Report 8B: Contact and expression

Cyril Matthey-Doret

19 novembre 2016

Introduction:

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.

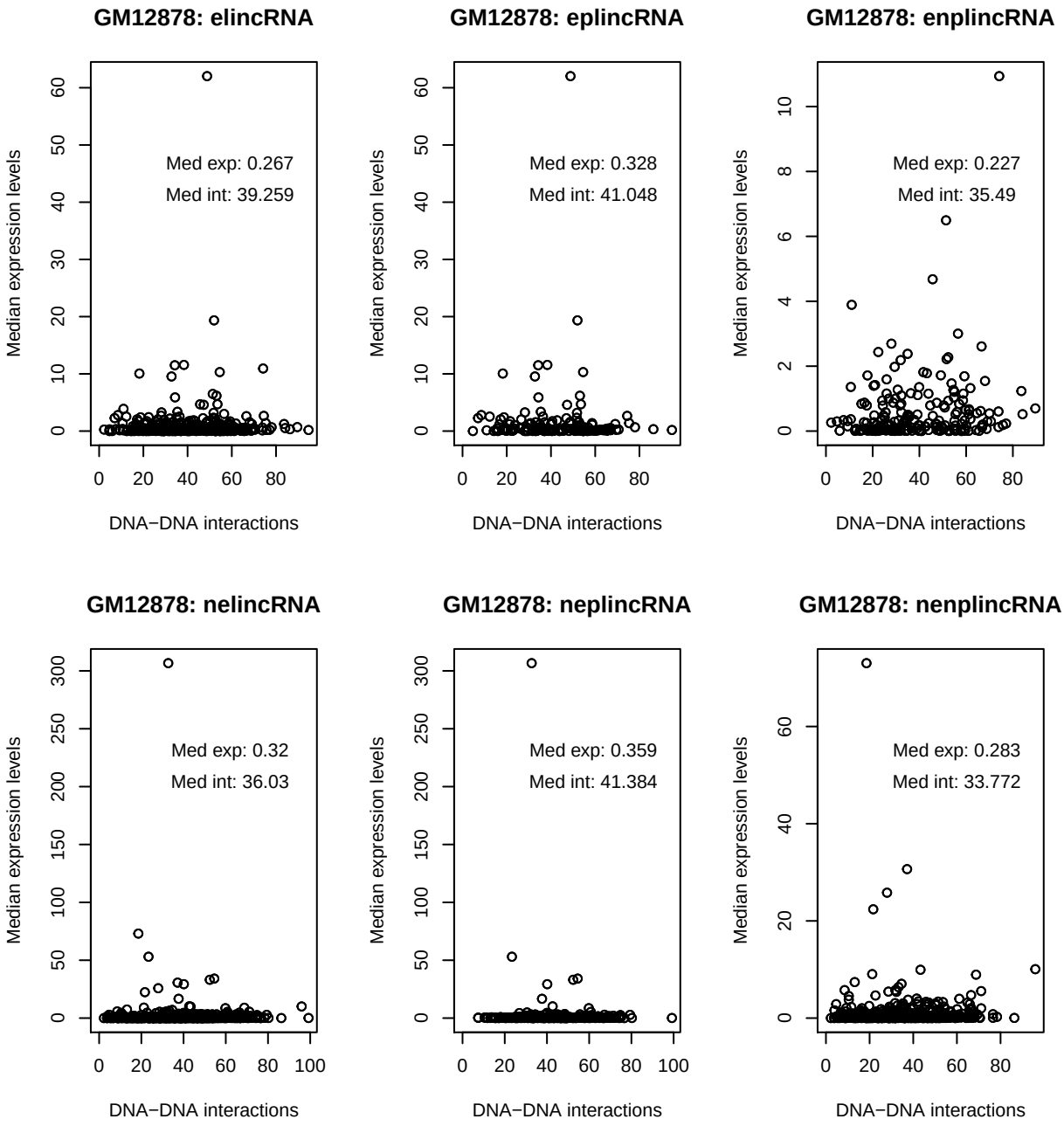
Results:

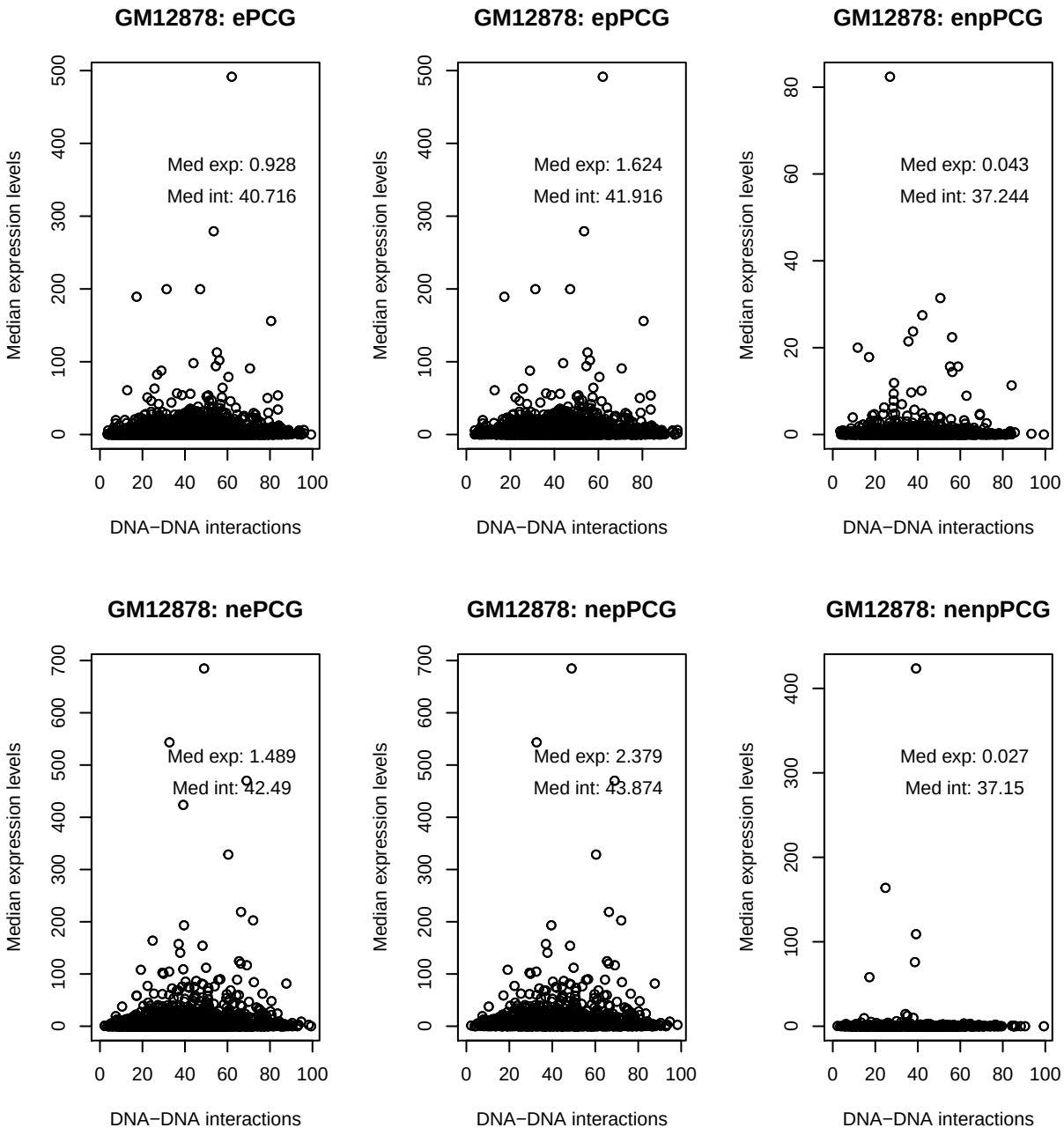
Overview of data:

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:

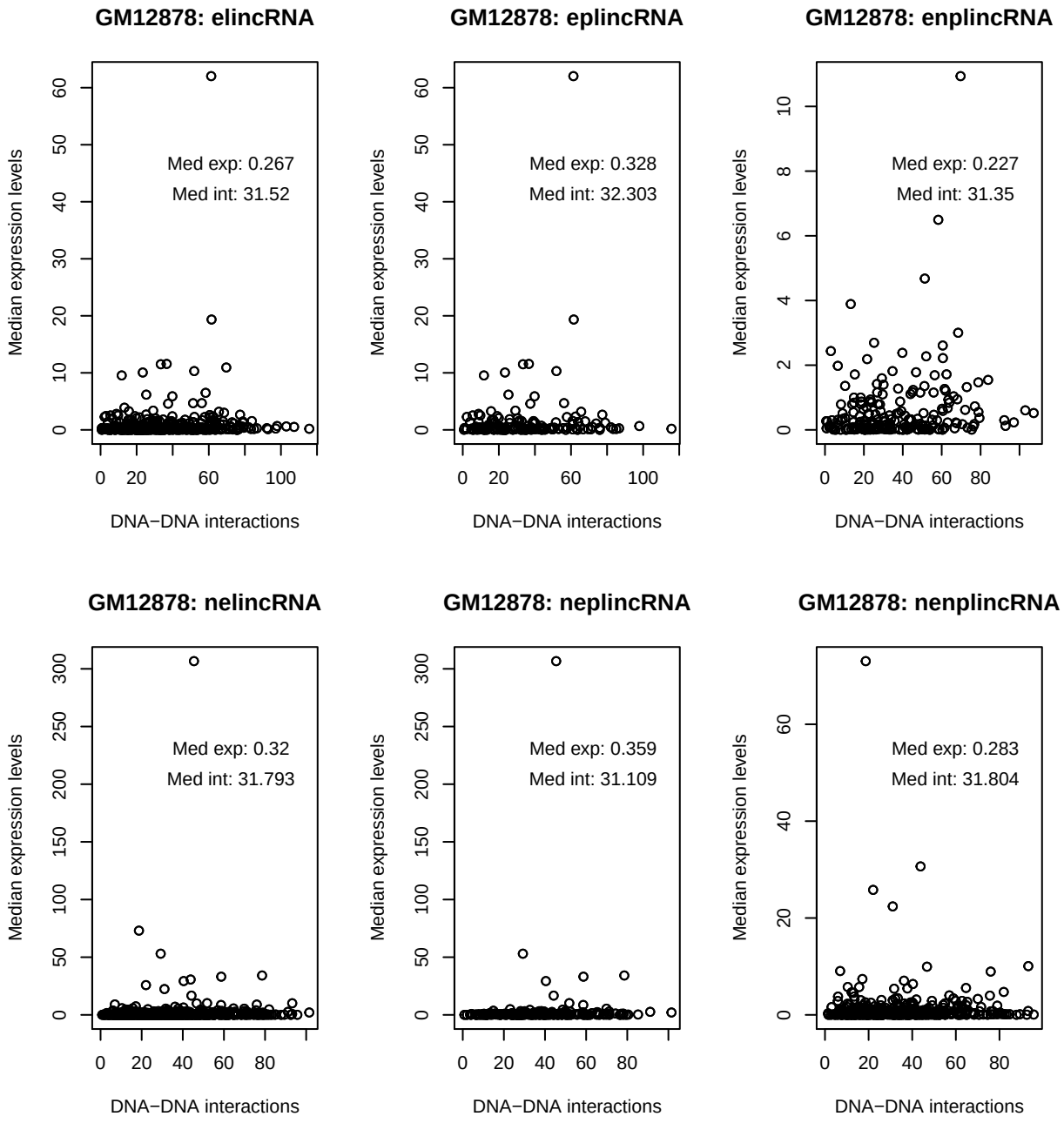
Here, the contact value of each gene is the mean contacts observed for the TAD it is in.

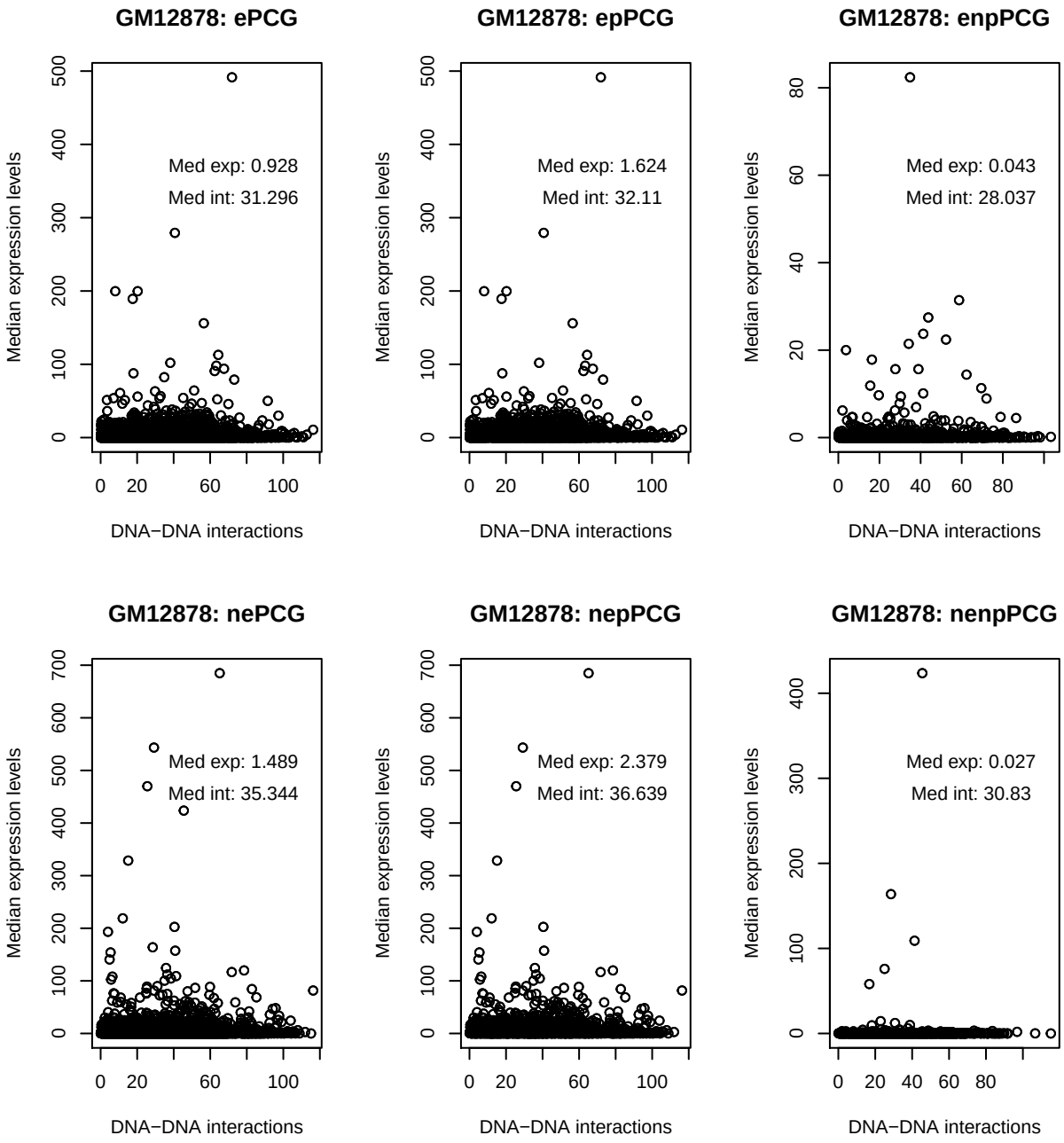

GM12878: elincRNA
Med exp: 0.267
Med int: 39.259
GM12878: eplincRNA
Med exp: 0.328
Med int: 41.048
GM12878: enplincRNA
Med exp: 0.227
Med int: 35.49
GM12878: nelincRNA
Med exp: 0.32
Med int: 36.03
GM12878: neplincRNA
Med exp: 0.359
Med int: 41.384
GM12878: nenplincRNA
Med exp: 0.283
Med int: 33.772
Median expression levels
DNA−DNA interactions


GM12878: ePCG
Med exp: 0.928
Med int: 40.716
GM12878: epPCG
Med exp: 1.624
Med int: 41.916
GM12878: enpPCG
Med exp: 0.043
Med int: 37.244
GM12878: nePCG
Med exp: 1.489
Med int: 42.49
GM12878: nepPCG
Med exp: 2.379
Med int: 43.874
GM12878: nenpPCG
Med exp: 0.027
Med int: 37.15
Median expression levels
DNA−DNA interactions

Gene versus TAD:

Here, I contacts are calculated between the gene and every position in its TAD.

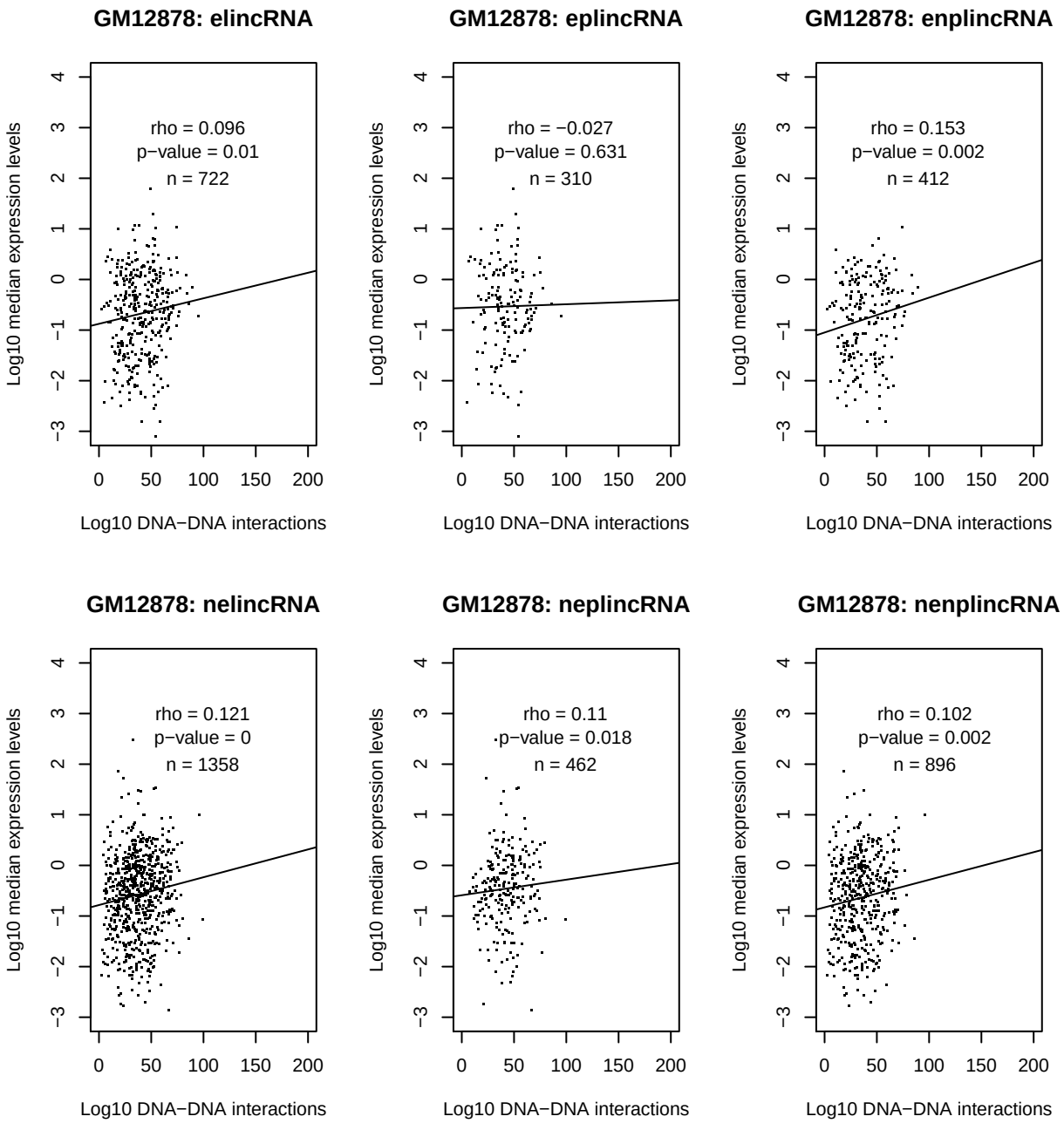


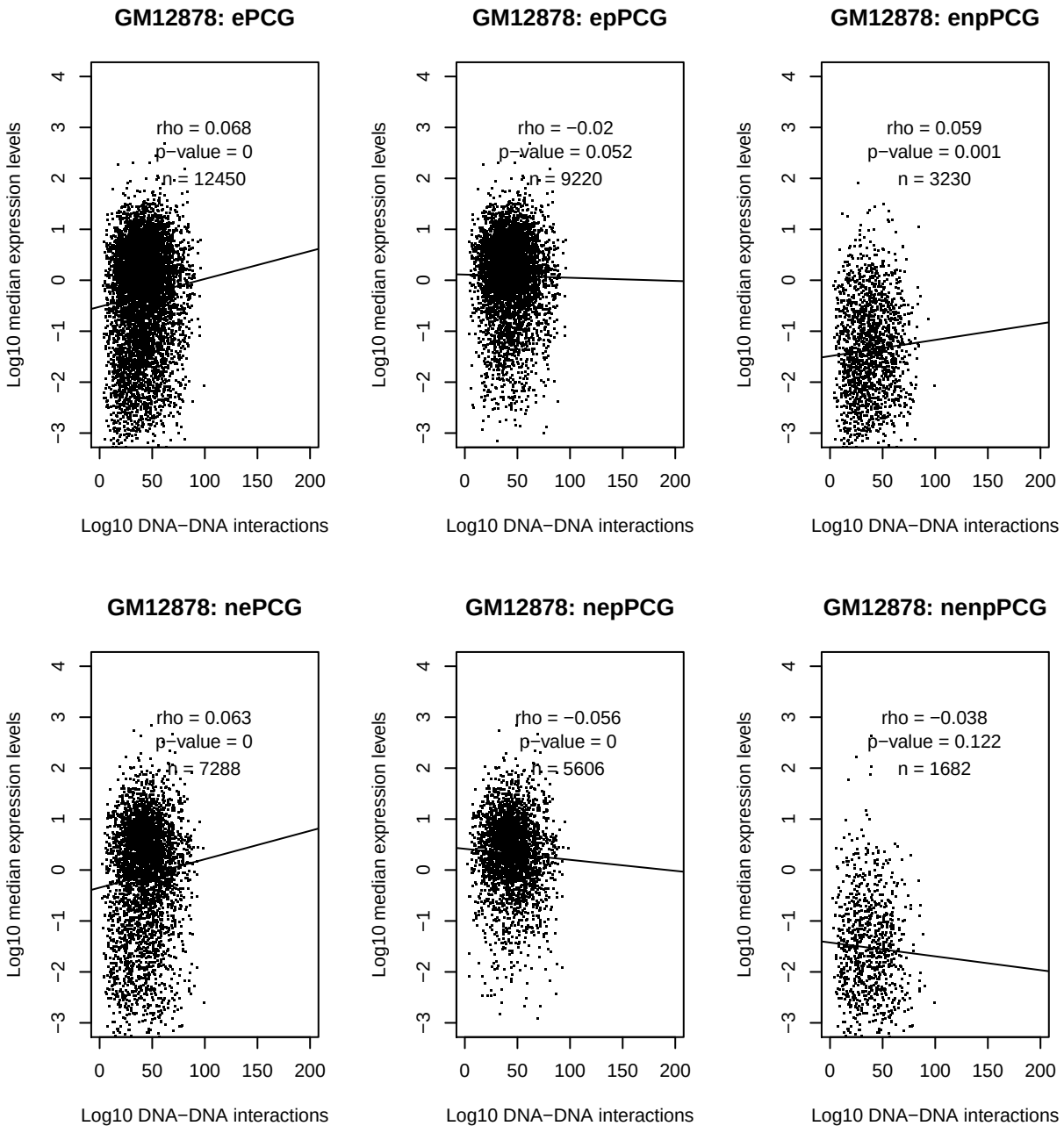

GM12878: ePCG
Med exp: 0.928
Med int: 31.296
GM12878: epPCG
Med exp: 1.624
Med int: 32.11
GM12878: enpPCG
Med exp: 0.043
Med int: 28.037
GM12878: nePCG
Med exp: 1.489
Med int: 35.344
GM12878: nepPCG
Med exp: 2.379
Med int: 36.639
GM12878: nenpPCG
Med exp: 0.027
Med int: 30.83
Median expression levels
DNA−DNA interactions

Correlations:

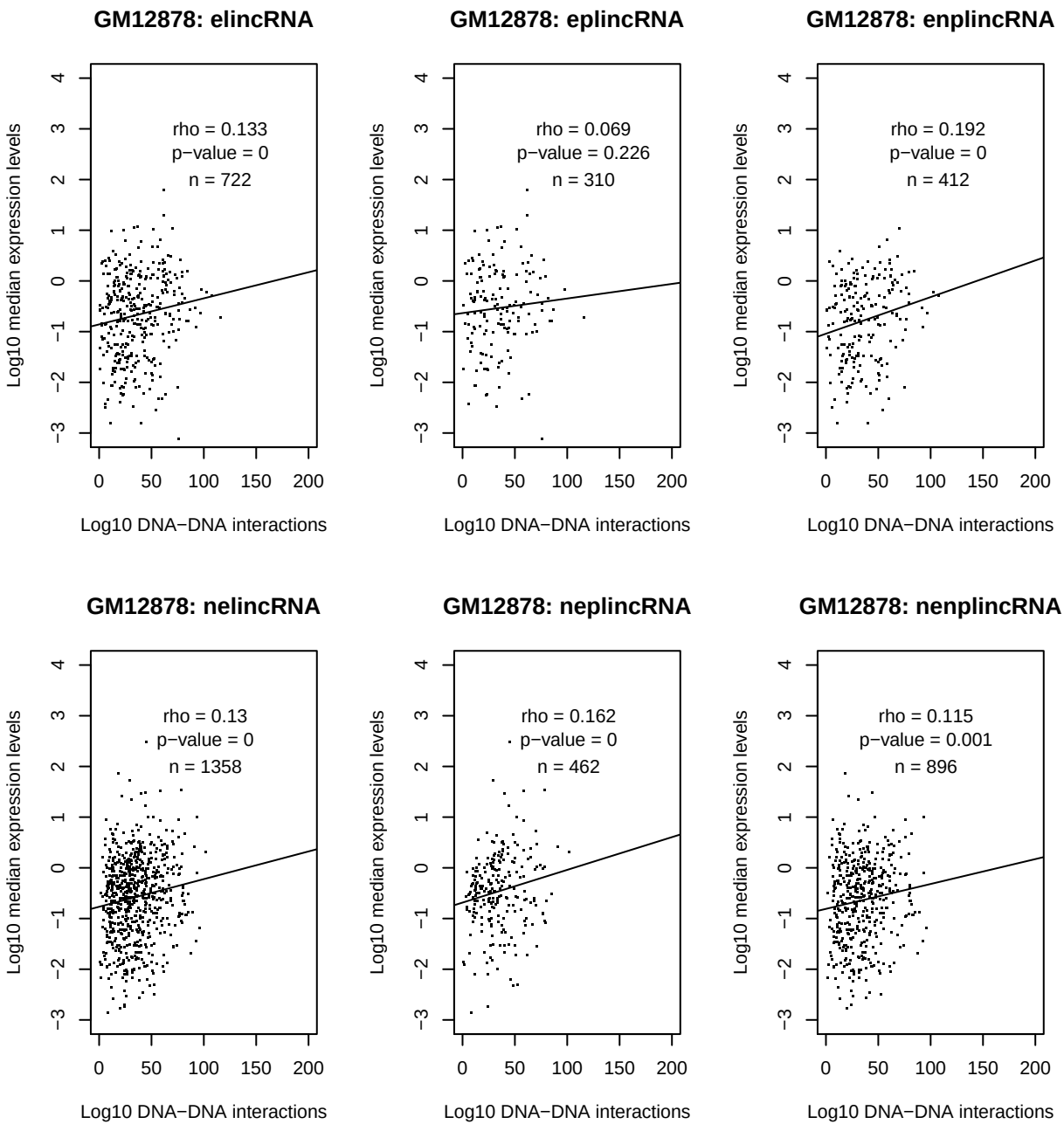
Spearman correlation was used to detect non-linear relationship between DNA contact and expression. The data was log-transformed for visualization.

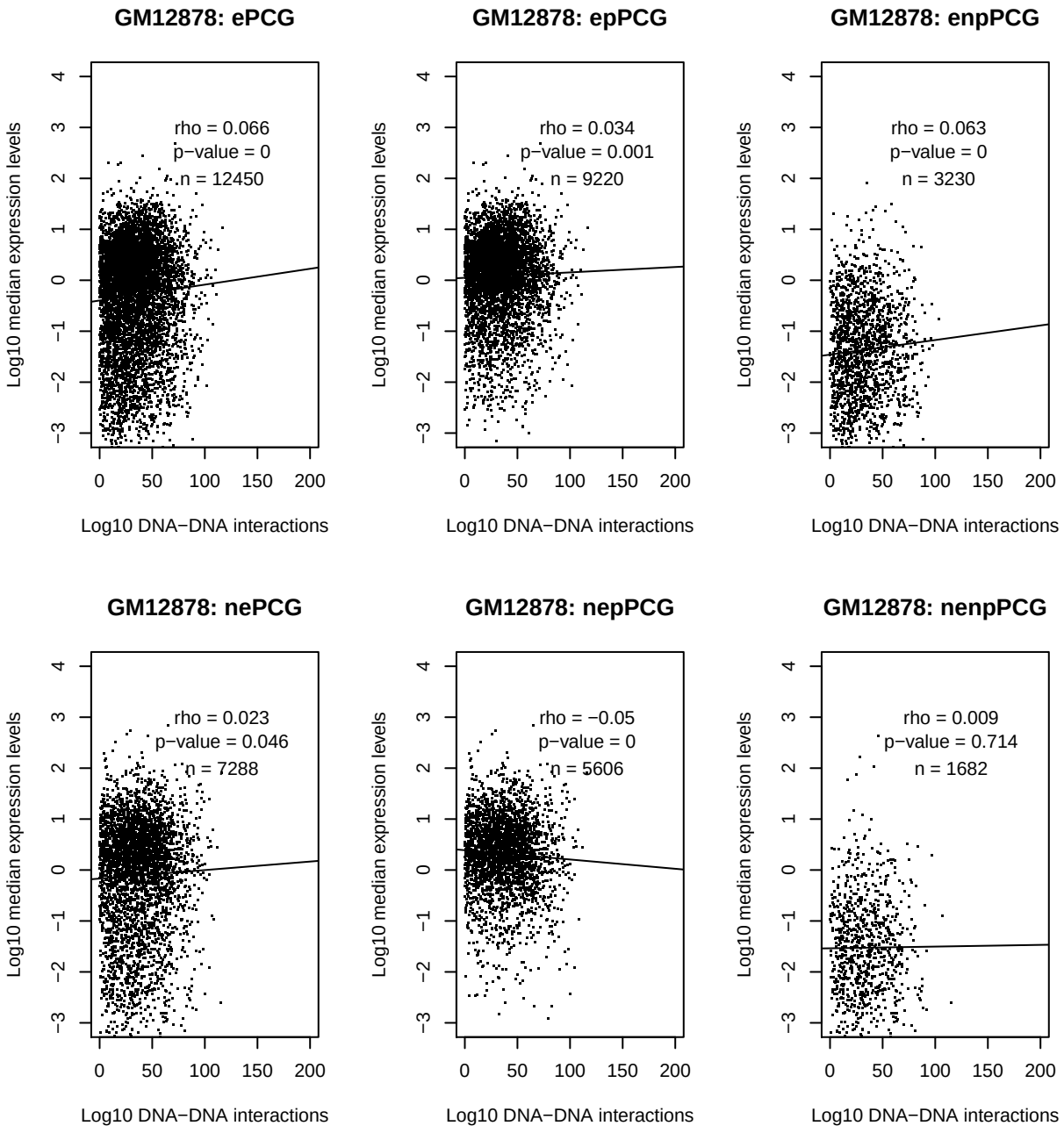
Mean TAD contacts:




GM12878: ePCG
rho = 0.068
p−value = 0
n = 12450
GM12878: epPCG
rho = −0.02
p−value = 0.052
n = 9220
GM12878: enpPCG
rho = 0.059
p−value = 0.001
n = 3230
GM12878: nePCG
rho = 0.063
p−value = 0
n = 7288
GM12878: nepPCG
rho = −0.056
p−value = 0
n = 5606
GM12878: nenpPCG
rho = −0.038
p−value = 0.122
n = 1682
Log10 median expression levels
Log10 DNA−DNA interactions

Gene versus TAD:




GM12878: ePCG
rho = 0.066
p−value = 0
n = 12450
GM12878: epPCG
rho = 0.034
p−value = 0.001
n = 9220
GM12878: enpPCG
rho = 0.063
p−value = 0
n = 3230
GM12878: nePCG
rho = 0.023
p−value = 0.046
n = 7288
GM12878: nepPCG
rho = −0.05
p−value = 0
n = 5606
GM12878: nenpPCG
rho = 0.009
p−value = 0.714
n = 1682
Log10 median expression levels
Log10 DNA−DNA interactions

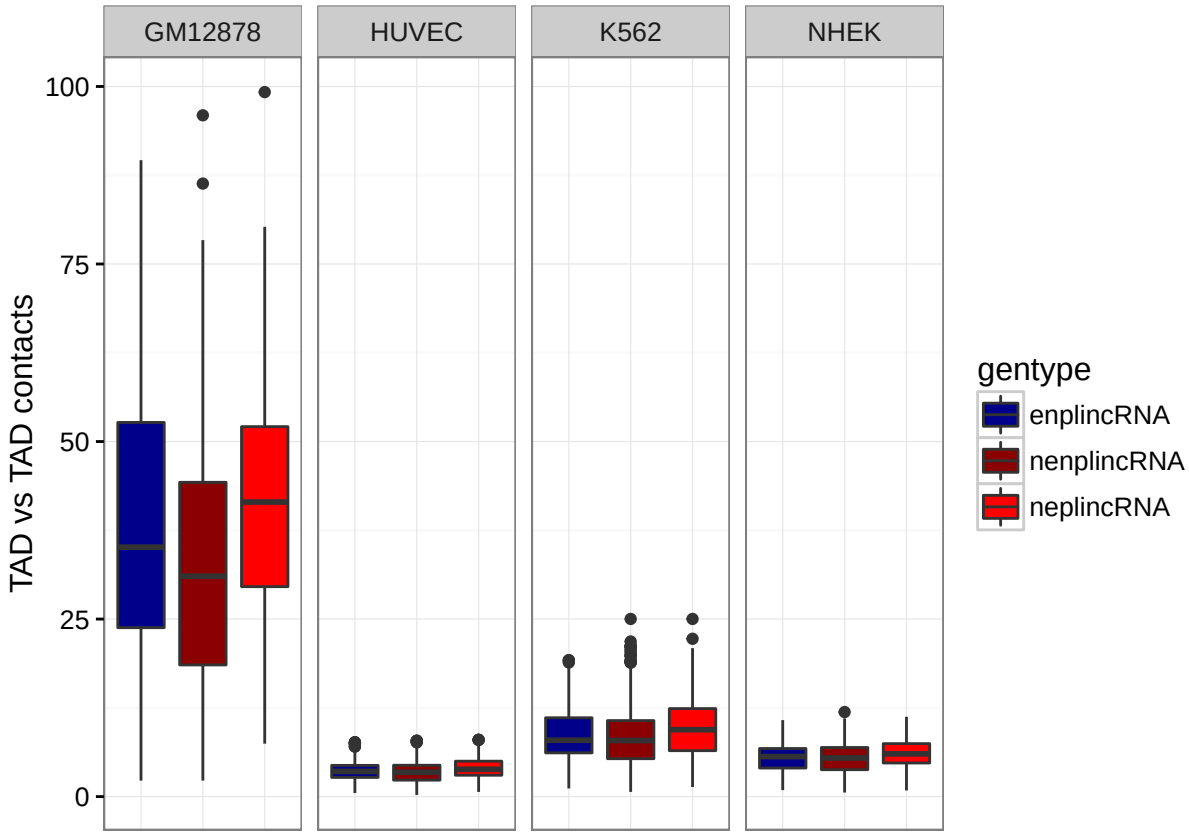
Stats for all cell lines

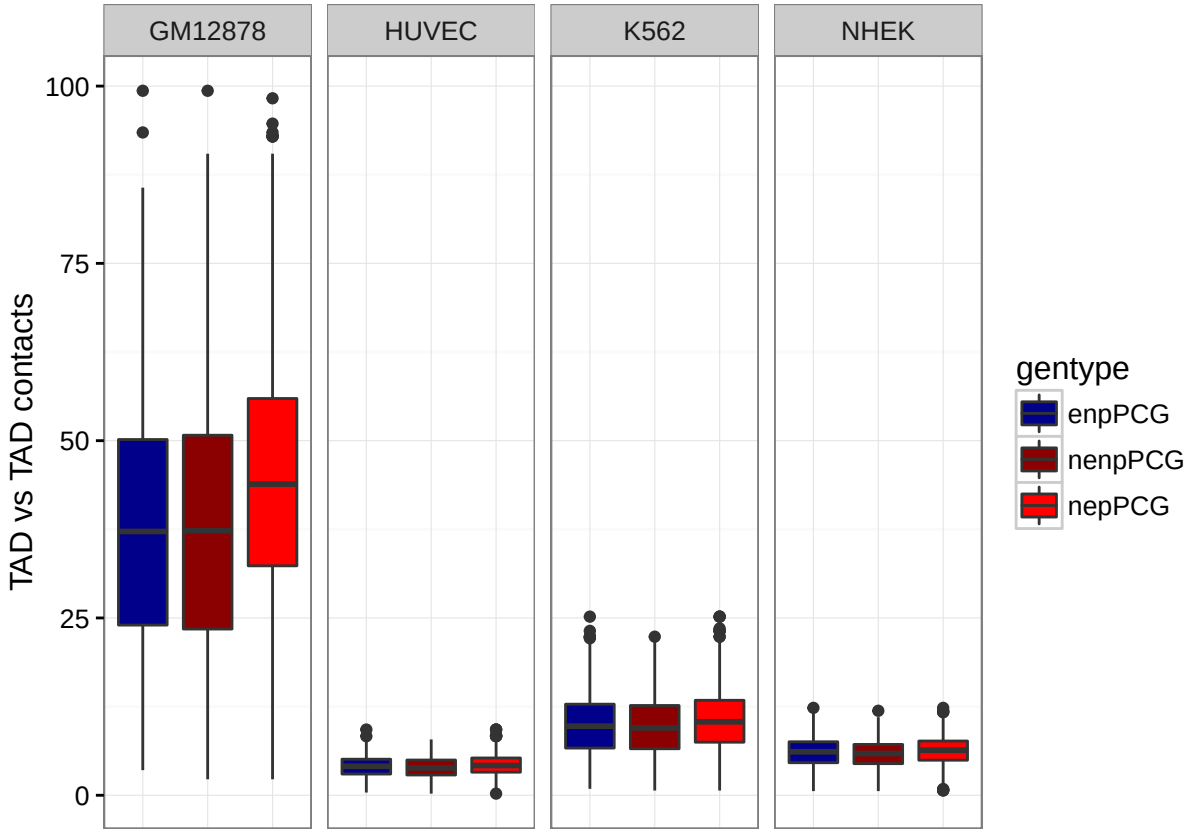
Correlations between TAD_TAD contact and expression. All values are obtained with Spearman correlation tests.

	cell_line	gentype	rho	p.value
20	NHEK	eplincRNA	-0.0053	0.9273
9	HUVEC	enplincRNA	-0.0062	0.9172
2	GM12878	eplincRNA	-0.0274	0.631
19	NHEK	elincRNA	-0.0502	0.2146
13	K562	elincRNA	-0.0572	0.1608
21	NHEK	enplincRNA	-0.1003	0.075
14	K562	eplincRNA	-0.2118	3e-04
7	HUVEC	elincRNA	0.0081	0.8484
23	NHEK	neplincRNA	0.0242	0.6108
8	HUVEC	eplincRNA	0.027	0.6575
15	K562	enplincRNA	0.0461	0.4138
22	NHEK	nelincRNA	0.0573	0.0334
24	NHEK	nenplincRNA	0.0746	0.0228
1	GM12878	elincRNA	0.0964	0.0095
18	K562	nenplincRNA	0.0975	0.0025
6	GM12878	nenplincRNA	0.1021	0.0022
5	GM12878	neplincRNA	0.1096	0.0184
4	GM12878	nelincRNA	0.1207	0
12	HUVEC	nenplincRNA	0.1295	1e-04
16	K562	nelincRNA	0.1373	0
10	HUVEC	nelincRNA	0.1404	0
17	K562	neplincRNA	0.1523	0.0011
11	HUVEC	neplincRNA	0.1529	0.0013
3	GM12878	enplincRNA	0.153	0.0018

Comparisons sacross cell lines.

Comparing gene-TAD contact across all cell lines. Metrics is gene_TAD contacts.


GM12878
HUVEC
K562
NHEK
TAD vs TAD contacts
100
75
50
25
0
gentype
enplincRNA
nenplincRNA
neplincRNA

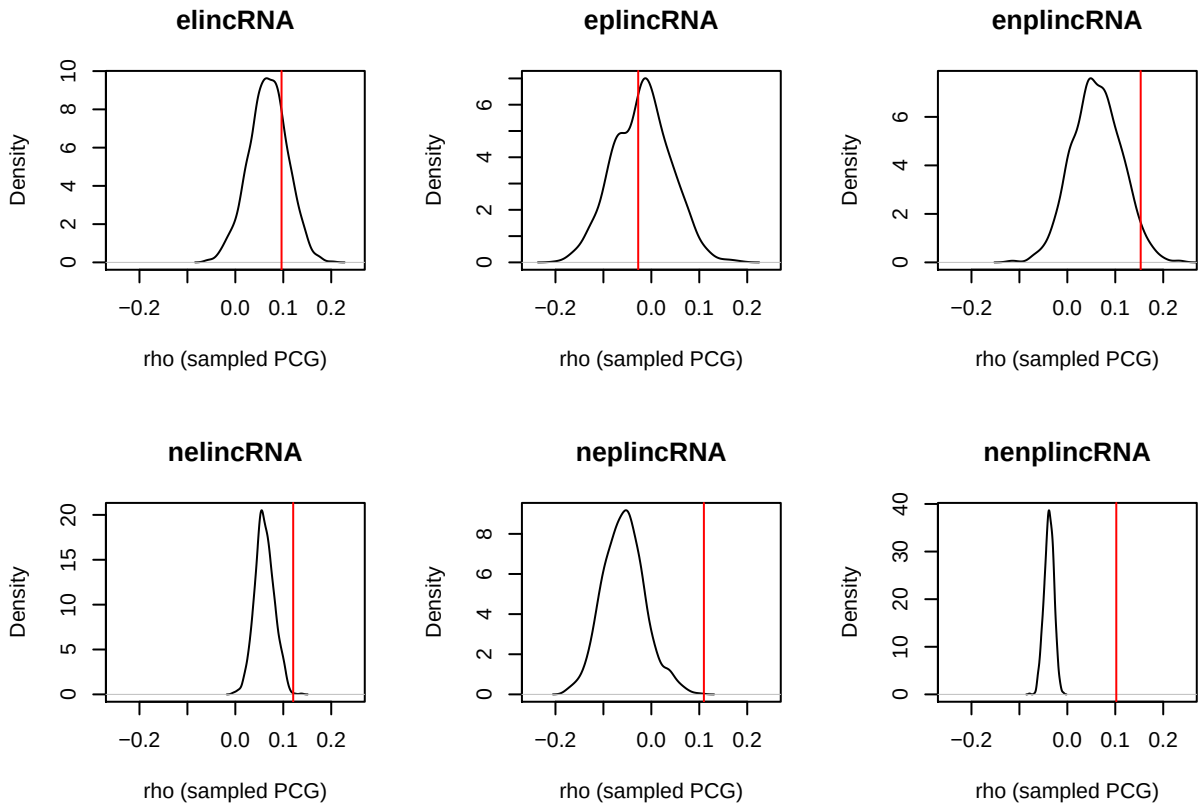

GM12878
HUVEC
K562
NHEK
TAD vs TAD contacts
100
75
50
25
0
gentype
enpPCG
nenpPCG
nepPCG

pval	cell.line	comp
1.17e-05	GM12878	enplincRNA ~ nenplincRNA
0.0299	GM12878	enplincRNA ~ neplincRNA
2.49e-13	GM12878	neplincRNA ~ nenplincRNA
0.064	HUVEC	enplincRNA ~ nenplincRNA
0.0073	HUVEC	enplincRNA ~ neplincRNA
6.73e-07	HUVEC	neplincRNA ~ nenplincRNA
0.192	K562	enplincRNA ~ nenplincRNA
0.00202	K562	enplincRNA ~ neplincRNA
5.81e-07	K562	neplincRNA ~ nenplincRNA
0.533	NHEK	enplincRNA ~ nenplincRNA
0.00184	NHEK	enplincRNA ~ neplincRNA
1.09e-05	NHEK	neplincRNA ~ nenplincRNA

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.

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.