



WebWise: Guide to The Institute for Genomic Research Web Site

Kim D. Pruitt

Genome Res. 1998 8: 1000-1004

Access the most recent version at doi:[10.1101/gr.8.10.1000](https://doi.org/10.1101/gr.8.10.1000)

References This article cites 1 articles, 1 of which can be accessed free at:
<http://genome.cshlp.org/content/8/10/1000.full.html#ref-list-1>

License

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

An advertisement banner with a teal background. On the left, the text reads "CRISPR and RNAi Genetic Screening. Your new superpower." In the center, there is a white box with the words "LEARN MORE" inside. On the right, there is a photograph of a woman wearing a red superhero mask and a red cape, and the Cellecta logo, which consists of a green molecular structure and the word "CELLECTA" below it.

To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Cold Spring Harbor Laboratory Press

WebWise: Guide to The Institute for Genomic Research Web Site

Kim D. Pruitt¹

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894 USA

This installment of the WebWise series reviews the final web site included in the scope of this series, namely The Institute for Genomic Research (TIGR; <http://www.tigr.org/>). TIGR, a nonprofit private research institute that was headed by Craig Venter until recently and is now headed by Claire Fraser, has made a significant contribution toward sequencing several genomes. This review focuses on the data, tools, and other resources currently available on this web site that pertain to the Human Genome Project (HGP). The human genome sequencing effort at TIGR has been impacted by some recent developments. Namely, the recent formation of Celera (<http://www.celera.com/>), a private-sector high-throughput shotgun sequencing facility headed by Craig Venter, shifts the human genome sequencing effort from TIGR to Celera. Furthermore, as reported in the News section of *Science*, TIGR was not among the U.S. HGP groups selected for continued funding by the National Institutes of Health (Pennisi 1998). Hence, the TIGR human genome sequencing effort will start to wind down; clone sequencing already in progress will be completed, but sequencing currently in early phases of the production pipeline will likely not be completed (M. Adams, pers. comm.). Some other HGP-related projects at TIGR, such as the BAC-end sequencing effort, retain Department of Energy (DOE) funding and will continue. Furthermore, the data, tools, and resources currently available on TIGR's web site will be maintained for the foreseeable future.

The site map (Fig. 1) is intended to provide a simple road map to data and tools pages. Therefore, it does not in-

clude all of the links provided on the TIGR site map; redundant links, links to other genomes, as well as some groupings of related links are depicted as a single link or are omitted from the site map. The main features of this web site, as well as the features of those sites already reviewed, are indicated in Table 1. As with the previous WebWise reviews, you may find it useful to point your Browser window to the TIGR web site while reading.

General Information

TIGR has made the necessary investment to design a polished, well-organized web site. The Home page (<http://www.tigr.org>) presents a set of stylized graphics that are linked to different sections, organized by category, of the web site. These links are then reiterated in plain text form directly below the graphic section. Unfortunately, the graphics are rather slow to download the first time you call up the Home page, but once they are cached in your computer memory the graphics download quickly and navigation is convenient. One nice feature of the graphics area is the underlying javascript incorporated into the code; brief text messages (describing the section linked to) are displayed in a central area as you move the mouse over the different sections of the graphic. The web site can be thought of as a series of concentric circles—the central point is the Home page with links pointing out to the next level, and so on. In general, navigation is easy as the first layers of the web site include a consistent set of navigation links in the side column. As you navigate farther away from the Home page, the side column navigation links are replaced with a link to the Home page and to the top page of the general section at hand. The top level section pages do include the entire set of

navigation links. Overall, the web site presents an agreeable “look and feel,” as the navigation links are consistently provided in a similar color and design style throughout the web site.

The Home page links that lead to information of more general interest are indicated in the top portion of the site map in Figure 1. Follow the About TIGR Home page link to reach some general information about the facility itself. This page includes links to pages presenting travel directions, staff publications, company officers and trustees, and a slide show of the TIGR facilities (exterior views only). Announcements of data releases are included on the What's New page. The Conferences, Education, and Training page includes sections on upcoming conferences sponsored by TIGR, a distinguished lecture series, education outreach, and training workshops. As expected, the Career Opportunities page presents a list of job postings, and the Other Links page includes useful links organized by general category, including news, genome centers, databases, and professional societies, to name a few. This page may be useful to some people, and its organization is quite effective. Links to each category are provided at the top of the page, or one can scroll down the page to review the whole collection of links.

Data

Follow the TIGR Database Home page link to access the human sequence data. In addition to the HGP sequencing effort, TIGR also has ongoing human cDNA transcript-mapping and BAC-end sequencing projects (follow the similarly named links on the TIGR Database page). Search interfaces are provided to search for the BAC-end sequence data by clone name or by sequence homology

¹E-MAIL pruitt@ncbi.nlm.nih.gov; FAX (301) 435-2433.

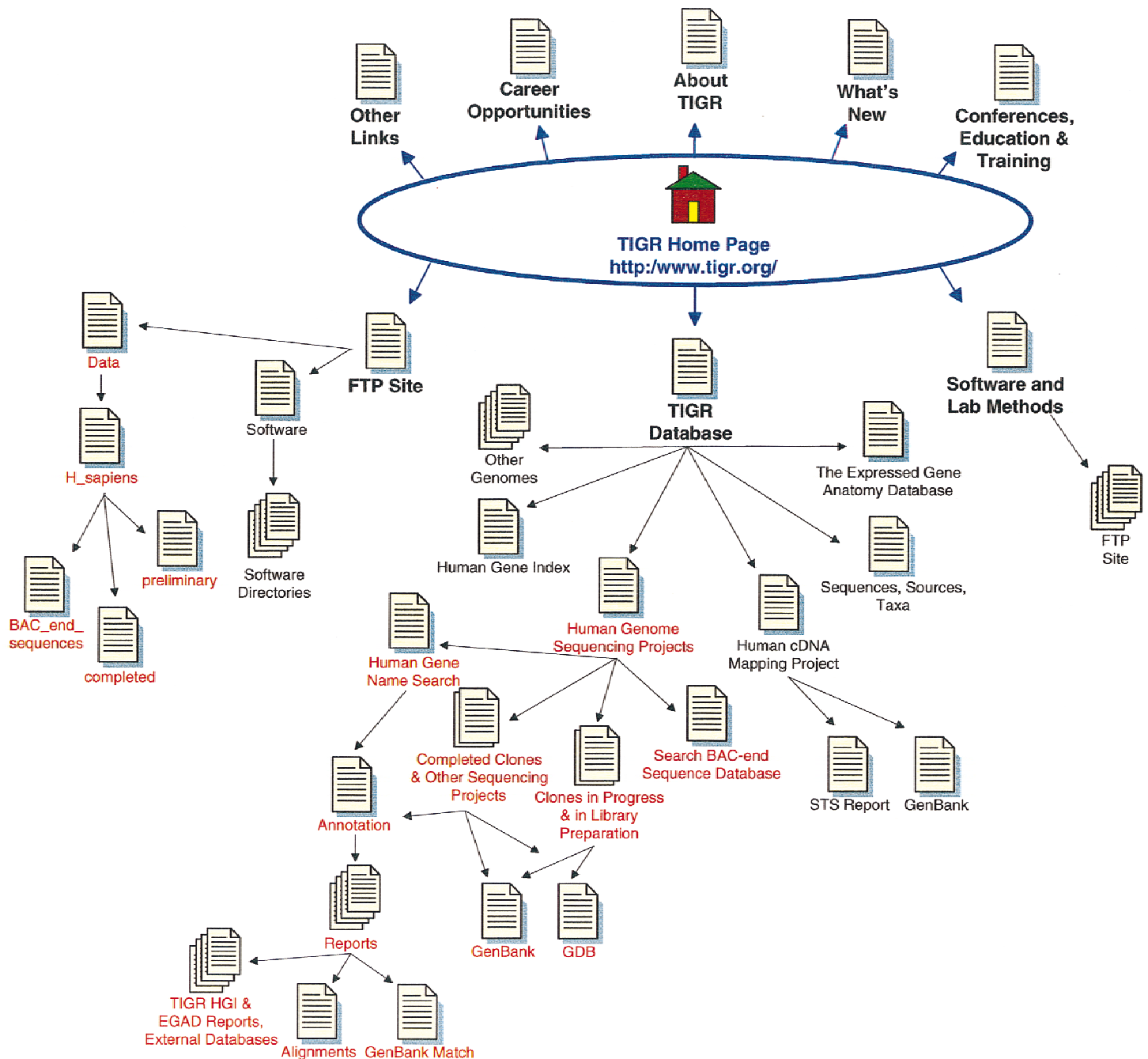


Figure 1 Site map for TIGR. The main links to pages discussed in the text are illustrated here. Links placed above the Home page icon are to general informational resources, whereas links located in the bottom portion are to the data and software tools. Some links available on the web site are not indicated here. Red text highlights those web pages presenting HGP sequencing and/or mapping data.

(http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_search.html). Although the homology search form was not tested for this review, the clone name search was tested using the value suggested on the web site—this search returns the BAC-end sequence data. The Human cDNA Mapping Project page (<http://www.tigr.org/tdb/hummap/hummap.html>) presents data on se-

quence-tagged sites (STSs) that match human transcripts. Several search forms are provided to search by STS number, gene ID number (either GenBank accession no. or internal TIGR ID), or chromosome. The first two search methods bring up an STS Report or a transcript report identifying STSs mapped to a given transcript, respectively. The STS Report details the protocol, sequence

data, and includes links to other synonymous sequences. The transcript report, which was tested using the default value provided in the search form, returns a page identifying TIGR's tentative human consensus (THC; see Tools, below) sequences and provides links to those pages as well as to the STS Report, GenBank, and the European Bioinformatics Institute (this link is outdated).

Insight/Outlook

Table 1. Features of the TIGR Web Site

Center		GSC	SC	SHGC	W/MIT	BCM	JENA	UTSW	UWGC	CGM	ACGT	JGI	TIGR	
Map Data	Static Map	●				●		●	●	●		●		
	Image Mapped	●	●	●			●	●			●			
	Tabular List			●	●	●								
	Clones Linked to Sequence		●	●	●	●		●			●			
Sequence Data	Download data from FTP or Web Site	●	●	●	●	●	●	●	●		●	●	●	
	Download a Database		●											
	Links to Public Databases		●		●	●	●		●	●	●	●	●	
	Update Frequency	Daily	●	●	●	●	●	●	●					●
		Weekly												
		Unknown*								●	●	●	●	
	Sequence Annotation	Graphic						●	●		●	●		●
Text		●					●	●		●				
Not Available		●	●	●	●	●			●			●		
Search Services	Similarity Searches	Performance	a	○	○	○	○	○	○	○	○	○	○	○
			b	○	○	○	○	○	○	○	○	○	○	○
			c	○	○	○	○	○	○	○	○	○	○	○
		d	●	●	○	○	○	○	○	○	○	○	○	○
		e	○	○	○	○	○	○	○	○	○	○	○	○
		f	○	○	○	○	○	○	○	○	○	○	○	○
	Quality of Output	a	○	○	○	○	○	○	○	○	○	○	○	○
		b	●	○	○	○	○	○	○	○	○	○	○	○
		c	○	○	○	○	○	○	○	○	○	○	○	○
	d	○	○	○	○	○	○	○	○	○	○	○	○	
e	○	○	○	○	○	○	○	○	○	○	○	○		
f	○	○	○	○	○	○	○	○	○	○	○	○		
Not Available			●	●	●			●	●	●	●	●		
Search the Maps		●					●							
Search for Sequences	●	●			●		●					●		
Search the Web Site	●		●	●	●		●			●		●		
Software	Documentation	a	○	○	○	○	○	○	○	○	○	○	○	
		b	●	○	○	○	○	○	○	○	○	○	○	
		c	○	○	○	○	○	○	○	○	○	○	○	
		d	○	○	○	○	○	○	○	○	○	○	○	
		e	○	○	○	○	○	○	○	○	○	○	○	
		f	○	○	○	○	○	○	○	○	○	○	○	
	Available from:	FTP Site	a	○	○	○	○	○	○	○	○	○	○	○
			b	○	○	○	○	○	○	○	○	○	○	○
			c	○	○	○	○	○	○	○	○	○	○	○
			d	○	○	○	○	○	○	○	○	○	○	○
e	○	○	○	○	○	○	○	○	○	○	○	○		
f	○	○	○	○	○	○	○	○	○	○	○	○		
Web Page Link	a	○	○	○	○	○	○	○	○	○	○	○	○	
	b	○	○	○	○	○	○	○	○	○	○	○	○	
	c	○	○	○	○	○	○	○	○	○	○	○	○	
	d	○	○	○	○	○	○	○	○	○	○	○	○	
e	○	○	○	○	○	○	○	○	○	○	○	○		
f	○	○	○	○	○	○	○	○	○	○	○	○		
Contact the Site				●	●			●	●	●	●	●		

The red circles indicate features that are available at this web site or the quality of a given feature within a general range of better (a) to worse (f). Sequence data are assessed for availability from an FTP site, availability in a database (such as ACeDB), whether archived sequences are linked directly to the public database records, the frequency of FTP site update, and whether any sequence annotation is provided in either a text or graphic format. Each web site is scored for the availability of various search services, including the ability to carry out similarity searches against the sequences in their database or perform a key word search of the map data, sequence data, or web site. Documentation and availability of software tools developed by the center is also indicated. (GSC) Washington University's Genome Sequencing Center; (SC) The Sanger Center; (SHGC) Stanford Human Genome Center; (W/MIT) Whitehead/MIT Genome Center; (BCM) Baylor College of Medicine Human Genome Center; (JENA) The Institute of Molecular Biology Genome sequencing Center Jena; (UTSW) McDermott Center for Human Growth and Development at University of Texas Southwest; (UWGC) University of Washington Genome Center; (CGM) Washington University Center for Genetics in Medicine; (ACGT) University of Oklahoma Advanced Center for Genome Technology; (JGI) Joint Genome Institute; (TIGR) The Institute for Genomic Research.

*Although the genome sequencing centers have agreed to update sequence data on a daily basis, the update frequency is not explicitly indicated on the web site.

The chromosome search returns a large list of links to TIGR THC reports.

The bulk of TIGR's human sequence data is accessed via the Human Genome Sequencing Projects page (<http://www.tigr.org/tdb/humgen/humgen.html>); TIGR, in collaboration with Caltech and the Los Alamos National Laboratory, has focused on mapping and sequencing the short arm of chro-

mosome 16. A project status table provides an overview of the amount of data completed, in a finishing stage, or in an early preparation stage. TIGR has generated >11 Mb of (redundant) finished sequence with another ~6.5 Mb in the pipeline.

The TIGR web site presents information about the sequence data; detailed map data are not included on this web

site and are instead available on their collaborators' web site. Although this is a good reflection of the nature of the collaboration, it has the unfortunate effect of dissociating the map and sequence data. These data are most useful when they are presented in a more integrated format. Although the TIGR web site does provide general location information (in terms of cytogenetic bands),

this information does not suffice if one wants to determine if the sequenced clones represent a contiguous stretch of sequence data, or if significant gaps still fall between the clones. Those interested in reviewing the extent of sequence data available for a given region must then look at the map data on the collaborators' web site and hope that they can ascertain what the correspondence is to the clones named on the TIGR web site.

Sequence data pertaining to clones in different stages of completion are available by following the links provided below the project status table. The data pages (links are termed Completed Clones and Other Sequencing Projects, Clones in Progress and in Library Preparation), as well as the subsequent pages, include a search form whereby one can search for a clone by name. This is a very useful feature to include if one is interested in monitoring the sequence progression of a particular clone of interest. The search form returns an annotation information page (see below) on the submitted clone. This is a more rapid approach, in contrast to tediously scrolling through tables of data, to determine if sequencing is finished for a given clone with the added benefit of being able to quickly ascertain if any potential ORFs occur. Although an Annotation report page is presented for unfinished clones, the annotation information is understandably only available once sequencing is finished (annotation pages described in more detail below). Hence, the clone search tool would be even more useful if links to the most up-to-date sequence data (on TIGR's web or FTP site) were also included on these pages. Indeed, the unfinished human sequence data does not appear to be currently available on TIGR's web site other than via links to GenBank records. For those interested in browsing the larger data set, follow the four links that lead to clone data (e.g., Completed Clones, etc.); data are displayed in tabular summaries of the sequencing effort as organized by status or project. The Other Sequencing Projects page (<http://www.tigr.org/tdb/humgen/other.html>) presents data available for a handful of clones located on other chromosomes (1, 2, 5, 8, 9, 12, 17, and 22).

As all of the sequence data pages are quite similar in organization—with the only difference being the understandable omission of links to annotation

data for clones still in progress—only one data page and its associated links is described here. The Completed Clones link calls up a page listing all human chromosome 16 clones in which sequencing has been completed to date (<http://www.tigr.org/tdb/humgen/completed.html>). This information is presented in tabular format and includes the Clone name, External DB (database), Clone size, Map position, and GenBank accession number columns. The data provided in the External DB and GenBank columns are hotlinked to the Genome Database (GDB) and GenBank records, respectively. Clone names are hotlinked to annotation pages that summarize the results of BLAST identity searches and ORF prediction programs.

The Annotation pages (e.g., see http://www.tigr.org/tigr-scripts/bac_scripts/bac_display.spl?bac_name=A-152E5) depict both a simple graphic illustrating the relative location and orientation of putative genes and a table indicating the gene names, if known. The gene names provided are based on the best match to sequences in the public databases. The genes depicted in the graphic and table are linked to the Reports page (e.g., see http://www.tigr.org/tigr-scripts/bac_scripts/bac_gene_display.spl?db=hug&gene_id=137&feat_type=mRNA_join). These reports present a more detailed graphic of the genomic organization as well as links to the best database match and to an alignment. The database match link, located in the top portion of the Reports page, is not formatted correctly; it points to the GenBank nucleotide database but provides a protein database ID number. The GenBank error page provides the correctly formatted link to the protein record, from which point you can also access the nucleotide record. The Percent Similarity link calls up an alignment between the TIGR sequence and the best match.

The genomic organization graphic presents the predicted organization of the TIGR sequence at the top, and additional views of the gene organization are shown below that. These additional organizations are derived from both the public database repository and from other TIGR resources that provide interpreted views of the public data. Each item included in the graphic is linked to the informational resource from which it was derived. Thus, a graphic may in-

clude links to the GenBank protein and nucleotide repositories as well as to TIGR resources such as the Expressed Gene Anatomy Database (EGAD) and the Human Gene Index (HGI). Links are provided to both HGI EST reports and to HGI THC reports (THCs represent assemblies of EST data). It is easy to determine where a given link leads to, as most of the text begins with the initial abbreviation of the resource (GP, GenBank protein; GB, GenBank nucleotide; EGAD and THC, TIGR resources). The only exception is that links to TIGR's HGI EST reports appear as a simple alphanumeric accession number that is not preceded with HGI.

One worthwhile alternative view of the human sequence data is available on the Human Gene Name Search page (http://www.tigr.org/tigr-scripts/bac_scripts/gene_name_search.spl?db=hug&name=all). This page presents all of the genes and putative genes identified to date by the TIGR chromosome 16 sequencing project. The initial page presents a list of all genes, and a search window is provided toward the top of the page whereby one can search for a gene of interest. The gene names are linked to an Annotation page, from which you can access additional data. Note that the gene identities are based on the best database match, and such identities cannot always be taken at face value. Nevertheless, if you suspect that a particular gene may be located on the short arm of chromosome 16, this set of pages provides a quick and easy interface to begin looking for said gene.

Tools

The TIGR web site offers a host of downloadable software tools as well as several resources presenting interpretive views of the raw sequence data. In addition it offers two tools with which to search the web site. The first search tool can be reached via the link at the bottom of the Home page, as well as on several other internal pages, and can be used to look for a general topic. This page, entitled Search the TIGR Web Site, presents a link to help on searching, in addition to the easy-to-use search form (<http://www.tigr.org/search/index.html>). The returned result page includes links to, and short descriptions of, pages matching the query. As mentioned in the Data section, a sequence data search form is

Insight/Outlook

also provided on the Human Genome Sequencing Project pages (search by clone name). This search engine is sensitive to format so you must type in clone names as they are presented on the TIGR web site or you may get a false-negative result. TIGR does have an easily accessed FTP directory (<ftp://ftp.tigr.org/pub/>; human sequence data are in the [data/h_sapiens](ftp://ftp.tigr.org/pub/data/h_sapiens/) subdirectory). Although the completed subdirectory does contain several files, surprisingly the preliminary sequence directory was completely empty at the time of this writing (ftp://ftp.tigr.org/pub/data/h_sapiens/preliminary/). Those preliminary sequences that have been submitted to GenBank can be retrieved through that resource, and the Clones in Progress status table includes some links to GenBank that should facilitate retrieval. Besides this omission, the main service omitted from TIGR's platter of resources is a human chromosome 16 BLAST server, so it is necessary to direct all human identity searches to a public database repository site.

The Software and Lab Methods page (<http://www.tigr.org/softlab/>) includes brief descriptions of several software tools, with links to either the appropriate FTP directory or, for some, to an additional information page. A handful of links to protocol descriptions are available at the bottom of this page. The software tools are freely available for non-profit, noncommercial use by the scientific community. These tools can be categorized into several types of tasks, including gene identification, identity searches, sequence assembly, sequence analysis, and annotation. Although extensive documentation or web-based demonstrations of these tools are not provided on the web site, references to published descriptions are included in some of the descriptions; tools were not downloaded and tested for this review. Further information on availability can be obtained by reading the copyright notice or contacting TIGR directly (e-mail contact links are provided on the bottom of most web pages).

The TIGR web site includes some value-added resources that utilize sequence data derived from GenBank as well as from TIGR's sequencing projects to provide an interpretive view of the underlying data. These resources, namely EGAD and HGI, represent efforts to present nonredundant views of tran-

scripts and genes, respectively. EGAD transcript sequences are derived from one to many sequence records found in the public databases. Transcript reports identify these sequences by accession number (linked to the appropriate public database record), and present the coding region and untranslated region (UTR) sizes in addition to the nucleotide and protein sequence data. Furthermore, some tissue distribution and cellular role information is available for many of the EGAD-identified transcripts. Information can be retrieved by searching for gene names or previously identified TIGR human transcript (HT) ID number, or by browsing through categorized cellular roles. The HGI resource is gene-based rather than transcript-based, but there is some overlap between these two resources, for instance, HGI reports of corresponding EGAD transcripts can be generated. In addition to functionally known genes, the HGI includes reports on putative new genes identified via computational assemblies of EST (expressed sequence tagged) sequences; the EST-based genes are termed THC sequences on TIGR's web pages. The HGI data set also includes expression information and information on ordering clones. Together, these two resources provide a considerable amount of information, but space and time considerations forestall providing a more expanded review of them at this time.

Conclusions

TIGR has committed the necessary resources to develop a polished, professional web site that is aesthetically pleasing and generally easy to navigate. Although the initial web browsing experience could be improved if the Home page graphics downloaded faster, more importantly, the TIGR web site presents a considerable amount of sequence data, software tools, and additional value-added resources of relevance to the HGP. For instance, the EGAD and HGI resources provide a good first step toward pulling together knowledge about sequence data and biological information such as expression and functional information. However, in light of these extra resources provided, it is rather surprising that the unfinished sequence data, a human sequence BLAST service, and contig map data are not provided on TIGR's web site. As TI-

GR's chromosome 16 sequencing effort is in the process of winding down, it is unlikely, in my opinion, that they will commit the resources at this time to add these features.

The TIGR web site provides a data-rich resource to the scientific community; it is reassuring to know that this service will not be impacted immediately by the fact that TIGR was not selected for continued NIH HGP funding. However, one can't help but speculate that as the Venter human sequencing torch passes from TIGR to Celera, some of TIGR's human genome resources may be transferred to the Celera web site (<http://www.celera.com>). Indeed, one anticipates that a new resource- and data-rich web site will become available to the scientific community in the not-too-distant future.

REFERENCES

- Pennisi, E. 1998. *Science* 281: 148-149.
- Next Month: The Final Installment—Overviews, Updates, and Conclusions