Tutorial section

VEGA, the genome browser with a difference

Abstract
The Vertebrate Genome Annotation (Vega) database is a community resource for browsing manual annotation from a variety of vertebrate genomes of finished sequence (http://vega.sanger.ac.uk). Vega is different from other genome browsers as it has a standardised classification of genes which encompasses pseudogenes and non-coding transcripts. The data is manually curated, which is more accurate at identifying splice variants, pseudogenes poly(A) features, non-coding and complex gene structures and arrangements than current automated methods. The database also contains annotation from regions, not just whole genomes, and displays multiple species annotation (human, mouse, dog and zebrafish) for comparative analysis. Vega encourages community feedback that results in annotation updates and manual annotation of finished vertebrate sequence.

Since completion of the draft human genome sequence in 2000\textsuperscript{1,2} and the subsequent finishing of this in 2003\textsuperscript{3} many different genome browsers have been developed to enable scientists to access genome data. The initial interpretation of the human genome was through automated annotation such as Ensembl\textsuperscript{4} and the UCSC genome browsers.\textsuperscript{5} There are currently limits to an automated approach for the analysis of genomes, for example in duplicated regions identifying unprocessed pseudogenes, and therefore there is still a need for manual intervention. As the genome sequence became finished, quality curated browsers such as MapView\textsuperscript{6,7} and the H-InvDB\textsuperscript{8,9} were developed.

The Vertebrate Genome Annotation (Vega) database\textsuperscript{10} is a community resource for browsing manual annotation from a variety of vertebrate genomes of finished sequence.\textsuperscript{11} Vega is based on the Ensembl schema, with gene objects shown in shades of blue, and also incorporates curation-specific data. The database allows users to view the manual annotation provided by the Havana group at the Wellcome Trust Sanger Institute (WTSI),\textsuperscript{12} IMB-Jena, the Joint Genome Institute, Genoscope and Washington University.

It currently contains the manual annotation of ten human chromosomes (6, 7, 9, 10, 13, 14, 20, 22, X and Y). As the genome sequencing centres publish the annotation and analysis of their chromosomes then the data will be accessible in Vega.

Why is Vega different from other browsers?

- It has a standardised classification of genes which encompasses pseudogenes and non-coding transcripts.
- PolyA sites/signals are annotated.
- The data are manually curated.
- The data are periodically updated.
- It contains annotation of haplotypes.
Single nucleotide polymorphisms (SNPs) are mapped to manual curation.

It is multispecies and small regions of finished sequence can be submitted and annotated as well as whole genomes.

It encourages community feedback and results in annotation updates.

**GENE CLASSIFICATION**
A standardised set of definitions has been used to categorise the annotation of the different gene features (Table 1). Irrespective of which category gene objects have been assigned to all annotated gene structures are supported by homology to cDNAs, expressed sequence tags (ESTs) or protein sequences.

**GENE NAMING**
It is important to use the correct gene nomenclature to maintain consistency in the annotation database, especially when comparing haplotypic or syntenic regions. The Vega annotators interact closely with the nomenclature committees from the Human Genome Organisation (HUGO, HGNC), Zebrafish Information Network (ZFIN) and Mouse Genome Database (MGD). If an approved symbol is not available for a gene locus, an interim identifier is used in the format of international clone identifier followed by number, eg RP11-695B14.2.

All loci and their associated transcripts and exons are given stable versioned database IDs (eg OTTHUMG00000021027) that are generated and tracked in the Otter database that underlies Vega (see Figure 1). Whenever a locus is edited the version number increases and the date of the change saved.

**MAIN FEATURES OF VEGA**
Manual annotation is currently more accurate at identifying splice variants, pseudogenes, polyadenylation features, non-coding genes, complex gene arrangements and clusters than automated methods.

**Splice variants** account for approximately 50 per cent of gene loci in finished chromosomes 9, 10 and X, with an average of 2.5 alternative transcripts per locus. Note the majority are non-coding but have canonical splice sites. Splice variants must be supported by splicing EST/cDNA evidence, but the presence of a coding sequence (CDS) is not essential. Hence the majority of variants are annotated without a CDS. ESTs and cDNAs from different species are also used as evidence to predict alternative transcripts as genome comparison studies have shown that gene structures are generally conserved between human and mouse.

**Pseudogenes** are defined as non-functional copies of genes and are categorised in Vega into unprocessed and processed pseudogenes (viewed in two shades of grey). They are generated by
either of two mechanisms: retrotransposition or duplication of genomic DNA. Those that arise from retrotransposition are called processed pseudogenes\(^\text{18}\) and have no 5' promoter sequence or introns but generally have an integrated poly(A) tail at the 3' end that often retains the poly(A) signal.

Unprocessed pseudogenes have arisen from genomic duplication and often have a structure that is very similar to the ancestral gene and may even splice correctly. The majority of pseudogenes of both types contain frameshifts and/or stop codons in the coding region. Pseudogenes are valuable in annotation as they have been implicated in human disease\(^\text{19}\) and can be used to study evolution.

**Poly(A) sites /signals** are annotated and may be browsed in Vega. Poly(A) signals are displayed in light red and poly(A) sites in dark red in contigview. Alternative polyadenylation appears to affect many higher eukaryotes, mainly in a tissue-dependent manner which may be implicated in disease.\(^\text{20}\) All poly(A) features are checked manually, using large numbers of ESTs marking out the 3' ends of genes and the fact that signals (of which there are 10 variants in human\(^\text{21}\)) are usually found within 60 bases of the poly(A) site.

**SNPs** can be viewed in ContigView and are mapped from the Glovar database\(^\text{22}\) onto the clones within Vega. Glovar contains all the data from dbSNP together with SNPs found from comparisons of the trace repository\(^\text{23}\) with the current genome build. Using Vega annotation, SNPs are classified as coding (red), untranslated region (pink), intronic (blue) or other (grey).

**ACCESSING AND QUERYING DATA**

As the Vega browser is based on Ensembl web code it has similar standard entry points such as keyword search and...
similarity searching (BLAST, SSAHA). ExportView can be used to download data in formats such as FASTA, Gene Feature Format (GFF) and flat files. There is also direct access to annotation via a distributed annotation server (DAS). If required, the Ensembl API24 can be used to perform more comprehensive searches of the Vega data. Also Vega genes mapped to the current genome assembly can be downloaded from Ensembl using Ensmart.

**MHC HAPLOTYPE ANNOTATION**

Unlike other browsers Vega can also contain annotations from regions, not just whole chromosomes. Regions available include the haplotype COX for the major histocompatibility complex (MHC) on human chromosome 6, with more haplotypes to follow.25

**ACCESSING MULTISPECIES ANNOTATION IN VEGA**

Vega can display multiple species annotation for comparative analysis. In the mouse annotation browser selected regions such as the Del36H deletion region on chromosome 13 and the insulin–dependent diabetes (IDD) susceptibility loci regions. The latter are annotated in both the reference mouse strain (C57BL/6) and the non-obese diabetic (Nod) strain.26

The zebrafish genome is being sequenced in its entirety at the Sanger Institute and Vega will be the main site for browsing the manually curated data. The reference is Tuebingen strain and Vega currently displays chromosomes/linkage groups 1–25 plus one artificial chromosome, U, that contains all clones with unknown chromosomal locations. The AB chromosome displays clones from the AB strain. Manual annotation is added on a monthly basis and clones which have not yet been annotated (displayed in grey) are shown with features from automated computational analysis (repeat masking, BLAST searches, etc).

Recently the finished sequence of the MHC (DLA) class II region from the dog breed Doberman has been annotated and is available in Vega.27 The sequence displays a high level of conservation with the human, cat and mouse class II region.

**COMMUNITY FEEDBACK**

Vega is a community annotation database and therefore to maintain up-to-date annotation it is essential to have feedback from researchers. A webform28 is available by which users can contact the Vega team to improve/correct annotation if there is additional evidence. Manual annotation of finished vertebrate sequence may also be submitted if it has been peer reviewed and/or meet the annotation standards.29

**FUTURE DEVELOPMENTS IN VEGA**

Currently available genome browsers often display different transcript structures for the same loci. In order to produce a single standard human gene set the Consensus CDS (CCDS) project has been set up between NCBI, USCS, Ensembl and the Havana group. The aim is to compare the human gene sets produced by RefSeq, Ensembl and Vega and then identify transcripts where the protein coding region is agreed on by all collaborators. These CDSs will be identified by stable CCDS identifiers in all the browsers.

In the near future manual annotation of the regions for the ENCODE project30,31 will be displayed in Vega. As mouse and zebrafish genomes reach completion it is hoped that the manually annotated orthologues may be browsed using multicontigview which is already available in Ensembl.

**Acknowledgments**

I gratefully acknowledge the help of Dr Jennifer Ashurst and Dr Laurens Wilming at the Wellcome Trust Sanger Institute.

Dr Jane Loveland
HAVANA Group,
Wellcome Trust Sanger Institute,
Wellcome Trust Genome Campus,
Hinxton CB10 1SA, UK
References


9. URL: http://www.h-invitational.jp/.


11. URL: http://vega.sanger.ac.uk

12. URL: http://www.sanger.ac.uk/HGP/Havana


22. URL: http://www.glovvar.org/

23. URL: http://trace.ensembl.org/

24. URL: http://www.ensembl.org/Docs/


28. URL: http://vega.sanger.ac.uk/helpdesk/index.html

29. URL: http://sanger.ac.uk/HGP/Havan/docs/guidelines.pdf


31. URL: http://www.genome.gov/1005107/