Report 8C: Contact and expression

Cyril Matthey-Doret

29 novembre 2016

Introduction:

This is a shortened, improved version of report 8B. patch note: * Only using TAD-TAD contacts * Only using 3 sets of genes for both lincRNAs and PCG: enp, nep and nenp * Improved some plots

Here, I investigate the relationship between gene expression and DNA-DNA contact across 4 different cell lines (GM12878, K562, HUVEC and NHEK). This analysis is performed across all LCL expressed lincRNAs and protein-coding genes, categorized by according to their overlap with enhancers and promoters. Contact per gene was calculated in 2 ways: Gene versus chromosome conmputes all interactions between the gene and the chromosome, while gene versus TAD only takes interactions between the gene and the TAD it belongs to.

NOTE 1: At the moment I only have expression data for GM12878, but I will add the other cell lines later. NOTE 2: Contact matrices for chromosome 9 for all cell lines have been normalised using SQRTVC instead of KR because the algorithm did not converge for chromosome 9 in K562 and vector was full of NAs.
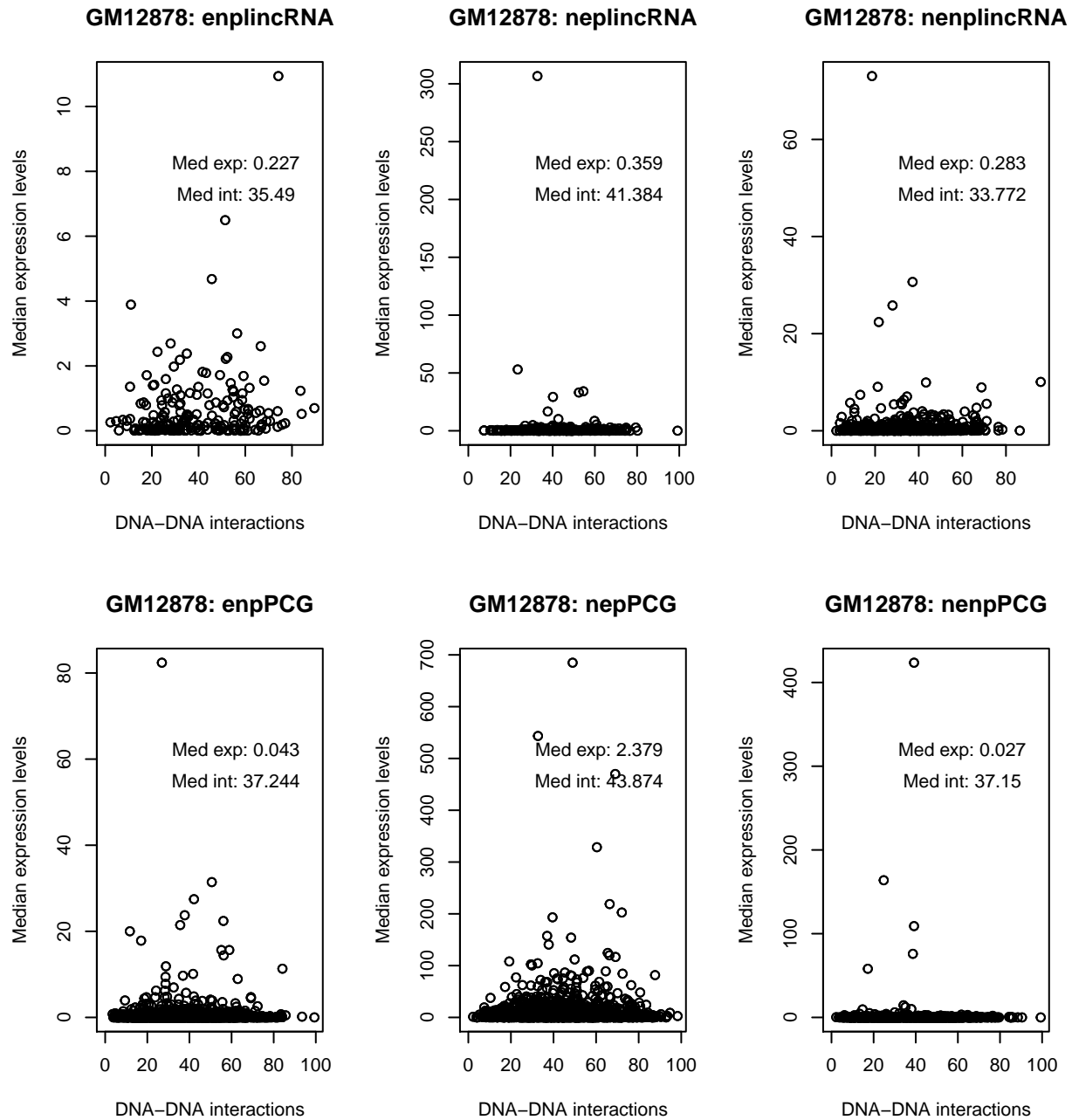
Results:

Overview of data:

A quick overview of the data shows that enhancer-associated genes have a lower expression and higher amounts of DNA contacts on average. Patterns in the data would be easier to detect after log transforming both variables.
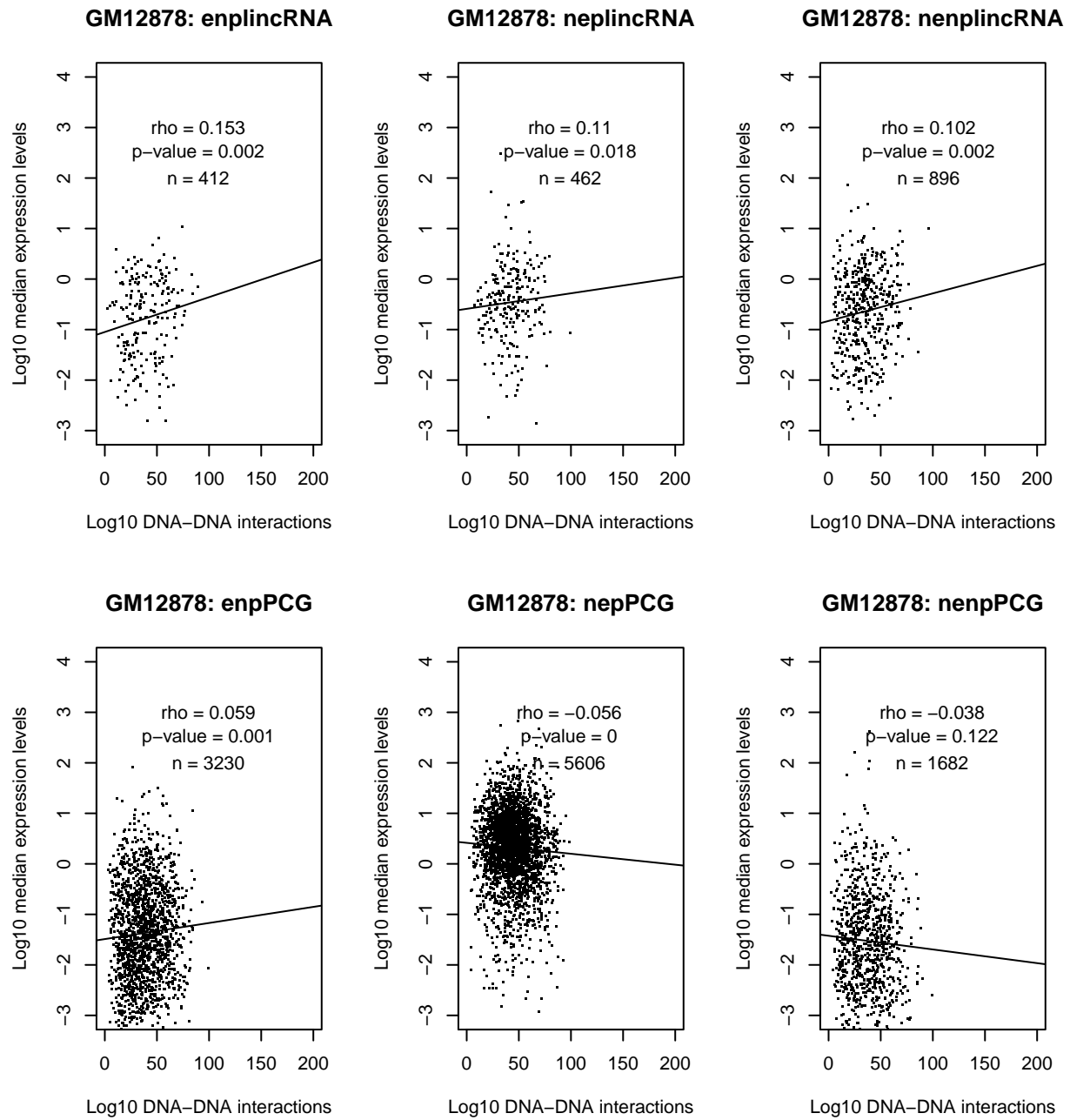
Mean TAD contacts:

Here, the contact value of each gene is the mean contacts observed for the TAD it is in.

Correlations:

Spearman correlation was used to detect non-linear relationship between DNA contact and expression. The data was log-transformed for visualization.

Mean TAD contacts:

GM12878: enplincRNA

rho = 0.153
p-value = 0.002
n = 412

Log10 median expression levels

Log10 DNA−DNA interactions

GM12878: neplincRNA

rho = 0.11
p-value = 0.018
n = 462

Log10 median expression levels

Log10 DNA−DNA interactions

GM12878: nenplincRNA

rho = 0.102
p-value = 0.002
n = 896

Log10 median expression levels

Log10 DNA−DNA interactions

GM12878: enpPCG

rho = 0.059
p-value = 0.001
n = 3230

Log10 median expression levels

Log10 DNA−DNA interactions

GM12878: nepPCG

rho = −0.056
p-value = 0
n = 5606

Log10 median expression levels

Log10 DNA−DNA interactions

GM12878: nenpPCG

rho = −0.038
p-value = 0.122
n = 1682

Log10 median expression levels

Log10 DNA−DNA interactions

3

Stats for all cell lines

Correlations between TAD_TAD contact and expression. All values are obtained with Spearman correlation tests.

|	cell_line	gentype	rho	p.value
4	HUVEC	enplincRNA	-0.0062	0.9172
10	NHEK	enplincRNA	-0.1003	0.075
11	NHEK	neplincRNA	0.0242	0.6108
7	K562	enplincRNA	0.0461	0.4138
12	NHEK	nenplincRNA	0.0746	0.0228
9	K562	nenplincRNA	0.0975	0.0025
3	GM12878	nenplincRNA	0.1021	0.0022
2	GM12878	neplincRNA	0.1096	0.0184
6	HUVEC	nenplincRNA	0.1295	1e-04
8	K562	neplincRNA	0.1523	0.0011
5	HUVEC	neplincRNA	0.1529	0.0013
1	GM12878	enplincRNA	0.153	0.0018

Comparisons across cell lines.

Contact

Comparing TAD-TAD contact across all cell lines. Metrics is TAD_TAD contacts.
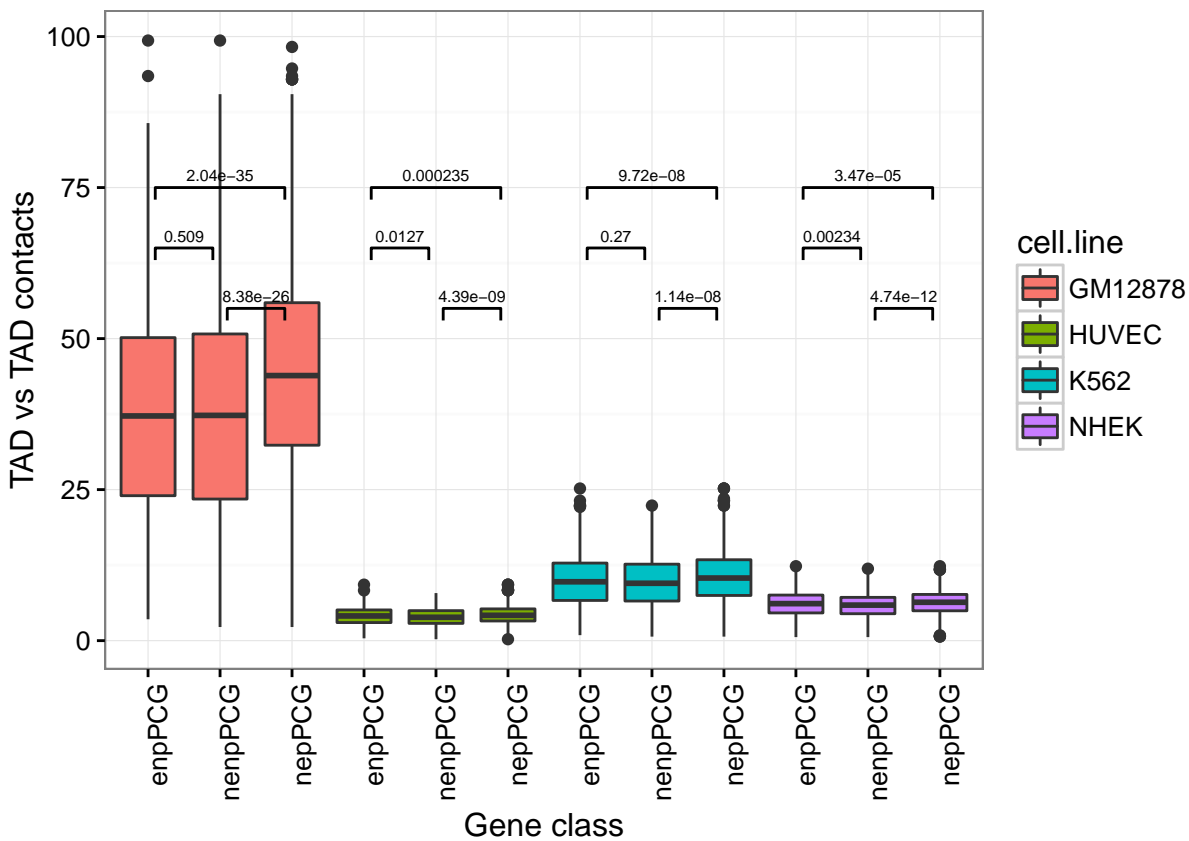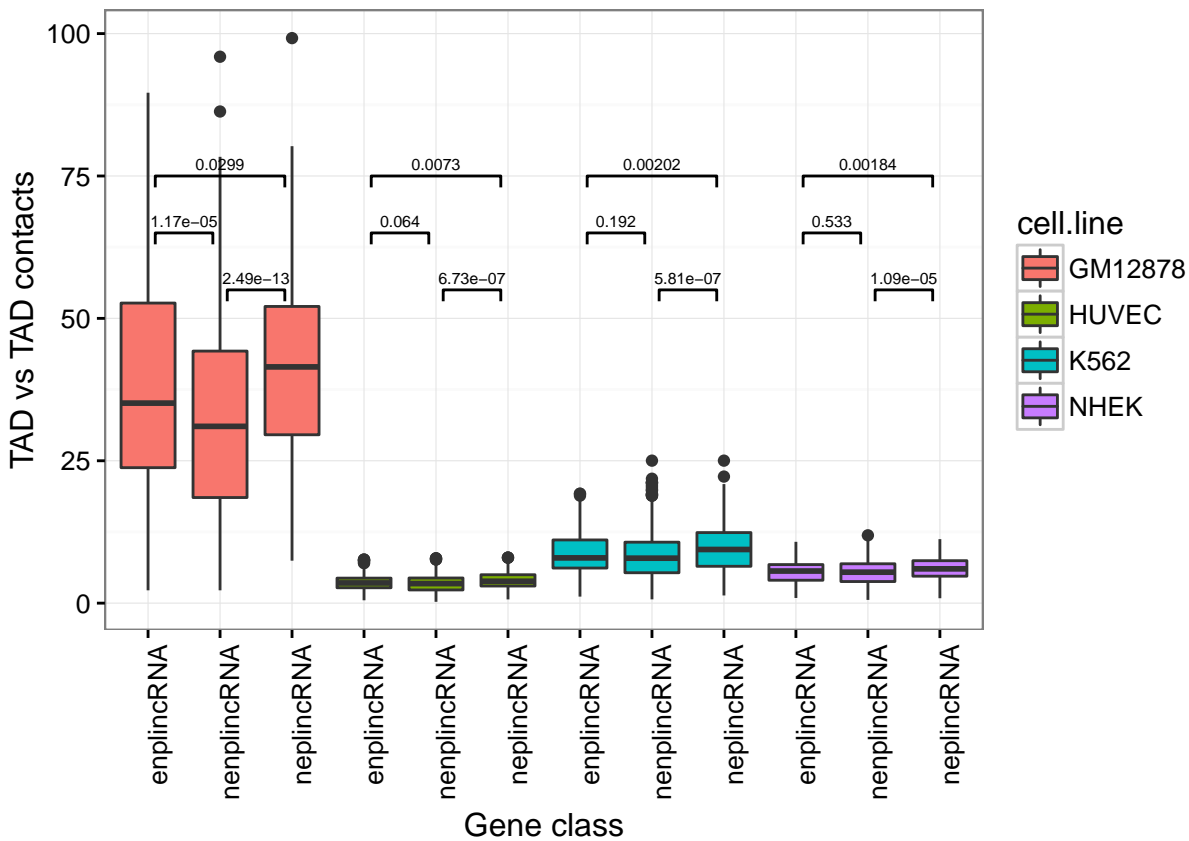
In HUVEC and K562, for lincRNAs, the contacts only differ when looking at comparisons between nep and nenp, indicating there is no difference between enhancer bound and non-enhancer bound. In GM12878 however, we also observe a significant difference between enp and nenp. In NHEK, there is also a significant difference between enp and nep, but not between enp and nenp.
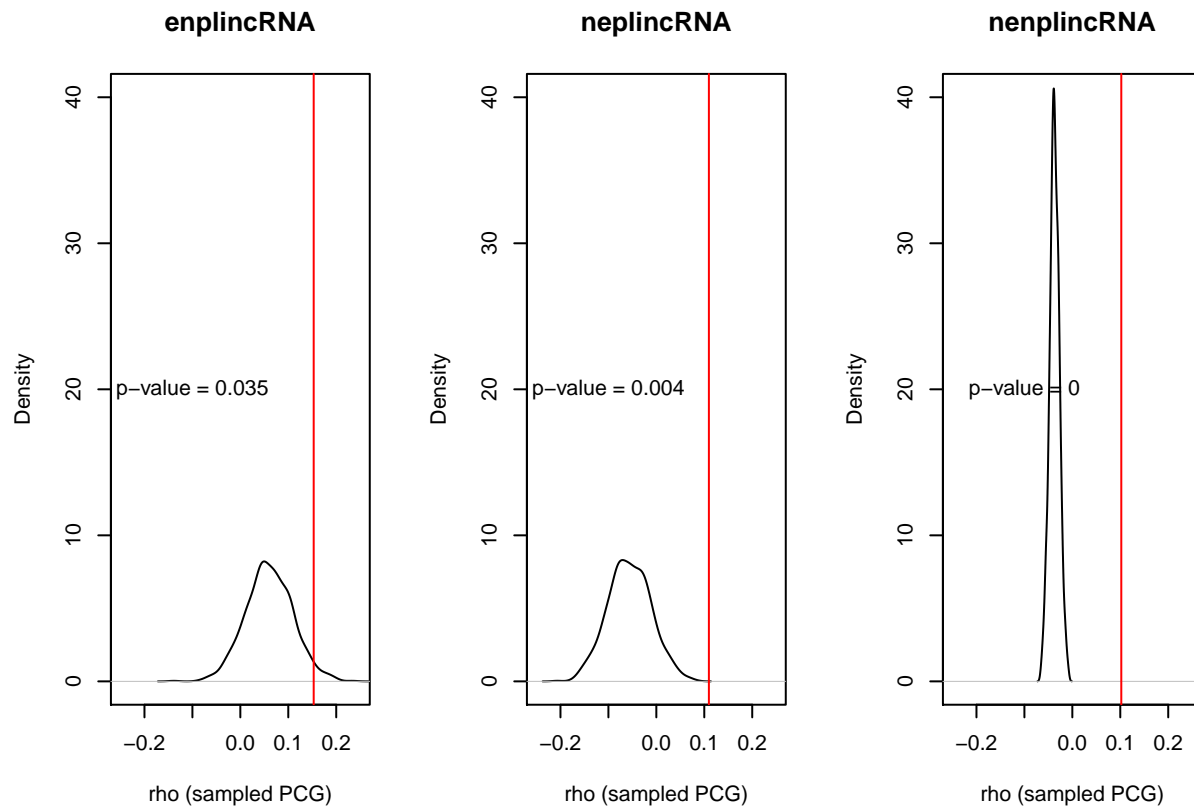
Expression

Same plots for expression across cell lines.
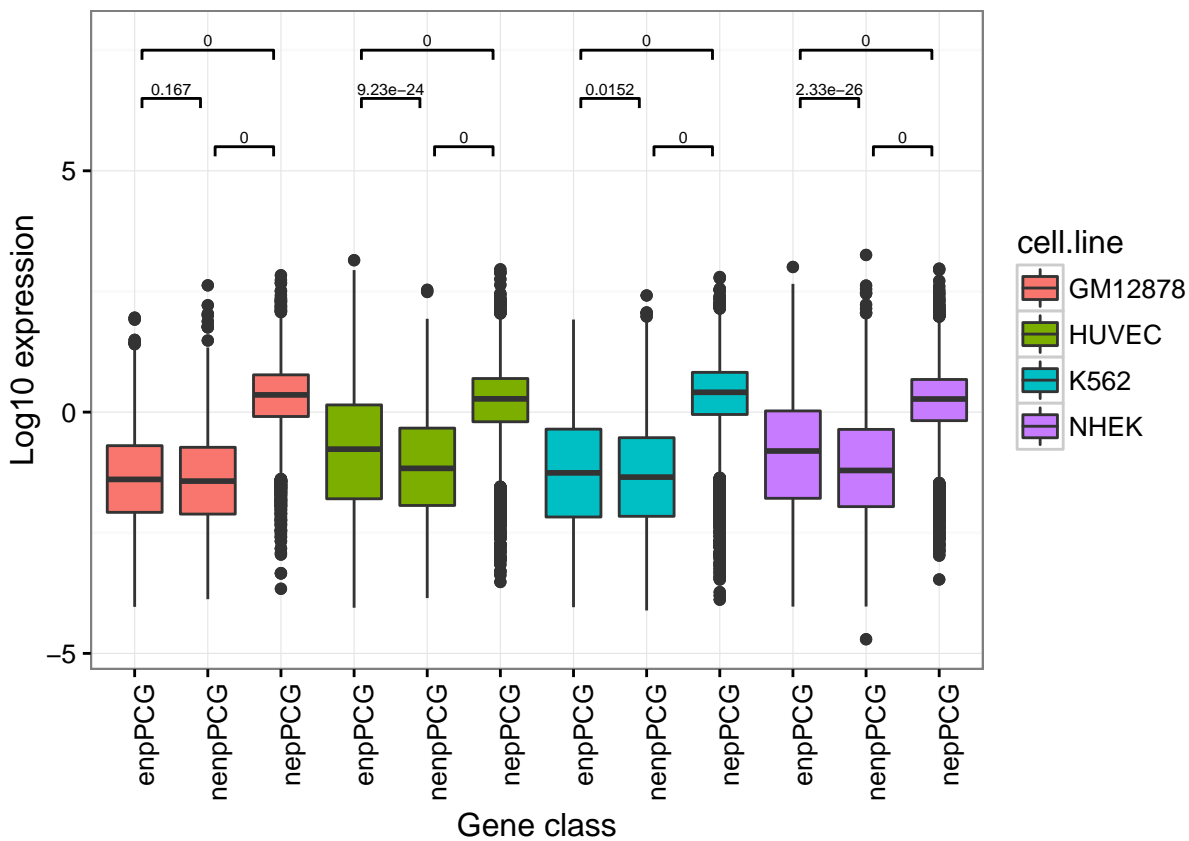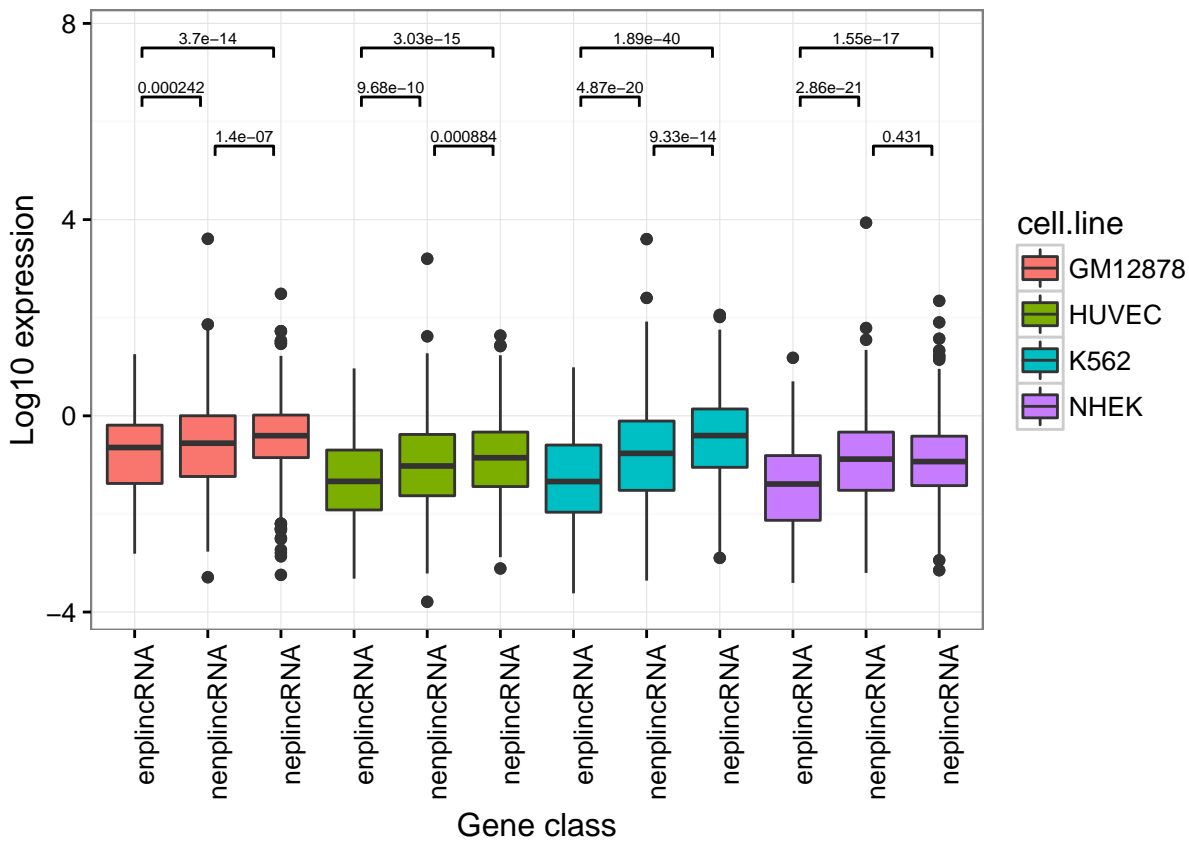
Random sampling comparisons

Here, I compare the contact-expression correlation coefficient of lincRNAs with randomly sampled PCG from the same category. I used 1000 simulations for each comparison and the number of sampled PCG in each simulation is equal to the number of lincRNAs comprised in the matching category. The red line corresponds to the rho observed in lincRNAs, while the densities are the rhos from sampled PCG. All correlation estimates are calculated using Spearman method.

Conclusions:

Correlations between contact and expression seem to be very weak and should be interpreted carefully. The striking differences in contact across cell lines is surprising. Due to experimental reasons ? normalization problem ?

The effect of promoter/enhancer overlap on contact in lincRNAs is also cell line dependant. Maybe linked to the nature of the cell line ? (GM12878=transformed lymphoblast, K562=cancer mesoderm, NHEK=normal skin, HUVEC=umbilical cells)

According to the random sampling comparison, the correlation between expression and contact is higher than would be expected in protein-coding genes for all categories of non-enhancer associated lincRNAs, but never for enhancer-associated lincRNAs.