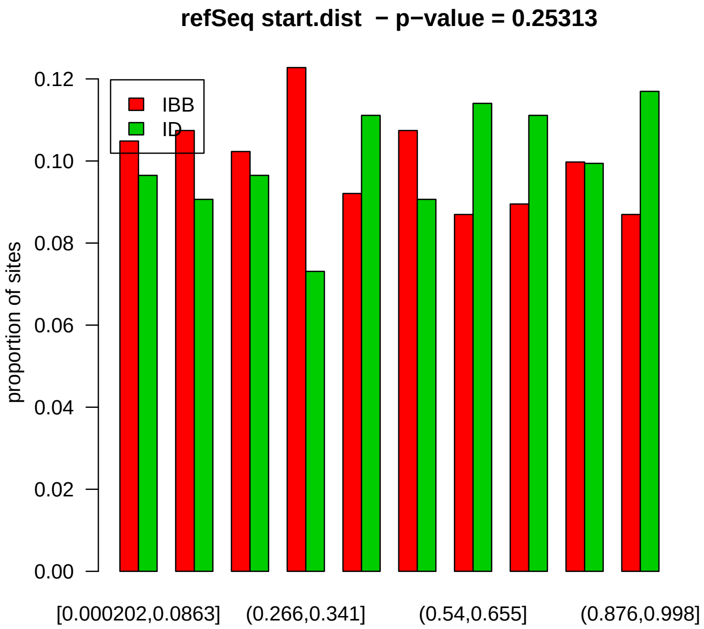
This plot studies the effect of nearness to the beginning of a transcript. For sites in genes, it is the distance to the start of the gene divided by the width of the gene. For other sites it is the distance from the site to the nearer gene if that gene boundary is also a transcription starting point.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


acembly start.dist − p−value = 0.060898
IBB
ID
proportion of sites
0.00
0.02
0.04
0.06
0.08
0.10
0.12
[0.00127,0.101]
(0.285,0.374]
(0.556,0.677]
(0.898,0.998]

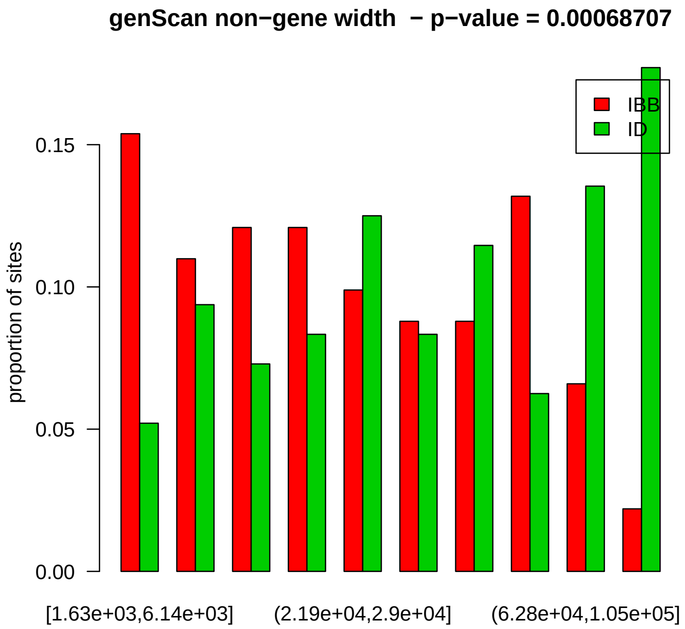
5.2 RefSeq Annotations

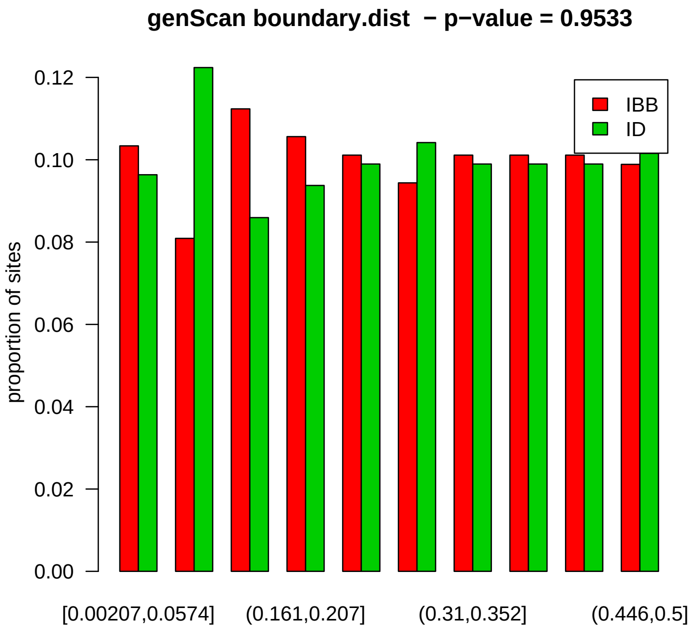


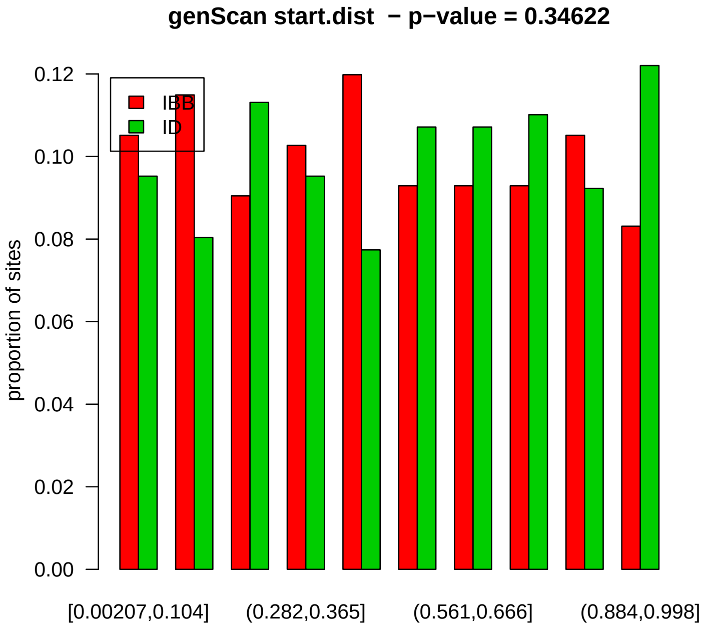

refSeq boundary.dist − p−value = 0.7536
proportion of sites
0.00
0.02
0.04
0.06
0.08
0.10
IBB
ID
[0.000202,0.0496]
(0.151,0.205]
(0.302,0.348]
(0.446,0.5]


refSeq start.dist − p−value = 0.25313
🟥 IBB
🟩 ID
proportion of sites
0.12
0.10
0.08
0.06
0.04
0.02
0.00
[0.000202,0.0863]
(0.266,0.341]
(0.54,0.655]
(0.876,0.998]

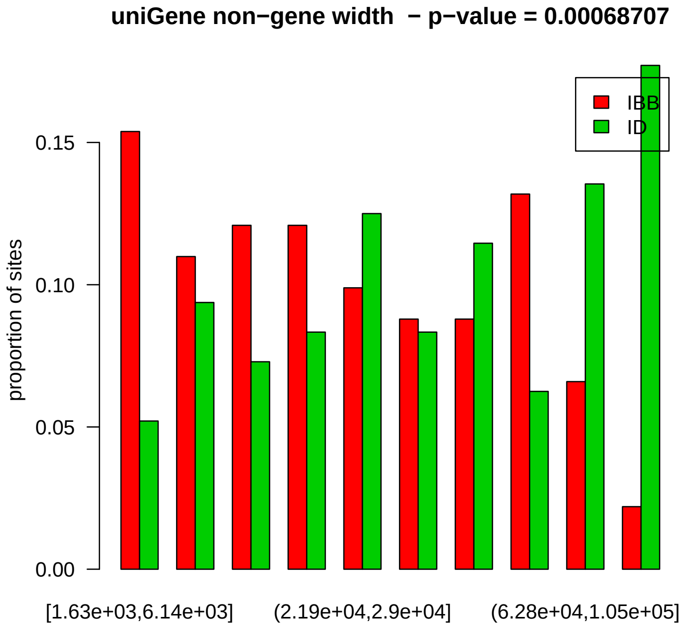
5.3 genScan Annotations

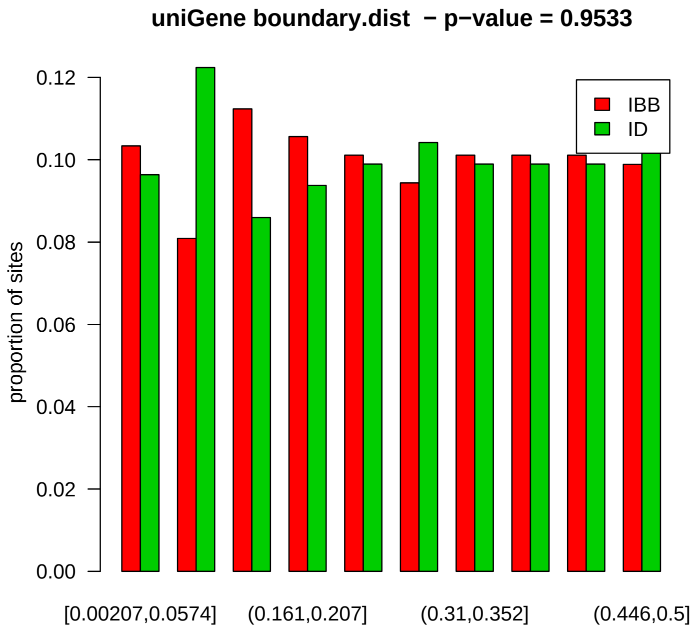


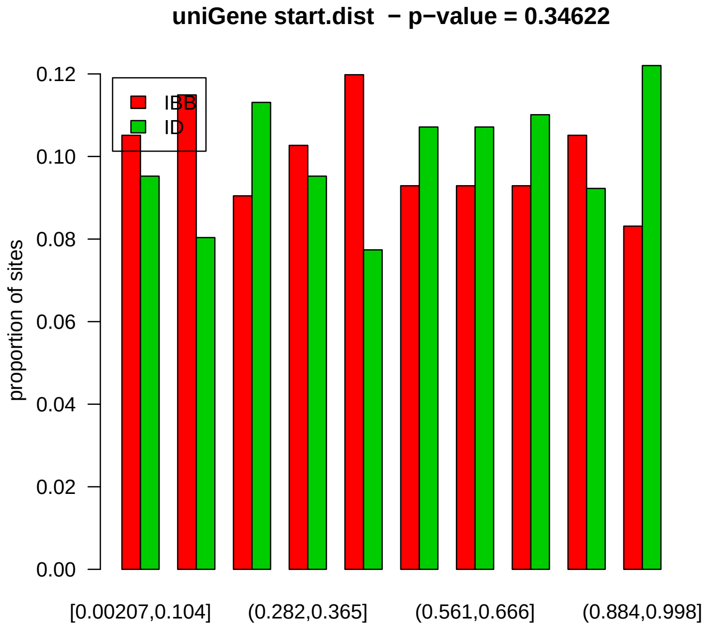

genScan boundary.dist − p−value = 0.9533
proportion of sites
0.12
0.10
0.08
0.06
0.04
0.02
0.00
🟥 IBB
🟩 ID
[0.00207,0.0574]
(0.161,0.207]
(0.31,0.352]
(0.446,0.5]


genScan start.dist − p−value = 0.34622
IBB
ID
proportion of sites
0.00
0.02
0.04
0.06
0.08
0.10
0.12
[0.00207,0.104]
(0.282,0.365]
(0.561,0.666]
(0.884,0.998]

5.4 uniGene Annotations



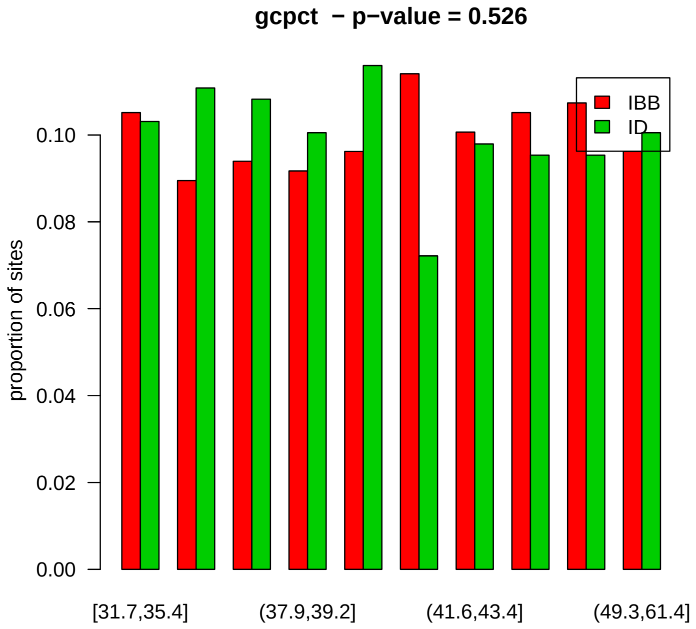

uniGene boundary.dist − p−value = 0.9533
0.12
0.10
0.08
0.06
0.04
0.02
0.00
proportion of sites
🟥 IBB
🟩 ID
[0.00207,0.0574]
(0.161,0.207]
(0.31,0.352]
(0.446,0.5]



6 GC content

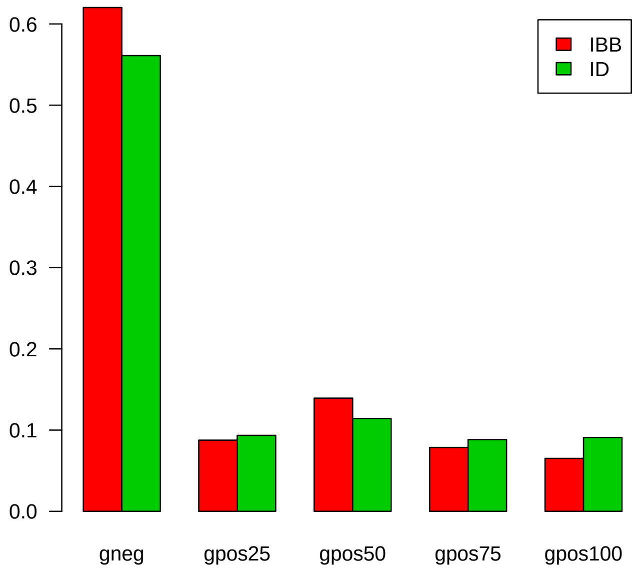
Here we study the effect of GC content on insertion. The GC content is taken from the Human Genome Draft at GoldenPath from the table `http://genome.ucsc.edu/goldenPath/14nov2002/database/gcPercent.txt.gz`.

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



7 Cytobands

Here we study the association of cytoband with insertion intensity. The data are obtained from
`http://genome.ucsc.edu/goldenPath/14nov2002/database/cytoBand.txt.gz`.



A formal test of significance attains a p-value of 0.41588.

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] Rogier. Versteeg, Barbera. D. C. van Schaik., Marinus. F. van Batenburg., Marco. Roos, Ramin. Monajemi, Huib. Caron, Harmen. J. Bussemaker, and Antoine. H. C. van Kampen. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res*, 13(9):1998–2004, Sep 2003.

[4] Xiaolin. Wu, Yuan. Li, Bruce. Crise, and Shawn. M. Burgess. Transcription start regions in the human genome are favored targets for MLV integration. *Science*, 300(5626):1749–1751, Jun 2003.