T1DBase: gene models

James E. Allen

27th April 2011

I recently wrote an overview of T1DBase, an online resource for the type 1 diabetes (T1D) research community (Hulbert et al. 2007; Burren et al. 2011). I shall now describe one of the more interesting of my contributions to the project, calculating and graphically displaying gene models. In the month since I first started writing this document, the 'Gene Models' section of T1DBase has changed, to become a 'Gene Overview', which looks slightly different (that'll teach me to take so long in writing this up). In particular, summary and consensus gene models are no longer displayed, and nor are gene models for older builds; these are still viewable via the T1DBase archive site. Here, I'll describe the work that I did, rather than what appears on the current site.

Genes and Gene Models

The first, and probably most difficult, step is to decide what a gene is in the first place. I started this work about 5 years ago, before the importance of non-protein-coding RNA was widely recognised, so the discussion here relates only to protein-coding genes, rather than RNA genes. I'm currently working on topics related to the evolution of RNA genes, so I think it'd be sensible to add them, which shouldn't actually be too difficult. But anyway, how do we define a protein-coding gene? One answer might be "a section of DNA that is responsible for the creation of a functional protein". But what if there are splice variants; which one do you choose, or do you merge them? And where do you get your information from? If from multiple sources, how do you deal with any conflicts in the delimitation of gene boundaries, or intron or UTR structure? And what if one source defines the same gene in multiple locations, perhaps on different chromosomes? You can deal with many of these issues by working with "gene models" rather than genes, where a gene model is a collection of structures for a single gene, from a single source. Figure 1 shows two gene models for the gene CTLA4, based on data from Ensembl and UCSC.

It is useful to have a common point of reference for gene models from different sources, and in T1DBase this is the RefSeq data from the NCBI. This ties in with what T1DBase considers to be a gene: in practical terms, something with an Entrez Gene ID. In previous versions of T1DBase, the NCBI gene is represented alongside each gene model in T1DBase, as a green box (Figure 1), but this is not shown in the current version of the site.

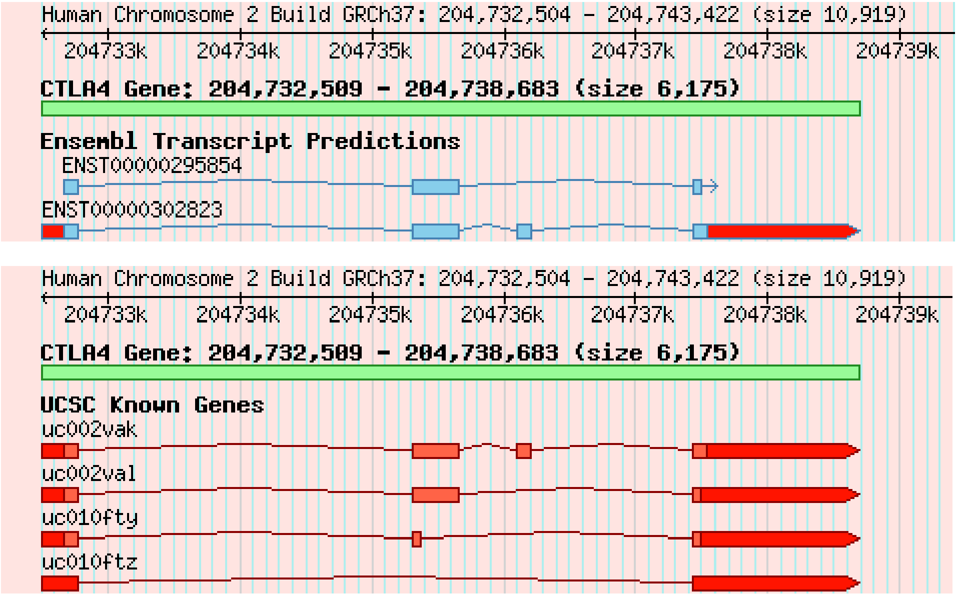
Human Chromosome 2 Build GRCh37: 204,732,504 - 204,743,422 (size 10,919)
204733k
204734k
204735k
204736k
204737k
204738k
204739k
CTLA4 Gene: 204,732,509 - 204,738,683 (size 6,175)
Ensembl Transcript Predictions
ENST00000295854
ENST00000302823
Human Chromosome 2 Build GRCh37: 204,732,504 - 204,743,422 (size 10,919)
204733k
204734k
204735k
204736k
204737k
204738k
204739k
CTLA4 Gene: 204,732,509 - 204,738,683 (size 6,175)
UCSC Known Genes
uc002vak
uc002val
uc010fty
uc010ftz

Figure 1.
Gene models for CTLA4, from Ensembl and UCSC. Exons and introns are shown by boxes and connecting lines, respectively, and UTRs are highlighted in red. The green box displays the gene according to NCBI RefSeq data.

Gene Models in T1DBase

When I started on the T1Dbase project, the site already displayed gene models based on data from a number of sources. My task was two-fold: to make the graphical display clearer and prettier, and to find a way to usefully summarise gene models, both within and between different sources. At this point it would be remiss not to acknowledge the importance of discussions with, and advice from, my colleagues on the project, i.e. my fellow authors on the Hulbert et al. (2007) paper. Also, I used libraries from the BioPerl project, chiefly Bio::DB::GFF, Bio::SeqFeature and Bio::Graphics::Panel.

Gene models in T1DBase are shown on each gene page (e.g. the CTLA4 gene page). I won't talk much about the graphical aspect of the work I did - I link to the code later, but it's rather tied into the T1DBase code base to be of general use (although by all means contact me if you would like some more information or assistance).

It's probably worth pointing out that T1DBase is not restricted to genes that are linked to type 1 diabetes, so if you are just after a nice display of gene models for a gene of interest, you can still use the T1DBase website. All of the NCBI genes are available, since you never know when a gene will be linked to diabetes (genes which *have* been linked will tend to have additional gene models). So if you want to look at gene models for any gene, based on data from 4 useful sources (CCDS, Ensembl, UCSC, and Vega), go to the T1DBase home page and type the gene name or ID into the search box in the top-right corner.

The most interesting aspect of the gene models in T1DBase relates to how multiple sources of data can eliminate spurious transcripts, and how data can be effectively summarised across all gene models.

Eliminating Incorrect Transcript Predictions

There are a few preparatory steps required before you can start weighing up transcripts from different sources, all of which are automated by a set of scripts that I wrote (and which continue to be maintained by the current T1DBase staff). These download the raw data for a given range of sources, species, and builds, convert them to GFF format, and load them into a Bio::DB::GFF database. One issue that arises with Ensembl and UCSC transcripts is that these sources assign Entrez Gene IDs to transcripts based on sequence similarity, irrespective of whether the gene and the transcript are at (roughly) the same location. This results in assignments which are wrong, so in order to sort out which Entrez Gene IDs go with which Ensembl/UCSC IDs, we cross-reference with RefSeq. A transcript is disregarded if it is on a different chromosome[1] to the RefSeq-defined gene; if it is on the same chromosome but a different strand; or if it is on the same chromosome and strand, but not within 100kb of the gene. This value of 100kb is chosen to be big enough to allow some variation, but to ignore infeasible large discrepancies; it is rather arbitrary, but works well in practice. (See the `resolve_id_ambiguities.pl` script for details.)

Having dealt with conflicts between the positions of transcripts within a data source, we turn our attention to comparisons across all of the data sources. We want to ascertain when a gene model from one or more sources is on a different chromosome or strand, or more than 1Mb distant (again, rather an arbitrary value for the distance, but sufficiently high that we are very unlikely to dismiss an accurate transcript). To start with, we need to decide on which sources we trust more than others (the discussion here relates to 4 sources, but the arguments are applicable to any sources). Ensembl and UCSC are comprehensive, but are largely automated; CCDS provides reliable positions for CDS regions but, by definition, no UTRs; and Vega is high-

[1] Chromosomes X and Y are considered to be the same chromosome.

quality, hand-curated data. Using this information, for the transcripts with conflicting positions we apply the following logic to all of the transcripts for that gene:

- If there is a single CCDS or Vega transcript, consider that to be the most likely position (termed the 'tentative' position), *unless* there are both CCDS and Vega transcripts, and they have conflicting positions.
- For all sources, across all transcripts, count how many support each conflicting position.
- If there are three or more conflicting positions, add the counts for all but the most supported position (MSP) together, so that we have two counts for comparison.
- If the MSP is only supported by one transcript, mark all of the transcripts as 'undecided', as there is insufficient evidence to automatically resolve the conflict.
- If the MSP has multiple transcripts, then compare it to the number of transcripts supporting other positions; if the ratio of the numbers is above a certain value, mark the MSP as 'accepted' and the unsupported transcripts as 'rejected'; otherwise, mark all transcripts as 'undecided'.
- The ratio in the previous step depends on whether a tentative position from CCDS or Vega exists. If not, the ratio is 2:1, i.e. the number of transcripts for the MSP must be at least twice the number of all other transcripts. If the tentative position exists, and agrees with the MSP, the ratio is 1:1 (a simple majority); in the case of disagreement, the ratio is 3:1 (a lot of support is required to outweigh a hand-curated position).

Once this process is complete, there are then three groups of transcripts, those that can be confidently used or ignored (i.e. the ones marked as 'accepted' or 'rejected', respectively), and those that require caution (the 'undecided' set). This information is loaded into the Bio::DB::GFF database, to enable useful queries and further data modification. The `generate_gene_models.pl` script automates the entire process of building gene models, from downloading the data from different sources, to evaluating when they do and do not agree.

Summary and Consensus Gene Models

As mentioned above, summary and consensus gene models are not shown on the current version of T1DBase, but can be viewed on the archive version of T1DBase, e.g.

CTLA4. So, having collated data on transcripts from multiple sources that are (approximately) in agreement, it is then useful to examine how much variation in the detail of those transcripts, for example, which exons are best supported. In 'summary gene models', the number of transcripts that support each base are plotted as a bar chart (Figure 2, top panel). Transcripts from different sources that have almost identical CDS regions (i.e. within 1 base in either direction) are shown as 'consensus gene models' (Figure 2, bottom panel); the bounds of the UTRs are allow to vary somewhat, proportional to the length of the gene. Hand-curated sources are considered sufficiently reliable to warrant a consensus gene model of their own.

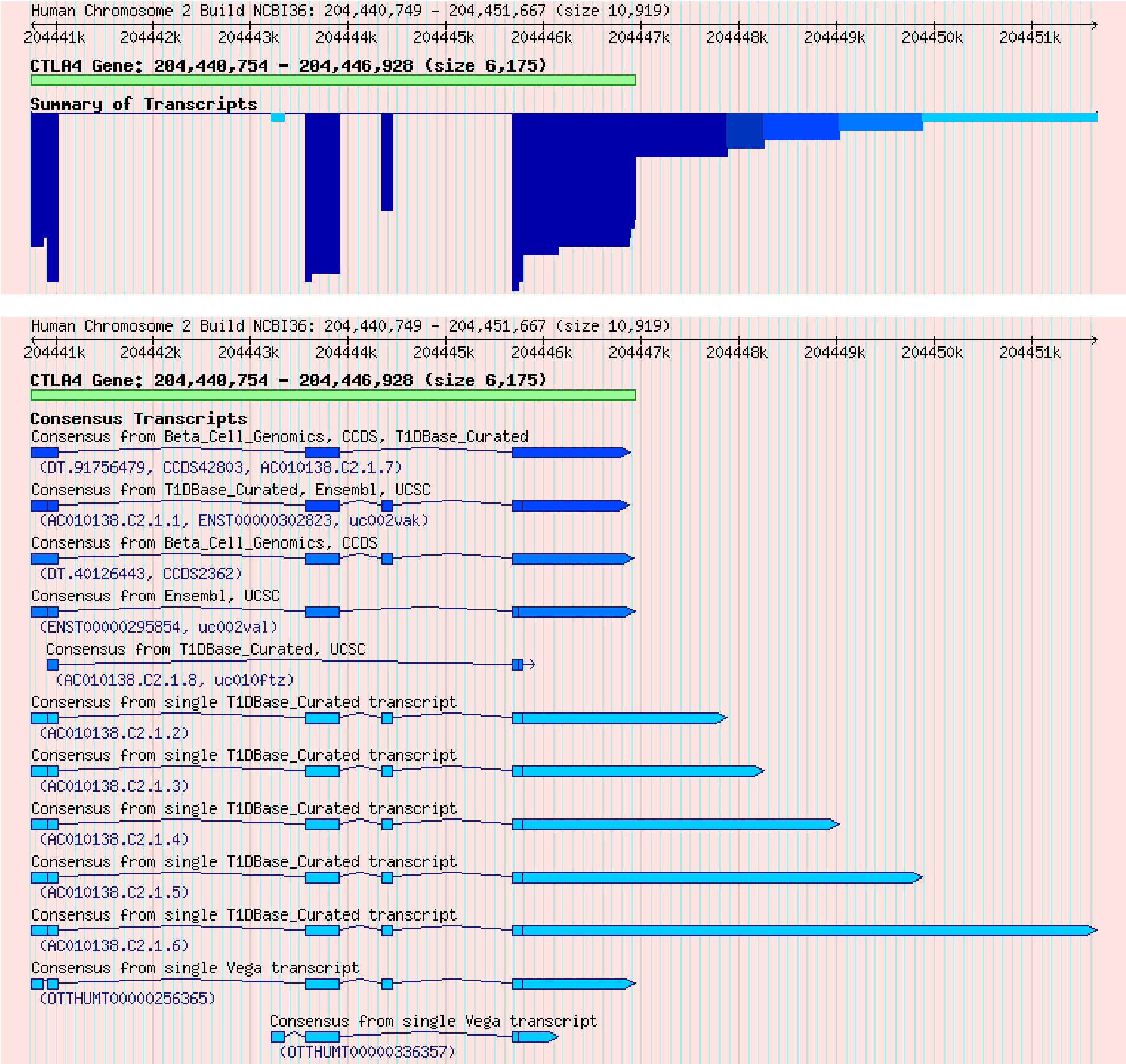

Figure 2.
Summary and consensus gene models for CTLA4. Exons and introns are shown by boxes and connecting lines, respectively. Darker shades of blue indicate greater support than lighter shades. The green box displays the gene span, which shows the bounds of the gene according to NCBI RefSeq data.

The summary and consensus models are calculated with Perl modules (see next section) that require some other T1DBase modules and configuration files. But the code is well commented and fairly generic, so I think only minor tweaks would be necessary to use it elsewhere; please let me know if you would like help in doing so.

Gene Models Code

All of the source code behind T1DBase is available under the GNU GPL (see the website for details), from a sourceforge subversion repository. The gene models Perl scripts are still under active development, but the associated GeneModel Perl modules that I wrote are no longer being used. I have included a copy of the GeneModel modules on my website, plus versions of the scripts that I have modified to work with a Windows installation of MySQL and a slightly different set of data sources than the current T1DBase site. As mentioned above, the GeneModel modules are integrated into the site, and won't work out of the box; please let me know if you want to use any of the functionality, and I would be delighted to help out.

Citing this Document

[If referring to this document, please cite its location on the Monkeyshines website: http://www.monkeyshines.co.uk/blog/archives/362]

Contact

James Allen: james@monkeyshines.co.uk

References

Burren OS, Adlem EC, Achuthan P, Christensen M, Coulson RMR, Todd JA. 2011. T1DBase: update 2011, organization and presentation of large-scale data sets for type 1 diabetes research. *Nucleic Acids Research* **39**(Database issue): D997-D1001.

Hulbert EM, Smink LJ, Adlem EC, Allen JE, Burdick DB, Burren OS, Cassen VM, Cavnor CC, Dolman GE, Flamez D et al. 2007. T1DBase: integration and presentation of complex data for type 1 diabetes research. *Nucleic Acids Research* **35**(Database issue): D742-746.