19. LocusLink: A Directory of Genes

Donna Maglott

Created: October 9, 2002
Updated: August 12, 2003

Summary

LocusLink organizes information from collaborating public databases and from other groups within NCBI to provide a locus-centered view of genomic information from human, mouse, rat, zebrafish, *Drosophila melanogaster*, and human immunodeficiency virus 1 (HIV-1). Each LocusLink record (one for each locus) consists of a collection of links to more information about a locus. The intent, therefore, is not to copy all information about a locus into one record; the intent is to provide a sufficient description of each linked item so that LocusLink can be searched and to provide enough connections for users to find related resources. As part of this process, a unique integer is assigned to each locus. This identifier can then be used by other resources to connect back to LocusLink.

Overview

LocusLink serves as a hub of information for loci from several model organisms: human, mouse, rat, fruit fly, and human immunodeficiency virus 1 (HIV-1). On a regular basis, “source” databases are checked for novel information about genes or other loci. If the record already exists in LocusLink, the novel information is added. Otherwise, a new record is created. Because many of the contributing databases are curated and because the LocusLink records are reviewed by NCBI staff, LocusLink can be considered a curated resource.

LocusLink gathers information from databases both within and external to NCBI. These are scanned using a combination of automatic and manual techniques. LocusLink collaborates with organism-specific and nomenclature databases, from which a combination of sequence data and names are used to initiate new LocusLink records. LocusLink also finds new records by a weekly review of submissions to GenBank and new versions of UniGene (which are released periodically). Each new LocusLink record is assigned a unique identifying number that is tracked, a LocusID.

LocusLink records are used in turn by UniGene, dbSNP, and the organism-specific and nomenclature databases. For example, a new LocusLink record created on the basis of a sequence submitted to GenBank can be the basis for a new entry in an organism-specific database. This new defining sequence of the LocusLink record is “BLASTed” against sequences from the same or other genomes, which identifies other GenBank records for the same gene or its homologs. When a related sequence is identified in a genome within the scope of LocusLink and that sequence has not yet been included in the appropriate genome-specific database, LocusLink makes the connection and reports it for others to use.

LocusLink also feeds new sequences into the Reference Sequence (RefSeq) project. The prerequisite for this process is that the sequence encodes a complete protein. For more details about how eukaryotic RefSeq mRNA and protein records are initiated, see Chapter 18 in this Handbook.

Although historically many of the GenBank sequences used to initiate LocusLink records represented characterized genes sequenced by individual research labs, an increasing number of records in LocusLink are generated as a part of NCBI's genome annotation pipeline (see Chapter 14). The IDs assigned to these records are termed “interim IDs” and are not tracked within the LocusLink database until they have been reviewed or a preponderance of evidence confirms that the gene prediction is real.

In summary, no matter whether the data are stored as a result of curation or computation, the central function of LocusLink is to establish an accurate connection between the defining sequence for a locus and other descriptors for that locus. With such a connection in place, it is possible to:

- Establish a RefSeq for that locus.
- Identify or validate putative orthologs (genes with a common ancestor) based on a combination of sequence similarity and conserved synteny (a stretch of chromosome in which the gene order is conserved across species).
- Support the NCBI annotation pipeline based on mRNA sequence alignment.

As will be discussed in more detail in the following sections, other NCBI resources make use of the LocusLink LocusID-to-sequence connection to provide appropriate nomenclature and other identifiers for sequences within their scope [see also Chapter 5 (dbSNP), Chapter 21 (UniGene and HomoloGene), and Chapter 20 (Map Viewer)].

How to Query LocusLink

The LocusLink homepage [http://www.ncbi.nlm.nih.gov/LocusLink/index.html] gives a brief introduction to the resource as well as information on new features. At the top of this and all LocusLink pages is a query bar that allows users to search not only LocusLink but also a selection of other resources within NCBI. Thus, if a query against LocusLink does not return a result, a different resource can be selected and searched without retyping the query term. Currently, these other resources include: OMIM, PubMed, Entrez Nucleotide, Entrez Protein, Human Map Viewer, UniGene, and UniSTS. The default for LocusLink queries is to search all of the available organisms, although a specific organism can be selected by using the pull-down menu in the search bar.

LocusLink supports several types of queries, including gene names, GenBank Accession numbers, and other resource-specific ID numbers such as UniGene cluster or STS marker IDs. Queries are not case sensitive, and retrievals are based on a word index, not a phrase index.

Punctuation is not removed; thus, entering the query *beta actin* will retrieve any record that contains the word “beta” and the word “actin” (processed as “beta AND actin”). *“beta actin”* will retrieve No Results, because the quotes are not removed, and look-up fails on *“beta”* or *actin?*. More advanced searches can limit queries to specific fields (Table 1) or by controlled terms (Table 2). Further information on options for formulating queries are documented in the Help [http://www.ncbi.nlm.nih.gov/LocusLink/help.html] pages.

Table 1. LocusLink query terms: field restrictions.

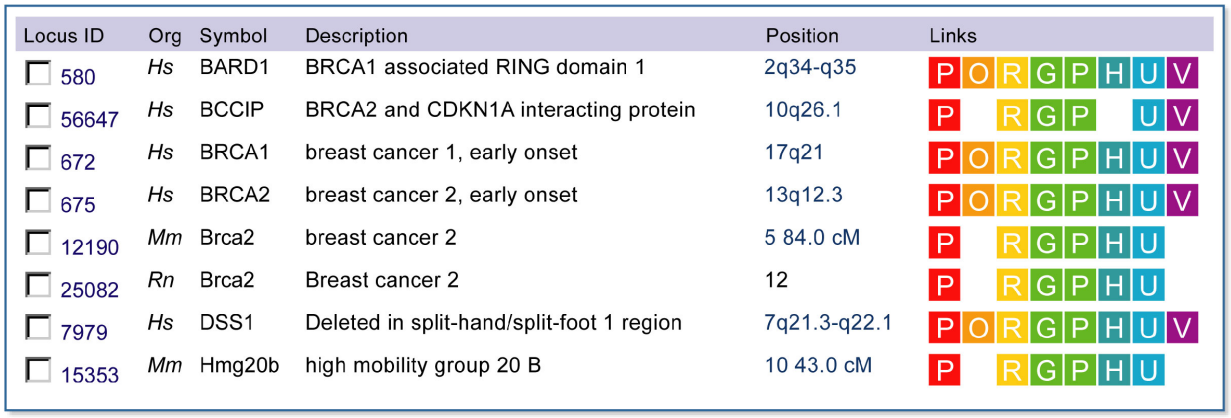
Field	Meaning	Example of search term[a]
[chr]	Chromosome number	21[chr]
[loc]	LocusLink ID	4292[loc]
[mim]	OMIM number	300200[mim]
[sym]	Gene symbol	abc*[sym]
[pm]	PubMed ID	123456[pm]
[ngi]	Nucleotide gi number	223344[ngi]
[pgi]	Protein gi number	1234567[pgi]

[a] No space is allowed between the value and the field name.

Table 2. LocusLink query terms: controlled terms.

Term	Meaning
disease_known	Human locus associated with a phenotype defined by a MIM number (may be only that phenotype)
has_homol	Associated with a HomoloGene record
has_omim	Associated with an OMIM record
has_refseq	Associated with a RefSeq
has_seq	Associated with nucleotide sequence
has_snp	Associated with a dbSNP record
seq_map	Either the gene or an STS in this record has been localized to the sequence-based map available from Map Viewer
type_dseg	A DNA segment. May include BAC or YAC ends.
type_gene_other	A gene that does not fall into the category type_gene_protein
type_gene_protein	A gene that encodes a protein product. Does not include unreviewed, putative genes based only on modeling; named genes that encode only part of a protein product, such as immunoglobulin variable, diversity, joining, or constant regions; or genes that exhibit somatic rearrangement.
type_pheno	Characterized as a mapped phenotype only
type_pseudo	A pseudogene
type_qtl	A phenotype characterized as a QTL only
type_region	A region on the genome. Examples are named gene clusters and viral integration sites.

A query on the LocusLink homepage returns a report page that is organized in a tabular format, as shown in Figure 1. The complete record is viewed by selecting **Locus ID** or **More** (for interim IDs from the annotation pipeline).



Locus ID	Org	Symbol	Description	Position	Links
580	*Hs*	BARD1	BRCA1 associated RING domain 1	2q34-q35	P O R G P H U V
56647	*Hs*	BCCIP	BRCA2 and CDKN1A interacting protein	10q26.1	P R G P U V
672	*Hs*	BRCA1	breast cancer 1, early onset	17q21	P O R G P H U V
675	*Hs*	BRCA2	breast cancer 2, early onset	13q12.3	P O R G P H U V
12190	*Mm*	Brca2	breast cancer 2	5 84.0 cM	P R G P H U
25082	*Rn*	Brca2	Breast cancer 2	12	P R G P H U
7979	*Hs*	DSS1	Deleted in split-hand/split-foot 1 region	7q21.3-q22.1	P O R G P H U V
15353	*Mm*	Hmg20b	high mobility group 20 B	10 43.0 cM	P R G P H U

Figure 1: Example of a LocusLink query results page. *LocusID* field: NCBI assigns a unique, stable LocusID to each locus. Selecting the number retrieves the detailed LocusLink report page. As part of NCBI's Genome Annotation Project [http://www.ncbi.nlm.nih.gov/genome/guide/build.html], some LocusLink records are generated that are likely to be temporary and which, therefore, are not represented by a stable ID. Such records are indicated by **More**, which can be selected to display the report page. *Org* field: a two-letter abbreviation for the organism in which this locus is described. The symbols (*Hs* for *Homo sapiens*, *Mm* for *Mus musculus*, and *Rn* for *Rattus norvegicus*) match the abbreviations used by UniGene [http://www.ncbi.nlm.nih.gov/UniGene/]. *Symbol* field: the official or alias symbol. In some cases, a symbol may be used for multiple loci. *Description* field: the gene name. *Position* field: the cytogenetic (human or rat) or genetic (mouse) location. *Links* field: small, color-coded *boxes* indicate when links are available for PubMed (P), OMIM (O), RefSeq (R), GenBank nucleotide (G), Protein (P), HomoloGene (H), UniGene (U), and variation data (V). Note: neither PubMed nor GenBank links are comprehensive. Use the **Related Articles** or **Related Sequences** in the **Links** menu to retrieve more records.

A LocusLink Report: The Details

This section is organized according to the features and subdivisions seen on a LocusLink report page, from top to bottom. We will illustrate the descriptions of the features by using screen shots from the LocusLink report of BRCA2 and CDKN1A interacting protein (BCCIP; LocusID 56647 [http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=56647]).

On each LocusLink report page, the first set of links is a table of contents for the sections included in the report being displayed, which are hyperlinked to the appropriate section. **Top of page** links to the top. The latter is useful if you want to gain access to the query bar to enter another query.

mRNA–Genomic Alignments Diagram

As part of the genome annotation process at NCBI, GenBank and RefSeq mRNAs are aligned to genomic contigs to determine the exon/intron organization of genes. The results of these calculations can be displayed using NCBI's Evidence Viewer [http://www.ncbi.nlm.nih.gov/Entrez/Genome/evvdoc.html].

The diagram at the top of a LocusLink report represents the intron/exon structure of the gene as determined by the Genome Annotation Pipeline (see Chapter 14). For example, the annotation for BCCIP looks like this (the details may change with each reannotation of the genome):

This abbreviated output displays the longest alignment, where tick marks indicate the positions of the exons. Genes with a single exon are represented as a solid thick bar with flanking horizontal lines. The diagram is hyperlinked to the Evidence Viewer, which in turn is linked to the appropriate Entrez Nucleotide records by the Accession number of each sequence. Mousing over the individual entries in the Evidence Viewer displays a definition line from that sequence record.

Quick Link Buttons

Buttons [http://www.ncbi.nlm.nih.gov/LocusLink/help.html#reports] are provided at the top of each report page to indicate what information may be available from other resources. In the case of BCCIP, the buttons are:

where (from left to right) they represent links to: PubMed publications, UniGene, Map Viewer, dbSNP variation database, the Genome Database, Ensembl, and UCSC Human Genome Browser. The buttons are used for connecting to related resources within NCBI or to external genomic databases. Others in this category include Ace Viewer, OMIM, MGC, MGI, and RGD. More links for any record may be listed at the end of the report page in the **Additional Links** section.

Nomenclature

Homo sapiens Official Gene Symbol and Name (HGNC)

BCCIP: BRCA2 and CDKN1A Interacting protein
LocusID: 56647

LocusLink uses symbols and gene names from official authority lists when available. If no connection to official nomenclature can be made, symbols and names are selected as available from the defining sequence record. If sequence and positional homology (synteny) suggest that a locus not named officially in one species is orthologous to a named gene in another species, the symbol from the ortholog may be included in the LocusLink record. If no symbol can be identified for a new locus, the letters LOC are prepended to the LocusID. Once an official or meaningful symbol has been identified, that LOC symbol is discontinued (because the record will still be searchable and identified by the LocusID itself).

Sources of Nomenclature

Information on the nomenclature authorities that collaborate with LocusLink can be viewed by following the Nomenclature [http://www.ncbi.nlm.nih.gov/LocusLink/LLnomen.html] link on the LocusLink homepage. Except for the human genome, data from these authorities are imported automatically and used either to correct existing LocusLink names or to create new records. For the human genome, the Human Gene Nomenclature Committee has direct access to the LocusLink database and inserts and updates records interactively.

Information Processing

Rules. In addition to official symbols and full names, LocusLink provides other symbols and names seen in publications and sequence records appropriate to the locus. These alternative names are not meant to be comprehensive and are usually reviewed only when the RefSeq is being reviewed.

Stability and Tracking. Although LocusLink does maintain some nomenclature history, the appropriate nomenclature committee performs this function more comprehensively. For individual LocusLink records, links to those committees are provided from the LocusLink report.

Update Frequency. Official nomenclature is modified when changes are available, ranging from daily (human, mouse) to weekly (zebrafish), or longer (fly, rat). Names provided by NCBI staff are available daily.

Methods. Data files are imported by FTP and analyzed through a combination of shell and Perl scripts, in conjunction with relational database tables used to hold the input data. Records with no data conflicts are updated automatically. If a flag is raised because of uncertainty about the relationships among a name, a sequence, and a record identifier, then an expert reviews that record. During review, either the record is corrected or the appropriate nomenclature committee (s) is contacted to resolve the discrepancies.

Interactions with Other Resources at NCBI

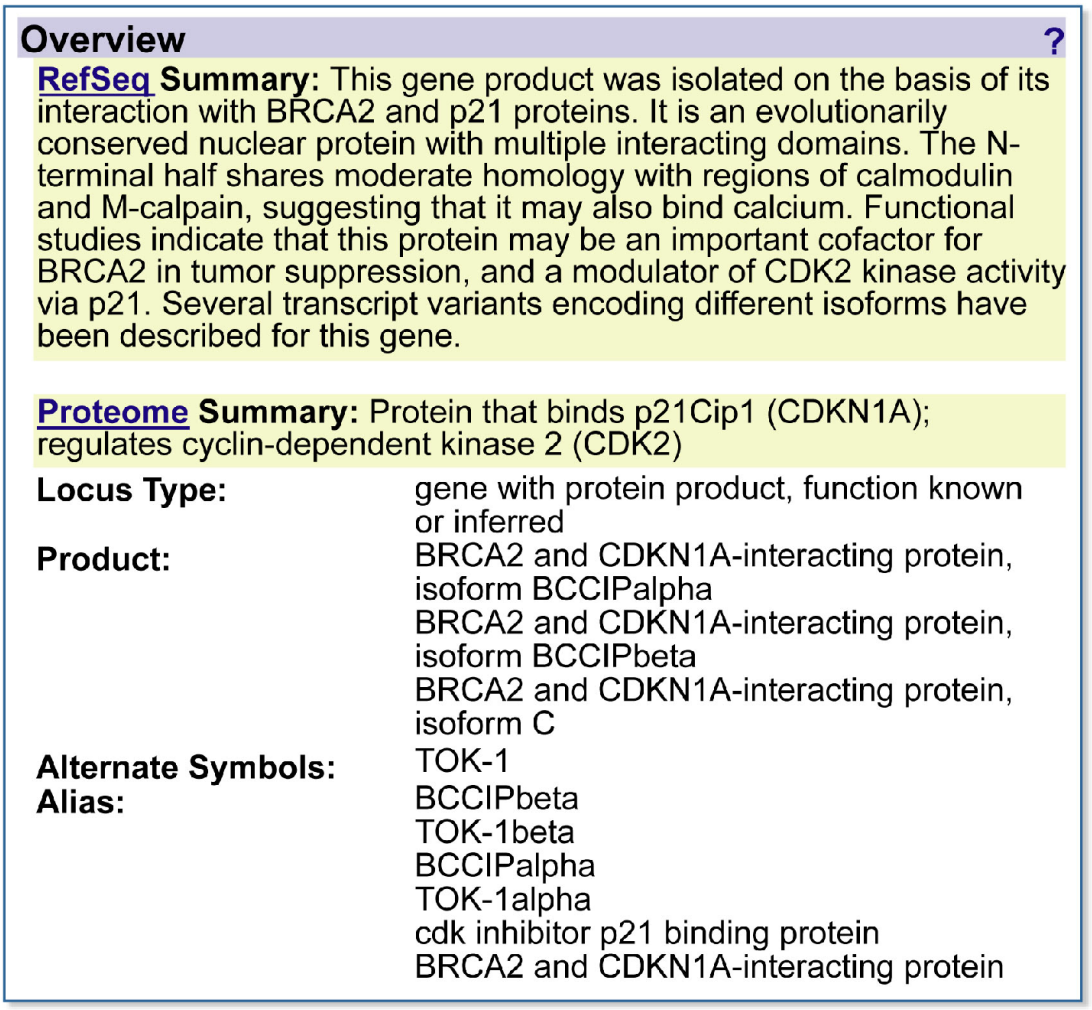
Several NCBI databases use the nomenclature maintained by LocusLink. These names are incorporated into other databases based primarily on name–LocusID–sequence connections, i.e., sequence comparisons identify similar sequences in LocusLink, all of which have a LocusID; from this, the associated nomenclature can be extracted and applied to the original sequence from the collaborating database (Table 3). Nomenclature data can be extracted from files available on the FTP site.

Table 3. Nomenclature interactions with NCBI resources.

Resource	Keys into LocusLink	Method of matching resource to LocusID
HomoloGene	mRNA	Alignment
Map Viewer	mRNA, gene feature	Alignment
UniGene	mRNA, protein gi	Clustering
UniSTS	mRNA	e-PCR
	Genomic annotation	e-PCR
	Marker name	Publications

Overview

This section of the LocusLink report page may include any or all of the following categories of information.



Overview ?

RefSeq Summary: This gene product was isolated on the basis of its interaction with BRCA2 and p21 proteins. It is an evolutionarily conserved nuclear protein with multiple interacting domains. The N-terminal half shares moderate homology with regions of calmodulin and M-calpain, suggesting that it may also bind calcium. Functional studies indicate that this protein may be an important cofactor for BRCA2 in tumor suppression, and a modulator of CDK2 kinase activity via p21. Several transcript variants encoding different isoforms have been described for this gene.

Proteome Summary: Protein that binds p21Cip1 (CDKN1A); regulates cyclin-dependent kinase 2 (CDK2)

Locus Type:	gene with protein product, function known or inferred
Product:	BRCA2 and CDKN1A-interacting protein, isoform BCCIPalpha BRCA2 and CDKN1A-interacting protein, isoform BCCIPbeta BRCA2 and CDKN1A-interacting protein, isoform C
Alternate Symbols:	TOK-1
Alias:	BCCIPbeta TOK-1beta BCCIPalpha TOK-1alpha cdk inhibitor p21 binding protein BRCA2 and CDKN1A-interacting protein

The **Summary** is written by RefSeq staff and/or by external contributors such as OMIM, Proteome, or Protein Reviews on the Web (PROW). These summaries provide a quick synopsis of what is known about the gene, the function of its encoded protein or RNA products, disease associations, spatial and temporal distribution, and so on.

Locus Type indicates the type of molecule that is the defining sequence for the Locus Link record. It is selected from the options listed in Box 1.

Product lists the known names of proteins associated with the locus. This is not an exhaustive list. The intent is to support data retrieval and to document usage; therefore, only some of the names from sequence databases or published literature are reported.

Alternate Symbols indicates other symbols or abbreviations. These symbols may be for the gene, the protein, or a disease phenotype.

Alias indicates other names.

Function

Information about the function of a gene and its RNA or protein products is gathered from several sources.

Function Submit GeneRIF (All Pubs) ?

GeneRIF: Gene References into Function:

11748848	• In BRACA2, two novel frame shift mutations were identified as 5073-507delCT and 6866delC.
11836363	• Unique de novo mutation of BRCA2 has been identified in a woman with early onset breast cancer.
11754111	• BRCA2 germline mutations appear to have a milder clinical phenotype when compared to non-BRCA2 mutations, since survival is higher among breast cancer patients carrying a BRCA2 mutation compared to sporadic breast cancers.

Gene Ontology™:

Term	Evidence	Source	Pub
• nucleus	E	NCBI	
• regulation of CDK activity	P	NCBI	

Other Ontologies:

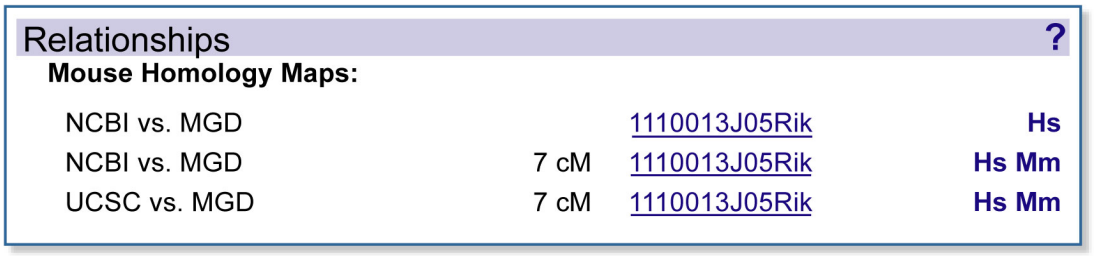
Term	Evidence	Source	Pub
• Nuclear	E	NCBI	

For human genes, links to OMIM, if available, are given under the heading Phenotype. For all genomes, links to the published literature are provided under the heading Gene References into Function (GeneRIF). Any user can submit a reference to a paper they think is important for a locus, but beginning in February 2002, these links are also supplied by MeSH indexers at the National Library of Medicine.

Increasingly, the Gene Ontology™ (GO) vocabulary terms are incorporated from GO's FTP site [ftp://ftp.geneontology.org/pub/go], and each term is linked to AmiGO [http://www.godatabase.org/cgi-bin/go.cgi], the GO database, which can be browsed or searched. When the literature citation(s) supporting the GO term is available, a link to PubMed (**pm**) is provided.

Relationships

This section reports other loci, and/or the proteins that they encode, that have a defined relationship to the locus being displayed.

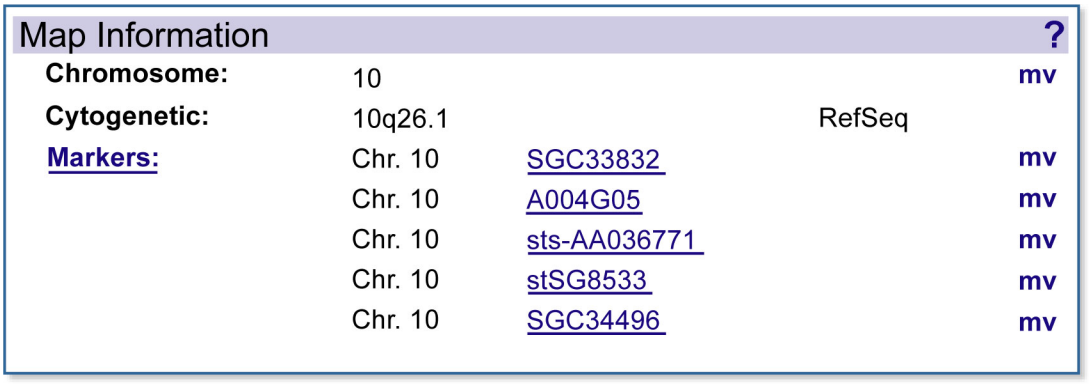


At present, this section includes: (*a*) reports of how interim loci (i.e., those identified by genome annotation processes) are related to other loci, based on the Accession number of the mRNA that was used to define the intron/exon organization; and (*b*) reports of human–mouse homologs, which are linked to the Human–Mouse Homology Map [http://www.ncbi.nlm.nih.gov/Homology/].

We plan to expand this section to include other types of pairwise relationships. For example, “overlap”, “interspersed”, and “protein binds” will refer to genes that overlap, that are contained in or contain another gene, and for which the encoded proteins interact, respectively. Although primarily for relationships within a species, when proteins of one species interact with those of another (e.g., infectious agents), such a relationship will be reported in this section as well.

Map Information

This section reports map data for the locus. The location listed is the same as that on the LocusLink query result page, but if there are any conflicts, additional locations may be reported, along with the source of the conflicting data and a link to that resource.



Map Information				?
Chromosome:	10			mv
Cytogenetic:	10q26.1		RefSeq	
Markers:	Chr. 10	SGC33832		mv
	Chr. 10	A004G05		mv
	Chr. 10	sts-AA036771		mv
	Chr. 10	stSG8533		mv
	Chr. 10	SGC34496		mv

For the human and mouse genomes, if no published map location has been identified and if the gene has been aligned to NCBI's genome assembly, the cytogenetic position is recalculated with each build. The conversion files used in the process are ISCN800_abc [ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/maps/mapview/ISCN800_abc] for human. This location is also displayed from Map Viewer. With each assembly of the genome, genes reported to be on one chromosome that appear to have better alignments on a different chromosome are identified. If marker and other data are consistent but distinct from the published map location, then the LocusLink report will be modified to conform to the evidence gained from analysis of the human genome.

Genetic and physical map positions are incorporated from the published maps used in Map Viewer. Rather than report all position data for any locus in any coordinate system, links are provided to Map Viewer via the **mv** link in the map section or indirectly through the marker names, which are linked to the UniSTS record.

Marker Data

In LocusLink, markers are defined as sequence tagged sites (STSs) associated with the locus. LocusLink reports markers either as the locus itself or as a marker that has some relationship with a gene. LocusLink does not store all of the markers available for a genome, which is the function of UniSTS. Some LocusLink records contain only markers. These are considered historical records; new records are added only if a contributing database reports a marker to represent a gene.

Information Sources

The marker data that LocusLink reports come primarily from any of the following paths: (1) a report from a genome-specific database that states that a marker is within a gene or locus; (2) for genes, a caluculation based on e-PCR using mRNA as the electronic template, that the marker is within a mRNA defined to be associated with genes; and (3) for genes and in genomes for which NCBI is making an assembly and/or providing annotation e-PCR based on placement of a marker within the range beginning 2.0 kb upstream of the most 5′ mRNA alignment of a gene feature and ending 0.5 kb downstream of the most 3′ mRNA alignment.

Information Processing

Rules. Markers are included in a locus report if they can be assayed by PCR, i.e., if a marker is a STS, and if, according to current computation, they are detected at no more than two locations in the genome being reported.

Stability and Tracking. Relationships between STS and gene records are not archived or tracked. New markers may be added to a gene report if more sequence is identified as being valid for a gene and the new marker "detects" that sequence by e-PCR. A marker may be removed from a report if the sequence that it detected is no longer considered valid for that gene.

Update Frequency. The LocusID-to-UniSTS marker relationship that is based on e-PCR from mRNA templates is calculated daily. The LocusID-to-UniSTS marker relationship that is based on genome annotation is recalculated with each genome build.

Sequence Information

A LocusLink record includes several categories of sequence information [http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=56647#refseq]. The RefSeq section provides information on accessions (nucleotide and protein), the GenBank accessions used to create the RefSeq, and the status (REVIEWED, PROVISIONAL, VALIDATED, PREDICTED). If alternate splice variants have been reviewed, a brief description of each is included. If there is a protein and it demonstrates significant sequence matches to domains defined in the Conserved Domain Database (CDD), then the domains are listed by name, with the score for the domain match. A link is also provided to the CDD browser. To facilitate identification of related proteins and structures, **BL** provides a link to BLink, a resource of precalculated BLAST searches for any given protein.

A second category in the RefSeq section reports sequences generated or annotated as part of an NCBI genome annotation project. The first Accession number(s) displayed is the genomic record (contig), as well as links to the genomic sequence containing the gene (**gb**), a graphic sequence view (**sv**), the Map Viewer display for that contig (**mv**), the Evidence Viewer (**ev**), and Model Maker (**mm**. Please note that the **mv** link in this section is contig based, whereas the **mv** link in the map section is based on the locus itself. These links are provided to make it easier to: retrieve a gene-specific region of a large chromosomal sequence, rather than finding that gene in a record that may be several megabases (**gb**); review the alignment evidence for a gene that has been annotated on a genomic assembly (**ev**); and construct your own mRNA model(s) in a genomic region based on mRNA and ESTs that align there (**mm**). The next Accession numbers listed are for sequences annotated on the genomic sequence. They may be the same as the curated RefSeq in the previous category, or they may be models, distinguished by the format of their Accession numbers (XM_000000, XP_000000, and XR_000000 refer to models). CDD domain matches and BLink links are also provided for model proteins.

The *GenBank Sequences* section provides a list of representative nucleotide and protein Accession numbers for the locus, each Accession number of which is linked to an Entrez sequence display. Proteins listed in this section also provide a link to BLink.

Additional Links

This section provides a list of sites that may have additional information. The OMIM and UniGene links are redundant, with the button links at the top of the page, but provide the respective ID numbers that are hidden in the button links. Other links shown here, such as GeneCards [http://bioinfo.weizmann.ac.il/cards/] or GeneTests · GeneClinics [http://www.geneclinics.org/], a resource containing genetic testing and disease information, are generated automatically based on links from OMIM provided through LinkOut (see Chapter 17). Other links are added by RefSeq curators as they review a gene or after suggestions by other authorities.

Maintenance and Reporting

LocusLink records can be categorized by two major criteria: the type of locus being described and the tracking maintained for the record (Table 4). All but the interim loci (ascribed to LocusLink records that are based on gene-prediction software only) are tracked, which means that if the record is discontinued or determined to be redundant with another, queries based on the ID representing the discontinued or secondary record will return a result. For interim loci, the interim identifier is not reused, but a query based on that identifier number will not return a result if that gene model is no longer being annotated. It is anticipated that the number of interim loci will decrease as more curation is applied to the records of the genome.

Table 4. LocusLink record tracking.

Maintenance	Type of locus
Tracked	Officially named genes and pseudogenes, for both nuclear and mitochondrial genomes, whether or not the final gene product is known to be a protein or a RNA.
Tracked	Probable protein-coding genes, defined by one or more mRNAs. The function of the encoded protein is not necessarily known.
Tracked	Mapped phenotypes, such as disease susceptibility loci or QTLs
Tracked	Gene segments (such as coding regions for variable regions of immunoglobulin or T-cell receptors)
Not tracked	Gene predictions from NCBI's genome annotation pipeline.

Retrieving Historical Data

LocusLink supports retrieval of inactive records in the following ways:

1. If a non-interim record has been discontinued and the record appeared to have been created in error (i.e., could not be merged into or made secondary to another record), then the withdrawn record can be retrieved. These records are clearly noted with the term “withdrawn” on the query result table and the report page. The report page also includes an explanation for why the record was discontinued.
2. If a locus record has been made secondary to another record, a query on the secondary ID will take you to the current record.

Merges are reported in the FTP file LocusID_history, and the status of all queryable LocusIDs is reported in the file LL_tmpl.gz (ftp://ftp.ncbi.nih.gov/refseq/LocusLink/ [ftp://ftp.ncbi.nih.gov/refseq/LocusLink/]).

New Records

Records are added to LocusLink in several ways:

- External resources and collaborators provide information on new, officially named genes and the sequences that define them.
- New sequences are released from GenBank/DDBJ/EMBL and are identified as being from a gene not yet described in LocusLink. If sequence alignments indicate that a new sequence matches an interim locus annotated in the Annotation pipeline, then the interim locus is converted to a “curated” one, and the sequence accession is added to that record.
- Communications from the public. LocusLink provides three mechanisms for users: (*a*) at the bottom of each report is a mail link to the NCBI Service Desk. It is helpful when the mail message contains a sequence accession and a published citation in addition to the description of the request; (*b*) The GeneRIF [http://www.ncbi.nlm.nih.gov/LocusLink/GeneRIFhelp.html] link within the blue function bar on any LocusLink report page can be used. For loci not yet in the database, there are generic **NEWENTRY** records established for human, mouse, rat, zebrafish, and fly. Thus, the user can use **NEWENTRY** as a query term, select the species that needs a new entry, and add the GeneRIF. Please note that GeneRIFs require a published citation with a PubMed ID; and (*c*) The Update Submission Form [http://www.ncbi.nlm.nih.gov/LocusLink/update.cgi] accessed from the LocusLink homepage. This form should be used to notify RefSeq, LocusLink, and OMIM staff of new genes, to suggest updates, or to report errors.

FTP Site

LocusLink can be obtained by FTP [ftp://ftp.ncbi.nih.gov/refseq/LocusLink] (Table 5). These files are refreshed when new data are available, which for most files is daily (weekdays).

Table 5. The LocusLink FTP site.

File or directory	Description
README	Documentation for the directory
HomologyMaps	Directory of data files used in the human/mouse comparative map
LL.out.gz	Tab-delimited file of descriptors for current LocusLink records
LL.out_dm.gz	Drosophila subset of LL.out
LL.out_dr.gz	Zebrafish subset of LL.out
LL.out_hs.gz	Human subset of LL.out
LL.out_mm.gz	Mouse subset of LL.out
LL.out_rn.gz	Rat subset of LL.out
LL_tmpl.gz	Current text file for displays on the LocusLink site
LocusID_history	Report of locus_id merges

File or directory	Description
homol_seq_pairs.gz	mRNA accession pairs determined by MegaBlast
loc2UG	Current LocusID/UniGene cluster conversion table
loc2acc	Current LocusID/GenBank accession report
loc2cit	Current LocusID/PubMed ID/MedLine ID report
loc2ref	Current LocusID/RefSeq accession report
loc2sts	Current LocusID/UniSTS ID relationships
mim2loc	Current LocusID/MIM ID relationships

The file LL_tmpl.gz represents the text file for displays of the complete LocusLink site and is in a semistructured tag:value format, which makes it challenging to parse. A subset of these data is available in tab-delimited format, either for all species covered by LocusLink (LL.out) or in species-specific files (e.g., LL.out.hs and LL.out.dm). These files report the LocusID, symbols, names, map location, and identity of the contributing database.

The file loc2UG (the LocusID/UniGene cluster conversion table) is refreshed with each UniGene build, and homology reports are refreshed with new genome annotation builds. These files can be used to obtain names and sequences connected to a locus.

Integration with Other Resources

The database supporting LocusLink houses more than only the unique loci identifiers for the genomes in its scope. It also records, whenever possible, the public sequence Accession numbers that define these loci and, along with its collaborators, applies several tests for data consistency. Thus, the relationships between LocusID and sequence Accession numbers are used by other databases at NCBI to convert information about mRNA or protein sequences to the LocusID and all other information associated with that LocusID (name, database cross-references, protein product, and others). These relationships are outlined in Table 6.

Table 6. Connections between LocusID and other NCBI resources.

Resource	Connection made	Basis
dbSNP	LocusID to Reference SNP ID	dbSNP accessions aligned to mRNAs, or within the boundary of 2000 nt upstream through 500 nt downstream of known exons
Map Viewer	LocusID to cytogenetic, genetic, or sequence position	Reports of cytogenetic position or calculated from sequence assembly; reports of genetic position, connecting LocusIDs to sequence position based on alignment of accessions associated with LocusIDs
RefSeq mRNAs and proteins	LocusID to mRNA and protein accessions	Calculated and curated LocusID/mRNA or protein_id relationships
UniGene	LocusID to UniGene cluster ID	Based on the LocusID/mRNA or protein_id relationships, identifying the cluster ID and requiring that the LocusID/UniGene cluster ID relationship be 1:1

Resource	Connection made	Basis
UniSTS	LocusID to UniSTS sts_id	Based on the LocusID/mRNA relationships or overlap in the genomic assembly, after positioning the STS by e-PCR

More Information on LocusLink

On LocusLink's static HTML pages, such as the homepage, there are links to general resources. Specific sites provide information for LocusLink FAQs [http://www.ncbi.nlm.nih.gov/LocusLink/LLfaq.html] and RefSeq statistics [http://www.ncbi.nlm.nih.gov/LocusLink/RSstatistics.html].

Box 1: List of possible locus types for LocusLink records.

- D segment
- RNA, ribosomal
- RNA, small cytoplasmic
- RNA, small nuclear
- RNA, small nucleolar
- RNA, transfer
- gene with no protein product
- gene with protein product, demonstrates somatic rearrangement
- gene with protein product, function known or inferred
- gene with protein product, function unknown
- gene, segment
- model, *ab initio*
- model, *ab initio*, with EST support
- model, supported by EST alignments
- model, supported by mRNA alignments
- model, supported by mRNA and EST alignments
- phenotype only
- pseudogene
- pseudogene, transcribed
- quantitative trait locus (QTL)
- region
- regulatory element
- repetitive element
- unknown