Genetics Update
Impact of the Human Genome Projects and Identification of a Stroke Gene

Mark J. Alberts, MD

Two articles were published in February 2001 that will have a significant impact on our understanding of human development, the pathogenesis of many human diseases, and the discovery of new therapies for many disorders. These articles deal with the mapping of the human genome. Two different entities, one a publicly traded company (Celera) and the other the Human Genome Project (HGP, sponsored and funded by NIH), published somewhat different versions of the human genome. The HGP began in 1990 (although extensive sequencing of the human genome began in 1995) and cost approximately $3 billion, while the Celera effort began in 1998. The HGP involved multiple laboratories in the United States and abroad. The 2 projects produced maps that differ from each other in terms of completeness, order of some genetic markers, and the ability to search the database for specific DNA sequences. A comparison of some features of both projects is in Table 1.

The challenge of sequencing the 3 billion base pairs of the human genome required the development of unique tools and approaches. Celera constructed a facility capable of high-throughput sequencing at a rate of 175,000 reads per day and conducted sequencing 24 hours a day, 7 days a week. The HGP divided the sequencing task among several large laboratories with demonstrated expertise in large-scale DNA sequencing. The strategy employed by the HGP focused on subcloning the human genome into bacterial artificial chromosomes (BAC), which were then sequenced and properly arranged. Each BAC could hold an insert of 150,000 bases of DNA and consists of 7 to 9 exons. The mean coding sequence for a human gene is 1340 base pairs. Between 20% and 40% of human genes have just 1 or 2 exons, while some have more than 20 exons. The gene with the most exons is titin, which has 234 exons. Genes are not evenly distributed throughout the human genome. Some chromosomes (17, 19, and 22) have a relative abundance of genes, while others (chromosomes 4, 13, 18, and X) have an apparent paucity of genes.

Each map produced several unexpected findings. Prior studies had estimated that humans have approximately 100,000 unique genes, and it was expected that these genome mapping projects would identify essentially all of these genes. Such was not the case, however. Results from the genome mapping projects have shown that humans have between 26,500 and 39,000 genes, compared with about 18,000 genes in a worm and 13,000 genes in a fly. This has implications for understanding evolutionary and developmental genetics. It may indicate that gene function and control is more important than simply having more genes producing more proteins as the explanation for the obvious differences between humans and other species.

The relatively small number of genes does not indicate a similarly small number of proteins. A single gene can undergo alternative splicing, thereby producing different proteins. Studies of messenger RNA have shown that there may be an average of 2.6 to 3.2 distinct transcripts per gene. Parts of separate genes can combine to form unique proteins. Proteins can undergo extensive posttranslational modifications that can change their function and activity. Studies comparing the human genome with other vertebrate genomes have shown an increase in genes related to neuronal function, developmental regulation, hemostasis, and immunologic function in the human genome.

A question sometimes asked is “Whose DNA was used for the sequencing?” The Celera group initially enrolled 21 donors, of whom 5 were selected for the DNA sequencing studies. The 5 individuals consisted of 3 females and 2 males: 2 Caucasians, 1 Asian-Chinese, 1 African-American, and 1 Hispanic. The DNA was obtained from whole blood in all cases, and additional DNA was obtained from semen samples. For the HGP, samples were obtained from a group of volunteers. However, all identifying information was removed from each sample; therefore, there is no information about the gender or ethnic/racial background of the donor samples sequenced by the HGP. For both the Celera project and the HGP, an Institutional Review Board reviewed the donation protocol, and in all cases the donors signed informed consent forms.

The opinions expressed in this editorial are not necessarily those of the editors or of the American Heart Association.

From the Department of Neurology, Northwestern University Medical School, Chicago, Illinois.

Correspondence to Mark J. Alberts, MD, Director, Stroke Program, Department of Neurology, Northwestern University Medical School, 710 N Lake Shore Dr, Room 1122, Chicago, IL 60611. E-mail malberts@northwestern.edu

(Stroke. 2001;32:1239-1241.)
© 2001 American Heart Association, Inc.

Stroke is available at http://www.strokeaha.org
In addition to identifying all human genes, the mapping projects have also produced a detailed map of thousands of genetic markers spaced fairly evenly throughout the genome. There are various types of markers that have been mapped as part of these projects, the most abundant of which are the single nucleotide polymorphisms, also referred to as SNPs. The human genome sequencing projects have identified approximately 2.1 millions SNPs. These markers are important for performing genetic linkage studies, which is one powerful method for determining exactly where on a chromosome a specific disease locus may reside. The availability of a map with numerous markers that are accurately mapped in terms of physical location provides researchers with a valuable tool for determining with high precision the location of a disease-causing gene. The researcher can then examine the genes that are close to the linked markers and determine whether one or more are good candidate genes for the disease under study. Since all of the genes have now been identified and mapped, the identification of appropriate candidate genes has been simplified greatly. This general approach has been used to identify several dozen pathogenic genes using the publicly available draft genome.

Another by-product of the genome sequencing effort is the identification of new targets for conventional drug therapy. Examples include the cloning of the β-secretase gene/protein that cleaves the amyloid precursor protein, 2 new dopamine receptors, a new serotonin receptor, 2 chains of the tubulin gene, and a glycine receptor. Using this knowledge to develop new medications that act on various cell receptors or extracellular proteins has been easier and more productive than direct gene therapy. One reason is that it is much more difficult to transport a specific gene into specific cells and have it act at the proper location in a complex genome, compared with “simply” blocking or activating a specific cell-surface receptor. Many of the most common and efficacious medications now in use act on cellular receptors or extracellular proteins, while few act on intracellular pathways or specific genes.

Identifying all of the human genes is but one step in understanding disease pathogenesis. We must now learn how these new proteins function (and malfunction) in novel and complex metabolic pathways. While the human genome sequence gives us a road map, much work remains in understanding where the road leads and what detours may exist. We are only beginning to understand the complex, multigenic disorders that are responsible for some of the most common maladies in society such as atherosclerosis, hypertension, diabetes, psychiatric disorders, and stroke. Correlating the thousands of individual genomic variations with some of these disorders will be extremely challenging due to the large number of variables within an individual and across entire populations.

In a separate but related project, a study by Stefasson and colleagues from deCode in Iceland report linkage and cloning of a unique stroke gene. The gene, on chromosome 5, appears to be important for endothelial function, although more details have not yet been published. One surprising finding is that the gene appears to be related to all types of ischemic stroke but is not linked to other types of vascular disease such as myocardial infarction or peripheral arterial disease. Perhaps this reflects some unique aspect of the vessels in and around the brain. Another possibility is that due to some unique genetic features of the Icelandic population, the results may not be generally applicable for all stroke types or in all populations. Further studies will be needed to address these issues.

The identification of a putative “stroke gene” is quite different from many recent studies of stroke genetics. To date, many genetic studies of stroke have dealt with specific polymorphisms that are overrepresented or underrepresented in the disease state compared with controls. There have been dozens of studies of disease-related polymorphisms, yet few have any functional significance in terms of gene function and few have been linked to a disease locus using rigorous genetic linkage techniques. One exception is the apolipoprotein E gene, in which specific genotypes have been strongly associated with cerebral amyloid angiopathy, cerebral hemorrhage, and recovery of function after a stroke. In general, caution is urged in interpreting such polymorphism studies, since past experience has shown that in many cases they do not identify causative genes.

There are examples of specific mutations that cause hereditary forms of stroke. A classic example is the CADASIL syndrome, where specific mutations in the NOTCH3 gene cause the disease. Mutations in the KRIT1 gene cause some forms of familial cavernous hemangiomas. Moyamoya disease has recently been linked to several genetic loci, although a specific gene or mutation has not yet been identified. Research is ongoing to define specific disease-causing genes and mutations for other common disorders such as familial aneurysmal subarachnoid hemorrhage, fibromuscular dysplasia, and lupus anticoagulants.

In summary, sequencing the human genome along with the mapping and characterization of thousands of genes is a dramatic step toward unraveling the many genetic disorders that cause familial cerebrovascular disease. The cloning of a putative stroke gene that is responsible for cerebral infarction is another major advance in this area. Much work is still needed to discern how these mutated proteins lead to the disease phenotype and how to treat or prevent strokes in susceptible individuals. These developments may also make

---

### Comparison of the Human Genome Project and the Celera Project

<table>
<thead>
<tr>
<th></th>
<th>Celera</th>
<th>HGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization</td>
<td>Public company</td>
<td>Government/NIH, DOE</td>
</tr>
<tr>
<td>Location</td>
<td>United States</td>
<td>United States, multinational</td>
</tr>
<tr>
<td>Start date</td>
<td>1998</td>
<td>1995</td>
</tr>
<tr>
<td>Cost</td>
<td>?</td>
<td>$3 billion</td>
</tr>
<tr>
<td>Total genome size</td>
<td>2.91 billion</td>
<td>3.2 billion</td>
</tr>
<tr>
<td>Percentage of finalized sequence</td>
<td>&gt;95%</td>
<td>90–94%</td>
</tr>
<tr>
<td>Sequencing strategy</td>
<td>Shotgun</td>
<td>BAC</td>
</tr>
<tr>
<td>No. of genes</td>
<td>26 500–39 000</td>
<td>30 500–35 500</td>
</tr>
<tr>
<td>SNPs</td>
<td>2.1 million</td>
<td>&gt;1.4 million</td>
</tr>
</tbody>
</table>

DOE indicates Department of Energy. Data adapted from References 1 and 2.
it possible to provide genetic counseling for at-risk indi-

ciduals. Considering the rapidly changing knowledge in this
area, neurologists and other healthcare professionals should
keep abreast of new developments, since it may directly
affect their clinical practices and their patients.

References
1. International Human Genome Sequencing Consortium. Initial sequencing
3. The International Human Genome Mapping Consortium. A physical map
4. The International SNP Map Working Group. A map of human genome
sequence variation containing 1.42 million single nucleotide poly-
5. Peltonen L, McKusick V. Dissecting human disease in the postgenomic
E, Agnarsson U, et al. A major susceptibility locus for common stroke
identified on chromosome 5q21. In press.
and abstracts of the 26th International Stroke Conference; February
14–16, 2001; Fort Lauderdale, Fla.
1784–1812.
9. McCarron MO, Nicoll JA. Apolipoprotein E genotype and cerebral
amyloid angiopathy-related hemorrhage. Ann NY Acad Sci. 2000;903:
176–179.
10. O’Donnell HC, Rosand J, Knudsen KA, Furie KL, Segal AZ, Chiu RI,
Ikeda D, Greenberg SM. Apolipoprotein E genotype and the risk of
240–245.
11. McCarron MO, Hoffmann KL, DeLong DM, Gray L, Saunders AM,
Alberts MJ. Intracerebral hemorrhage outcome: apolipoprotein E geno-
12. Joutel A, Dodick DD, Parisi JE, Cecillon M, Tournier-Lasserve E,
Bousser MG. De novo mutation in the Notch3 gene causing Cadasil.
Cecillon M, Marechal E, Joutel A, Bach JF, Tournier-Lasserve E. Trunc-
cating mutations in CCM1, encoding Krit1, cause hereditary cavernous
H, Inoue T, Ikezaki K, Matsumata T, Fukui M. Linkage of familial
moyamoya disease (spontaneous occlusion of the circle of Willis) to
15. Inoue TK, Ikezaki K, Sasazuki T, Matsumata T, Fukui M. Linkage
15:179–182.

Key words: genetics ▪ stroke
Genetics Update: Impact of the Human Genome Projects and Identification of a Stroke Gene
Mark J. Alberts

*Stroke*. 2001;32:1239-1241
doi: 10.1161/01.STR.32.6.1239

*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2001 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/32/6/1239

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Stroke* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to *Stroke* is online at:
http://stroke.ahajournals.org//subscriptions/