Initial analyses of the human genome: does it reveal anything new?

Rebecca Lawrence, News & Features Editor

Detailed searches of the current draft of the human genome sequence for new candidate genes and paralogues of disease genes have had varying successes[1–9]. Disappointingly, in several instances, these searches have revealed little, if any, new information that has not already been discovered using traditional techniques.

These studies have been published in the 12 February 2001 issue of *Nature* alongside a paper by the International Human Genome Sequencing Consortium on the initial sequencing and analysis of the human genome[10]. The nine data-mining papers cover a variety of topics: cancer[1], addiction[2], gene expression[3], immunology[4], evolutionary genomics[5], membrane trafficking[6], the cytoskeleton[7], the cell cycle[8] and the circadian clock[9].

The results of these attempts to mine the current draft of the human genome have greatly varied in their success to obtain useful information. For example, no paralogues of known tumour suppressor genes were found from a search for the proteins predicted from the draft sequence[1]. Similarly, few novel cyclins and no cyclin-dependent kinases (Cdks) were discovered[8]. By contrast, workers examining the molecular components involved in addiction were able to identify several new candidate genes[2] and other teams have found many insights into the evolutionary genomics of the human genome[5] and, more specifically, into the evolution of transport vesicles[9]. Similarly, a team examining the genomics of immunology found several novel proteins[4], and more than 2000 hypothetical genes that encode transcriptional activators were discovered[3]. The lack of success in finding new genes in some fields could either be because these new genes are currently in a fragmented state in the draft sequence or just because all the genes in these areas have already been found using traditional methods.

Caution required?

Many of the teams were cautious about the information they gained from the draft sequence. Although the searches located numerous genes previously identified by traditional methods (such as cytoskeletal genes[7], tumour suppressor genes[1], clock loci[9], cyclins and Cdks[8]), many other known markers[4] and genes[7] were not found. There are several possible reasons for this, such as the fact that many of these genes might be currently represented as fragments that have not yet been assembled into full-length coding sequences. Furthermore, there are currently many gaps in the gene annotation.

There was also disappointment that comparing sequenced genomes produced little information about the cell cycle in general and was unable to help explain how differences have evolved between the cell cycles of different organisms[8]. A study searching for clock genes identified new clock gene candidates but the team conducting this study commented on the lack of information regarding their expression patterns or function[9].

Other concerns came from a study that mapped pairs of sequences (derived from cDNA libraries constructed from normal and neoplastic tissue samples) from the same cDNA clone to the genome. As expected, most of these mapped to the same position in the genome, but a few pairs mapped to two different parts of the genome, suggesting a significant rate of false positives[1].

Several limitations to the draft sequence have also been highlighted by these preliminary searches[3]. For example, the existence of a related gene sequence does not automatically mean that there is a corresponding protein as the sequence could be a non-expressed pseudogene. Furthermore, it is often difficult to decipher whether two related genes are simultaneously expressed in one cell or differentially expressed. A further complicating factor is that many factors are components of multi-subunit complexes, making analysis difficult without associated biochemical studies. Moreover, relatively few clones have been sequenced from most of the libraries and many of the libraries are not normalized[1].

Consortium analysis more promising

A more global analysis of the draft human genome sequence by the Consortium itself has produced more promising results[10]. Their rapid search of the draft sequence produced 286 potential paralogues of disease genes, demonstrating a potential for the rapid identification of disease gene paralogues *in silico*. Furthermore, a search for paralogues of the classic drug target proteins in the draft sequence identified 18 putative novel paralogues, which included possible dopamine receptors, purinergic receptors and insulin-like growth receptors. These all either

1359-6446/01/$ – see front matter PII: S1359-6446(01)01715-9

matched at least one expressed sequence tag (EST) or contained long open reading frames (ORFs), and all showed homology spanning multiple exons separated by introns, and are therefore not pseudogenes. These could therefore represent interesting new candidate drug targets.

Predicted impact of the completed sequence

Although there are a number concerns over the impact of the draft human genome sequence as a resource to identify novel genes and paralogues of disease genes, there is more optimism about the benefits of the completed sequence, which from the latest information available, has been estimated to contain 31,000–32,000 genes[10]. Suggestions from the teams reported here include an increase in the knowledge of the number of clock genes in mammals and a significant impact on the study of circadian output systems[9], and identification of many addiction vulnerability genes[2]. Production of a reliable annotation of the genome and 'clean' databases of human genes and proteins should enable rigorous analysis of the genome to discover its evolution[5]. Meanwhile, it has been suggested that research into the genomics of immunology are most likely to be successful using 'forward genetics' that start from the altered immunological trait and identify the causative gene, rather than using 'reverse genetics' by systematically knocking out every immunologically expressed gene[4].

However, there are some that are less optimistic about the impact of the completed human genome sequence. They suggest that, although it should reveal useful targets for the development of new therapies and will be essential for the understanding of the cytoskeleton and associated molecular motors, it is unlikely to provide insight into molecular mechanisms due to the complexity of proteins[7]. Similarly, the diversity of the structure and function of cancer genes and the fact that many close relatives of known important cancer genes are not mutated in cancers means that searching for paralogues of these known genes is unlikely to be effective in identifying new candidate cancer genes[1]. Success in this area is likely to require an increase in the cancer genome sequence data available. Most agree that to get significantly more out of the human genome sequence, there needs to be a marked improvement in our ability to turn raw sequence data into biological knowledge[8].

There are even some concerns that having the full genome sequences could have a negative effect on research by delaying studies on completing the mechanistic analyses required to understand physiology[7].

Conclusions

The overall consensus seems to be that some initial information can be gained from the draft sequence in certain fields. However, much care needs to be exercised when interpreting this information due to incomplete and sometimes incorrect assembly of some raw genome data in the draft sequence. Furthermore, considerable work is still required to correctly identify all protein-coding segments and to remove intronic sequences. However, after completion of the human genome sequence, most groups agree that with the development of new technologies and strategies, it should provide a source of much more valuable information for the identification of novel targets and ultimately, for the development of new therapies.

References

1 Futreal, P.A. *et al.* (2001) Cancer and genomics. *Nature* 12 Feb
2 Nestler, E.J. and Landsman, D. (2001) Learning about addiction from the human draft genome. *Nature* 12 Feb
3 Tupler, R. *et al.* (2001) Gene expression and the human genome. *Nature* 12 Feb
4 Fahrer, A.M. *et al.* (2001) A genomic view of immunology. *Nature* 12 Feb
5 Li, W-H. *et al.* (2001) Evolutionary analyses of the human genome. *Nature* 12 Feb
6 Bock, J.B. *et al.* (2001) A genomic perspective on membrane compartment organization. *Nature* 12 Feb
7 Pollard, T.D. (2001) The human genome, the cytoskeleton and motility. *Nature* 12 Feb
8 Murray, A.W. and Marks, D. (2001) Can sequencing shed light on cycling? *Nature* 12 Feb
9 Clayton, J.D. *et al.* (2001) Keeping time with the human genome. *Nature* 12 Feb
10 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 12 Feb

What do YOU think the impact of the human genome sequence will be on drug discovery?

Have you had any experiences of data mining the draft sequence, successful or otherwise?

Do you have comparable experiences of data mining genome sequences from other species?

Please send your comments to Dr Rebecca Lawrence, News & Features Editor, *Drug Discovery Today*
e-mail: Rebecca.Lawrence@current-trends.com

Publication of letters is subject to editorial discretion.