BIOINFORMATICS

Vol. 19 no. 13 2003, pages 1597–1605
DOI: 10.1093/bioinformatics/btg219

A comparative analysis of HGSC and Celera human genome assemblies and gene sets

Shuyu Li[1,†], Gene Cutler[1,†], Jane Jijun Liu[1,†], Timothy Hoey[1], Liangbiao Chen[2], Peter G. Schultz[3], Jiayu Liao[3,*] and Xuefeng Bruce Ling[1,*]

[1]Tularik, Inc. 112 Veterans Blvd, South San Francisco, CA 94080, USA, [2]Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, People's Republic of China and [3]The Genomic Institute of Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, CA 92121, USA

Received on December 21, 2002; revised on March 11, 2003; accepted on March 26, 2003

ABSTRACT

Motivation: Since the simultaneous publication of the human genome assembly by the International Human Genome Sequencing Consortium (HGSC) and Celera Genomics, several comparisons have been made of various aspects of these two assemblies. In this work, we set out to provide a more comprehensive comparative analysis of the two assemblies and their associated gene sets.

Results: The local sequence content for both draft genome assemblies has been similar since the early releases, however it took a year for the quality of the Celera assembly to approach that of HGSC, suggesting an advantage of HGSC's hierarchical shotgun (HS) sequencing strategy over Celera's whole genome shotgun (WGS) approach. While similar numbers of *ab initio* predicted genes can be derived from both assemblies, Celera's Otto approach consistently generated larger, more varied gene sets than the Ensembl gene build system. The presence of a non-overlapping gene set has persisted with successive data releases from both groups. Since most of the unique genes from either genome assembly could be mapped back to the other assembly, we conclude that the gene set discrepancies do not reflect differences in local sequence content but rather in the assemblies and especially the different gene-prediction methodologies.

Contact: xling@tularik.com

INTRODUCTION

In February 2001, the International Human Genome Sequencing Consortium (HGSC) and Celera Genomics simultaneously published descriptions of the sequencing, assembly, analysis, and gene annotation of the human genome (IHGSC, 2001; Venter *et al*., 2001). Although both teams identified approximately 30 000 human genes (IHGSC, 2001; Venter *et al*., 2001), a direct comparison of the Celera and HGSC (Ensembl) data sets revealed relatively little overlap between their novel predicted genes (Hogenesch *et al*., 2001). Our previous parallel analysis (Li *et al*., 2002) of the two genome assemblies showed that there are major fundamental differences between these two data sets, in the numbers, identities, and properties of predicted genes derived from these sequences, and that assembly-level differences must be at least partly responsible for the gene set discrepancies. In addition, the recent re-analyses (Myers *et al*., 2002; Waterston *et al*., 2003; Adams *et al*., 2003) of Celera's genome assembly debated how much of an impact Celera's use of the public-domain genome data had on its assembly. In order to provide an up-to-date status report of the human genome sequencing efforts, understand how the genome assemblies have been evolving since their initial releases, and compare the different assembly approaches and their resulting gene data sets, we have collected the majority of HGSC and Celera assembly releases and performed a systematic comparative analysis.

METHODS

Sequence databases

HGSC and Celera database of assemblies and transcriptomes, released from May 2000 to July 2002, were collected and summarized in Table 1. A total of nine HGSC human genome assemblies (June 2000, July 2000, September 2000, October 2000, December 2000, April 2001, August 2001, December 2001, April 2002) were downloaded from http://www.genome.ucsc.edu/#Downloading. Ensembl curated gene sets (Ensembl 0.8.0, Ensembl 1.0.0, Ensembl 1.2.0, Ensembl 3.26 and Ensembl 5.28) were downloaded from ftp.ensembl.org. Five Celera human genome assembly releases (R20, R25h, R26b, R26f and R26i) and four Celera gene sets (R25e, R25h, R26b, R26k) were licensed from subscription of the Celera Discovery System by GNF and analyzed by GNF (The Genomic Institute of Novartis

*To whom correspondence should be addressed.
† The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

Table 1. HGSC and Celera genome assembly and gene set release history

Release date	Assembly		Curated genes	
	HGSC (UCSC)	Celera	HGSC (Ensembl)	Celera
05-2000		R18, R19		
06-2000	06-2000	R20, R21		
07-2000	07-2000	R22, R23		
08-2000		R24		
09-2000	09-2000		E−0.8.0	
10-2000	10-2000			
11-2000		R25e		
12-2000	12-2000		E−1.0.0	R25e
01-2001		R25h		R25h
04-2001	04-2001		E−1.1.0	
07-2001		R26b		R26b
08-2001	08-2001		E−1.2.0	
10-2001				R26d
11-2001				R26e
12-2001	12-2001			
01-2002		R26f	E−3.26	R26f, R26h
03-2002			E−4.28	
04-2002	04-2002			R26j
05-2002		R26i	E−5.28	
06-2002				R26k

The release dates and release names (where applicable) are shown for the HGSC and Celera genome assembly releases analyzed in this study. The Ensembl and Celera gene set releases are also shown.

Research Foundation). Human RefSeq sequences were obtained by FTP from ftp.ncbi.gov/refseq/H_sapiens. The PFAM 7.0 Hidden Markov Model (HMM) database was obtained by FTP from ftp.genetics.wustl.edu/pub/eddy/pfam 7.0/. The Research Genetics cDNA database was obtained by FTP from ftp://ftp.resgen.com/pub/sv_libraries/RG_Hs_seq_ver_101100.txt. 07-2002 RefSeq database was downloaded from NCBI ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/ site.

Local genome database setup and configuration

UCSC annotation databases hg4, hg5, hg6, hg7, hg8, hg10, and hg11, corresponding to the September 2000, October 2000, December 2000, April 2001, August 2001, December 2001 and April 2002 UCSC genome assemblies respectively, were downloaded, and imported into a local relational database. The UCSC relational database schema is available online at http://genome.ucsc.edu/goldenPath/gbdDescriptions.html

Ensembl databases were set up and configured on local servers following instructions from http://www.ensembl.org/Docs/ and personal communications with Ensembl colleagues (ensembl-dev@ebi.ac.uk). Data sets were downloaded from Ensembl and imported to a local relational database.

BLAT to map sequences onto genome assemblies

Gene sequences were mapped onto genome assemblies using the BLAT program (Kent, 2002) directly (local BLAT server setup) or indirectly (locally installed UCSC genome database with pre-computed BLAT results). In the UCSC genome database, chromosome locations are stored in the all_est or all_mrna tables of which the qName column stores RG Genbank accession numbers. The BLAT server setup and homology search were performed using instructions from UCSC. BLAT analysis was run using an identity threshold of 95% over at least 40 bp as described at UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start&org=human). These criteria have been previously determined to give optimal sensitivity, specificity, speed for genomic searches (Kent, 2002). Similar results were obtained when sequences were mapped by running BLAT or by querying pre-computed BLAT results from UCSC database.

Gene prediction

Predicted gene sets were derived from the HGSC and Celera genome assemblies by running the GENSCAN algorithm (Burge and Karlin, 1997) with its default settings. Full-length gene sets, were derived from these total gene sets by selecting all predicted genes for which GENSCAN identified 5′ promoter and the 3′ poly-A signal sequences.

BLAST comparative data analysis

Sequence comparison was performed using the NCBI BLAST algorithm (Altschul *et al.*, 1997): BLASTN for gene–gene

comparisons (E-value $< 1 \times 10^{-5}$, at least 98% identity over 100 bp) and BLASTX for gene/SWISS-PROT comparisons (E-value $< 1 \times 10^{-5}$).

PFAM domain analysis

The PFAM 7.0 database release, containing 3360 HMMs, was used to analyze gene sets for their protein domain content. For this analysis, the HMMER software package (Eddy, 1998) or its compatible implementations from Paracel (http://www.paracel.com) and TimeLogic (http://www.timelogic.com) were run on a Linux computing cluster (150 CPUs, Linux Networks), a Paracel GENEMATCHER machine, and a TimeLogic Decypher machine, respectively.

RESULTS AND DISCUSSION

Gene-based quality assessment of HGSC and Celera genome assemblies

Multiple releases of human genome assemblies and their associated predicted gene sets from HGSC (International Human Genome Sequencing Consortium, UCSC, Ensembl) and Celera are listed in Table 1 based on release dates. These data sets were the basis for comparing the HGSC and Celera genome assemblies and analyzing how they have changed over time. Genome assemblies can vary due to differences in local sequence content as well as long-range differences due to differing sequence assembly. As a gauge of the quality and completeness of the draft local sequence content in both genome assemblies, we used the BLAT algorithm (Kent, 2002) to map the large Research Genetics human cDNA sequence database (RG, 41 472 sequences) against the genome assemblies (Fig. 1). Since a positive BLAT hit only requires a match of 40 bp, this analysis should be largely insensitive to global assembly issues. We have observed a gradual increase in the number of mapped RG sequences with both HGSC and Celera assemblies, leveling off for both at around 97%. These results suggest that the HGSC and Celera assemblies have had similar local sequence content since their early releases.

Gene sets derived from the genome assemblies can vary due to differences in local sequence, global assembly, and the particular gene-prediction pipelines used. Since genes can span large sequence lengths, all gene prediction algorithms, to some extent, will be sensitive to sequence coverage and assembly issues. To eliminate variability due to differing gene-prediction pipelines, GENSCAN was used to generate two sets of genes from multiple releases of both genome assemblies. The full-length GENSCAN genes subsets were extracted from the full sets, including only those GENSCAN predictions containing both 5′ promoter and 3′ poly-adenylation signal sequence predictions. Since long-range sequence discontinuity in the assemblies can lead GENSCAN to predict partial genes that would lack 5′ promoter and/or 3′ poly-adenylation signal sequences, this full-length subset can be used to probe the quality of the genome assembly.

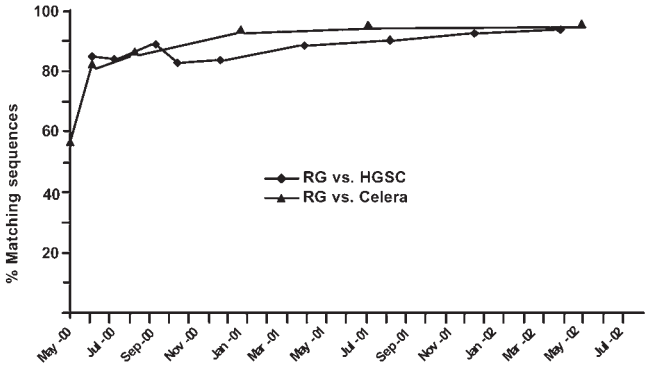

Fig. 1. BLAT mapping of Research Genetics sequences to HGSC and Celera genome assemblies. The percentages of sequences from the Research Genetics sequence database which give positive BLAT hits against various releases of the HGSC and Celera genome assemblies are plotted.

The total and full-length GENSCAN HGSC gene counts as well as the Celera full-length GENSCAN gene counts all showed modest and gradual increases over time (Fig. 2A). In contrast, the total GENSCAN gene counts for the Celera assemblies started out at levels more than twice as high as the HGSC gene sets, and only came down to comparable levels in the July 2001 release. Since gene prediction depends on not only local sequence content but also on long-range assembled sequence, we believe that the initially high total GENSCAN gene numbers for Celera were due to sequence fragmentation resulting in many individual genes being split into separate GENSCAN predictions. This apparent Celera genome fragmentation, perhaps due to gaps or assembly errors, may indicate a disadvantage of Celera's whole genome shotgun (WGS) sequencing approach (Huson *et al*., 2001; Myers *et al*., 2002) compared to HGSC's hierarchical shotgun (HS) approach (IHGSC, 2001).

Both the Ensembl gene build system (Hubbard *et al*., 2002) and Celera's Otto pipeline (Venter *et al*., 2001) use various forms of evidence including homology to known proteins, ESTs and *ab initio* gene prediction with algorithms including GENSCAN (Burge and Karlin, 1997). Ensembl is more dependent on known human proteins from SPTREMBL, GENSCAN predictions, and gene prediction HMMs while Celera uses more data from outside of its genome such as cross-genome homology and even the Ensembl gene set (Venter *et al*., 2001, reference 62). Analyzing the length distributions of the Ensembl and Celera gene sets (Fig. 2B) shows a large decrease in short Celera genes accompanied by increases in the numbers of longer genes over time, similar but more pronounced than what is seen with the HGSC genes. A similar trend is seen with the GENSCAN-predicted gene sets (data not shown), further reinforcing the notion that initial

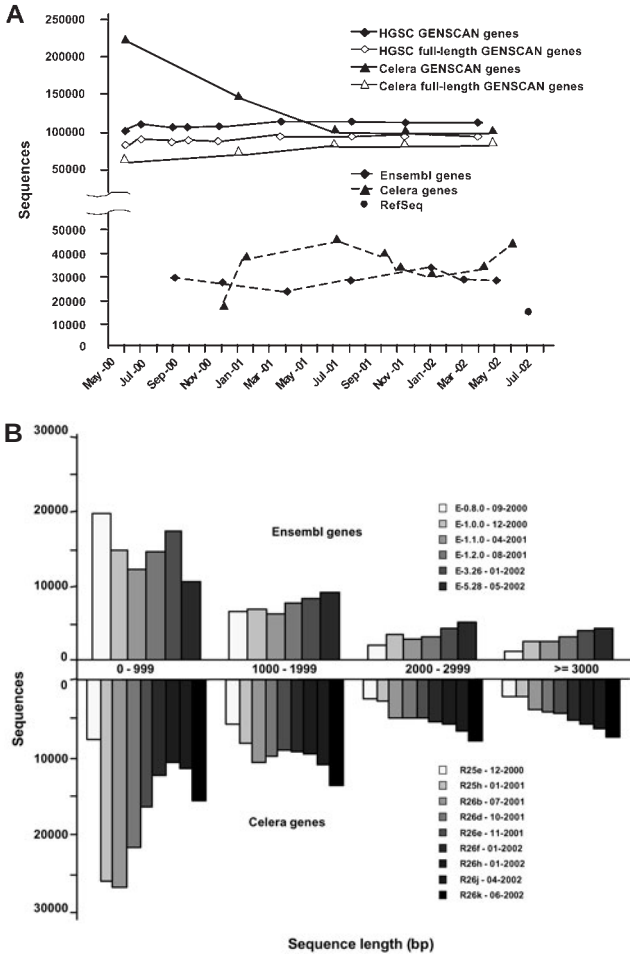

Fig. 2. Gene set distributions from multiple HGSC and Celera genome releases. (**A**) The numbers of pipeline-derived genes from various releases of Ensembl and Celera gene sets along with the numbers of total GENSCAN-predicted genes and full-length GENSCAN-predicted genes derived from various releases of the HGSC and Celera genomes are plotted based on release dates. For reference, the number of human genes in the July 2002 release of RefSeq is also shown. (**B**) Multiple Ensembl and Celera gene sets were analyzed based on gene length. The numbers of sequences from each release that lie in the given gene-length bins are shown.

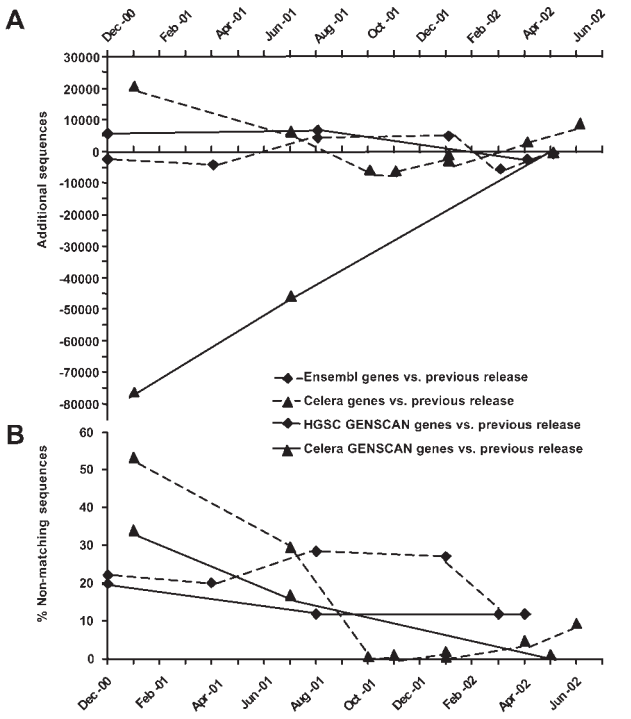

Fig. 3. Gene changes within gene sets across multiple releases. (**A**) The changes in the numbers of genes in the Ensembl and Celera gene sets and in the HGSC- and Celera-derived GENSCAN genes between successive genome releases are plotted. (**B**) Genes in successive gene sets were compared to genes in the previous gene sets using BLAST. The percent of genes that did not match any sequence in the previous gene set are plotted for each gene set group.

Celera assembly releases may have had comparatively high levels of fragmentation. Interestingly, the latest two Celera gene sets released show a reversal of this gene-length trend, with increasing numbers of short genes concomitant with an increase in total gene number.

Within-group gene set comparisons

An alternate way to look at changes in the assembly and gene data set is to compare the genes derived from each genome assembly release with those from the previous release. BLAST (Altschul *et al*., 1997) comparative analysis of genes with those from previous releases identified new genes as those sequences that did not match any sequence in the previous set. The analysis of the HGSC GENSCAN gene sets shows a 10–20% level of new gene content per gene set (Fig. 3), consistent with the modest increases in gene number (Fig. 2) and sequence coverage (Fig. 1) already observed. In contrast, the Celera GENSCAN gene sets show an initially high level of new GENSCAN gene content being added (30–40%) concomitant with a large decrease in gene number, a trend that has diminished in the most recent genome releases, where very few new GENSCAN genes appear to be present. The large gene count of the initial Celera GENSCAN set and its decrease over the course of time correlates with the decrease of the initial large fraction of short (<1 kb) Celera genes (Fig. 2), suggesting that the levels of fragmentation seen in the initial releases decrease overtime. The pattern of changes in the Celera Otto genes in successive releases is even more dramatic: more than 50% of the genes in the January 2001 gene set release cannot be found in the previous December 2000 release. By October 2001, however, virtually no new genes were being added. Interestingly, new gene addition can again be observed in the recent Celera releases, occurring in the same releases where the total gene number (Fig. 2A) and

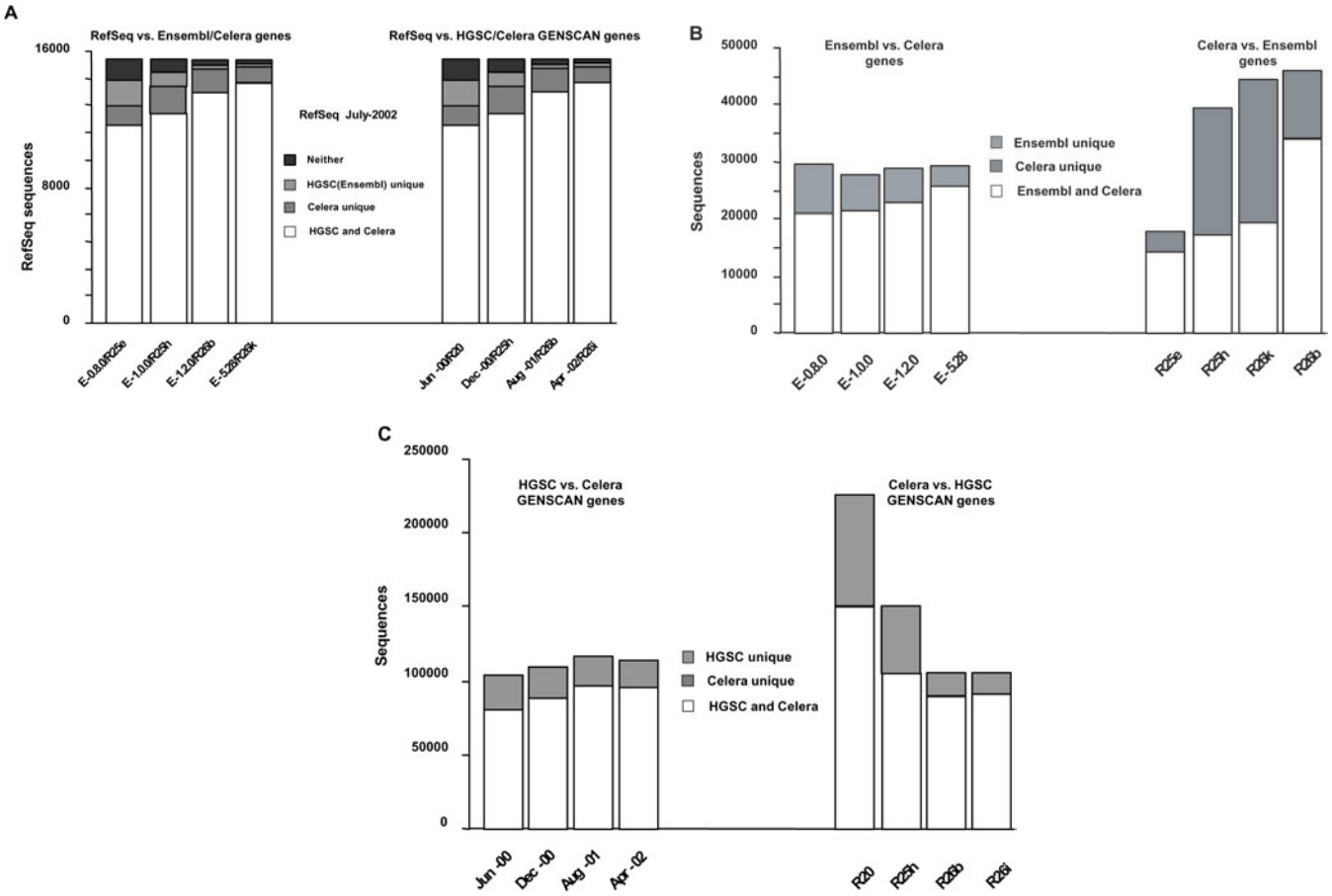

Fig. 4. Gene set comparisons between groups. (**A**) Human RefSeq genes were compared to multiple Ensembl and Celera gene sets using BLAST. The numbers of RefSeq sequences that matched both gene sets, only the Ensembl gene set, only the Celera gene set, and neither gene set are plotted on the left. The human RefSeq genes were also compared to multiple HGSC- and Celera-derived GENSCAN gene sets using BLAST. The distribution of matching sequences is plotted on the right. (**B**) Ensembl genes from multiple releases were compared with the corresponding Celera gene set releases using BLAST. The numbers of matching and non-matching (Ensembl-unique) sequences are plotted on the left. Similarly, Celera genes were compared with the corresponding Ensembl gene sets using BLAST and the numbers of matching and non-matching (Celera-unique) sequences are plotted on the right. (**C**) HGSC-derived GENSCAN genes and Celera-derived GENSCAN genes were compared with each other using BLAST in both directions as in (B). The numbers of sequences found in both gene sets, HGSC-unique sequences, and Celera-unique sequences are plotted.

the short gene number (Fig. 2B) rebound. Since neither the genome content nor quality appears to have changed much in these releases, we believe that this recent trend is likely due to changes in Celera's gene-prediction pipeline.

RefSeq-based quality assessment of ensembl and celera gene sets

The NCBI RefSeq database (Maglott *et al*., 2000; Pruitt and Maglott, 2001), derived from Genbank sequences and the published literature, provides a non-redundant view of the current knowledge about human genes, transcripts and proteins. We evaluated the quality and comprehensiveness of the *in silico* predicted gene sets, by comparing them to the human RefSeq database with BLAST. Comparing RefSeq to multiple Ensembl and Celera pipeline gene sets, and to HGSC and Celera GENSCAN gene sets reveals that, even with the earliest releases, greater than 75% of RefSeq genes can be found in some form in gene sets from both groups (Fig. 4A). Small fractions of RefSeq genes could be matched only to genes from HGSC, only to Celera genes, or to neither gene set. Over the course of time, the numbers of unmatched RefSeq genes and those matching only HGSC have significantly decreased. At the same time, the Celera gene set continues to have a modest number of RefSeq genes not found in Ensembl, suggesting that the Celera gene set can be more comprehensive than the Ensembl data set with respect to RefSeq. Similar BLAST results were obtained after a permissive sequence clustering approach (Hogenesch *et al*., 2001) was applied to eliminate sequence redundancy in all RefSeq, HGSC and Celera gene sets (data not shown). Because RefSeq (07-2002 release,

15 740 genes) contains far fewer genes than Ensembl and Celera, more efforts are needed in order to complete RefSeq as a gene reference standard.

Between-group gene set comparisons

Much has been made of the concordance between the gene numbers of the initial HGSC and Celera gene releases (IHGSC, 2001; Venter *et al*., 2001) and the subsequent observations that each set actually contained many unique genes (Hogenesch *et al*., 2001; Li *et al*., 2002). We have repeated this analysis across multiple gene set releases. Comparing Ensembl to Celera genes shows that the fraction of Ensembl-unique genes ranges from 29% initially to 12% in the most recent release analyzed (Fig. 4B), indicating that most of the Ensembl genes can find matches in the Celera set. The reverse comparison, Celera compared to Ensembl, reveals that the fraction of Celera-unique genes decreased from an initial 56 to 26% in the most recent analyzed release. The large increase in Celera-unique genes in R25h release coincided with the large increase in total gene number (Fig. 2A) consisting largely of short genes (Fig. 2B). Similar results were obtained when redundancy was removed from the data sets (data not shown).

To discriminate between changes in actual sequence information versus changes in gene-prediction pipelines, this analysis was repeated with the GENSCAN-derived gene sets. The HGSC versus Celera GENSCAN gene set comparison (Fig. 4C) looks much like the Ensembl versus Celera pipeline gene comparison (Fig. 4B), with approximately 16% of the HGSC genes being unique. In contrast, the Celera versus HGSC GENSCAN-gene comparison shows an initially high number (33%) of Celera-unique genes, decreasing to a fraction (13%) similar to the number of unique HGSC GENSCAN genes. The difference between these results and the pipeline-gene comparison suggests that the unique gene content of the Celera pipeline gene set cannot be explained by fundamental differences in the genome assemblies.

To further characterize the HGSC and Celera-unique gene sets, we mapped the unique genes back to the genome assemblies from which they came as well as to that of the other group using BLAT. Nearly all of the sequences from all four unique gene sets can be mapped to both genome assemblies of the same or similar release date (Fig. 5A). This again confirms that genome content, specifically the local sequence content, is very similar between both assemblies. Since the differences between Ensembl and Celera gene sets are much larger than that observed between HGSC and Celera GENSCAN gene sets, we can conclude that the gene-building process, including human curation, must have contributed more to the observed gene set difference than the different genome sequencing and assembly processes.

In order to estimate how likely the unique Ensembl or Celera genes are to represent true genes, we compared the unique pipeline genes to the large SWISS-PROT protein database using BLAST with moderate stringency (E-value $= 1e-5$).

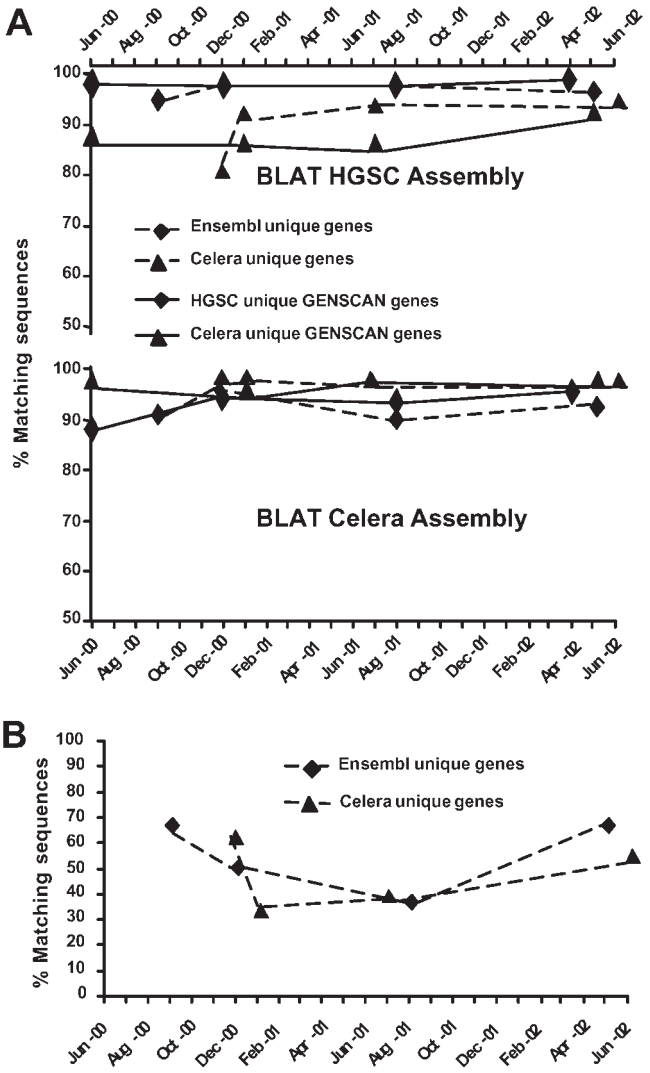

Fig. 5. Analysis of HGSC and Celera-unique genes. (**A**) Sequences that were unique to the Ensembl, Celera, HGSC-derived GENSCAN, and Celera-derived GENSCAN gene sets based on BLAST analysis (Fig. 4) were mapped back to the genome assembly from which they were derived as well as to the other genome assembly using BLAT. The percentages of sequences from each unique set, which could be mapped to either genome assemblies, are plotted. (**B**) The unique Ensembl and Celera genes were compared with the SWISS-PROT database using a moderate-stringency BLAST analysis. The percentages of sequences from both sets for which homologous sequences could be identified in SWISS-PROT are plotted.

While more than 60% of some of the earlier unique gene sets appear to have no significant homology to any protein sequence in SWISS-PROT, analysis of the most recent gene sets shows that 55% of Celera-unique genes and 68% of Ensembl unique genes have known protein homologs (Fig. 5B). Using SWISS-PROT homology matches as a rough estimate of the likelihood that predicted genes are real, it

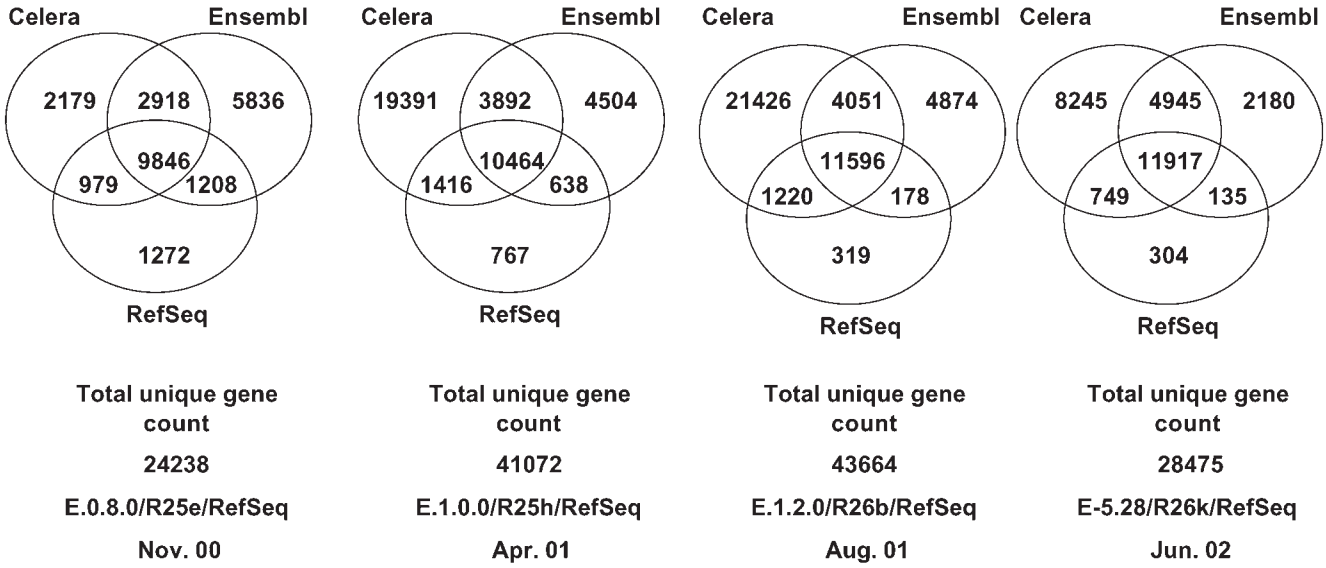

Fig. 6. Estimation of non-redundant gene count. For the releases shown, the Ensembl and Celera gene sets were combined along with the human subset of RefSeq. This combined gene set was clustered via a permissive clustering algorithm. The resulting gene cluster number represents the total number of unique genes in the Ensembl, Celera and RefSeq gene sets that could be resolved by our BLAST analysis.

appears that a large fraction of the unique genes from both data sets are likely to be real.

Total number of protein-coding genes—lower bound estimation

As shown in Figure 2A, the Ensembl gene sets have consistently been comprised of around 30 000 sequences, while the Celera gene set has varied in the range of 20 000–45 000 sequences. Interestingly, the two latest Celera gene sets analyzed show an increase in gene number, bringing the total well above that of the Ensembl gene set. To put these numbers in perspective, the human component of RefSeq (Maglott *et al*., 2000; Pruitt and Maglott, 2001) contains much fewer genes (07-2002 release, 15 740 genes) than either of these two gene sets.

In order to estimate the total gene number, the Ensembl, Celera and RefSeq gene sets were combined into a large superset. Following an all-to-all BLAST comparison, redundant sequences were removed with a permissive clustering algorithm (Hogenesch *et al*., 2001). The resulting gene cluster number represents the total number of unique genes in the Ensembl, Celera and RefSeq gene sets that could be resolved by our BLAST analysis. Different Ensembl and Celera releases were combined with RefSeq and processed to analyze how this total gene number has changed over time, increasing from an initial 24 238 to over 40 000 and then down to 28 475 (Fig. 6). The non-redundant gene number we computed here should represent a lower bound for the true human gene count: our BLAST threshold cannot distinguish between the nearly identical paralogs that are found in some gene families; this approach omits genes that were missed by both Ensembl and Celera gene identification processes. This analysis of multiple gene sets together, coupled with the removal of redundancy, allows us to make a more complete estimate of the total human genome gene content than has previously been described (IHGSC, 2001; Venter *et al*., 2001).

Gene set domain content analysis

Similar to the SWISS-PROT homology analysis (Fig. 5B), protein domain profiling should provide an indirect measure of the quality of the genome-derived gene sets. The drawback of this analysis is that it can only analyze genes that contain already known protein domains. We used the PFAM 7.0 (Bateman *et al*., 2002) database of domain models to look at the comparative domain content of gene sets from the HGSC and Celera genome assemblies. Figure 7A shows the numbers of PFAM models that have an excess of matches against various releases of either the Ensembl or Celera gene sets. In early releases, many more PFAM models had more matches against the Ensembl gene set than against the Celera gene set. However in recent releases, the domain content of the Celera gene set has increased dramatically relative to Ensembl. In contrast, when the GENSCAN gene sets are analyzed (Fig. 7B), while the gap has narrowed, the HGSC genes continue to contain more domain matches than the Celera GENSCAN genes. Similar to the SWISS-PROT homology analysis (Fig. 5B), this domain analysis should provide an approximate measure of the quality of the genome-derived gene sets. The GENSCAN-derived gene set numbers suggest that over time the Celera genome assembly has approached the quality of the HGSC assembly. Given the similarity of local sequence content between the HGSC and Celera assemblies,

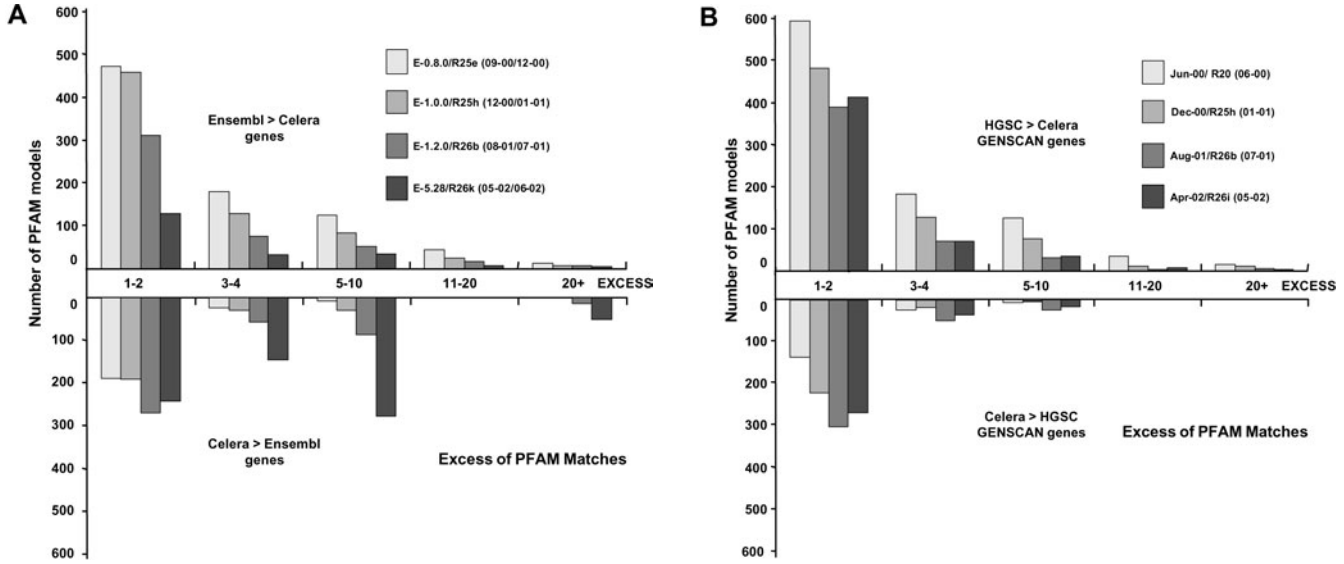

Fig. 7. Domain profiling of HGSC and Celera gene sets. The domain content of multiple HGSC and Celera gene sets was analyzed by performing a search of these gene sets with the PFAM database. For each PFAM domain model, the number of hits against each pair of HGSC and Celera gene sets was identified. The numbers of PFAM models that have an excess of hits against HGSC are plotted in the upper section, based on how large the excess of HGSC hits was. Similarly, the numbers of PFAM models that have an excess of hits against Celera are plotted in the lower section based on the number of excess Celera hits. (**A**) PFAM analysis of Ensembl and Celera gene sets. (**B**) PFAM analysis of HGSC and Celera assembly GENSCAN gene sets.

this PFAM analysis of the GENSCAN gene sets supports the idea that the HGSC HS approach may have had advantages over the Celera WGS approach. The significant difference in PFAM matches to the recent Celera pipeline gene sets, in contrast, suggests that Celera has been able to add many new gene types to their gene set that would not otherwise be identified by *ab initio* gene prediction, making their gene annotation efforts more comprehensive than that of Ensembl.

DISCUSSION

Numerous reports comparing the HGSC and Celera genome assemblies (Aach *et al*., 2001; Olivier *et al*., 2001; Li *et al*., 2002; Xuan *et al*., 2003) and gene sets (Hogenesch *et al*., 2001) have been made since the simultaneous publication of the two genomes in February 2001 (IHGSC, 2001; Venter *et al*., 2001). The analysis presented here suggests that the initial HGSC genome assembly, although containing a similar amount of genomic sequence information as the Celera genome assembly, was in a much better state of assembly. This is not entirely unexpected as whole genome shotgun sequencing, the technique used by Celera, is more challenging to assemble than HGSC's hierarchical shotgun approach. Over the course of two years, however, Celera has made up for the shortcomings of their initial assemblies with newer assemblies that have approached the quality of HGSC's draft genomes. Since the Ensembl gene build system predicts genes through GENSCAN, homology, and gene prediction HMM methods, the quality and quantity of their gene predictions should mirror the quality of the genome assembly, as we have observed. In contrast, Celera uses a richer gene prediction pipeline named Otto that places greater emphasis on cross-species genome comparisons, EST homology, and curated gene set homology (Venter *et al*., 2001). By incorporating information in addition to its genome sequence, Celera has been able to generate a larger, more unique gene set. While many of the predicted genes unique to both the Ensembl and Celera gene sets are likely to be proven not to be *bona fide* genes (Fig. 5B; Hogenesch *et al*., 2001), we expect that a significant number of them will be validated when the full content of the human transcriptome is finally determined.

ACKNOWLEDGEMENTS

We thank Jim Kent (UCSC) and the members of the Ensembl project (UK) for various technical assistance and help in HGSC genome database setup, and Tularik/GNF Bioinformatics and IT staff for outstanding computational support. The authors are also grateful to Drs Greg Peterson and Zheng Pan for critical discussions.

REFERENCES

Aach,J., Bulyk,M.L., Church,G.M., Comander,J., Derti,A. and Shendure,J. (2001) Computational comparison of two draft sequences of the human genome. *Nature*, **409**, 856–859.

Adams,M.D., Sutton,G.G., Smith,H.O., Myers,E.W. and Venter,J.C. (2003) The independence of our genome assemblies. *Proc. Natl Acad. Sci. USA*, **100**, 3025–3026.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Hogenesch,J.B., Ching,K.A., Batalov,S., Su,A.I., Walker,J.R., Zhou,Y., Kay,S.A., Schultz,P.G. and Cooke,M.P. (2001) A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell*, **106**, 413–415.

Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Huson,D.H., Reinert,K., Kravitz,S.A., Remington,K.A., Delcher,A.L., Dew,I.M., Flanigan,M., Halpern,A.L., Lai,Z., Mobarry,C.M. *et al*. (2001) Design of a compartmentalized shotgun assembler for the human genome. *Bioinformatics*, **17**, S132–S139.

International Human Genome Sequencing Consortium (IHGSC) (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Li,S., Liao,J., Cutler,G., Hoey,T., Hogenesch,J., Cooke,M., Schultz,P. and Ling,X. (2002) Comparative analysis of human genome assemblies reveals genome-level differences. *Genomics*, **80**, 138.

Maglott,D.R., Katz,K.S., Sicotte,H. and Pruitt,K.D. (2000) NCBI's LocusLink and RefSeq. *Nucleic Acids Res.*, **28**, 126–128.

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*, **19**, 19.

Olivier,M., Aggarwal,A., Allen,J., Almendras,A.A., Bajorek,E.S., Beasley,E.M., Brady,S.D., Bushard,J.M., Bustos,V.I., Chu,A. *et al*. (2001) A high-resolution radiation hybrid map of the human genome draft sequence. *Science*, **291**, 1298–1302.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.

Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

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.

Waterston,R.H., Lander,E.S. and Sulston,J.E. (2003) More on the sequencing of the human genome. *Proc. Natl Acad. Sci. USA*, **100**, 3022–3024.

Xuan,Z., Wang,J. and Zhang,M.Q. (2003) Computational comparison of two mouse draft genomes and the human golden path. *Genome Biol.*, **4**, R1.