Report 3: Redefining boundaries

Cyril Matthey-Doret

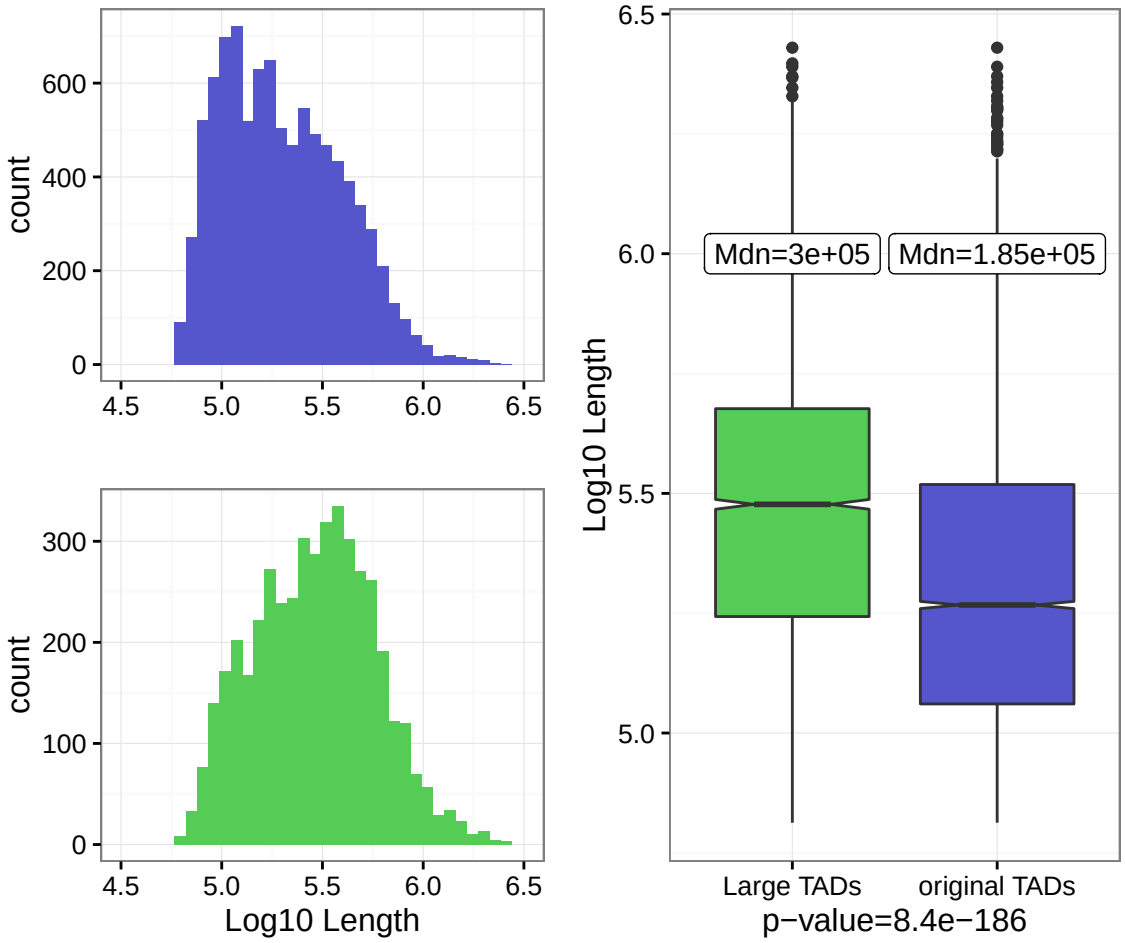
11 octobre 2016

Introduction

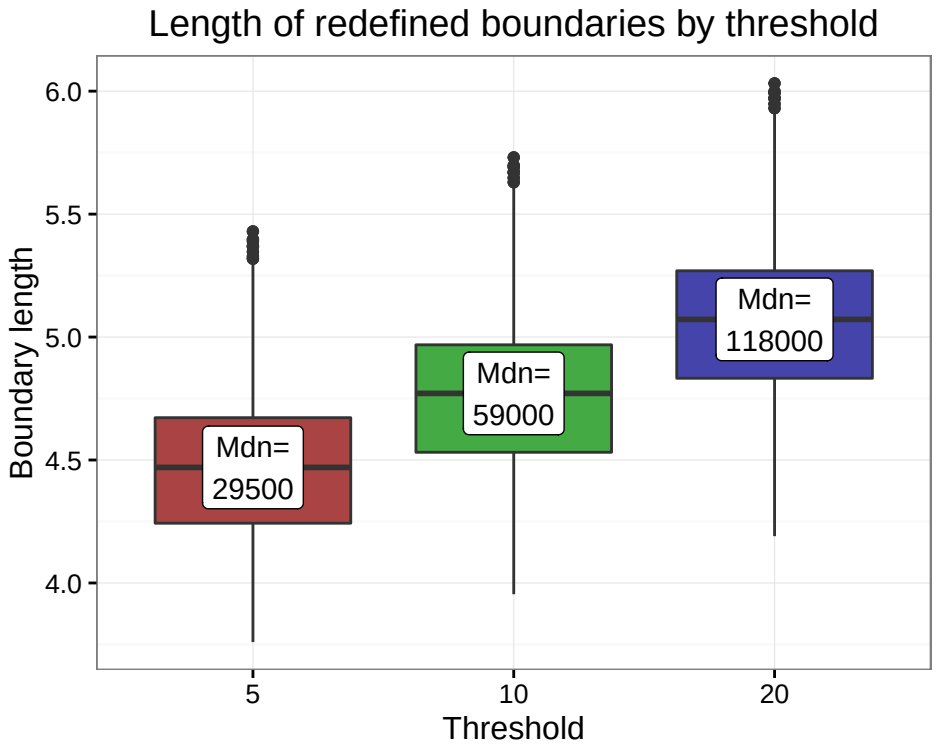
Visualising and quantifying the distribution of TAD lengths, or the areas of ovelap between lincRNAs and TADs might help finding a better threshold for the definition of TAD boundaries. This report aims to provide support for the definition of those boundaries.

1. Length distribution of TADs, boundaries and gaps

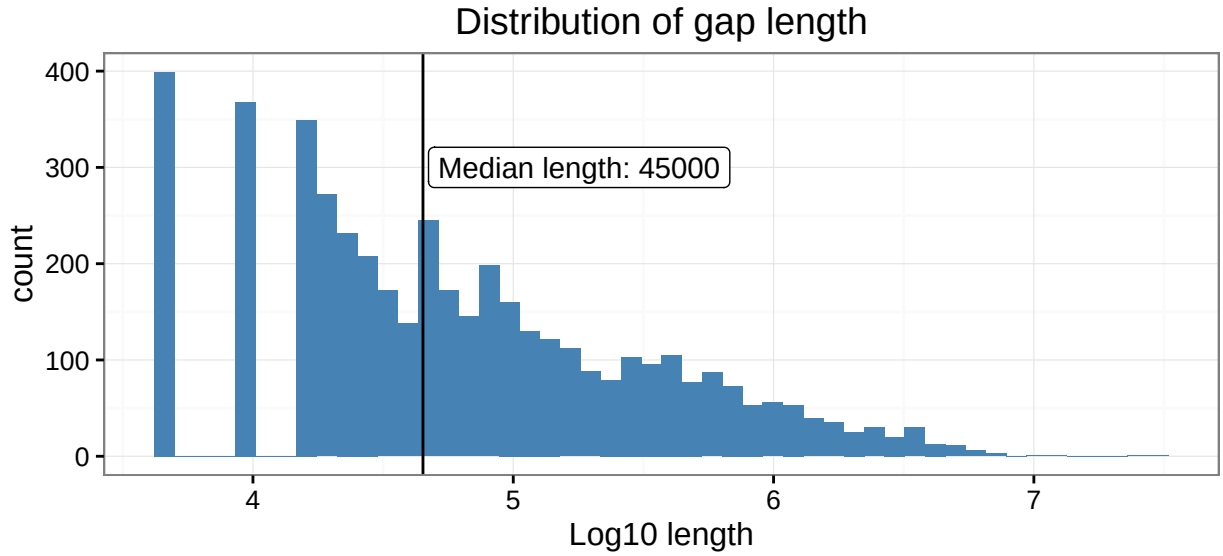
- Length of TADs: TADs often contained many smaller TADs which causes the boundaries to overlap and makes it harder to distinguish between TAD-bound and non-TAD-bound RNAs. All overlapping TADs were merge so that only large domains would be considered and there would not be any boundaries contained inside another domain. The mean of the length distribution for the newly generated TADs (Large TADs) is significantly higher.



- Length of TAD boundaries: TAD boundaries have been redefined so that they are more flexible. Again, three different thresholds have been used to define them, however a boundary cannot overlap another. A TAD boundary will always reach only as far as the smallest value between the threshold and half the gap between its TAD and the neighbour.



- Length of gaps: The distribution of gap length show that the gaps are generally smaller than the TADs. The median length of gaps is 6.67 times smaller than for TADs. The discrete distribution observed on a log scale is caused by the approximations that are used for defining the border of TADs.

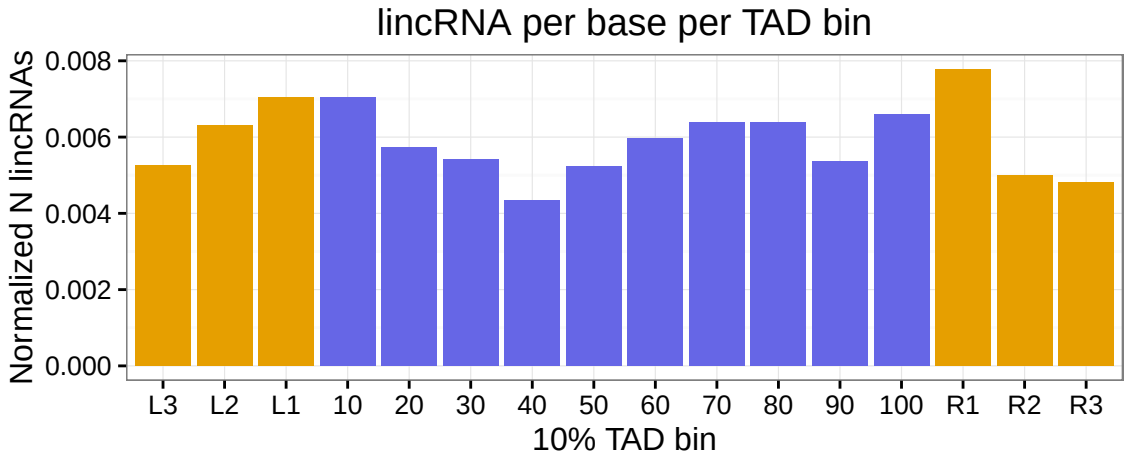


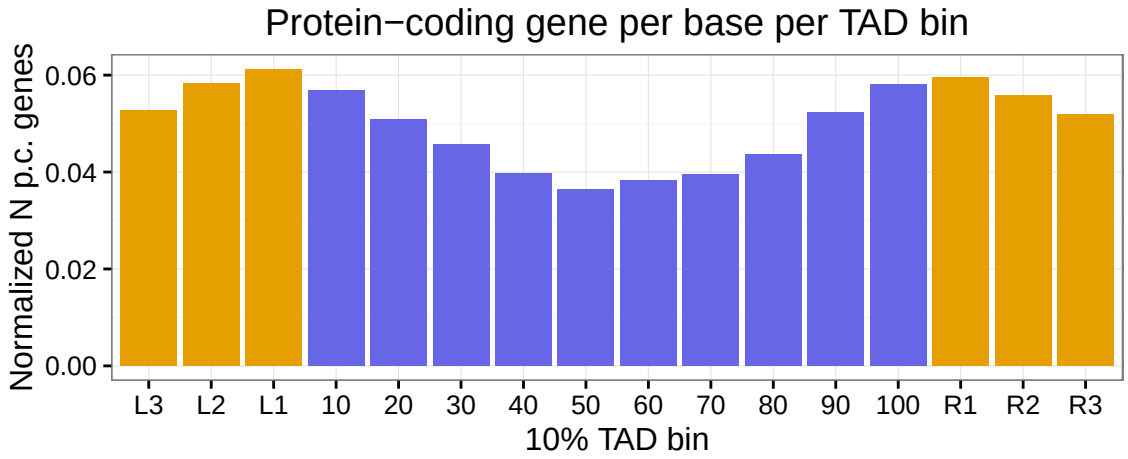
2. Overlaps between RNA and areas of TADs

TADs were divided into 10 bins, plus 3 outer bins on both sides. The length of each inner bin is 10% of its TAD length. For outer bins, the length was adjusted so that it would not overlap with the next TAD. If half the distance until the next TAD was larger than 30% (i.e. 3 bins of 10%) of TAD length, the size of outer bins remains unchanged. Otherwise, half the gap between neighbouring TADs is split into the 3 bins, meaning each bin is resized to one sixth of its TAD length.

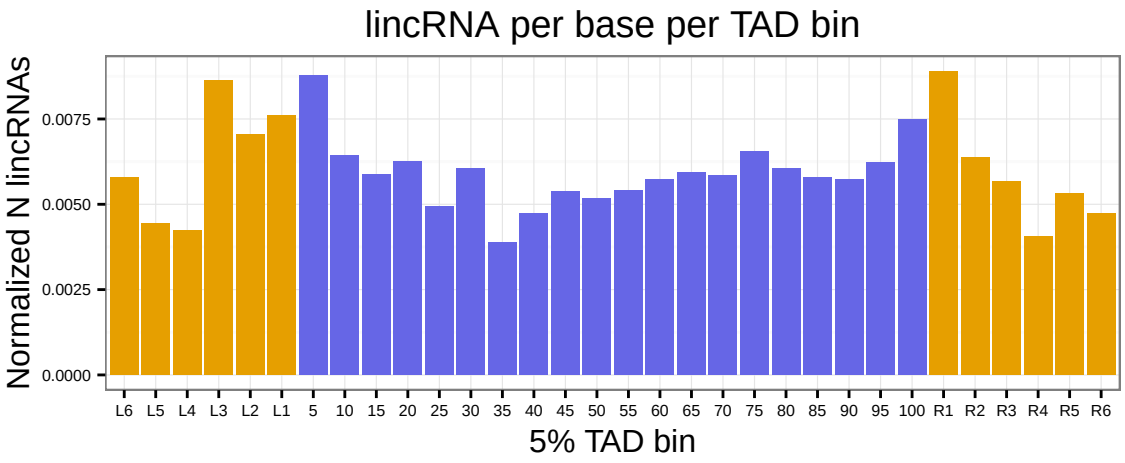
In order to know in which area of TADs lincRNAs and protein coding genes are more likely to be expressed, all of these genes with at least 25% of their sequence overlapping a bin were counted. The total number of lincRNAs or protein-coding genes that overlap with each bin was then calculated. Those steps were performed at once using BEDtools intersect version 2.25.

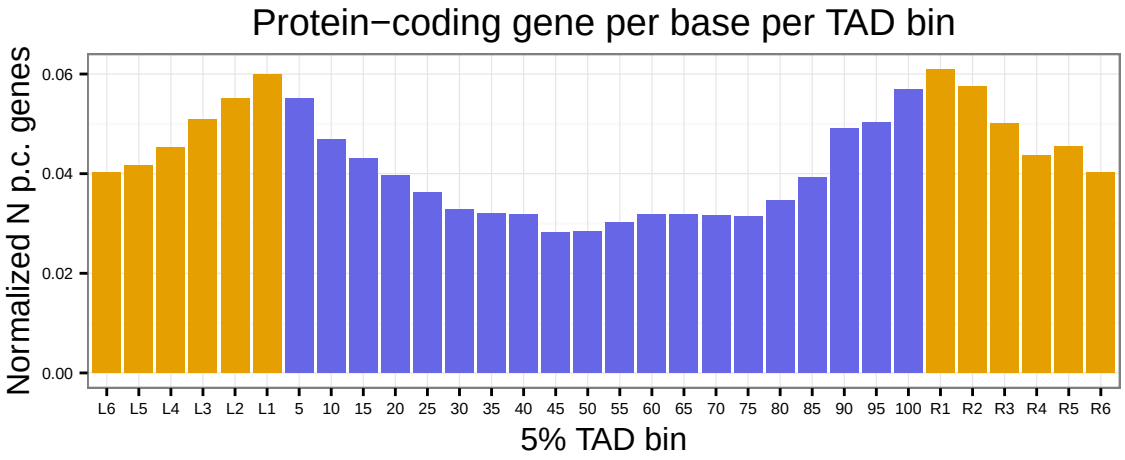
Using the raw number of genes overlapping each bin would be biased since longer TADs have have longer bins, and those bins will have more weight on the plot. The barplots below show the values after they have been normalized by bin length. The bins depicted in orange are located outside TADs, while the ones in blue are inside. All bins beginning by L are on the left (before the TAD), while those beginning with R are on the right (after the TAD).

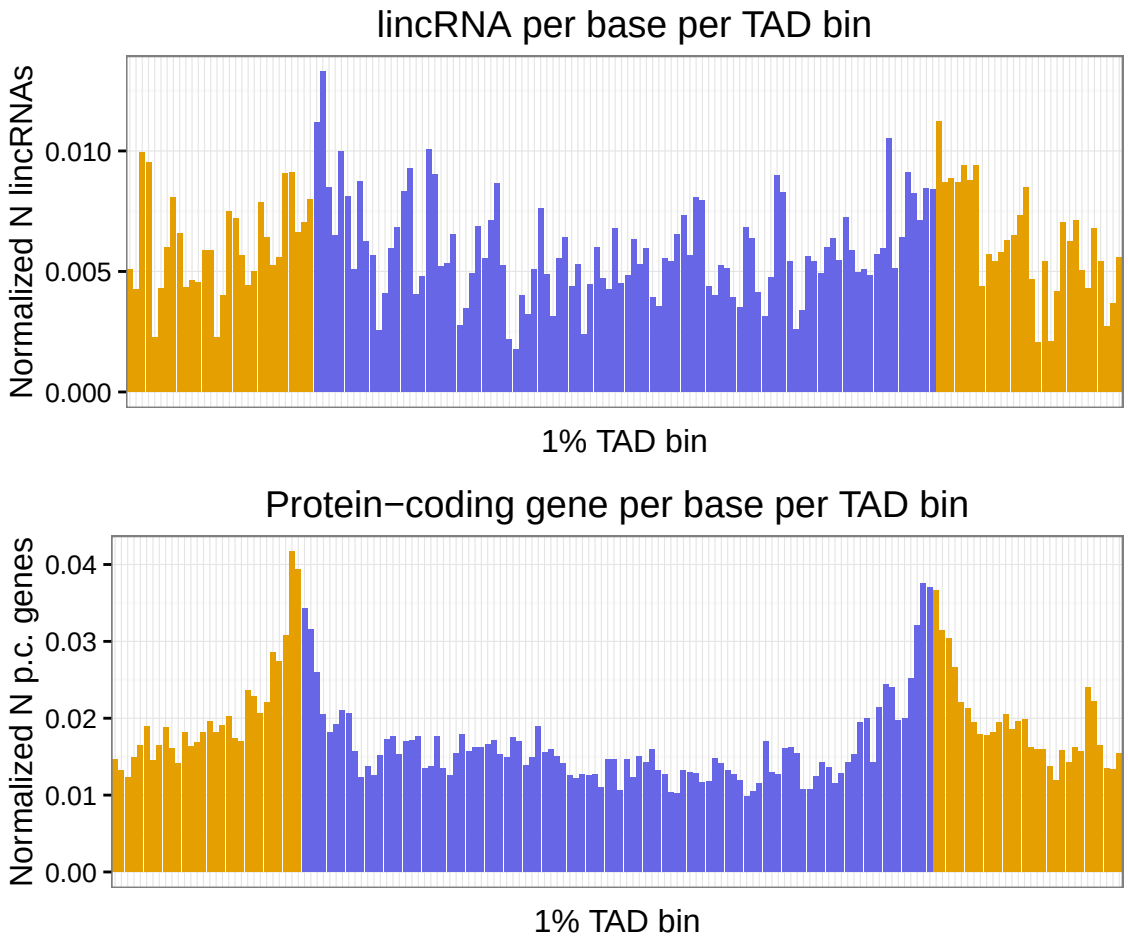




Below are the same barplots, using respectively 5% and 1% for binwidth instead of 10%. Having different resolutions for the detection of genes location in TAD boundaries will help defining a better threshold for the TAD boundaries.

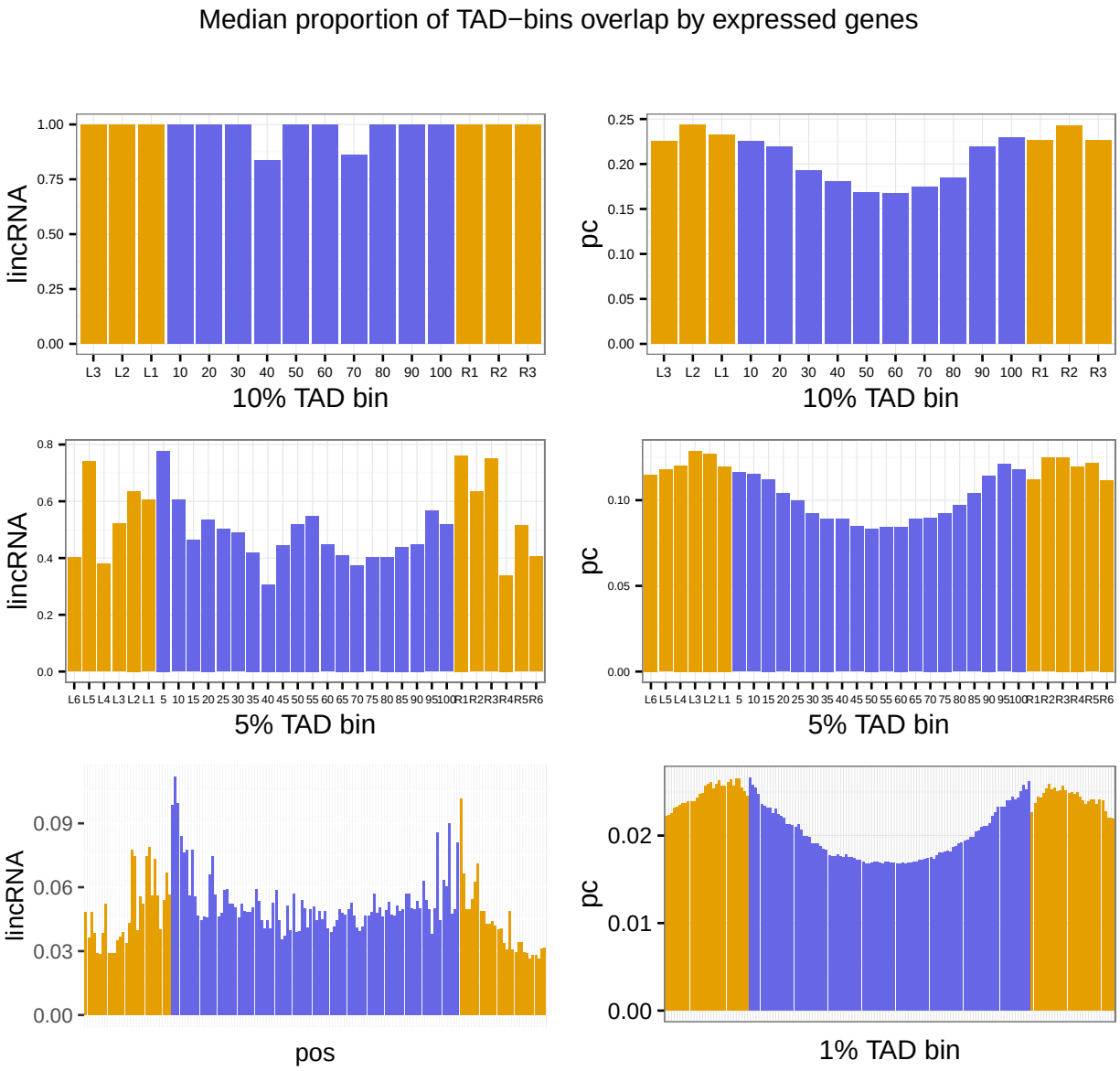

lincRNA per base per TAD bin
Normalized N lincRNAs
0.0075
0.0050
0.0025
0.0000
L6 L5 L4 L3 L2 L1 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 R1 R2 R3 R4 R5 R6
5% TAD bin


Protein−coding gene per base per TAD bin
Normalized N p.c. genes
0.06
0.04
0.02
0.00
L6 L5 L4 L3 L2 L1 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 R1 R2 R3 R4 R5 R6
5% TAD bin



3. Length of overlap per bin

The distribution of median overlap length betweeen RNAs and bins may help define TAD boundaries. This is done by looking at all overlaps between TAD bins and RNAs without any minimum overlap requirement, and computing the median overlap length for each bin. The values are then normalized by dividing overlap length by the length of the RNA. The resulting values represent the portion of genes length that overlap a bin.



Bins located at the borders of TADs tend to contain higher portion of the sequence of protein-coding genes. It seems to be the case for lincRNAs as well, but the lincRNAs also have regular drops inside the TADs. These regular drops might be linked to the small length of lincRNA compared to protein coding genes (median length: 2125 vs 31128.5). They could also be caused by smaller TAD's located inside TADs which have disappeared when they were merged together.

The 10% almost all contain 100% of the lincRNAs sequence, suggesting they are much larger, but only between 16.7 and 24.4% of protein-coding genes. The 5% bins contain between 30.7 and 77.8% of lincRNAs sequence, and between 8.3 and 12.9% of protein-coding genes. 1% bins contain betwteen 2.6 and 11.2% of lincRNA, and between 1.7 and 2.7% of protein-coding genes.

4. Determination of TAD boundaries:

From the results obtained in **2**, it seems 10% bins are larger than the area that is enriched in expressed genes and may be too large to be used as TAD-boundaries. On the other hand, the 1% bins capture well the differences, but using 1% TAD boundaries would not englobe the whole area. The best threshold among those 3 may be 5%, which would offer a fair trade off between size and accuracy.

Results from **3** yield the following summary statics for sequences overlapping 5% bins:

```
summary(medover_len5$lincRNA)  # Proportion of lincRNAs sequence overlapping 5% TAD bins.
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3070  0.4096  0.4956  0.5099  0.5762  0.7784
```

```
summary(medover_len5$pc)  # Proportion of protein-coding genes sequence overlapping  5% TAD bins.
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.08333 0.09169 0.11200 0.10700 0.11960 0.12890
```

Next: How to choose minimum overlapping requirement.