# Computational Gene Finding in the Human Genome: How many genes do we have?

**Steven Salzberg**
**Horvitz Professor of Computer Science**
**Director, Center for Bioinformatics and Computational Biology**
**University of Maryland**

After more than 40 years of research on the question, we still do not know precisely how many genes are contained in the human genome. Over time, the estimate has steadily shrunk, from more than 1 million in the 1960s to somewhere around 20,00-25,000 today. Many scientists believed that the sequence of the human genome would finally answer this question definitively, but the publication of the genome in 2001(1,2) merely served to highlight our uncertainty. Further complications have arisen as we discovered many elements in the human genome that appear to have a function, but whose function is as yet unknown (3).

Bioinformaticists and geneticists have developed a wide range of methods for identifying genes and applied them to the human genome and many other species. Although the accuracy of these methods continues to improve, we have not yet converged on an estimate of the number of human genes. Indeed, the best methods do not even agree on the precise exon-intron structure of most genes, although groups including ours continue to refine their algorithms (4) and much progress has been reported. Even as progress continues on the algorithmic side, however, improvements in sequencing technology are contributing to an acceleration in the amount of DNA sequence data available to the public. The uncertainty over the gene count for humans is far greater for most other genomes, which now include hundreds of animals, plants, fungi, and single-celled eukaryotes, including many pathogenic organisms. For many of these species we are only just beginning to understand their biology, despite the fact that we already have their genomes. Much more research – both biological and computational – will be needed to keep pace with the ever-growing list of genomes and the rapidly expanding DNA sequence databases.

1. The Sequence of the Human Genome. J.C. Venter et al., *Science* 291 (2001), 1304–1351.
2. Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium, *Nature* 409 (2001), 860-921.
3. The ENCODE (ENCyclopedia Of DNA Elements) Project. The ENCODE Project Consortium, *Science* 306 (2004), 636-640.
4. JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. J.E. Allen, W.M. Majoros, M. Pertea, and S.L. Salzberg. *Genome Biology* 7 (2006). Suppl 1:S9.