doi:10.1006/geno.2002.6824, available online at http://www.idealibrary.com on IDEAL

Comparative Analysis of Human Genome Assemblies Reveals Genome-Level Differences

Shuyu Li,[1] Jiayu Liao,[2,*] Gene Cutler,[1] Timothy Hoey,[1] John B. Hogenesch,[2] Michael P. Cooke,[2] Peter G. Schultz,[2] and Xuefeng Bruce Ling[1,*]

[1] *Tularik, Inc., Two Corporate Drive, South San Francisco, California 94080, USA*
[2] *The Genomic Institute of Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121, USA*

**To whom correspondence and reprint requests should be addressed. Fax: (858) 812-1502. E-mail: liao@gnf.org. Fax: (650) 825-7400. E-mail: xling@tularik.com.*

Previous comparative analysis has revealed a significant disparity between the predicted gene sets produced by the International Human Genome Sequencing Consortium (HGSC) and Celera Genomics. To determine whether the source of this discrepancy was due to underlying differences in the genomic sequences or different gene prediction methodologies, we analyzed both genome assemblies in parallel. Using the GENSCAN gene prediction algorithm, we generated predicted transcriptomes that could be directly compared. BLAST-based comparisons revealed a 20–30% difference between the transcriptomes. Further differences between the two genomes were revealed with protein domain PFAM analyses. These results suggest that fundamental differences between the two genome assemblies are likely responsible for a significant portion of the discrepancy between the transcript sets predicted by the two groups.

Celera Genomics and the International Human Genome Sequencing Consortium (HGSC) simultaneously published the description of the human genome sequencing, analysis, and gene annotation [1,2]. Although both teams identified approximately 30,000 human genes [1,2], a direct comparison of the Celera and HGSC (Ensembl) data sets revealed little overlap between their novel predicted genes [3]. Questions arose as to whether this observed difference is due to discrepancies in the underlying raw sequence data, the resultant genome assemblies, or the independent gene prediction methodologies used by both groups.

To distinguish between these possibilities, we have carried out a comparative analysis of the HGSC genomes (Ensembl 1.0.0, Ensembl 1.1.0, and Ensembl 1.2.0; performed at Tularik, Inc.) and the Celera genome (CHGD_assembly_R25h; performed at the Genomics Institute of the Novartis Research Foundation) using the GENSCAN [4] gene prediction program to generate corresponding predicted transcriptomes. GENSCAN, which was a key component of both the Celera and HGSC gene prediction pipelines, predicts both partial and full-length transcripts. GENSCAN full-length transcripts are defined as those for which GENSCAN predicts a promoter region, one or more exons, and a polyadenylation signal. This analysis revealed that the Celera transcriptome (150,571) has more predicted transcripts than that of HGSC (Ensembl 1.0.0; 109,083). The results for the more recent HGSC genome releases (Ensembl 1.1.0, Ensembl 1.2.0) gave very similar results and are therefore not shown here. A detailed analysis of these GENSCAN-predicted transcripts found that Celera (71,721) has fewer full-length gene predictions than does HGSC (87,295). A BLAST [5]-based comparison of all GENSCAN transcripts (threshold of ≥ 98% identity over at least 100 nucleotides) showed that 80% of predicted HGSC genes have at least one matching sequence in the Celera GENSCAN predictions, whereas 70% of Celera predictions have at least one overlapping sequence in the HGSC set. These results demonstrate that significant discrepancies exist even between Celera and HGSC assembly-derived gene sets predicted with the exact same methodology.

To understand the impact of these transcriptome differences on the derived proteomes, we have analyzed the predicted translations of these sequences for the presence of known protein domains using the PFAM [6] 7.0 set of Hidden Markov Models (HMMs) (3360 models, hit threshold E value 1×10^{-10}). The differences between the number of hits for each protein domain model in the HGSC and Celera predicted gene sets were plotted in Fig. 1 for the 1495 models that had hits (data for searches with E values of 1×10^{-5} or 1×10^{-2} gave similar results and are not shown). Of all the matching PFAM models, a large percentage have more matches (47%) in the HGSC-derived gene set than in the Celera-derived genes. This is more than the number of models that matched both data sets equally (30%), and more than twice the number that had excess matches in the Celera data (22%). This analysis further supports the conclusion that the genome assemblies had a significant impact on the predicted transcript sets.

This parallel analysis of the genome assemblies released by the HGSC and Celera teams provides strong evidence that there are major fundamental differences between these two

0888-7543/02 $35.00

data sets in the numbers, identities, and properties of predicted genes derived from these sequences. Based on this, we conclude that these sequence-level differences must be at least partly responsible for the discrepancies in the previous findings [3]. Along with the recent re-analysis [7,8] of Celera's genome assembly [1], this report provides further evidence that the whole genome approach and the hierarchical shotgun sequencing approach yielded different genomes.

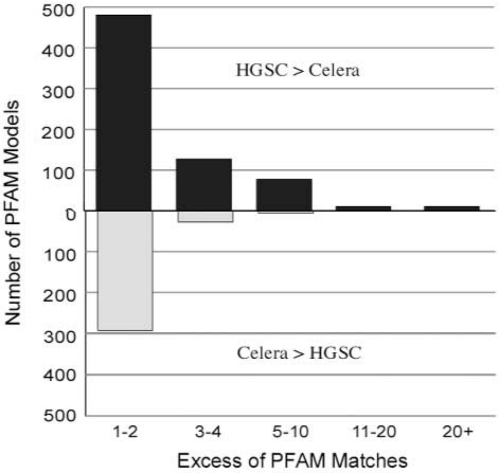

FIG. 1. PFAM domain profiling of Celera and HGSC derived transcriptomes. The x-axis represents the excess of matches per PFAM model in the HGSC versus Celera data sets. The y-axis represents the number of models that fall into each category. Upward bars represent PFAM models, which have more hits in the HGSC data set. Downward bars represent PFAM models, which have more hits in the Celera data set.

RECEIVED FOR PUBLICATION JUNE 6; ACCEPTED JUNE 19, 2002.

REFERENCES

1. Venter, J. C., *et al.* (2001). The sequence of the human genome. *Science* **291:** 1304–1351.
2. Lander, E. S., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
3. Hogenesch, J. B., *et al.* (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106:** 413–415.
4. Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.
5. Altschul, S. F., *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.
6. Bateman, A., *et al.* (2002). The Pfam protein families database. *Nucleic Acids Res.* **30:** 276–280.
7. Waterston, R. H., Lander, E. S., and Sulston, J. E. (2002). On the sequencing of the human genome. *Proc. Natl. Acad. Sci. USA* **99:** 3712–3716.
8. Myers, E. W., Sutton, G. G., Smith, H. O., Adams, M. D., and Venter, J. C. (2002). On the sequencing and assembly of the human genome. *Proc. Natl. Acad. Sci. USA* **99:** 4145–4146.