Web Note A

To test the statistical significance of the paralogons we identified in the human genome, comparisons were made to the results from 1000 simulations where the gene order was shuffled before the paralogon-finding algorithm was applied (Table 2). Shuffling retains characteristics of the real genome, such as gene family sizes and chromosome sizes, but any paralogons detected in shuffled data must be artifacts. The numbers of blocks containing three or more paralog pairs ($sm \geq 3$) are consistently higher in the real data than in the simulations, and are statistically significant by both parametric and non-parametric tests (Table 2). This deviation is more marked for the larger blocks. The number of paralogons containing at least 6 duplicated genes in the human genome (96) is over 50 standard deviations greater than the mean observed in the shuffled genomes (2.56).