Introduction: putting it together

doi:10.1038/ng965

In its short history, the Human Genome Project (HGP) has provided significant advances in the understanding of gene structure and organization, genetic variation, comparative genomics and appreciation of the ethical, legal and social issues surrounding the availability of human sequence data. One of the most significant milestones in the history of this project was met in February 2001 with the announcement and publication of the draft version of the human genome sequence[1]. The significance of this milestone cannot be understated, as it firmly marks the entrance of modern biology into the genome era (and not the post-genome era, as many have stated). The potential usefulness of this rich databank of information should not be lost on any biologist: it provides the basis for 'sequence-based biology', whereby sequence data can be used more effectively to design and interpret experiments at the bench. The intelligent use of sequence data from humans and model organisms, along with recent technological innovation fostered by the HGP, will lead to important advances in the understanding of diseases and disorders having a genetic basis and, more importantly, in how health care is delivered from this point forward[2].

Although this flood of data has enormous potential, many investigators whose research programs stand to benefit in a tangible way from the availability of this information have not been able to capitalize on its potential. Some have found the data difficult to use, particularly with respect to incomplete human genome draft sequence information. Others are simply not sufficiently conversant with the seeming myriad of databases and analytical tools that have arisen over the last several years. To assist investigators and students in navigating this rapidly expanding information space, numerous World Wide Web sites, courses and textbooks have become available; many individuals, of course, also turn to their friends and colleagues for guidance. We have prepared this Guide in that same spirit, as an additional resource for our fellow scientists who wish to make use (or better use) of both sequence data and the major tools that can be used to view these data. The Guide has been written in a practical, question-and-answer format, with step-by-step instructions on how to approach a representative set of problems using publicly available resources. The reader is encouraged to work through the examples, as this is the best way to truly learn how to navigate the resources covered and become comfortable using them on a regular basis. We suggest that readers keep copies of the Guide next to their computers as an easy-to-use reference.

Before embarking on this new adventure, it is important to review a number of basic concepts regarding the generation of human genome sequence data. This review does not discuss the chronological development of the HGP or provide an in-depth treatment of its implications; the reader is referred to *Nature*'s Genome Gateway (http://www.nature.com/genomics/human/) for more information on these topics.

Current status of human genome sequencing

Sequencing of the human genome is nearing completion. The target date for making the complete, high-accuracy sequence available is April 2003, the 50th anniversary of the discovery of the double helix[3]. As we go to press, however, the work is still a mosaic of finished and draft sequence. A sequence becomes finished when it has been determined at an accuracy of at least 99.99% and has no gaps. Sequence data that fall short of that benchmark but can be positioned along the physical map of the chromosomes are termed 'draft'. Currently, 87% of the euchromatic fraction of the genome is finished and less than 13% is at the draft stage.

Even in this incomplete state, the available data are extremely useful. This usefulness was apparent early on, leading the International Human Genome Sequencing Consortium (IHGSC) to pursue a staged approach in sequencing the human genome. The first stage generated draft sequence across the entire genome[1]. The project is now well advanced into its second stage, with draft sequence being improved to 'finished quality' across the entire genome, a necessarily localized process. As a result, and as it has been presented to date, the human genome sequence is an evolving mix of both finished and unfinished regions, with the unfinished regions varying in data quality. As the data are initially made available in raw form, with subsequent refinement and improvement, and because data of different quality are found in different places in the genome, users must understand the kinds of data presented by the various tools available.

Determining the human sequence: a brief overview

As with all systematic sequencing projects, the basic experimental problem in sequencing lies in the fact that the output of a single reaction (a 'read') yields about 500–800 bp[1,4]. To determine the sequence of a DNA molecule that is millions of bases long, it must first be fragmented into pieces that are within an order of magnitude of the read size. The sequence at one or both ends of many such fragments is determined, and the pieces are then 'assembled' back into the long linear string from which they were originally derived. A number of approaches for doing this have been suggested and tested; the most commonly used is shotgun sequencing[4]. The application of shotgun sequencing to the multimegabase- or gigabase-sized genomes of metazoans is still evolving. A small number of strategies are currently being evaluated, for example, hierarchical or map-based shotgun sequencing, whole-genome shotgun sequencing and hybrid approaches. These approaches are described in detail elsewhere[4].

The IHGSC's human sequencing effort began as a purely map-based strategy and evolved into a hybrid strategy[1]. The 'pipeline' that the IHGSC used to generate the human sequence data involved the following steps.

1. Bacterial artificial chromosome (BAC) clones were selected, and a random subclone library was constructed for each one in either an M13- or a plasmid-based vector.

2. A small number of members of the subclone library (usually 96 or 192) were sequenced to produce very-low-coverage, single-pass or 'phase 0' data. These data were used for quality control and can be found in the Genome Survey Sequence division of The DNA Database of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL) and GenBank (of the National Center for Biotechnology and Information; NCBI).

3. If a BAC clone met the requisite standard, subclones were derived and sufficient sequence data generated from these to provide four- to fivefold coverage (that is, enough data to represent an average base in the BAC clone between four and five times). This is known as 'draft-level' coverage, and permits the assembly

NCBI reference sequences

The data release and distribution practices adopted by the HGP participants have led not only to very early, pre-publication access to this treasure trove of information, but also to a potentially confusing variety of formats and sources for the sequence data. To address this and other issues, the NCBI initiated the RefSeq project (http://www.ncbi.nlm.nih.gov/locuslink/refseq.html).

The goal of the RefSeq effort is to provide a single reference sequence for each molecule of the central dogma: DNA, the mRNA transcript, and the protein. The RefSeq project helps to simplify the redundant information in GenBank by providing, for example, a single reference for human glyceraldehyde-3-phosphate dehydrogenase mRNA and protein, out of the 14 or so full-length sequences in GenBank. Each alternatively spliced transcript is represented by its own reference mRNA and protein. The RefSeq project also includes sequences of complete genomes and whole chromosomes, and genomic sequence contigs. The human genomic contigs that NCBI assembles, which form the basis of the presentations in the different genome browsers, are part of the RefSeq project. Most RefSeq entries are considered provisional and are derived by an automated process from existing GenBank records. Reviewed RefSeq entries are manually curated and list additional publications, gene function summaries and sometimes sequence corrections or extensions.

Reference sequences are available through NCBI resources, including Entrez, BLAST and LocusLink. They can be easily recognized by the distinctive style of their accession numbers. NM_###### is used to designate mRNAs, NP_###### to designate proteins and NT_###### to designate genomic contigs. The NCBI and UCSC use alignments of the mRNA RefSeqs with the genome to annotate the positions of known genes. Ensembl aligns mRNA RefSeqs to the genome. The NCBI also provides model mRNA RefSeqs produced from genome annotation. These are derived by aligning the NM_ mRNAs and other GenBank mRNAs to the assembled genome and then extracting the genomic sequence corresponding to the transcripts. The resulting model mRNA and model protein sequences have accession numbers of the form XM_###### and XP_######. As the XM_ and XP_ records are derived from genomic sequence, they may differ from the original NM_ or GenBank mRNAs because of real-sequence polymorphisms, errors in the genomic or mRNA sequences or problems in the mRNA/genomic sequence alignment. A complete list of types of RefSeqs, along with details on how they are produced, is available from http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html.

of sequence using computer programs that can detect overlaps between the random reads from the subclones, yielding longer 'sequence contigs'. At this stage, the sequence of a BAC clone could typically exist on between four and ten different contigs, only some of which were ordered and oriented with respect to one another. The BAC 'projects' were submitted, within 24 hours of having been assembled, to the High-Throughput Genomic Sequences (HTGS) division of DDBJ/EMBL/GenBank[5], where each was given a unique accession number and identified with the keyword 'htgs_draft'. (The DDBJ, EMBL and GenBank are members of the International Nucleotide Sequence Database Collaboration, whose members exchange data nightly and assure that the sequence data generated by all public sequencing efforts are made available to all interested parties freely and in a timely fashion.) Less-complete high-throughput genomic (HTG) records are also known as 'phase 1' records. As the sequence is refined, it is designated 'phase 2'. In the context of a BLAST search at the NCBI, these sequences would be available in the HTGS database.

4. In late 2000, the draft sequence of the entire human genome was assembled from the sequence of 30,445 clones (BAC clones and a relatively small number of other large-insert clones). This assembled draft human genome sequence was published in February 2001 and made publicly available through three primary portals: the University of California, Santa Cruz (UCSC), Ensembl (of the European Bioinformatics Institute; EBI) and the NCBI. The use of all three of these sites to obtain annotated information on the human genome sequence is the primary subject of this guide.

5. Subsequent to the generation and publication of the draft human genome sequence, work has continued towards finishing the sequencing. The final stage initially targeted draft-quality BAC clones. For each of these clones, enough additional shotgun sequence data are obtained to bring the coverage to eight- to tenfold, a stage referred to as 'fully topped-up'. The data from each fully topped-up BAC are reassembled, typically resulting in a smaller number of contigs (often in just a single contig) than at the draft level. The new assembly is again submitted to the HTGS division as an update of the existing BAC clone, now identified with the keyword 'htgs_fulltop'. The accession number of the clone stays the same, and the version number increases by one (AC108475.2, for example, becoming AC108475.3).

6. At this stage, there are, even for clones comprising a single contig, typically some regions that are of insufficient quality for the clone to be considered finished. If this is the case, the fully topped-up sequence is analyzed by a sequence finisher (an actual person) who collects, in a directed manner, the additional data that are needed to close the few remaining gaps and to bring any regions of low quality up to the finished sequence standard. While the clone is worked on by the finisher, the HTGS entry in GenBank is identified by the keyword 'htgs_activefin'. Once work on the clone has been completed, the keyword of the HTG record is changed to 'htgs_phase3', the version number is once again increased, and the record is moved from the HTGS division to the primate division of DDBJ/EMBL/GenBank. In the context of a BLAST search at NCBI, these finished BAC sequences would now be available in the nr ("non-redundant") database.

7. The finished clone sequences are then put together into a finished chromosome sequence. As with the initial draft assemblies, there are a number of steps involved in this process that use map-based and sequence-based information in calculating the maps. The final assembly process involves identifying overlaps between the clones and then anchoring the finished sequence contigs to the map of the genome; details of the process can be found on the NCBI web site (http://www.ncbi.nlm.nih.gov/genome/guide/build.html).

Initially, both the UCSC and NCBI groups generated complete assemblies of the human genome, albeit using different approaches. As noted on the UCSC web site, the NCBI assembly tended to have slightly better local order and orientation, whereas the UCSC assembly tended to track the chromosome-level maps somewhat better. Rather than having different assemblies based on the same data, IHGSC, UCSC, Ensembl and NCBI decided that it would be more productive (and obviously less confusing) to focus their efforts on a single, definitive assembly. To this end, and by agreement, the NCBI assembly will be taken as the reference human genome sequence. It is this NCBI assembly that is displayed at the three major portals covered in this guide.

Annotating the assemblies

Once the assemblies have been constructed, the DNA sequence undergoes a process known as annotation, in which useful sequence features and other relevant experimental data are coupled to the assembly. The most obvious annotation is that of known genes. In the case of NCBI, known genes are identified by simply aligning Reference Sequence (RefSeq) mRNAs (see box), GenBank mRNAs, or both to the assembly. If the RefSeq or GenBank mRNA aligns to more than one location, the best alignment is selected. If, however, the alignments are of the same quality, both are marked on to the contig, subject to certain rules (specifically, the transcript alignment must be at least 95% identical, with the aligned region covering 50% or more of the length, or at least 1,000 bases). Transcript models are used to refine the alignments. Ensembl identifies 'best in genome' positions for known genes by performing alignments between all known human proteins in the SPTREMBL database[6] and the assembly using a fast protein-to-DNA sequence matcher[7]. UCSC predicts the location of known genes and human mRNAs by aligning RefSeq and other GenBank mRNAs to the genome using the BLAST-like alignment tool (BLAT) program[8]. In addition to identifying and placing known genes onto the assemblies, all of the major genome browser sites provide *ab initio* gene predictions, using a variety of prediction programs and approaches.

Genome annotation goes well beyond noting where known and predicted genes are. Features found in the Ensembl, NCBI and UCSC assemblies include, for example, the location and placement of single-nucleotide polymorphisms, sequence-tagged sites, expressed sequence tags, repetitive elements and clones. Full details on the types of annotation available and the methods underlying sequence annotation for each of these different types of sequence feature can be found by accessing the URLs listed under Genome Annotation in the Web Resources section of this guide. At UCSC, many of the annotations are provided by outside groups, and there may be a significant delay between the release of the genome assembly and the annotation of certain features. Furthermore, some tracks are generated for only a limited number of assemblies. For an in-depth discussion of genome annotation, the reader is referred to an excellent review by Stein[9] and the references cited therein. This review, along with the Commentary in this guide, also provides cautions on the possible overinterpretation of genome annotation data.

The data—and sometimes the tools—change every day

The steps outlined in the previous section should emphasize that the state of the human genome sequence will continue to be in flux, as it will be updated daily until it has actually been declared 'finished'. (Finished sequence is properly defined as the "complete sequence of a clone or genome, with an accuracy of at least 99.99% and no gaps"[2]. A more practical definition is that of "essentially finished sequence," meaning the complete sequence of a clone or genome, with an accuracy of at least 99.99% and no gaps, except those that cannot be closed by any current method.) The reader should be mindful of this, not just when reading this guide, but also, when referring back to it over time. Similarly, the tools used to search, visualize and analyze these sequence data also undergo constant evolution, capitalizing on new knowledge and new technology in increasing the usefulness of these data to the user.

Over the next year, sequence producers will continue to add finished sequence to the nucleotide sequence databases, and the NCBI will continue to update the human sequence assembly until its ultimate completion. The human genome sequence will, however, continue to improve even after April 2003, as new cloning, mapping and sequencing technologies lead to the closure of the few gaps that will remain in the euchromatic regions. It is hoped that such technological advances will also allow for the sequencing of heterochromatic regions, regions that cannot be cloned or sequenced using currently available methods.

The sequence-based and functional annotations presented at the three major genome portals will certainly continue to evolve long after April 2003. Computational annotation is a highly active area of research, yielding better methods for identifying coding regions, noncoding transcribed regions and noncoding, non-transcribed functional elements contained within the human sequence.

Accessing human genome sequence data

Although each of the three portals through which users access genome data has its own distinctive features, coordination among the three ensures that the most recent version and annotations of the human genome sequence are available.

Ensembl (http://www.ensembl.org) is the product of a collaborative effort between the Wellcome Trust Sanger Institute and EMBL's European Bioinformatics Institute and provides a bioinformatics framework to organize biology around the sequences of large genomes[7]. It contains comprehensive human genome annotation through *ab initio* gene prediction, as well as information on putative gene function and expression. The web site provides numerous different views of the data, which can be either map-, gene- or protein-centric. Ensembl is actively building comparative genome sequence views, and presents data from human, mouse, mosquito and zebrafish. In addition, numerous sequence-based search tools are available, and the Ensembl system itself can be downloaded for use with individual sequencing projects.

The UCSC Genome Browser (http://genome.ucsc.edu) was originally developed by a relatively small academic research group that was responsible for the first human genome assemblies. The genome can be viewed at any scale and is based on the intuitive idea of overlaying 'tracks' onto the human genome sequence; these annotation tracks include, for example, known genes, predicted genes and possible patterns of alternative splicing. There is also an emphasis on comparative genomics, with mouse genomic alignments being available. The browser also provides access to an interactive version of the BLAT algorithm[8], which UCSC uses for RNA and comparative genomic alignments.

Given its Congressional mandate to store and analyze biological data and to facilitate the use of databases by the research community, the NCBI (http://www.ncbi.nlm.nih.gov) serves as a central hub for genome-related resources. NCBI maintains GenBank, which stores sequence data, including that generated by the HGP and other systematic sequencing projects. NCBI's Map Viewer provides a tool through which information such as experimentally verified genes, predicted genes, genomic markers, physical maps, genetic maps and sequence variation data can be visualized. The Map Viewer is linked to other NCBI tools—for example, Entrez, the integrated information retrieval system that provides access to numerous component databases.

Although we have chosen to illustrate each example using resources available at a single site, almost all the questions in this guide can be answered using any of the three browsers. The informational sidebars that follow some of the questions provide pointers on how to format the search at other sites. Furthermore, the three sites link to each other wherever possible. Examples presented in this Guide rely on the data and genome browser interfaces that were available in June 2002. As new versions of the genome assembly and viewing tools will come online every few months, the specifics of some of the examples may change over time. Regardless, the basic strategies behind answering the questions in the examples will remain the same. This underscores the importance of readers working through the examples at their own computers so that they may understand and be able to navigate these public databases. The readers are encouraged to explore the alternative methods for answering the questions.

Browser problems?

In following the question-and-answer portion of this guide, some readers may find that their web browsers are not be able to render the web pages properly. If this occurs, do one or more of the following:

1. Install the most recent version of either Netscape Navigator or Internet Explorer.
2. Increase the amount of memory available to the web browser.
3. Try a different web browser. In general, Macintosh users who seek to gain access to these three genome portals will see better performance with Internet Explorer.