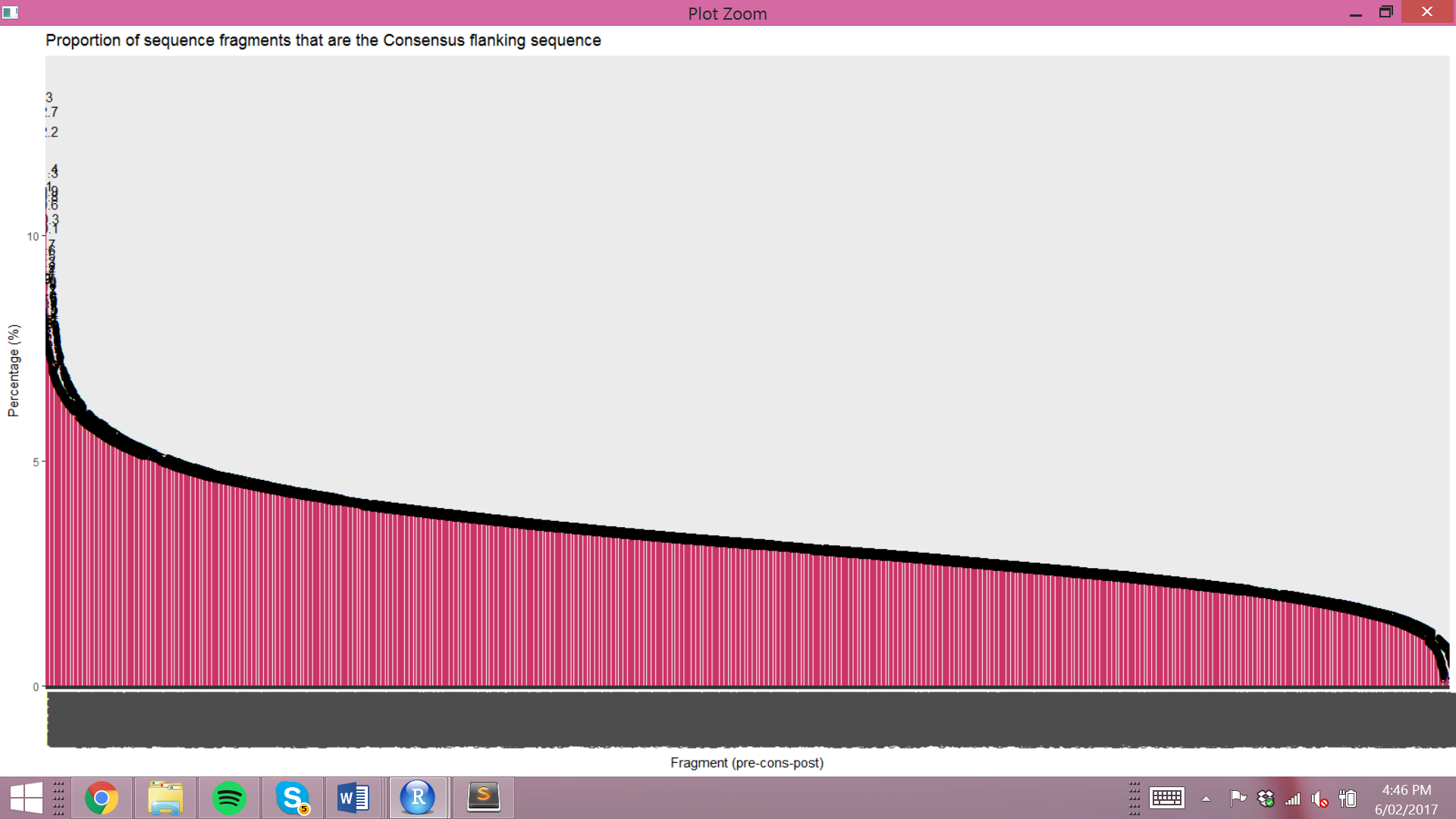
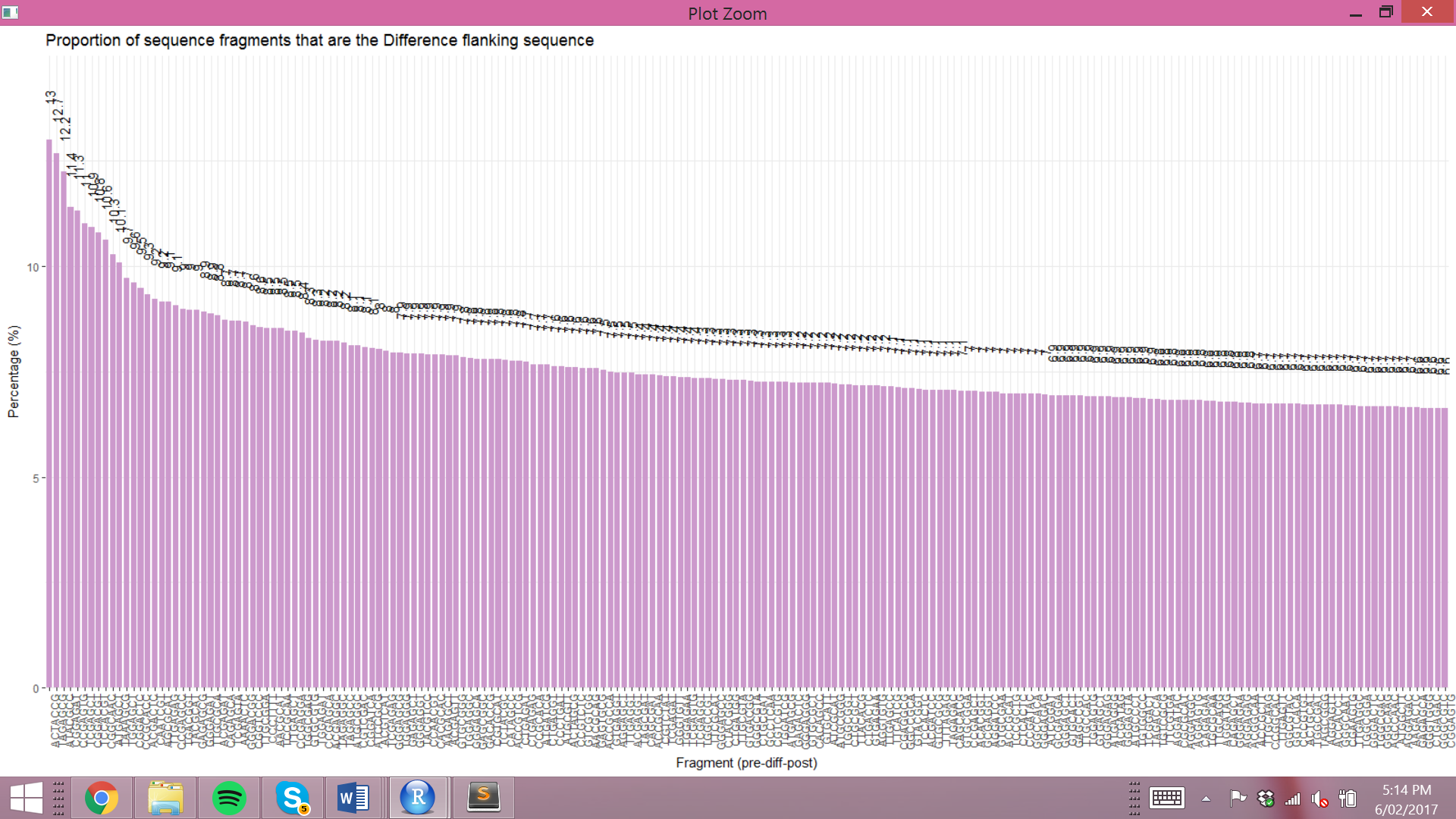
Normalised pcp – whole, 16370



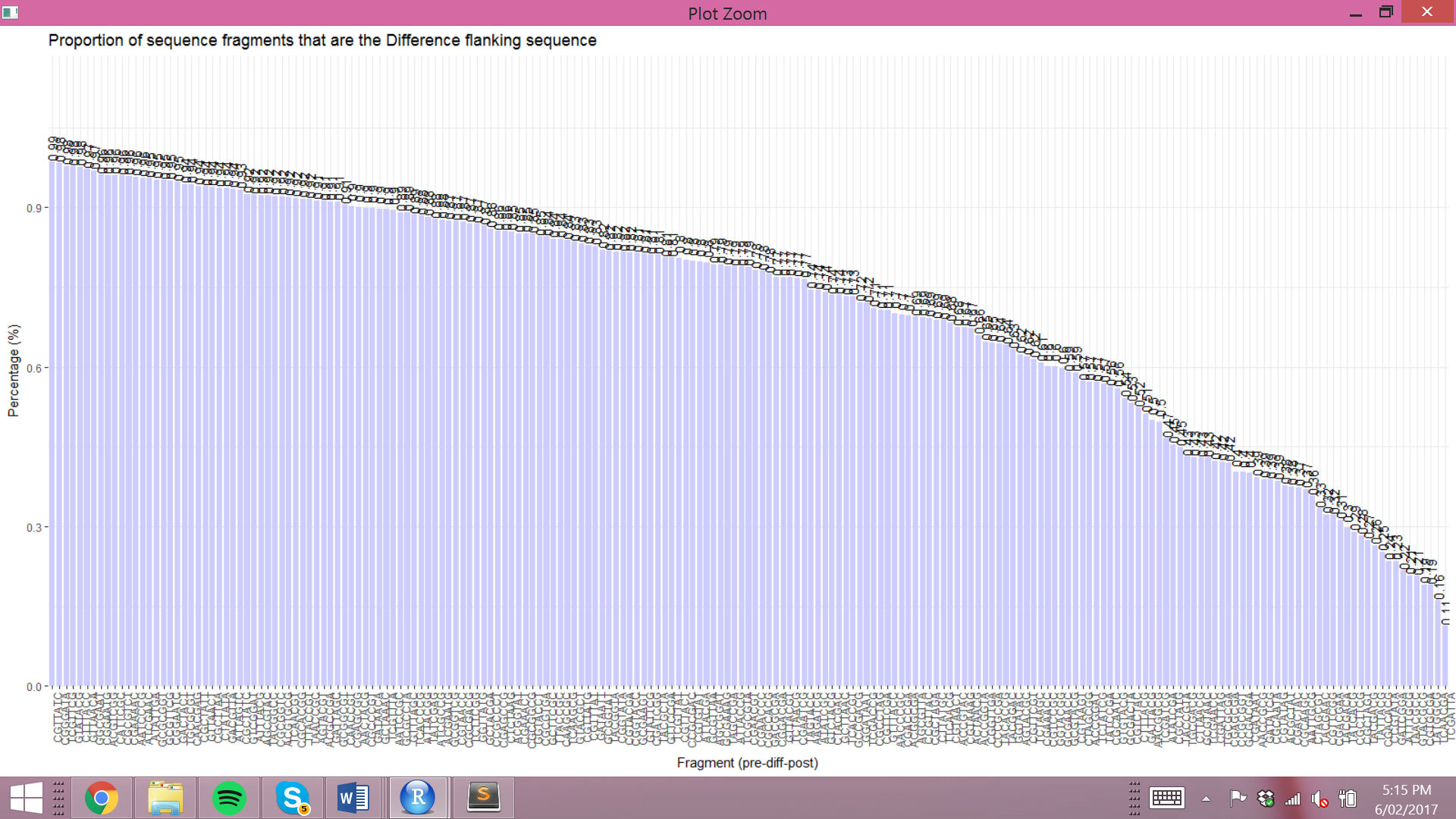
* Normalised by fragment occurrence across all sequences
* positively skewed distribution and steep negative gradients at the extremities, small fraction of k-mers that have a high and low error rate, vast majority around 2-5%, expected 3.2%
* how many below 3.2%?

Normalised – first 200



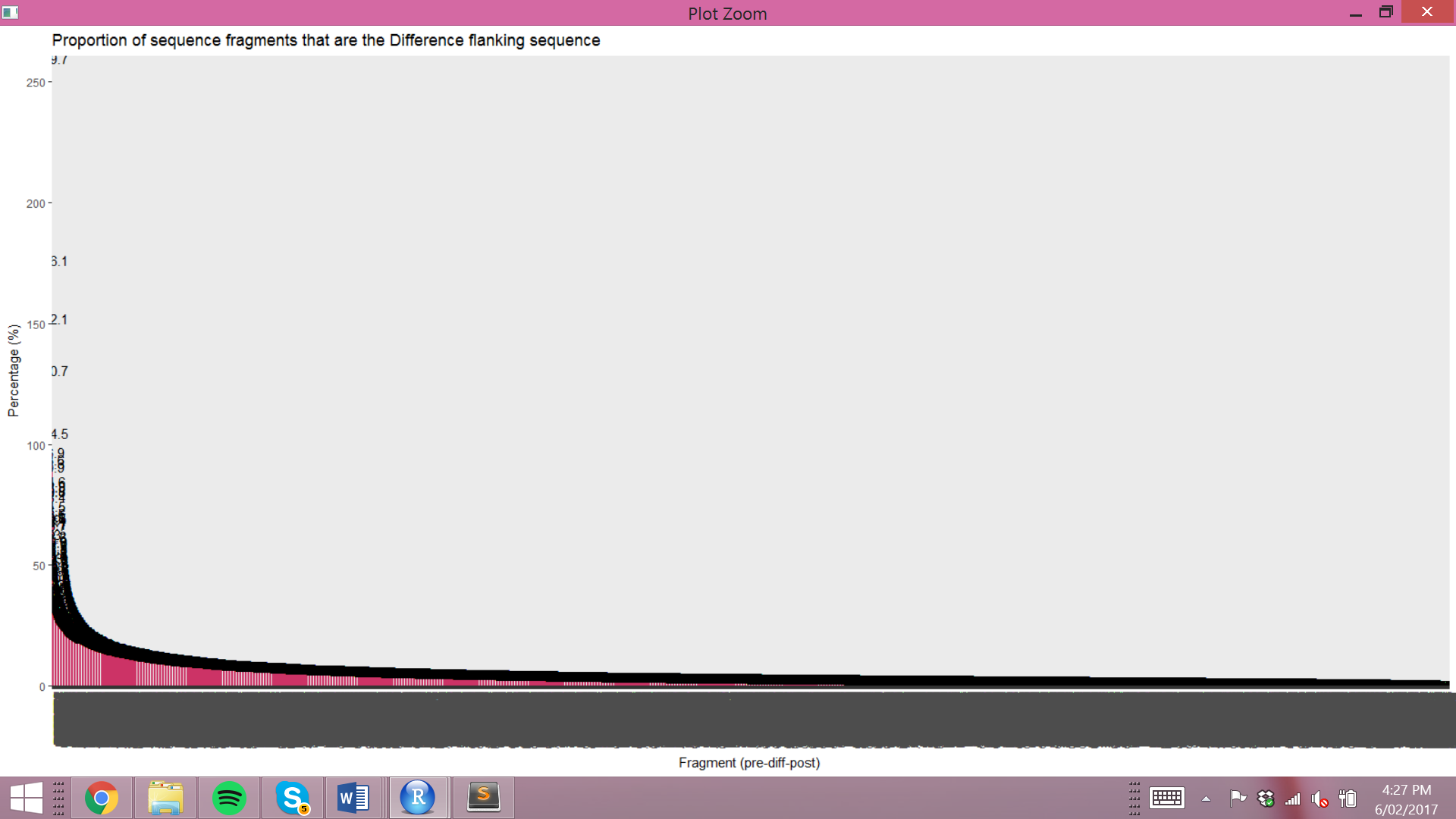
* ACTACCG is the highest error rate of 13%, 10% above the sample error rate
* mean 7.67%
* median 7.29%
* range 6.36

Normalise pcp – last 200



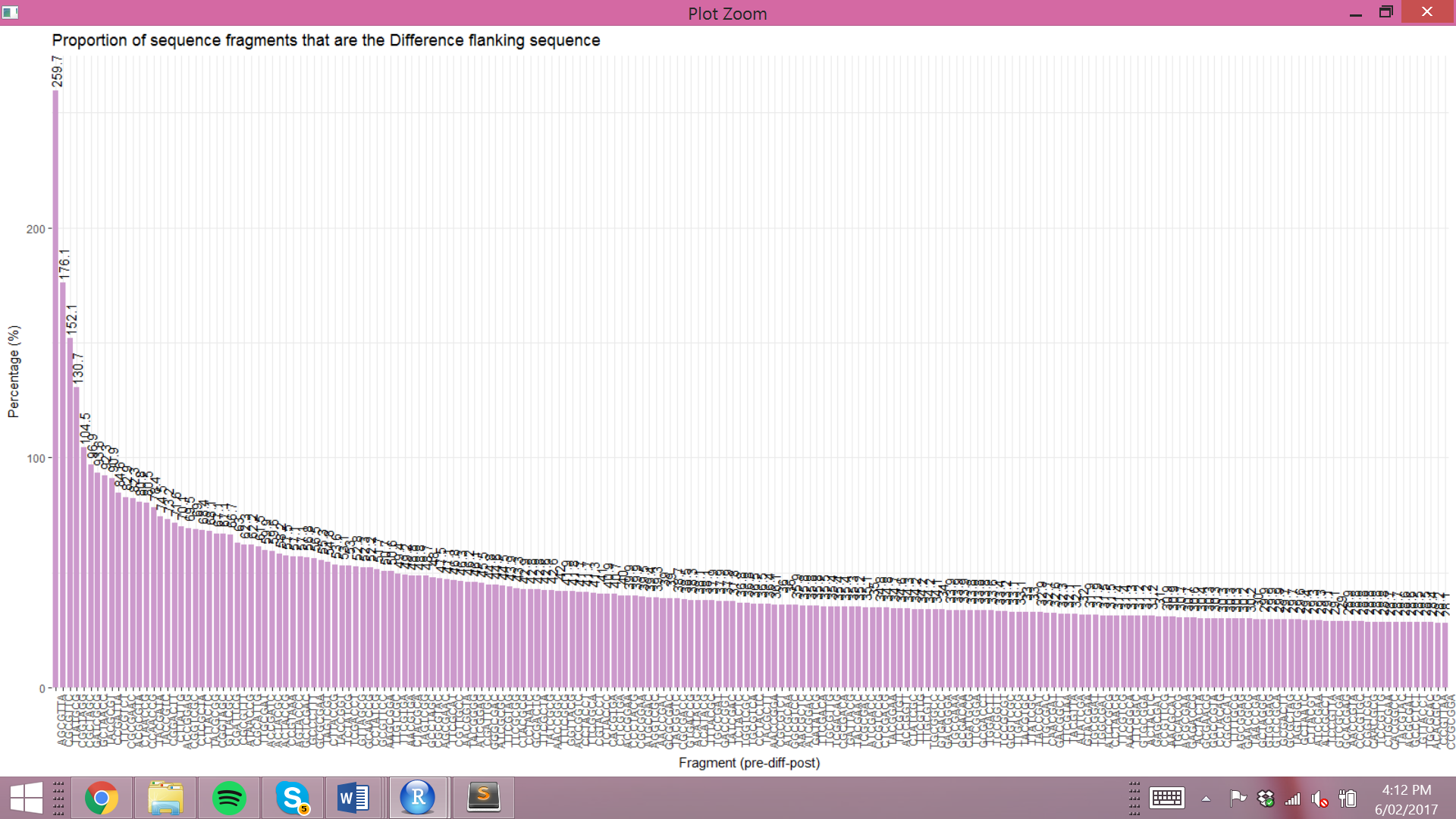
* TCGATTA lowest error rate of 0.11%, a 30th of the sample error rate
* Range 0.87, much smaller than above
* Mean 0.71%
* Median 0.79%.

Normalised pdp – whole, 16384



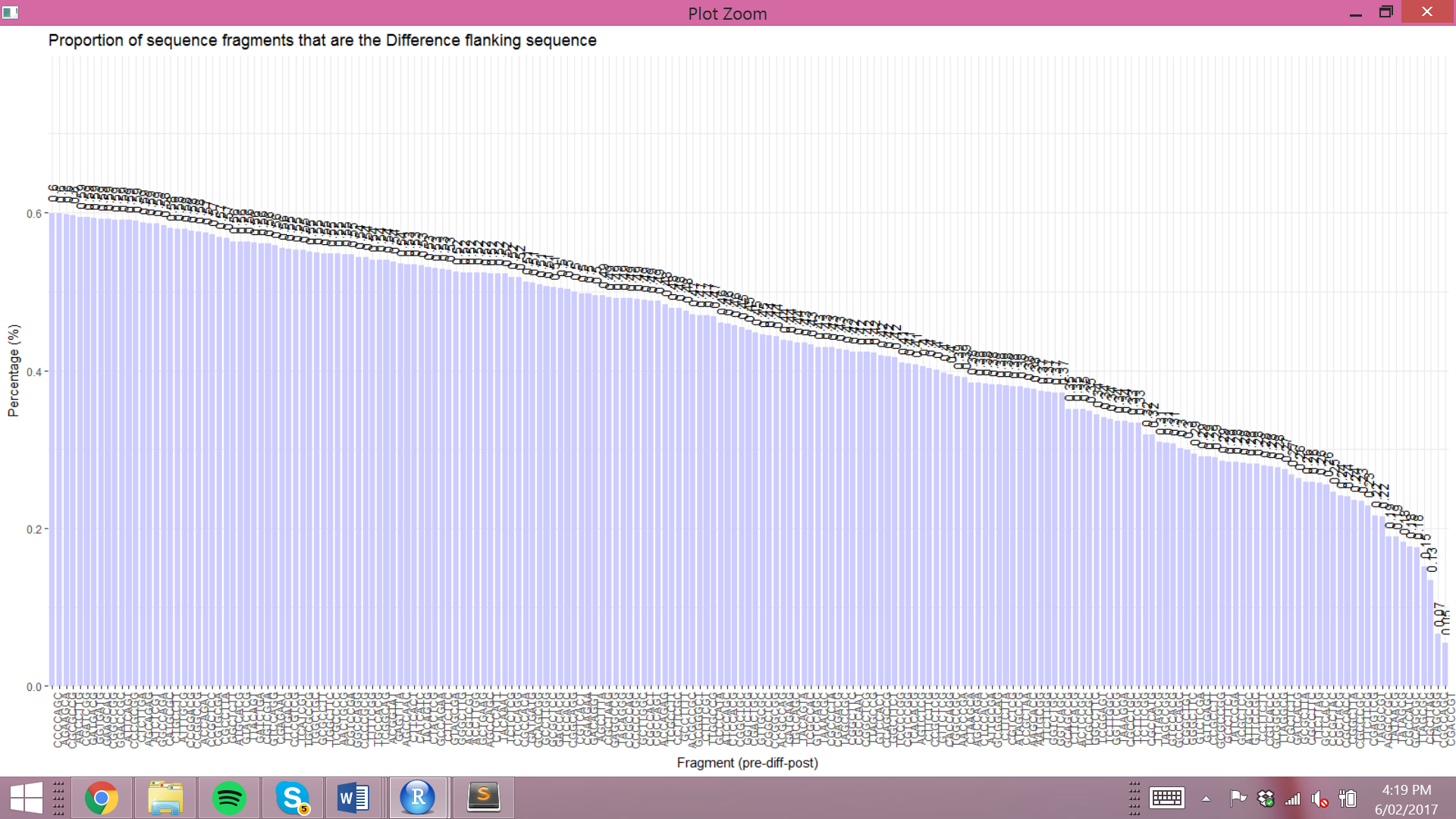
* Difference character of the error at the centre of the fragment
* Sample consists of the resultant consensus sequences, so these fragments can have an error rate above 100%, meaning the fragment occurs in clusters of reads more so than in the consensus sequences, and is often changed when developing the consensus.
* decreasing exponential

Normalised pdp – first 200



* first 80th, similar shape
* A few fragments have much higher error rates than the rest. Five fragments have an error rate above 100%.
* mean 45.53%
* median 36.65%
* range 231.65%
* AGCGTTA has the highest error rate of 259.7%, fragment is read and changed 2.6 times more than it occurs in the consensus sequences, is an outlier, the next fragment error rate is 176.1%, nearly 100 less.
* 0.7% error rate in consensus flanking sequence

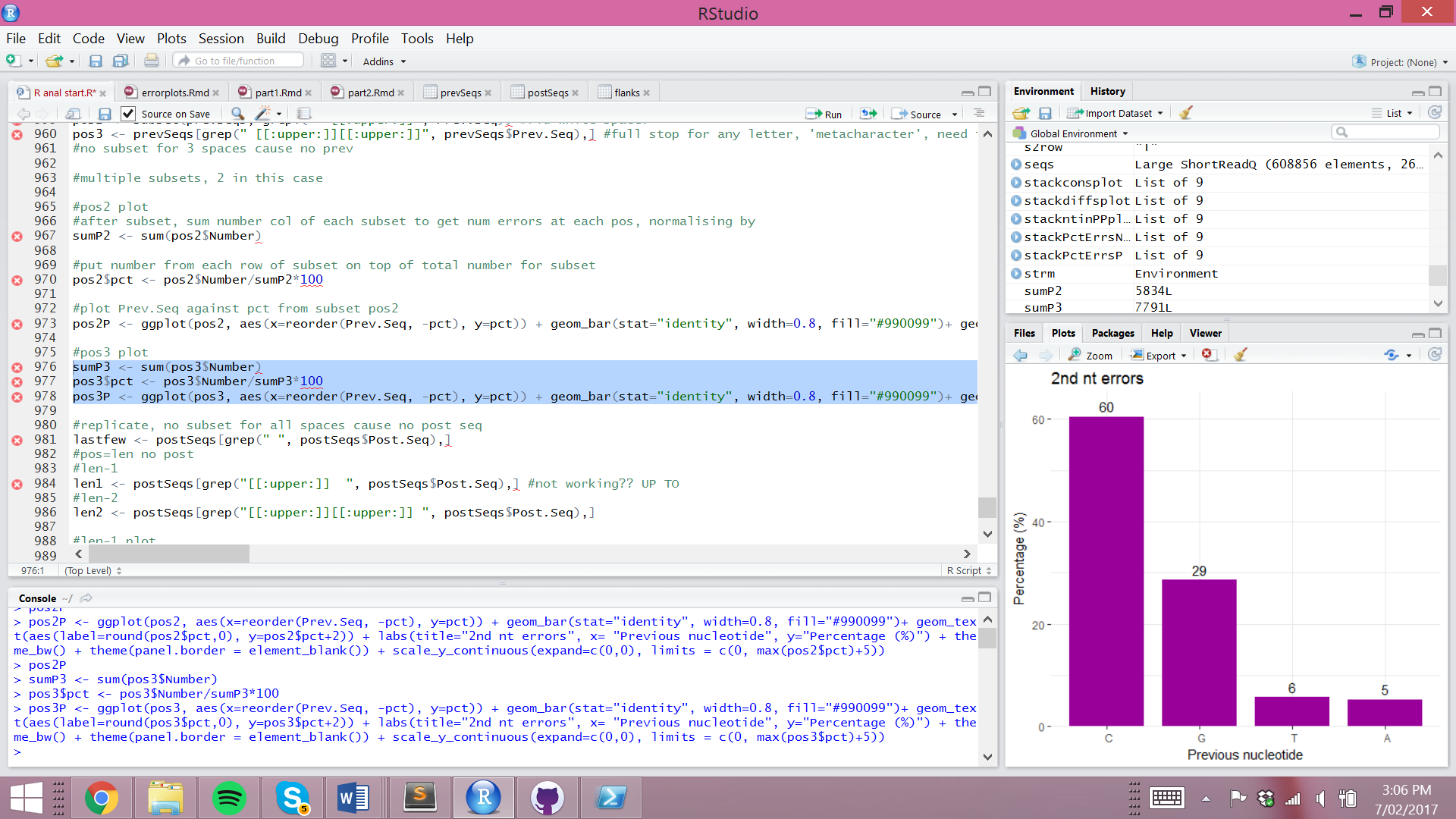
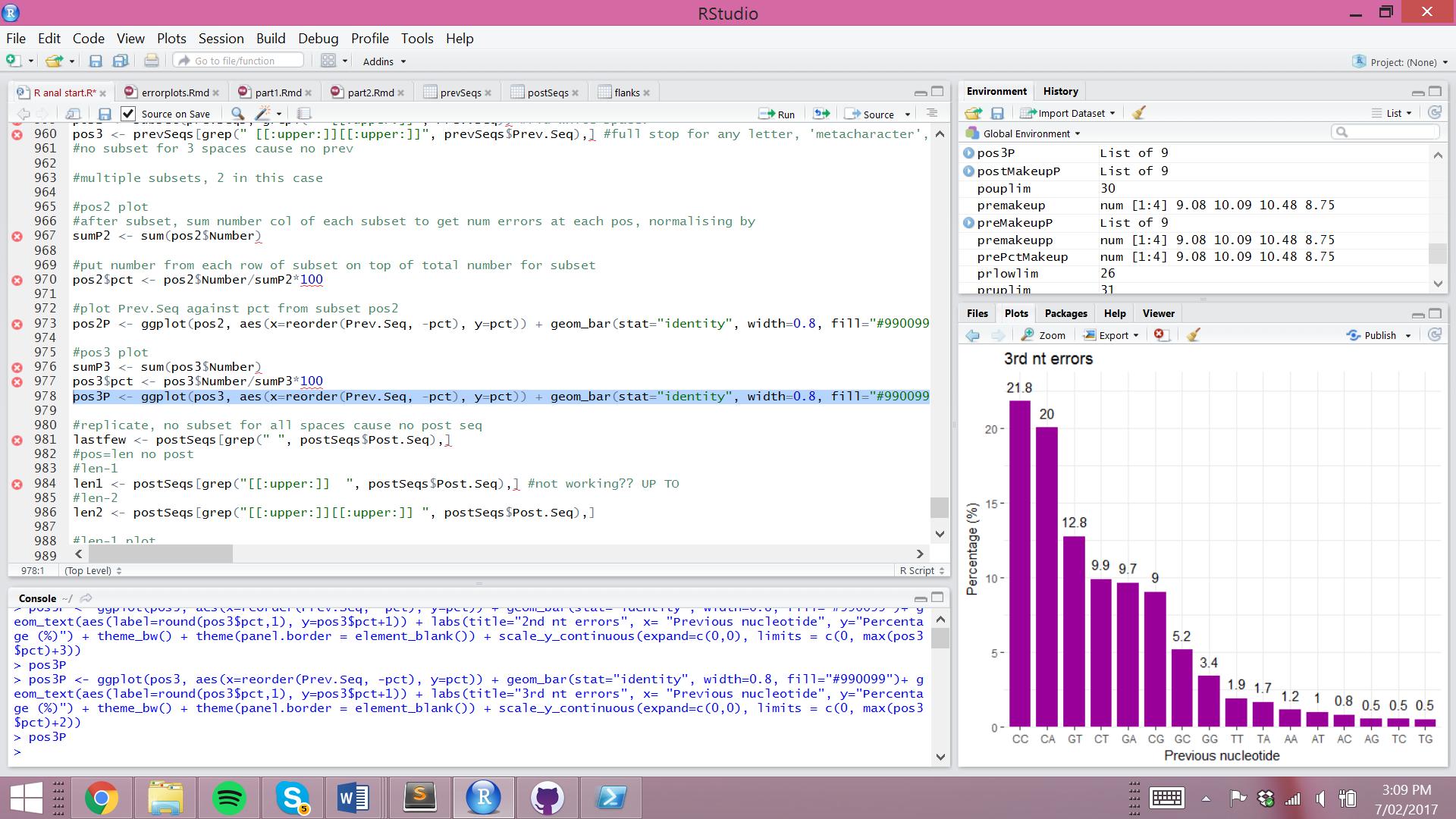
Normalised pdp – last 200



* range only 0.54
* mean 0.43%
* median 0.45%
* Lowest error rate is CGACCGT at 0.05%.

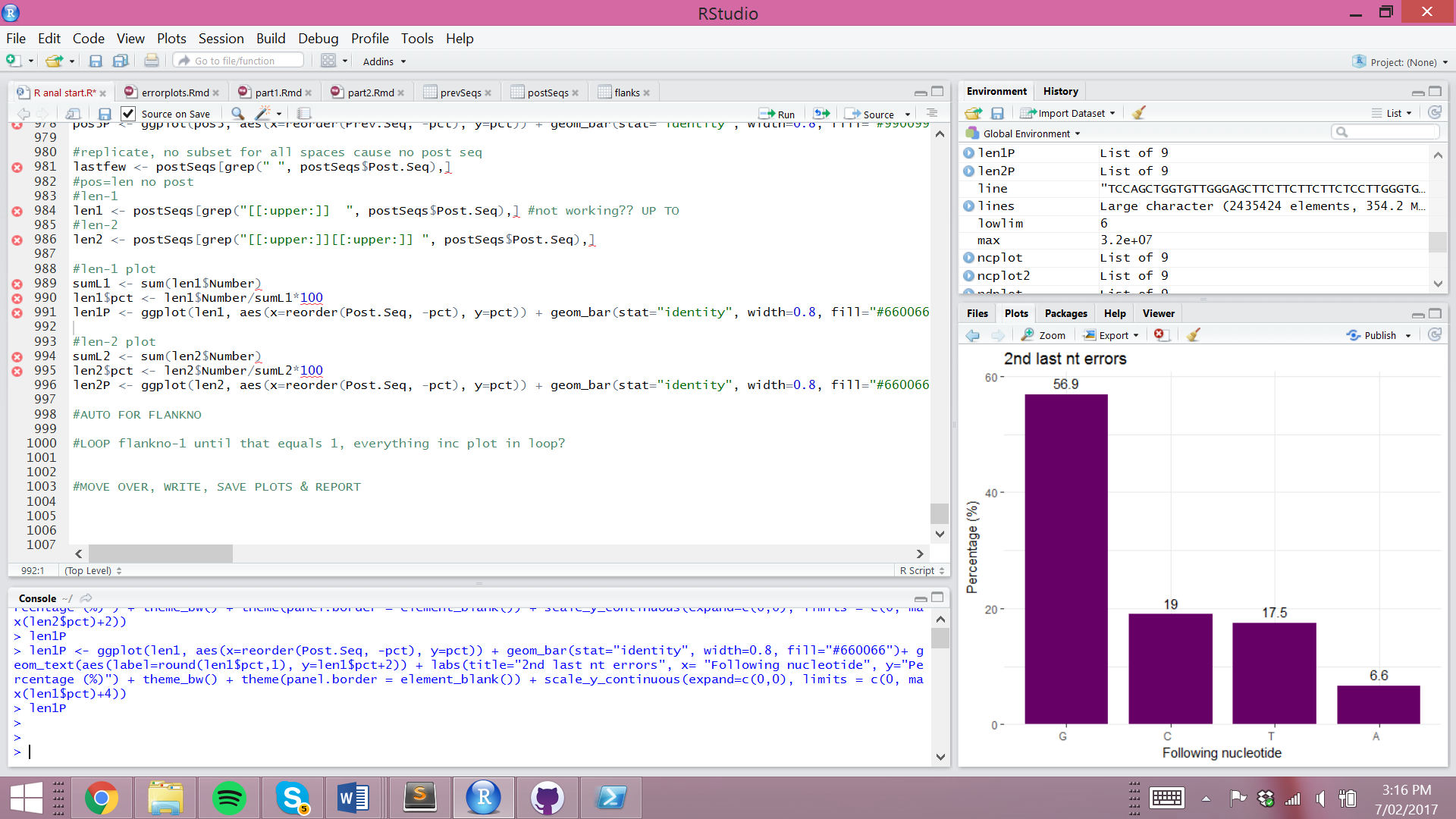
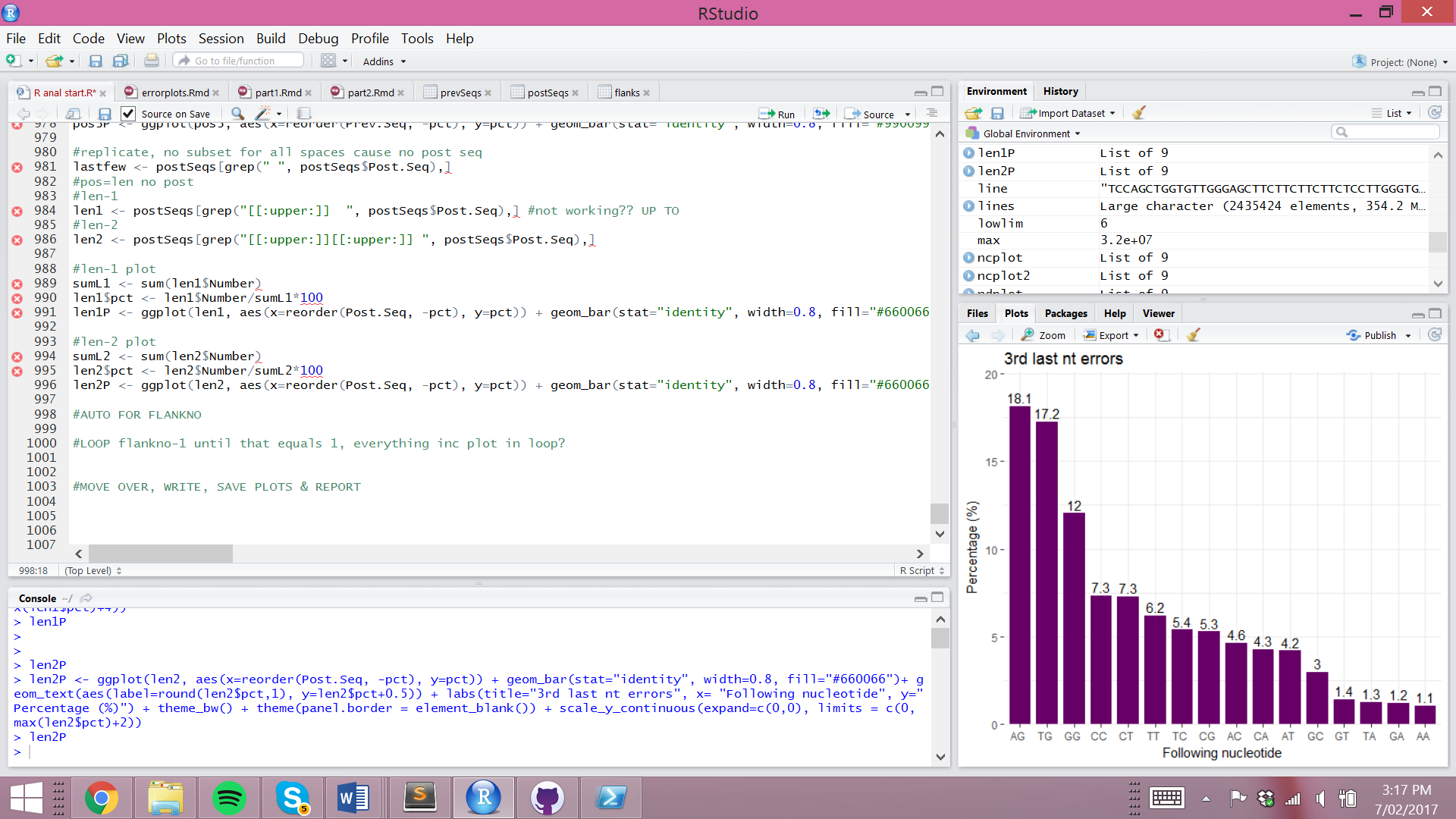
Errors at sequence extremities

* These errors have shorter pre or post-error sequences, shorter than the input flank number
* Normalised, percentage of all errors that occur at that position in the sequence

* pre-error sequences of errors at the start of the read sequence
* Position 2, one previous nucleotide
* 60% of these errors had C as the previous nucleotide
* Are sequences starting with C more error prone?
* T and A occurred only 6 and 5% of the time
* Position 3, has 2 previous nucleotides
* CC occurs most often with 21.8%
* AG, TC, TG are lowest at 0.5%.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* post-error sequence from errors at the end of sequences
* Position 2 have one nucleotide as their post-error sequence
* G was post-error nucleotide 60% of the time
* Are sequences ending with G more error-prone?
* Position 3 have 2 nucleotides as their post-error sequence
* AG occurs most often at 18.1% and AA least often at 1.1%.
* more errors at the end (position 2 - 5834, position 3 - 7791, 3rd in – 20857, 2nd in – 27411, 4x)