Chapter 18

Quantitative Aside 18.1--Data Overload in Comparative Genomics

As the number of sequences available for different species has increased, evolutionary biologists are revising phylogenies. Researchers in this field are encountering many of the same challenges with data overload as genomics researchers trying to assemble vast amounts of DNA sequence into whole genomes. A significant limiting facture is computational capacity.

For *n* species, one estimate of the number of trees that can be drawn is:

$$(2n-3)!! = \frac{(2n-3)!}{2^{n-2}(n-2)!}\text{, for } n \geq 2$$

Each '!' stands for a factorial. Multiply each positive integer less than or equal to one. Consider the difference in the number of possible trees if 10 versus 100 species were being used for tree construction. Now imagine what it would take to make a tree from 500,000 plants.

Visualizing phylogenetic trees for a large number of trees is also challenging. It would take a giant sheet of paper, more than five times the height of the Empire State Building, to print out a phylogeny of all 500,000 extant plant species if the species names were typed in 10 point font.