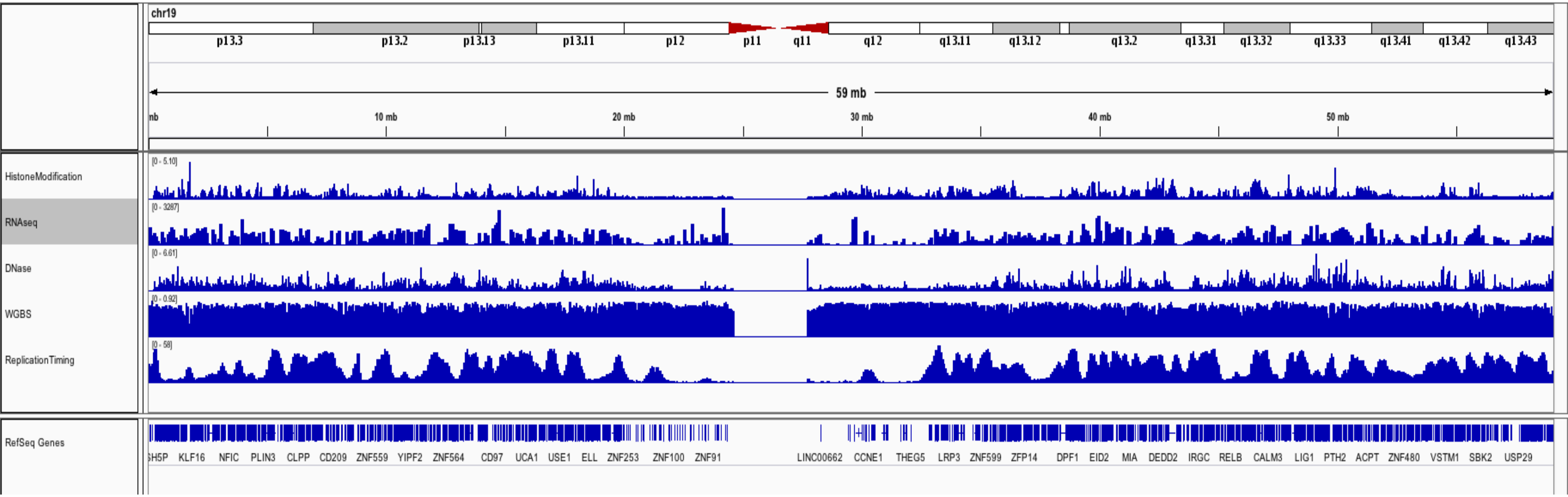
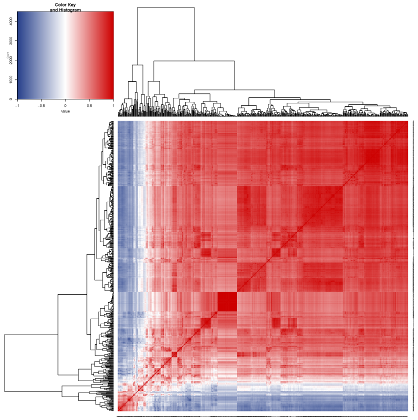
BMR Estimation using ENCODE data

JZ & DL

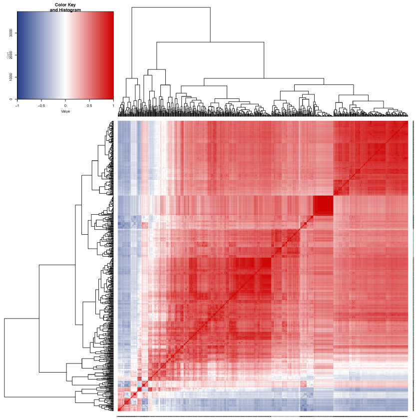
Genomic features have very different range and resolution



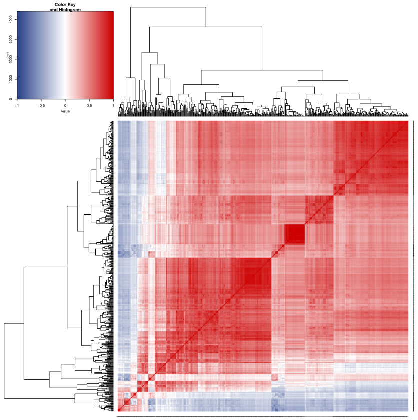
- Replication timing is generally at way large resolution than RNAseq & DHS
- Correlation analysis at bins with different length will be confounded by variant density
- But averaging over large range will sometimes destroy the signal structures.



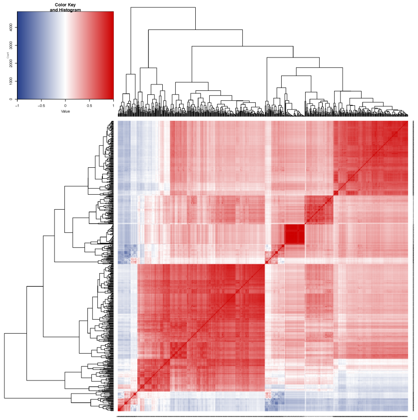
1mb



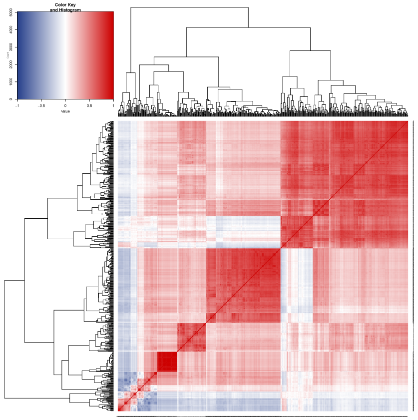
100kb



50kb



10kb



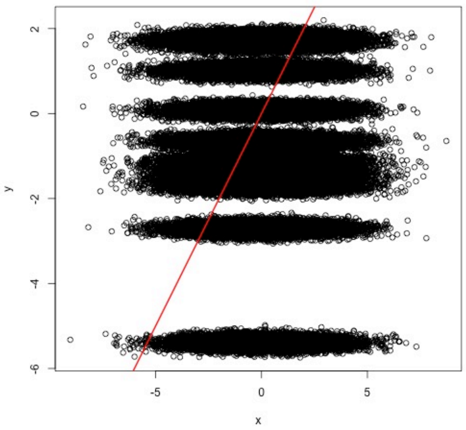
5kb

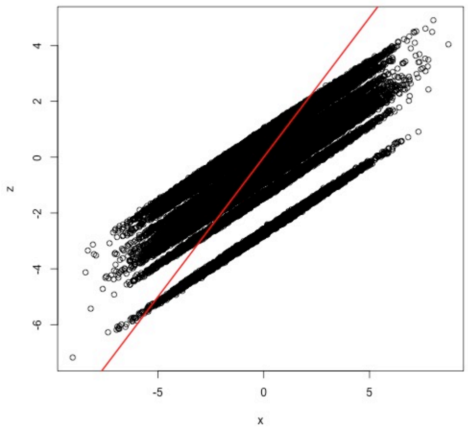
A simple example of multi-resolution issue

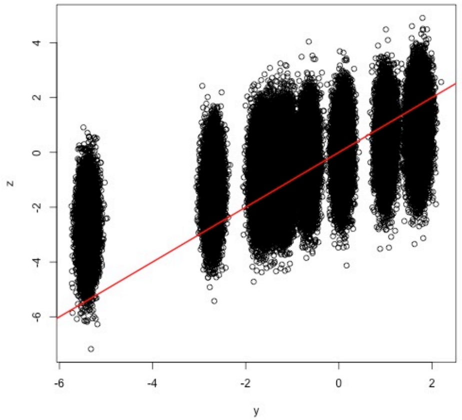
Z = 0.5x + 0.5y
X: Gaussian(0,2) at 1 bp resolution
Y: k1+k2
k1:Gaussian(0,2) at 10 bp resolution,
k2 Gaussian(0,0.1) at 1 bp local fluctuation

	Cor(x,z)	**Cor(y,z)**
1bp	0.693384	0.718598
100bp	0.202546	0.976674

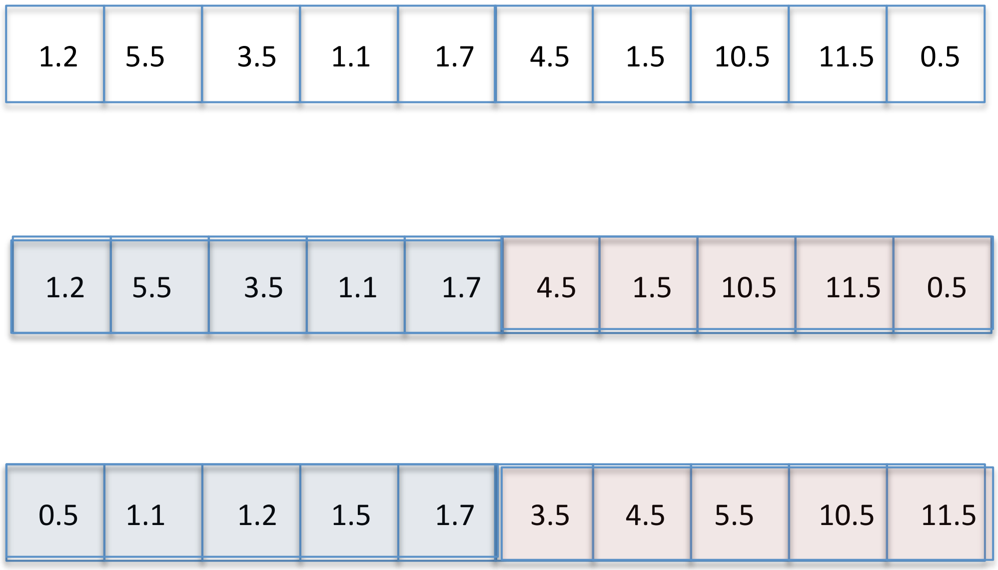
y1	**y2**	**y3**	**y4**	**y5**	**y6**	**y7**	**y8**	**y9**	**y10**
K1,1+ k2,1	K1,1+ k2,2	K1,1+ k2,3	K1,1+ k2,4	K1,1+ k2,5	K1,1+ k2,6	K1,1+ k2,7	K1,1+ k2,8	K1,1+ k2,9	K1,1+ k2,10
x1	x2	x3	x4	x5	x6	x7	x8	x9	x10



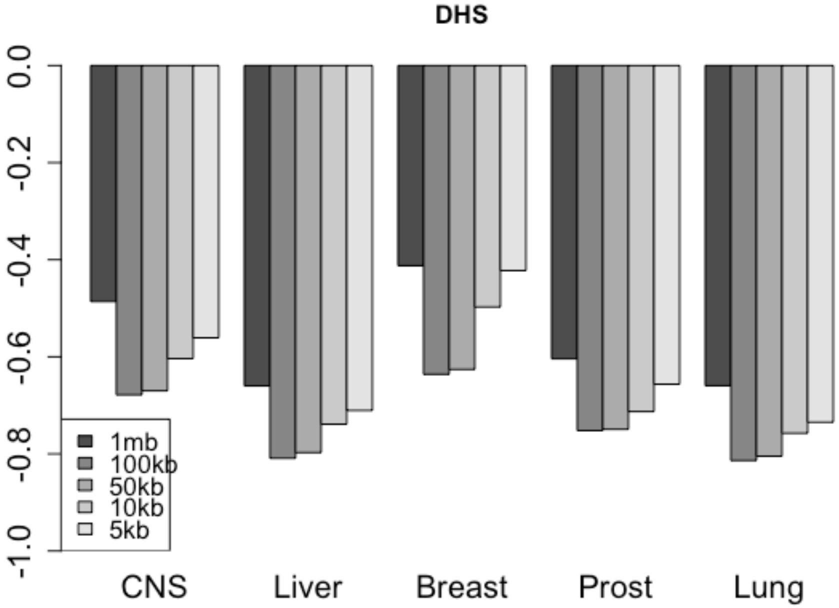




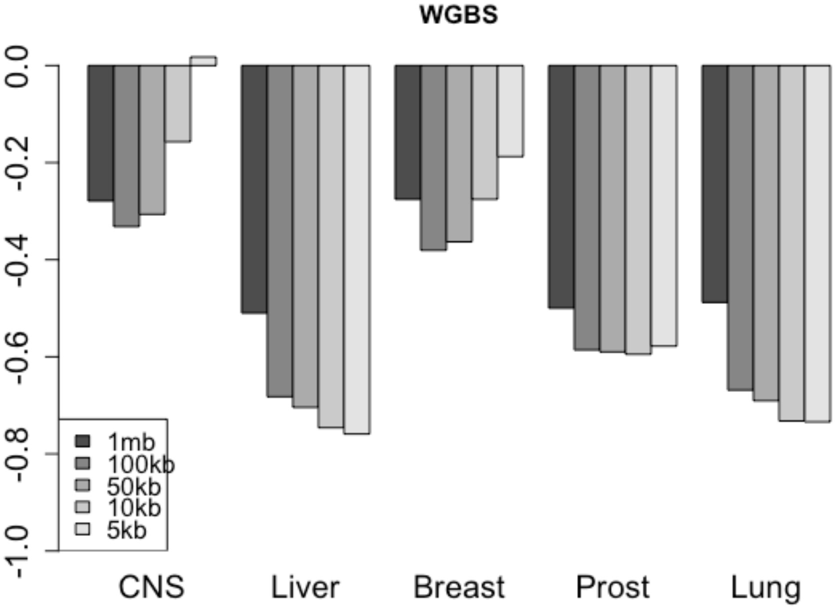
How to avoid the confounding from variant number



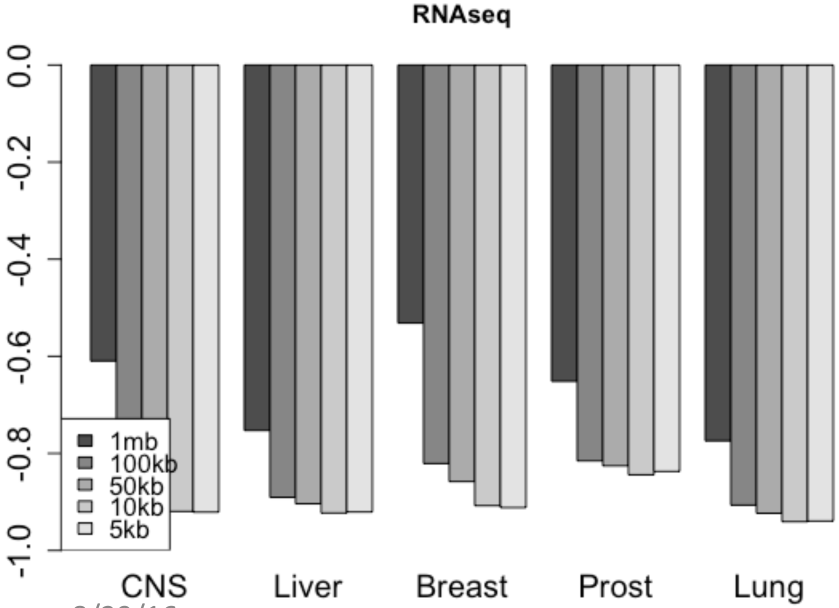
1.2	5.5	3.5	1.1	1.7	4.5	1.5	10.5	11.5	0.5	Original
1.2	5.5	3.5	1.1	1.7	4.5	1.5	10.5	11.5	0.5	5bp resolution
0.5	1.1	1.2	1.5	1.7	3.5	4.5	5.5	10.5	11.5	1bp resolution



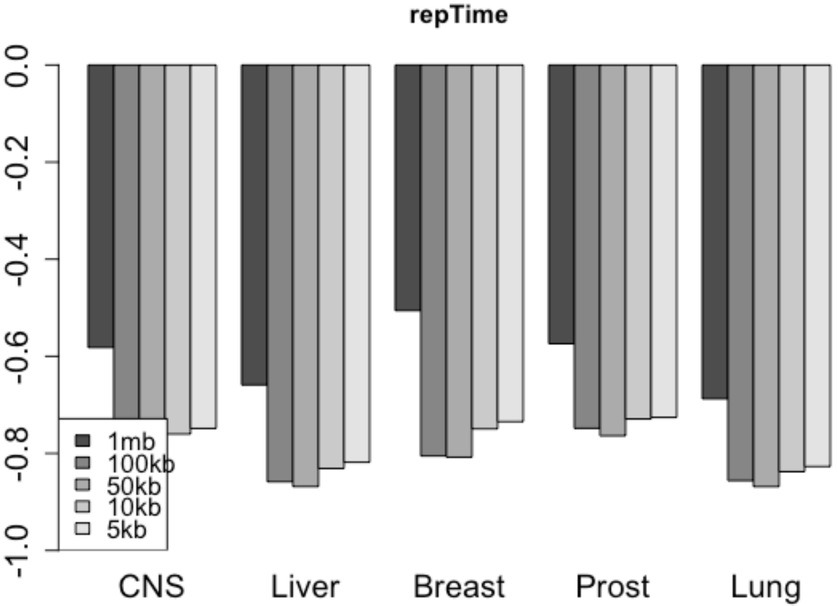
DHS



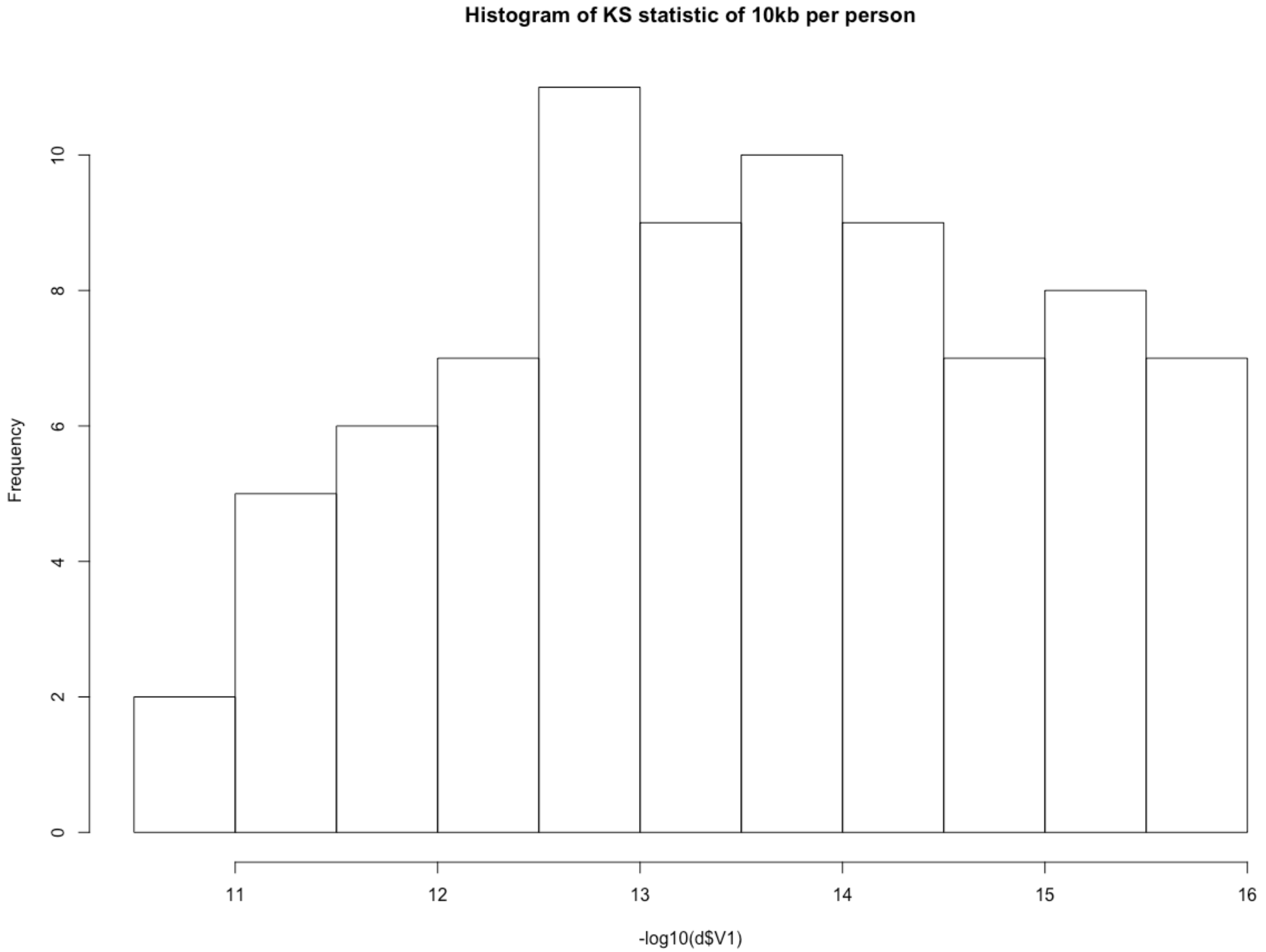
WGBS



RNAseq



repTime

Histogram of KS statistic of 10kb per person
Frequency
0
2
4
6
8
10
11
12
13
14
15
16
-log10(d$V1)