Report 6: Redirection

Cyril Matthey-Doret

28 octobre 2016

Introduction

The largely different gene length between protein-coding genes and lincRNAs made it more difficult than expected to find appropriate thresholds to call TAD-bound and non TAD-bound genes. The focus of the project has thus been redirected to investigate the role of enhancer-associated lincRNAs (elincRNAs) in the organization of TADs.

Methods

- **Definition of TAD-boundaries**: TAD boundaries are now only extended towards the inside of TADs at 5% of TAD length and when TAD boundaries overlap, the largest ones are removed. This prevent the boundaries of large encompassing TADs from masking those of smaller TADs inside.
- **Definition of enhancer-associated genes**: LincRNAs and protein-coding genes whose promoter regions (or promoter region + gene body) was overlapping enhancer elements (defined by histone modification marks) were defined as elincRNAs or epc-genes. The others are considered non-enhancer-associated. Promoter regions have been defined as the region of 1kb around the TSS of each gene.

Gene type	region	Original set	Overlaps	enhancer-associated genes
lincRNA	pro	2510	323	274 (10.9%)
lincRNA	pro+bod	2510	1544	611 (24.3%)
protein-coding	pro	14846	3411	3019 (20.3%)
protein-coding	pro+bod	14846	34808	8486 (57.2%)

- **Generation of bins**: TADs have been split into bins of 5%. This was done after removing all large TADs overlapping smaller ones. Note the difference with the definition of TAD-boundaries, where only large overlapping boundaries have been removed.
- **Testing enrichment at TAD boundaries**: The GAT framework is used to test if enhancer-bound lincRNA and protein-coding genes are enriched at TAD boundaries. For both gene types, it is done in two ways: testing enrichment of enhancer-bound and non-enhancer bound genes in the different bins, and testing enrichment of the different bins in enhancer-bound and non-enhancer bound genes.
- **Transcription factors binding sites**: Chip-seq data for SMC1, SMC3, RAD21, SA1, SA2 and CTCF from the ENCODE website was used.

Enrichment at TAD boundaries

Several GAT tests were performed using different combinations of segments, annotations and workspace. Both nucleotide overlap and number of overlapping segments were used as a measure:

Workspace	Segment	Annotation
Whole genome	[n]e(linc | pc)pr[b]	(bins | CTCF & cohesin bs)
Whole genome	(bins | CTCF & cohesin bs)	[n]e(linc | pc)pr[b]
intergenic	[n]elinc pr[b]	(bins | CTCF & cohesin bs)
intergenic	(bins | CTCF & cohesin bs)	[n]elinc pr[b]
all pc genes	[n]epc pr[b]	(bins | CTCF & cohesin bs)
all pc genes	(bins | CTCF & cohesin bs)	[n]epc pr[b]
expressed lincRNAs	elinc pr[b]	(bins | CTCF & cohesin bs)
expressed lincRNAs	(bins | CTCF & cohesin bs)	elinc pr[b]

where:

bs = binding sites

e = enhancer bound

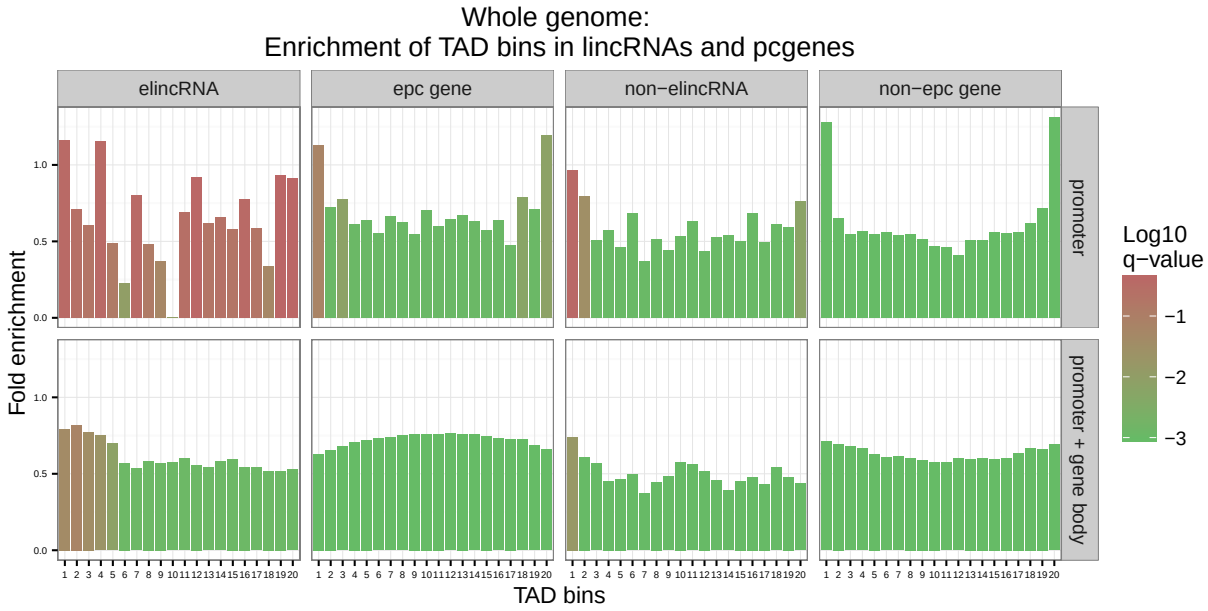
ne = non-enhancer bound

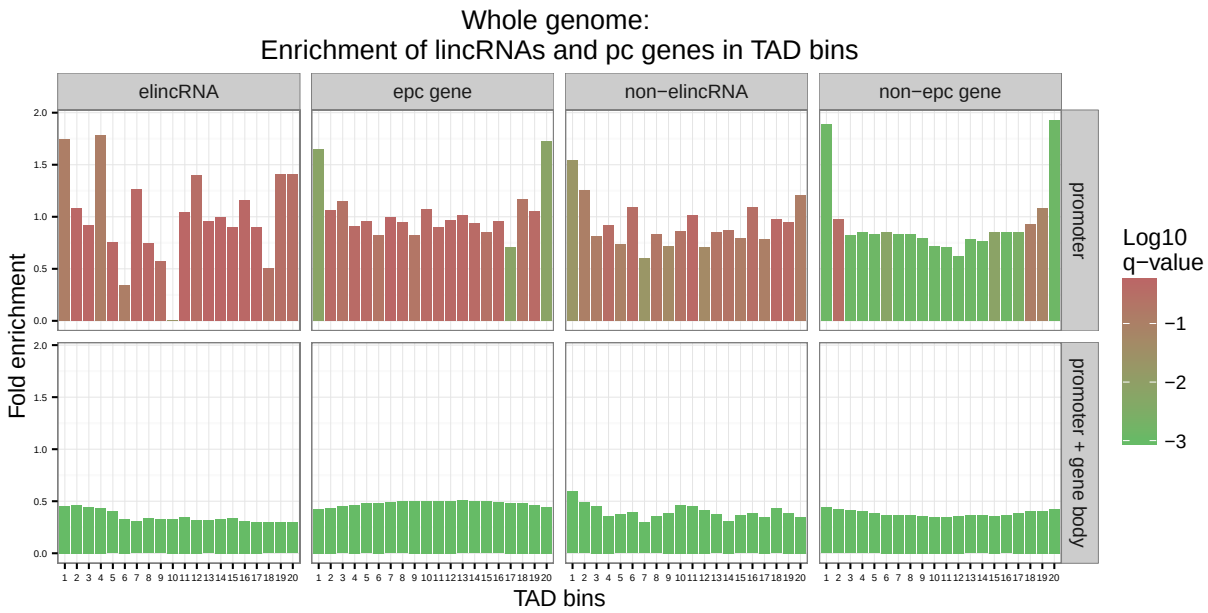
pr = promoter region

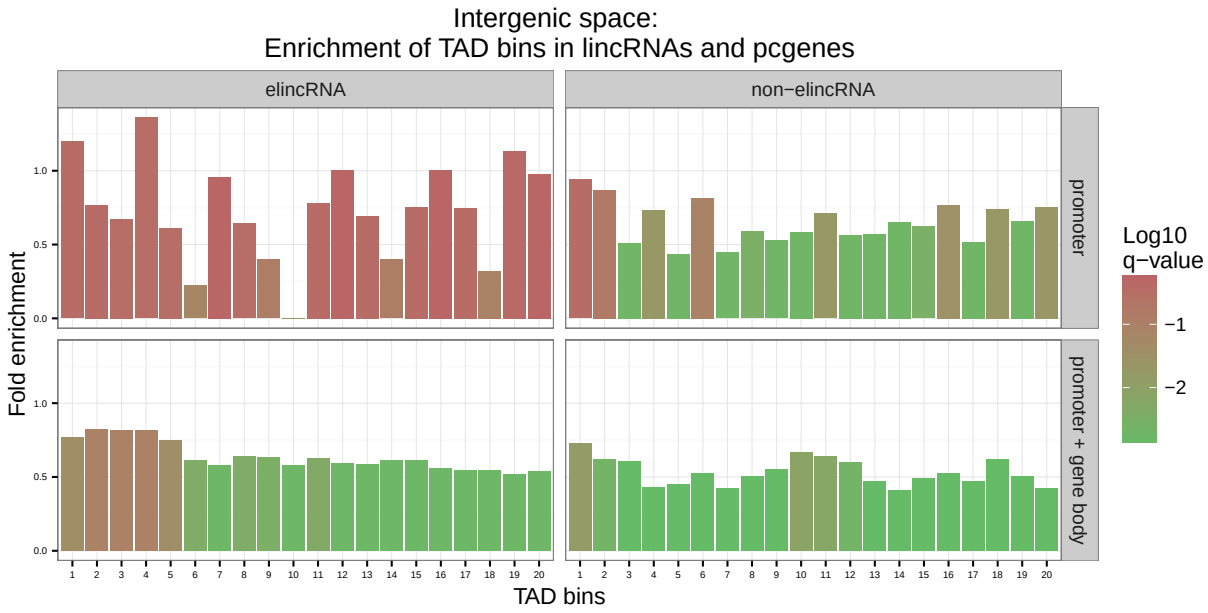
prb = promoter region and body

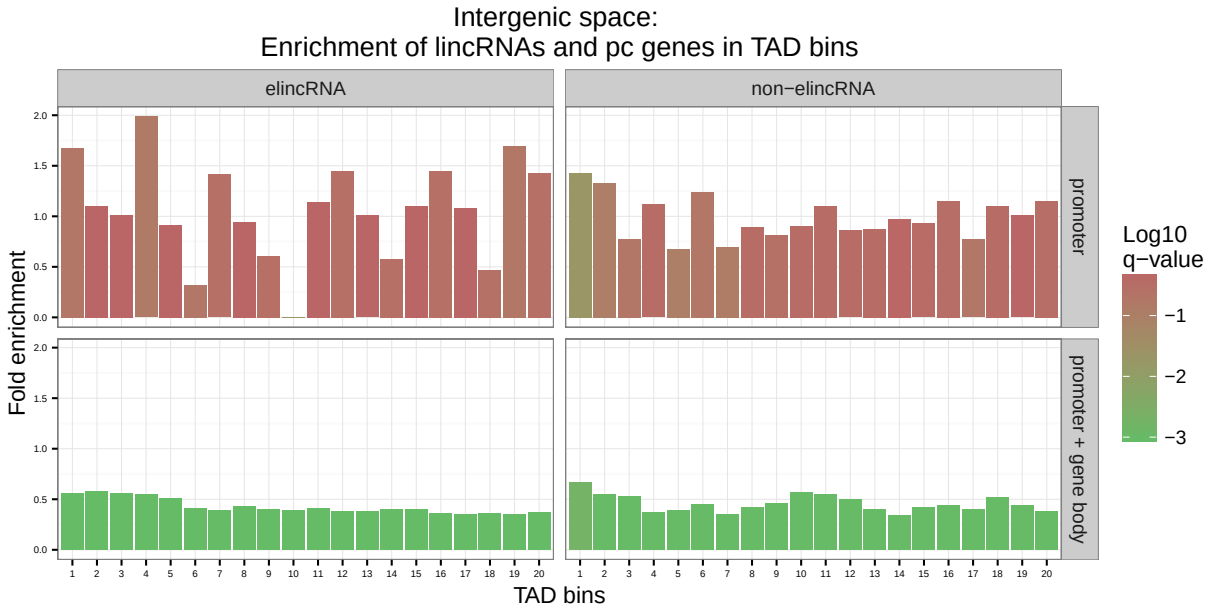
Nucleotides overlap

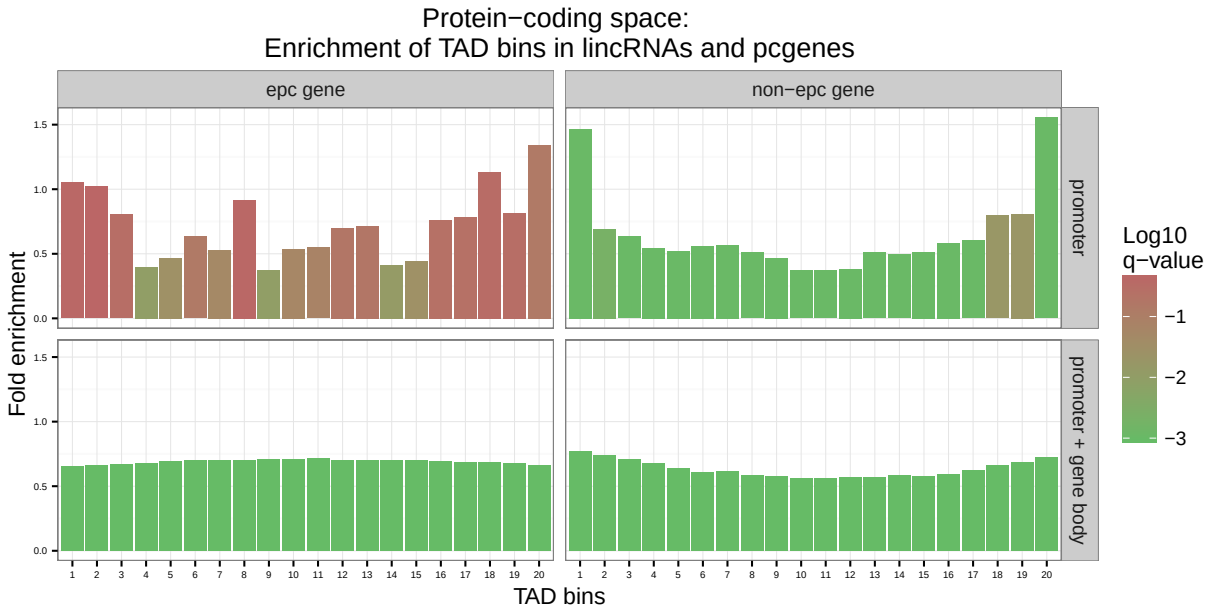
Analyses using TAD bins

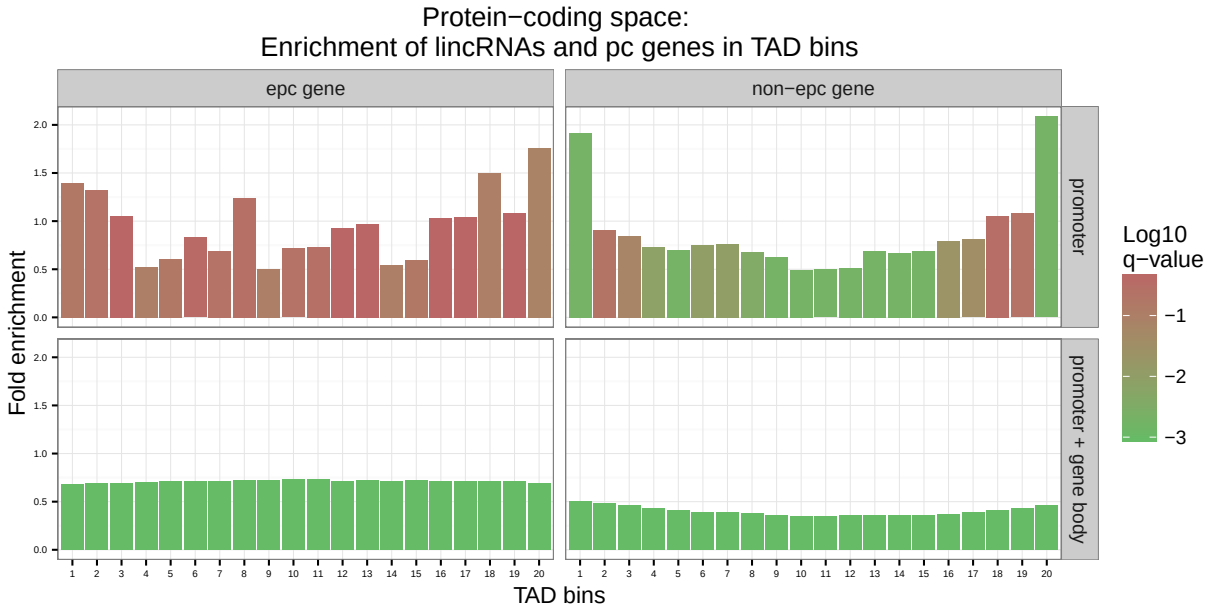


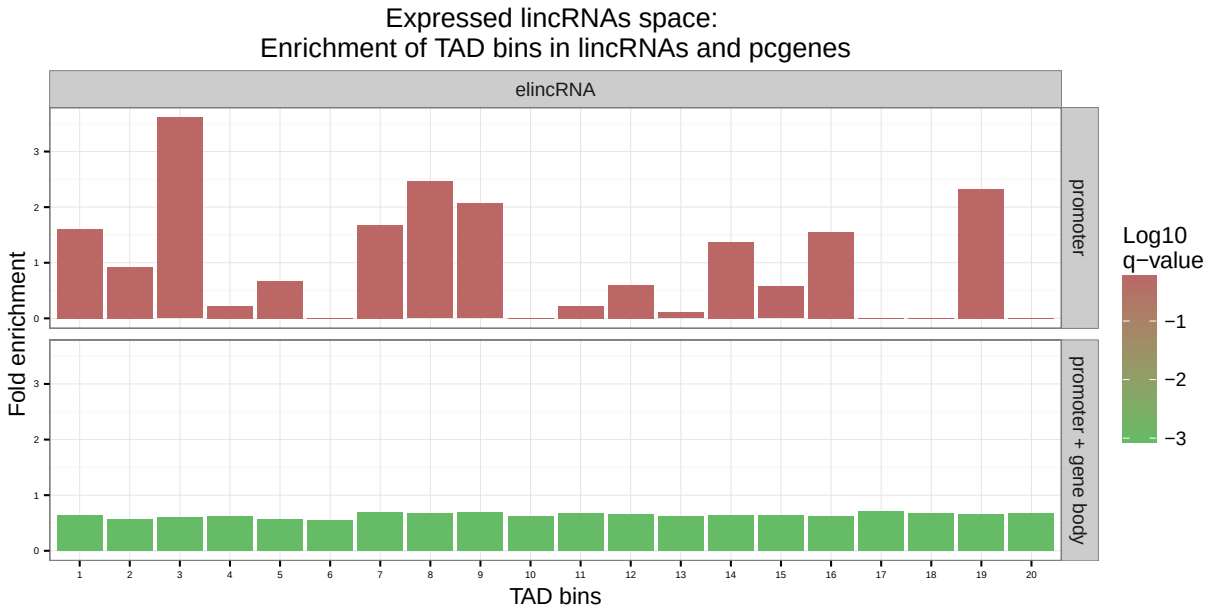


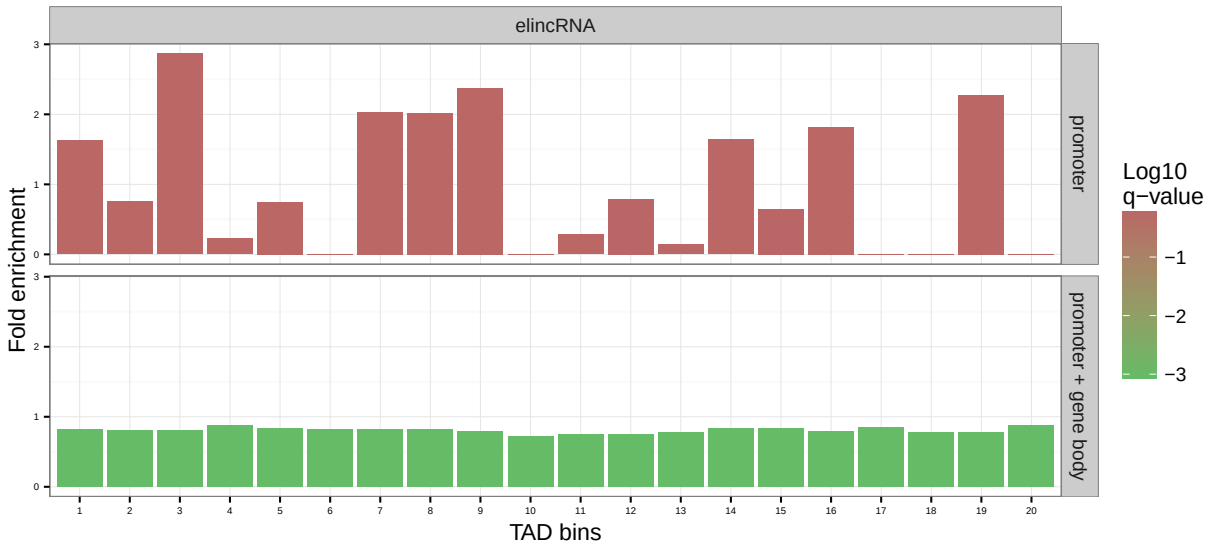

Intergenic space:
Enrichment of TAD bins in lincRNAs and pcgenes
elincRNA
non−elincRNA
promoter
promoter + gene body
Fold enrichment
TAD bins
Log10
q−value
−1
−2


Intergenic space:
Enrichment of lincRNAs and pc genes in TAD bins
elincRNA
non−elincRNA
promoter
promoter + gene body
Fold enrichment
TAD bins
Log10
q−value
−1
−2
−3

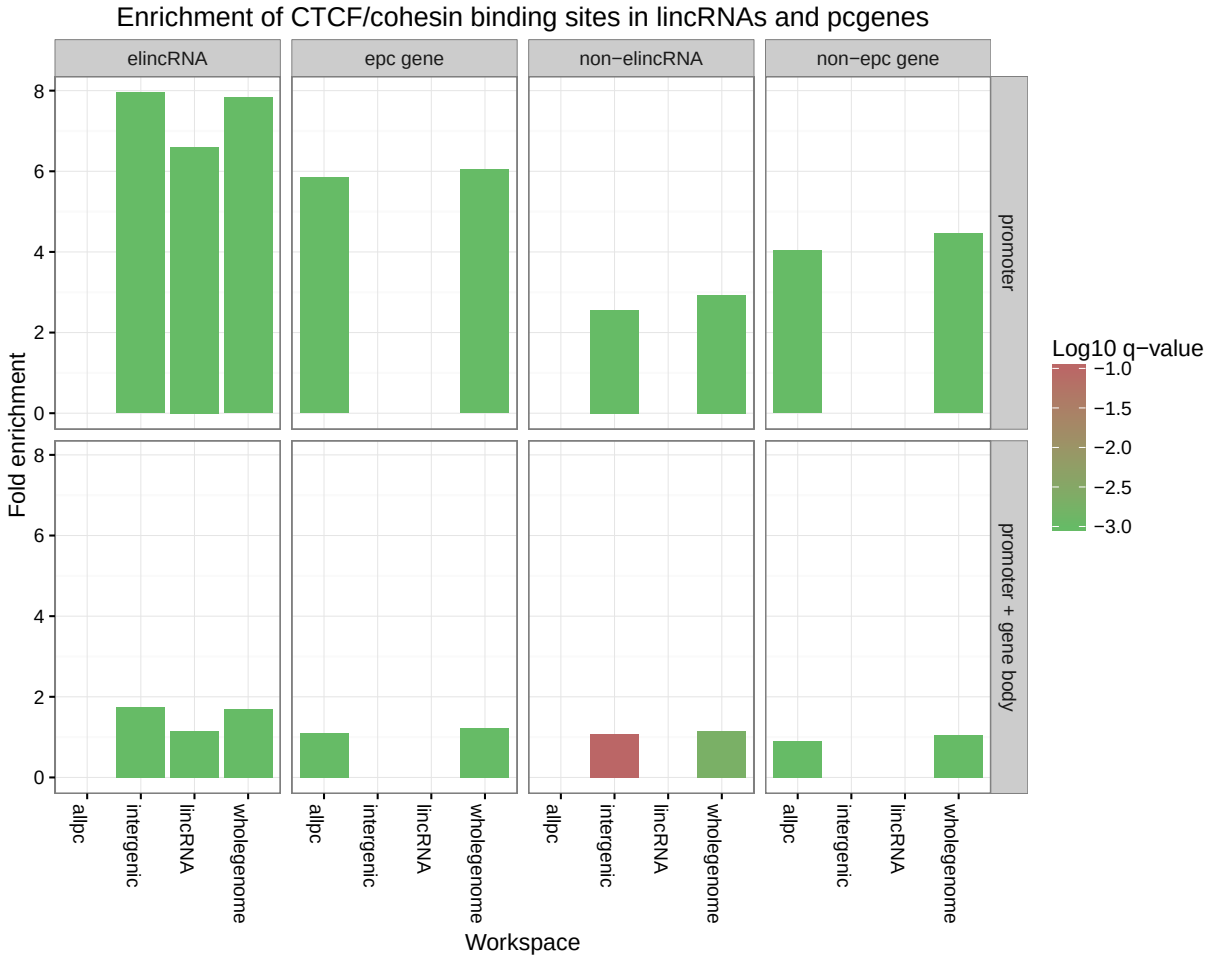

Protein−coding space:
Enrichment of TAD bins in lincRNAs and pcgenes
epc gene
non−epc gene
promoter
promoter + gene body
Fold enrichment
TAD bins
Log10
q−value
−1
−2
−3

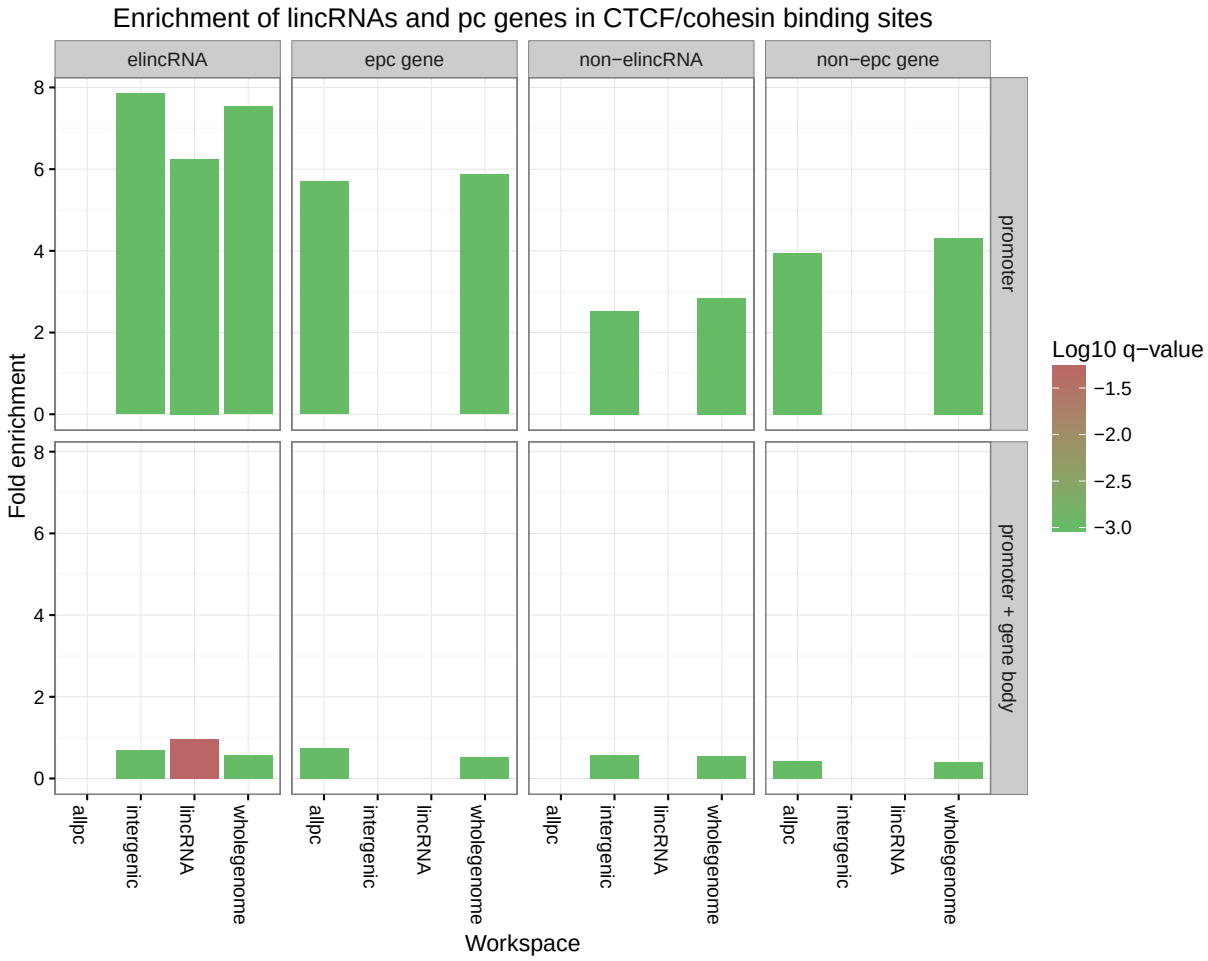

Protein−coding space:
Enrichment of lincRNAs and pc genes in TAD bins
epc gene
non−epc gene
promoter
promoter + gene body
Fold enrichment
TAD bins
Log10
q−value
−1
−2
−3


Expressed lincRNAs space:
Enrichment of TAD bins in lincRNAs and pcgenes
elincRNA
promoter
promoter + gene body
Fold enrichment
TAD bins
Log10
q−value
−1
−2
−3


Expressed lincRNAs space:
Enrichment of lincRNAs and pc genes in TAD bins
elincRNA
promoter
promoter + gene body
Fold enrichment
TAD bins
Log10
q−value
−1
−2
−3

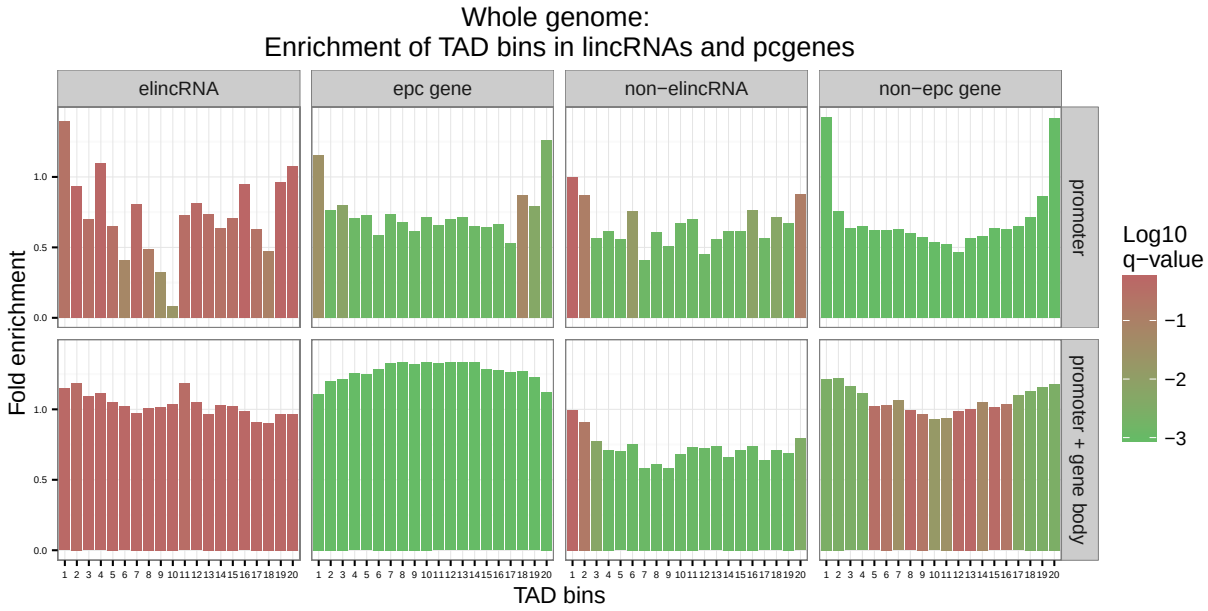
Analyses using CTCF/Cohesin binding sites

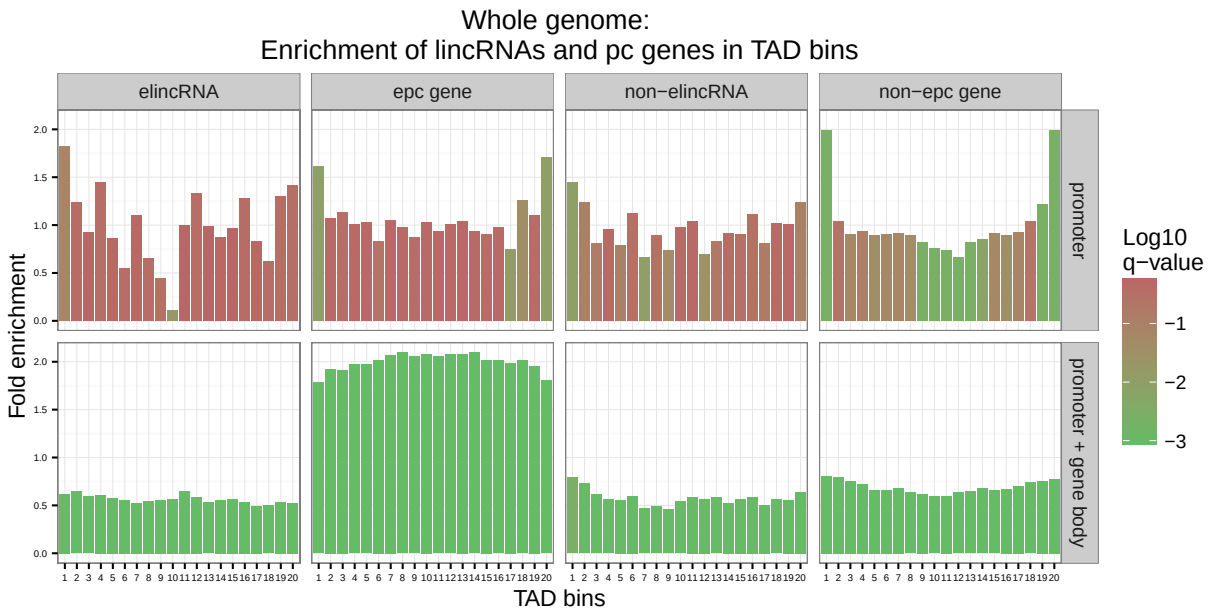


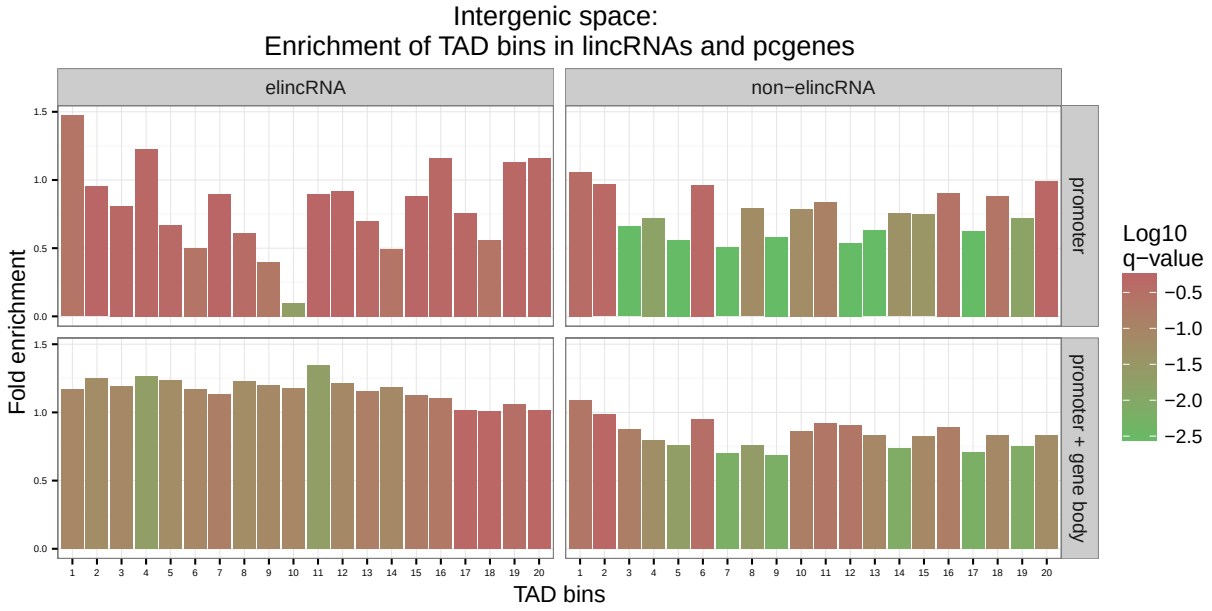

Enrichment of lincRNAs and pc genes in CTCF/cohesin binding sites
elincRNA
epc gene
non−elincRNA
non−epc gene
promoter
promoter + gene body
Fold enrichment
0
2
4
6
8
allpc
intergenic
lincRNA
wholegenome
Workspace
Log10 q−value
−1.5
−2.0
−2.5
−3.0

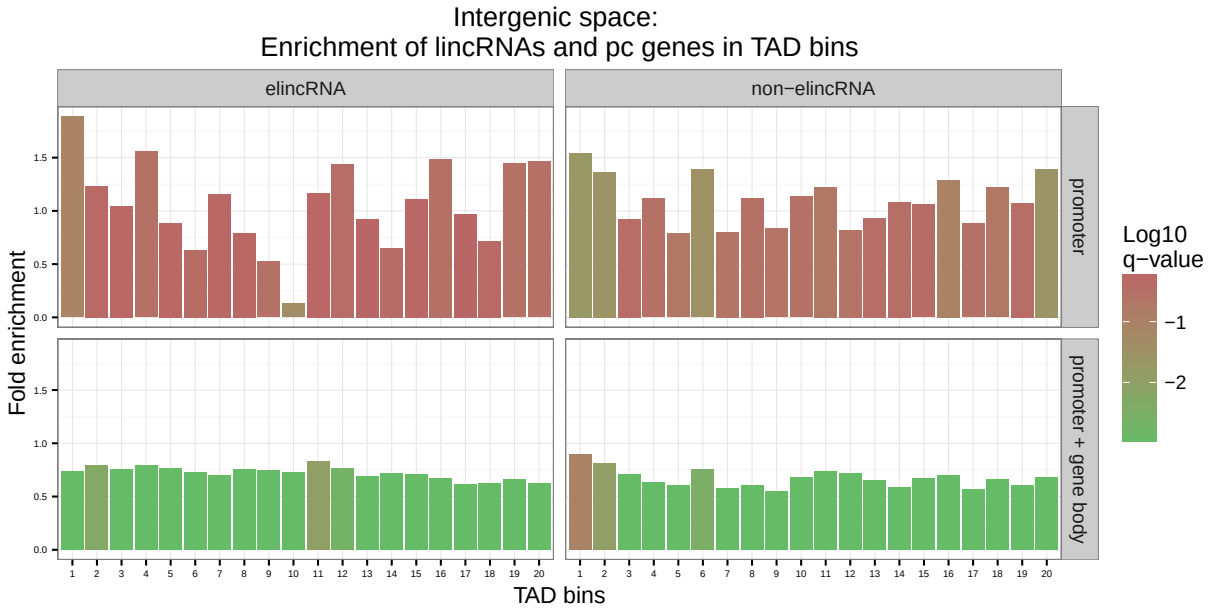
Segments overlap

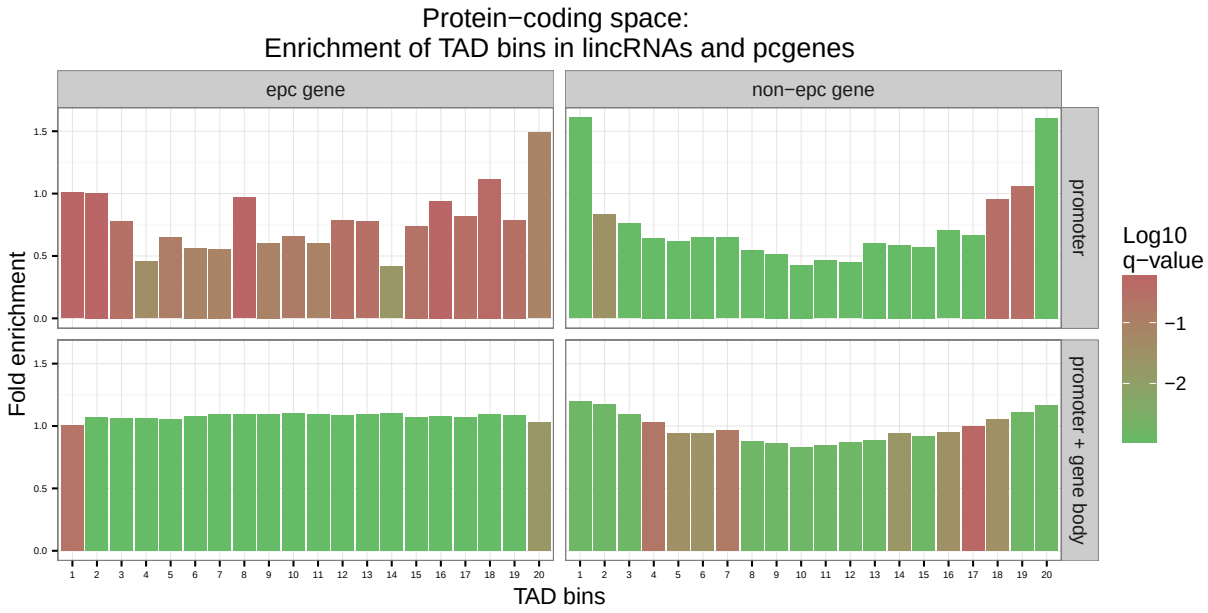
Analyses using TAD bins

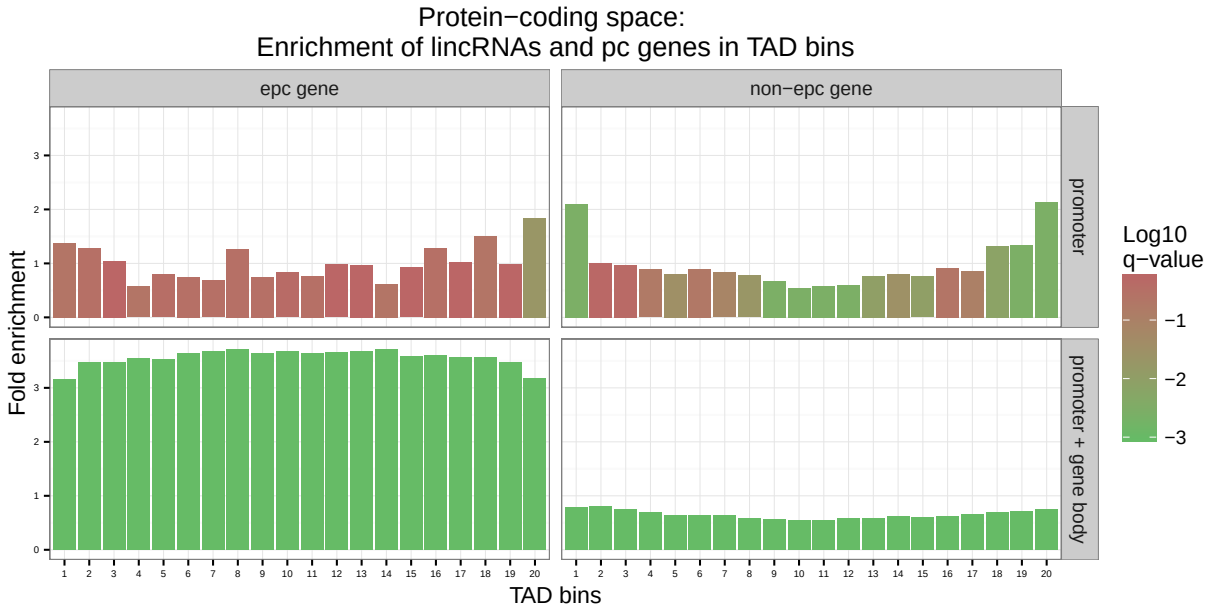


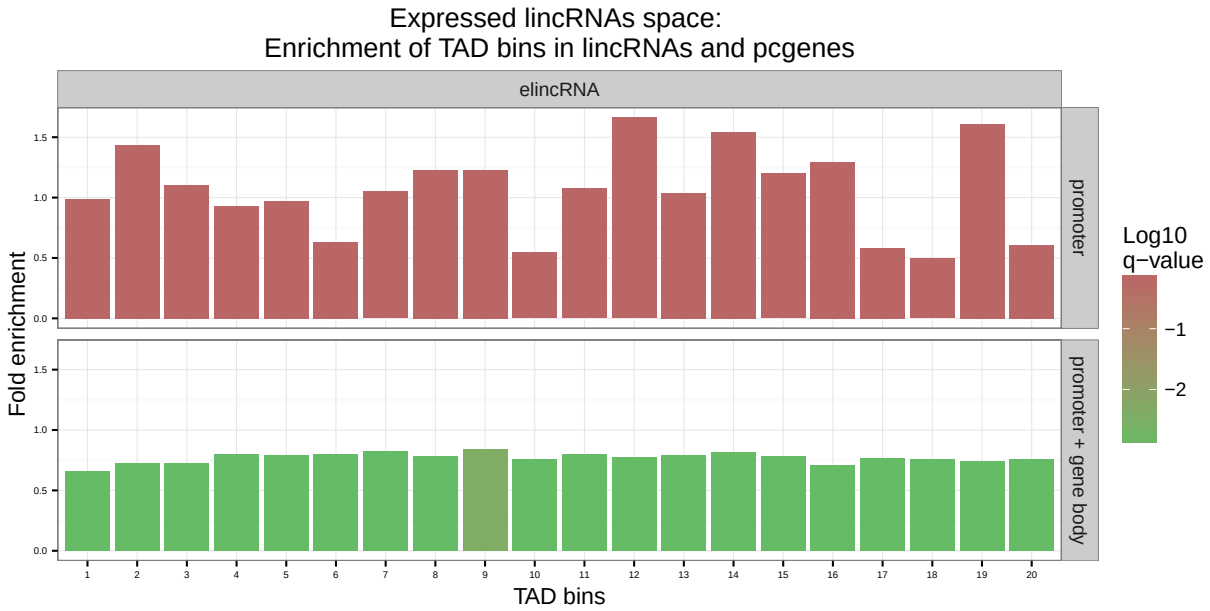


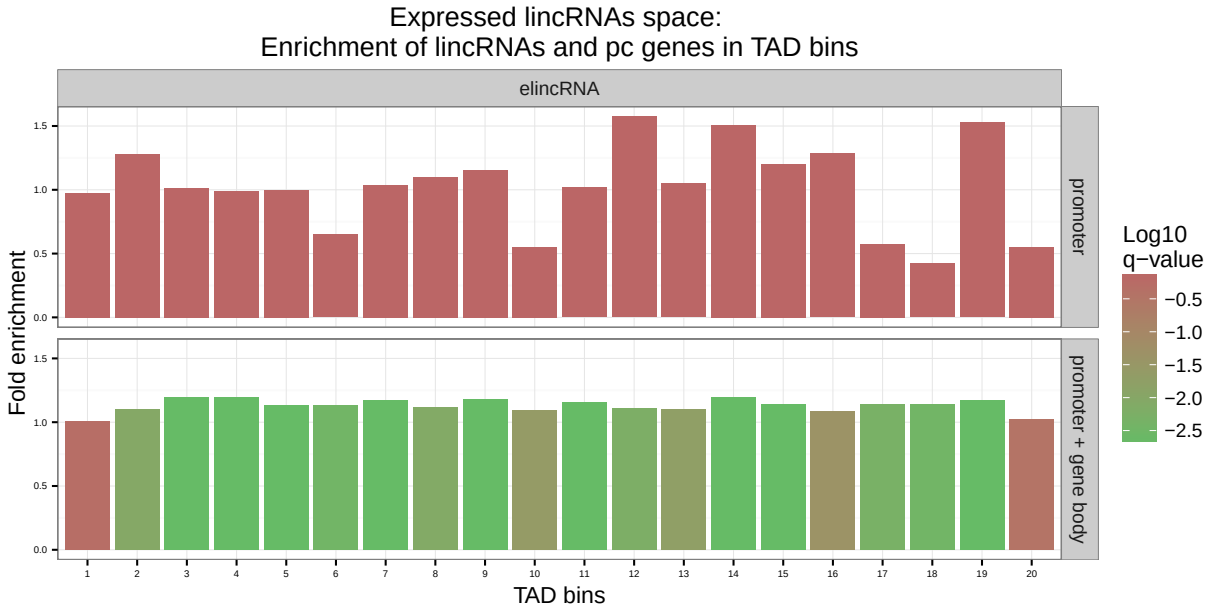

Intergenic space:
Enrichment of TAD bins in lincRNAs and pcgenes
elincRNA
non−elincRNA
promoter
promoter + gene body
Fold enrichment
TAD bins
Log10
q−value
−0.5
−1.0
−1.5
−2.0
−2.5


Intergenic space:
Enrichment of lincRNAs and pc genes in TAD bins
elincRNA
non−elincRNA
promoter
promoter + gene body
Fold enrichment
TAD bins
Log10
q−value
−1
−2

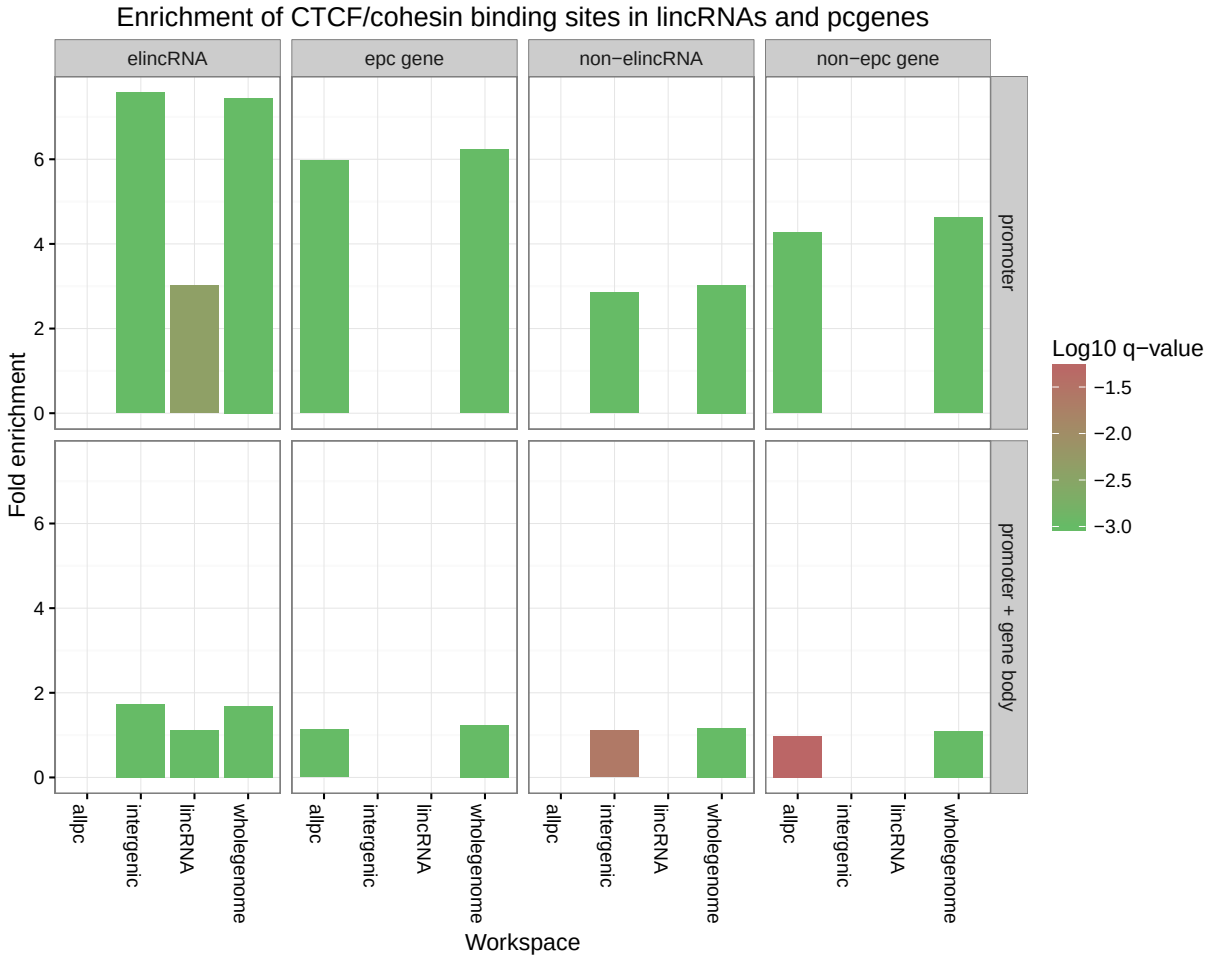

Protein−coding space:
Enrichment of TAD bins in lincRNAs and pcgenes
epc gene
non−epc gene
promoter
promoter + gene body
Fold enrichment
TAD bins
Log10
q−value
−1
−2

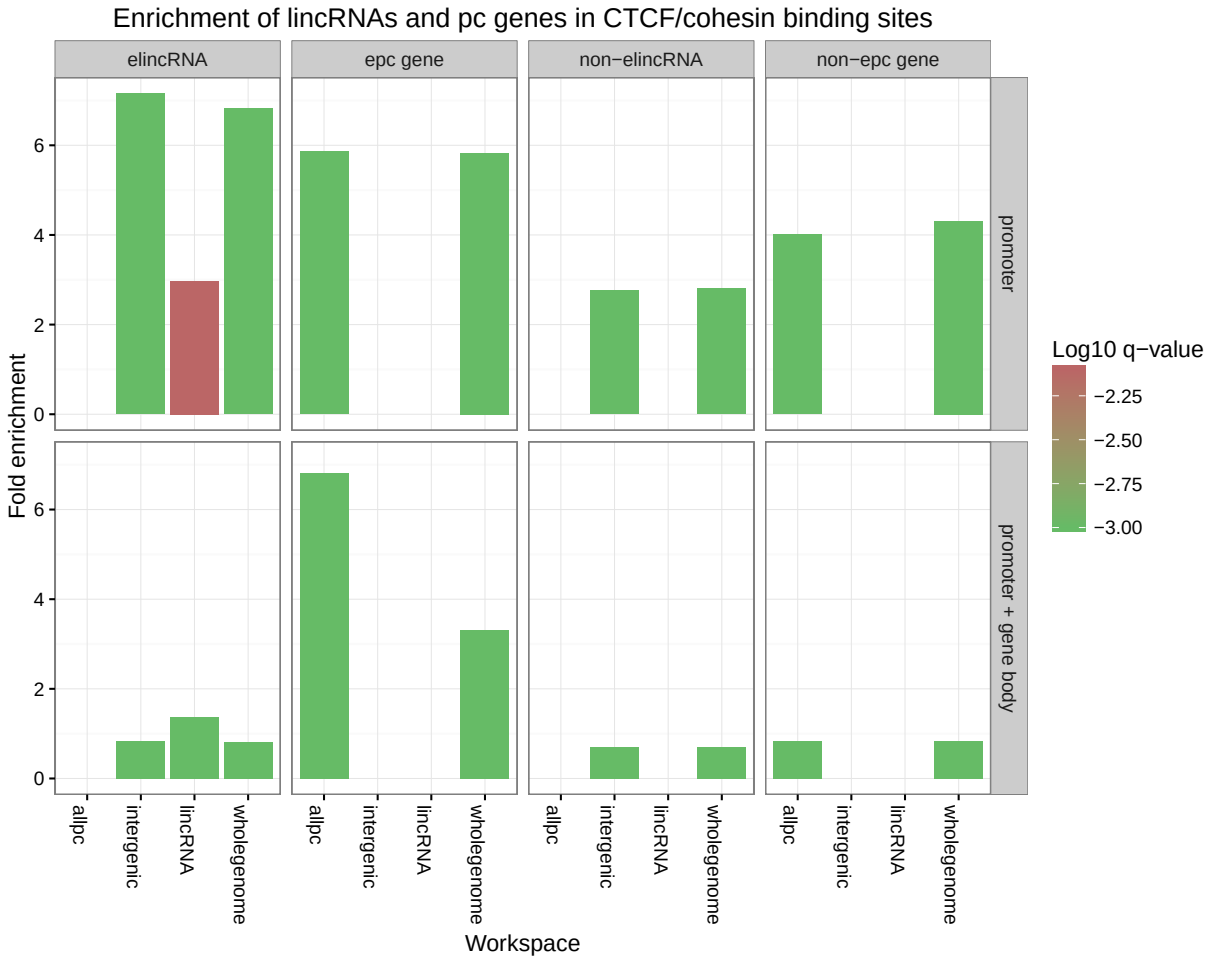

Protein−coding space:
Enrichment of lincRNAs and pc genes in TAD bins
epc gene
non−epc gene
promoter
promoter + gene body
Fold enrichment
TAD bins
Log10
q−value
−1
−2
−3


Expressed lincRNAs space:
Enrichment of TAD bins in lincRNAs and pcgenes
elincRNA
promoter
promoter + gene body
Fold enrichment
TAD bins
Log10
q−value
−1
−2


Expressed lincRNAs space:
Enrichment of lincRNAs and pc genes in TAD bins
elincRNA
promoter
promoter + gene body
Fold enrichment
TAD bins
Log10
q−value
−0.5
−1.0
−1.5
−2.0
−2.5

Analyses using CTCF/Cohesin binding sites





Conclusions

Whole genome:

epc and non-epc genes promoters are enriched at TAD borders and depleted indide TADs. When taking the gene bodies in addition of the promoters, the trend is weaker.

It is hard to detect patterns with nelinc and elincRNAs promoters alone as the p-values returned are often high. When taking the gene bodies into account, the p-values get better but there seem to be an overall depletion of genes throughout TAD. elincRNAs seem to be especially depleted in the middle of the TADs. The depletion seems to be less intense on the left border of TADs. These issues may be caused by the weak number of lincRNAs compared to pc genes.

Intergenic space:

Same kind of pattern, the p-values are higher. Note the strong depletion on bin 10 is still present for elincRNAs promoters only while nelincRNAs seem to be enriched at this position.

Binding sites:

elincRNAs promoters are strongly (6-8 fold) enriched in CTCF/cohesin binding sites while nelincRNAs promoters are only enriched 2-3 folds. The promoters of epc genes are less enriched than those of elincRNAs (~6 folds). Note that among all "gene bodies + promoters", those of elincRNAs are those with the strongest enrichment.