GeneRetriever: software to extract all genes and transcripts in between two genetic markers to assist design of human custom microarrays

Mathieu Clément-Ziza¹, Yehuda Brody², Arnold Munnich¹, Stanislas Lyonnet¹, and Claude Besmond¹

¹Hôpital Necker Enfants-Malades, Paris, France and ²Bar Ilan University, Ramat Gan, Israel

BioTechniques 39: 180-184(August 2005)

Identifying the genes that are responsible for human genetic disorders with complex inheritance patterns (e.g., multigenic diseases) has proven to be more difficult than anticipated. Indeed, when the mode of inheritance is unknown, classical parametric linkage studies are not relevant. Nonparametric linkage analysis can help approximate the candidate loci, but this type of analysis often leads to the identification of large intervals (10–20 cM) that may contain hundreds of genes, thus rendering the candidate gene approach rather tedious (1). Adding an expression screening may significantly reduce the number of candidate genes. Microarray expression studies of the genes located within such genetic intervals should be performed on relevant tissues (2). The design of a custom microarray containing probes of all the genes located in the intervals of interest is required to carry out such experiments.

Designing such microarrays involves gathering much additional information concerning these genes, in particular, name, transcript accession numbers, or nucleotide sequences. Collecting these data manually is time-consuming and very error-prone. GeneRetriever is a Perl-based data mining tool developed to automate, accelerate, and secure the process of locally retrieving user-chosen comprehensive information about human genes or transcripts located between two genetic markers.

As annotation strategies are specific for each database, we imple-

mented a database parameter entry that allows collection of data from either the National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov) or Ensembl (www.ensembl.org) databases (3,4). Then, several options make it possible to define which data should be included in the returned gene/transcript table. These options are clustered into three parts: (i) gene-specific data; (ii) transcript-specific data; and (iii) expression analysis data.

Gene-specific options include database (either NCBI or Ensembl)

identifier, Hugo Gene Nomenclature Committee symbol (5), gene description, DNA strand (plus or minus), type of gene (known or predicted), cytogenetic localization, summary of functional annotations, Entrez Gene identifier (6), web address of either NCBI or Ensembl gene page, and the number of transcript variants.

Additional transcript information is optionally available, such as Ensembl transcript or RefSeq accession numbers. Structural data, including transcript size, number of exons, and size of the longest exon, can also be added to the query. These data may be useful when working with predicted genes; for instance, the relevance of predicted genes composed of one single exon of <100 bp can be questioned. Nucleotide sequences can also be gathered; in addition, the retrieved size of the 3' end of transcript sequences can be user-defined in order to facilitate the design of transcript-specific oligonucleotides for microarray spotting.

Since GeneRetriever was developed to help the design of a custom gene expression study, the list of genes can be directly linked either to the available

Figure 1. GeneRetriever graphic-user interface. In this example, data are gathered from NCBI database between markers D3S1603 and D3S1544. The size of every transcript is retrieved, and transcripts longer than 100 nucleotides are analyzed.

GeneChip® probe sets (Affymetrix, Santa Clara, CA, USA) (7) or to the available TaqMan® Gene Expression Assay identifiers (Applied Biosystem, Foster City, CA, USA), depending on whether data are retrieved from

Ensembl or NCBI databases, respectively.

As all studies do not necessarily require exhaustive exploration of genes and transcripts, we implemented simple filtering options. Genes may be

filtered according to the type of gene (whether known only or predicted only), the strand (plus or minus strand), and the number of transcript variants. Transcripts can be filtered according to their size, the number of exons, and the availability of GeneChip probe sets or TaqMan Gene Expression Assay identifiers.

Entry parameters, output options, and filtering capabilities are selected using a graphic user interface (Figure 1) built on top of Perl/Tk. Data are retrieved from the NCBI databases via HTML requests using the LWP::Simple module to access web contents and transfer relevant information to the local machine. Powerful Perl pattern-matching capabilities are then put to task to parse them. Data are gleaned from Ensembl databases using the Ensembl Perl API (8).

Wherever the database data are gathered from, the retrieval follows the same strategy. GeneRetriever starts by retrieving the position of the genetic markers on the physical map. A list of unique database-specific identifiers of the genes located between the two user-chosen genetic markers is then collected. Each gene is then explored iteratively. If transcript data are requested, a list of transcript identifiers is retrieved during the gene data scanning, and each transcript is explored iteratively. Structural data of transcripts such as size, number of exons, and size of the longest exon are computed from the chromosomal coordinates extracted for each transcript. Finally, the sequence of each transcript is retrieved from the user-defined database (Figure 2). Buffering and output file writing operations are performed at each step of the execution of the program, so that system crashes or Internet connection failures do not affect already retrieved data.

GeneRetriever returns a text file containing tabulated data, which can be easily exported to any spreadsheet or database software, such as Microsoft® Excel® or FileMaker® Pro, if more filtering is needed. Collected GeneChip probes sets or TaqMan Gene Expression Assay identifiers can be piled and used for the design of a custom GeneChip or a TaqMan low density array, respectively. Finally, the collected sequences

Figure 2. GeneRetriever algorithm. For simplicity, only the main steps are presented. Different shapes have been used for each of the steps (see Chart key).

should be exported to an external program [i.e., OligoArray (9)] to design appropriate oligonucleotide probes (i.e., desirable secondary structure and uniqueness with respect to the entire genome) for microarray spotting. Even though some programs are able to perform similar extractions [i.e., Ensembl MartView (8) or RRE (10)], they do not offer the same functionalities as GeneRetriever, and some queries available in GeneRetriever are not addressed.

We have used GeneRetriever to collect the data needed to design a human custom GeneChip that contains all the genes and corresponding transcripts located between the markers D19S899 and D19S417. All options were checked, and no filtering was performed. This query successively launched at different times of the day takes a few minutes at most to complete. This procedure, especially when repeated with different sets of markers, as it is often the case in genetics laboratories, saves considerable time as compared with manually performing the same process or using distant resources.

ACKNOWLEDGMENTS

GeneRetriever is distributed under the GNU General Public License. The program, documentation, and source code are freely available at www.necker.fr/GeneRetriever. M.C.-Z. is a recipient of a fellowship from the French ministry of education and research. This study has been funded by grants from the Groupement d'Intérêt Scientifique (GIS) maladies rares and Association de Recherche contre le Cancer (ARC).

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES

1. Gabriel, S.B., R. Salomon, A. Pelet, M. Angrist, J. Amiel, M. Fornage, T. Attie-Bitach, J.M. Olson, et al. 2002. Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nat. Genet.* 31:89-93.
2. Wayne, M.L. and L.M. McIntyre. 2002. Combining mapping and arraying: An approach to candidate gene identification. *Proc. Natl. Acad. Sci. USA* 99:14903-14906.
3. Wheeler, D.L., T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, D.M. Church, M. DiCuccio, R. Edgar, et al. 2005. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 33(Database Issue):D39-D45.
4. Birney, E., T.D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, et al. 2004. An overview of Ensembl. *Genome Res.* 14:925-928.
5. Povey, S., R. Lovering, E. Bruford, M. Wright, M. Lush, and H. Wain. 2001. The HUGO Gene Nomenclature Committee (HGNC). *Hum. Genet.* 109:678-680.
6. Maglott, D., J. Ostell, K.D. Pruitt, and T. Tatusova. 2005. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 33(Database Issue):D54-D58.
7. Liu, G., A.E. Loraine, R. Shigeta, M. Cline, J. Cheng, V. Valmeekam, S. Sun, D. Kulp, and M.A. Siani-Rose. 2003. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.* 31:82-86.
8. Stabenau, A., G. McVicker, C. Melsopp, G. Proctor, M. Clamp, and E. Birney. 2004. The Ensembl core software libraries. *Genome Res.* 14:929-933.
9. Rouillard, J.M., M. Zuker, and E. Gulari. 2003. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.* 31:3057-3062.
10. Lazzarato, F., G. Franceschinis, M. Botta, F. Cordero, and R.A. Calogero. 2004. RRE: a tool for the extraction of non-coding regions surrounding annotated genes from genomic datasets. *Bioinformatics* 20:2848-2850.
1. Gabriel, S.B., R. Salomon, A. Pelet, M. Angrist, J. Amiel, M. Fornage, T. Attie-Bitach, J.M. Olson, et al. 2002. Segregation

Received 19 January 2005; accepted 25 April 2005.

Address correspondence to Claude Besmond, Hôpital Necker, Inserm U393, Tour Lavoisier, 2nd floor, 149 rue de Sèvres, 75015 Paris, France. e-mail: besmond@necker.fr

To purchase reprints
of this article, contact
apfeffer@BioTechniques.com