Appendix 7: Perl program to identify regions of co-ordinated replication

A purpose-written perl program was used to find the optimal segmentation of the replication timing (RT) data. Suppose a chromosome contains n RT signals arranged in genome order. Within each segment, starting at coordinate i and ending at coordinate j, we define the score S_{ij} equal to the sum of squared deviations of the RT values from the mean RT signal μ_{ij} for the segment. The optimal segmentation pattern (ie the number of segments and coordinates of segment boundaries) is chosen which minimises a function, W_n, based on the sum of segment scores plus a penalty score B for each segment transition. Let W_k be the score of the optimal segmentation for coordinates 1 through k. Then $W_0 = 0$ and $W_k = \min_{i<k} \{ W_{i-1}+B+S_{ik} \}$ for all $k>0$. The degree of segmentation is controlled by the value of B. The optimal segmentation is found by backtracking from the terminal value W_n. The statistical significance of W was determined by re-running the program on 1000 permuted data sets in which the order of observed RT signals was shuffled. The P-value for the test of the null hypothesis that the observed segmentation score could have arisen by chance is estimated as the proportion of times the permuted W score exceeded the observed score.