GIST: A web tool for collecting gene information

MARC K. HALUSHKA, DEBRA J. MATHEWS, JEFFREY A. BAILEY, AND ARAVINDA CHAKRAVARTI

Department of Genetics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106

Halushka, Marc K., Debra J. Mathews, Jeffrey A. Bailey, and Aravinda Chakravarti. GIST: A web tool for collecting gene information. Physiol. Genomics 1: 75–81, 1999.—As the human genome is sequenced and annotated, an important step in future genetic studies of complex traits and diseases will be the identification of relevant candidate genes. To enable such compilations, it would be useful to collate all necessary and available genetic information for each candidate gene. To this end, we have created a web tool (http://genome.cwru.edu/gist/gist.html) to allow the rapid cataloging of currently available genetic data. This tool, called GIST (or “Gene Information Search Tool”), allows an investigator to search the major genomic databases containing gene and marker information from a single query point. To prove the utility of GIST, a catalog of 150 hypertension candidate genes was created. This resource collates all available nucleotide and amino acid sequence data, expression data, chromosomal map location, and genetic marker interval for each gene, collected from on-line databases. These data can be used to guide genetic studies of hypertension.

hypertension; candidate gene; complex genetic disease; genomics; Internet resource

A common problem in studies of complex disease is the difficulty in acquiring the relevant genetic data for rational choice of candidate genes. This has led to the use of only some of the many relevant genes and markers in genetic studies of association and linkage. A primary explanation for the noncomprehensive nature of candidate gene compilations is the lack of easy availability of informatics tools to identify all relevant candidate genes. Although pertinent information such as nucleotide sequence and chromosomal map location for genes can be found in a host of genomic databases, there is no simple way to tailor this information for specific complex genetic disease studies. Harnessing this genomic information is necessary if we are to thoroughly evaluate all potential candidate genes for complex genetic diseases. As the human genomic reference sequence gets completed and then annotated, this task will become more difficult without the appropriate computational tools.

Complex genetic diseases can arise from both environmental and genetic factors (20). Although some environmental effects contributing to human diseases such as diabetes and hypertension are understood, our current knowledge of specific genes contributing to these same diseases is sorely lacking (17). Current methods for gene discovery that have had tremendous success in Mendelian disorders, such as positional cloning, have failed to uncover the genes regulating complex genetic diseases. However, two methods that are beginning to show utility in complex disease studies are genomewide nonparametric linkage analysis (genome scanning) and candidate gene association studies (5, 20). Recently, association studies have gained favor because of the expected reduced sample sizes needed to detect genes of small effect relative to linkage mapping (24). These two types of studies also differ in that a candidate gene association study can implicate or reject the role of a particular gene in a biological process, whereas linkage scans can only implicate genomic regions but not specific genes. To perform ideal association studies we need to identify all available and pertinent candidate genes. But what makes a gene a candidate for a disease? Currently, we define candidate genes as all genes that may contribute to the pathophysiology of disease through their roles in pathways regulating the disease-related phenotypes. For many complex genetic diseases such as diabetes, hypertension, and schizophrenia, the number of genes fulfilling this criterion can be in the several hundreds.

Various types of genomic information are useful for defining candidate genes. For example, a gene's chromosomal map location is useful for assessing candidacy by virtue of proximity to known disease quantitative trait loci (QTLs) implicated by linkage scans. Also, knowledge of a gene's tissue expression pattern is pertinent to a postulated role in a given tissue or cell line. Gene sequences, both cDNA and genomic, are critical for understanding the structure of a gene and for marker development. Finally, for association studies, compilation of all known polymorphic markers (single nucleotide polymorphisms [SNPs] and/or microsatellite markers) can be used to test the individual genes for association. Genetic polymorphism data will ultimately be the most useful for the study of genes. Unfortunately, there has been little effort at comprehensive discovery of polymorphic markers in genes, although this is changing rapidly (3, 8, 14). To generate these markers, most discovery programs are initiated at the nucleotide level from available DNA sequence data.
Acquiring and assembling genomic data for a large number of genes from the many different databases is not a trivial task. To ease this process, we have created a web-based search tool called GIST, the Gene Information Search Tool (Fig. 1). This web page is a one-site query tool for multiple databases. Its frames format has been designed to allow the quick assembly of information from a variety of databases (Table 1). It allows searching both by genes and by microsatellite markers from multiple information sources. To test the utility of this tool, we built an on-line candidate gene list for hypertension containing 150 genes.

Essential hypertension is a leading cause of morbidity and mortality in the United States and is an important target of contemporary biomedical research. Hypertension is clinically defined as a systolic blood pressure greater than 140 mmHg and/or a diastolic blood pressure greater than 90 mmHg (17), and it shows familiality with estimates of heritability ranging from 60% in twin studies to 25% in family studies (29). The familial resemblance of hypertension fits a model in which no single locus contributes overwhelmingly to the trait but, rather, multiple genes and environmental factors, each with a small cumulative effect, contribute to the hypertensive state (15, 25).

A variety of biological/biochemical pathways, composed of potentially hundreds of genes, including the renin-angiotensin system, hypothalamus-pituitary axis, and hormonal regulation, control an individual’s blood pressure via homeostatic mechanisms. Some genes in these pathways, such as that encoding angiotensinogen, are well described; some, such as that encoding chymase, have not been well studied; and many, if not most, are yet to be discovered. However, it is clear that some of these genes must be involved in hypertension given their unequivocal role from physiological studies. Evidence for a genetic role in hypertension has arisen from the study of several genes, including angiotensinogen, α-adducin, and the β3 subunit of G protein (4, 16, 28), through candidate gene association studies. Although the biological implication that interindividual genetic variation in these genes leads to variation in blood pressure is not complete (6, 18, 19), this small collection of genes implicates remarkably different pathways. Therefore, to responsibly characterize hypertension candidate genes, we have “cast a wide net” for identifying a variety of genes from diverse pathways that work directly or indirectly to control blood pressure.
Table 1. Databases that can be searched by GIST

<table>
<thead>
<tr>
<th>Gene searches</th>
<th>Marker searches</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCBI—PubMed</td>
<td>CEPH—D</td>
</tr>
<tr>
<td>NCBI—GenBank</td>
<td>CEPH—AFM</td>
</tr>
<tr>
<td>NCBI—UniGene</td>
<td>NCBI—PubMed</td>
</tr>
<tr>
<td>ExPASy—SwissProt</td>
<td>CEPH—YAC Maps</td>
</tr>
<tr>
<td>GDB—Genome Data Base</td>
<td>WI-MIT—Whitehead/MIT</td>
</tr>
<tr>
<td>NCBI—Entrez Protein Division</td>
<td>GDB—Genome Data Base</td>
</tr>
<tr>
<td>Weizmann Institute—GeneCard</td>
<td>SHGC—Stanford Human Genome Center</td>
</tr>
<tr>
<td>HUGO Nomenclature Committee</td>
<td>CHLC—Cooperative Human Linkage Center</td>
</tr>
<tr>
<td>NCBI—Entrez Nucleotide Division</td>
<td>dbSTS—Database of Sequence Tagged Sites</td>
</tr>
</tbody>
</table>

GIST, gene information search tool; NCBI, National Center for Biotechnology Information; HUGO, Human Genome Organization; RH, radiation hybrid; CEPH, Centre d’Etudes Polymorphismes Humains; AFM, Association Francaise contre les Myopathies; YAC, yeast artificial chromosome; MIT, Massachusetts Institute of Technology.

RESULTS

To test the usefulness of GIST, we built a hypertension candidate gene database that could be viewed through the Internet. We performed a search of the hypertension literature, selecting genes based on prior genetic, animal, or physiological studies, and identified 150 of the most common and likely candidate genes. Although there are certainly scores of other genes that might be deemed as candidates, our discovery was limited to these 150 genes, as explained in discussion. A complete listing of these 150 genes and the information gathered on each is available at the web site http://genome.cwru.edu/candidates/candidates.html (Fig. 2). These 150 hypertension genes have been sorted by gene name, gene symbol, and functional group and have been hyperlinked to separate web pages for each gene. Figure 3 shows the available data for each gene and the databases queried to obtain the information. Using GIST, we were successful in quickly acquiring and collating the necessary genomic information from a variety of databases. The types of information and relative success rates for collection of each type of information are discussed below.

The full gene name, the official gene nomenclature symbol, and possibly a common gene symbol(s) were identified for each gene. Using the gene name or symbol, we queried seven genomic databases. OMIM records were obtained for 147 of the 150 genes. OMIM accession numbers on the individual gene web pages were hyperlinked to provide direct access to the OMIM record. For all 150 genes, two or three GenBank records were obtained for each gene: the first represents DNA sequence from the gene itself, either as a whole or partial genomic segment or as a cDNA; the second accession number is for a nucleotide-sequenced large-insert clone, generally not annotated, on which part, or all, of a gene resides; the third represents a protein record for the gene. These records are also hyperlinked to the accession number to the full GenBank sequence record.

The available GenBank sequence was next parsed into the categories of 5′ (promoter and 5′-UTR), exonic, intronic, and 3′ (3′-UTR and flanking 3′) sequence to represent how much DNA information was known for...
each gene. DNA sequence totaling over 3.6-megabases was obtained for the 150 genes. The most significant variable determining the amount of sequence available for a gene was whether it is contained in a large-insert clone (bacterial artificial chromosome (BAC), P1-derived artificial chromosome (PAC), cosmid, etc.) that has been sequenced as part of the Human Genome Project: parts of 37 genes were contained on large-insert clones, representing 80% of all sequence data available. At the time the list was generated (the data set was frozen as of January 9, 1999), many of these sequences were on unordered “phase 1” clones. Phase 1 sequencing indicates the sequencing record is incomplete and that multiple unordered “contigs” of the larger clone exist. Therefore, approximations of available sequence length, to the nearest 10 kb, for 5′-UTR, introns, and 3′-UTR were made. Approximately one-fourth of the 150 genes (i.e., 36 genes) have been sequenced as cDNAs with no genomic structure information. The remaining 77 genes have various amounts of genomic sequence available.

Tissue expression data were available for 118 candidate genes from the UniGene database, based on the specific library from which the gene, or an expressed sequence tag (EST), was sequenced (13). A gene is assumed to be expressed in a tissue if its cDNA can be sequenced from a specific library derived from that tissue. This method clearly underestimates expression patterns of genes, due to limitations in the depth of sequencing from many tissue-specific EST libraries. This information can be used to identify tissues and cell lines that could produce the transcribed mRNA product. Many candidate genes appear to have ubiquitous expression, as 27 of the candidate genes were found to be transcribed in 10 or more tissues.

The chromosomal location of each gene was also determined. All but two genes had been assigned to at least a specific chromosome, but most genes have detailed cytogenetic map locations. The use of GeneMap 98 allowed the placement of 77% of genes (116) within a radiation hybrid mapped interval between two polymorphic microsatellite markers (9). Genes were located on all chromosomes except on chromosome 15 and the Y chromosome. We found that the pattern of distribution of these genes on the chromosomes differs from that expected for the distribution of a random set of 148 genes (χ² = 53.5, 22 df, P < 0.0002). To determine if this was an effect of gene family clustering, all
clusters were reduced to a single representative in the analysis; the distribution remained nonrandom ($\chi^2 = 44.6$, 22 df, $P < 0.003$), albeit less significant, after the removal of 10 genes. This nonrandom distribution may be the result of a selection bias for these 150 genes arising from human or model organism blood pressure QTL studies, which in turn have encouraged increased study of genes located in these areas. Indeed, we found 17% percent of all genes were located within a linkage group from a human or model organism QTL for blood pressure or a blood pressure intermediate phenotype (12, 15, 21).

**DISCUSSION**

GIST. GIST was designed to reduce the tedious process of consulting multiple on-line databases for information regarding a specific gene (Fig. 1). GIST is a simple JavaScript and HTML application that allows queries of all major genetic databases from a single site. Three types of searches can be performed with GIST: a gene search, a human-only gene search, and a marker search. To use GIST, the name of the gene or marker of interest is entered, followed by selection of an appropriate database (listed in Table 1) to send an Internet query to the database for the gene entered. For human-only queries, the query will be sent with a tag to return human data only, where such data exists. Twelve different databases are accessible from the query window, including several specific for microsatellite marker-based searches.

GIST is not meant to replace more thorough searches that can be performed at specific web sites. More importantly, GIST is not a database. It is simply a one-site query tool for multiple databases, having been designed to allow quick surveys of information available about a gene of interest and to get the “gist” of what is known about a gene or marker. GIST can be used for a variety of additional applications beyond building candidate gene lists, including performing simple searches of genes of interest in multiple databases and identifying microsatellite marker information such as heterozygosity and a marker’s prior use in genetic studies.

Candidate gene list. Using GIST, we have rapidly created an on-line resource for hypertension candidate genes. In preparing this resource, two aspects of list construction were important: the genes themselves and the information obtained for each. To build this list, genes were selected from a literature search based on several different lines of evidence. The strongest criteria used were human genetic studies (linkage and association analyses) and gene-specific transgenic or other animal models that have associated blood pressure-altering phenotypes. These two criteria account for only 30–40 genes (7, 15). We next selected genes from pathways regulating vasculature remodeling/repair, glucose homeostasis, cardiac function, kidney small molecule/ion transport, and neural and hormonal blood pressure pathways. The list we generated represents only a subset of all hypertension candidate genes, but as such it is the largest publicly available resource of genetic information focused on a specific complex genetic disease.

Nearly all of the information displayed at the web site was obtained from on-line databases. The one exception is the inclusion of hyperlinks to papers describing genetic studies of linkage or association for those genes in which studies have been attempted, obtained from a recent hypertension review article (15). The pace at which new gene information becomes available is remarkable. The data for this list were updated twice in an 18-month period after the initial list was created. Each time, ~20% more sequence, mapping, or expression information became available. Currently, 80% of all the data fields have been filled, and this percentage will certainly improve over the next several months because of the continuing ramp-up of the Human Genome Project sequencing effort (8). The information compiled for each gene is sufficient for projects designed to find polymorphic markers (SNPs) in gene regions and to compare genes relative to QTLs based on map position. However, none of this information allows hierarchical ranking of these genes as being more or less likely candidate genes for hypertension. That is, very little biologically relevant information relating to disease is available in genomic databases; but this is likely to change in the future. We predict that the information available from alternative expression pattern studies of these genes in hypertensive and normotensive subjects as determined by microarray systems (10) will allow the ranking of genes based on
the observed extent of difference between the two groups, as is currently being performed in the cancer community with the Cancer Genome Anatomy Project initiative.

We expect this on-line resource to be beneficial to both the hypertension research community and other complex genetic disease research communities. We believe this resource initiates a common starting point for gene discovery in hypertension. For other communities, this resource can serve as a template for additional candidate gene lists.

We thank Drs. Myriam Fornage and Perry Halushka for helpful discussions about hypertension genes, Crystal Rorbaugh for initial help in data collection, and Daniel Dickinson for computing assistance.

This study was supported by research funds from Case Western Reserve University, University Hospitals of Cleveland, and the National Heart, Lung, and Blood Institute as part of the GenNet Network of the Family Blood Pressure Program (U10-HL-54466). Address for reprint requests and other correspondence: A. Chakravarti, Dept. of Genetics BRB 721, Case Western Reserve Univ., 10900 Euclid Ave., Cleveland, OH 44106-4955 (E-mail: axc39@pop.cwru.edu).

REFERENCES


http://physiogenomics.physiology.org
