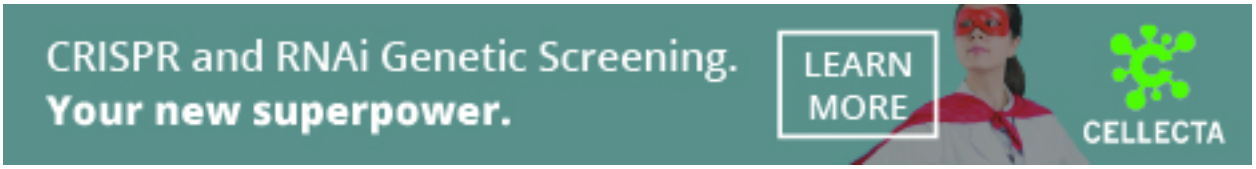
GeneLynx Mouse: Integrated Portal to the Mouse Genome

Boris Lenhard, Claes Wahlestedt and Wyeth W. Wasserman

Genome Res. 2003 13: 1501-1504
Access the most recent version at doi:10.1101/gr.951403



GeneLynx Mouse: Integrated Portal to the Mouse Genome

Boris Lenhard,[1,3] Claes Wahlestedt,[1] and Wyeth W. Wasserman[2]

[1]*Center for Genomics and Bioinformatics, Karolinska Institutet, 17177 Stockholm, Sweden;* [2]*Centre for Molecular Medicine and Therapeutics, University of British Columbia, Vancouver B.C., V5Z 4H4, Canada*

GeneLynx Mouse is a meta-database providing an extensive collection of hyperlinks to mouse gene-specific information in diverse databases available via the Internet. The GeneLynx project is based on the simple notion that given any gene-specific identifier (e.g., accession number, gene name, text, or sequence), scientists should be able to access a single location that provides a set of links to all the publicly available information pertinent to the specified gene. The recent climax in the mouse genome and RIKEN cDNA sequencing projects provided the data necessary for the development of a gene-centric mouse information portal based on the GeneLynx ideals. Clusters of RIKEN cDNA sequences were used to define the initial set of mouse genes. Like its human counterpart, GeneLynx Mouse is designed as an extensible relational database with an intuitive and user-friendly Web interface. Data is automatically extracted from diverse resources, using appropriate approaches to maximize the coverage. To promote cross-database interoperability, an indexing utility is provided to facilitate the establishment of hyperlinks in external databases. As a result of the integration of the human and mouse systems, GeneLynx now serves as a powerful comparative genomics data mining resource. GeneLynx Mouse can be freely accessed at http://mouse.genelynx.org.

Prolific genome and transcript sequencing efforts have motivated diverse bioinformatics efforts to organize and analyze gene-specific data. Gene indices have emerged as an efficient means to deliver relevant data to laboratory scientists on the frontlines of biomedical research. By collating gene-specific data into a cohesive presentation, gene portals expand and accelerate data access for biologists. As with general-purpose Internet portals (e.g., Yahoo, http://www.yahoo.com), a broad range of indices has emerged catering to diverse research communities. The most mature systems focus on well-studied species, including humans (GeneCards, GeneLynx) and diverse model organisms (MGI, FlyBase, SGD, AceDB; Table 1). More recently, a number of generically implemented systems have been introduced to support access to data from multiple species (LocusLink, SWISS-PROT, TIGR, EuGenes, Ensembl, AllGenes, SOURCE). Research and development of gene indices is an integral component of efforts to capitalize on the biomedical opportunities created by the successful genome projects.

Research related to gene indices attempts to expand user access to data, improve visualization and analysis procedures, and deliver efficient performance. The GeneLynx project (Lenhard et al. 2001) initially produced a gene-centric portal to the human genome with an interface and functionality that emphasized ease of use and rapid access to gene-specific data in numerous, but highly diverse and disparate Web resources. The power of a gene index such as GeneLynx is inherently dependent on the availability of nucleotide sequence information. Rapid progress in the sequencing and analysis of the mouse genome (Waterston et al. 2002) has resulted in an increased demand for GeneLynx-like access to mouse gene information. From a biomedical research perspective, the principal challenge in developing such a resource is effectively integrating human and model organisms gene data in a manner that illuminates ideas and motivates further investigation.

In this article, we present GeneLynx Mouse, a gene-centric portal based on the Fantom2 collection of 60770 mouse full-length cDNAs. In addition to the popular functionality of the human portal, GeneLynx Mouse introduces a novel set of features to facilitate comparisons between orthologous mouse and human genes.

RESULTS AND DISCUSSION

Gene-Based Clusters

GeneLynx Mouse (release 1.0 RC1) contains 37,086 clusters corresponding to RIKEN representative transcripts set, each of which aims to correspond one transcriptional unit. The clusters are based on the nonredundant mouse cDNA set drawn from the Fantom2 RIKEN full-length cDNA collection. In contrast to the human GeneLynx system, no cluster rearrangements are introduced in the semiautomated GeneLynx curation process. Rare conflicts identified in the quality control process are incorporated into the RIKEN clustering efforts for future releases.

Although an overwhelming majority of the clusters qualify as gene-based, a certain number of RIKEN cDNAs has been labeled "unclassifiable" by the curators. We retained those cDNAs and the corresponding single-member clusters. With the completion of the curation process, those clusters will be put in a separate category in GeneLynx, reducing the total number of gene-based clusters.

Infrastructure and User Interface

To maintain consistency for users, the interface of GeneLynx Mouse is essentially identical to its human counterpart (Lenhard et al. 2001). It includes text- and sequence-based search

[3]Corresponding author.
E-MAIL Boris.Lenhard@cgb.ki.se; FAX 46-8-32-48-26.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.951403.

Table 1. Leading Gene Indices and Related Resources on the Internet

Resource	Organism	URL	Reference
GeneCards	human	http://bioinfo.weizmann.ac.il/cards/	Rebhan et al. 1997
GeneLynx	human, mouse	http://www.genelynx.org	Lenhard et al. 2001
AllGenes	human, mouse	http://www.allgenes.org	
Standard SOURCE	human, mouse, rat	http://source.stanford.edu	
MGI	mouse	http://www.informatics.jax.org/mgihome	
FlyBase	fruit fly	http://www.flybase.org	FlyBase Consortium, 2003
SGD	*Saccharomyces cerevisiae*	http://genome-www.stanford.edu/Saccharomyces	Cherry et al. 1998
LocusLink	multiple	http://www.ncbi.nlm.nih.gov/LocusLink	Pruitt and Maglott 2001
SWISS-PROT	multiple	http://www.expasy.org/sprot	O'Donovan et al. 2002
EuGenes	multiple	http://iubio.bio.indiana.edu:8089/	Gilbert 2002
TIGR Gene Indices	multiple	http://www.tigr.org/tdb/tgi	Blake et al. 2002
Ensembl	multiple	http://www.ensembl.org/	Hubbard et al. 2002
UCSC Genome Bioinformatics	human, mouse, rat	http://genome.ucsc.edu	Kent et al. 2002

engines, batch retrieval of identifiers, communal curation tools, and a standardized protocol for submission of new resources. For the latter, it is now possible to submit a new general or species-specific resource. The addition of new resources is deliberately simple and lends itself to rapid incorporation of new data resources as they become available.

Links to External Resources

The set of hyperlinks to gene-specific databases is expanded by an iterative procedure. Newly captured gene-specific identifiers often enable access to additional databases. As the list of synonymous terms extends, the collection of gene-specific links nears the maximum. The annotation data produced for the Fantom2 mouse full-length cDNA set (Okazaki et al. 2002), were also incorporated, as the majority (two-thirds) of the cDNAs are not yet cross-referenced in other public databases. The linking statistics are presented in Table 2. We expect the number of links to increase dramatically once the new Fantom2 cDNAs are released into global sequence databases and incorporated into other bioinformatics projects.

Data Integration for Comparative Genomics

The greatest opportunities presented by the new mouse data are in the area of comparative genomics. By integrating the human and mouse GeneLynx systems, we have produced a unique resource for the investigation of similarities and differences between orthologous genes. Several resources and approaches were combined to match mouse genes to putative human orthologs:

1. Curated ortholog lists were obtained from the Mouse Genome Informatics Web site (http://www.informatics.jax.org). These data were directly incorporated as a new resource for the human and mouse systems. The curated orthology links were sufficient to map 5485 gene pairs (out of a total of 7450), using only official and unambiguous unofficial gene symbols. Methods for matching curated gene pairs based on other types of identifiers are under development.
2. Cross-genome mapping was performed with the fast BLAT sequence comparison tool (Kent 2002); in this approach, we used both standard and translated BLAT searches of the human genome assembly with the mouse cDNA collection, and detected each cDNA that uniquely mapped to a specific human gene. This analysis identified an additional 5144 orthology relationships.
3. Because BLAST is more sensitive than is BLAT, we performed a reciprocal BLAST search. The mouse cDNAs were compared with the human genome to identify the most similar segment. For those segments annotated as human genes, the corresponding human cDNA sequence was compared with the mouse genome. For those cases in which the last search maps to the same mouse locus as the original mouse DNA, we considered this "best reciprocal match" of the human and mouse genes as sufficient to

Table 2. Linking Statistics for the Explicitly Linked Resources in GeneLynx Mouse 1.0 RC1

Resource	Number of items linked to GeneLynx	Number of linked GeneLynx records
RIKEN cDNAs	60,746	33,409
UniGene, CGAP	16,414	16,052
cDNAs (GB/EMBL/DDBJ)	44,102	13,686
KEGG (genes)	7,624	7,657
KEGG (pathways)	77	791
Genomic sequences	8,842	2,984
LocusLink	12,045	10,863
Ensembl (genes)	12,865	12,930
Ensembl (transcripts)	15,211	12,930
RefSeq (NM)	18,963	10,376
MGI (MGD)	19,729	17,638
SWISS-PROT	4,881	4,972
TrEMBL	18,970	12,190
PIR	4,035	2,648
GenPept collection	38,600	13,968
PDB	308	140
HSSP	1,155	4,011
InterPro	2,241	10,452
PRINTS	650	3,744
BLOCKS/PRODOM	1,965	2,395
PFAM	1,746	8,395
SBASE	79,997	2,857
PROSITE	1,086	7,107
ENZYME DB/WIT/Brenda	432	857
Homologs (UniGene)	14,016	14,265
Homologs (LocusLink)	6,362	5,091
Homologs (nucleotide seqs)	73,600	24,193
Homologs (protein seqs)	14,991	11,469
ESTs	1,745,916	14,905

define an orthologous pair. We used nucleotide BLAST on masked sequences, and eliminated those cases in which the e-value for the reciprocal match was greater than 1e-50. In a small number of cases, this assumption of orthology will be incorrect, but curation efforts will continue to refine the list of orthologs as new information becomes available.

To facilitate comparisons between the orthologs, a new comparative view of links is introduced for the GeneLynx system (Fig. 1). We strove to retain the look and feel of the original GeneLynx record page that users found readable and easy to use, adopting a two-column side-by-side format, color-coded for easier orientation and extensible to eventually accommodate additional species. It enables users to compare the knowledge bases for orthologous genes and to identify missing information. For instance, protein domains or gene ontology terms for a gene can be inferred by reading them directly from the annotations of the better-annotated ortholog. This feature, to our knowledge, is unique to GeneLynx. A comparison of GeneLynx features with those of related mouse gene indices is available at http://www.genelynx.org/MOUSE/TECHNICAL/.

Future Developments

We expect to see an increase in the use of GeneLynx as a comparative genomics portal. To that end, future development must emphasize closer coupling with comparative genomics analysis tools. For a specific example, one should be able to launch utilities for pair-wise comparisons, which would be invoked with the desired pairs of orthologous sequences (genomic, cDNA, or protein) from the pair-wise GeneLynx view. We are in the process of coupling it with our tool for the comparative analysis of functional elements in regulatory regions in genes (Consite, www.phylofoot.org/consite), as well as other successful comparative genomics analysis resources.

METHODS

Linking GeneLynx Clusters to External Resources

The gene-centric clusters are based on the clusters produced by the RIKEN team on the Fantom2 set of full-length mouse cDNAs. GeneLynx Mouse 1.0 uses the cluster freeze of July 24, 2002, the one used for analyses in the Fantom2 report (Okazaki et al. 2002).

During the development of the resource, a test build was performed without any intervention to locate possible problems in the establishment of links. The majority of problems were the obvious violations of the one-to-one correspondence between the Fantom2 clusters and other gene-centric or gene-product–centric resources. Most of these problems involve mouse cDNA sequences labeled as "unclassifiable" during the human curation of the set. The unclassifiable sequences either could not be mapped to the mouse genome assembly or have

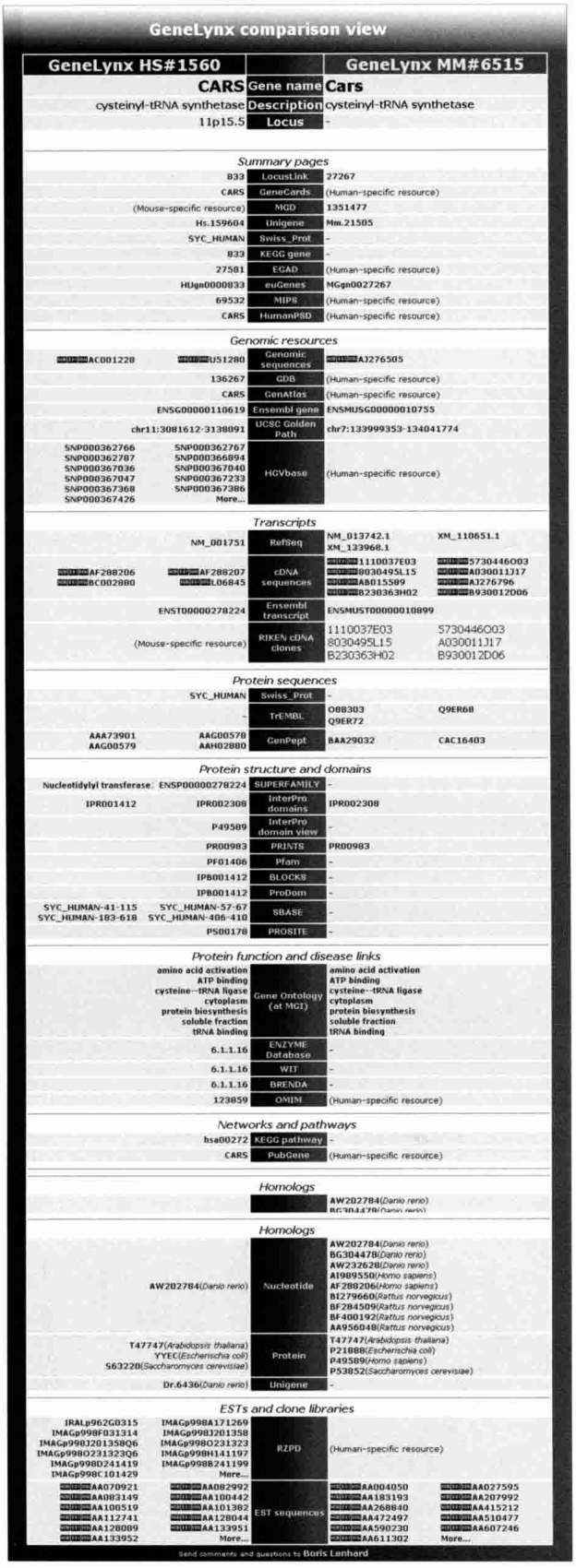


Figure 1 GeneLynx comparison view: an example of the side-by side view of the GeneLynx entries for human and mouse cysteinyl-tRNA synthetase. This case demonstrates the utility of viewing orthologous genes in parallel, as relevant information can be obtained for a more thoroughly studied species. Mouse CysRS does not have a SWISS-PROT entry, and its protein structure and function links are therefore also lacking. However, the missing information is available for the human ortholog and is directly applicable to the mouse gene.

no sensible open reading frame; or contain repeats, transposons, or fusion products.

We expect those clusters to be further modified by expert curation, and subsequent changes will be incorporated in future builds of GeneLynx Mouse.

Limited Cluster Curation

The semiautomated curation process originally applied to the curation of human GeneLynx clusters (Lenhard et al. 2001) was modified to take advantage of the availability of human and mouse genome assemblies. All cDNA sequences were mapped to the corresponding genome assembly by using BLAT, and were examined closely if they do not map unambiguously to a single locus in the genome. In the case of GeneLynx Mouse, we have chosen not to rearrange the curated Fantom2 clusters; instead, we plan to include the modifications made to the official Fantom2 clustering schema into future builds of GeneLynx. The clusters that we considered problematic were flagged as such and were excluded from the downstream process of link creation. They are still available via sequence and keyword search mechanisms, but instead of a full set of external resources, they contain a warning, possibly with the curator's note describing the problem.

Platform and Availability

The resource back-end is written in object-oriented Perl (Perl 5.6.1; Wall et al. 2000), with extensive use of BioPerl modules (Stajich et al. 2002). The data is stored in a relational database (MySQL 3.23.51, http://www.mysql.com). BLAT (Kent 2002) was used in its client/server implementation (gfclient/gfServer) and run via a set of in-house OO Perl adapter modules (B. Lenhard and P. Engström, unpubl.). For BLAST searches, we currently use NCBI Blast 2.0.14 (available at ftp://ftp.ncbi.nlm.nih.gov), using BLASTALL (BLASTN or TBLASTN) with default parameters unless noted otherwise. The GeneLynx Mouse database files are available on request. XML files will become publicly available at a later date, when the curated clusters stabilize.

Genome Assemblies

For all analyses and data integration purposes, we used the most recent public human and mouse genome assemblies, as available at http://genome.ucsc.edu: human assembly hg12 (June 2002), and mouse assembly mm2 (February 2002). Newer assemblies will be used as soon as they become available.

ACKNOWLEDGMENTS

We are indebted to Yoshihide Hayashizaki for initiating the collaboration between the RIKEN Genome Exploration Research Group and the Center for Genomics and Bioinformatics, and to Yasushi Okazaki for valuable advice on using the RIKEN data. This project was supported by funds from the Karolinska Institutet and the Pharmacia Corporation.

REFERENCES

Blake, J.A., Richardson, J.E., Bult, C.J., Kadin, J.A., and Eppig, J.T. 2002. The Mouse Genome Database (MGD): The model organism database for the laboratory mouse. *Nucleic Acids Res.* **30:** 113–115.

Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., et al. 1998. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* **26:** 73–79.

FlyBase Consortium. 2003. The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.* **31:** 172–175.

Gilbert, D.G. 2002. euGenes: A eukaryote genome information system. *Nucleic Acids Res.* **30:** 145–148.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30:** 38–41.

Kent, W.J. 2002. BLAT: The BLAST-like alignment tool. *Genome Res.* **12:** 656–664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

Lenhard, B., Hayes, W.S., and Wasserman, W.W. 2001. GeneLynx: A gene-centric portal to the human genome. *Genome Res.* **11:** 2151–2157.

O'Donovan, C., Martin, M.J., Gattiker, A., Gasteiger, E., Bairoch, A., and Apweiler, R. 2002. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform.* **3:** 275–284.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.

Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene–centered resources. *Nucleic Acids Res.* **29:** 137–140.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. 1997. GeneCards: Integrating information about genes, proteins and diseases. *Trends Genet.* **13:** 163.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., et al. 2002. The bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12:** 1611–1618.

Wall, L., Christiansen, T., and Orwant, J. 2000. *Programming Perl.* O'Reilly & Associates, Sebastopol, CA.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

WEB SITE REFERENCES

http://mouse.genelynx.org; GeneLynx Mouse.

http://www.informatics.jax.org; Mouse Genome Informatics.

http://www.genelynx.org/MOUSE/TECHNICAL/; Technical info for GeneLynx Mouse.

http://www.phylofoot.org/consite; Web site for detecting transcription factor binding sites in genomic sequences using phlogenetic footprinting.

http://www.mysql.com; MySQL database homepage.

http://genome.ucsc.edu: human assembly hg12; UCSC Genome Informatics Web site (genome browsers, BLAT search, sequence, and annotation downloads).

Received November 5, 2002; accepted in revised form February 14, 2003.