Published online 14 November 2007 Nucleic Acids Research, 2008, Vol. 36, Database issue D753–D760
doi:10.1093/nar/gkm987

The vertebrate genome annotation (Vega) database

L. G. Wilming*, J. G. R. Gilbert, K. Howe, S. Trevanion, T. Hubbard and J. L. Harrow

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK

Received September 15, 2007; Revised October 18, 2007; Accepted October 19, 2007

ABSTRACT

The Vertebrate Genome Annotation (Vega) database (http://vega.sanger.ac.uk) was first made public in 2004 and has been designed to view manual annotation of human, mouse and zebrafish genomic sequences produced at the Wellcome Trust Sanger Institute. Since its initial release, the number of human annotated loci has more than doubled to close to 33 000 and now contains comprehensive annotation on 20 of the 24 human chromosomes, four whole mouse chromosomes and around 40% of the zebrafish Danio rerio genome. In addition, we offer manual annotation of a number of haplotype regions in mouse and human and regions of comparative interest in pig and dog that are unique to Vega.

INTRODUCTION

Currently only three vertebrate genomes, human, mouse and zebrafish, are being fully sequenced and finished to a quality which merits manual annotation. Although labour intensive and relatively slow compared with automatic annotation methods, manual annotation provides an invaluable reliable reference resource that can be used to predict gene structures on low coverage genomes from other vertebrate species. The Vega database is the central repository for the majority of genome sequencing centres to deposit their annotation of human chromosomes. Unlike other browsers, Vega only displays a manually annotated gene set on the latest chromosome assemblies, which are often more up-to-date than the reference genome assembly generated by NCBI. Currently, the human database contains twenty chromosomes annotated by eight different sequencing centres. The Havana Group at the Wellcome Trust Sanger Institute (WTSI) is updating the annotation through its involvement in the consensus-coding sequence (CCDS) collaboration with UCSC, NCBI and Ensembl (http://www.ncbi.nlm.nih.gov/CCDS/) which aims to produce a reference set of protein-coding gene annotation across the entire human genome.

The four mouse chromosomes (2, 4, 11 and X) sequenced at WTSI have been virtually fully annotated and can be browsed through Vega. The rest of the mouse genome is being annotated on a gene-by-gene basis as part of the mouse CCDS collaboration.

The Zebrafish genome, which is being fully sequenced and manually annotated at the WTSI in collaboration with Zfin (1), currently features eight completely annotated chromosomes.

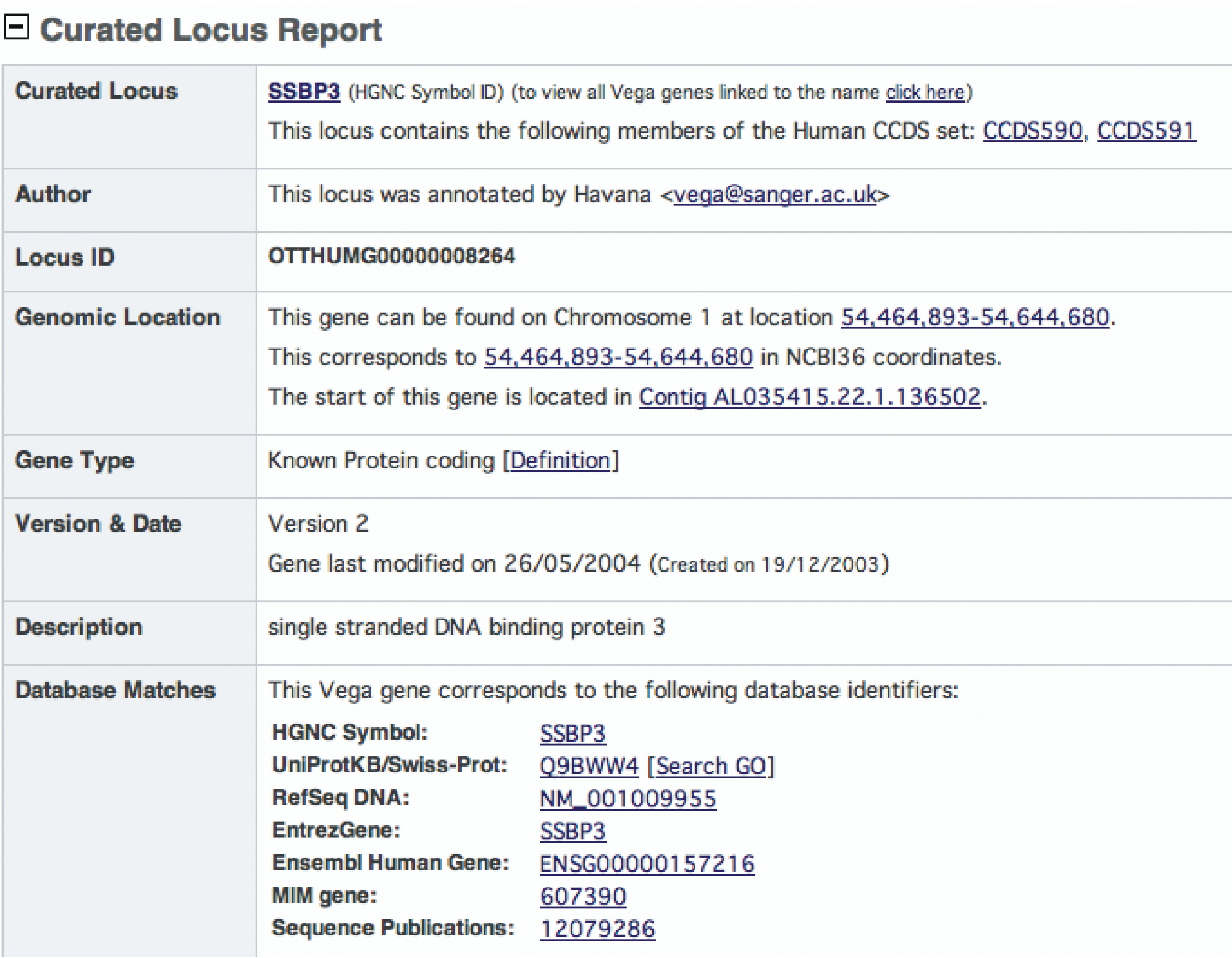
In addition to full genomes, and unlike other browsers, Vega also displays small finished regions of interest from genomes of other vertebrates, human haplotypes and mouse strains. Currently this comprises the finished sequence and annotation of the major histocompatability complex (MHC) from different human haplotypes, and dog and pig [the latter of which is currently otherwise only available in very limited form in Ensembl Pre! (http://pre.ensembl.org/Sus_scrofa/index.html)]. Additionally there is mouse NOD (non-obese diabetes) strain annotation of IDD (insulin-dependent diabetes) candidate regions and two more pig regions.

Improvements and progress in Vega since 2004

All three complete genomes (mouse, human and zebrafish) now contain a view of all the chromosomes in the Karyotype View and the annotation progress of each chromosome is highlighted with grey shading. Since the original Vega publication in 2005 (2), the number of human gene loci annotated has more than doubled to almost 33 000 (June 2007 release), close to 19 000 of which are predicted to be protein coding. Four chromosomes (2, 4, 5 and 11) remain to be fully manually annotated to the Havana standard and these will be completed as part of the CCDS collaboration and the whole-genome extension of the ENCODE project (see below). Since annotation is continually re-evaluated on a gene-by-gene basis, every locus is versioned and the date of creation and last update can now be viewed by the user on the curated locus report page (GeneView, see Figure 1).

The CCDS project aims to produce a set of protein-coding transcripts that is agreed upon by the RefSeq group at the NCBI, the Havana and Ensembl groups at the Wellcome Trust Genome Campus and the Genome Informatics group at the UCSC. Though originally limited to human genes, the project now includes mouse. As part of the collaboration, we are comprehensively annotating (i.e. including all coding and non-coding variants) each

*To whom correspondence should be addressed. Tel: +44 1223 496843; Fax: +44 1223 496802; Email: lw2@sanger.ac.uk



⊟ Curated Locus Report

Curated Locus	SSBP3 (HGNC Symbol ID) (to view all Vega genes linked to the name click here) This locus contains the following members of the Human CCDS set: CCDS590, CCDS591
Author	This locus was annotated by Havana <vega@sanger.ac.uk>
Locus ID	**OTTHUMG00000008264**
Genomic Location	This gene can be found on Chromosome 1 at location 54,464,893-54,644,680. This corresponds to 54,464,893-54,644,680 in NCBI36 coordinates. The start of this gene is located in Contig AL035415.22.1.136502.
Gene Type	Known Protein coding [Definition]
Version & Date	Version 2 Gene last modified on 26/05/2004 (Created on 19/12/2003)
Description	single stranded DNA binding protein 3
Database Matches	This Vega gene corresponds to the following database identifiers: **HGNC Symbol:** SSBP3 **UniProtKB/Swiss-Prot:** Q9BWW4 [Search GO] **RefSeq DNA:** NM_001009955 **EntrezGene:** SSBP3 **Ensembl Human Gene:** ENSG00000157216 **MIM gene:** 607390 **Sequence Publications:** 12079286

Figure 1. Part of the GeneView Locus Report showing versioning information and CCDS and nomenclature (in this case HGNC) information and links. Edited from http://vega.sanger.ac.uk/Homo_sapiens/geneview?gene = OTTHUMG00000008264&db = core.

human and mouse CCDS locus to provide a solid basis for comparison with RefSeq. In the process, this supplies up-to-date annotation to previously annotated sequences and novel annotation to unannotated sequence. Where appropriate, Vega transcript and gene records (TranscriptView, GeneView) have links to the CCDS gene records in the CCDS database at the NCBI (http://www.ncbi.nlm.nih.gov/CCDS/) (Figure 1).

As part of the ENCODE project (3,4), Havana have comprehensively annotated the target genes (1% of human genes) in human and mouse and updated the annotation following both experimental and computational feedback from the GENCODE project (5–7). In human Vega, ENCODE regions are marked in ContigView (users may have to switch on the relevant track in the 'Decorations' menu).

Vega transcript objects are also shown, in a separate track, in Ensembl Detailed View (tracks named 'Vega Havana Gene' and 'Vega External Gene'; the user may have to switch these tracks on in the 'Features' menu). In order to eliminate redundancy in the Ensembl transcript track and highlight commonality, Ensembl and Vega have started to match protein-coding transcripts between the two datasets and only present a single transcript if within a given locus a Vega and an Ensembl transcript are identical. These transcripts (and loci containing them) are coloured gold and labelled 'Merged Known Protein Coding' or 'Common Known Protein Coding' in Ensembl ContigVew (Figure 2). The project is currently limited to human genes annotated by Havana, but is expected to include Havana-annotated mouse genes in Ensembl version 48 (December 2007 release).

In preparation for the zebrafish genome paper (which will be based on genome assembly Zv8), all mRNA entries in the Zfin database (http://zfin.org/) have been aligned to the current Zv7 assembly and those that map have been annotated (currently 6157). On an ongoing basis, known mRNAs are being mapped and annotated as new finished genomic sequence becomes available. To remove artificial duplications, annotation from the previous mixed-strain library genomic clones has been moved to a reference assembly constructed from a single double-haplotype

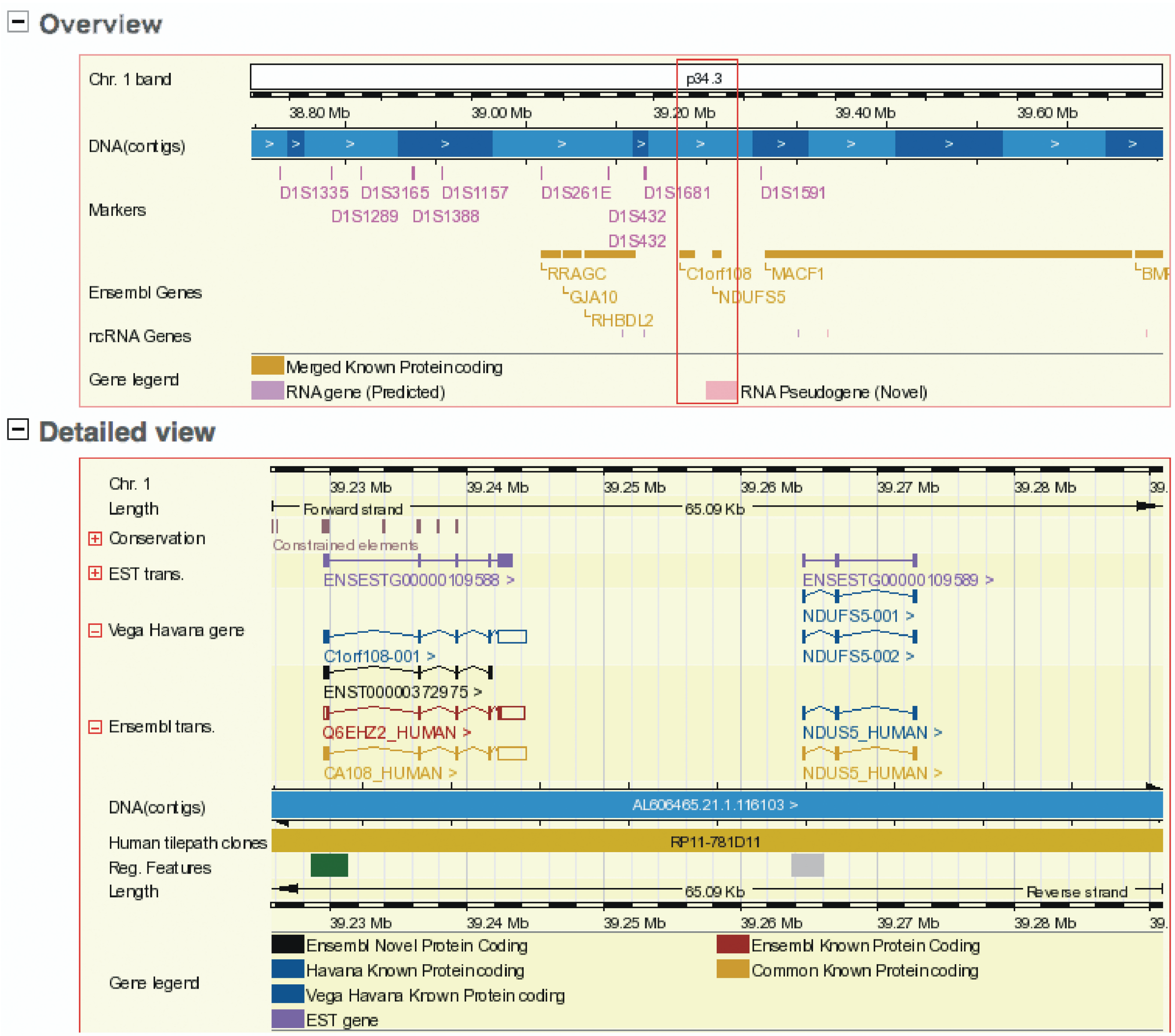


Figure 2. Ensembl release 46 page showing the gold-coloured Ensembl-Vega gene merge loci (top: Overview) and transcripts (bottom: Detailed view). Edited from http://www.ensembl.org/Homo_sapiens/contigview?c = 1:39255078;w = 63203.

Tübingen strain individual. The original clones are still visible in ContigView and annotation can be compared between the two in MultiContigView (Figure 3) using the 'Comparative' menu in Detailed View.

We collaborate closely with the MGI group at The Jackson Laboratory (http://www.informatics.jax.org/) regarding mouse gene sets and their nomenclature. Genes are cross-linked between Vega and MGI: Vega GeneView pages link to MGI locus records and vice versa. A similar collaboration is in place with the HGNC (http://www.genenames.org/) for human genes (Figure 1) and Zfin (http://zfin.org/) for zebrafish.

For the first time a large region of the porcine genome, 8.2 Mb of chromosome 17 sequence orthologous to human chromosome 20q13 and mouse 2, has been made available (8). The region has been used to assess the sequencing methodology for the pig genome (8). As both the pig sequence and the orthologous human and mouse sequences have been annotated by Havana, users can compare the sequences in Vega's MultiContigView. In addition to the chromosome 17 sequence, Vega presents the pig MHC region, located on chromosome 7 (9) (see below), and the region of pig chromosome 6 containing the LRC (leukocyte receptor complex) genes (10,11). Their orthologous regions in human have been annotated by Havana, so again, they can be viewed in Vega alongside human sequence, and, in the case of the MHC, dog as well.

Below, a selection of new projects, where the data have been first released in Vega, are described in more detail.

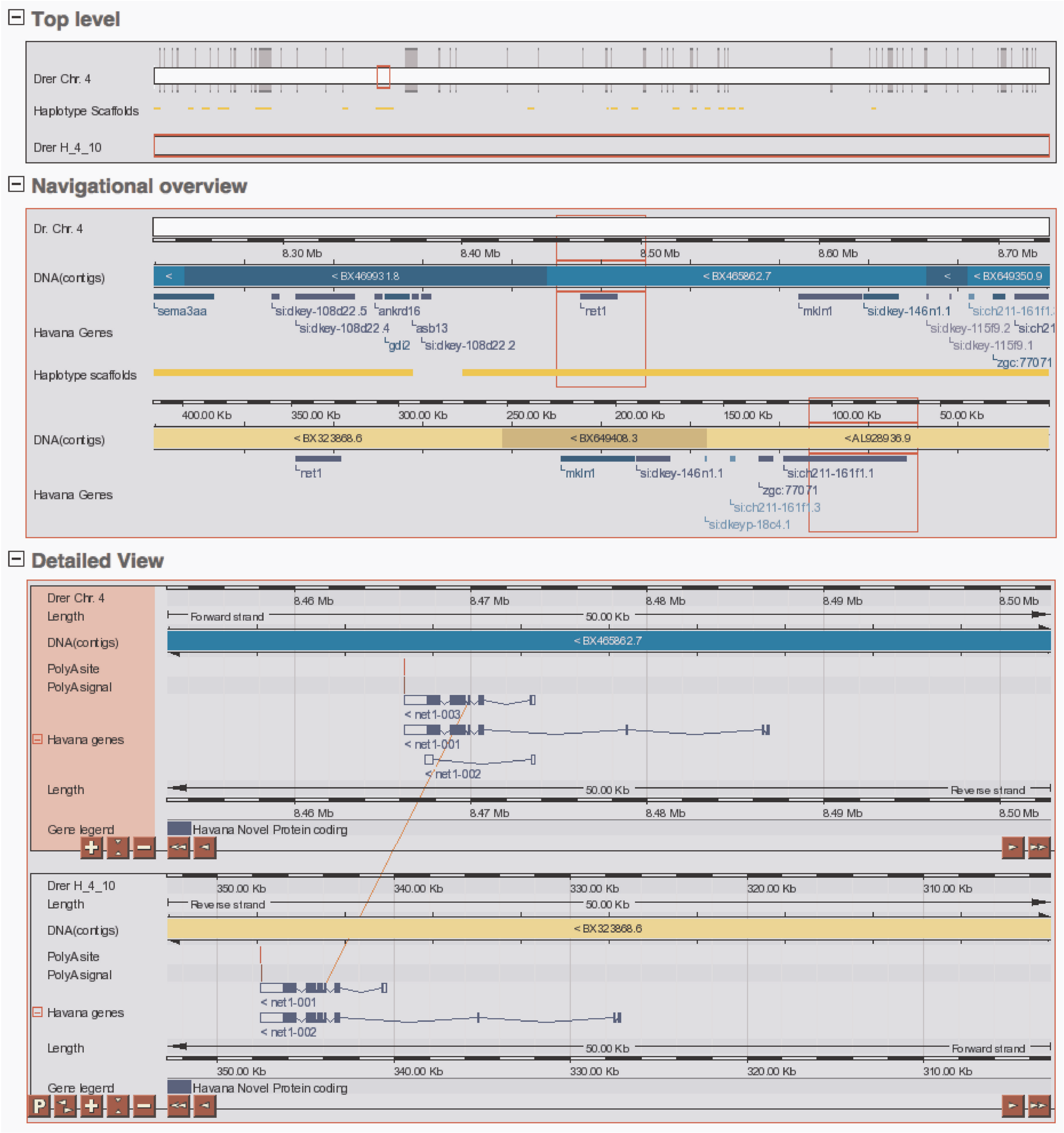


Figure 3. Zebrafish haplotype clones are marked in yellow (Top level and Navigational overview panels above). In MultiContigView, annotation can be shown on both reference and haplotype simultaneously with lines linking homologous genes (Detailed View panel above). This view is accessible by choosing the desired second dataset from the 'View alongside' menu from the left-hand menu/navigation bar in ContigView (not shown). Edited from http://vega.sanger.ac.uk/Danio_rerio/multicontigview?sr1 = H_4_11;s1 = Danio_rerio;c = 4:8468245;w = 44415.

Mouse genes targeted for knockout. The WTSI is producing annotation for both the EUCOMM (European Conditional Mouse Mutagenesis) (http://www.eucomm.org/) and KOMP (Knock-Out Mouse Project) (http://www.nih.gov/science/models/mouse/knockout/) efforts. These two projects aim to generate a comprehensive resource of (conditional) knockout (KO) alleles in mouse embryonic stem cells. The target genes can be viewed in Vega as a KO track. Transcript models shown in this track are the transcripts produced in KO mice where target exon(s) (also shown) have been deleted; the resulting coding transcripts are subject to nonsense-mediated decay.

Mouse diabetes (IDD) candidate regions. Mouse strain NOD (non-obese diabetic) is a model for identifying genes involved in IDD (insulin-dependent diabetes) (12–14). We are annotating candidate regions, in parallel, in the reference BL/6 strain and the NOD strain in order to compare the two strains and detect differences that may be relevant to type I diabetes susceptibility. Reference and NOD strain annotation can be viewed alongside each other in Vega MultiContigView.

MHC haplotype and comparative MHC. The primary aim of the human MHC Haplotype Project (15–17) is to provide a comprehensively annotated reference sequence of a single, HLA-homozygous MHC haplotype and to use it as a basis against which we could assess variations from seven other similarly homozygous cell lines, representative of the most common MHC haplotypes in the European population. Through the Vega database users can access gene annotation of the eight MHC haplotype sequences as it becomes available, providing a valuable public resource and a means of integrating annotation and variation data. As mentioned earlier, canine (Doberman breed) (18) and porcine MHC regions have been sequenced and annotated as well, allowing for a direct comparison of the region between three different organisms and between a number of human haplotypes (Figure 4).

Additional classification and improved annotation of alternatively spliced variants. Our locus classification classes were developed to aid standardization of the annotation of gene features by different groups across the human genome and were initially developed through a series of workshops (http://www.sanger.ac.uk/HGP/havana/hawk.shtml). However, as the transcript diversity appears to present a complex landscape for each locus, we have introduced an in-depth classification at the transcript level to aid interpretation of their functionality. As mentioned above, the Havana group produced the reference annotation for the ENCODE project as part of the GENCODE collaboration. As part of this project, all coding transcripts were analysed by the Biosapiens consortium which examined the structural viability of each protein by various methods (19). On feedback from the consortium we have started to classify our coding transcripts into the following four categories:

(i) **Known CDS**: identical to SwissProt entry or RefSeq NP protein.
(ii) **Novel CDS**: shares >60% of its coding length with Known CDS, has cross-species or gene family support for its structure or a Pfam domain structure identical to Known CDS.
(iii) **Putative CDS**: shares <60% of its coding length with Known CDS, has novel first or last coding exon or lacks cross-species or gene family support for its structure.
(iv) **NMD**: if the CDS (following the appropriate reference CDS) of a transcript finishes >50 bp from a downstream splice site, the transcript is tagged as being subject to nonsense-mediated decay (NMD)

Further more, transcript variants for which a CDS cannot be assigned confidently, are classified into the following main types:

(i) **Transcript**: does not qualify for any of the specific types below.
(ii) **Retained intron**: relative to an appropriate reference variant, transcript contains intronic sequences not due to alternative splice sites.
(iii) **Putative**: up to three exons, supported by only up to two ESTs (from same or other species).
(iv) **Non-coding**: for known non-coding genes only.
(v) **Antisense**: for known antisense genes only (i.e. genes that have a published regulatory/expression/functional relationship with the gene on the opposite strand, such as mouse Nespas).
(vi) **IG segment**: for known immunoglobulin gene fragments only (e.g. the IGL cluster on human chromosome 22 or the Trav cluster on mouse chromosome 14).

As far as we are aware, Vega is the only place to find large-scale annotation of putative NMD targets [though there is a database of SNP-induced NMD targets (20) (http://variome.kobic.re.kr/SNP2NMD/)]. The full description of the current locus and transcript classification classes can be found at: http://vega.sanger.ac.uk/info/about/gene_and_transcript_types.html.

Generating the database for the Vega website

As mentioned in Ashurst *et al.* (2), the data released via the Vega website is produced by merging two in-house databases at the Sanger Institute: the pipeline database containing the genome assembly and alignments of features (mRNAs, proteins and ESTs, gene predictions, etc.) to that assembly, and the Otter annotation database containing the manual annotation. The Vega website runs from an Ensembl (21–23) schema database, the version of which is, as far as possible, kept synchronized with that of the Ensembl website. This strategy of keeping closely synchronized with Ensembl has advantages such as facilitating maintenance of the website—new features developed for Ensembl can sometimes become available to Vega with little or no development time being required. However, the schema difference between the Otter annotation database (which is based on a version of the schema originating from 2003 and positions genes on clones instead of chromosomes) and the Vega website database is significant for the Vega release process: the genes have to be mapped from clones onto chromosomes, and data has to be moved from legacy tables into core Ensembl schema tables. Whilst there have been numerous improvements to this process over the four year life of Vega, this step does remain a bottleneck in the release process. For this, and for other reasons, we are currently in the process of migrating the Otter annotation database onto the current Ensembl schema (see Future Plans section). However, the frequency of release of the website

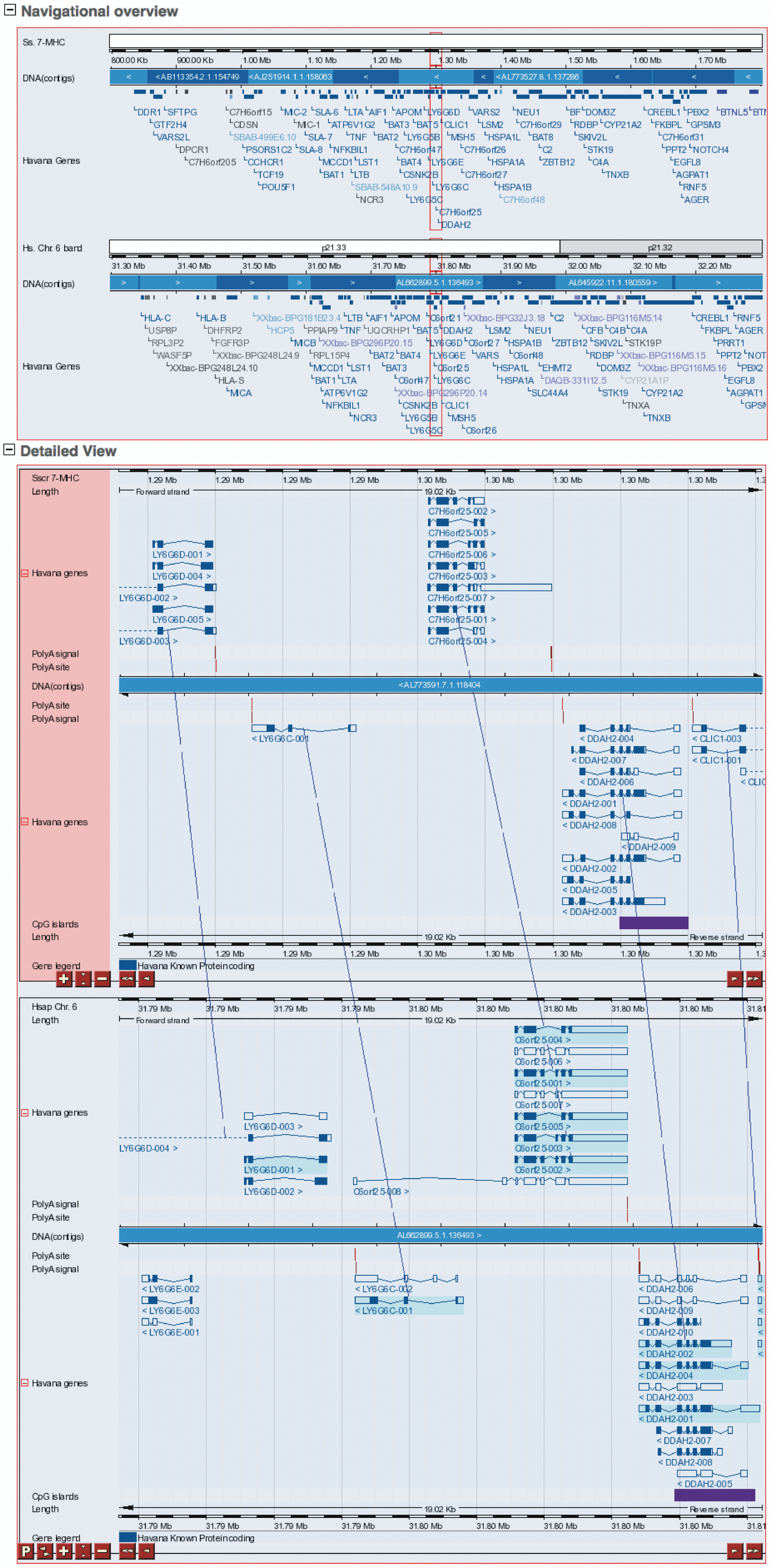


Figure 4. MultiContigView of a region of the pig and dog MHC. Lines link computationally determined orthologues in the Detailed View. This view is accessible by choosing the desired second dataset from the 'View alongside' menu from the left-hand menu/navigation bar in pig ContigView (not shown). Edited from http://vega.sanger.ac.uk/Sus_scrofa/multicontigview?sr1 = 6;s1 = Homo_sapiens;c = 7-MHC:1296785;w = 19183.

will always be limited by the requirement to generate additional data required for the functionality of each specific release of the website. These data include the Compara database that allows for the comparative analysis in Vega, the files used for sequence (BLAST and SSAHA) and text (Exalead) searching, and updates to help documentation and other information.

Accessing and querying data

Most of the Vega annotation data can be accessed via Ensembl through its BioMart system (24,25) for data queries. Furthermore, genomic, transcript and protein sequences can be easily exported in several formats from the various Views (for example 'Export cDNA' or 'Export peptide' from the menu obtained by clicking on gene cartoons in the Detailed View or Basepair View panels in ContigView). We also have Blast and SSAHA services available for alignments of user's query sequences against Vega transcripts, proteins or genomic sequence and users can download Fasta files from the Vega FTP site (ftp://ftp.sanger.ac.uk/pub/vega).

Feedback and submitting data

In order to maintain and enhance the quality and coverage of our annotation, the Havana team is always interested in feedback, collaboration and high-quality external data. Please feel free to contact us at vega@sanger.ac.uk for feedback and queries or contact the corresponding author to discuss collaborations and data submissions.

Future plans

A significant development in the near future will be the migration of the Otter annotation database to a near-current version of the Ensembl schema. This should increase the release frequency and allow us to present the most recent data to the community. It will also improve versioning, searching, dealing with exceptions (e.g. selenocysteine), and mapping features across clone boundaries. In addition, in the longer term we are aiming for much of the data that is currently generated after merging the pipeline and annotation databases, such as the location of protein domains on translations, links between Vega genes/transcripts and external databases (such as MGI), karyotype images, etc. to be incorporated into the annotation database.

We will continue adding mouse, and updating human, CCDS annotation in collaboration with the NCBI and UCSC. Other ongoing collaborations are Ensembl-Vega gene merges, refining and extending ENCODE annotation with the ENCODE and GENCODE consortia and refining nomenclature and annotation with HGNC, MGI and Zfin. Maintenance and updating of existing annotation in human, mouse and zebrafish is ongoing, as is general (non-project related) *de novo* annotation.

ACKNOWLEDGEMENTS

We thank the Havana, human Chromosome 22, Genoscope, Washington University, Baylor College of Medicine, Broad Institute, Joint Genome Institute, RIKEN and Hospital for Sick Children genome annotation groups for providing the annotation data that currently incorporated into Vega. We also thank the MHC haplotype consortium for providing sequence and analysis of the MHC haplotypes. We thank the Ensembl Project for the software that is the basis of the Vega website and the Otter client/server. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust Sanger Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Sprague,J., Bayraktaroglu,L., Clements,D., Conlin,T., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Mani,P. *et al.* (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res.*, **34**, D581–D585.
2. Ashurst,J.L., Chen,C.K., Gilbert,J.G., Jekosch,K., Keenan,S., Meidl,P., Searle,S.M., Stalker,J., Storey,R. *et al.* (2005) The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.*, **33**, D459–D465.
3. ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
4. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
5. Guigo,R., Flicek,P., Abril,J.F., Reymond,A., Lagarde,J., Denoeud,F., Antonarakis,S., Ashburner,M., Bajic,V.B. *et al.* (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome. Biol.*, **7** (Suppl 1), S2. 1–31.
6. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome. Biol.*, **7** (Suppl 1), S4. 1–9.
7. Tress,M.L., Martelli,P.L., Frankish,A., Reeves,G.A., Wesselink,J.J., Yeats,C., Olason,P.L., Albrecht,M., Hegyi,H. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.
8. Hart,E.A., Caccamo,M., Harrow,J.L., Humphray,S.J., Gilbert,J.G., Trevanion,S., Hubbard,T., Rogers,J. and Rothschild,M.F. (2007) Lessons learned from the initial sequencing of the pig genome: comparative analysis of an 8 Mb region of pig chromosome 17. *Genome. Biol.*, **8**, R168.
9. Renard,C., Hart,E., Sehra,H., Beasley,H., Coggill,P., Howe,K., Harrow,J., Gilbert,J., Sims,S. *et al.* (2006) The genomic sequence and analysis of the swine major histocompatibility complex. *Genomics*, **88**, 96–110.
10. Horton,R., Coggill,P., Miretti,M.M., Sambrook,J.G., Traherne,J.A., Ward,R., Sims,S., Palmer,S., Sehra,H. *et al.* (2006) The LRC haplotype project: a resource for killer immunoglobulin-like receptor-linked association studies. *Tissue Antigens*, **68**, 450–452.
11. Sambrook,J.G., Sehra,H., Coggill,P., Humphray,S., Palmer,S., Sims,S., Takamatsu,H.H., Wileman,T., Archibald,A.L. *et al.* (2006) Identification of a single killer immunoglobulin-like receptor (KIR) gene in the porcine leukocyte receptor complex on chromosome 6q. *Immunogenetics*, **58**, 481–486.
12. Eaves,I.A., Wicker,L.S., Ghandour,G., Lyons,P.A., Peterson,L.B., Todd,J.A. and Glynne,R.J. (2002) Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. *Genome Res.*, **12**, 232–243.
13. Prochazka,M., Leiter,E.H., Serreze,D.V. and Coleman,D.L. (1987) Three recessive loci required for insulin-dependent diabetes in nonobese diabetic mice. *Science*, **237**, 286–289.
14. Wicker,L.S., Todd,J.A. and Peterson,L.B. (1995) Genetic control of autoimmune diabetes in the NOD mouse. *Annu. Rev. Immunol.*, **13**, 179–200.

15. Allcock,R.J., Atrazhev,A.M., Beck,S., de Jong,P.J., Elliott,J.F., Forbes,S., Halls,K., Horton,R., Osoegawa,K. *et al.* (2002) The MHC haplotype project: a resource for HLA-linked association studies. *Tissue Antigens*, **59**, 520–521.
16. Stewart,C.A., Horton,R., Allcock,R.J., Ashurst,J.L., Atrazhev,A.M., Coggill,P., Dunham,I., Forbes,S., Halls,K. *et al.* (2004) Complete MHC haplotype sequencing for common disease gene mapping. *Genome Res.*, **14**, 1176–1187.
17. Traherne,J.A., Horton,R., Roberts,A.N., Miretti,M.M., Hurles,M.E., Stewart,C.A., Ashurst,J.L., Atrazhev,A.M., Coggill,P. *et al.* (2006) Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. *PLoS Genet.*, **2**, e9.
18. Debenham,S.L., Hart,E.A., Ashurst,J.L., Howe,K.L., Quail,M.A., Ollier,W.E. and Binns,M.M. (2005) Genomic sequence of the class II region of the canine MHC: comparison with the MHC of other mammalian species. *Genomics*, **85**, 48–59.
19. The BioSapiens Network of Excellence. (2005) Research networks: BioSapiens: a European network for integrated genome annotation. *Eur. J. Hum. Genet.*, **13**, 994–997.
20. Han,A., Kim,W.Y. and Park,S.M. (2007) SNP2NMD: a database of human single nucleotide polymorphisms causing nonsense-mediated mRNA decay. *Bioinformatics*, **23**, 397–399.
21. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.
22. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
23. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
24. Gilbert,D. (2003) Shopping in the genome market with EnsMart. *Brief Bioinform.*, **4**, 292–296.
25. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. *et al.* (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.