The Merck Gene Index project

Alan R. Williamson

The Merck Gene Index project (MGIP) fills an important niche in the Human Genome Project by directly identifying genes through sequences of their transcripts and placing in the public domain a set of EST sequences and associated clones for the uniquely expressed human genes. The MGIP promotes the unrestricted exchange of human genomic data, and facilitates progress in biomedical research by reducing duplication of efforts, speeding the identification of disease-related genes and furthering our understanding of disease processes. The project is stimulating biological research, which in turn is the driving force in drug discovery today.

The Merck Gene Index project (MGIP) should be seen in the context of the Human Genome Project (HGP), which is probably the most important project ever conceived of and carried through in science. The HGP is the first 'big' biology project that has ever been undertaken; it is an international project shared among the major scientific nations and data produced are placed directly into freely accessible public databases.

The objective of the HGP is to produce a single reference sequence of the human genome and also to sequence the genomes of several model organisms. These sequences will contain the information to enable an understanding of the basis of life and, in particular, the basis of human life. Hence, the HGP is the most fundamental study of man ever undertaken. The public's expectations of this project are extremely high. Moreover, unlike many other scientific projects, where scientists are often criticized for not looking at the implications of the project beforehand, the HGP is continually being analysed in great detail with extensive speculation as to how it will impact humanity.

One result of the HGP will be the identification of each of the estimated 70,000–100,000 human genes and their locations on the genome map. This will eventually be achieved by the sequencing of the complete genome and its interpretation, although this is unlikely to be completed for several years[1]. An alternative and more immediate solution to determining the set of expressed human genes is presented by large-scale partial sequencing of random cDNA clones[2–5]. With this approach it should be possible, in principle, to identify all human genes through their expression as mRNA molecules. This technology, termed expressed sequence tag (EST) sequencing, is applied in two commercial efforts (Human Genome Sciences, Rockville, MD, USA and Incyte Pharmaceuticals, Palo Alto, CA, USA). The objective of these companies is to place a complete set of expressed human genes in private databases and to establish proprietary and intellectual property positions on the set of human genes. Access to these databases will be restricted to the companies' partners and collaborators, with no access for the majority of academic and industrial researchers.

Public demand

It became clear to us at Merck in mid-1993 that such private domain databases would cause much apprehension in most research labs. Our response was to devise and initiate the MGIP (Ref. 6) and thus stimulate a great increase in the rate of public sequencing of expressed sequences. The specific goals of the MGIP were to:

- Coordinate the development of a standard human expressed gene collection
- Develop an organized distribution system for all of the clones
- Characterize those clones through partial sequencing of both the 5′ and 3′ end

Alan R. Williamson (former Vice President of Research Strategy at Merck Research Laboratories), 8 Wyngrave Place, Knotty Green, Beaconsfield, Buckinghamshire, UK HP9 1XX. tel: 144 1494 677561, e-mail: arwill37@aol.com

 1359-6446/99/$ – see front matter PII: S1359-6446(99)01303-3

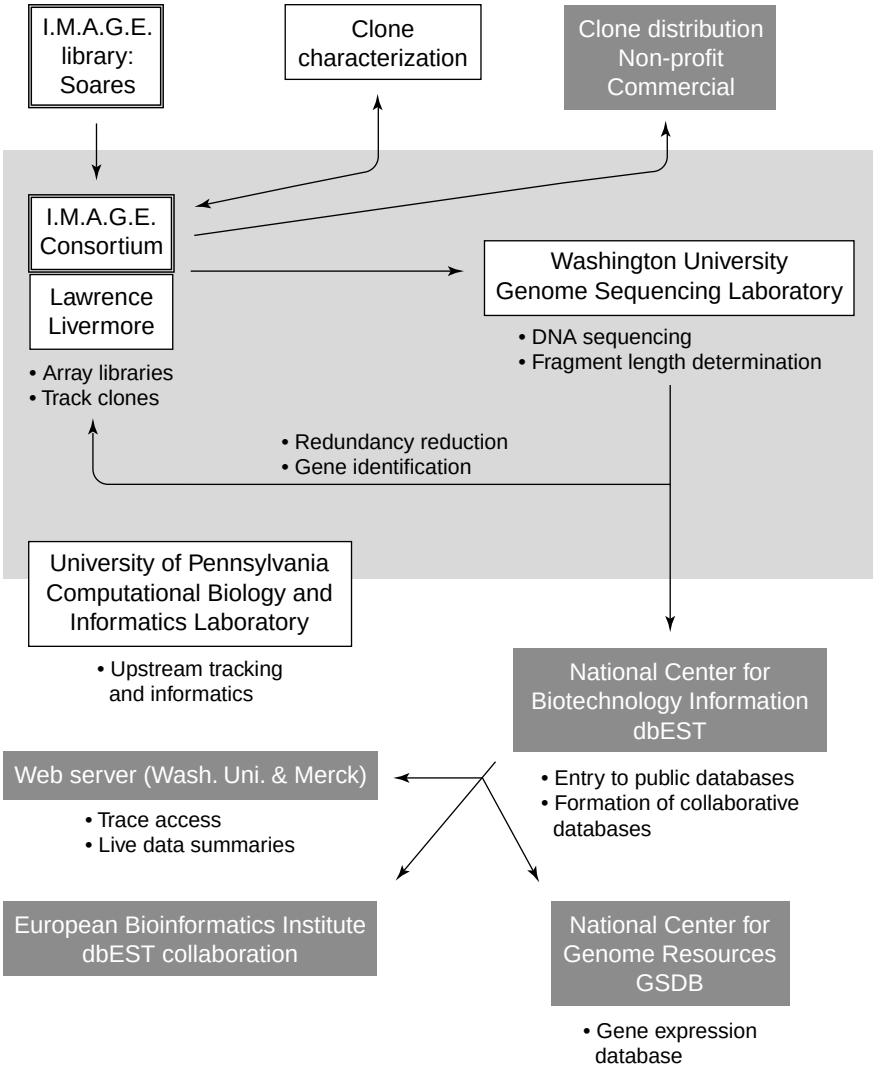

Figure 1. *Logistical plan of the collaborations involved in the MGIP. The central grey area depicts LENS database management by the University of Pennsylvania.*

cDNA clones, as a publicly available resource. The project was conceived as a multicenter collaboration organized and managed by Merck[6]. Collaborations were established primarily with the I.M.A.G.E. (Integrated Molecular Analysis of Genomes and their Expression)[7] consortium and with the Genome Sequencing Center at Washington University (St Louis, MO, USA). The roles of these and other collaborating institutions and the relationships between them are shown in Fig. 1. The cDNA libraries are produced mainly by M. Bento Soares, an I.M.A.G.E. member, based at Columbia University (NY, USA) in 1994 but now at the University of Iowa (Iowa City, IA, USA). Soares is a leader in the production of high-quality cDNA libraries and especially of normalized cDNA libraries[8,9] (see below). The libraries are arrayed, through collaboration with the laboratory of Greg Lennon, also of the I.M.A.G.E. consortium, then at the Lawrence Livermore National Laboratory (Livermore, CA, USA) and recently moved to GeneLogic (Columbia, MD, USA). Arrayed cDNA clones are sent to the Genome Sequence Center (Robert Waterston, Director) for single pass sequencing from each end. The sequence data are quality controlled by the informatics team at Washington University and sent with minimal delay, usually 48 h, directly to the public EST division (dbEST) of the GenBank, a Sybase database managed by Mark Boguski at the National Center for Biotechnology Information (NCBI, Bethesda, MD, USA). The dbEST stores the results of pairwise comparison between all EST sequences and all known protein and nucleic acid sequences, and is updated on a nightly basis[10].

The latter point is very important because almost all of the EST sequencing prior to the MGIP, and indeed most of it that has been undertaken since, has been by 5′-end sequencing. The logic of tagging each clone at both the 3′ and the 5′ end is based on the fact that the cDNA clones used are primed from the 3′ poly-A sequence and so each cDNA molecule contains an untranslated 3′-end sequence that provides a unique identifier for each gene. Thus, the unique set of 3′ EST sequences can be used to index each human gene, while the corresponding 5′ ESTs may add information from the coding region of the genes.

Merck announced in September 1994 the plan to develop a collection of ESTs of human genes, with associated

Additionally, Washington University makes the original ABI sequence trace files for all ESTs sequenced available from their Web site. The Computational Biology and Informatics Laboratory at the University of Pennsylvania School of Medicine manages the LENS database that provides integration of data and monitors consistency. The

Box 1. Commercial vendors of I.M.A.G.E. data

I.M.A.G.E. distributors are:

- American Type Culture Collection, Manassas, VA, USA
- Genome Systems, St Louis, MO, USA
- Research Genetics, Huntsville, AL, USA
- UK Human Genome Mapping Project Resource Centre, Hinxton, UK
- Resource Center of the German Human Genome Project, Berlin, Germany

For information on line:

- I.M.A.G.E. (http://www-bio.llnl.gov/bbrp/image/image.html)
- Merck (http://www.merck.com/)
- Washington University (http://genome.wustl.edu/est/esthmpg.html)
- UniGene

sequenced I.M.A.G.E. clones are also released to five commercial vendors (Box 1) for distribution to the public, at reasonable fees.

The first MGIP sequence data were sent directly from Washington University to the public databases on 15 February 1995. During the first year, the MGIP contributed >260,000 sequences from >150,000 clones to the public EST database dbEST, and the flow has continued at a rate of 4000–5000 EST sequences per week (Fig. 2). The contributions as of December 1998 are summarized in Box 2. The addition of nearly one million EST sequences to Genbank has fueled the exponential growth of the database (Fig. 3), and at this point the MGIP data set accounts for 79% of the human ESTs reported (>1,200,000) to the dbEST.

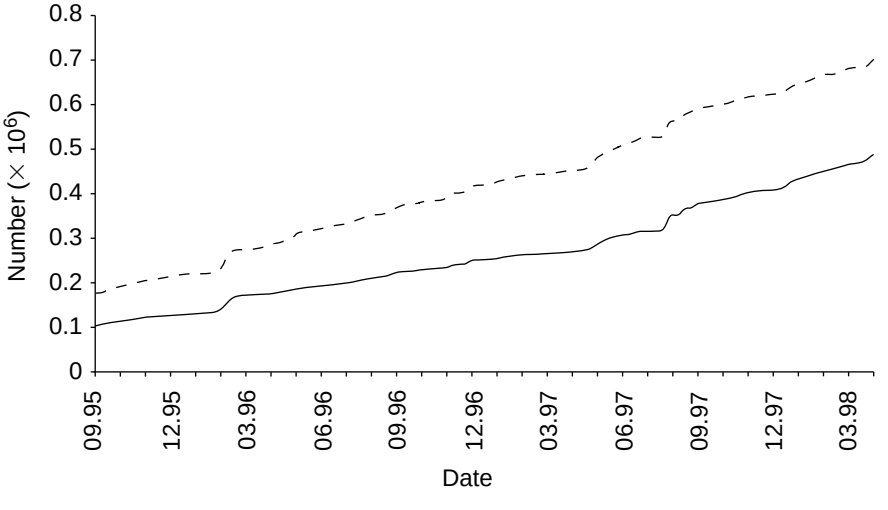

Figure 2. *Flow of data from Washington University to the dbEST since inception of the MGIP.*

MGIP philosophy

The MGIP was conceived on the basis of the clear distinction between patentability and accessibility. Merck wished to make data on human expressed gene sequences accessible for biological research everywhere.

Whether one is in a large pharmaceutical company or a major government or academic institution, one can say accurately that most research is performed elsewhere. The goals of individual research groups may vary but in each case progress comes from integration of the efforts of many groups. Academics see research as a path to knowledge and knowledge as a worthy end in itself; there is healthy competition among academics for priority of discovery. Pharmaceutical companies see research as a way to discover novel therapeutic agents; companies compete to produce and bring to market the best of such agents. Each company needs to be poised to spot any new finding that may be exploited to discover a novel therapeutic agent and also to be able to enter a given therapeutic field at the right time to gain a competitive advantage. Research tools are generally 'pre-competitive' for the pharmaceutical companies.

The MGIP generates research tools that are of equal value to researchers in academia and the biotechnology

Box 2. Summary of data from the MGIP

MGIP contribution to the dbEST (summary data December 1998):

- 647,842 clones sampled from 144 cDNA libraries
- 978,722 EST sequences:
  - 608,127 MGIP (47 libraries)
  - 370,595 CGAP (97 libraries)

MGIP contribution is 79% of human ESTs in the dbEST

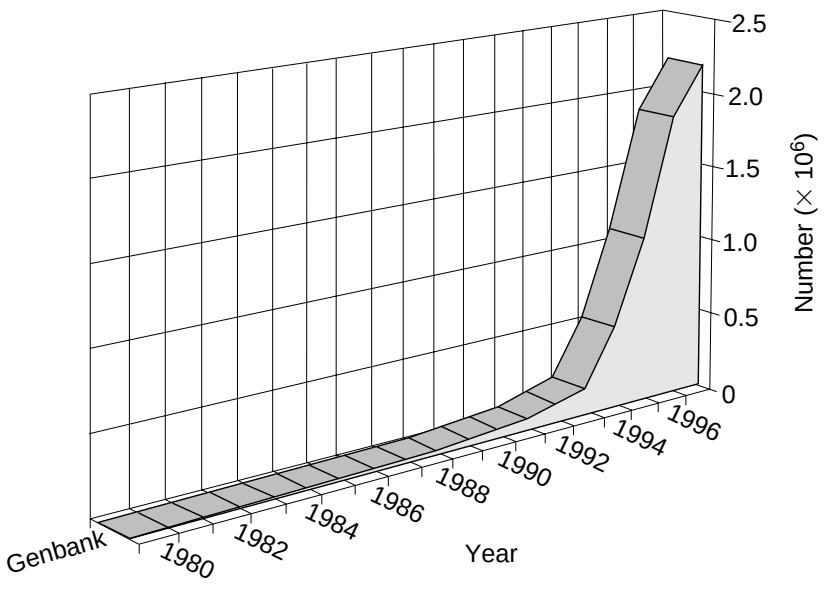

Figure 3. *Accumulation of EST data in the dbEST.*

and pharmaceutical industry. The MGIP was not intended either to erode the value of privately held databases or to reduce any company's prospect of securing intellectual property rights if the claims to utility are deemed appropriate. Sequence information alone does not constitute a patentable invention, so the contribution of ESTs to the public domain does not equate with putting patentable research tools into the public domain. On the contrary, making the EST data freely available should stimulate patentable inventions stemming from subsequent elucidation of the entire sequence, function and utility of each gene.

The set of genes for each species is a coded description of the biology of that organism, but we are profoundly ignorant when it comes to reading DNA sequences to reveal the biology. Illuminating that area of ignorance will take the combined talents and efforts of the world's scientists for years to come. For this reason, the DNA sequences – the tools of the biological researcher – should not be owned in a restricted manner that would slow the progress of scientific understanding.

MGIP scientific plan

In addition to providing a publicly accessible EST data set, the MGI Project aimed to improve upon the commercial EST projects that were already ongoing. The two key advantages in the design of the MGIP are:

- Use of normalized cDNA libraries
- Generation of two ESTs from each cDNA clone (i.e. from the 5′ and 3′ end, for reasons discussed above)

In normalized libraries the representation of each expressed gene is equalized so that rare and common transcripts are present at a similar frequency. The use of normalized libraries has greatly reduced the redundancy in sequencing and correspondingly has increased the efficiency of the project (Table 1). Analysis of the first 300,000 ESTs generated by the MGIP shows that normalization reduces significantly the relative abundance of redundant cDNA clones, without resulting in the complete removal of members of gene families[11].

By comparing the design of the MGIP with the EST project conducted by The Institute for Genomic Research (TIGR, Rockville, MD, USA)[12] (Table 1), three points become clear. First, that normalization of libraries increases the efficiency with which novel sequence can be found; second, that a 3′ EST provides a unique identifier for each clone; and third, that surveying many fewer libraries, each in much greater depth, reveals more fully the diversity of genes expressed in each library.

Cancer Gene Anatomy Project (CGAP)

The CGAP is a project launched by the National Cancer Institute (NCI, Bethesda, MD, USA) of the National Institute of Health, with the aim of establishing an index of all genes that are expressed in tumors[13]. As most genes are probably expressed among tumors, the CGAP target is not substantially different from that of the MGIP. The CGAP is in fact an extension of the MGIP using the same scientific plan and working with the same collaborators at Washington University and the I.M.A.G.E. consortium. The main difference is in the use of cDNA libraries from a wide variety of tumor tissues. The CGAP has contributed ESTs from twice as many libraries as have been sampled for the MGIP but with many fewer clones per library being sequenced (Table 1). The CGAP aim is to generate a Tumor Gene Index (TGI). Given the acknowledged pleiotropic expression of most genes[14] (and see below) the TGI may not differ significantly from either the MGI or UniGene set (see below).

A common aim of the MGIP and the CGAP is eventually to sequence all complete transcripts, starting with the complete inserts in the longest cDNA clone corresponding to

Table 1. Design of MGI, CGAP and TIGR EST projects

	MGI	CGAP	TIGR
Libraries	Normalized 47 libraries Directional oligi dT Primed	Normalized and non-normalized 93 libraries Directional oligi dT Primed	Non-normalized 310 libraries Directional oligi dT Primed
Sequencing	3′ and 5′ EST for each clone 604,154 ESTs	3′ and 5′ EST for each clone 258,753 ESTs	5′ EST for each clone 174,472 ESTs
Average per library	12,832 ESTs	2535 ESTs	562 ESTs
Range	115–85,982 reads <200 in 1 library >1000 in 40 libraries	109–47,662 reads <200 in 4 libraries >1000 in 53 libraries	Not reported <200 in 115 libraries >1000 in 55 libraries

each unique transcript. It is expected that many of these inserts will not cover the complete transcribed gene sequence. The limiting factor is that technologies do not exist for development of libraries composed mainly of full-length cDNAs and suitable methods for generating such libraries need to be developed.

Analysis of MGIP data

To make effective use of the EST sequence data flowing into the dbEST, sequences must be organized into gene classes so that scientists can 'mine' the gene class data in the context of related genomic data. Analysis of the EST sequence data to generate a 'gene index' uses procedures for automated sequence clustering. Algorithms for such clustering continue to be improved as the data are accumulating. Merck generates and supports the MGI for use by in-house researchers[15–17]. Probably other pharmaceutical companies are also generating in-house analyses of the publicly available data. Incyte produce a combined analysis of the public data plus their internal data and that product is out-licensed non-exclusively – many pharmaceutical companies subscribe to the Incyte databases. The NCBI produces a publicly available clustering analysis called UniGene[18]. The current MGI and UniGene analyses yield 55,067 and 60,502 transcript clusters, respectively. With all such clustering analyses one must caution that the numbers derived are not directly equated to the number of genes because there is known to be some alternative splicing in 3′ untranslated regions (UTRs).

The Merck Gene Index is a non-redundant set of clones and sequences, each representing a distinct gene, constructed from the 3′ EST sequences present in the dbEST. An initial analysis of the data and the construction of the MGI as well as a browser for accessing the Index have been described[16,17]. The MGI is constructed iteratively from all index-quality 3′ ESTs – index quality control eliminates poor quality sequences and sets a minimum of 100 base-pairs in length for an acceptable EST.

The Indexing Algorithm compares each index-quality 3′ EST against the Index and if it is equivalent to an Index entry, the EST is placed into that Index class; if it is novel with respect to the Index, a new class is created with that 3′ EST as its representative sequence. Incremental runs of the indexing algorithm are performed nightly on any new EST data arriving in GenBank updates. Upon completion of an incremental run, results are loaded into the internal relational database that underlies the MGI browser. About one in every five clones is being included in the evolving Merck Gene Index (Fig. 4).

The MGI Browser is an easily extensible World Wide Web-based system for mining the MGI and related genomic data[17], and it integrates data from a variety of sources and storage formats. Data currently integrated include LENS cDNA clone and EST data, dbEST protein and non-EST nucleic acid similarity data, Washington University sequence chromatograms, Entrez sequence and Medline

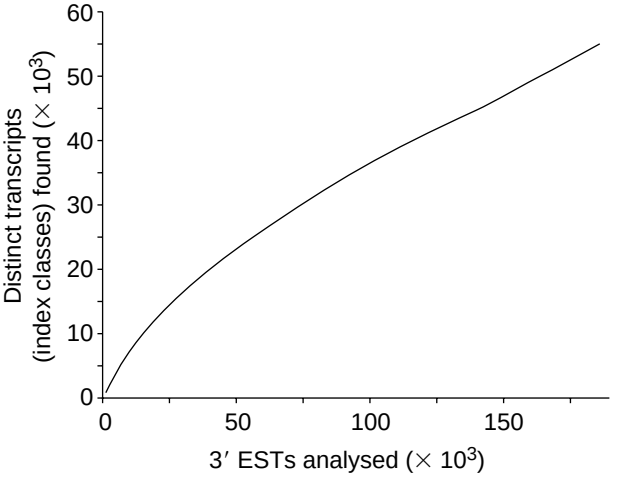

Figure 4. *Rate of discovery of novel transcript clusters in the MGI.*

entries, and UniGene gene clusters. As the field of genomics is generating data at an explosive rate, the MGI browser has been specifically designed to facilitate the addition of many additional types of data, both Merck proprietary and publicly available data.

The MGI is coordinated with the associated cDNA clones as individual clones, sets of clones and as high-density gridded filters. It is planned that all of the cDNA clones incorporated into the Index will be re-sequenced to verify their identities.

UniGene is an experimental system for automatically clustering all GenBank sequences, of which ESTs are the major part, into a non-redundant set of gene-oriented clusters[18]. Each UniGene cluster contains sequences that represent a unique gene, and this is annotated with related information such as the tissue types in which the gene has been expressed and the map location generated by mapping sequence-tagged sites (STSs) made from the UniGene classes. The NCBI provides the UniGene Web browser to facilitate public access (http://www.ncbi.nlm.nih.gov/UniGene/index.html).

One important analytical function, which is tracked as the project progresses, is the percentage of novel EST sequences being found in each library. This provides a guide to whether or not sequencing should go deeper into that library. The fetal liver/spleen cDNA library (Soares_fetal_liver_spleen_1NFLS) has been the most productive; it has produced 85,982 EST reads from 50,629 clones yielding 11,083 transcript clusters, and contains 21.9% diversity and 1750 (3.5%) unique transcript clusters.

Coverage of human genes

One frequently asked question is what proportion of all human genes is covered by the data so far? As we do not know exactly how many human genes make up a complete set it is possible only to estimate the coverage.

The number of human genes has been estimated to be $\geq$70,000 (Refs 1,14). One basis for this estimate is the prediction that mammals probably have four to six times as many genes as *Caenorhabditis* and *Drosophila*, because a significant component of the mammalian increase may have occurred by polyploidization. It is hypothesized that the evolution of mammalian genomes has involved at least two whole genome duplications of an ancestral genome that was equivalent in size to the contemporary fly and worm genomes (each estimated to contain ~18,000 genes)[14]. In this context it is interesting to note a theoretical estimate of the number of human genes based on gene prediction in long contiguous regions of genomic sequence. In 27 Mb of high-quality contiguous human genomic sequence, the Genscan program predicts 530 genes, which extrapolates to a total of ~58,000 human genes (R. Waterston, pers. commun.).

Using 70,000 as the number of human genes and an estimate of about 58,000 clusters (averaging the MGI and UniGene estimates), the current data would cover 83% of human genes.

An estimate of the coverage of known genes is provided by comparing the MGI EST sequences with the set of known, positionally cloned human genes that are mutated in human disease states. As of December 1997, 91% (83/91) of these genes are represented in the dbEST – an increase from 38% (12/32) in 1994.

The surest estimate of the fraction of human genes sampled by EST sequencing will be afforded when a large portion of the genome sequence is available; a low pass or 'rough draft' sequence would be most informative. A preliminary estimate of this type comes from mapping the ESTs onto currently available, contiguous regions of the human chromosomal DNA sequence. On a well-studied 235 Kb segment from the DiGeorge region, there are 11 known genes. Of these 11 genes, eight (72%) are hit by ESTs. In addition, four ESTs hit regions that were not previously known to be genes.

It is impossible to estimate the number of genes that are rarely expressed (e.g. at a certain time during development or upon a certain environmental challenge). Classical genetic studies in mice and fruit flies have shown that many genes have pleiotropic effects, and more recent studies of gene expression support the concept that nearly all gene products will be found to be expressed and utilized at multiple sites and times during development[14]. Pleiotropic gene expression in human cells and tissues is consistent with the low percentage of novel ESTs found in any cDNA library (current range 0–11.9%, although only eight in 144 libraries have >5% novel ESTs). Clearly, the EST strategy should be capable of revealing rarely expressed genes in appropriate normalized libraries.

Utility of gene sequences

The increased public accessibility of the sequence data and cDNA clones afforded by the MGIP has provided a starting point for research that will ultimately lead to new targets for drug design, expressed proteins with potential therapeutic value and disease-related genes that may then become the focus for gene therapy. By making these basic research tools broadly available to the biomedical community on an unrestricted basis, the probability of breakthrough discoveries should increase. Moreover, it will also provide plenty of opportunities and preserve incentives for future gene-based product development.

Although some human genes will have utility as therapeutic or diagnostic agents, it is likely that the utility of most genes and ESTs will be purely as basic research tools. Application of such research tools will lead to increased understanding of the molecular basis of disease and identification of new enzymes or receptors as potential novel drug targets.

Gene mapping[19] and novel gene discovery are two particularly fertile areas stimulated by dbEST usage. One physical map of the human genome was constructed by making use of 3000 EST sequences[20]. More recently, the Radiation Hybrid Map Consortium organized by the Human Genome Organization (HUGO) has generated a higher resolution transcript map using about 30,000 EST sequences as markers[21]. This mapping consortium expects to extend the map, using up to 50,000 ESTs, to an accuracy of 0.5 Mb (C. Aufray, pers. commun.), making a candidate gene approach more viable than the traditional positional cloning efforts for mapping disease genes. The MGI initiative thus promises to help change the way in which the genetic basis of disease is addressed and understood.

EST data are already being used to study human polymorphisms. A study of size and sequence polymorphisms in the transcribed trinucleotide repeat D2S196E is of significance because this polymorphism can be used for demonstrating microsatellite instability and loss of heterozygosity in colorectal tumors[22].

The EST database, particularly 3′ ESTs, will also be a starting point for searching for single nucleotide polymorphisms (SNPs)[23]. SNPs are estimated to occur at a frequency of about one in 1 Kb between any two individual human genomes, and these biallelic markers will be more frequent when populations are compared. A proportion of SNPs will be found in genes, referred to as coding-SNPs or cSNPs, and they will be more likely to occur in UTRs – hence the value of the 3′ ESTs. SNPs will be important research tools for genetic association studies and have the potential to be used in the development of diagnostic, prognostic and therapeutic approaches to diseases that have an underlying genetic component.

For drug discovery one of the key uses of the MGI data lies in identification of novel molecular targets for therapeutic intervention. Some of the ways in which the data are being used are described below.

Exploring gene families of interest

ESTs are being identified corresponding to families of genes of therapeutic interest, particularly classes of genes that are known to have yielded drug targets such as the G-protein coupled receptors, nuclear receptors, ion channels and various enzyme families. Examples of new family members identified using EST data include a novel human γ-aminobutyric acid-type receptor gene[24], genes involved in signal transduction (e.g. inositol-1,3,4-trisphosphate 5-6-kinase[25] and p38b MAP kinase[26]), a gene encoding a protein involved in DNA alkylation-damage repair[27], two novel and apparently tissue-specific mono(ADP-ribosyl)transferases[28], two novel members – LRP5 and LRP6 – of the low-density lipoprotein receptor family[29,30] and >10 new human chemokines from the previously known families (termed C, CC and CXC, based on the spacings of N-terminal cysteine residues), plus the first member of a novel chemokine subfamily possessing the CXXXC cysteine spacing[31].

Differential gene expression

One of the most powerful uses of genomics lies in techniques for identifying genes that are differentially expressed in a given disease state, for example, an appropriate tissue type under pathological conditions compared with its healthy physiological state. Both hybridization and sequence-based methods[32–36] can be applied to determine differential expression of genes by using as markers cDNA clones corresponding to unique ESTs.

Identifying 'disease genes'

Genetic evidence from a human disease can be a compelling validation of a given target mechanism. Mapping ESTs to genomic regions containing disease-linked genes can help to identify mutations in candidate genes that may lead to disease susceptibility. Examples of discoveries made using EST data include a novel gene (presenilin-2)[37] associated with Alzheimer's Disease and LRP5 and LRP6 (Refs 29,30) associated with type 1 diabetes.

Non-human gene homologs

Identifying animal species homologs of a known human target protein allows the rapid evaluation of the potential utility of a given species as a disease model or an animal model for pharmacodynamic studies. For example, gene 'knock-out' in mice can yield valuable data on the role of a given gene in physiology and pathology.

Human homologs for genes identified in other species

Identifying homologies with functionally interesting genes derived from model organisms (e.g. *Saccharomyces cervisaea*, *C. elegans* and *D. melanogaster*) can elucidate the function of the human gene. Examples of genes with human homologs include the yeast OGG1 gene that is involved in the repair of oxidative DNA damage[38] and the

dystrophin-related phosphoprotein found at the electric organ post-synaptic membrane of *Torpedo californica*[39]. There are also three new mammalian ATP-binding transporter genes[40] and the putative functional identification of 66 human genes by homology of ESTs with *Drosophila* mutant phenotypes[41].

Conclusions

The uses of MGI data illustrated in this review point the way for the future. As the EST data become more comprehensive, and as full-length cDNA sequences are generated, the breadth of the utility of the MGI will increase. Expression analysis data will be generated and annotated to the Index providing much better information than the mere tissue source of the ESTs currently available. Protein expression analysis using two-dimensional gel electrophoresis with coupled MS/MS sequence analysis is already being interpreted by reference to the expressed sequence data[42–44] and these will be important tools in the future.

The MGI data are facilitating genomics research worldwide. The results of genomics are informing the choices of drug targets and their validation. A map of human SNPs, particularly cSNPs, discovered using the MGI data should prove to be an essential tool for pharmacogenetic and toxicogenetic studies in drug development. Moreover, the predicted power of ESTs is being fully realized through the open public access afforded by the MGIP.

ACKNOWLEDGEMENTS

The MGIP is the result of teamwork. Keith Elliston played a key role in the conception and management of the project. The project has been possible though the dedicated efforts of Jeffrey S. Aaronson, Kamil Ali-Jackson, Wendy Bailey, Mary Bartkus, Werten Bellamy, Richard A. Blevins, Joseph A. Borkowski, Barbara A. Eckman, Oliver Johnson, Anthony Starks, Jeffrey Sturchio, Eileen Undercoffler and Cindy Zarsky with the support and vision of Merck senior management. Key external collaborators making the MGIP possible include Robert Waterston, Rick Wison, Marco Marra and LaDeana Hillier (Washington University), Bento Soares (Columbia University), Greg Lennon (I.M.A.G.E.), Mark Boguski and Carolyn Tolstoshev (NCBI), Chris Overton and Mark Gibson (University of Pennsylvania, Philadelphia, PA, USA) and Ken Fasman (Genome Database, Johns Hopkins University, Baltimore, MD, USA).

REFERENCES

1 Collins, F.S. (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 10821–10823
2 Brenner, S. (1990) *Ciba Found. Symp.* 146, 6–12
3 Okubo, K. *et al.* (1992) *Nat. Genet.* 2, 173–179
4 Adams, M.D. *et al.* (1993) *Nat. Genet.* 4, 256–267
5 Adams, M.D. *et al.* (1993) *Nat. Genet.* 4, 373–380
6 Williamson, A.R. *et al.* (1995) *J. Natl Inst. Health Res.* 7, 61–63
7 Lennon, G.G. *et al.* (1996) *Genomics* 33, 151–152
8 Soares, M.B. *et al.* (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 9228–9232
9 Bonaldo, M.F., Lennon, G. and Soares, M.B. (1996) *Genome Res.* 6, 791–806
10 Boguski, M.S., Lowe, T.M.J. and Tolstoshev, C.M. (1993) *Nat. Genet.* 4, 332–333
11 Hillier, L. *et al.* (1996) *Genome Res.* 6, 807–828
12 Adams, M.D. *et al.* (1995) *Nature* 377, 3–174
13 Strausberg, R.L., Dahl, C.A. and Klausner, R.D. (1997) *Nat. Genet.* Apr. 15 Spec No. 415–416
14 Miklos, G.L.G. and Rubin, G.M. (1996) *Cell* 86, 521–529
15 Blevins, R. *et al.* (1995) *Computer Appl. Biosci.* 11, 667–673
16 Aaronson, J.S. *et al.* (1996) *Genome Res.* 6, 829–845
17 Eckman, B.A. *et al.* (1998) *Bioinformatics* 14, 2–13
18 Schuler, G. *et al.* (1996) *Science* 274, 540–546
19 Wilcox, A.S. *et al.* (1991) *Nucleic Acids Res.* 19, 1837–1843
20 Hudson, T.J. *et al.* (1995) *Science* 270, 1945–1954
21 Deloukas, P. *et al.* (1998) *Science* 282, 744–746
22 Haddad, L.A., Fuzikawa, A.K. and Pena, S.D.J. (1997) *Hum. Genet.* 99, 796–800
23 Collins, F.S., Guyer, M.S. and Charkravarti, A. (1997) *Science* 278, 1580–1581
24 Whiting, P.J. *et al.* (1997) *Biochem. Soc. Trans.* 25, 817–819
25 Wilson, M.P. and Majerus, P.W. (1996) *J. Biol. Chem.* 271, 11904–11910
26 Wei, Y-F. *et al.* (1996) *Nucleic Acids Res.* 24, 931–937
27 Jiang, Y. *et al.* (1996) *J. Biol. Chem.* 271, 17920–17926
28 Braren, R. *et al.* (1997) *Adv. Exp. Med. Biol.* 419, 163–168
29 Brown, S.D. *et al.* (1998) *Biochem. Biophys. Res. Commun.* 248, 879–888
30 Hey, P.J. *et al.* (1998) *Gene* 216, 103–111
31 Wells, T.N.C. and Peitsch, M.C.T. (1997) *J. Leukocyte Biol.* 61, 545–550
32 Velculescu, V.E. *et al.* (1995) *Science* 270, 484–487
33 Zhao, N. *et al.* (1995) *Gene* 156, 207–213
34 Marshall, A. and Hodgson, J. (1998) *Nat. Biotechnol.* 16, 27–31
35 Southern, E.M. (1996) *Trends Genet.* 12, 110–115
36 Soares, M.B. (1997) *Curr. Opin. Biotechnol.* 8, 542–546
37 Rogaev, E.I. *et al.* (1995) *Nature* 376, 775–778
38 Arai, K. *et al.* (1997) *Oncogene* 14, 2857–2861
39 Sadoulet-Puccio, H.M. *et al.* (1996) *Hum. Mol. Genet.* 5, 489–496
40 Allikmets, R. *et al.* (1995) *Mamm. Genome* 6, 114–117
41 Banfi, S. *et al.* (1996) *Nat. Genet.* 13, 167–174
42 Yates, J.R., 3rd, McCormack, A.L. and Eng, J. (1996) *Anal. Chem.* 68, 534A–540A
43 McCormack, A.L. *et al.* (1997) *Anal. Chem.* 69, 767–776
44 Leffers, H. (1996) *Electrophoresis* 17, 1713–1719