Creating Feature Space

Spiro Stilianoudakis

August 16, 2018

Loading Libraries

```
## Loading required package: stats4

## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, cbind, colMeans,
##     colnames, colSums, do.call, duplicated, eval, evalq, Filter,
##     Find, get, grep, grepl, intersect, is.unsorted, lapply,
##     lengths, Map, mapply, match, mget, order, paste, pmax,
##     pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce,
##     rowMeans, rownames, rowSums, sapply, setdiff, sort, table,
##     tapply, union, unique, unsplit, which, which.max, which.min

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:base':
##
##     expand.grid

## Loading required package: IRanges

## Loading required package: GenomeInfoDb

## Warning: package 'caret' was built under R version 3.4.4

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.4.4

## Warning: package 'data.table' was built under R version 3.4.4
```

```
##
## Attaching package: 'data.table'

## The following object is masked from 'package:GenomicRanges':
##
##     shift

## The following object is masked from 'package:IRanges':
##
##     shift

## The following objects are masked from 'package:S4Vectors':
##
##     first, second

## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:caret':
##
##     cluster

## Loading required package: splines

## Loaded gbm 2.1.3

## Warning: package 'pROC' was built under R version 3.4.4

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:IRanges':
##
##     cov, var

## The following objects are masked from 'package:S4Vectors':
##
##     cov, var

## The following object is masked from 'package:BiocGenerics':
##
##     var

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

##
## Attaching package: 'plyr'

## The following object is masked from 'package:IRanges':
##
##     desc

## The following object is masked from 'package:S4Vectors':
##
##     rename

## Warning: package 'dplyr' was built under R version 3.4.4
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Warning: package 'knitr' was built under R version 3.4.4
```

Reading in data

Adding genomic annotations from Merlot folder (for 1kb bins)

3D subcompartments

DGV

GERP

nestedRepeats

super_enhancers

UCNE

VMR

BroadHMM

Combined

DNase I

Histone Modifications

Adding Chromosome information to the data

Saving the data

Adding distance from region to TAD boundary for each feature

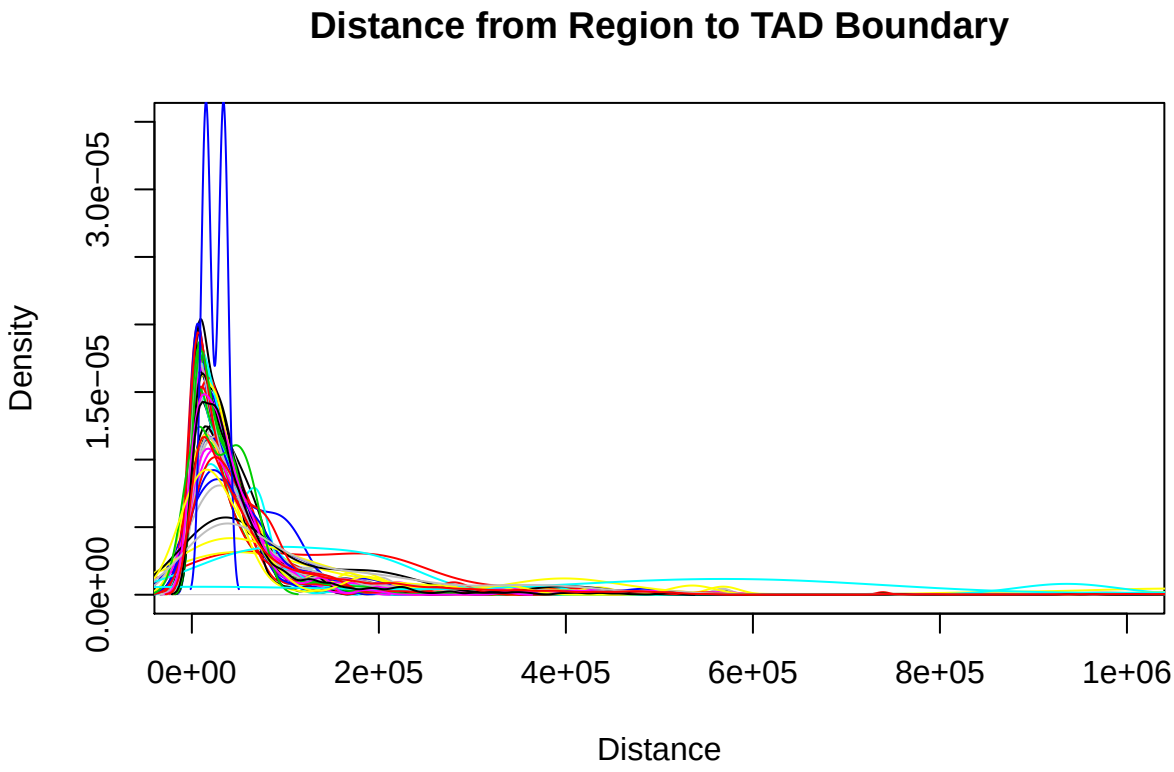
Distance Summaries for each feature from region to TAD boundary

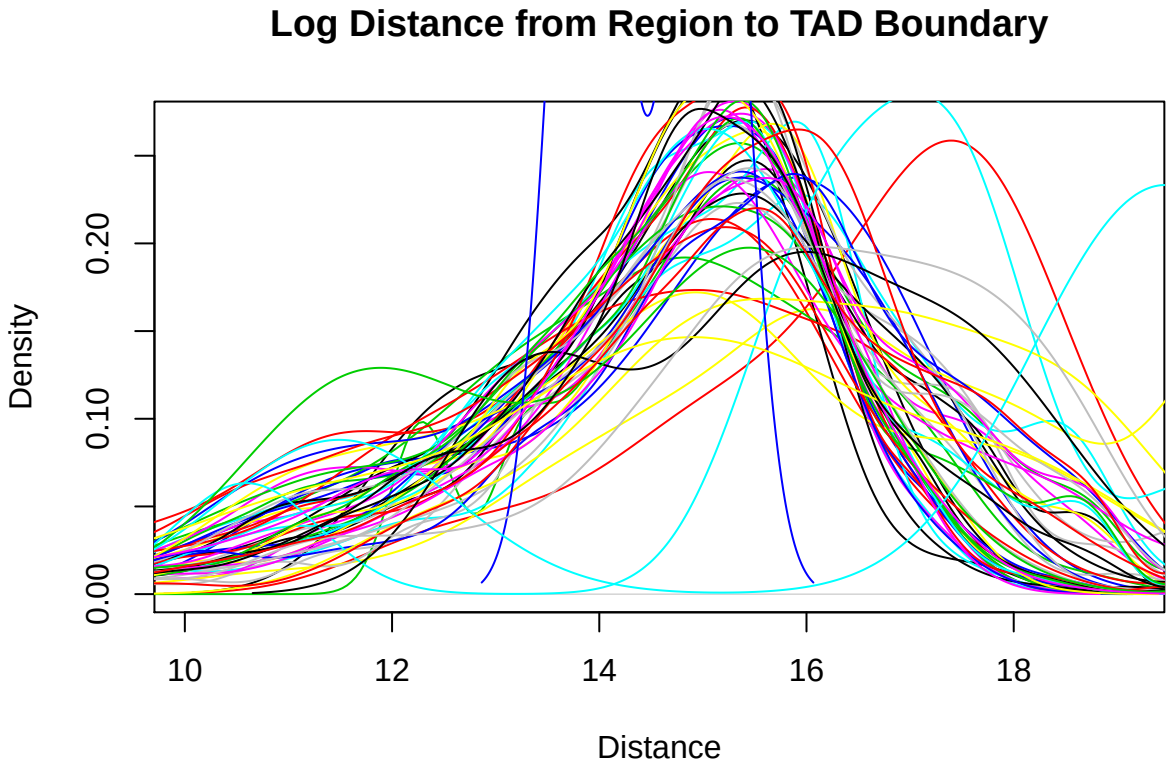
Feature	Mean	Median	Range	MeanLog	MedianLog	RangeLog
A_gr_center	39665.7	34999.0	115000	14.7	15.1	14.9
B_gr_center	141189.5	129999.0	420000	16.4	17.0	16.7
binslist10_center	96994.5	40000.0	1380000	14.8	15.3	20.4
complex_gr_center	50040.1	42910.0	104223	14.9	15.4	14.7

Feature	Mean	Median	Range	MeanLog	MedianLog	RangeLog
deletion_gr_center	98865.4	51675.5	1331261	15.4	15.7	18.3
DNA_gr_center	76622.6	39493.0	1275877	15.1	15.3	18.3
duplication_gr_center	325242.6	91199.0	1360691	16.6	16.5	18.4
gain_loss_gr_center	144578.7	47116.5	1303066	15.5	15.5	18.3
gerp_gr_center	71268.0	37490.0	1326011	15.0	15.2	19.3
Gm12878_ActivePromoter_gr_center	31754.1	19945.0	257999	13.8	14.3	18.0
Gm12878_CTCF_gr_center	45905.9	23575.0	552599	14.2	14.5	19.1
Gm12878_DNaseI_gr_center	39589.3	26996.5	485999	14.3	14.7	18.9
Gm12878_E_gr_center	34787.4	25685.0	245564	14.2	14.6	15.9
Gm12878_H2az_gr_center	38613.6	28500.0	574185	14.4	14.8	17.1
Gm12878_H3k27ac_gr_center	40193.7	28350.0	526308	14.4	14.8	17.0
Gm12878_H3k27me3_gr_center	70755.4	40267.0	574169	15.2	15.3	17.1
Gm12878_H3k36me3_gr_center	33895.6	24578.0	574862	14.2	14.6	17.1
Gm12878_H3k4me1_gr_center	38835.9	28107.0	574810	14.4	14.8	17.1
Gm12878_H3k4me2_gr_center	37168.9	26267.0	256760	14.2	14.7	16.0
Gm12878_H3k4me3_gr_center	38795.3	27034.0	566289	14.3	14.7	17.1
Gm12878_H3k79me2_gr_center	37284.7	27591.0	256381	14.4	14.8	16.0
Gm12878_H3k9ac_gr_center	35621.8	26750.0	569711	14.2	14.7	17.1
Gm12878_H3k9me3_gr_center	76702.5	40780.0	1337180	15.1	15.3	18.4
Gm12878_H4k20me1_gr_center	54420.1	33127.5	573873	14.7	15.0	17.1
Gm12878_Heterochromlo_gr_center	74497.6	36522.0	1357864	14.9	15.2	18.4
Gm12878_Insulator_gr_center	61610.6	24672.0	1340599	14.4	14.6	20.4
Gm12878_PF_gr_center	37028.0	23523.5	1328737	14.1	14.5	18.3
Gm12878_PoisedPromoter_gr_center	50145.1	34145.0	239156	14.7	15.1	15.9
Gm12878_R_gr_center	76504.8	35895.5	1374244	14.9	15.1	20.4
Gm12878_RepetitiveCNV14_gr_center	105972.4	35499.0	1232100	15.2	15.1	18.2
Gm12878_RepetitiveCNV15_gr_center	92939.9	28072.0	1057900	14.7	14.8	18.0
Gm12878_Repressed_gr_center	59569.6	39849.0	558754	15.1	15.3	17.1
Gm12878_StrongEnhancer4_gr_center	35078.2	27699.0	333200	14.3	14.8	16.3
Gm12878_StrongEnhancer5_gr_center	45489.3	30000.0	1132300	14.6	14.9	18.1
Gm12878_T_gr_center	75862.1	35005.5	1370679	14.9	15.1	20.4
Gm12878_TSS_gr_center	32064.3	22608.5	207402	13.9	14.5	15.7
Gm12878_TxnElongation_gr_center	38780.0	26872.0	344903	14.4	14.7	16.4
Gm12878_TxnTransition_gr_center	36674.4	28749.0	188136	14.4	14.8	15.5
Gm12878_WE_gr_center	41395.5	29156.0	332091	14.6	14.8	16.3
Gm12878_WeakEnhancer6_gr_center	44759.3	27753.0	1328454	14.4	14.8	18.3
Gm12878_WeakEnhancer7_gr_center	46266.1	30972.0	1133599	14.6	14.9	20.1
Gm12878_WeakPromoter_gr_center	39419.3	21833.0	1328646	14.0	14.4	18.3
Gm12878_WeakTxn_gr_center	42599.6	28555.0	1225699	14.4	14.8	20.2
insertion_gr_center	87069.7	47619.0	488680	15.4	15.5	16.9
inversion_gr_center	206239.2	121922.0	934330	16.5	16.9	17.8
line_gr_center	98573.8	44988.5	1355431	15.3	15.5	19.4
low_complexity_gr_center	158201.9	65460.0	1046216	16.2	16.0	18.0
LTR_gr_center	93192.2	39546.0	1374059	15.2	15.3	18.4
mobile_element_insertion_gr_center	99756.4	60939.0	412222	15.7	15.9	16.7
novel_sequence_insertion_gr_center	70321.8	44422.0	548194	15.2	15.4	17.1
other_gr_center	30278.3	33473.0	65993	14.0	15.0	14.1
RC_gr_center	24488.0	24488.0	18734	14.5	14.5	13.0
satellite_gr_center	617541.8	569605.5	1325206	17.5	19.1	18.3
se_GM12878_gr_center	39182.6	29581.0	335474	14.3	14.9	16.4
sequence_alteration_gr_center	83753.3	29191.0	534983	14.5	14.8	17.0
simple_repeat_gr_center	124784.4	78848.5	946792	15.9	16.3	17.9

Feature	Mean	Median	Range	MeanLog	MedianLog	RangeLog
SINE_gr_center	58907.1	32164.0	1339319	14.8	15.0	18.4
VMR_gr_center	105935.3	43624.0	1374836	15.3	15.4	18.4

Plots


Distance from Region to TAD Boundary
Density
0.0e+00
1.5e−05
3.0e−05
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
Distance


Log Distance from Region to TAD Boundary
Density
0.00
0.10
0.20
10
12
14
16
18
Distance