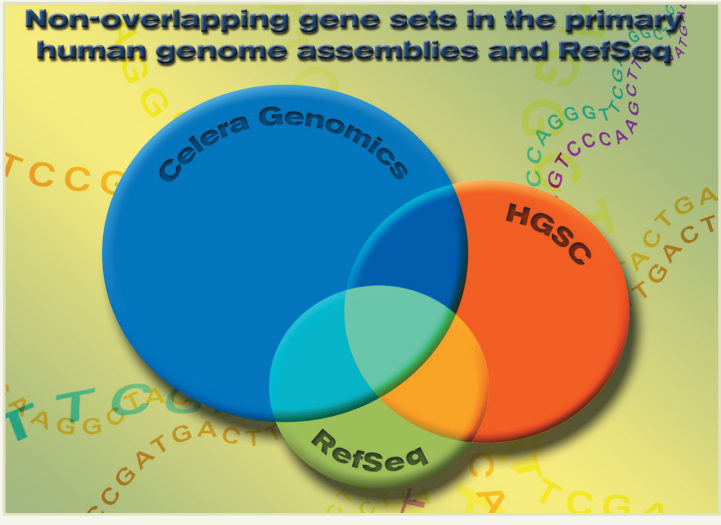
Tularik Unlikely to Subscribe to Celera's Human Genome Sequence

Companies looking to mine the human genome for unique, druggable targets will have greater success using the sequence provided by the International Human Genome Sequencing Consortium (HGSC) than the proprietary, and more expensive, version offered by Celera Genomics Group, Rockville, Md., according to research published by Tularik Inc., South San Francisco, Calif., and the Genomic Institute of Novartis Research Foundation (GNF), San Diego.

With much fanfare, Celera and HGSC released their respective drafts of the human genome sequence in 2001. The drafts showed that the human genome contained roughly 30,000 genes, leading many observers to conclude that the sequences were largely in agreement with each other. Not true, says Bruce Ling, PhD, director of research informatics at Tularik. There were quite a few nonoverlapping novel genes in the 2001 versions, and the number of such genes fluctuated over the succeeding years. According to Ling, the phenomenon of non-overlapping gene sequences in the two genome assemblies persists even to today.



Tularik and GNF are interested in discovering whether the unique genes predicted by the Celera genome sequence are necessary to develop therapies. "We wanted to know [if there] are novel genes in the genome between these two data sets," says Ling. "[It is important to discover] if the public domain data set has all the novel genes, or if we can mine out all the novel genes. If not, then we may need to subscribe to [Celera's genome], because we need to be on the cutting edge. From this conclusion, we don't really need [Celera's sequence]."

Celera derived its version of the human genome sequence using its proprietary whole genome sequencing approach, which reassembles sequence fragments from across the entire genome using highly sophisticated computing power. The HGSC sequence was derived with a hierarchical shotgun approach that breaks the genome into manageable parts that are cloned into bacterial artificial chromosomes (BACs). The HGSC sequence, which has been updated in several releases, is open to the public. In contrast, Celera , which has also released several genome assemblies, provides access only to companies that purchase a subscription to its dataset.

The Tularik and GNF groups collected almost all of the assembly releases and gene sets and, in a high-throughput computing environment that Ling says was invaluable to the success of the project, performed a parallel analysis of both genome assemblies using the BLAT, GENSCAN, and BLAST algorithms. The Tularik group, led by Ling, was responsible for gathering the transcriptome data from the HGSC databases; the GNF group subscribed to and analyzed the Celera data. Both sets of data were then compared to the highly validated RefSeq database of individually sequenced genes. An extended analysis of the multiple datasets is scheduled to be published in the September 2003 issue of *Bioinformatics*.

"If the prediction methods are different, of course you could get different things," says Ling. "That's why we used only one prediction algorithm, [with] no human intervention. If the two genomes are similar, you should get similar results. But they are not."

In a recent paper published in *Genomics*, the Tularik and GNF groups reported a 20 to 30% difference between the transcriptomes using the GENSCAN filter. The authors claim that the results suggest a fundamental difference between the public and the Celera genome assemblies is likely responsible for much of the discrepancy.

Celera disagrees. "The differences between the Celera and public genome sequences are less than 5%," says Tony Kerlavage, PhD, senior director of bioinformatics applications at Celera. "We feel highly confident that our order is correct." He says, "We're not touting to our customers that they are getting a proprietary genome. Our product is . . . the aggregation, integration, quality, and classification of data that make it a valuable research tool."

Over the past year and a half, the controversy spilled into the *Proceedings of the National Academy of Sciences U.S.A.* with charges by the leaders of the HGSC that the Celera genome sequence was derived largely because of Celera's access to the public data, and that the effectiveness of whole-genome sequencing is still unclear. Celera responded that the charge is invalid because it is based on flawed assumptions. According to the Celera group, its genome assembly is more accurate than the public data because it contains better long-range contiguity and less redundant data.

"It's not just an intellectual game," says Ling of the debate over whether to develop drugs based on the public data only. "It's also very practical for a pharmaceutical company to gain a competitive edge."

— Chris Dickey, DrPH
Editor in chief