Software

Integrating computationally assembled mouse transcript sequences with the Mouse Genome Informatics (MGI) database

Yunxia Zhu*, Benjamin L King*, Babak Parvizi†, Brian P Brunk‡, Christian J Stoeckert Jr‡, John Quackenbush§, Joel Richardson* and Carol J Bult*

Addresses: *Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME 04609, USA. †Invitrogen Corporation, 1610 Faraday Ave, Carlsbad, CA 92008, USA. ‡Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104, USA. §The Institute for Genomic Research, Rockville, MD 20850, USA.

Correspondence: Yunxia (Sophia) Zhu. E-mail: yz@informatics.jax.org

Published: 3 February 2003

Genome ***Biology*** 2003, **4**:R16

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/2/R16

Received: 9 October 2002
Revised: 27 November 2002
Accepted: 19 December 2002

Abstract

Databases of experimentally generated and computationally derived transcript sequences are valuable resources for genome analysis and annotation. The utility of such databases is enhanced when the sequences they contain are integrated with such biological information as genomic location, gene function, gene expression and phenotypic variation. We present the analysis and results of a semi-automated process of connecting transcript assemblies with highly curated biological information for mouse genes that is available through the Mouse Genome Informatics (MGI) database.

Rationale

The volume and diversity of expressed sequence tag (EST) data in the public databases makes them an important resource for gene identification, genome annotation and comparative genomics. The value of EST data is enhanced when the sequences are clustered (on the basis of sequence overlap) to reduce redundancy. In some cases these sequence clusters can be used to generate a consensus sequence that represents a virtual transcript. Examples of electronic transcript data resources include UniGene [1], TIGR Gene Indices [2,3], DoTS [4] and STACK [5]. Each of these resources differs in the methods used to reduce the redundancy in EST sequence data and in how the data are represented. For example, UniGene uses pairwise sequence comparisons to group and partition EST and other transcript sequences from GenBank into gene-orientated clusters with no consensus sequence. The other three resources (TIGR Gene Indices, DoTS, and STACK) cluster sequences from ESTs and known transcripts and then assemble the members of each cluster to produce a consensus representation of the transcripts. The algorithms and/or parameters used to guide the clustering and assembly process for the virtual transcript resources are similar, but not identical. These resources also differ with respect to the number of species for which EST assemblies are available. For example, the STACK database includes only human sequence data, DoTS has both human and mouse assemblies, and TIGR Gene Indices maintains separate electronic transcript databases for over 50 species. In contrast to computational approaches to transcript analysis and representation, the Mammalian Gene Collection (MGC) [6] and the RIKEN Mouse Encyclopedia projects [7] are systematically generating full-length cDNA clones with the aim of having at least one full-length clone reagent and sequence for every human (MGC) and mouse (MGC, RIKEN) gene.

The Mouse Genome Informatics (MGI) database [8] provides integrated access to genetic, genomic and biological

data for the laboratory mouse. The MGI database represents an integrated platform to which several related projects contribute, including the Mouse Genome Database (MGD) [9], the Gene Expression Database (GXD) [10], the Mouse Genome Sequence (MGS) project [11], the Mouse Tumor Biology (MTB) database [12], and the Gene Ontology (GO) project [13]. MGI provides access to gene annotation and nomenclature, mapping, nucleotide and protein sequences, mammalian gene homology, gene expression, phenotypes, allelic variants and mutants and strain data. The information in the MGI database is updated daily by professional scientific curators who extract relevant data from the scientific literature and other sources. MGI staff curate associations between genes and nucleotide and protein sequences in collaboration with other database groups, including SWISS-PROT [14], RIKEN [7], and the National Center for Biotechnology Information's LocusLink [15,16]. Genes in MGI are given unique, permanent accession ids to facilitate stable cross-references with other databases even when such information as gene name, functional annotation, and so on changes over time [11]. Table 1 shows a summary of some of the MGI database content.

Table 1

Selected database content statistics for the MGI information resource (as of 11 October 2002).

Category	Number
References	74,845
Genetic markers	51,398
Genes	31,708
Genetic markers mapped	41,342
Genes mapped	22,645
Curated mouse/human orthologs	7,566
Genes with molecular probes and segments data	25,672
Number of genetic markers with molecular polymorphisms	12,718
Number of genes with molecular polymorphisms	3,599
MGI markers with GenBank sequence associations	29,144
Genes with SwissProt-TrEMBL protein sequences	13,633

The database content of MGI is updated daily. The current database content statistics can be found at the MGI FTP site (MGI Data and Statistical Reports). MGI contains information on genetic markers (such as sequence-tagged site (STS) markers), genes and other genomic features.

The utility of both experimentally and computationally derived transcript resources are greatly enhanced when the transcripts are associated with well curated biological knowledge about the genes with which the transcripts are associated [11]. However, manual curation of computationally derived transcript data is not feasible because the underlying data for these resources are constantly changing. Therefore, we have developed a semi-automated curation process to create and update associations between constantly changing electronic transcript databases and the genes represented in the MGI database. Associations are based on GenBank sequence accession identifiers shared between MGI genes and transcript clusters/assemblies Although associations between the genes in MGI and the electronic transcripts could also be made on the basis of sequence similarity, the use of shared accession ids is faster and avoids inconsistencies in sequence-to-gene associations that arise from highly similar sequence among members of multigene families.

Transitive associations between MGI genes and assembled transcripts

To establish associations between MGI genes and the Institute for Genomic Research (TIGR) mouse gene index tentative consensus sequence (TCs) or DoTS mouse transcript assemblies (DTs), we used GenBank sequence accession identifiers (GB) that are associated with genes in the MGI database and are also component sequences of transcript assemblies as bridges. All gene-to-transcript associations are represented as a set of graphs (Figure 1). Nodes of the graph are members of a group of interrelated MGI genes and transcript assemblies, and each edge is a group of GenBank sequence accession identifiers that are shared by the related MGI gene and the transcript assembly.

We first generated a set of GenBank sequence accession identifiers that have trusted associations (that is they have been manually curated) with mouse genes represented in MGI. A daily report of the associations of MGI markers and GenBank sequences, MRK_Sequence.rpt, is available from the MGI public FTP site [17]. We removed sequences associated with more than one gene object in MGI (for example, large cloned inserts containing multiple genes) to avoid confounding multiple genes to sequence associations. After this filtering step, the relationships of MGI genes to GenBank sequences were maintained in a dictionary data structure, MGI-GB, with MGI accession numbers as keys and GenBank accession identifiers as values.

A second dictionary, GB-MGI (GB as keys and MGI as values), was generated by reversing the keys and values of MGI-GB. A report with all DT identifiers and their constituent GenBank sequences accession identifiers, musDoTS_rel5_accessionsPerAssembly.dat.gz file (Release 5.0, 19 August 2002), was downloaded from CBIL's website [18].

A report containing TIGR Mouse Gene Index TC identifiers and their constituent GenBank sequence identifiers was generated from TIGR Mouse Gene Index Release 9.0 (October 1, 2002) (Geo Pertea, personal communication). This report was used to create two dictionaries, TC-GB (TC as keys and GB as values) and GB-TC (GB as keys and TC as values) to map TCs to GenBank sequence accession ids.

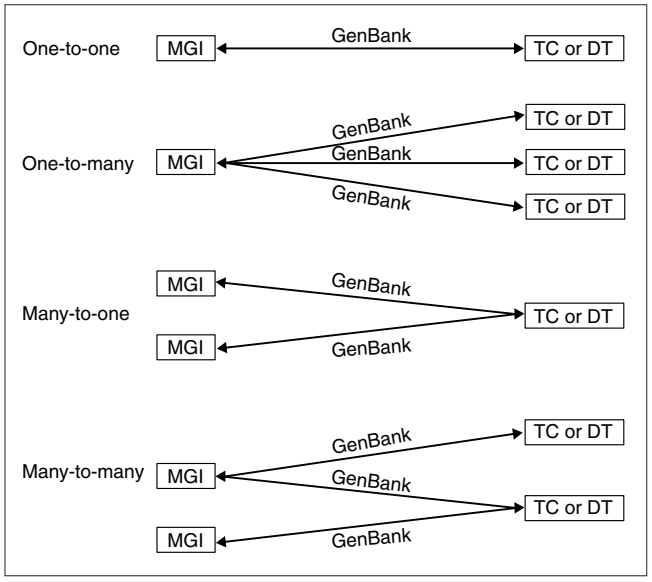

Figure 1
Association of MGI genes with TIGR mouse TCs or DoTS mouse DTs through the shared references of GenBank accession identifiers can be represented as a set of graphs. The associations can be classified into four categories: one-to-one, one-to-many, many-to-one, and many-to-many.

There were many more GenBank sequences (mostly ESTs) in the TIGR Mouse Gene Index than in the MGI database because the TIGR Mouse Gene Index was built on all available GenBank sequences, and MGI curated only a subset of them (mostly mRNA and RefSeq sequences). Only GenBank sequence accession identifiers that appear in the MGI database were retained in TC-GB and GB-TC because those sequences will bridge the transitive associations between MGI genes and TCs. Sequences associated with more than one TC were removed.

Two of the dictionaries described above, MGI-GB and GB-TC, were used to link MGI genes to TCs on the basis of shared GenBank sequences. A dictionary, MGI-TC-via-GB (MGI as keys and TC as values), was used to maintain the gene-to-transcript associations and their supporting GenBank sequences. For each GenBank accession identifier in MGI-GB, the related TC identifier in GB-TC was retrieved and added as a value to the MGI accession identifier key, keeping the GenBank accession identifiers as a line of evidence in support of the link. When more than one GenBank sequence supports the same MGI-to-TC link, all were attached to the same TC. Figure 2a shows examples of links from MGI genes to TCs with the supporting GenBank sequences. Of MGI genetic markers with GenBank sequences, 71.3% (20,772 out of 29,144) of were linked to one or more TCs in this analysis. The majority of MGI markers with no TC associations have only one GenBank sequence, which is either singleton mRNA/EST sequence with no TC accession number assigned or DNA sequence excluded from building TIGR Mouse Gene Index. Table 2 summarizes the associations between MGI markers and TIGR Mouse Gene Index TCs.

Two other dictionaries, TC-GB and GB-MGI, were used to link TCs to MGI genes in the same way. A dictionary, TC-MGI-via-GB (TC as keys and MGI as values), was used to maintain links from TCs to MGI genes and the supporting

(a)

MGI:101764	TC608273::R74993::U89527::R74987::R74988::R74992
MGI:101784	TC577815::AK012622 TC601026::AK009706::AF076623::AA166324::BC022629::C78523
MGI:2142452	TC639728::AA960159
MGI:96610	TC639728::U47283::Y00769::U37029::X15202

(b)

TC567945	MGI:1919829::AU022477
TC635728	MGI:87904::J04181
TC635741	MGI:87904::U89400::AA709861::M12481::X03672::AA590859::X03765
TC639728	MGI:2142452::AA960159 MGI:96610::U37029::X15202::Y00769::U47283

Figure 2
Examples of MGI-to-TC and TC-to-MGI associations with supporting GenBank sequences. **(a)** MGI genes may associate with zero, one or more TCs. Each association is supported by one or more GenBank sequences that are shared by the MGI gene and the related TC. For example, the association of MGI gene *Nes* (nestin; MGI:101784) with TC577815 is supported by AK012622 and with TC601026 is supported by AK009706, AF076623, AA166324, BC022629 and C78523. **(b)** TCs may associate with zero, one or more MGI genes. Each association is supported by one or more GenBank sequences that are shared by the TC and the related MGI gene.

Table 2

Statistics of associations between MGI genes and transcript assemblies (as of 11 October 2002)

Datasets	TIGR TCs	DoTS DTs
Sequences used to build TCs and DTs	2,611,422	2,495,338
Sequences included in the assemblies (excluding singletons)	2,254,999	2,044,540
Assemblies (excluding singletons)	105,520	128,341
GenBank sequences shared by MGI markers and assemblies	43,200	52,754
MGI genes linked to assemblies through GenBank sequences	20,783	24,340
Assemblies linked to MGI genes through GenBank sequences	20,942	25,799

GenBank sequences. Figure 2b shows examples of links from TCs to MGI genes with the supporting GenBank sequences. 19.8% (20,942 out of 105,520) of TCs (excluding singletons) were linked to one or more MGI genes.

The same approaches were also used to associate MGI genetic markers to DoTS DTs. A report with all DT identifiers and their constituent GenBank sequence accession identifiers was downloaded from CBIL's website (Release 5.0, 19 August, 2002). The musDoTS_1-7-02_containedIds.dat.gz file can be downloaded from this site [18]. The report lists both DoTS assemblies (excluding singletons) and singletons. We included only assemblies in our analysis. Statistics of associations of MGI markers and DTs are shown in Table 2. The analysis linked 83.5% (24,340 out of 29,144) of MGI markers with sequence information to 20.1% (25,799 out of 128,341) of DTs. It is not surprising that only about 20% of DTs or TCs can be associated with genes in MGI because the majority of the assemblies are composed solely of EST sequences and the MGI curation processes focus primarily on collecting and curating associations with genomic and mRNA sequence data. There are a total of 26,440 DTs (including singletons) with mRNA sequences and 20,908 of them have MGI associations. The remaining 5,532 DTs with mRNA sequences might represent alternative transcripts of MGI genes, known genes not yet represented by MGI, or novel genes. We will evaluate the component sequences of these DTs and incorporate them into MGI database over time through manual curation.

Classification of the relationships between MGI genes and transcript assemblies

We used bipartite graphs to represent the relationships between MGI genes and TCs (or DTs) (Figure 1). All the related MGI genes and TCs (or DTs) were represented as a node of one graph, which links MGI genes and TCs (or DTs) when they share common GenBank sequences. The accession identifiers of GenBank sequences that support the links were attached to MGI or TC identifiers. The relationships between MGI genes and TCs (or DTs) were categorized into subsets of one-to-one, one-to-many, many-to-one, and many-to-many associations (Table 3).

The majority of the associations were one-to-one relationships: 16,996 MGI-to-DT and 13,451 MGI-to-TC. Among these, a large number of MGI genes (9,509 in MGI-to-DT and 5,742 in MGI-to-TC associations) have only single GenBank sequence. The remaining MGI genes in one-to-one category have two or more GenBank sequences associated with them. The one-to-one associations between MGI genes and TCs/DTs suggests that these genes have only one form of transcript or that the data needed to detect transcript variants is not yet available in public databases.

One-to-many associations between genes and transcripts are related to transcript diversity

The TIGR Mouse Gene Index and DoTS databases are transcript orientated. That is, the sequence clustering and assembly process seeks to generate distinct assemblies for every form of transcript. The MGI database is gene-centric and associates transcripts from the same locus to a single gene object in the database. Therefore, in many cases, there are multiple TCs/DTs associated with a single gene in the MGI database. The average numbers of DTs/TCs per MGI gene among the one-to-many associations were 2.29 and 2.24, respectively.

Multiple TCs/DTs associated with a single MGI gene often represent alternatively spliced transcripts. For example, *Ncam1* (neural cell adhesion molecule 1) in the MGI database (MGI:97281) was associated with five TCs (TC549908, TC582634, TC582635, TC640342, TC640343) and with five DTs (DT.487850, DT.87072470, DT.87072472, DT.97397085, and DT.97411237). *Ncam1* is known to exist in three prominent protein isoforms encoded by at least four different transcripts generated from alternative splicing [19]. At least

Table 3

Classification of associations between MGI genes and both DT and TC gene indices (as of 11 October 2002)

Datasets	TIGR	DoTS
One-to-one MGI gene to assembly	13,451	16,996
One-to-many MGI gene to assembly*	1,975	2,522
Many-to-one MGI gene to assembly†	1,932	1,675
Many-to-many MGI gene to assembly‡	454	531

*The link of one MGI gene to multiple assemblies is counted as one association. †The link of multiple MGI genes to one assembly is counted as one association. ‡The link of multiple MGI genes to multiple assemblies is counted as one association.

eight distinct mRNAs have been identified by a variety of analyses, and 24 potential transcripts have been proposed [20]. These one-to-many associations caused by alternative splicing (and/or alternative poly(A) addition sites, see below) were confirmed by mapping the TCs or DTs to the mouse genome assembly using BLAT search at the University of California Santa Cruz (UCSC) genome browser [21]. Figure 3a shows how these virtual transcripts align to the *Ncam1* gene on mouse chromosome 9.

Multiple TCs or DTs could also represent products of transcription from alternative promoters and/or polyadenylation sites of a single gene. For example, *Dtna* (dystrobrevin alpha) in the MGI database (MGI:106039) was associated with five TCs (TC546133, TC569762, TC577975, TC590157, TC633947) and with five DTs (DT.50316348, DT.60101497, DT.87050693, DT.91340878 and DT.91393353). The *Dtna* gene has three promoters that are active in tissue-specific manner [22]. Figure 3b clearly shows multiple transcripts from three promoters of *Dtna* gene on mouse chromosome 18. Experimental results suggested that *Ncam1* contains more than one poly(A) addition site and produces transcripts with different 3′ untranslated regions [19]. Figure 3a demonstrates that multiple virtual transcripts with different 3′ ends align the *Ncam1* gene on mouse chromosome 9.

Another explanation for one-to-many MGI gene to DTs/TCs associations is multiple site-specific recombination or DNA rearrangement that occurs normally in certain cell types. For example, the *Igh-VS107* (immunoglobulin heavy chain (S107 family)) locus in MGI (MGI: 96490) was associated with four TCs (TC632874, TC632875, TC632877 and TC643641) and with three DTs (DT.94166135, DT.94209475 and DT.94398318). All above TCs/DTs and sequences associated with *Igh-VS107* were mapped to the same locus on chromosome 12 using the UCSC genome browser (data not shown). The sequence differences of *Igh-VS107* transcripts are readily explained by normal DNA rearrangements (V(D)J recombination) [23]. Another example is that *H2-Eb1* (histocompatibility 2, class II antigen E beta) in the MGI database (MGI:95901) was associated with four TCs (TC575977, TC608775, TC638140 and TC640785) and with two DTs (DT.493389 and DT.55100612). The *H2-Eb1* gene contains a recombination hotspot, which has a predominant role in generating different recombinants through meiotic crossing-over within the I region of the mouse major histocompatibility complex (MHC) [24]. All above TCs/DTs except

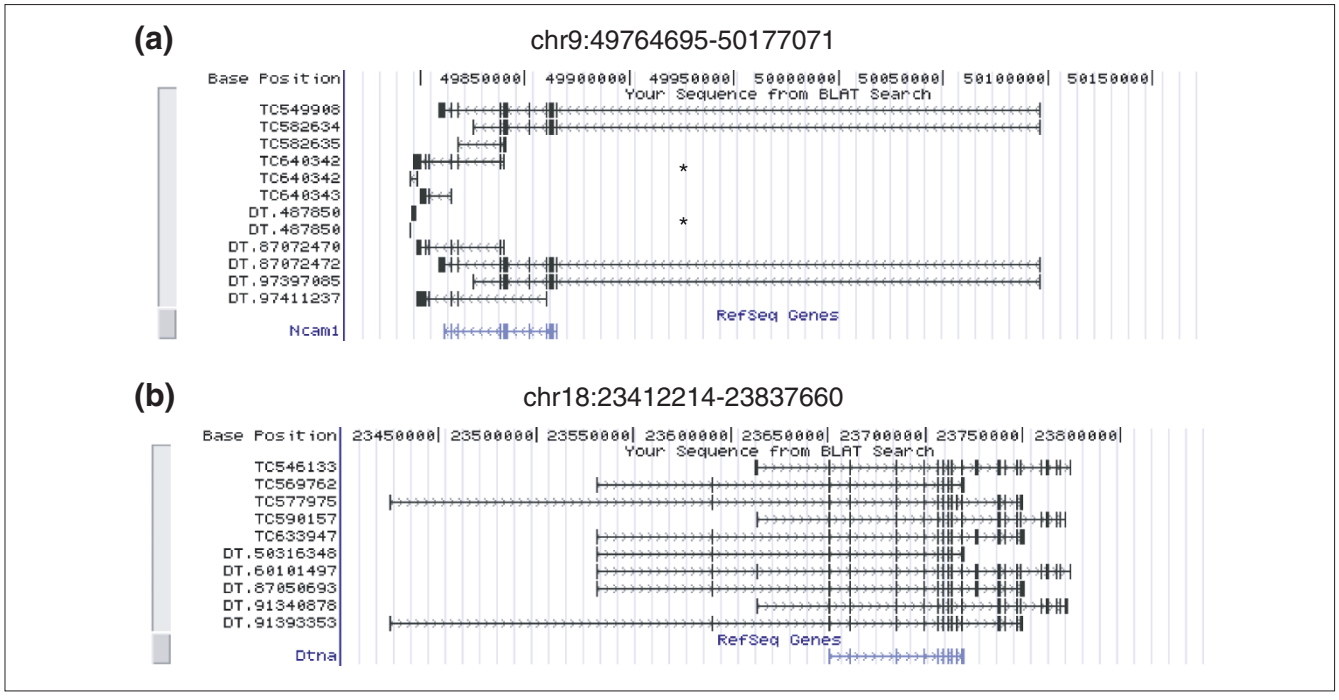

Figure 3
Transcripts can be aligned to the mouse genome assembly using BLAT search at the UCSC Genome Browser. Aligning regions (usually exons) are shown as black blocks. The aligning regions are connected by lines representing gaps (usually spliced-out introns), with arrowheads indicating the direction of transcription. **(a)** The alignment of TCs and DTs associated with MGI gene *Ncam1* (MGI:97281) to the annotated *Ncam1* gene on chromosome 9 shows that alternative spliced exons and alternative poly(A) addition sites cause multiple transcripts from one gene. The tracks of *TC640342 and *DT.487850 are matches with lower percentage identity over a shorter region of the sequence. **(b)** The alignment of TCs and DTs associated with the MGI gene *Dtna* (MGI:106039) to the annotated *Dtna* gene on chromosome 18 demonstrates that three alternative promoters are actively used, as suggested by published experimental results.

TC575977 and DT.55100612 were mapped to annotated *H2-Eb1* gene on chromosome 17 using the UCSC genome browser (data not shown). TC575977 and DT.55100612 were associated to *H2-Eb1* through one GenBank sequence AK012147, which was mapped to chromosome 5. Further analysis indicated that AK012147 was incorrectly associated with *H2-Eb1*.

Non-biological explanations may also explain some of the one-to-many associations among genes and transcripts in our analysis. For example, low-quality sequence data or problematic sequences that are not filtered out before clustering and assembly can cause errors in the sequence assemblies. Another possible explanation is that nucleotide and protein sequences are occasionally associated with the wrong gene in MGI. This will always be a challenge to the database community when both completeness (including as much data as possible in the database) and accuracy (associating every sequence to the right gene) are goals. Fortunately, non-biological reasons for one-to-many associations between transcripts and genes only account for a small percentage of the whole dataset based on our experience of ongoing internal quality control and manual check of portions of the data in this analysis.

The many-to-one associations between genes and transcripts are evidence for over-clustering or gene redundancy in MGI

In the analysis reported here, 14.8% (4,302 out of 29,144) of MGI genes with sequence information were involved in many-to-one gene to TC transcript associations and 12.4% (3,621 out 29,144) in many-to-one gene to DT transcript associations. The average numbers of MGI genes per DT/TC were 2.16 and 2.23, respectively. First, some of the many-to-one associations are due to sequence clusters that contain mistakenly grouped similar sequences from closely related genes (paralogs). For example, sequences from 14 members of the defensin-related cryptdin gene family were clustered into one TC (TC611932) and to one DT (DT.94272645). These genes share similar structure and sequence. Their mRNAs are distinguished by a 45-nucleotide 5′ untranslated sequence (UTS) encoded completely by the first exon [25]. Second, genes adjacent to each other in the same chromosomal location were occasionally clustered together and assembled into one sequence because their transcripts overlap each other. For example, the 3′ end of *Stk11* (serine/threonine kinase 11; MGI:1341870) is in very close proximity to the 3′ end of a functionally unrelated gene *Dos* (downstream of *Stk11*; MGI:1354170) and it seems that overlapping transcripts of the two genes are produced [26]. Both *Stk11* and *Dos* were linked to one DT (DT.493186) because sequences associated with both genes were clustered and assembled together. Third, there are rare cases of polycistronic transcripts in mammalian genomes. For example, *Snrpn* (small nuclear ribonucleoprotein N; MGI:98347) and *Snurf* (*SNRPN* upstream reading frame; MGI:1891236) are expressed as bicistronic *Snurf-Snrpn* transcript [27] and both of them were associated with one single TC (TC619385) and one single DT (DT.535946) in our analysis. Finally, many-to-one associations can be caused by uncorrected gene redundancy (one gene represented by multiple entries) in the MGI database. The majority of the redundant records are the result of genes in MGI that are represented solely by EST sequences. As these redundancies are identified in MGI they are corrected.

Many-to-many associations between genes and transcripts could be the result of any combination of many-to-one and one-to-many associations

In the analysis reported here, we had 531 MGI-DT and 454 MGI-TC many-to-many associations. There were 4.1% (1,202 out of 29,144) of MGI genes with sequence information and 1,355 DTs involved in many-to-many MGI gene to DT transcript associations and 3.6% (1,044 out 29,144) of MGI genes with sequence information and 1,127 TCs in many-to-many MGI gene to TC transcript associations. The average numbers of MGI genes per DT/TC were 2.26 and 2.30, respectively, and the average numbers of DTs/TCs per MGI gene were 2.55 and 2.48, respectively. The majority of the many-to-many associations had only two MGI genes and two DTs/TCs. The group with the largest number of MGI genes in MGI-DT associations included eight paired-Ig-like receptor A genes (*Pira1*, *Pira2*, *Pira3*, *Pira4*, *Pira5*, *Pira7*, *Pira10*, *Pira11*) and two DTs (DT.87053023 and DT.94272531). DNA blot analysis indicated the presence of multiple paired-Ig-like receptor A genes in the genome, and cDNA sequencing analysis suggested 0.2-4.7% frequency of overall nucleotide variations [28]. The group with the largest number of MGI genes in MGI-TC associations included six eosinophil-associated ribonuclease (*Ear1*, *Ear2*, *Ear3*, *Ear8*, *Ear9* and *Ear10*) and RNA guanylyltransferase and 5′-phosphatase (*Rngtt*) and four TCs (TC561767, TC569331, TC557280 and TC557281). The mouse Ear family has at least 13 members, 11 functional genes and 2 pseudogenes [29]. The genes within this family share a common genomic structure that is conserved with primate Ear genes. The mouse Ear gene family forms four unique clades (*Ear1/2/3/8/9/10* genes form subfamily A). The members of each clade share a high degree of sequence identity. Transcripts from *Ear1/2/3/8/9/10* were over-clustered into one TC (TC561767). TC569331 was associated with *Ear2* because of shared EST sequence AA510162 and associated with *Rngtt* because of shared GenBank sequence AK002922. Sequence analysis indicated that AA510162 encodes *Rngtt* instead of *Ear2*. Further analysis suggested that a typographical error in the publication [30] caused the reported EST sequence associated with *Ear2* to be AA510162 instead of AA510161. This many-to-many association can be resolved into one many-to-one association (six Ear genes to TC561767) caused by over-clustering and one one-to-many association (one MGI gene *Rngtt* to three TCs) caused by

transcript diversity. The group with the largest number of DTs in MGI-DT associations included 18 DTs and three immunoglobin heavy-chain genes (*Igh-4*, *Igh-VJ558* and *Igh-V*). The group with the largest number of TCs in MGI-TC associations included 25 TCs and three immunoglobin heavy-chain genes (*Igh-4*, *Igh-VJ558* and *Igh-1*). The complexity of the many-to-many associations demonstrates the challenges of creating links between genes and electronic transcripts and highlights the caveats that users of these resources must keep in mind.

Comparison of DoTS and TIGR Mouse Gene Index

For the one-to-one MGI-DT and MGI-TC associations, the number of shared sequences between DT and TC pairs linked to the same MGI gene varies

Our analysis reported 16,996 MGI-DT and 13,451 MGI-TC one-to-one associations. A total of 11,126 MGI genes had both TC and DT one-to-one associations. Among these, all except eight pairs of TC and DT had one or more GenBank sequences in common (see Table 4 for details). There are 1,305 MGI genes, whose associated DT and TC pairs had exactly the same component sequences. The assemblies that were identical in the number of component sequences were generally small clusters. Of the assemblies associated with the 1,305 MGI genes described above, 496 had only two component sequences, 635 had three to five component sequences, 136 had six to ten component sequences, and 38 had more than ten but less than 33 component sequences. Most TC and DT pairs associated with the same MGI genes have one or more sequences in common. The number of shared sequences between DT and TC varies widely, ranging from one to more than 1,000. The maximum number of shared GenBank sequences is 1,157 between DT.91337061 and TC615398 both associated with MGI gene *Svs2* (seminal vesicle protein, secretion 2; MGI:1858275). These pairs generally have very different numbers of component sequences, ranging from only two to more than a few thousand sequences. TCs generally have larger numbers of sequences per cluster than do DTs. The maximum number of component sequences for DT and TC are 1,592 (DT.537719) and 9,193 (TC619155) respectively.

For the one-to-many MGI-DT and MGI-TC associations, DoTS and TIGR Mouse Gene Index did not consistently cluster the GenBank sequences

There are 2,522 MGI genes associated with multiple DTs, and 1,975 MGI genes with multiple TCs. And 1,475 MGI genes had both MGI-to-TC and MGI-to-DT one-to-many associations. We considered all TCs or DTs associated with the same MGI genes as different forms of transcripts and grouped them together. We compared the identity and grouping of the component sequences between the TC group and its corresponding DT group. We included only the sequences curated in the MGI database in the comparison because they are mostly high-quality mRNA sequences and should be reliably clustered. There were only 245 pairs of the TC group and DT group associated with the same set of MGI curated GenBank sequences, which were also clustered in the same way. The remaining pairs differ either in their set of associated GenBank sequences or in the way of sequence clustering.

Table 4

Comparison of the constituent sequences of TCs and DTs (as of 11 October 2002)

Category	Number
DT and TC pairs analyzed*	11,126
DT and TC that have the same constituent sequences†	1,305
DT is a subset of TC†	1,416
TC is a subset of DT†	736
DT and TC assemblies that share one sequence	148
DT and TC assemblies that share 2-4 sequences	466
DT and TC assemblies that share 5-9 sequences	709
DT and TC assemblies that share 10-99 sequences	4,890
DT and TC assemblies that share 100 or more sequences	1,448
DT and TC assemblies that share zero sequence	8

*Only those with one-to-one relationship to the same MGI genes were compared. †These were not included in the count of DT and TC with shared sequences.

The differences between the two electronic transcript databases are likely to be due to the different criteria used by the two groups for clustering and assembly of EST and mRNA sequences. One possibility is different degrees of trimming poor-quality sequences from the ends of ESTs (C.J.S., personal communication). Less trimming in DoTS build might result in more assemblies than TIGR TCs. In testing, fewer larger assemblies were generated when trimming was not limited. Limited trimming was chosen in attempt to preserve better representation of differentially processed transcripts in DoTS build. The comparison of the two databases using curated data from MGI as a reference provides some measures to evaluate and improve computational methods.

Utility of the analysis

The association of MGI genes with electronic transcript assemblies supplies biological context to the computationally assembled transcripts and allows researchers to access these data from biological as well as sequence perspectives. The curation process described here permits us to rapidly build high-confidence associations between MGI genes and electronic transcript sequences. The results reveal the complications that can arise from the clustering process as well as errors in the MGI database. The assessment of the results

will provide measures to evaluate and improve the EST-assembly protocols and to check the quality of gene representation in the MGI database.

Access to the links between the MGI database and the TIGR and CBIL electronic transcript databases

Only associations between MGI genes and TCs/DTs that are supported by non-conflicting evidence (one-to-one and one-to-many associations) are accessible from the web browsers for these resources. The links from MGI genes to TCs and DTs are available from the MGI gene detail pages. The links from TCs to MGI genes are available from TIGR's TC report page and through another TIGR database resource, RESOURCERER [31]. The links from DTs to MGI genes are available from Allgenes's DT report page [32]. Users can query for related DTs by MGI gene accession identifiers or symbols. The data files for MGI-DT/TC associations are available from MGI public FTP site [17]. These data will be updated after each build of TIGR's Mouse Gene Index and CBIL's DoTS database or after every major change in MGI databases.

Additional data files

The original datasets from TIGR (from TIGR Mouse Gene Index Release 9.0 (1 October 2002)), DoTS (from DoTS mouse assembly Release 5.0 (19 August 2002)) and MGI (11 October 2002) are available as additional data files. Links from MGI to the DOTS (one-to-one, one-to-many, many-to-one and many-to-many) and TIGR (one-to-one, one-to-many, many-to-one and many-to-many) electronic transcript associations from the analysis done on 11 October 2002 are also available as additional data files. The most recent data files for MGI-DT/TC associations can be obtained from MGI public FTP site [17].

Acknowledgements

This work was supported by the Department of Energy (FG02-99ER62850) and NIH/NHGRI (HG0030-P1). The comments of Martin Ringwald, Molly Bogue, and Dong (Donnie) Qi helped to improve the clarity of the manuscript. The authors thank Jim Kadin and David Miers (MGI Software Group) and Geo Pertea (TIGR) for their technical assistance.

References

1. Boguski MS, Schuler GD: **ESTablishing a human transcript map.** *Nat Genet* 1995, **10:**369-371.
2. Quackenbush J, Liang F, Holt I, Pertea G, Upton J: **The TIGR gene indices: reconstruction and representation of expressed gene sequences.** *Nucleic Acids Res* 2000, **28:**141-145.
3. Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J: **The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species.** *Nucleic Acids Res* 2001, **29:**159-164.
4. **DoTS: a database of transcribed sequences for human and mouse genes** [http://www.cbil.upenn.edu/downloads/DoTS]
5. Christoffels A, van Gelder A, Greyling G, Miller R, Hide T, Hide W: **STACK: sequence tag alignment and consensus knowledge-base.** *Nucleic Acids Res* 2001, **29:**234-238.
6. **The Mammalian Gene Collection** [http://mgc.nci.nih.gov]
7. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara, A, Fukunishi Y, Konno H, *et al.*: **Functional annotation of a full-length mouse cDNA collection.** *Nature* 2001, **409:**685-690.
8. **Mouse Genome Informatics** [http://www.informatics.jax.org]
9. Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT: **The Mouse Genome Database (MGD): the model organism database for the laboratory mouse.** *Nucleic Acids Res* 2002, **30:**113-115.
10. Ringwald M, Eppig JT, Begley DA, Corradi JP, McCright IJ, Hayamizu TF, Hill DP, Kadin JA, Richardson JE: **The Mouse Gene Expression Database (GXD).** *Nucleic Acids Res* 2001, **29:**98-101.
11. Bult CJ, Richardson JE, Blake JA, Kadin JA, Ringwald M, Eppig JT, the Mouse Genome Informatics Staff: **MGI: Mouse Genome Informatics in a New Age of Biological Inquiry.** In *Proceedings of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering.* Los Alamitos, CA: IEEE Computer Society; 2000:29-32.
12. Naf D, Krupke DM, Sundberg JP, Eppig JT, Bult CJ: **The Mouse Tumor Biology Database: a public resource for cancer genetics and pathology of the mouse.** *Cancer Res* 2002, **62:**1235-1240.
13. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
14. **SWISS-PROT** [http://www.ebi.ac.uk/swissprot]
15. Maglott DR, Katz KS, Sicotte H, Pruitt KD: **NCBI's LocusLink and RefSeq.** *Nucleic Acids Res* 2000, **28:**126-128.
16. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140.
17. **MGI Data and Statistical reports** [ftp://ftp.informatics.jax.org/pub/reports/index.html]
18. **DoTS download** [http://www.cbil.upenn.edu/downloads/DoTS/release_5]
19. Barbas JA, Chaix JC, Steinmetz M, Goridis C: **Differential splicing and alternative polyadenylation generates distinct NCAM transcripts and proteins in the mouse.** *EMBO J* 1988, **7:**625-632.
20. Santoni MJ, Barthels D, Vopper G, Boned A, Goridis C, Wille W: **Differential exon usage involving an unusual splicing mechanism generates at least eight types of NCAM cDNA in mouse brain.** *EMBO J* 1989, **8:**385-392.
21. **UCSC Genome Browser-Mouse Genome Assembly February, 2002** [http://genome.ucsc.edu/]
22. Holzfeind PJ, Ambrose HJ, Newey SE, Nawrotzki RA, Blake DJ, Davies KE: **Tissue-selective expression of alpha-dystrobrevin is determined by multiple promoters.** *J Biol Chem* 1999, **274:**6250-6258.
23. Ferguson SE, Rudikoff S, Osborne BA: **Interaction and sequence diversity among T15 VH genes in CBA/J mice.** *J Exp Med* 1988, **168:**1339-1349.
24. Zimmerer EJ, Passmore HC: **Structural and genetic properties of the Eb recombinational hotspot in the mouse.** *Immunogenetics* 1991, **33:**132-140.
25. Huttner KM, Selsted ME, Ouellette AJ: **Structure and diversity of the murine cryptdin gene family.** *Genomics* 1994, **19:**448-453.
26. Smith DP, Spicer J, Smith A, Swift S, Ashworth A: **The mouse Peutz-Jeghers syndrome gene Lkb1 encodes a nuclear protein kinase.** *Hum Mol Genet* 1999, **8:**1479-1485.
27. Gray TA, Saitoh S, Nicholls RD: **An imprinted, mammalian bicistronic transcript encodes two independent proteins.** *Proc Natl Acad Sci USA* 1999, **96:**5616-5621.
28. Kubagawa H, Burrows PD, Cooper MD: **A novel pair of immunoglobulin-like receptors expressed by B cells and myeloid cells.** *Proc Natl Acad Sci USA* 1997, **94:**5261-5266.
29. Cormier SA, Larson KA, Yuan S, Mitchell TL, Lindenberger K, Carrigan P, Lee NA, Lee JJ: **Mouse eosinophil-associated ribonucleases: a unique subfamily expressed during hematopoiesis.** *Mamm Genome* 2001, **12:**352-361.
30. McDevitt AL, Deming MS, Rosenberg HF, Dyer KD: **Gene structure and enzymatic activity of mouse eosinophil-associated ribonuclease 2.** *Gene* 2001, **267:**23-30.
31. **RESOURCERER 5.0** [http://pga.tigr.org/tigr-scripts/magic/r1.pl]
32. **AllGenes** [http://www.allgenes.org]