RefSeq and LocusLink: NCBI gene-centered resources

Kim D. Pruitt* and Donna R. Maglott

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health,
Building 38A Room 6N605, 8600 Rockville Pike, Bethesda, MD 20894 USA

Received October 2, 2000; Accepted October 4, 2000

ABSTRACT

Thousands of genes have been painstakingly identified and characterized a few genes at a time. Many thousands more are being predicted by large scale cDNA and genomic sequencing projects, with levels of evidence ranging from supporting mRNA sequence and comparative genomics to computing *ab initio* models. This, coupled with the burgeoning scientific literature, makes it critical to have a comprehensive directory for genes and reference sequences for key genomes. The NCBI provides two resources, LocusLink and RefSeq, to meet these needs. LocusLink organizes information around genes to generate a central hub for accessing gene-specific information for fruit fly, human, mouse, rat and zebrafish. RefSeq provides reference sequence standards for genomes, transcripts and proteins; human, mouse and rat mRNA RefSeqs, and their corresponding proteins, are discussed here. Together, RefSeq and LocusLink provide a non-redundant view of genes and other loci to support research on genes and gene families, variation, gene expression and genome annotation. Additional information about LocusLink and RefSeq is available at http://www.ncbi.nlm.nih.gov/LocusLink/.

BACKGROUND

LocusLink maintains descriptive information about loci including nomenclature, database identifiers (ID), disease associations, map positions and sequence accessions. These associations are then used to compute connections to other NCBI resources, to both facilitate navigation and enhance opportunities for new discoveries. LocusLink data are continuously augmented and reviewed through the combined efforts of NCBI staff and successful collaborations with the Human Gene Nomenclature Committee (HGNC) (1), Online Mendelian Inheritance in Man (OMIM) (2), Mouse Genome Database (MGD) (3), Rat Genome Database (RGD), FlyBase (4) and Zebrafish (ZFIN) (5) groups. This compilation is freely distributed via the LocusLink web site and by FTP. LocusLink queries can be based on text terms (such as a protein or disease name), gene symbols, sequence accessions and database IDs (e.g., MIM or EC numbers). More complex operations such as wild cards (e.g., ABC*), field restriction (e.g., 123456[mim]) and booleans (e.g., ABC* AND 17[chr]) are also supported.

The NCBI reference sequences (RefSeq) provide standards for complete genomic nucleic acids, assembled contigs, transcripts and proteins (see Table 1). RefSeq records are derived from GenBank and the literature to provide a non-redundant set of sequences that facilitate sequence identification and information retrieval (6). Human, mouse and rat RefSeq mRNAs, and their corresponding proteins, are the focus of this discussion; these sequences can be accessed through the NCBI Entrez retrieval system (7), BLAST (8), FTP and LocusLink.

EXPANDED AND NEW FEATURES

LocusLink

The locus-to-sequence association maintained in LocusLink is used to provide reciprocal connections among LocusLink, RefSeq and several content-rich resources at NCBI including the Conserved Domain Database (CDD), GenBank (9), Genes and Disease, HomoloGene, Map Viewer, OMIM, PubMed, dbSNP (10), dbSTS and UniGene (11) (Table 2). In addition, nucleotide and protein sequences and OMIM records are linked to LocusLink reports via the NCBI LinkOut service. Following ‘LinkOut’ at the upper right side of an Entrez record is a convenient method to access a wealth of information about that gene. Links between LocusLink and Genes and Disease, HomoloGene, RefSeq, dbSNP and UniGene are based on shared identifiers stored by these resources (e.g., MIM numbers or GenBank accessions). Links to dbSTS and CDD are computed based on electronic PCR (ePCR) (12) and BLAST, respectively. And links to the NCBI Map Viewer are based on sequence identity, cytogenetic location or genetic map data. URLs to these resources are provided in Table 3.

Several organisms and features have been added to the LocusLink web site over the past year and are listed below. LocusLink provides links to each resource whenever additional information is available.

Conserved domains: CDD domains annotated on RefSeq proteins are summarized in the product section of LocusLink reports. This section also includes the protein name, alternate names and enzyme commission (EC) numbers.

GeneRIF: LocusLink now supports community-generated functional annotation. LocusLink reports display submitted GeneRIFs (Gene References Into Function) and include a link to the submission form. Submissions consist of a publication

*To whom correspondence should be addressed. Tel: +1 301 435 5898; Fax: +1 301 480 2290; Email: pruitt@ncbi.nlm.nih.gov

Table 1. RefSeq accessions, sequence type, processing method and categories

Accession format	Type	Method	Category
NC_123456	Genomic	Curated	Genomic molecules, available in Entrez Genomes (mitochondrion, viral and bacterial genomes, chromosomes)
NT_123456	Genomic	Assembled contigs	Genome annotation
NM_123456	mRNA	Computed	Predicted
		Curated	Provisional
		Curated	Reviewed
NG_123456	Genomic	Curated	Gene region
NP_123456	Protein	Computed; curated	Full-length proteins associated with curated nucleotide sequences
XM_123456	mRNA	Gene prediction	Genome annotation
XP_123456	Protein	Gene prediction	Genome annotation

Table 2. LocusLink: distributed data and links to additional information

Data type	Collection method	Source
Official gene symbol	Curated	Nomenclature
Official gene name	Curated	Nomenclature
GenBank accessions	Curated	Nomenclature, OMIM, NCBI
Protein names	Curated	Nomenclature, OMIM, NCBI
Map position	Computed; curated	Genome specific databases, OMIM, NCBI
Alias symbols	Curated	Nomenclature, OMIM, NCBI
Phenotype	Curated	OMIM
NCBI Resource Links:		
CDD	Computed	NCBI
dbSTS	Computed	NCBI
dbSNP	Computed	NCBI
Genes and Disease	Computed	NCBI
HomoloGene	Computed; curated	NCBI
Homology Map	Computed; curated	MGD, NCBI
Map Viewer	Computed; curated	NCBI
OMIM	Curated	OMIM
PubMed	Computed; curated	Nomenclature, OMIM, NCBI
RefSeq	Computed; curated	NCBI
UniGene	Computed; curated	NCBI
External links[a]:		
GDB	Computed	
GeneCards	Computed	
GeneClinics	Computed	
Gene family resources	Curated	
Mutation web sites	Curated	
Nomenclature web sites	Curated	
LinkOut	Submitted links	NCBI, research community
GeneRIF	Submitted	Research community

[a]External links are based on suggestions from our collaborators, the research community, ongoing review of available resources by NCBI staff and computational analysis to identify common identifiers (e.g., GDB links).

Table 3. URLs for NCBI resources

NCBI resource	URL
CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
Homology Map	http://www.ncbi.nlm.nih.gov/Homology/
FTP: LocusLink	ftp://ncbi.nlm.nih.gov/refseq/LocusLink/
FTP: RefSeq mRNA/protein	ftp://ncbi.nlm.nih.gov/refseq/
GenBank	http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html
Genes and Disease	http://www.ncbi.nlm.nih.gov/disease/
HomoloGene	http://www.ncbi.nlm.nih.gov/HomoloGene/
LinkOut	http://www.ncbi.nlm.nih.gov/entrez/query/static/linkoutoverview.html
LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
Map Viewer (Human)	http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch?
OMIM	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
PubMed	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
RefSeq	http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
dbSNP (Variation)	http://www.ncbi.nlm.nih.gov/SNP/
dbSTS	http://www.ncbi.nlm.nih.gov/dbSTS/index.html
UniGene	http://www.ncbi.nlm.nih.gov/UniGene/

(PubMed ID) and brief comment summarizing the functional information supported by the publication.

HomoloGene: HomoloGene reports UniGene clusters that are either calculated to be homologous, based on sequence identity, or reported to be homologous. HomoloGene also reports the 'best' homologous sequence pairs. HomoloGene includes all genomes in UniGene (human, mouse, rat and zebrafish), as well as *Drosophila melanogaster* mRNAs.

Maps: LocusLink now provides links to the Map Viewer and the Human/Mouse Homology map, as appropriate. The Map Viewer presents graphical views of various maps at the whole genome and single chromosome levels. This resource is currently available for *D.melanogaster*, *Homo sapiens* and *Mus musculus*. Views include presentation of genes, STS markers, SNPs and disease phenotypes along the chromosomes. The Human/Mouse Homology maps display genes, thought to be orthologous, ordered by MGD's genetic map and the human sequence map.

Organisms: LocusLink now includes fruit fly, human, mouse, rat and zebrafish genes. The query interface supports retrieving results for all species, restricting to a single organism and displaying query results as a 'Brief' or expanded 'Summary' display.

Reference sequences: The RefSeq section of the LocusLink report has been expanded to include mitochondrial records, genomic contigs and modeled transcripts and proteins produced by the NCBI Human Genome Annotation effort. RefSeq data for a locus may include curated RefSeq records representing known transcripts, a genomic contig and modeled transcripts that overlap, but are distinct from, the known transcripts. Table 1 indicates the different accession number formats and categories they represent.

RefSeq

RefSeq transcript and corresponding protein records are now produced through two independent processes. The first process, which relies heavily on manual curation, provides a non-redundant dataset of transcripts and proteins of known genes for human, mouse and rat; these 'known' genes have some supporting evidence for the existence of the gene although the protein function may not be clear. The second process is NCBI's Human Genome Annotation project, which will be described in detail elsewhere. Briefly, working draft and finished human genome sequence is assembled into contigs (13); gene predictions methods are used to produce the modeled mRNA and protein RefSeq records (see Table 1). The modeled transcripts and proteins may have different degrees of mRNA or EST evidence, or may be predicted *ab initio*.

RefSeq records include links to LocusLink where additional information is often available. The organisms and features added to RefSeq records over the past year are summarized below.

Category: RefSeq mRNA and protein records are now provided in four categories (previously termed 'Status' in LocusLink reports): (i) genome annotation, (ii) predicted, (iii) provisional and (iv) reviewed. The genome annotation category includes contigs, modeled mRNAs and corresponding modeled proteins. Predicted records represent genes of unknown function that are supported by mRNA sequence generated in a full-length insert cDNA sequencing project, homology or ESTs; the location of the protein product may be predicted in these records. Provisional and reviewed RefSeq records largely represent genes with known or inferred function. Provisional RefSeqs are not yet reviewed, whereas reviewed RefSeq records have been individually reviewed by NCBI staff and are richly annotated. Reviewed records include publications, a gene description and annotation that may not be on the provisional RefSeq or original GenBank records. Furthermore

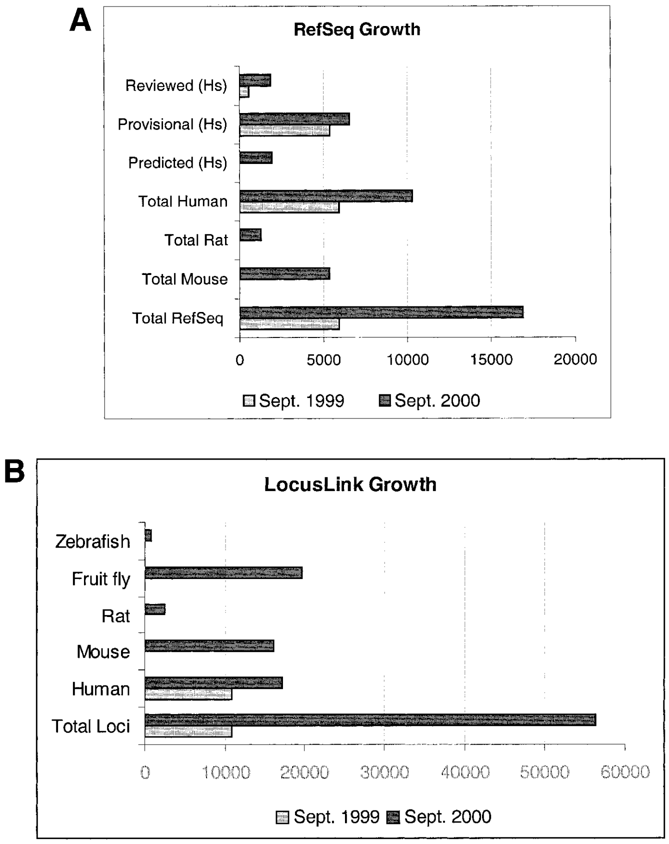

Figure 1. RefSeq and LocusLink growth over a 1-year period. (**A**) The number of RefSeq records for human, mouse and rat for mid-September of 1999 and 2000. These numbers include genes for which multiple reference sequences are provided to represent splice variants and their products. As of September 2000, 10 301 human reference sequence mRNAs (and their corresponding proteins) are provided for 9954 genes. The number of human RefSeq records in three categories is also indicated (Hs, human); Genome Annotation reference sequence numbers are not included. Rat, mouse and predicted human RefSeq records became available after September 1999. (**B**) Loci available in LocusLink for September of 1999 and 2000. Mouse, rat, zebrafish and fruit fly genomes were added after September 1999. Of the 17 214 known human loci, 14 103 have some associated sequence data (including ESTs), and 9954 loci have at least one mRNA and protein reference sequence.

the sequence may be modified, relative to the source GenBank sequence, to extend the UTR, provide a transcript variant or remove contaminating vector or linker sequence. By including sequence data derived from more than one GenBank record and from the literature, reviewed RefSeq records may provide a full-length mRNA sequence that is not available in a single GenBank record.

Conserved domains: BLAST analysis is used to compare RefSeq proteins to the NCBI CDD; significant domain matches are annotated on the protein record and are linked to the CDD resource where additional information is available. Reviewed RefSeq proteins may include additional domain annotation derived from the literature.

Links: RefSeq records include links that facilitate access to additional information. Links are provided to: (i) the record(s) from which the mRNA (and hence protein) was derived, (ii) LocusLink, (iii) OMIM, (iv) MGD, (v) RGD and (vi) CDD, as appropriate. In addition, an overview of sequences that share sequence identity is available through the 'Related sequences' link at the top right of RefSeq (and GenBank) records.

Organisms: The mRNA/protein RefSeq resource now provides sequence standards for human, mouse and rat. Review of all three organisms to verify the initial 'gene name-to-sequence' association and to provide highly annotated reviewed RefSeq records is ongoing.

Variation: Variation features are automatically added to Genome Annotation RefSeq records using the data available in dbSNP. This expanded annotation will also be added to RefSeq records representing known genes.

GROWTH AND MAINTENANCE

The information available in LocusLink and RefSeq is continually being reviewed and augmented. As illustrated in Figure 1, the number of RefSeq records tripled, and LocusLink expanded ~5-fold over a 1-year period. New or updated RefSeq records are made publicly available on a daily basis. The LocusLink web site is updated weekly; a subset of data on the FTP site updated daily and the remainder is updated weekly.

REFERENCES

1. White,J.A., McAlpine,P.J., Antonarakis,S., Cann,H., Eppig,J.T., Frazer,K., Frezal,J., Lancet,D., Nahmias,J., Pearson,P., Peters,J., Scott,A., Scott,H., Spurr,N., Talbot,C.,Jr. and Povey,S. (1997) Guidelines for human gene nomenclature (1997). HUGO Nomenclature Committee. *Genomics*, **45**, 468–471.
2. Hamosh,A., Scott,A.F., Amberger,J., Valle,D. and McKusick,V.A. (2000) Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.*, **15**, 57–61.
3. Blake,J.A., Eppig,J.T., Richardson,J.E. and Davisson,M.T. (2000) The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. The Mouse Genome Database Group. *Nucleic Acids Res.*, **28**, 108–111. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 91–94.
4. FlyBase Consortium (1999) The FlyBase database of the Drosophila Genome Projects and community literature. The FlyBase Consortium. *Nucleic Acids Res.*, **27**, 85–88.
5. Westerfield,M., Doerry,E., Kirkpatric,A.E. and Douglas,S.A. (1999) Zebrafish informatics and the ZFIN database. *Methods Cell Biol.*, **60**, 339–355.
6. Pruitt,K.D., Katz,K.S., Sicotte,H. and Maglott,D.R. (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.*, **16**, 44–47.
7. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
8. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
9. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.
10. Sherry,S.T., Ward,M. and Sirotkin,K. (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.*, **8**, 677–679.
11. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
12. Schuler,G.D. (1997) Sequence mapping by electronic PCR. *Genome Res.*, **7**, 541–550.
13. Jang,W., Chen,H.C., Sicotte,H. and Schuler,G.D. (1999) Making effective use of human genomic sequence data. *Trends Genet.*, **15**, 284–286.