126–128 *Nucleic Acids Research, 2000, Vol. 28, No. 1*

NCBI's LocusLink and RefSeq

Donna R. Maglott, Kenneth S. Katz, Hugues Sicotte and Kim D. Pruitt*

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

Received September 2, 1999; Revised and Accepted October 4, 1999

ABSTRACT

The NCBI has introduced two new web resources—LocusLink and RefSeq—that facilitate retrieval of gene-based information and provide reference sequence standards. These resources are designed to provide a non-redundant view of current knowledge about human genes, transcripts and proteins. Additional information about these resources is available on the LocusLink web site at http://www.ncbi.nlm.nih.gov/LocusLink/

BACKGROUND

The LocusLink and RefSeq databases were initiated to address data-access problems resulting from significant increases in both sequence data and the number of web sites relating information about genes. For example, it is increasingly difficult to identify unambiguously which sequence—of the many publicly available—is an appropriate, complete representative of a given mRNA or protein. Inversely, given an mRNA or protein sequence, it can also be a challenge to determine the official name or symbol for the gene from which the sequence was derived. And once a gene symbol or name is known, identifying other web resources that include information about that gene of interest may be very time-consuming. In its role as a web directory, LocusLink provides a single point-of-access to a variety of gene-specific information sources including web resources and RefSeq. RefSeq provides a non-redundant data set of reference sequences representing transcripts and proteins of known genes. RefSeq records include links to LocusLink, thereby facilitating making connections among sequence data, gene names and related biological information. The LocusLink and RefSeq resources establish reference sequences and stable database identifiers (LocusID) that can be used in variation, mutation and expression analyses.

SCOPE

LocusLink

LocusLink offers a simple query interface to retrieve information about human genes and some non-gene loci. It supports text-based queries by using official nomenclature provided through collaboration with the Human Gene Nomenclature Committee (HGNC; http://www.gene.ucl.ac.uk/nomenclature/) (1), as well as cytogenetic locations, aliases and historical names for both a gene and its products. LocusLink provides direct connections to related information available from several resources at NCBI (Table 1) as well as to external web sites including the Genome Database (GDB; http://gdbwww.gdb.org/), the Human Gene Mutation Database (HGMD; http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html) (2), GeneCard (http://bioinfo.weizmann.ac.il/cards/), GeneClinics (http://www.geneclinics.org/), and locus- or gene family-specific web sites. Some of the links to NCBI resources listed in Table 1 are represented by icons that, when displayed, give an immediate indication that additional information is indeed available. The goal of the PubMed and GenBank/GenPept (3) links is not to be comprehensive, but to establish sufficient connections to facilitate information retrieval via NCBI's ENTREZ (4) 'related sequences' or 'related publications' links or through BLAST (5). LocusLink also provides a unique stable identifier for each locus (LocusID).

RefSeq

Although the goal of RefSeq in general is to provide reference sequences representing chromosomes, transcripts and proteins, discussion here is restricted to the subset of human mRNAs and proteins. A RefSeq record is made for an mRNA if the function of the gene product has been studied, and if the sequence of the complete coding region is known. Separate RefSeq records are made for experimentally supported alternate transcripts and their products. The sequence presented in a RefSeq record is usually derived from available GenBank records, although additional information is at times added from the literature or from communications with the research community. RefSeq records are provided in one of two states, either provisional or reviewed. Records initially released as provisional include much of the annotation from the GenBank record used as the source, but incorporate gene and protein names, PubMed links, summary text, and map and chromosome data from LocusLink when available (Table 2). Provisional records are subjected to a manual curation and review process, with the reviewed record being the end product. The reviewed record might differ from the original provisional record by including: (i) more extensive 5′ and 3′ untranslated regions derived from other GenBank records or the literature, (ii) additional mRNA and/or protein features, (iii) more publications and (iv) a summary text describing the gene. Table 2 lists additional annotation that may be added to provisional and reviewed RefSeq records. RefSeq records can be distinguished from GenBank records by the inclusion of a REFSEQ statement in a COMMENT field, and by the unique format of the accession number. The first three characters of the RefSeq mRNA and protein accession numbers are NM_ and NP_, respectively, followed by six numerals (e.g. NM_000280, NP_000337).

*To whom correspondence should be addressed. Tel: +1 301 435 5898; Fax: +1 301 480 9241; Email: pruitt@ncbi.nlm.nih.gov

Table 1. LocusLink connections to resources at NCBI

Resource	Icon[a]	Information provided
dbSNP	V	Variation
dbSTS		Markers and matching sequences and maps
GenBank	G	Sequence[d]
OMIM[b]	O	Descriptions of disorders and genes
PDB Neighbors		Structures of related proteins
PROW[c]		Description of a protein
PubMed	P	Literature citations[d]
RefSeq	R	Reference sequence transcripts and proteins
UniGene	U	Sequence, related proteins from model organisms, expression (CGAP) and radiation hybrid maps

[a]Icons displayed on the LocusLink query result and alphabetic listing pages.
[b]Online Mendelian Inheritance in Man.
[c]Protein Reviews on the Web.
[d]Not comprehensive.

Table 2. Enhanced annotation in RefSeq nucleotide records

Sequence record feature	Data source	Content
CDS – /product=	reviewer	Preferred protein name
COMMENT – COMPLETENESS	reviewer	Complete; complete 5′, complete 3′
COMMENT – REFSEQ	calculated	Identifies GenBank source sequence(s)
COMMENT – Summary:	reviewer	Summary of gene and this product
COMMENT – Transcript Variant	reviewer	Description distinguishing alternate transcripts
DEFINITION	HGNC	Official gene name and symbol[a]
gene – /db_xref=LocusID	LocusLink	Unique, stable ID; linked to LocusLink
gene – /db_xref=MIM:	OMIM	Unique, stable ID; linked to OMIM record
gene – /gene=	HGNC	Official symbol[a]
LOCUS	HGNC	Official symbol[a]
polyA_signal; polyA_site	reviewer	mRNA terminus indicated when sufficient data available
REFERENCE	HGNC; reviewer	Publication citations of relevance to the gene
source – /chromosome; /map	LocusLink	Integrated from GDB, OMIM

[a]An interim gene symbol and name are used if an official symbol/name is not yet available.

ACCESS

RefSeq records can be retrieved by text word queries (gene or protein names or symbols, accession numbers, etc.) or by sequence homology. LocusLink (see Table 3 for URLs) and ENTREZ both support accessing RefSeq records by text. BLAST-based sequence queries must be done against the nucleotide or protein nr databases. The RefSeq records in a BLAST query result can be readily identified by the 'ref' prefix and the distinct accession number format described above. More query details and examples are provided in the LocusLink and RefSeq help and FAQ pages available from the LocusLink home page.

LocusLink and RefSeq records are also freely available on the NCBI FTP site (see Table 3). Note that RefSeq records are not in GenBank and must be downloaded separately.

SEARCHING

Comprehensive descriptions of query strategies and navigation from LocusLink and RefSeq are provided from the LocusLink home page. Please note there are multiple sites within NCBI that include links to LocusLink and RefSeq by specific identifiers. These include Online Mendelian Inheritance in Man (OMIM; http://www.ncbi.nlm.nih.gov/omim/), UniGene (http://www.ncbi.nlm.nih.gov/UniGene/), GeneMap'99 (http://www.ncbi.nlm.nih.gov/genemap/) and dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) (6).

MAINTENANCE

LocusLink and RefSeq records are created and maintained by an ongoing process as described by Pruitt *et al.* (7) and on the

Table 3. LocusLink and RefSeq URLs

Web page	URL
LocusLink Home	http://www.ncbi.nlm.nih.gov/LocusLink/
LocusLink Help Documentation[a]	http://www.ncbi.nlm.nih.gov/LocusLink/help.html
LocusLink FAQ[b]	http://www.ncbi.nlm.nih.gov/LocusLink/LLfaq.html
LocusLink Statistics	http://www.ncbi.nlm.nih.gov/LocusLink/statistics.html
About RefSeq	http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
RefSeq FAQ[b]	http://www.ncbi.nlm.nih.gov/LocusLink/RSfaq.html
RefSeq Statistics	http://www.ncbi.nlm.nih.gov/LocusLink/RSstatistics
LocusLink ftp site	ftp://ncbi.nlm.nih.gov/refseq/LocusLink/
RefSeq ftp site	ftp://ncbi.nlm.nih.gov/refseq/

[a]Definitions of terms and content, query help.
[b]FAQ: frequently asked questions.

LocusLink web site. The LocusLink web pages are currently refreshed weekly. RefSeq records may be modified at any time based either on text changes (nomenclature), or by replacing a provisional record with a reviewed one (maintaining the same accession number, but changing the version number and sequence ID numbers if the sequence data has changed).

CONTACT

Questions, comments and suggestions can be emailed to info@ncbi.nlm.nih.gov . We welcome collaborations with and contributions from the research community.

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. (1978) *Genomics*, **45**, 468–471.
2. Cooper,D.N., Ball,E.V. and Krawczak,M. (1998) *Nucleic Acids Res.*, **26**, 285–287.
3. Benson,D.A. (1999) *Nucleic Acids Res.*, **27**, 12–17. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 15–18.
4. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) *Methods Enzymol.*, **266**, 141–162.
5. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
6. Sherry,S.T. (2000) *Nucleic Acids Res.*, **28**, 352–355.
7. Pruitt,K.D., Katz,K.S., Sicotte,H. and Maglott,D.R. (2000) *Trends Genet.*, in press.