Genome analysis

The UCSC Known Genes

Fan Hsu[1,*], W. James Kent[1], Hiram Clawson[1], Robert M. Kuhn[1],
Mark Diekhans[1] and David Haussler[2]

[1]Center for Biomolecular Science and Engineering, School of Engineering and [2]Howard Hughes
Medical Institute University of California Santa Cruz Santa Cruz, CA 95064, USA

ABSTRACT

The University of California Santa Cruz (UCSC) Known Genes dataset is constructed by a fully automated process, based on protein data from Swiss-Prot/TrEMBL (UniProt) and the associated mRNA data from Genbank. The detailed steps of this process are described. Extensive cross-references from this dataset to other genomic and proteomic data were constructed. For each known gene, a details page is provided containing rich information about the gene, together with extensive links to other relevant genomic, proteomic and pathway data. As of July 2005, the UCSC Known Genes are available for human, mouse and rat genomes. The Known Genes serves as a foundation to support several key programs: the Genome Browser, Proteome Browser, Gene Sorter and Table Browser offered at the UCSC website. All the associated data files and program source code are also available. They can be accessed at http://genome.ucsc.edu. The genomic coverage of UCSC Known Genes, RefSeq, Ensembl Genes, H-Invitational and CCDS is analyzed. Although UCSC Known Genes offers the highest genomic and CDS coverage among major human and mouse gene sets, more detailed analysis suggests all of them could be further improved.

Contact: fanhsu@soe.ucsc.edu

INTRODUCTION

The UCSC Genome Browser (Kent *et al.*, 2002; Karolichik *et al.*, 2003), which was developed in conjunction with the assembly and publication of the first Human Draft Genome (International Human Genome Sequencing Consortium, 2001), has become a popular website for biomedical communities around the world. The number of its annotation datasets, or tracks, continues to grow each year.

During the earlier stage of Genome Browser development, there were only a few annotation tracks in its Genes and Gene Prediction section. The section included a few gene prediction tracks and a RefSeq Gene track, each having its own limitations. Different gene prediction programs often produce different results. The NCBI RefSeq (Pruitt *et al.*, 2005) offers a high-quality gene set, but because it is produced by an extensive manual curation process it has limitations on its coverage and timeliness of availability.

In Addition, direct links between RefSeq genes and Swiss-Prot proteins were not available. Hence we decided to develop an automated process to construct the UCSC Known Genes dataset based on the latest protein data from Swiss-Prot/TrEMBL (Bairoch *et al.*, 2005), now also known as UniProt, and the associated mRNA data from GenBank (Benson *et al.*, 2005).

While there are various different definitions of what constitutes a gene, we chose to limit our gene set to protein coding genes and require each gene be substantiated by at least a transcript (either a GenBank mRNA or a NCBI RefSeq) and a UniProt protein. We relied upon UniProt's comprehensive cross-references between the proteins and their associated GenBank mRNAs to build our initial candidate gene set. Alternative splicing isoforms are included as different entries, as long as they are represented by a UniProt protein and a transcript. The initial candidate gene set is further ranked and processed to select the best representative protein/mRNA for each gene and duplicates with identical CDS structure removed.

The result of this effort is the UCSC Known Genes: a comprehensive gene set based mostly upon experimental data. The set can be built automatically in a relatively short time. Since its first introduction in early 2003, UCSC Known Genes are now available for several assembly releases of three major genomes, human, mouse and rat. As shown in Figure 1, the UCSC Known Genes dataset has also become a central foundation for key genomic and proteomic applications, such as the UCSC Genome Browser, Proteome Browser (Hsu *et al.*, 2005), Gene Sorter (Kent *et al.*, 2005), and Table Browser (Karolchick *et al.*, 2004), offered at the UCSC bioinformatics web site, genome.ucsc.edu. Extensive cross-reference links to other gene-related data available on the web are also compiled and presented for each Known Gene.

Since the start of our effort, several other gene sets besides RefSeq from NCBI have become available: Ensembl Genes from EMBL-EBI, H-Invitational Gene Database (HInv-DB) of JBIRC and CCDS (the Consensus Coding DNA Sequence) from EBI, NCBI, UCSC and WTSI. We present comparison between UCSC Known Genes and other gene sets in the ANALYSIS section.

METHODS

Raw protein data files are downloaded from UniProt and parsed to create a set of structured relational database tables. A cross-reference table between
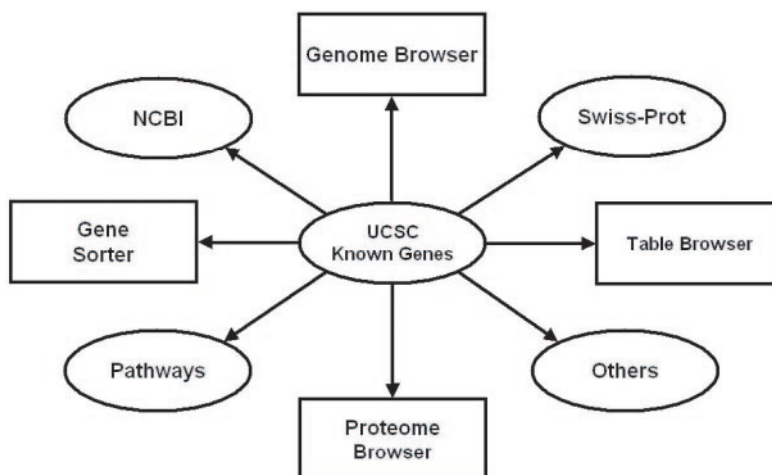
---

*To whom correspondence should be addressed.

Fig. 1. The UCSC Known Genes dataset serves as a foundation for many key programs, e.g. Genome Browser, Gene Sorter, Proteome Browser, and Table Browser, at the UCSC web site. It also has extensive cross-references to other databases like GenBank/NCBI, Swiss-Prot/TrEMBL, KEGG, BioCyc, and many others.

protein IDs and GenBank mRNA IDs is created from this UniProt data. The existing GenBank mRNA sequences are aligned with their corresponding proteins using BLAT to select the best representative mRNAs for each protein. The resulting protein–mRNA pairs with their mRNA genomic alignments and CDS structures are sorted and filtered to remove redundancy and invalid short sequences. Finally, RefSeq genes having only DNA evidence, which escaped the above process, are added to form the final results as the UCSC Known Genes. More details of this process are described in this section.

A high-level flowchart of the UCSC Known Genes build process is shown in Figure 2. The process consists of the following four sub-processes, as depicted in different colors:

- build protein databases (green)
- get mRNA alignments (red)
- select and prune known genes (blue)
- add DNA-based RefSeq (magenta)

Protein databases construction

The Swiss-Prot and TrEMBL database flat files are downloaded from UniProt at ftp://us.expasy.org/databases/uniprot/knowledgebase/. These files are parsed into 29 relational tables into the database, *swissProt*.

Two cross-reference tables, spXref2 and spXref3, are created from the Swiss-Prot/TrEMBL data. The spXref2 table contains rows of the accession and display IDs of proteins and their external reference databases and the accession numbers of the external database entries. The spXref3 table contains rows of accession and display IDs of proteins, their descriptions, division numbers and HUGO gene symbols and gene descriptions if available.

In addition to Swiss-Prot/TrEMBL and HUGO data, cross-reference data to other protein databases, e.g. InterPro, Superfamily, NCBI Taxonomy, pFam and Ensembl are also compiled and stored in the *proteins* database. Both the *swissProt* and *proteins* databases are used during Known Genes data set construction and at run-time to support the UCSC Genome Browser, Proteome Browser, Gene Sorter, and Table Browser.

mRNA alignments

The UCSC Genome Browser has a collection of MySQL genome databases, one for each genome release. For example, *hg17* is the genomic database for the May 2004 Human Assembly. The mRNA and RefSeq data in those databases are updated every night by importing data from GenBank

(Benson *et al.*, 2005) and then aligned to the base genomes using BLAT (Kent, 2002).

At the beginning of the Known Genes build process, a snapshot of all the mRNA alignments of a genome is filtered to select mRNAs that satisfy a more stringent criterion, i.e. to have at least 40% of its sequence aligned with the genome and having at least 97% sequence identity. For hg17, the filter process removes 14 148 mRNA alignments. These filtered mRNA alignments constitute the initial candidates for UCSC Known Genes. They are converted to genePred format to become UCSC Genome Browser gene tracks.

Selection and pruning known genes

Select best mRNA for each protein. A set of candidate mRNA alignments are selected from the filtered mRNAs that aligned to the genome and have corresponding protein sequences in the Swiss-Prot/TrEMBL data as determined via the cross-reference table spXref2. For hg17, this list contains 74 290 entries.

Each Swiss-Prot/TrEMBL protein may cross-reference multiple mRNAs as its supporting evidence. We tried to pick the best mRNA among all those referenced as the representative mRNA and designate it as a Known Gene. The protein and mRNA sequences in each protein–mRNA pair are aligned using BLAT and then a composite score is calculated for each protein–mRNA pair using the following formula:

Score = mRNA length + date in months since January 1970 $*$ 2 − number of mismatches $*$ 50.

This formula favors longer and newer mRNAs and includes weighted penalty against mismatches of the alignment between the mRNA and base genomes. The relative weights of each term were decided empirically, which seems to work well. We did not conduct a comprehensive search for an optimum combination of the weights.

The mRNA with the highest score is selected as the representative mRNA for the protein. This task is fairly computational intensive because of the number of alignments needed. The UCSC 1000-CPU cluster computer system, KiloKluster, was used to complete this task in ∼45 min. For hg17, the result list has 46 485 entries.

Reduce redundancy and duplication

Sometimes multiple proteins cross-refer to a single mRNA. For those cases, we keep only one protein to represent the gene and remove the redundant proteins from the Known Genes set.
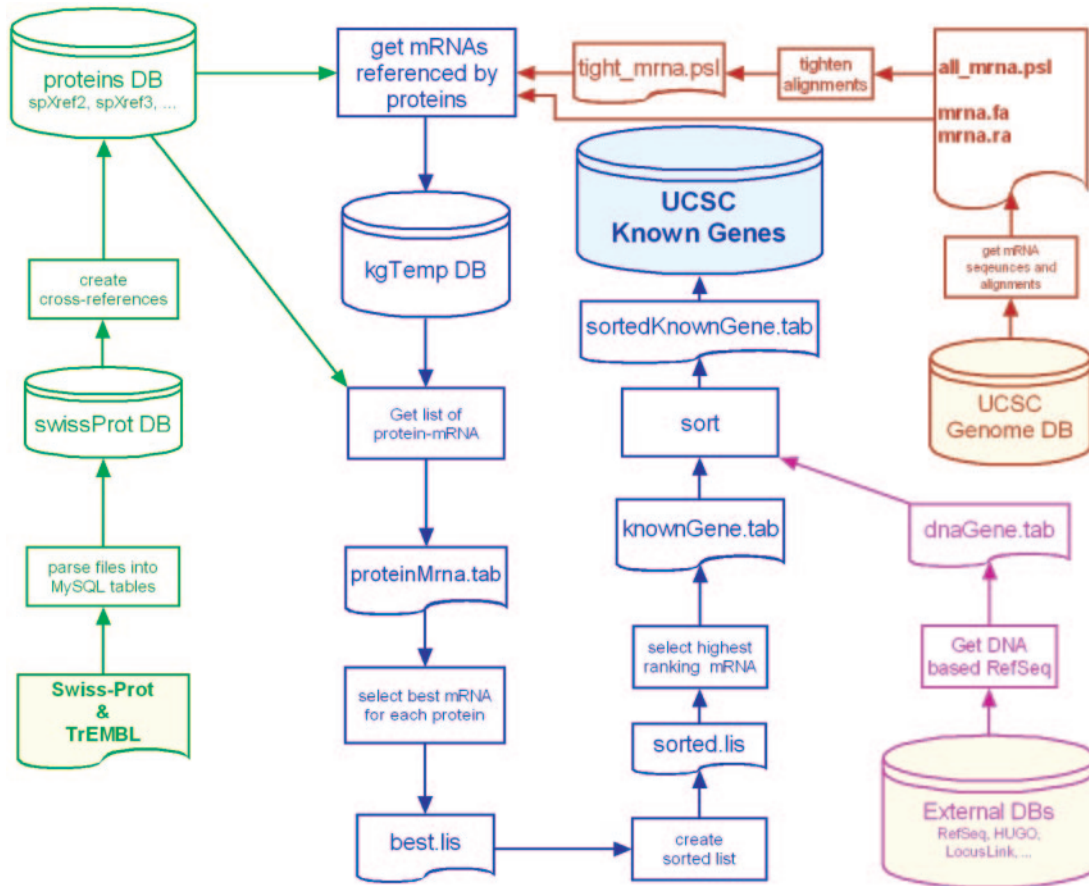
Fig. 2. The high-level flowchart of UCSC Known Genes build process.

A sorted list based on the priority (defined below), transcript length, mRNA date, mRNA ID, protein ID and a computer-generated unique alignment ID is pruned by removing duplicates having identical chromosome number, start and ending positions of coding sequence. Short entries, with total mRNA coding sequence length <50 nucleotides, and questionable protein–mRNA pairs whose coding sequences cover <50% of the corresponding proteins are also removed. For hg17, the resulting table contains 43 597 entries.

Proteins that have corresponding PDB entries are assigned the highest priority. The remaining proteins in Swiss-Prot are then assigned a medium priority. The lowest priority is assigned to TrEMBL entries. This scheme is to ensure that if an mRNA has multiple proteins to choose from, the preference goes to proteins with known 3D structures first, then it would favor Swiss-Prot proteins, which usually have gone through more extensive curation than proteins in TrEMBL.

Add DNA-based RefSeqs

During the initial development of the UCSC Known Genes process, we discovered that several hundred RefSeq genes were not covered by the resulting Known Genes dataset. Further analysis shows that most of the missing RefSeq genes have only DNA evidence without any mRNA evidence. These escaped inclusion owing to the fact that the initial candidates in our process include only mRNA alignments. RefSeq genes without supporting mRNA evidence are added to the set of UCSC Known Genes for completeness. For hg17, there are 741 RefSeq genes without supporting mRNA evidence.

Final loading to Genome database

The finished set of UCSC Known Genes for hg17 has 44 338 entries.

Figure 3 shows the UCSC Known Genes track display together with RefSeq, Ensembl Genes and H-Invitational gene tracks.

CROSS REFERENCES TO OTHER DATA

The UCSC Known Genes have been used as the underpinning base for many other programs and database tables at the UCSC website. Four major cross-reference tables are described here and many more can be found in the database schema specification file, all.joiner.

kgXref

The kgXref table is the central cross-reference table for the UCSC Known Genes. Each record has the following fields:

- kgID, ID of a Known Gene
- mRNA, GenBank accession number of the gene's representative mRNA
- spID, Swiss-Prot/TrEMBL protein accession number
- spDisplayID, Swiss-Prot/TrEMBL display ID
- geneSymbol, HUGO or other gene symbol
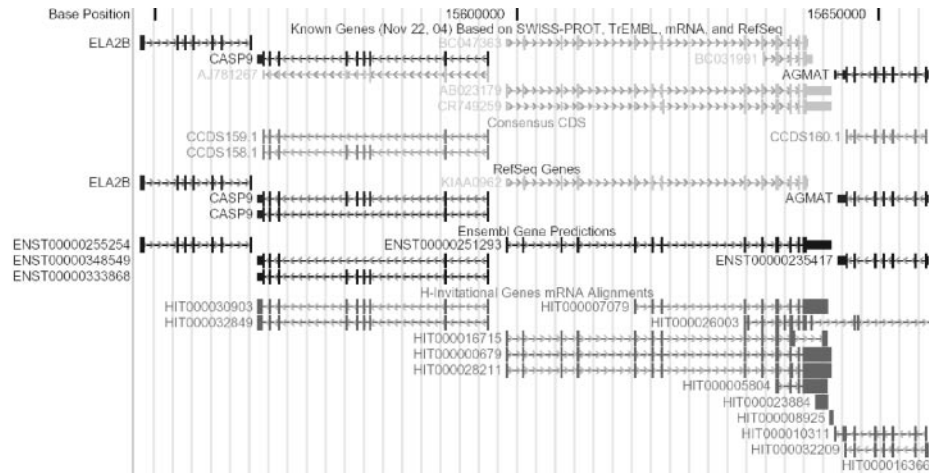- refseq, NCBI RefSeq ID

Fig. 3. The UCSC Known Genes track display.

- proAcc, NCBI protein accession number of the RefSeq
- description, description of the Known Gene.

If a corresponding HUGO (Wain *et al*., 2004) entry can be found for a Known Gene (via the Swiss-Prot/TrEMBL protein ID), the HUGO gene symbol is used as the gene symbol for this Known Gene and the HUGO RefSeq (if available) is recorded in the refseq field of the kgXref table. If there is no corresponding HUGO entry, but a corresponding RefSeq can be found (via mRNA), the gene symbol of this RefSeq will be used. Otherwise, the representative mRNA accession number is used as the gene symbol.

Not every Known Gene has a corresponding RefSeq. LocusLink (now Entrez) data are downloaded from NCBI and an mrnaRefseq table is constructed to store RefSeq–mRNA pairing info for each RefSeq. If a Known Gene's representative mRNA is referenced by a RefSeq, this RefSeq with its corresponding NCBI protein accession number are recorded in the refseq and protAcc fields of the kgXref table. If the gene has a corresponding entry in HUGO, the gene description from HUGO is adopted as description for the Known Gene; otherwise the description for the corresponding protein from Swiss-Prot/TrEMBL is used.

kgAlias, kgProtAlias and kgSpAlias

Genes and proteins are often referred to by different names, symbols and IDs. To facilitate the various cross-reference and searching requirements of many of our own programs and requests from users, we created three comprehensive alias tables, kgAlias, kgProtAlias and kgSpAlias.

The kgAlias table contains just two columns, the Known Gene ID and gene alias. Each Known Gene may have multiple aliases, such as HUGO gene symbol, alternate names and also withdrawn gene names, with the understanding that some users probably still refer to their favorite genes by their old names. The IDs and gene symbol of the RefSeq entries referred to by Known Genes are also added as part of the aliases. Finally, all gene names associated with a Known Gene protein found in Swiss-Prot/TrEMBL are added as gene aliases as well. Each Known Gene–alias pair forms a separate row in the table. Figure 4a shows the relationship between the kgAlias and kgProtAlias tables to various other gene IDs and protein IDs.

Similarly, the kgProtAlias table contains three columns, the Known Gene ID, Swiss-Prot/TrEMBL display ID and protein alias. The protein aliases encompass Swiss-Prot/TrEMBL display IDs, accession numbers, secondary accessions, corresponding PDB IDs if available and the NCBI protein accession numbers of the RefSeq entries referenced by Known Genes.

Recently we added the kgSpAlias table to support the UCSC Proteome Browser for easy searching, i.e. the user can enter either a gene or protein ID/name to search for the desired protein. This table contains three columns; the Known Gene ID, Swiss-Prot/TrEMBL accession number and an alias column which contains both the gene aliases and protein aliases from the kgAlias and kgProtAlias tables.

Other Known Genes related cross-reference tables

In addition to the four cross-reference tables described above, there are a few dozen additional cross-reference tables that are directly related to Known Genes. These tables are used by the hgGene program, which produces the Known Gene Details page, and the UCSC Gene Sorter program. Most of these tables are built as part of the Gene Sorter build process. Figure 4b shows a conceptual entity relationship diagram of the data contained in many database tables related to UCSC Known Genes. A complete list of these tables can be found in the schema specification file, all.joiner, or via the UCSC Table Browser. For example, clicking the 'Filter' button of the UCSC Table Browser shows the following list of 46 tables that are cross linked to UCSC Known Genes:

|		
1.	goaPart	GO objects
2.	all_mrna	Summary info about mRNA alignment
3.	bioCycPathway	BioCyc Pathway to Known Gene cross reference
4.	blastKGPep01	Known Genes protein used for Human Protein tracks of other organisms
5.	blastKGRef01	KG cross-references used for Human Protein tracks of other organisms
6.	ceBlastTab	Tab-delimited blast output file of *Caenorohabditis elegans* proteins
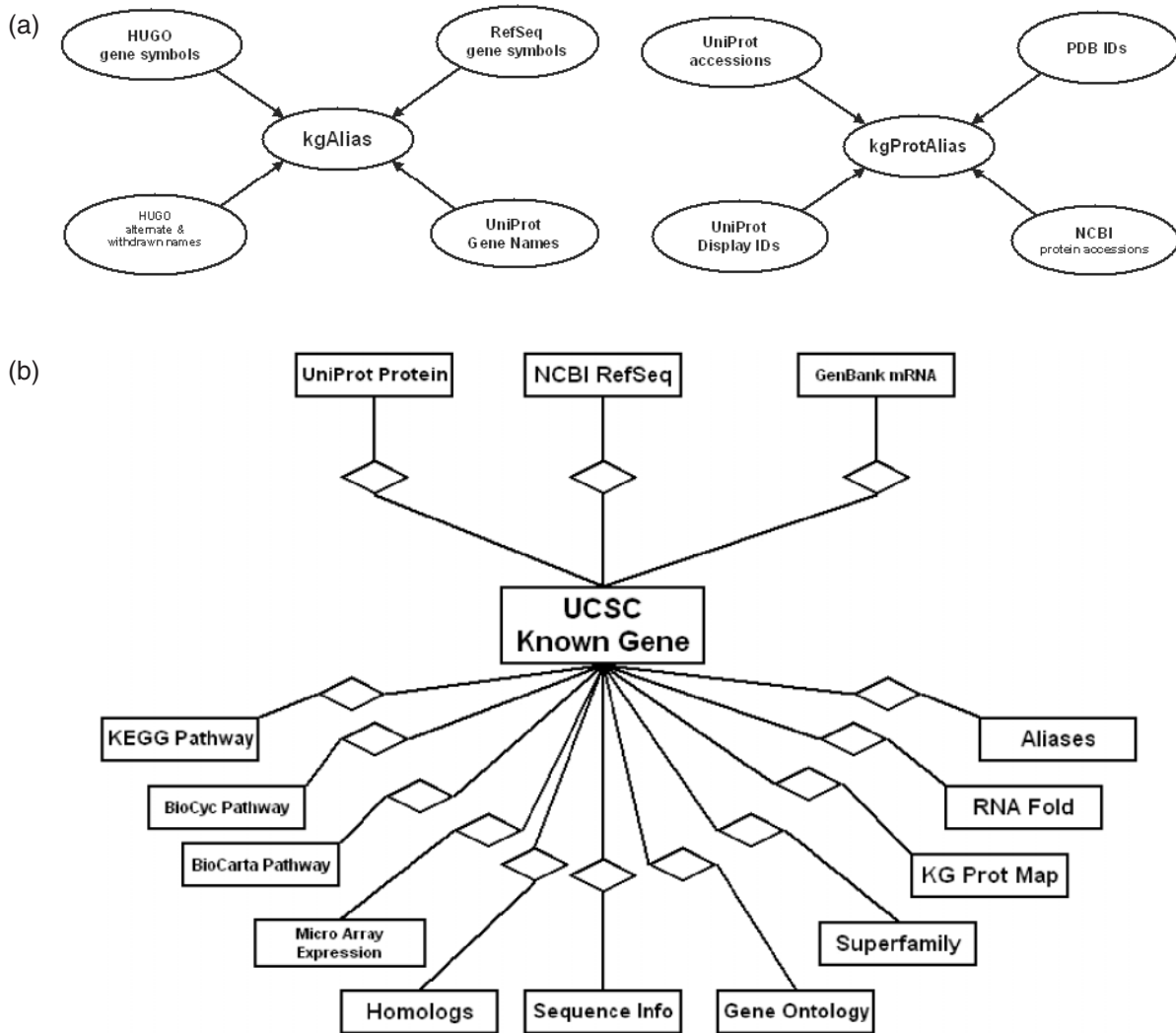
Fig. 4. (**a**) The relationship of kgAlias and kgProtAlias to other gene and protein IDs. (**b**) Conceptual entity relationship diagram of UCSC Known Gene.

7.	dmBlastTab	Tab-delimited blast output file of *Drosophila melanogaster* proteins	16.	gnfU95Distance	Distance between two genes in expression space
8.	drBlastTab	Tab-delimited blast output file of Zebrafish proteins	17.	imageClone	For use with image consortium's cumulative_plate files
9.	dupSpMrna	Duplicated mRNA/protein IDs corresponding to a Known Gene mRNA	18.	keggPathway	KEGG pathway cross reference
10.	foldUtr3	Info about folding of RNA into secondary structure	19.	kgAlias	Link together a Known Gene ID and a gene alias
11.	foldUtr5	Info about folding of RNA into secondary structure	20.	kgProtAlias	Link together a Known Gene ID and a protein alias
12.	gbCdnaInfo	Links together various info associated with a GenBank mRNA or EST	21.	kgSpAlias	Link together a Known Gene ID and either a gene alias or a protein alias
13.	gbSeq	GenBank sequence	22.	kgXref	Link together a Known Gene ID and a gene alias
14.	gbStatus	GenBank sequence status	23.	knownBlastTab	Tab-delimited blast output file
15.	gnfAtlas2Distance	Distance between two genes in expression space	24.	knownCanonical	Describes the canonical splice variant of a gene
			25.	knownGeneLink	Known Genes Link table, currently storing DNA based entries only

26.	knownGeneMrna	A predicted peptide—linked to a predicted gene
27.	knownGenePep	A predicted peptide—linked to a predicted gene
28.	knownIsoforms	Links together various transcripts of a gene into a cluster
29.	knownToEnsembl	Links to Ensembl Genes
30.	knownToLocusLink	Map known gene to some other id
31.	knownToPfam	Links to PFam
32.	knownToRefSeq	Map known gene to some other id
33.	knownToSuper	Links to Superfamily
34.	knownToU133	Map known gene to some other id
35.	knownToU133Plus2	Map Known Genes to U133Plus2 Microarray probes
36.	knownToU95	Map known gene to some other id
37.	mmBlastTab	Tab-delimited blast output file of mouse proteins
38.	mrnaOrientInfo	Extra information on ESTs—calculated by polyInfo program
39.	mrnaRefseq	Cross-reference table between refseq and mRNA IDs based on LocusLink
40.	rnBlastTab	Tab-delimited blast output file of rat proteins
41.	scBlastTab	Tab-delimited blast output file of yeast proteins
42.	seq	Link to sequences
43.	spMrna	The best representative mRNA for a protein
44.	spOldNew	cross-reference between old and new Swiss-Prot IDs
45.	uniProtAlias	UniProt Alias
46.	displayId	UniProt Display ID

The definition of each table can be found in individual .sql files at our download server (see the 'Downloading source code and data' section for details).

UCSC KNOWN GENE DETAILS PAGE

Besides being displayed as one of the major tracks in the Genome Browser main display, each UCSC Known Gene also has an accompanying details page. This details page contains rich and comprehensive information about the chosen gene together with extensive links to other additional web resources. Major sections of this page are shown in Figure 5a–d.

At the top of the details page, Figure 5a, the gene symbol and its description are shown, followed by the representative mRNA and the protein ID. If a corresponding RefSeq entry exists for this gene, the RefSeq summary is also shown. Additional links to other sections of the details page are also shown.

A Quick Links section contains links to the following:

- Genome Browser (UCSC)
- Proteome Browser (UCSC)
- Gene Sorter (UCSC)
- Swiss-Prot (ExPASy/SIB)
- Entrez (NCBI)
- PubMed (NCBI/NLM)

- OMIM (NCBI and Johns Hopkins University)
- GeneLynx
- GeneCards (Weizmann Institute of Science)
- CGAP (NCI/NIH)
- Source (Stanford University)
- MGI (Jackson Labs)

Various comments and descriptive text about the protein encoded by this gene are copied here from Swiss-Prot/TrEMBL and shown. A Sequence section provides links to detailed displays of genomic, mRNA and protein sequences.

A microarray expression section, Figure 5b, shows the absolute and relative expression levels of RNAs in various tissues using several different microarrays, e.g. the GNF Gene Expression Atlas 2 data provided by the Genomics Institute of the Novartis Research Foundation (GNF), which contain two replicates each of 79 human tissues run over Affymetrix microarrays (U133A and GNF1H). As is standard with microarray data red indicates over expression in the tissue and green indicates under expression.

The protein domain and structure information section, Figure 5c, shows links to InterPro and Pfam for further domain information. If there are known or predicted 3D structures, small picture stamps of the structures are shown. Clicking on them will bring the user to PDB (Berman *et al.*, 2000) or ModBase (Pieper *et al.*, 2004) for more detailed protein structural information.

As shown in Figure 5d, a Homolog section contains links to other genomes if homologs are found (determined by best protein sequence match using BLASTP). A Gene Ontology section lists the molecular function, biological process and cellular component information about this gene based on data from the Gene Ontology (GO) (Gene Ontology Consortium, 2004) database. An associated mRNA section lists all mRNAs that overlap with the Known Gene for at least 12 bases. A Pathway section shows links to three major pathway databases, KEGG, NCI/BioCarta and BioCyc where available. Finally a section at the end presents a link for the user to learn more about the UCSC Known Genes.

DOWNLOADING SOURCE CODE AND DATA

The UCSC Genome Browser, Proteome Browser, Gene Sorter and Table Browser are all CGI-based web applications written in C and supported by MySQL databases. The entire source code of these programs and many other utilities are available at http://hgdownload.cse.ucsc.edu/admin/jksrc.zip, which are free for academic, government and non-profit organizations but a license is required for commercial use.

The shell script, KGprocess.sh, which was used to automatically generate the knownGene table and many other associated tables, can be found in the source tree directory src/hg/protein/. The processes to produce other downstream database tables used by the Gene Sorter and Proteome Browser are documented in various files, e.g. makeHg17.doc, makeMm5.doc and makeRn3.doc. These documents are found in the source directory src/hg/makeDb/

All the table definitions and actual data of genomic and proteomic databases used by the major applications mentioned above can be downloaded at http://hgdownload.cse.ucsc.edu/downloads.html

(a)

Home - Genomes - Genome Browser - Gene Sorter - Blat - PCR - Tables - FAQ - Help

Human Gene LEP Description and Page Index

Description: leptin (obesity homolog, mouse)
Representative mRNA: BC060830 **Protein:** P41159 (OB_HUMAN)
RefSeq Summary: This gene is similar to the mouse obesity gene (ob). The protein encoded by this gene is secreted by white adipocytes. In the mouse study, mutations in this gene are linked to severe and morbid obesity.

Page Index	Quick Links	SwissProt Comments	Sequence	Microarray	Protein Structure
Other Species	GO Annotations	mRNA Descriptions	Pathways	Methods	

Quick Links to Tools and Databases

Genome Browser	Proteome Browser	Gene Sorter	SwissProt	LocusLink	Entrez Gene
PubMed	OMIM	GeneLynx	GeneCards	CGAP	Stanford SOURCE
Jackson Labs					

Comments and Description Text from SwissProt

ID: OB_HUMAN
DESCRIPTION: Leptin precursor (Obesity factor) (Obese protein).
FUNCTION: May function as part of a signaling pathway that acts to regulate the size of the body fat depot. An increase in the level of LEP may act directly or indirectly on the CNS to inhibit food intake and/or regulate energy expenditure as part of a homeostatic mechanism to maintain constancy of the adipose mass.
SUBUNIT: Interacts with SIGLEC6.
SUBCELLULAR LOCATION: Secreted.
DISEASE: Defects in LEP may be a cause of autosomal recessive obesity [MIM:601665].
SIMILARITY: Belongs to the leptin family.
DATABASE: NAME=R&D Systems' cytokine mini-reviews: LEP;
WWW="http://www.rndsystems.com/asp/g_sitebuilder.asp?bodyId=213".

Sequence

Genomic (chr7:127,475,283-127,491,632)	mRNA (may differ from genome)	Protein (167 aa)

(b)

Microarray Expression Data

Expression ratio colors: [red high/green low ▾] [Submit]

GNF Expression Atlas 2 Data from U133A and GNF1H Chips

Fig. 5. The details page of a UCSC Known Gene.

(c)

Protein Domain and Structure Information

InterPro Domains: Graphical view of domain structure
IPR009079 - Four-helical cytokine
IPR000065 - Obesity factor

Pfam Domains:
PF02024 - Leptin

Protein Data Bank (PDB) 3-D Structure

1AX8 - X-ray

ModBase Predicted Comparative 3D Structure on P41159

| Front | Top | Side |

The pictures above may be empty if there is no ModBase structure for the protein. The ModBase structure frequently covers just a fragment of the protein. You may be asked to log onto ModBase the first time you click on the pictures. It is simplest after logging in to just click on the picture again to get to the specific info on that model.

(d)

Homologous Genes in Other Species (BLASTP Best Hit)

Mouse

Gene Ontology (GO) Annotations with Structured Vocabulary

Molecular Function:
GO:0005179 hormone activity

Biological Process:
GO:0006112 energy reserve metabolism
GO:0007165 signal transduction
GO:0007267 cell-cell signaling

Cellular Component:
GO:0005576 extracellular
GO:0005615 extracellular space

Descriptions from all associated mRNAs.

U43653 - Human obese protein (ob) mRNA, complete cds.
BC060830 - Homo sapiens leptin (obesity homolog, mouse), mRNA (cDNA clone MGC:71704 IMAGE:30333193), complete cds.
AF008123 - Homo sapiens obese protein (ob) mRNA, complete cds.
D49487 - Homo sapiens obese mRNA, complete cds.
U18915 - Human obese (ob) mRNA, complete cds.
BC069323 - Homo sapiens leptin (obesity homolog, mouse), mRNA (cDNA clone MGC:96888 IMAGE:7262097), complete cds.
BC069452 - Homo sapiens leptin (obesity homolog, mouse), mRNA (cDNA clone MGC:96900 IMAGE:7262109), complete cds.
BC069527 - Homo sapiens leptin (obesity homolog, mouse), mRNA (cDNA clone MGC:96912 IMAGE:7262121), complete cds.

Biochemical and Signalling Pathways

BioCarta from NCI Cancer Genome Anatomy Project
h_leptinPathway - Reversal of Insulin Resistance by Leptin

UCSC Known Genes Methods, Credits, and Data Use Restrictions

Click here for details.

Fig. 5. *Continued*

ANALYSIS

The genomic coverage of UCSC Known Genes is compared with other gene sets using the featureBits analysis utility. As shown in Table 1, out of 2 866 216 770 bases of human genome (May 2004 Assembly), 2.293% bases are covered by Known Genes exons (including UTRs), which is 31% more than the RefSeq coverage of 1.746%. In comparison, the Ensembl Gene Predictions track covers 1.985% and the H-Invitational Genes (Imanishi *et al.*, 2004) track covers 2.171% of the total genome. We also included the gene set (CDS only) from the CCDS (Consensus CDS) project (Pruitt,K., Maglott,D., Harrow,J., Diekhans,M., Birney,E., Baertsch,R., Haussler,D., Hubbard,T., Searle,S., Siepel,A., Ostell,J. and Durbin,R., manuscript in preparation, http://www.ncbi.nlm.nih.gov/CCDS/), a highly curated but conservative subset of RefSeq, with a total genomic coverage of 0.704%.

Similar analysis is performed on the coverage of the coding regions (CDS only) of different gene sets. As shown in Table 2, the CDS regions of RefSeq cover 1.018% of human genome, while UCSC Known Genes and Ensembl genes each offers about 15% more CDS coverage, at 1.183 and 1.182%.

The number of entries in human UCSC Known Genes track is 44 338, which is almost double of the number for RefSeq, 23 111. Visual examination of the Known Genes track compared with the RefSeq track shows that typically for a given gene locus, the Known Genes often include multiple entries with different exon structures. Compared with a manual curation process, e.g. RefSeq, the current automated Known Genes build process could not differentiate and identify true alternative splicing isoforms at a gene locus among multiple transcripts of slight variations or imperfect exon structures. As part of the data we generate for the UCSC Gene Sorter (Kent *et al.*, 2005), Known Genes are clustered based on their genomic positions and one canonical gene is assigned to each cluster. This result is stored in the knownCanonical table, which has 21 680 entries, comparable with the total number of RefSeq entries of 23 105.

In Table 1, we also used RefSeq as the gold standard and compared it with UCSC Known Genes and other gene sets. In terms of RefSeq bases covered, UCSC Known Genes has the highest overlap at 94.4%. Ensembl Genes overlap coverage with RefSeq is 92.1% and H-Invitational gene set covers 65.1% of RefSeq. When we compared the coding regions only, we found both UCSC Known Genes and Ensembl genes have high coverage on RefSeq CDS regions at 97.6 and 96.5%, as shown in Table 2.

Table 3 shows the comparison of UCSC Known Genes with other gene sets. In terms of the percent genome bases covered, the UCSC Known Genes missed 0.098% when compared with RefSeq bases, 0.221% when compared with Ensembl Gene bases and 0.754% when compared with H-Invitational Gene bases. On the other hand, the Known Genes also offers additional bases, which are not covered by RefSeq (0.645%, 6.5 times higher than the number of RefSeq bases missed by Known Genes), or not covered by Ensembl genes (0.529%, 2.4 times higher) and not by H-Invitational (0.876%). Table 4 shows the comparison of CDS only regions. H-Invitational was not included because of incomplete CDS data. Significantly more CDS regions were covered by Known Genes when compared with RefSeq. When compared with Ensembl Genes, each offers similar amount of unique CDS regions.

Table 1. Genomic coverage of gene sets, human (May 2004 release)

CDS and UTR	RefSeq	Ensembl Genes	H-Invitational	UCSC Known Genes
Gene Count	23 105	33 666	44 811	44 338
Percentage of RefSeq	100.0	145.7	193.9	191.9
Base Coverage (%)	1.746	1.985	2.171	2.293
Pecentage of RefSeq	100.0	113.6	124.3	131.3
Overlap with RefSeq	1.746	1.612	1.137	1.649
Percentage of RefSeq	100.0	92.3	65.1	94.4

Table 2. Genomic coverage of gene sets, CDS only human (May 2004 release)

CDS Only	RefSeq	CCDS	Ensembl Genes	UCSC Known Genes
CDS Base Coverage (%)	1.018	0.704	1.182	1.183
Percentage of RefSeq	100.00	69.20	116.10	116.20
Overlap with RefSeq (%)	1.018	0.704	0.982	0.994
Percentage of RefSeq	100.00	69.20	96.50	97.60

Table 3. Genomic coverage comparison, CDS and UTR, human (May 2004 release)

CDS and UTR	RefSeq	Ensembl genes	H-Invitational
Percentage of genome bases unique to UCSC Known Genes	0.645	0.529	0.876
Percentage of genome bases unique to the other gene set	0.098	0.221	0.754

Table 4. Genomic coverage comparison, CDS only, human (May 2004 release)

CDS only	RefSeq	CCDS	Ensembl genes
Percentage of genome bases unique to UCSC Known Genes	0.188	0.492	0.115
Percentage of genome bases unique to the other gene set	0.023	0.013	0.114

We are pleased to see that our automated process could capture close to 95% of genomic bases of human RefSeq genes. We also noticed the significant difference between UCSC Known Genes and the H-Invitational Genes. Using the Table Browser, we pulled out

Table 5. Genomic coverage of gene sets, mouse (May 2004 release)

	RefSeq	Ensembl gene	UCSC Known Genes
Gene count	18 103	31 035	41 208
Percentage of RefSeq	100.0	171.4	227.6
Base coverage (%)	1.562	2.002	2.184
Percentage of RefSeq	100.0	128.2	139.8
Overlap with RefSeq	1.562	1.388	1.486
Percentage of RefSeq	100.0	88.9	95.1

Table 7. Genomic coverage of gene sets, rat (June 2003 release)

	RefSeq	Ensembl gene	UCSC Known Genes
Gene count	7166	28 545	8348
Percentage of RefSeq	100.0	398.3	116.5
Genome base coverage (%)	0.523	1.210	0.523
Percentage of RefSeq	100.0	231.4	100.0
Overlap with RefSeq	0.523	0.392	0.396
Percentage of RefSeq	100.0	75.0	75.7

Table 6. Genomic coverage comparison, mouse (May 2004 release)

	RefSeq	Ensembl gene
Percentage of genome bases unique to UCSC Known Genes	0.699	0.498
Percentage of genome bases unique to other gene set	0.077	0.316

Table 8. Genomic coverage comparison, rat (June 2003 release)

	RefSeq	Ensembl gene
Percentage of Genome bases unique to UCSC Known Genes	0.127	0.159
Percentage of Genome bases unique to other gene set	0.127	0.846

H-Invitational genes that are not covered by the UCSC Known Genes and found most of them having only mRNA evidence without direct protein evidence. These genes are not expected to show up in UCSC Known Genes because our process requires the existence of a representative protein–mRNA pair of a gene be considered as a UCSC Known Gene candidate. Likewise, out of a total of 57 401 human proteins listed in Swiss-Prot/TrEMBL, ∼18% (10 462) of them were not included in our initial list of Known Gene candidates because they do not have any external reference to Genbank mRNAs. These findings suggest that all the major human gene lists included in our analysis may not be complete.

Similar analyses were done on mouse and rat genomes. The results are listed in Tables 5 and 6, and Tables 7 and 8, respectively. For mouse genome, the Known Genes have comparable RefSeq coverage (95.1%). For rat genome, the RefSeq genomic coverage itself is relatively low (0.523%) and only 75.7% of the RefSeq bases are covered by UCSC Known Genes. This probably owes to the fact that the UCSC Known Genes for rat had not been updated for more than a year and the base genome data of rat (June, 2003) was a relatively incomplete one.

FUTURE DIRECTIONS

The coverage of the UCSC Known Genes dataset could be further extended. We would like to develop additional processing steps to accomplish this. For example, we could include the large number of Swiss-Prot/TrEMBL proteins, which do not have supporting mRNA evidence listed under the external cross-reference section, as Known Genes candidates. Likewise, we would also like to consider including the full-length cDNAs in other sources, e.g. H-Invitational and MGC, which are not covered by the Swiss-Prot/TrEMBL external cross-references as part of our initial gene candidates dataset.

We would also like to implement additional selection criteria to mimic expert curators' thought processes to eliminate false positives or redundant data to increase the quality of the resulting dataset.

Although the build time, about a day, of the UCSC Known Genes for a genome is relatively short because of process automation and utilization of the UCSC Kilo Kluster, the Known Genes are not updated as frequently as we would like to owing to practical resource constraints and need of time to update and QA the associated downstream annotation data. Nevertheless, we would like to further automate the processing steps of those downstream annotation data so that we can update the UCSC Known Genes and the associated downstream annotations more frequently.

CONTACTING US

The mailing list genome@cse.ucsc.edu provides a forum for questions and discussion about the UCSC Genome Browser, Proteome Browser, Table Browser, Gene Sorter and databases. Users may subscribe to this list at http://www.cse.ucsc.edu/mailman/listinfo/genome. For announcement of new releases and features, a lower volume mailing list, genome-announce@soe.ucsc.edu is available. To report problems on accessing the website, servers or mirror sites, or for correspondence inappropriate for the public forum, send email to genome-www@cse.ucsc.edu

ACKNOWLEDGEMENTS

REFERENCES

Bairoch,A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucl. Acids Res.*, **33**, D154–D159.

Benson,D.A. *et al.* (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.

Berman,H.M, Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., Bourne,P.E. (2000) The Protein Data Bank, *Nucleic Acids Res.*, **28**, 235-242.

Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource, *Nucleic Acids Res.*, 32, D258-D261.

Hsu,F. *et al.* (2005) The UCSC Proteome Browser. *Nucleic Acids Res.*, **33**, D454–D458.

Imanishi,T. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones, *PLoS Biology,* **2,** 856–875.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Karolchik,D. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.

Karolchik,D. *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Kent,W.J. *et al.* (2005) Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.*, **15**, 737–741.

Kent,W.J. *et al.* (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

Pieper,U., Eswar,N., Braberg,H., Madhusudhan,M.S., Davis,F.P., Stuart,A.C., Mirkovic,N., Rossi,A., Marti-Renom,M.A., Fiser,A., Webb,B., Greenblatt,D., Huang,C.C., Ferrin,T.E. and Sali,A. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids. Res.*, **32**, D217–D222.

Pruitt,K.D. *et al.* (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.

Wain,H.M. *et al.* (2004) Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res.*, **32**, D255–D257.