

Tutorial

Bases and spaces: resources on the web for accessing the draft human genome

Colin Semple

Address: Medical Genetics Section, Department of Medical Sciences, The University of Edinburgh, Molecular Medicine Centre, Western General Hospital, Edinburgh EH4 2XU, UK. E-mail: Colin.Semple@ed.ac.uk

Published: 16 October 2000

Genome **Biology** 2000, **1**(4):reviews2001.1–2001.5

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2000/1/4/reviews/2001>

© Genome**Biology**.com (Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

Much is expected of the draft human genome sequence, and yet there is no central resource to host the plethora of sequence and mapping information available. Consequently, finding the most useful and reliable human genome data and resources currently available on the web can be challenging, but is not impossible.

Nice press release, shame about the data

The entire sequence of the human genome is not expected to be finished for some time, and gaps are expected to persist into 2003 [1]. In the meantime, the genome exists in 'draft' form: multiple segments of sequence in which we have high confidence, placed relative to one another by mapping information of lower confidence. Many biologists study particular regions of the genome, such as those involved in positional cloning of disease genes, and this type of work is greatly accelerated by having most of the sequence of the region of interest. The draft human genome now includes this information for most of the genome. Unfortunately, no single resource unites the available human genomic sequences with their locations and their gene content, but by combining the varied resources currently available it is possible to devise strategies that fully exploit the draft genome data. So what resources and information are available so far and where can we find them? Note that the databases and resources mentioned in this article are underlined and the corresponding URLs are listed in Table 1, and are linked online.

Raw sequence data

As recently as 1996, the entire GenBank database contained around 0.65 Gb of DNA sequence; but the draft human genome sequence alone runs to more than 3.08 Gb. Most of the draft sequence is present in GenBank as unfinished,

fragmentary BAC (bacterial artificial chromosome) sequences. These consist of a number of non-overlapping, arbitrarily ordered, fragments, or 'contigs', which have been artificially concatenated to produce a single sequence entry for each BAC. Typically, each contig within a BAC is separated from the next by a large number of bases, labeled 'N'. All unfinished BAC entries are subject to irregular updates until they are finished, and this might alter the number and size of the contigs they contain. The most straightforward web interface for retrieving BAC sequence (and various other types of data) is Entrez at the National Centre for Biotechnology Information (NCBI), which also includes substantial online documentation.

Most BAC sequence entries contain information about the BAC in the 'DEFINITION' field near the top of the Entrez display. For example, the DEFINITION field in the Entrez entry AP001002 contains the BAC clone name (678K21) and the cytogenetic band to which it has been localized (11q14). Many entries give much less annotation; for example, at present, Entrez entry AC007104 provides no clone name, nor even the clone library, and gives the location as simply 'chromosome 4'. I will discuss ways to find a more specific location for this clone below, but we can retrieve the clone name using a little-known feature of Entrez. Under the 'Display' pull-down menu simply select 'ASN.1' (which is a sequence format used internally at NCBI) and redisplay the entry. The sequence data are now unreadable, but near the

Table 1

Website	URL
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
CEPH-Genethon	http://www.cephb.fr/ceph-genethon-map.html
electronic-PCR (e-PCR)	http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi
Ensembl	http://www.ensembl.org/
Entrez	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Nucleotide
Entrez entry AP001002	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=8117673&dopt=GenBank
Entrez entry AC007104	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=5523795&dopt=GenBank
Entrez <i>Homo sapiens</i> genome view	http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch?chr=hum_chr.inf&query
EuroGeneIndexes	http://corba.ebi.ac.uk/EST/egi.html
FPC	http://www.sanger.ac.uk/Software/fpc
GenBank	http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html
GeneMap'99	http://www.ncbi.nlm.nih.gov/genemap99/page.cgi?F=Home.html
Genome Database (GDB)	http://www.gdb.org/
The Genome Channel	http://compbio.ornl.gov/tools/channel/index.html
Human Accession Map	http://genome.wustl.edu:8021/pub/gsc1/fpc_files/freeze_2000_06_15/MAP/
Human BAC Ends	http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_intro.html
Human Gene Index	http://www.tigr.org/tdb/hgi/index.html
Human Genome BAC map	http://genome.wustl.edu/gsc/human/Mapping/
NIX	http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/
RepeatMasker	http://www.genome.washington.edu/UWGC/analysistools/repeatmask.htm
Sputnik	http://www.abajian.com/sputnik/
STACK	http://www.sanbi.ac.za/Dbases.html
TIGR Gene Indices	http://www.tigr.org/tdb/tgi.shtml
TNG4 radiation hybrid map	http://www-shgc.stanford.edu/Mapping/Marker/RHTNG4index.html
UniGene	http://www.ncbi.nlm.nih.gov/UniGene/index.html
Working Draft Sequence	http://genome.ucsc.edu

top of the file is the clone name '301J10'. The same information is retrievable from the 'XML' and 'Graphics' display formats, but not under the default GenBank format.

A related site, the [Human BAC Ends](#) site at The Institute for Genomic Research (TIGR), provides access to more than 743,000 end sequences from 470,000 BAC clones. (Typically, end sequences consist of several hundred base pairs from the clone ends.) It is possible to search the sequences with either a clone name or a sequence of interest. As the unfinished BAC sequences in GenBank do not always contain the sequences from the BAC ends, the BAC end sequences may provide extra sequence data for a clone of interest. In addition, the end sequences can help to identify the fragments of unfinished BAC sequences that represent the ends of the clone. One caveat is that, as usual, the annotation of these sequences should be treated with a certain

degree of caution, because clone ends have been known to be attributed to the wrong clone [2]. Conveniently, the BAC end sequences at TIGR are provided with any repetitive subsequences masked (they are replaced with runs of the letter X). Repetitive sequences are a recurring problem in dealing with genomic sequence, particularly interspersed repeats (regions of very similar sequence descended from various classes of transposable elements) [3]. Interspersed repeats often span hundreds or thousands of bases and so can appear as spurious overlaps between genomic sequence fragments. The excellent program [RepeatMasker](#) does a good job of masking both interspersed and simple repeats. Simple repeats are stretches of sequence made up of units consisting of one or more bases, which may be repeated hundreds of times. They can be used as genetic markers, for example, in disease association studies, so finding them and annotating them properly is an important task. The [Sputnik](#) program

provides a fast and elegant method for annotating simple repeats, giving each repeat's location, classification (on the basis of repeat unit length - dinucleotides, trinucleotides and so on) and sequence.

Ideally, it would be desirable to retrieve the genomic sequence of a region of interest defined by the user, rather than multiple segments restricted to the size of BAC clones. A heroic, preliminary assembly of the draft genome sequence is available on the [Working Draft Sequence](#) site from David Haussler's group at the University of California, Santa Cruz (UCSC). Although this assembly contains over 200,000 gaps as well as some misassemblies and incorrectly ordered sequences, as it is updated with more sequence data it will become an important resource. The information is incorporated into the [Entrez Homo sapiens genome view](#) at NCBI, which is a graphical viewer designed to integrate sequence data with mapping information from various sources. Again, this NCBI interface will be a potent tool when more sequence data are available; it is already the best integration of data for finished chromosomes such as 21 and 22.

Expressed sequence data

Before the flood of genomic sequence from the Human Genome Project, full sequences were available for only a small proportion of human genes. Most human genes were represented only by expressed sequence tags (ESTs; fragments of mRNA sequences). Various efforts have been made to cluster overlapping EST sequences to give a longer representative sequence for each gene [4]. The most comprehensive of these efforts is the human [UniGene](#) database at the NCBI, in which ESTs and mRNAs from [GenBank](#) that share overlapping subsequences have been grouped together into clusters. [UniGene](#) can be searched either with [UniGene](#) cluster accession numbers or with [GenBank](#) sequence accession numbers for ESTs or mRNAs. Clusters are linked to related mapping, sequence and expression data at the NCBI, and each cluster should represent a separate gene. As [UniGene](#) is automatically and regularly generated, it often contains errors. One serious problem is chimeric clusters, produced as a consequence of sequencing from chimeric clones (artifactual cDNAs that contain sequences from two different genes). TIGR also maintains a clustered EST database, called the [Human Gene Index](#) (HGI), which has more stringent clustering criteria than [UniGene](#). Another human expressed sequence database, called [STACK](#), is held at the South African National Bioinformatics Institute (SANBI). In [STACK](#), expressed sequences are separated with respect to tissue of origin before clustering, and an attempt is made to represent differently spliced transcripts of the same gene. Unlike [UniGene](#), the [STACK](#), [HGI](#) and [EuroGeneIndexes](#) sites produce consensus sequences for clusters. Transcribed sequence databases are also available for species other than human; [UniGene](#) holds data for mouse, rat and zebrafish

and there are [TIGR Gene Indices](#) for various other species. The [EuroGeneIndexes](#) at the European Bioinformatics Institute (EBI) also contain expressed sequence clusters for a number of non-human species. It is worth remembering that all expressed sequence databases will contain repetitive sequences because much of the sequence is from untranslated regions of genes.

Mapping data

The fragmentary human genome sequence is of little use without some idea of how the pieces fit together, so a map is needed that relates distinct landmarks, or sequence tagged sites (STSs), around the genome. Different types of mapping data provide maps of different resolutions. Genetic maps, based on the frequency of recombination events between STSs, are of relatively poor resolution - on the order of hundreds, or more often thousands, of kilobases. Physical mapping techniques can resolve STSs only tens of kilobases apart. In the early stages of the Human Genome Project, an important task was to construct a high-resolution genetic map of the genome, and the [Genome Database](#) (GDB) was set up to curate such data. Genetic mapping data allowed genomic regions to be broadly defined, and efforts proceeded to physical mapping for finer distinctions. Various physical mapping projects have confirmed the physical order of genetic maps and extended genome maps to include further STSs and transcribed sequences. A physical map of the genome based on overlapping YAC (yeast artificial chromosomes) contigs was among the first to be published and the data are available from [CEPH-Genethon](#). One of the most important physical mapping techniques to emerge has been radiation hybrid (RH) mapping [5]. RH maps are orderings of STSs based on assay scores of the STSs against a whole genome radiation hybrid 'panel'. Such panels consist of hybrid cell lines that contain different fragments of human genomic DNA. Each STS is assayed against each cell line to discover whether it is present in the genomic fragments particular to that cell line. The pattern of presence or absence in the cell lines making up a panel constitutes the retention pattern of the STS, and, by comparing STS retention patterns, the distance between STSs can be estimated. In this way, the [TNG4 radiation hybrid map](#) was generated at the Stanford Human Genome Centre (SHGC) and provides an average of 60 kb resolution across the genome. Such impressive estimates of resolution must be tempered, however, by the ambiguity that often accompanies RH-derived marker ordering. Comparisons of STS orders in sequenced regions of the genome with orders derived from RH maps suggest that RH map orders may be wrong up to 50% of the time [6]. A consortium of RH mapping centres has produced a transcript map of the genome based on RH mapping data, named [GeneMap'99](#), which is accessible at the NCBI. The STS content of a sequence of interest can be determined online using the [electronic-PCR](#) (e-PCR) program at NCBI. This is a rapid sequence-search algorithm

that searches your sequence for occurrences of the STS sequences in [GenBank](#).

An important new source of mapping data has become available with the release of the draft genome: the fingerprint analysis of BAC clones for the genome project at Washington University Genome Sequencing Centre (WUGSC). The [human genome BAC map](#) provides the highest resolution human mapping data yet made available and is likely to do so until publication of the full human genome. The overlaps between clones are calculated using the program [FPC](#) on the basis of clone restriction fragment patterns or fingerprints. The resulting contigs are estimated to cover 97-98% of the genome. The fingerprint analysis has also been extended to show the sequence accession numbers for those clones that have been sequenced forming a [Human Accession Map](#).

Genomic sequence annotation

Once a region of the genome has been sequenced, the immediate concern is to identify the genes, if any, that are present. Broadly speaking, the computational annotation of genomic sequence proceeds by two methods: *ab initio* gene prediction, and detection of similarity. Strictly defined, *ab initio* prediction of genes relies on the presence of compositional biases in genomic sequence that are characteristic of exons. Similarity to known transcribed or protein sequences can be used as further evidence of the accuracy of an *ab initio* prediction. Many gene prediction programs combine these types of evidence and show considerable success in detecting genes [7,8]. Computational predictions must be treated with caution, however, before they have been confirmed at the bench.

The [Ensembl](#) database aims to provide a basic level of computational annotation for the draft genome. It localises BAC sequences in the genome according to a combination of mapping data and runs the sequences themselves through an 'analysis pipeline'. This pipeline consists of repeat masking the sequence, processing it with a gene prediction program called Genscan and then searching the predicted genes against sequence databases. Predicted genes that match known genes become [Ensembl](#) genes and are stored in the searchable Ensembl database. [The Genome Channel](#) is an analogous pipeline system that gives more detailed annotation, including CpG islands (areas of DNA that have a relatively high cytosine and guanine content), poly-adenylation sites and gene predictions from more than one gene prediction program. With rather more effort, it is possible to get very detailed annotation for a genomic sequence of interest through the [NIX](#) interface at the Human Genome Mapping Project Resource Centre (HGMP). Sequences submitted to [NIX](#) are processed by a variety of programs that detect repetitive regions, exons, tRNA genes, promoters, CpG islands, poly-adenylation sites and similarity to known proteins or transcribed sequences. The NIX interface is only available to

registered HGMP users but it is possible for academic scientists to register without charge.

Putting the data to work

Although I am a computational biologist, most of my work involves collaboration with molecular biologists generating real data at the bench. I find that, after a hard day in their labs, people very rarely ask me to discuss the available draft genome resources. Their problems are invariably specific to a small number of genes or genomic regions. Where in the genome is gene X? What is in genomic region Y? These are the commonest questions, and I suggest generic approaches to answering them below. Even in the best-case scenario, however, where gene X is well characterized and already mapped, there will often be additional information to be extracted. What are the neighbouring genes and what is their relative order and orientation? What non-coding features (regulatory elements, pseudogenes and repetitive regions) lie in the vicinity?

Where in the genome is gene X?

As the draft sequence is estimated to cover more than 90% of the genome, the chances of finding part or all of gene X in unfinished BAC sequence are high. If the available sequence of gene X contains any non-coding DNA, it should first be masked using [RepeatMasker](#). A [BLAST](#) search of the sequence of gene X against the section of the database that contains the draft sequence is all that is necessary to find the relevant BACs. Using the NCBI Advanced [BLAST](#) site it is possible to limit the search to human draft genome sequence by selecting the 'htgs' database and 'Homo sapiens' in 'Advanced options' (Figure 1). Assuming the sequence quality is good for gene X, the [BLAST](#) output should show at least one segment of BAC sequence that is almost identical (greater than or equal to 98% identical is a reasonable rule of thumb) over a reasonable stretch of gene X. In the absence of any good match to a BAC sequence, the best option is to [BLAST](#) search gene X against human EST sequences (the 'human ests' database at NCBI) and search [UniGene](#) with matching EST accession numbers, because many [UniGene](#) clusters contain mapped ESTs. If gene X is found within a BAC sequence, the BAC should be repeat masked and submitted to [e-PCR](#) at NCBI, which will often provide one or more STSs. These STSs may be localized to a genetic or RH map using their accession numbers to search either the [GDB](#) or the Stanford RH maps. If the BAC that contains gene X does not contain any STSs, it can be used, after masking repeats, to search the htgs database again to discover overlapping BACs. Again, the intention is to find identical sequences, allowing for sequencing errors, and a reasonable rule-of-thumb measure of 'identical' is a stretch of greater than 1 kb showing greater than or equal to 98% identity in the BLAST output. Overlapping BACs may be annotated as coming from the same chromosome as gene X, or the first BAC and can be submitted to [e-PCR](#) and assigned a location. Supporting

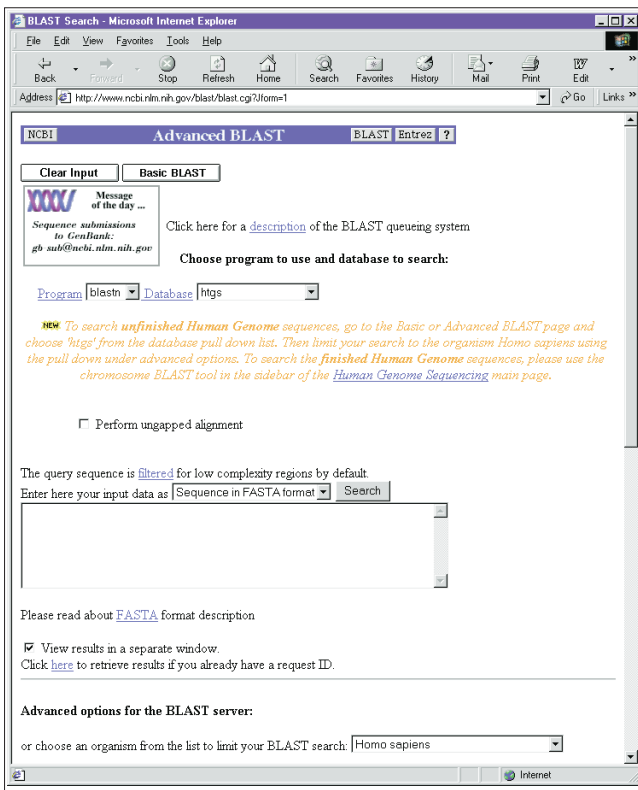
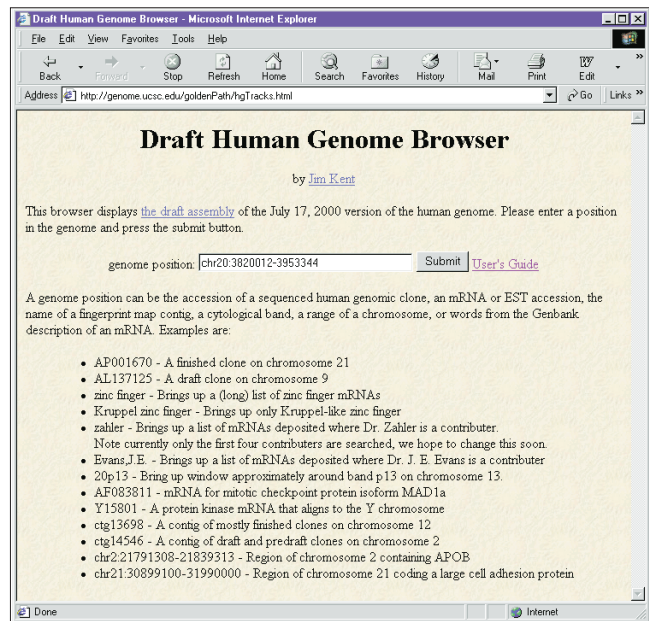


Figure 1
Advanced BLAST. The htgs database can be selected (near the top of the page) and 'Homo sapiens' can be selected in the advanced options (near the bottom).

evidence for a collection of overlapping BACs can be obtained from the [Human Accession Map](#) and the [Working Draft Sequence](#) site at UCSC (Figure 2). If these two resources include the BACs but do not show them to overlap, one should be suspicious.

What is in genomic region Y?

Determining what is in genomic region Y involves a process similar to mapping gene X. The more sequence we start with from region Y the better. In the worst case, we might only have one STS sequence that is known to be from region Y but is not part of any known transcript. As in the section 'Where in the genome is gene X', after repeat-masking the STS we can use it to search the htgs database using [NCBI Advanced BLAST](#) for matching genomic sequence. This first stage is the most troublesome; as STSs are only a few hundred bases long it is desirable to have some corroborating evidence to back up any apparent match to a BAC. This can come in the form of mapping data. Other markers or genes near the STS on genetic or RH maps would be expected to appear in the sequence of the apparently matching BAC or in BACs that overlap with it. Once we have reliably placed the starting STS sequence in a BAC, the task is to build up a contig of overlapping BACs around it, as in the section 'Where in the genome



is gene X'. The [Human Accession Map](#) at WUGSC and [Working Draft Sequence](#) site at UCSC can be used as guides to choosing overlapping BACs that extend your contig furthest. Many BAC sequences, particularly those in earlier stages of sequencing, contain cloning vector sequences that can generate spurious [BLAST](#) matches. It is possible to use [RepeatMasker](#) to also mask vector sequences but this is not an option offered on the [RepeatMasker](#) web server. If a BAC sequence generates a large number of [BLAST](#) matches then the sequence should be searched against the entire sequence database ('nr' at NCBI [BLAST](#)) to look for the presence of bacterial sequence. It is important to remember that the word contig is actually rather inappropriate here, because most BAC sequences are fragmented and incomplete. Generally BAC sequences are 150-200 kb long when complete, so it is possible to estimate roughly the amount of missing sequence. This process should eventually result in a list of BAC accession numbers that represents most of the sequence in region Y. The accession numbers that result can be used to search [Ensembl](#) to give the minimum number and identities if genes in region Y. More detailed analysis, including the identification of non-coding sequence features, can be carried out using [NIX](#). This strategy is probably only practical for investigating modestly sized regions of, say, less than 1 Mb; for larger regions, for example, a chromosomal band, it is easier to approach the process as an automated task - ask your friendly local bioinformaticist.

The web resources described here are those I find most useful, and are a 'snapshot' as of September 2000. They will, of course, be subject to change or updates as more information

becomes available. Other people will have their own opinions and methods on how to use web resources for accessing the draft human genome - there is no substitute for experience.

References

1. Roach JC, Siegel AF, van den Engh G, Trask B, Hood L: **Gaps in the human genome project.** *Nature* 1999, **401**:843-845.
2. Zhao S, Malek J, Mahairas G, Fu L, Nierman W, Venter JG, Adams MD: **Human BAC ends quality assessment and sequence analyses.** *Genomics* 2000, **63**:321-332.
3. Smit AF: **Interspersed repeats and other mementos of transposable elements in mammalian genomes.** *Curr Opin Genet Dev* 1999, **9**:657-663.
4. Bouck J, Yu W, Gibbs R, Worley K: **Comparison of gene indexing databases.** *Trends Genet* 1999, **15**:159-162.
5. **Radiation hybrid mapping information.** [<http://compgen.rutgers.edu/rhmap/>]
6. Agarwala R, Applegate DL, Maglott D, Schuler GD, Schaffer AA: **A fast and scalable radiation hybrid map construction and integration strategy.** *Genome Res* 2000, **10**:350-364.
7. Burge CB, Karlin S: **Finding the genes in genomic DNA.** *Curr Opin Struct Biol* 1998, **8**:346-354.
8. Semple C: **Gene prediction: the end of the beginning.** *Genome-Biology* 2000, **1**:reports4012.1-4012.3.