The human genome and our genes: how much are we missing?

Steven L. Salzberg, Ph.D.
Bloomberg Distinguished Professor of Biomedical Engineering, Computer Science, and Biostatistics
Director, Center for Computational Biology
Johns Hopkins University
http://salzberg-lab.org

How many genes do we have? The Human Genome Project was launched with the promise of revealing all of our genes, the "code" that would help explain our biology. The publication of the genome in 2001 provided only a very rough answer to the question of how many genes we have, and a highly-fragmented draft genome sequence. For more than a decade following, the number of protein-coding genes steadily shrank, but the invention of RNA sequencing revealed a vast new world of splice variants and RNA genes.

In this talk, I will review where we've been and where we are today, and I will describe our use of an unprecedentedly large RNA sequencing resource to create a comprehensive new human gene database, containing thousands of novel genes and gene variants. I will describe how we created new, more efficient algorithms to assemble 10,000 human RNA-seq experiments containing nearly 900 billion reads, and then to create a comprehensive new human gene catalog, called CHESS, that contains thousands of novel genes and gene variants. I will also discuss recent breakthroughs that have finally allowed the human genome itself to be completed, and how that effort has revealed hundreds of new genes that were previously hidden in the gaps.