Manual Annotation and the Vega Genome Browser

Laurens G. Wilming
Informatics Department
Wellcome Trust Sanger Institute, UK

1 Introduction

Traditional large-scale genome sequencing projects have generated a wealth of sequence data over the years. With the new generation sequencing technologies currently being employed, even small-scale projects will be producing more data than all the traditional large-scale projects combined and in only a fraction of the time [G10KCOS, 2009; Kaiser, 2008; Siva, 2008]. To be of any value, this sequence needs to be annotated with gene loci and associated features. Most of this new sequence data is or will be available for any scientist to use, but as many will lack the time, resources, and expertise to process and interpret large quantities of sequence data, providing comprehensive annotation of this bare genomic sequence is important. Essentially, there are two ways of annotating genes and other features: automatic and manual. Researchers can employ these methods themselves or rely on centrally provided general resources, such as Ensembl and RefSeq, or more specialised resources such as Gramene or Dictybase. Most resources depend on automated annotation methods. Automatic annotation can be as simple as using ab initio gene prediction software such as FGENES [Solovyev et al., 1995], GENSCAN [Burge & Karlin, 1997], MZEF [M. Q. Zhang, 1997], or Augustus [Stanke et al., 2004; Stanke et al., 2006]. However, this approach is not always satisfactory, as it results in considerable over- and under-prediction [Harrow et al., 2006; Rogic et al., 2001]. A more successful approach is to combine ab initio predictions with homology data (e.g., protein, mRNA and EST matches). This is used to a notable effect by GeneWise [Birney & Durbin, 2000] and GenomeScan [Yeh et al., 2001] in Ensembl [Hubbard et al., 2002] and Otto [Venter et al., 2001], respectively. While these processes improve over time and may be sufficient for annotation of low-coverage genomes (at least most protein coding genes are located, even if their predicted structure is not necessarily correct), for a reference genome, manual annotation is required. This is especially true for regions of a genome that have been subjected to extensive rearrangements or duplications. Clusters of tandemly duplicated genes or segmental duplication of groups of non-related genes still, and possibly always will, confound accurate and complete annotation by these methods. Automatic protocols look at sequences in isolation, although some use comparative data [Korf et al., 2001; Meyer & Durbin, 2004]. In contrast, human annotators can look at genes and sequences in any chosen context, such as other parts of the same genome, another genome, or other family members. They can also pull in data from any number of sources, such as literature, personal communications, and personal expertise. The human and vertebrate analysis and annotation (Havana) group at the Wellcome Trust Sanger Institute provides this kind of manual annotation for human, mouse, and zebrafish genomic sequence.

To create reliable, comprehensive, high-quality annotation on at least a set of reference genomes is now more important than ever. Many of the new genomes that will be released are actually variations of
Figure 1: Havana annotation system pipeline and dataflow. Information flow of Havana manual annotation, showing how the information resulting from manual annotation propagates via many collaborations. NOMENCLATURE = human, mouse, and zebrafish nomenclature committees; GRC = Genome Reference Consortium; CCDS = Consensus Coding Sequence collaboration; IKMC = International Knockout Mouse Consortium.

reference genomes (17 strains of mice, and human 1000 genomes project [Kaiser, 2008; Siva, 2008]) or are closely related to them (e.g., gorilla). This means that existing reference annotation will be mapped or projected onto these new genomes. Quality manual reference annotation also contributes to genomes where projection is not a feasible option, that is, genomes that need de novo automatic annotation. The Havana manual annotation of protein coding regions is either submitted directly to the UniProt protein sequence database or indirectly via Ensembl (http://www.ensembl.org) through the Ensembl/Vega gene merge. Vega (http://vega.sanger.ac.uk) is the Ensembl-derived database or genome browser specialising in Havana manual annotation [Ashurst et al., 2005] (see the Conclusion section). As Ensembl automatic annotation uses protein homologies to build its gene models [Hubbard et al., 2002], the manual annotation contributes in a roundabout way even in genomes without it. In an even more roundabout contribution, our annotation is used as reference for testing, comparing and refining both algorithms and experiments [Guigo et al., 2006; Harrow et al., 2006]. Figure 1 depicts the information flow of Havana annotation. The input is an analysis pipeline derived from the Ensembl pipeline and consists of homology searches against protein and nucleotide sequence, protein domain and repeat databases, gene predictions, and raw sequence analysis (CpG islands, tandem repeats). Additionally, external annotation is pulled in from Ensembl and RefSeq. We can also display any number of distributed annotation system (DAS) sources [Dowell et al., 2001], such as predicted pseudogenes or non-canonical introns.

In the following section, I will illustrate the advantages of manual annotation utilising the tools briefly described above, using particularly challenging examples of features, genes, and regions that are not well-served by automatic annotation.

2 Discussion

2.1 Annotation Aids Assembly Accuracy

The quality of genomic sequence assembly and the sequence itself is improved by manual annotation [Taudien et al., 2004]. In the course of annotation, we have detected potential sequencing or finishing errors
because they affect the coding region (CDS) of a gene, introducing a frameshift or in-frame stop codon not supported by any transcripts. In the vast majority of cases, the genomic sequence is in error, mostly due to compression artefacts in homonucleotide stretches. On some occasions, the discrepancy is actually due to polymorphism (e.g., single nucleotide polymorphism, SNP, or deletion insertion polymorphisms, DIP), effectively knocking-out the gene in certain individuals and not in others. Often this occurs in regions of known polymorphic activity (e.g., human MHC region on chromosome 6) or in clusters of similar genes where one can expect some level of functional redundancy [Yngvadottir et al., 2009]. Anytime the source genomic material comes from a mix of individuals, as is often the case, especially with older projects, there is the possibility of encountering instances of variation involving not only single nucleotides but also larger stretches of sequence. During annotation of human chromosome 20, we found clones that could not be assembled properly, as evidenced by the incorrect arrangement and structure of genes, such as SIRPD and others, in that area. Further investigation revealed that this was caused by a large DIP (33 kb) with an estimated frequency of the insertion allele of 37% in a panel of 174 Caucasians [Deloukas et al., 2001]. The affected genes are part of a cluster of related genes located in the area, and not all family members need to be functional. A slightly different example is the ABO blood group gene on human chromosome 9. It was contained on two overlapping genomic clones (in this case bacterial artificial chromosomes, BACs) in the then-current assembly. However, the BACs represent different alleles for this gene. Thus, the gene, as it

Figure 2: Extensive rearrangements between mouse and human in the mouse brown deletion region. The brown deletion region mouse chromosome 4 and the equivalent regions on human chromosome 9 are shown as grey, purple, and blue areas. At syntenic breakpoints (red stars) in the human chromosomes, synteny with other mouse chromosomes is also indicated (orange, yellow, green, and red areas). Named genes at the edges of the syntenic blocks are indicated for reference (black arrows). Blue arrows indicate the two parts of the novel tandemly duplicated mouse gene cluster not present in human. Green arrows indicate the inverted Frmd3 gene (dark green arrow) and the mouse-specific 3 end Frmd3 duplication (light green arrow, dark purple area). Figure is not to scale; black circle indicates centromere.
presented itself on the assembly, contained a 5’ end of one allele mated to the 3’ end of another.

Genomic sequence errors and polymorphic events that may otherwise have escaped detection are both brought to our attention, incidental to the process of manual annotation. For known RefSeq genes, the initial detection of discrepancies adversely affecting translation is largely automated, but the evaluation of these is still a manual process. When it concerns human and mouse genomic sequence (and soon zebrafish as well), this is conducted in the context of the Genome Reference Consortium (GRC) (http://www.ncbi.nlm.nih.gov/projectsgenomeassemblygrc/) (see comment by Dolgin [Dolgin, 2009]). A collaboration among the Wellcome Trust Genome Campus (Sanger Institute and EBI), the Washington University Genome Center and the NCBI, the GRC are the custodians of the human and mouse reference genomic sequence tasked with correcting errors and presenting the most accurate and representative reference sequence to the community.

2.2 Duplications: Rearrangements and Segmental Duplications

Some chromosomal regions have undergone extensive evolutionary rearrangements, often combined with segmental duplications. Small, rearranged genomic segments are likely to remain undetected by large-scale cross-species alignments. Thus, here, gene annotation can contribute by filling in the gaps. Duplications often result in pseudogenisation, and as pseudogenes can deliver vital clues about the history and evolution of chromosomes, it is imperative that they are annotated as part of the comprehensive annotation of genomic sequence.

The brown deletion region on mouse chromosome 4 is orthologous to human chromosome 9 with respect to the broad gene content. However, the human equivalent of the contiguous stretch of mouse sequence, roughly between genes Ras3 and Bnc2, is distributed over distinct and distant blocks of sequences on chromosome 9 on either side of the centromere (Figure 2). Additionally, mouse has a small cluster comprising a set of ten tandemly duplicated novel genes plus two pseudogenes derived from these that are not present in human or rat [Smyth et al., 2006]. The genes are extremely conserved between each other with a translation that had no homology to the protein databases at the time. Due to this lack of protein homology, the two pseudogenes were only discovered after annotation of the expressed copies and alignment of the translation to the genomic sequence. Clearly, this is an active region with respect to chromosomal breakage, resulting in large and small evolutionary rearrangements. One of the smaller ones is an evolutionary or synteny breakpoint between the Frmd3 and Jmjd2c genes generating a novel gene from the duplication of the 3’ end and inversion of the Frmd3 gene in mouse and rat (green arrows in Figure 2).

Human chromosome 17 contains the Charcot-Marie-Tooth Disease (CMTD) and Smith Magenis Syndrome (SMS) associated regions, the equivalent of which can be found on mouse chromosome 11. Whereas both orthologous regions are contiguous, the internal order of sequence blocks is completely different between the two species, with at least ten blocks (some only containing a single gene) rearranged. Half of the synteny breakpoints occur at the regions of segmental duplication in human (coloured bounding boxes in Figure 3). Most of these duplications contain mainly pseudogenes, which are not often predicted by automated annotation systems or incorrectly predicted as coding genes. As mentioned earlier, pseudogenes provide vital clues in discovering the evolutionary origins and history of rearranged regions. The human genomic structure of this region is the result of a series of complex duplications, insertions, and pseudogenisation events. A keratin cluster, containing mostly pseudogenes and a variety of other genes, appears three times in this region. The ancestral cluster of active genes is located on the q arm of the chromosome (black box labelled A4 in Figure 3; true gene content not shown due to space constraints). In studying the gene content of the three copies in the CMTD/SMS region, one can propose a sequence of duplication and insertion events involving the ancestral cluster and the ancestral LGALS and NOS32A genes on 17q (yellow box
Figure 3: Extensive rearrangements between mouse and human in the human CMTD1 and SMS region. Synteny between human chromosome 17 and mouse chromosome 11 is indicated by grey, blue, and yellow areas. Genes are placed on human sequence (see legend). Segmental duplications on the human chromosome are indicated by variously shaped and coloured bounding boxes. Note that many of the syntenic breakpoints coincide with these duplications. Linked diamonds indicate sets of genes and the corresponding unprocessed pseudogenes, further indicating rearrangements and duplications. Double hash marks separate non-contiguous regions. Figure is not to scale.

Figure 4: Extensive rearrangements between human and mouse around the TPTE genes and pseudogenes. The various instances of TPTE genes are indicated by red rectangles (genes) or ovals (pseudogenes). The coding versions of the gene (one in mouse, two in human) are marked with an arrow head. Various shaped bounding boxes enclose equivalent gene sets (variously coloured rectangles) between human and mouse. Note that while mouse has single copies of all the genes shown, many genes are duplicated and/or pseudogenised in human. Note also that two separate contiguous blocks in mouse changed into six blocks in human.

labelled B3 in Figure 3), giving rise to clusters A2 and A3 containing B2 and B3, respectively, and to cluster A1. Further evidence of extensive rearrangement activities is the presence of a number of gene-pseudogene pairs or groups distributed throughout the region (linked diamonds in Figure 3).

Another example of extensive rearrangement and duplication-insertions in human versus mouse involves the ortholog of mouse Tpte. Mouse has one Tpte gene on chromosome 8 (indicated in Figure 4). Human has eight copies of the gene, mostly pseudogenes (only two, TPTE and TPTE2, are active), distributed over three chromosomes. The copies on chromosome 13 are distributed amongst four groups of genes located megabases apart. Several of these flanking genes, which in mouse are single-copy and active, have also been duplicated and pseudogenised (Figure 4). Judging from the genomic context of the Tpte gene (enclosed in a circle in Figure 4), the ancestral copy has become a pseudogene in human.

On human chromosome 10, we find several genes (coding and transcribed pseudo) with a complex aetiology generated by a succession of rearrangements, joining disparate fragments of genomic sequence
containing parts of the BMS1L and CTGLF genes. Interestingly, in one of the examples, a transcript of the CTGLF2 gene contains a large protein coding 3’ exon that is a repurposed or exapted CENTG2 processed pseudogene, representing the last ten exons of the parent CENTG2 gene (black ovals at the bottom in Figure 5). Notice the difference in annotation between Ensembl and manual, especially for the pseudogenes.

2.3 Duplications: Clusters and Tandem Duplications

In the above section I discussed segmental duplications mostly involving groups of different genes. However, tandemly duplication gene clusters pose their own unique challenges, both with regards to annotating as well as assembling the genomic sequence. Clusters containing multiple copies of tandemly duplicated genes can be difficult to assemble if the copies are very similar. A whole genome shotgun sequence is especially sensitive to the issue of tandem duplications and assembly algorithms are liable either to collapse multiple nearly identical copies into a small number of copies or to leave them as islands in a sea of gaps. Even clone-based assemblies often have gaps in clusters like these, gaps that can be difficult to close. The problem is aggravated by copy number polymorphism, as shown for example in the GAGE cluster on human chromosome X [Ross et al., 2005] and the Mup (major urinary protein) gene cluster on mouse chromosome 4 [Mudge et al., 2008]. Between both individuals and strains, the latter cluster is thought to be highly variable [Bennett et al., 1982; Beynon et al., 2002; Clissold & Bishop, 1982]. Combined with the high level of conservation between the copies of the genes, assembly is a challenge. At the time of writing, there are still gaps and ‘floating’ clones in this particular region of the otherwise finished mouse chromosome. Detailed analysis through annotation of the BACs can help determine whether clones purportedly overlapping, do indeed overlap, and whether floating clones can indeed not be placed. The Mup gene cluster is not present on human chromosome 9, except for a single pseudogene. Due to their highly repetitive nature and frequently high levels of conservation among family members, tandemly duplicated gene clusters are generally poorly annotated by automatic means. Sometimes there is a ‘mix-and-match’ approach to the predicted genes (exons from adjacent genes are combined to form fictitious genes covering more than one locus), and more distant family members can be missed out because of a lack of transcriptional evidence. On the other hand, pseudogenes of single exon genes, such as histone genes, or unprocessed pseudogenes of cluster gene family members, are frequently incorrectly predicted to be coding genes. Hist1 histone family genes form a cluster on mouse chromosome 13 and human chromosome 6. Given that all members of a given histone type have virtually identical protein sequences, the initial nomenclature automatically assigned to these genes by, for example, Ensembl is often incorrect: the same name can be attached to multiple genes, and/or the incorrect name is used. The annotation of all the histone genes and pseudogenes in this cluster in both human and mouse allowed the Human Gene Nomenclature Committee (HGNC), and subsequently Mouse Genome Informatics (MGI), to assign new rationalized names and symbols to the genes. Annotation also revealed that in Marzluff et al. [Marzluff et al., 2002], part of the cluster was incorrectly assembled or aligned to the human sequence, changing the perceived gene content and arrangement of the cluster. In comparing mouse and human, it was due to annotation of the pseudogenes that a pattern of groups of histone genes emerged as core repeat units, suggesting a duplication mechanism involving successive duplications-insertions of the core unit or sub-parts thereof (coloured boxes in Figure 6) [Mallon et al., 2004].

Located within the histone cluster, the butyrophilin (Btn) cluster has undergone expansion in human. On human chromosome 6, there is a cluster of seven BTN genes, while mouse has only two on chromosome 13, which can be paired to direct orthologs in human (shaded area in Figure 6). Between the Btn and histone gene clusters is a vomeronasal receptor gene cluster, which massively expanded in mouse from five genes in human (all pseudogenes) to sixty-seven in mouse (including thirty-four pseudogenes). Tandemly duplicated single-exon genes such as these pose a particular challenge with respect to determining the nature of genes as
Figure 5: Genes with complex relationships, combining exons from separate ancestral genes. Three panels show three separate parts of chromosome 10. Coding transcripts (with green exons), non-coding transcripts (red), and pseudogenes (light green, narrow) were built based on protein and transcript evidence. Note that the long transcript in the middle panel shares exons with both the CTGLF2 gene and the BMS1LP4 pseudogene. It does not have a viable CDS. Linked coloured curly brackets indicate equivalent exons. Note some discrepancies between Ensembl predictions (blue coding region, pink UTR) and manual annotation, especially for BMS1LP4, which Ensembl predicts as protein coding because it skips over the pseudogene event (frameshift).
Figure 6: The BTN and HIST1 gene clusters on human chromosome 6 and mouse 13. Coloured rectangles indicate various gene families (see legend). Coloured bounding boxes indicate the proposed histone gene duplication units with conserved content and orientation, where smaller ones are subsets of larger ones through the deletion of single genes. Since the human-mouse evolutionary divergence, genes have been deleted, inserted or pseudogenised in one species or the other. Note that the histone gene cluster is interrupted by the butyrophilin gene cluster (shaded area). In mouse, the vomeronasal receptor gene cluster is located between the left histone gene cluster and the butyrophilin gene cluster. Figure is not to scale and only shows relevant family genes (interspersed non-family genes and other clusters are not shown).

Figure 7: The SERPINB gene cluster on human chromosome 6 and mouse 13. Genes from the three serpin b subfamilies located in this cluster (1, 6 and 9) are shown (see legend). Box marks the proposed ancestral unit that duplicated-inserted in mouse. Note that in mouse, the gene order is conserved (outside the boxed area), but the individual subfamily members have multiplied and often pseudogenised. Gene marked with an arrow head is partial, lacking its 3' end. Figure is not to scale and only shows relevant family genes (interspersed non-family genes are not shown).

pseudogenes or protein coding genes. To determine whether a premature stop codon close to the consensus stop codon would render the gene a pseudogene (and not a slightly shorter family member) requires looking at the effect on protein domains, for example.

Similar to the situation with the histone cluster, only after annotating all the SERPINB genes, including the pseudogenes, on human chromosome 6 and mouse 13, can we get a picture of how the cluster evolved. The ancestral arrangement in human of one of each of the three subfamilies (plus possibly an additional copy of B9, pseudogenised in human and only partially duplicated in mouse) expanded to a cluster in mouse (twenty-seven genes and pseudogenes spread across the three subfamilies) (Figure 7) [Mallon et al., 2004].

Segmental duplications and rearrangements are important events in evolution, adding complexity and variety [Coghlan et al., 2005; Eichler, 2001; Freeman et al., 2006; Mehan et al., 2004; Perry, 2008; Rodin et al., 2005; Sharp et al., 2005]. Disease-related and species-specific development regions of human and mouse
chromosomes can display high levels of inter- and intra-species polymorphism and variation [Antonell et al., 2005; Conrad et al., 2010; Dunham et al., 1989; Graubert et al., 2007; C. J. Shaw & Lupski, 2004; She et al., 2008; Stewart et al., 2004; Traherne et al., 2010; F. Zhang et al., 2009]. Careful study and annotation of these regions in a comparative fashion adds greatly to our understanding of their aetiology and how they contribute to diseases. The CMT region on human chromosome 17 discussed above is a case in point. The region is subject to copy number variation (CNV) [Weterman et al., 2010], and accurate annotation of this region in human and mouse is essential for the interpretation of these variations. Similarly, the beta defensin gene cluster on human chromosome 8 is subject to CNV [Hollox et al., 2003] and benefited from manual annotation [Amid et al., 2009]. In many ways, comparative and variation studies go hand-in-hand. To obtain the best results from the variation studies, the datasets being compared are ideally annotated to the same standard. Moreover, it is easier and it improves the annotation of complex regions, such as, for example, Major Histocompatibility Complex (MHC) regions [Debenham et al., 2005; Horton et al., 2008; Renard et al., 2006], when it is conducted on different genomes in parallel. Thus, there is a mutually beneficial relationship between comparative and variation analysis and manual annotation.

2.4 Splice Variants and Unusual Gene Structures and Arrangements

Splice variants are a main contributor to biological diversity and to complexity in higher organisms, with the vast majority of human and mouse genes undergoing variant splicing. They are subject to tissue-specific regulation and inter- and intra-species variation and are involved in disease and pharmacology [Ben-Dov et al., 2008; Blekhman et al., 2010; Carling et al., 2009; Coulombe-Huntington et al., 2009; Ghigna et al., 2008; Kim et al., 2008; Korner & Miller, 2009; Park & Graveley, 2007; Tazi et al., 2009; E. T. Wang et al., 2008; Z. Wang & Burge, 2008; Yeo et al., 2005; Zhao et al., 2009].

A combination of splice variants and overlapping genes contributed to the incorrect automatic annotation of the group of genes bracketed by APOM and LY6G5B (Figure 8). At the time of manual annotation, in Entrez Gene (then LocusLink), four of these genes, namely, APOM, C6orf47, CSNK2B, and LY6G5B were merged into a single APOM locus. Note that the single-exon C6orf47 gene is located on the opposite strand relative to the other three genes. Its incorporation was possibly due to the confusion with the novel transcript gene that overlaps the C6orf47 gene on the opposite strand and was probably considered to be a splice variant. Apart from this overlap of opposite strand genes, several of the other genes have transcripts that overlap on the intron or exon level (Figure 8). However, the most interesting feature, and a major contributor to the incorrect automatic annotation, is the overlap between transcripts of the CSNK2B gene and its neighbour LY6G5B. A variant transcript from the CSNK2B gene shares exons and splice sites with the LY6G5B gene (pink arrow head 1 in Figure 8); however, the shared exons are translated in a different reading frame in the two transcripts. Thus, while there is structural overlap (identity even) on the nucleotide level, there is no homology at the protein level. Intriguingly, a second (partial) variant transcript joins all LY6G5B exons to one of the more 3’ exons of CSNK2B, but this time in frame with the existing reading frames (pink arrow head 2 in Figure 8). A different case of transcripts joining exons from two genes can be found amongst variants of the human ZNF695 gene that share the first exon and CpG island with that of an upstream ZNF670 gene (pink arrow heads in Figure 9). Notice the discrepancy between gene predictions (i.e., Ensembl, Genscan, Fgenesh) and the manual annotation.

Variant splicing is unfortunately an area underrepresented by automated gene predictors. Although Ensembl does predict some splice variants, it does not nearly detect all, and tends to construct variant structures that are an amalgam of different variants that do not necessarily go together in reality. Automatic algorithms also rarely correctly predict variant transcripts that use rare non-consensus splice sites [Mount, 2000]. Human annotators can annotate variants in detail and with greater accuracy, and this is beneficial in the study
Figure 8: Densely packed genes on human chromosome 6 with variant transcripts linking two genes. Transcripts linking CSNK2B and LY6G5B are indicated with pink arrow heads. Transcript 1 shares the translation of the CSNK2B exons, but the translation for the last two exons differs from that of LY6G5B (even though the splice sites are identical). Transcript 2 shares both CSNK2B and LY6G5B translations in-frame. Note the CSNK2B variant with its first exon in an intron of BAT4 on the opposite strand. At the top, the most upstream first APOM exon overlaps the most upstream first BAT3 exon. Note also the novel transcript gene overlapping the C6orf47 gene on the opposite strand.

Figure 9: Two zinc finger protein genes with variants linking the two. Transcripts linking the first exon of ZNF670 with exons from ZNF695 are indicated by pink arrowheads. The second exons of both genes share some homology. Both genes have a CDS that starts in exon one, but because the ZNF670 start is not in frame with the second exon of ZNF695, the linking transcripts do not have a CDS. Note that Ensembl only partially predicts the structure of these genes, due to the fact that the CDS on the first exon is only a few basepairs long.
on cross-species conservation. New techniques such as RNA-seq (next-generation high-throughput sequencing of transcripts) and protein mass spectrometry will be of great benefit to detecting and confirming splice variants and unusual transcripts such as those discussed above [Hiller et al., 2009; Jiang & Wong, 2009; Mortazavi et al., 2008; Nagalakshmi et al., 2010; Richard et al., 2010; L. Wang et al., 2010; Z. Wang et al., 2009; Wilhelm et al., 2010].

I have discussed only a fraction of the numerous unusual gene variants, structures, and configurations we have observed over the years. As automated annotation methods have difficulty correctly processing these types of genes, manual annotation is required to unpick these complex genes and regions.

2.5 Pseudogenes

Pseudogenes can tell much about how gene clusters evolve, how regions duplicate and rearrange, and what makes one species different from another [Balasubramanian et al., 2009; Niimura & Nei, 2005; X. Wang et al., 2006; Young et al., 2005; Z. Zhang & Gerstein, 2004]. They can even be used to find novel splice variants and the evolutionary history of parent genes [Shemesh et al., 2006], can be expressed and/or be functional [Andersen et al., 2004; Balakirev & Ayala, 2003; Edgar, 2002; Harrison et al., 2005; Hirotsune et al., 2003; Korneev et al., 1999; Mitrovich & Anderson, 2005; Suo et al., 2005; Yano et al., 2004; Zheng et al., 2007; Zheng et al., 2005], or be present as pseudogenes but only in certain haplotypes or species [Guo et al., 1998; Uhrberg, 2005]. In evolutionary terms, pseudogenes can also act as a reservoir from which new genes can be constructed (CENTG2, mentioned earlier (Figure 5), and [Harrison & Gerstein, 2002]. One of the strengths of manual annotation is its ability to distinguish gene from pseudogene, find pseudogenes and find potentially transcribed pseudogenes, neither of which can be easily done by automated systems except for the most obvious cases. However, Ensembl does predict some pseudogenes. Also, Torrents et al. [Torrents et al., 2003] and Khelifi et al. [Khelifi et al., 2005] applied pseudogene detection protocols to the vertebrate genomes. Nevertheless, despite advances in pseudogene detection algorithms [Coin & Durbin, 2004; Karro et al., 2007], manual annotators retain the edge. Automated systems benefit from higher sensitivity but at the cost of lower specificity resulting from both higher false positive and false negative rates [Nelson, 2004]. Cases such as retrotransposed pseudogenes with an intact CDS invariably need the attention of manual annotators. An example of a challenging pseudogene annotation is a rare instance of nested pseudogenes and repeats that can be found on human chromosome 10 (Figure 10). Through a succession of insertions, the current configuration of an Alu repeat inserted into an MSTB repeat inserted into an RPL29 pseudogene inserted into an ADIPOR1 pseudogene arose. The latter pseudogene independently received an Alu repeat insertion. Cases like this and the examples discussed in the Rearrangements sections above highlight the importance of manual annotation of pseudogenes.
3 Conclusion – Automatic Versus Manual Annotation: Mutual Benefits

In this chapter, I have highlighted the strengths of manual annotation with several examples of challenging gene structures, organisations, and clusters. Software for automatic annotation does of course improve continuously (partially due to the ever increasing amount of experimental and manual annotation data) [Brent & Guigo, 2004], and systems are invented (and often become obsolete) at a steady pace (GAIA [Bailey et al., 1998], DNannotator [Liu et al., 2003], EAnnot [Ding et al., 2004], FIGENEX [Gouret et al., 2005], AIR [Florea et al., 2005]). However, while programs such as Ensembl and EAnnot work ostensibly the same as human annotators, that is, aligning expressed sequences to the genome to build gene structures, they do not have the same flexibility to check whether the source data are correct to start with. Many mRNAs, from both large-scale projects such as KIAA [Nomura et al., 1994], FLJ [Hattori et al., 2000], DKFZ [Wiemann et al., 2001], and MGC [Strausberg et al., 2002], as well as individual projects, have incorrect CDSs annotated. What appears entirely reasonable on a processed mRNA sequence in isolation may not look right in the context of an exon structure on genomic sequence. Errors can be as simple as an open-ended 5’ end of the CDS, where in genomic context it is clear that the start is at the downstream ATG. More troublesome are CDSs in what in fact are 3’ UTRs, invariably in mRNAs that are 5’ incomplete and represent only or mostly 3’ UTR. These are troublesome because their translations are in the protein databases, leading to the creation of false positives by, for example, Ensembl, and causing confusion in the scientific community.

Access to Havana manual annotation of vertebrate reference genomes (zebrafish, mouse, and human), is mainly through the Vega genome browser (http://vega.sanger.ac.uk) [Wilming et al., 2008]. The Vega browser contains annotation not found elsewhere, such as the MHC and LRC regions of the pig [Renard et al., 2006; Sambrook et al., 2006] and gorilla, the MHC region in dog [Debenham et al., 2005] and wallaby [Siddle et al., 2009], diabetes candidate regions in different mouse strains [Steward et al., 2010], and the LRC region in different human haplotypes [Horton et al., 2006]. Annotation not unique to Vega, that is, large-scale whole chromosome annotation, is also visible in the Ensembl browser (http://www.ensembl.org) as either Havana transcripts or Ensembl/Havana gene merge transcripts. The latter are transcripts that are identical between Havana and Ensembl annotation. Via the Ensembl database, Havana annotation is visible in the UCSC Genome Browser (http://genome.ucsc.edu). Furthermore, through our collaborations with various groups (see below), reciprocal links have been established between our annotation and the HGNC (http://www.genenames.org), MGI (http://www.informatics.jax.org), and zebrafish (http://zfin.org) nomenclature databases.

Naturally, automatic annotation plays an important role in the world of sequence analysis. It is not possible to perform manual annotation on all the genomic sequences available now and in the future, as it is expensive and time-consuming. Due to the time-intensive nature of manual annotation, updating the annotation is also an issue. However, the sheer complexity and limitless variability of biological systems, with exceptions to the rules that are next to impossible to capture or predict reliably, means that for gold standard reference annotation, we need manual annotation. Human annotators can judge any event or occurrence on its merits and in context. Many instances, especially those involving duplications, rearrangements and variability, require a human eye for the optimal result. Manual annotation makes particular sense for the finished genomic sequence of reference model organisms, which is very expensive and time-consuming to produce and warrants a similar investment in manual annotation. However, for the majority of genomes from low-coverage whole genome shotgun projects and new generation sequencing projects, we need automated systems. These systems benefit from the manual annotation because the latter propagates through numerous collaborations. We are part of the Consensus Coding Sequence (CCDS) consortium, a collaboration between the Sanger Institute and the EBI (Havana and Ensembl), the NCBI (RefSeq), and the UCSC (Genome Browser), which aims to produce a consensus set of coding transcripts for human [Pruitt et al.,
2009]. We work closely with the human (HGNC) [Bruford et al., 2008; Povey et al., 2001], mouse (MGI) [Maltais et al., 2002; D. Shaw, 2004; Wain et al., 2003], and zebrafish (Zfin) [Mullins, 1995; Sprague et al., 2003] nomenclature committees to ensure proper and systematic naming of annotated loci. Finally, manual annotation feeds back into both experiments and algorithms, and vice versa, because of our involvement in the GENCODE project [Harrow et al., 2006]. This project spans different disciplines, ranging from experimental to purely computational, looking at genomes, genes, and their products from many different angles. It is due to the various feedback loops and close cooperation between manual annotation, automatic annotation and nomenclature, protein, and gene databases that both automatic and manual annotation have made great strides. Most of the examples I discussed in the Discussion section have greatly improved in the automatic annotation databases due to this.

With manual annotation, we aim to provide a comprehensive gold standard reference set as a foundation for the research community to build on.

Acknowledgements

I would like to thank my colleagues from the Havana group for supplying the annotation examples mentioned here. I would especially like to thank Jennifer Harrow, Adam Frankish, and Jonathan Mudge for their helpful discussions.

References


