XH-pi frequency vs resolution

Emma Rand

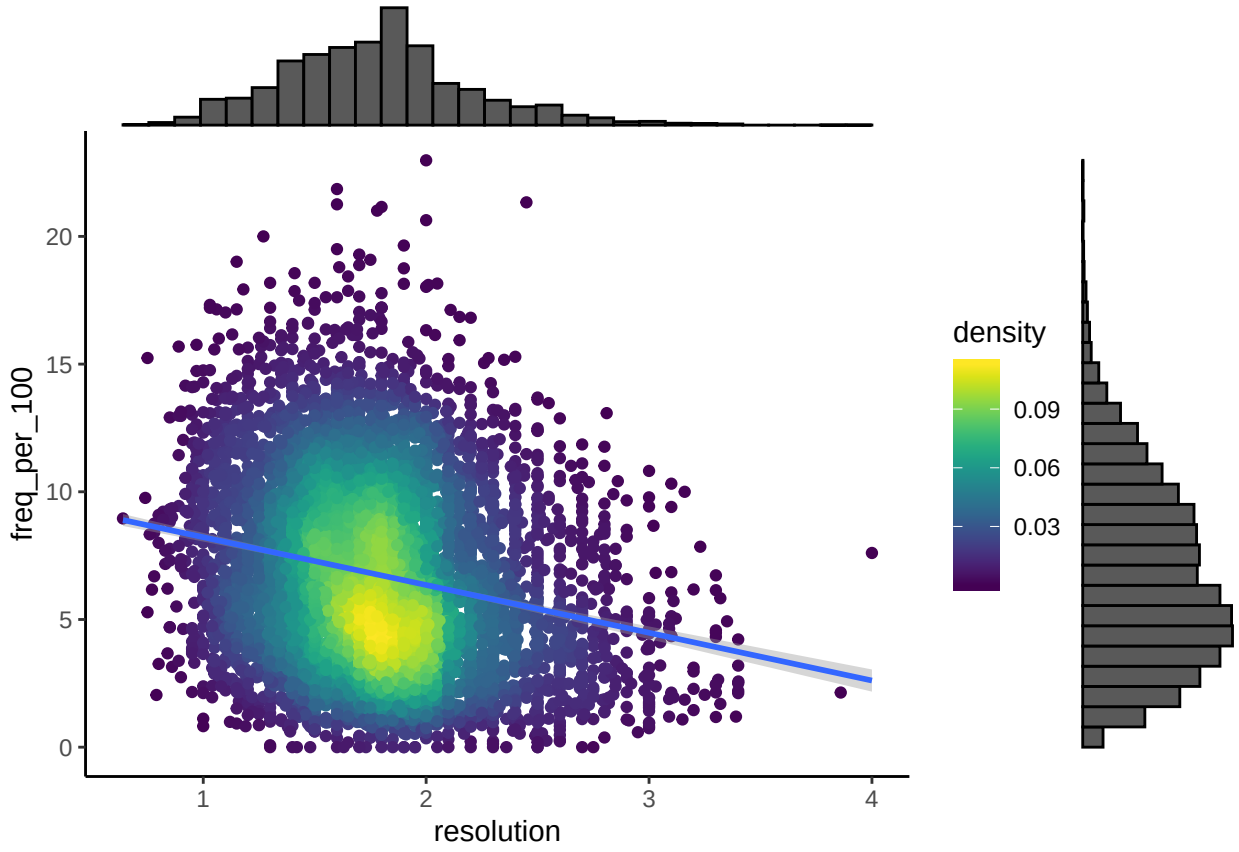
22/08/2019

Contents

These data have 7306 structures.

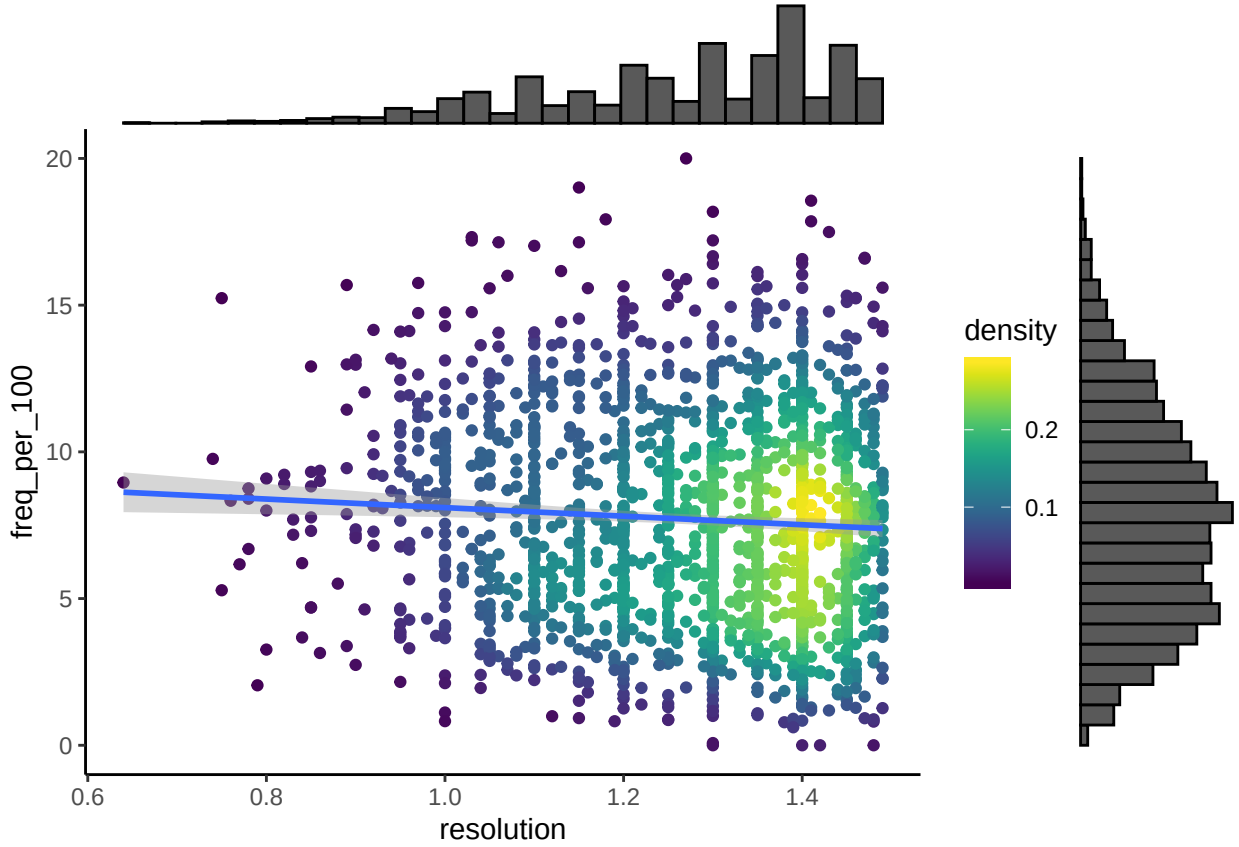
The linear model of `freq_per_100 ~ resolution` has the equation: freq_per_100 = 10.0917671 -1.8706521 * resolution

The sample size makes it highly significant but the R squared value is low (0.0475239) i.e., very little of the variation in the number of interactions can be explained by the resolution.

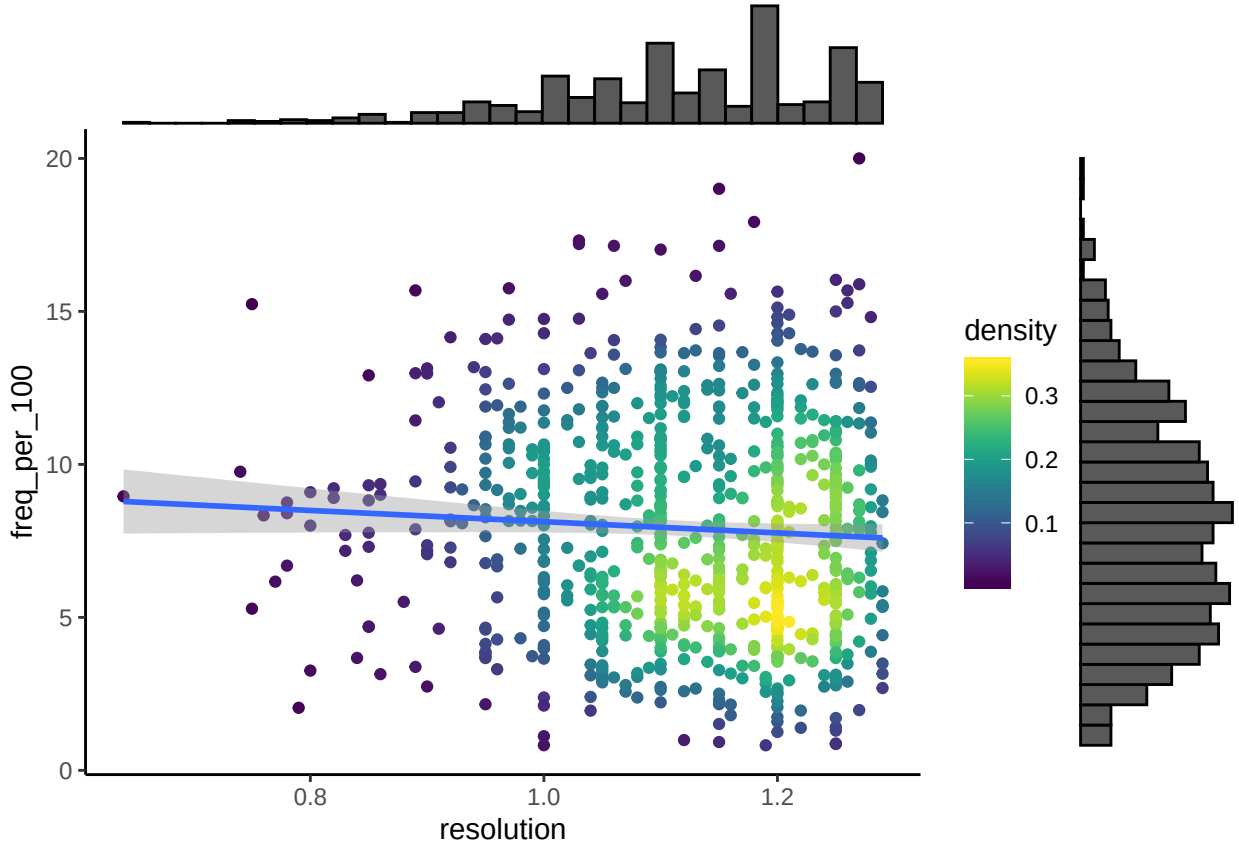


When filtering the data for resolutions less than 1.5A there are 1708 structures and the linear model of `freq_per_100 ~ resolution` has the equation: freq_per_100 = 9.5570776 -1.4576012 * resolution

which is still significant with an R squared value of (0.0043161).



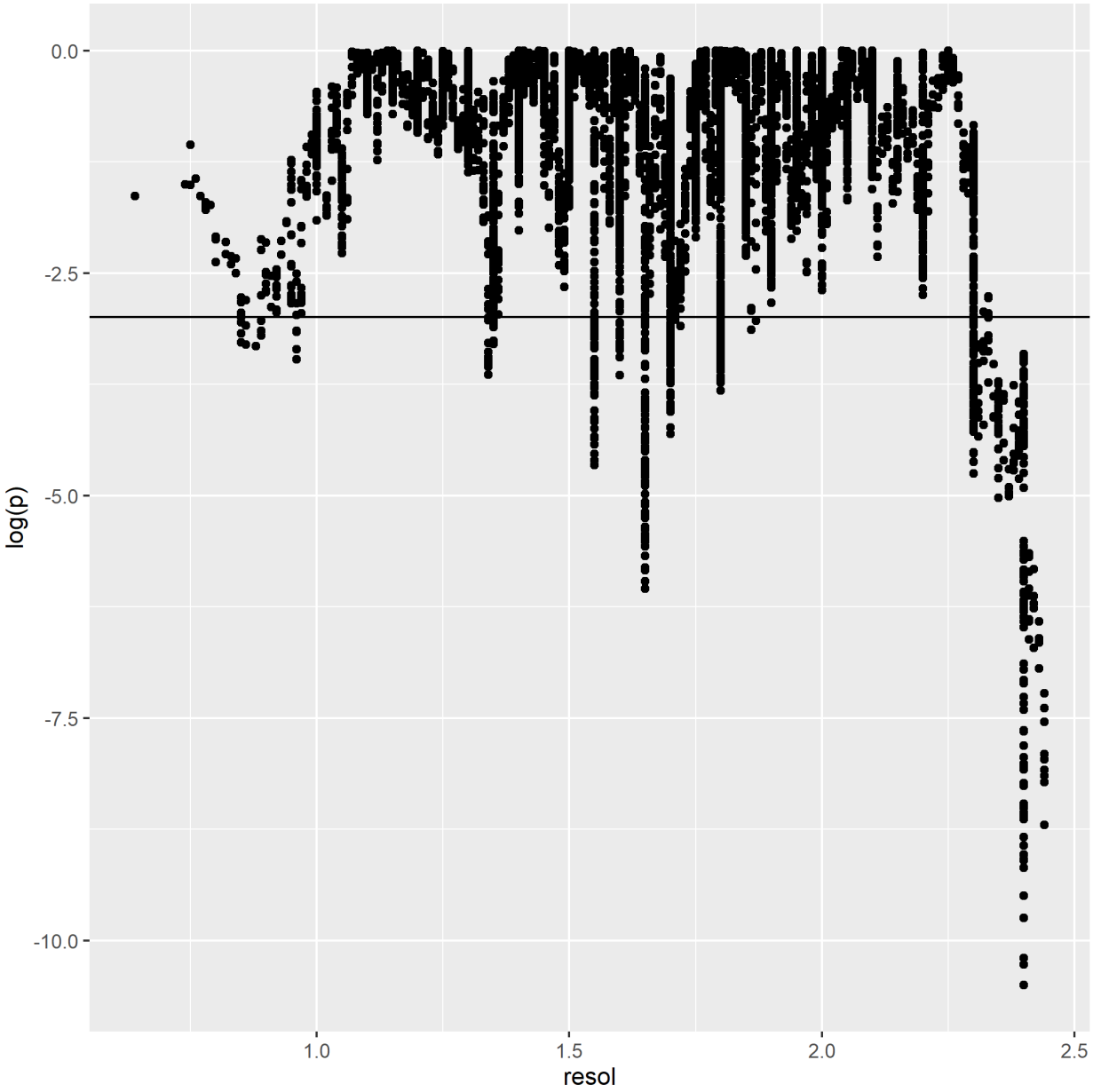
At a cut-off resolution of 1.3A there are 777 and the relationship disappears.



1 Moving Window regression

1.1 Fixed first-500 window of sorted resolutions

Regressions were carried out for a moving window of 500 observations, i.e., the size of resolution window is dependent on the number of observations

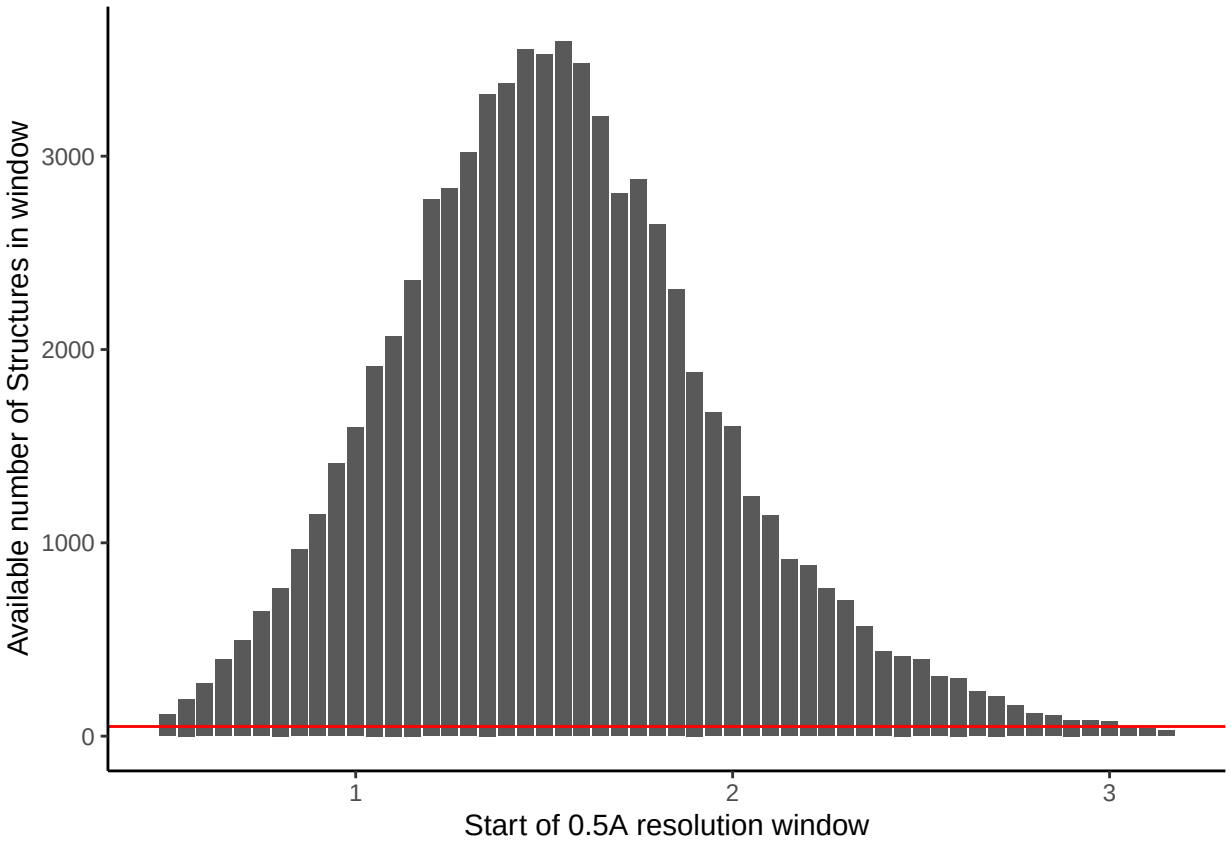


1.2 Fixed resolution window

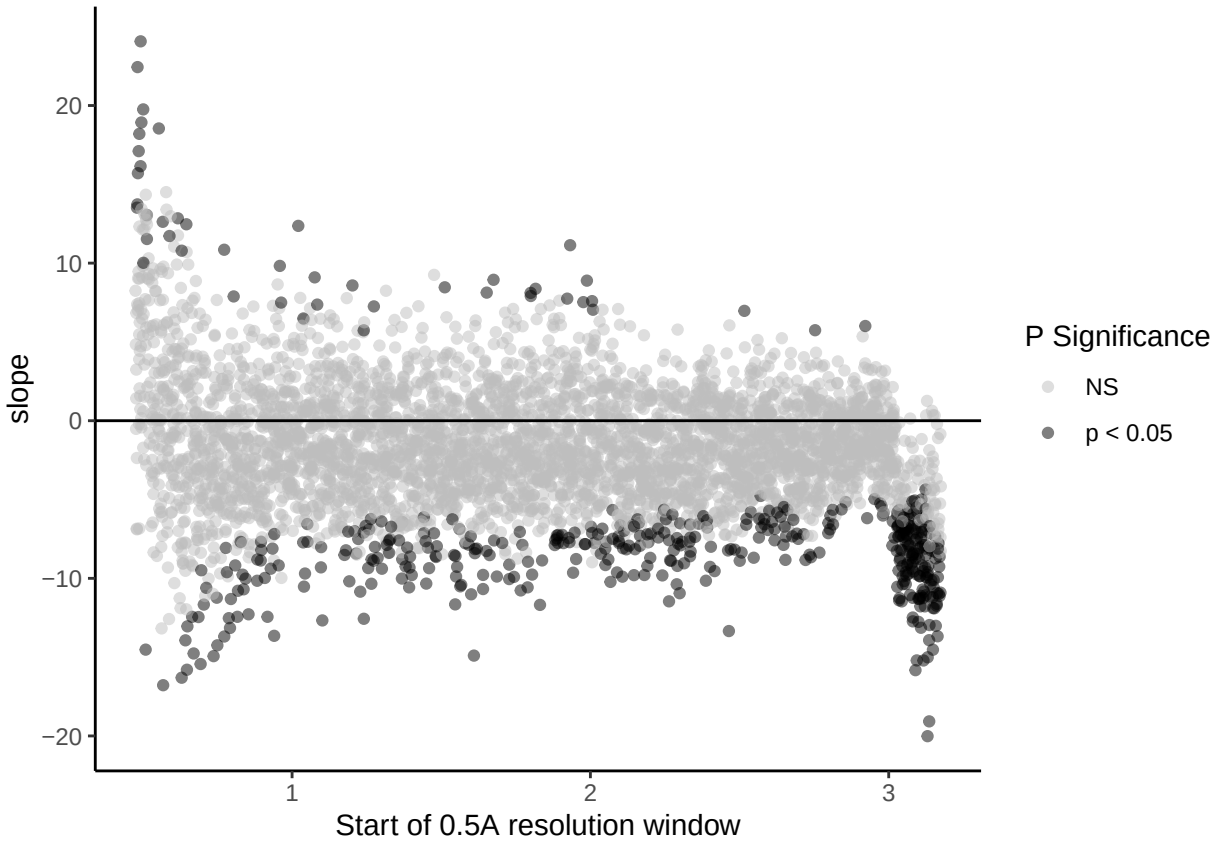
Regressions were calculated from a random sample (with replacement) of 50 observations drawn from a 0.5A window moving by 0.05A. 80 random samples were drawn for each window. Values collected:

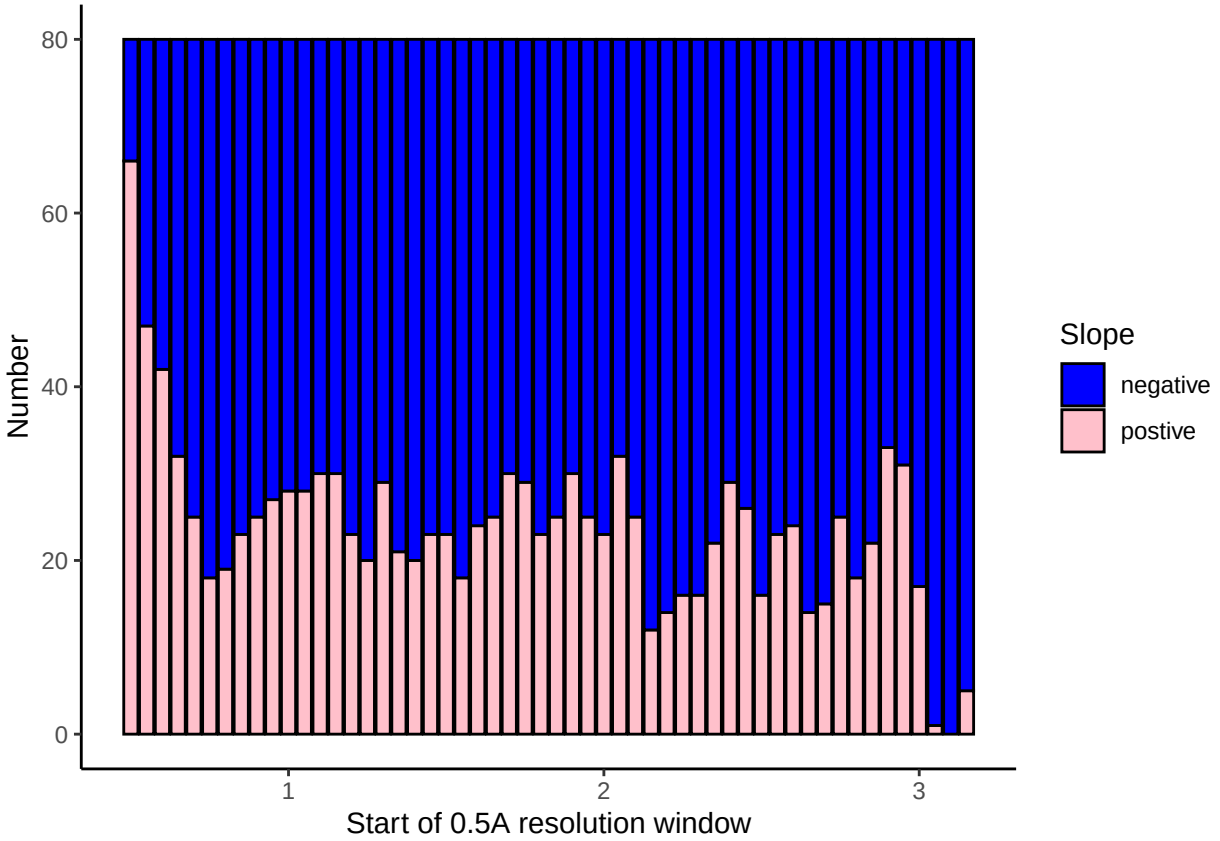
- slope of the linear regression
- standard error on the slope
- p value obtained from testing the slope against zero
- False discovery rate - the p value adjusted by the Benjamini & Hochberg method.

Size of the population available in each sample window. The red line indicates the size of the samples being drawn: 50

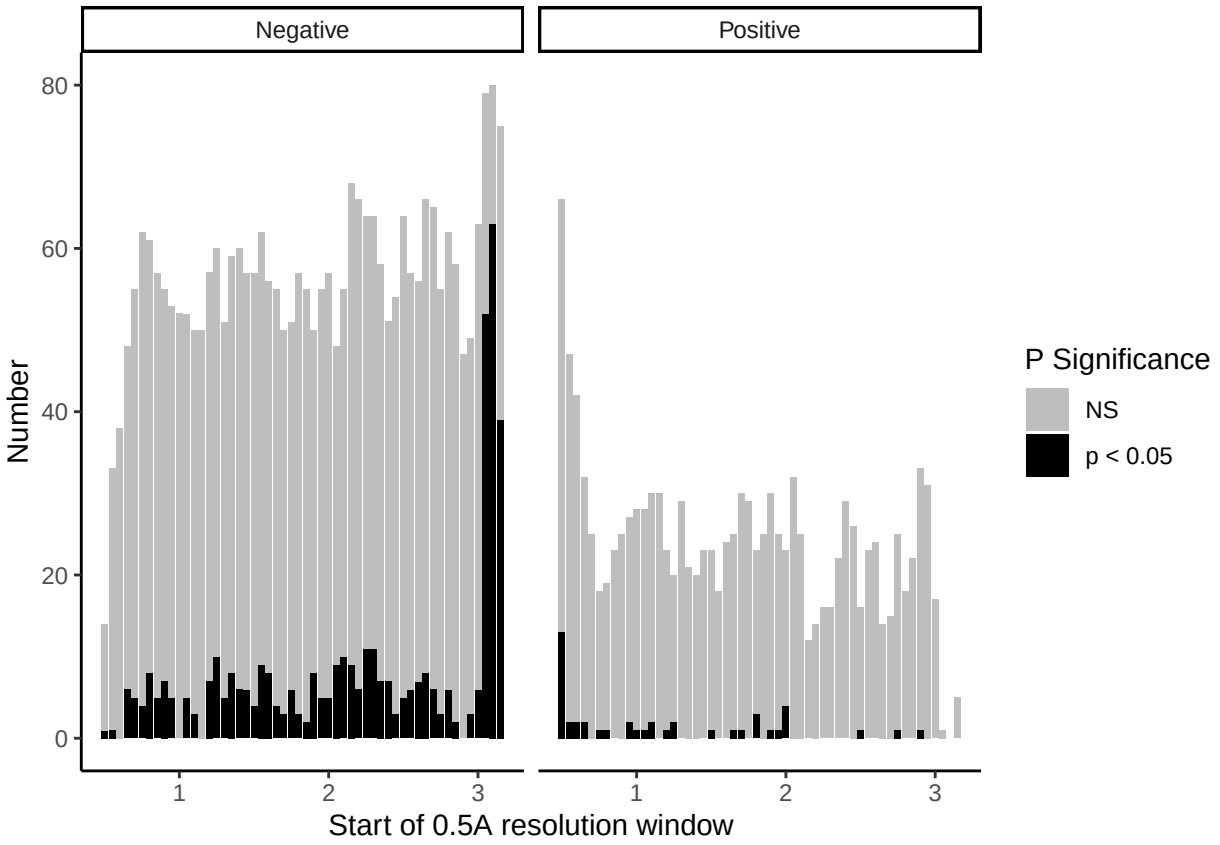


There are more negative than positive slopes and the large majority of the significant slopes are negative ($p < 0.05$).


20
10
0
−10
−20
slope
1
2
3
Start of 0.5A resolution window
P Significance
NS
p < 0.05

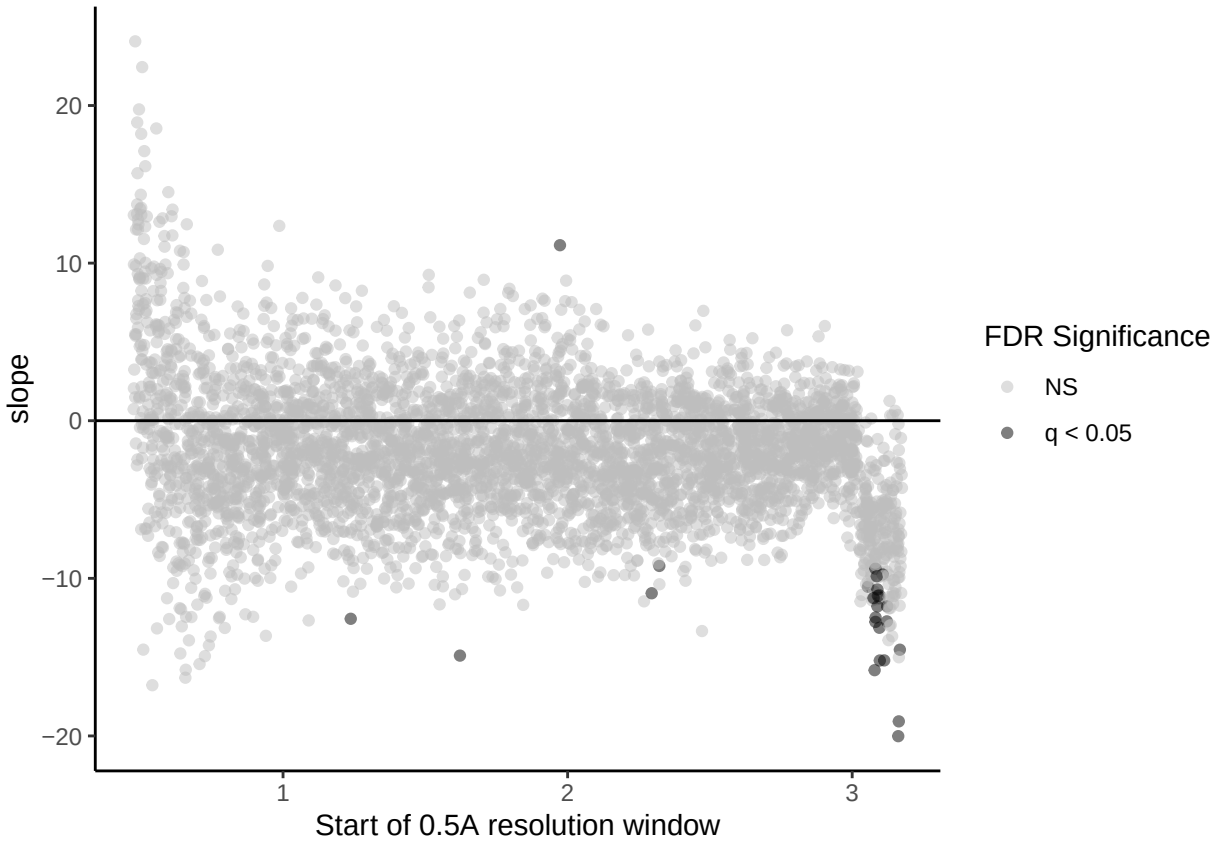


However, it is not the case that significant negative slopes start appearing at some resolution. A cut-off is not obvious.



If we considered the adjusted p-value which controls the false discovery rate (i.e., corrects for the number of tests done) the significant slopes are mainly negative but it is still the case that they appear at all resolutions.

(remember these are random samples and the outcomes with differ at little each time)


20
10
0
−10
−20
slope
1
2
3
Start of 0.5A resolution window
FDR Significance
NS
q < 0.05