Assembling the transcriptome and re-assessing the human gene catalog: How many genes do we have?

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 human genome in 2001 provided only a very rough answer to this question. 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 first describe the computational methods that made this analysis possible: the HISAT system for spliced aligment of NGS reads, a successor to the TopHat spliced aligner; and the StringTie program for assembly and quantitation of RNA-seq data, a successor to the Cufflinks transcript assembler. I will then describe how we have used these systems to assemble ~10,000 human RNA-seq experiments containing nearly 900 billion reads, and then used the results to create a comprehensive new human gene catalog, called CHESS, that contains thousands of novel genes and gene variants.

This talk describes joint work with Mihaela Pertea Ph.D., Daehwan Kim, Ph.D., Ales Varabyou, Geo Pertea, and others.