S-plot2: Rapid visual and statistical analysis of genomic sequences

Kalesinskas *et al.*

Supplemental File 1

Calculating the divergence between species. When comparing a single chromosomal sequence of length M to itself, each window i of length w was compared to every other window in the same chromosomal sequence resulting in an $M/w \times M/w$ matrix. For each window i within the chromosome sequence, the average r^2 value, A_i, across all other windows within the chromosome was calculated. (We refer to the set of the averages A_i for all $i < M/w$ as "HA" for human chromosome comparisons, "CA" for chimpanzee chromosome comparisons, and "RA" for rhesus chromosome comparisons.)

Similarly, we can apply the same method to compare two homologous chromosomes belonging to different species. If we oversimplify the process of species divergence to a single point in time (thus ignoring subsequent gene flow), one could make the assumption that the chromosomal sequences are essentially identical such that w_i(HA)=w_i(CA). As such, the calculations described here of an individual chromosome compared to itself would be indiscernible from the comparison of the chromosome to its homolog. Post-speciation, the two individual genomes would begin to diverge. As such, the divergence between the human and chimpanzee homologous chromosome can be quantified by the cross-species comparison value and the intra-species comparison.

Comparison post-speciation of human and chimpanzee is calculated for each window i in human as w_i(HA)-w_i(HCA) or the average value for that window compared to all windows within the same chromosome sequence less the average value for that window compared to all windows within the homologous chromosome in chimpanzee. (Here HCA signifies the average r^2 value from the comparison of the homologous human and chimpanzee chromosome.) If w_i(HA)-w_i(HCA) > 0, one may conclude that the window i in the human genome is exhibiting a composition more similar to the remainder of the human chromosome in which it is located than it does to the homologous chimpanzee chromosome. If w_i(HA)-w_i(HCA) < 0, one can infer that the window i in the human genome is exhibiting a composition more similar to its homologous chimpanzee chromosome than the rest of the human chromosome in which it is located. Thus, a window exhibiting either a negative or positive value is indicative of divergence; discerning between positive and negative values provides better discrimination on the directionality of molecular evolution. The sign of the value w_i(HA)-w_i(HCA), be it positive or negative, presents a number of possible alternative hypotheses.

Given the situation in which w_i(HA)≠w_i(CA), one of three scenarios presents itself: 1) the variation in k-mer usage observed is due to nucleotide changes incorporated in the human genome, 2) the variation in k-mer usage observed is due to nucleotide changes incorporated in the chimpanzee genome, or 3) the variation in k-mer usage observed is due to nucleotide changes incorporated in both genomes. The comparison of each human and chimpanzee chromosome to its homologous rhesus macaque chromosome, thus calculating HRA (human and rhesus) and CRA (chimpanzee and rhesus) values, can shed some light. If the value of w_i(CA) is more similar to its rhesus homolog than it is to its human homolog, one can assume that window i in chimpanzee has undergone less change at the nucleotide level than has the human window. Correspondingly, if w_i(HA) is more similar to its rhesus homolog than chimpanzee homolog a similar assumption can be made of the window in human. If HA is more similar to its chimpanzee homolog (HCA) than either HA is to its rhesus homolog (HRA) or the chimpanzee homologous chromosome CA is to CRA, one can hypothesize that the divergence observed

between the human and chimpanzee homologous chromosomes is post-speciation of its last common ancestor and rhesus macaque. This divergence can be the result of compositional changes within: 1) the human-chimpanzee last common ancestor, 2) the human and chimpanzee genomes independently, or 3) the rhesus macaque genome. Analysis of the genomes of additional Old World monkeys and apes can aid in ascertaining the lineage(s) attributing to the observed divergence.

It is important to note that this alignment free-approach does not necessitate one to identify which window in a human chromosome corresponds to which window in a chimpanzee chromosome. Rather, it provides a means of comparing a window based solely on the knowledge of the chromosome in which it is located. A comparison of each human chromosome to all other human chromosomes (w=100,000 and k=6) was conducted revealing that each human chromosome was more similar to its homologous chromosome in chimpanzee than it was to all of the other human chromosomes (results not shown); this is as expected given the fact that each chromosome has its own history as well.