Review

Identifying protein-coding genes in genomic sequences

Jennifer Harrow*, Alinda Nagy[†], Alexandre Reymond[‡], Tyler Alioto[§], Laszlo Patthy[†], Stylianos E Antonarakis[¶] and Roderic Guigó[§]

Addresses: *Wellcome Trust Sanger Institute, Wellcome Trust Campus, Hinxton, Cambridge CB10 1SA, UK. [†]Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, H-1113 Budapest, Hungary. [‡]Center for Integrative Genomics, Genopode Building, University of Lausanne, CH-1015 Lausanne, Switzerland. [§]Centre de Regulació Genòmica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, E-08003 Barcelona, Catalonia, Spain. [¶]Department of Genetic Medicine and Development, University of Geneva Medical School and University Hospitals of Geneva, Geneva 1211, Switzerland.

Correspondence: Roderic Guigó: Email: roderic.guigo@crg.es

Published: 30 January 2009

Genome Biology 2009, **10**:201 (doi:10.1186/gb-2009-10-1-201)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/1/201

Abstract

The vast majority of the biology of a newly sequenced genome is inferred from the set of encoded proteins. Predicting this set is therefore invariably the first step after the completion of the genome DNA sequence. Here we review the main computational pipelines used to generate the human reference protein-coding gene sets.

The genome sequence is an organism's blueprint: the set of instructions dictating its biological traits. The unfolding of these instructions is initiated by the transcription of the DNA into RNA sequences. According to the standard model, the majority of RNA sequences originate from protein-coding genes; that is, they are processed into messenger RNAs (mRNAs) which, after their export to the cytosol, are translated into proteins. While the importance of noncoding RNAs has come to the fore over the past ten years [1-5], proteins are still assumed to be the main functional and structural players in the cell. The delineation of the complete set of protein-coding genes and their alternative splice forms is, therefore, essential to the task of translating the information in the sequence of the genome into biologically relevant knowledge. This is not a trivial task, as illustrated by the fact that many years after the first drafts of the human genome sequence became available [6-8], uncertainty remains regarding the exact number of protein-coding genes [9], a number that might actually vary between individuals - and even between cells within the same individual - as extensive structural variation has been reported in the human genome [10-12].

Even the concept of a 'gene' is under revision. Genes have long been regarded as discrete entities located linearly along chromosomes, but recent investigations have demonstrated extensive transcriptional overlap between different genes. Specifically, genomic regions from otherwise distinct and apparently well characterized protein-coding loci (which may be very far apart in linear genomic space) often appear to combine to produce transcripts with the potential for encoding novel protein species [13,14].

Despite all these caveats, delineating the set of protein-coding genes is invariably the first step taken after completing the DNA sequencing of a genome. Indeed, the vast majority of the biology of a genome is initially inferred from the set of proteins that genome is predicted to encode. The gene-finding problem has consequently attracted wide attention within the field of bioinformatics. Since the early work of Fickett [15], in which methods were developed to distinguish coding from noncoding regions, a plethora of strategies have been explored and many methods developed to elucidate the exonic structure of genes in eukaryotic genomes. Figure 1 summarizes the main avenues of research. The technical details underlying these computational methods are reviewed in [16-18] and the references and URLs for the methods are given in Additional data file 1. Here we will focus on the strategies being applied to delineate a number of reference human gene sets - the ones most widely used by researchers in biology - and to assess their quality and

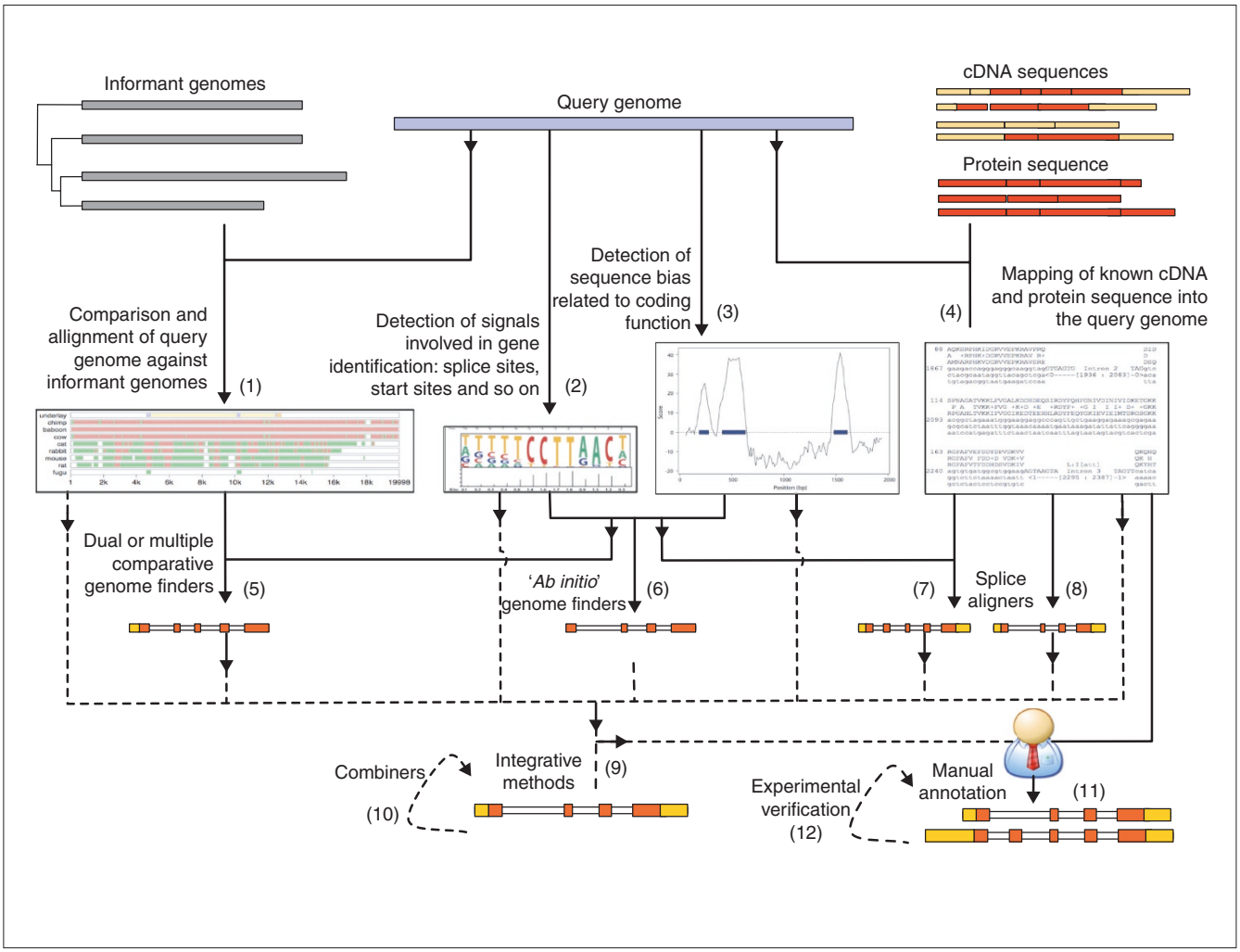

Figure 1

Gene-finding strategies. Given a genome DNA sequence, information on the location of genes and transcripts can be obtained from different sources: conservation with one or more informant genomes (1); intrinsic signals involved in gene specification, such as start and stop codons and splice sites (2); the statistical properties of coding sequences (3); and, most importantly, known transcript sequences (either full-length cDNAs or partial ESTs) and protein sequences (4). Over the past two decades, a plethora of programs and strategies has been developed to combine these sources of information to obtain reliable gene predictions. The 'intrinsic' evidence from sequence signals and statistical bias can be combined (using a variety of frameworks often related to hidden Markov models [59]), to produce gene predictions (6). These programs are often referred to as *ab initio* or *de novo* gene finders. They are the programs of choice in the absence of known transcript or protein sequences or phylogenetically related genomes. If related genome sequences are available, the intrinsic information can be combined with patterns of genomic sequence conservation using programs often referred to as comparative (or dual- or multi-genome) gene finders (5). With these programs, maximum resolution is achieved when the compared genomes are at a phylogenetic distance such that there is maximum separation between the conservation in coding and noncoding regions. To increase resolution, programs have been developed that use multiple informant genomes. The most sophisticated use an underlying phylogenetic tree to appropriately weight sequence conservation depending on evolutionary distance. If cDNA and EST sequences are available, these often take priority over other sources of information. The initial map of the transcript or protein sequences onto the genome, which can be obtained using a variety of tools, including sequence-similarity searches, is refined using more sophisticated 'splice alignment' algorithms, whose explicit splice-site models allow more precise alignment across gaps corresponding to introns (8). Alternatively, cDNA and protein information can be fed into an *ab initio* gene-finder algorithm to give information on the exons included in the prediction (7). Often, cDNA and protein evidence is only partial; in such cases, the initial reliable gene and transcript set may be extended with more hypothetical models derived from *ab initio* or comparative gene finders, or from the genome mapping of cDNA and protein sequences from other species. Pipelines have been derived that automate this multi-step process (9). More recently, programs have been developed that combine the output of many individual gene finders (10). The underlying assumption in these 'combiners' is that consensus across programs increases the likelihood of the predictions. Thus, predictions are weighted according to the particular features of the program producing them. The most general frameworks allow the integration of a great variety of types of predictions - not only gene predictions, but also predictions of individual sites and exons. Despite all the developments in computational gene finding, the most reliable and complete gene annotations are still obtained after the initial alignments of cDNA and proteins onto the genome sequence are inspected manually to establish the exon boundaries of genes and transcripts (11). This is the task carried out by the HAVANA team at the Sanger Institute. The initial manual annotation can be refined even further by subsequent experimental verification of those transcript models lacking sufficiently strong evidence, as in the GENCODE project (12). Examples of gene-prediction programs (with references and URLs) corresponding to each strategy outlined here are provided in Additional data file 1.

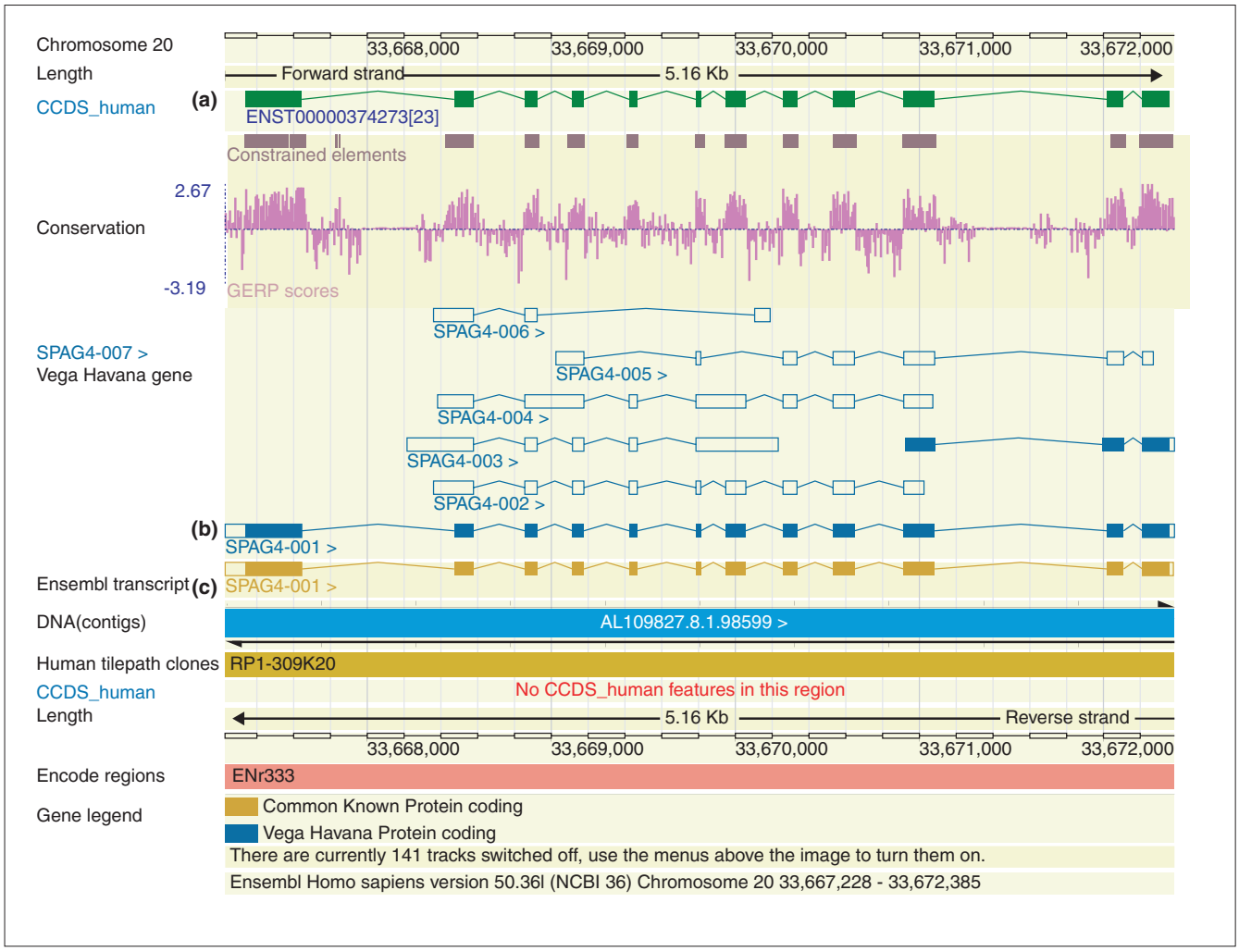

Figure 2
ENSEMBL browser. The ContigView page of the Ensembl browser representing the *SPAG4* gene locus on chromosome 20 within the Encode region ENr333. **(a)** The green transcript represents the CCDS coding region agreed on by the CCDS consortium. **(b)** The blue transcripts are the Vega transcripts, which are manually annotated by the HAVANA group and are a mixture of coding (solid blues) and noncoding (blue outline) transcripts. **(c)** Finally, the gold transcript represents the coding transcript on which the HAVANA and Ensembl annotations agree.

completeness. In addition, as transcript sequences (complete or partial cDNAs) are among the most reliable evidence used to annotate genes, we will also review a number of recent surveys of the transcriptional activity of the human genome. These, carried out using a variety of high-throughput technologies, have consistently reported a wealth of transcriptional activity in the human genome that had apparently not been captured through the large cDNA sequencing projects of the past two decades. It is now apparent that current gene and transcript annotation sets cover only a fraction of the total transcriptional output of the human genome.

Human reference gene sets

Since the publication of the draft human genome sequence in 2001 [6,7], a number of human gene reference sets have been created using either computational prediction or manual annotation or a mixture of the two methods. The Ensembl project was initially set up to warehouse and annotate the large amount of unfinished genomic data being produced as part of the public human genome project, as well as to provide browser capacity for both sequences and annotations (Figure 2). Ensembl has expanded and now generates automatic predictions for more than 35 species. The Ensembl gene build process is based on alignments of protein and cDNA sequences to produce a highly accurate gene set with a low rate of false positives [19].

Another genome browser supplying sequence and annotation data for a large number of genomes is the University of California, Santa Cruz (UCSC) genome browser database [20]. In April 2007, UCSC released an improved version of

their 'Known Gene Set' for the human genome and included putative noncoding RNAs as well as protein-coding genes. Each entry in this set requires the support of a GenBank entry and at least one other line of evidence, except for curated cDNAs, which require no other evidence.

Manual annotation still plays a significant part in annotating high-quality finished genomes. Currently, the National Center for Biotechnology Information (NCBI) reference sequences (RefSeq) collection provides a highly (manually) curated resource of multi-species transcripts, including plant, viral, vertebrate and invertebrate sequences [21,22]. These are, as their name indicates, transcript-oriented and usually rely on full-length cDNAs for reliable curation, although the dataset also contains predictions using expressed sequence tags (ESTs) and partial cDNAs aligned against genomic sequence using the Gnomon prediction program [23]. Manually reviewed RefSeq nucleotide sequences begin with the reference NM identifier whereas unreviewed predictions have the XM identifier. When a new genome is initially sequenced, researchers usually use the RefSeq data set to identify genes that are missing or identify genomic rearrangements within genes, as RefSeq is used internationally as a standard for genome annotation [21]. RefSeq is a very reliable, but also conservative, gene reference set. Other reference sets usually include RefSeq, but extend it substantially. For instance, the UCSC 'Known Genes' has 10% more protein-coding genes, approximately five times as many putative coding genes and twice as many splice variants as RefSeq.

A different approach to manual gene annotation is to annotate transcripts aligned to the genome and take the genomic sequences as the reference rather than the cDNAs. This is how the HAVANA group at the Wellcome Trust Sanger Institute produces its annotation on vertebrate sequence. Currently, only three vertebrate genomes - human, mouse and zebrafish - are being fully finished and sequenced to a quality that merits manual annotation [24]. The finished genomic sequence is analyzed using a modified Ensembl pipeline [25], and BLAST results of cDNAs/ESTs and proteins, along with various *ab initio* predictions, can be analyzed manually in the annotation browser tool Otterlace. The advantage of genomic annotation compared with cDNA annotation is that more alternative spliced variants can be predicted, as partial EST evidence and protein evidence can be used, whereas cDNA annotation is limited to availability of full-length transcripts. Moreover, genomic annotation produces a more comprehensive analysis of pseudogenes. One disadvantage, however, is that if a polymorphism occurs in the reference sequence, a coding transcript cannot be annotated, whereas cDNA annotation can select the major haplotypic form and is, therefore, not limited by a reference sequence.

In 2006, the groups mentioned above (NCBI (RefSeq), UCSC, the Wellcome Trust Sanger Institute (HAVANA) and Ensembl) identified a need to collaborate and produce a consensus gene set for the human reference genome as there was still no official agreement between the different databases on the human protein-coding genes. Referred to as the Consensus Coding Sequence Set (CCDS) [26], it currently contains only those coding transcripts that are equivalent in each database's gene build from start codon to stop codon. The latest human CCDS release (May 2008) contains 20,151 consensus coding sequences representing 17,052 genes. For the first time, this provides researchers with a consistent reliable gene set that has been derived independently from a combination of manual and automated annotation by three groups (Ensembl, NCBI and HAVANA) and quality checked at the UCSC. The protein-coding genes that differ between the gene sets of the different groups and cannot be merged automatically will be re-examined manually and either rejected or added to the consensus set if they get a unanimous vote from the groups at NCBI, UCSC and HAVANA.

Complementary to the CCDS project is the GENCODE project [27]. The GENCODE consortium [28] was initially formed to identify and map all protein-coding genes within the regions selected in the framework of the ENCODE project [29,30], representing 1% of human genome sequence. This was achieved by a combination of initial manual annotation by HAVANA, computational predictions and experimental validation, and the consequent refinement of the annotation on the basis of these experimental results. The project has been funded in 2008 to annotate the whole reference human genome sequence and experimentally verify a number of putative loci. The scaled-up annotation includes identification of pseudogenes and noncoding loci supported by transcript evidence. The initial manual annotation is compared with automated predictions to highlight inconsistencies based on comparative analysis or new transcript data. It is expected that, upon completion in 2011, this gene set will become the standard human gene reference set.

Assessing the annotation

The issue obviously arises of the reliability of the reference sets. Usually, the experimentally verified manual annotations, such as those produced by GENCODE, are considered the most exhaustive and reliable reference human gene sets. Based on 'bona fide' cDNA sequences, the annotated gene models are, in these cases, generally correct - although issues still remain because, on occasion, the same cDNA sequence can be mapped into the human genome through alternative exonic structures. Completeness is more difficult to assess, because it is unclear how representative of the complete human transcriptome the current set of cDNA sequences is.

To assess the completeness of GENCODE, the EGASP community experiment was organized in 2005 [31]. In this experiment a number of computational predictions were evaluated against the GENCODE annotation. Then, a subset

of high-confidence computational predictions that were not present in the annotation was tested by reverse transcription-polymerase chain reaction (RT-PCR) on a panel of human tissues. Only a handful of these predictions could be verified, strongly suggesting completeness of the GENCODE annotation (with respect to computational predictions of protein-coding genes). A second goal of EGASP was to assess to what extent purely computational methods can reproduce the slow and expensive manual annotations. In this regard, EGASP results indicated that although computational methods are quite accurate in identifying protein-coding exons with an overall accuracy of more than 80% (in terms of both the fraction of real exons correctly identified and the fraction of predicted exons that are real), finding the complete transcript structure is more challenging, with the most accurate methods correctly predicting only about 60% of the annotated protein-coding transcripts. This indicates that computational methods cannot yet totally replace human expertise in gene annotation.

After mapping a cDNA to the genome, the protein-coding status of the transcript needs to be assessed, and the boundaries of the eventual coding regions precisely delimited - so that it is possible to identify the correct amino acid sequence of the protein, from which most of the biology of the transcript will be inferred. As direct evidence of the existence of the protein is generally absent, the criterion often used to annotate a transcript as protein-coding is the existence of an open reading frame (ORF). However, this criterion has recently been put in question by a number of methods developed to assess the quality of protein-coding gene annotations. These are based on the principle that gene models that conflict with our current knowledge about functional protein-coding genes are incorrect. Thus, the rationale of the method of Clamp *et al*. [9] is that functional protein-coding genes are subject to purifying selection, and are therefore expected to show evolutionary conservation. The authors used two types of measures for the assessment of evolutionary conservation of predicted human genes: reading frame conservation (RFC; based on the observation that indels do not affect significantly the size of functional proteins) and codon substitution frequency (CSF; based on the observation that the patterns of nucleotide substitution in functional protein-coding genes is different from that observed on random DNA). In their analysis of a number of human gene reference sets, Clamp *et al*. [9] identified around 1,200 human 'orphans': ORFs that lack homology with known genes. Both RFC and CSF analysis revealed that the behavior of many of these human orphans is essentially indistinguishable from that of matched random controls, and is very different from that of non-orphan protein-coding genes. From these results, the authors concluded that, overall, about 15% of the entries in the gene catalogs investigated are not valid protein-coding genes.

While the quality-control method of Clamp *et al*. [9] can distinguish protein-coding genes from non-coding sequences, it is less suitable for identifying gene predictions that are only partially correct. Indeed, if an annotated gene misses one or more exons, or a fraction of one exon, it may still display the expected evolutionary characteristics of protein-coding genes. To find errors in the annotated protein-coding genes, the MisPred approach [32-35] uses several criteria that hold for different subsets of correctly folded, correctly localized, functionally competent protein molecules. Hypothetical proteins that violate any of these rules are judged to be nonfunctional and the corresponding coding regions to be misidentified. For example, one of the quality-control tools of this approach is based on the observation that the number of residues in closely related members of a globular protein domain family usually falls within a relatively narrow range. Accordingly, proteins containing domains that consist of significantly larger or smaller numbers of residues than closely related members of the same family may be suspected to be nonviable and the corresponding genes to be mispredicted. Several quality-control tools in MisPred address the issue of whether the predicted protein is able to reach the cellular compartment where it could be properly folded, stable and functional. The rationale of these tools is that mislocalized proteins are usually misfolded, unstable and nonfunctional. For example, predicted proteins that contain extracellular domains but lack sequence signals that could direct these domains to the extracellular space are likely to be misfolded and nonfunctional. Analyses of predicted human sequences with MisPred tools revealed that 2.3% of Ensembl entries (v41) and 3.4% of proteins predicted by the NCBI's Gnomon pipeline are likely to be mislocalized and/or misfolded as they lack appropriate sequence signals or they contain domains that deviate from normal size [32].

In a similar spirit, the EPipe pipeline [36] of the BioSapiens consortium incorporates a variety of tools to assess the structural and functional properties of hypothetical proteins. Analysis of the GENCODE peptides with these tools revealed that many of the potential alternative gene products have markedly different structure and function from their constitutively spliced counterparts. For the vast majority of these alternatively spliced forms, there is little evidence that they have a role as functional proteins, and many splice variants encode abnormal proteins that are mislocalized and/or misfolded [33].

Alternative splicing and protein complexity

Alternative splicing is common in mammalian genomes, and it has been suggested to be a means of increasing protein complexity from a limited number of genes. Therefore, any complete gene set should include annotation of the protein-coding variants. Detailed cDNA mapping into the genome, as in the GENCODE annotation, reveals that alternative splicing is widespread, affecting more than 86% of multi-exon gene

loci [27] with an average of 5.7 transcript variants per locus. While this is a proportion of alternative-splicing events much larger than that in other human reference gene sets, the use of novel high-throughput methods that concatenate and sequence the 5′ tags of transcripts (cap analysis gene expression (CAGE) [3]) or sequence paired 5′ and 3′ cDNA ends (5′ paired-end ditags (5′PETs) [37]) has revealed that traditional methods based on cDNA clone sequencing were not fully surveying the complexity of mammalian transcriptomes. Similarly, the (re)analysis of gene-trapping sequences has unveiled thousands of novel transcripts [1]. Tiling oligonucleotide arrays that monitor the expression of the non-repeated fraction of the genome have consistently identified many more transcribed fragments than previously anticipated [38,39]. The combination of all these experimental approaches in the frame of the ENCODE project [30] surprisingly showed that more than 90% of the genome is transcribed as primary RNA [29], with at least 15% being incorporated into processed transcripts. Many such novel transcripts map to protein-coding loci, as revealed by experiments in which RACE (rapid amplification of 3′ ends) products originating in these loci were hybridized onto tiling arrays. When applied to the ENCODE regions, these experiments yielded as many novel as annotated exons [13]. Often these exons corresponded to tissue-specific 5′ distal transcription start sites (TSSs) [13]. These distal TSSs map hundreds of kilobases upstream of the currently annotated TSS and often overlap with a 5′-positioned gene, suggesting extensive overlap between protein-coding loci (Additional data file 2). Next-generation sequencing will further enhance our capacity to sequence the transcriptome of the cell (RNAseq). Indeed, preliminary results demonstrate that RNAseq can detect 25% more genes than microarrays can and that a third of the sequences emanate from unannotated regions [40-45].

Interestingly, only a small fraction of these novel transcripts seem to have protein-coding capacity - often through transcription-induced chimeras that fuse two different ORFs that may be encoded by genes far apart in the genome [13,46,47]. Instead, the majority correspond to 'novel' noncoding RNA classes, such as transcribed pseudogenes [48-50], antisense transcripts [51-53] and structured RNAs [54,55], that might regulate transcription and/or translation. For example, Watanabe *et al.* [56] recently described precursor transcripts of small interfering RNAs (siRNAs) that are derived from transcribed pseudogenes. Other yet-unannotated RNAs appear to be processed into short RNAs, some of which, like the 'promoter-associated sRNAs' (PASRs) and 'termini-associated sRNAs' (TASRs), are coupled to the expression state of protein-coding genes [2,57]. Finally, it was postulated that some of these novel transcripts might be the outcome of interchromosomal transcript chimerism: that is, chimeric transcripts resulting from the proximity of active genes in so-called transcription factories [58].

In summary, recent technological developments and large-scale whole-genome analyses have shown that mammalian transcriptomes are composed of a swarming mass of different overlapping transcripts, sometimes originating from both strands, and suggest that only a small fraction of the transcriptional complexity has been discovered. Little evidence exists, however, that the majority of this transcript complexity leads to protein complexity. In fact, all evidence suggests otherwise - that the human protein-coding gene set is near consolidation. Thus, the 5.7 average transcripts per coding locus annotated in GENCODE translates to only 1.7 proteins per locus (because a large fraction of transcript variation corresponds to noncoding transcripts or accumulates in the untranslated regions of coding transcripts) [27]. Moreover, if the GENCODE proteins flagged as problematic by the protein-assessment methods discussed above are ignored, there are barely 1.3 annotated proteins per locus - a somewhat unexpected return to one of the founding axioms of molecular biology: Beadle and Tatum's 'one gene one protein' principle. The discrepancy between a complex, variable and largely unexplored population of RNA molecules and a relatively small, stable, and well defined population of proteins constitutes one of the challenges that molecular biology needs to address to fully elucidate cellular function.

Additional data files

Additional data file 1 contains a table listing software used for gene prediction and annotation. The programs are categorized according to the sources of information utilized and each listing includes a literature reference and URL where the software may be obtained. This list is meant to be representative rather than comprehensive. Additional data file 2 contains a figure showing novel transcripts discovered through a combination of directed RACE and hybridization onto tiling arrays.

Acknowledgements

This work was carried out as part of the BioSapiens project. The BioSapiens project is funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2003-503265. AR, SA and RG also acknowledge support from grants U01HG003150 and U01HG003147 from the National Human Genome Research Institute, NIH. RG acknowledges support from grant BIO2006-03380 from the Spanish Ministry of Education and Science. AR and SA acknowledge support from the EU AnEUploidy project, and the NCCR Frontiers in Genetics. LP thanks the Hungarian National Office for Research and Technology for partial support under grant no. RET14/2005. JA's work is supported by the Wellcome Trust.

References

1. Roma G, Cobellis G, Claudiani P, Maione F, Cruz P, Tripoli G, Sardiello M, Peluso I, Stupka E: **A novel view of the transcriptome revealed from gene trapping in mouse embryonic stem cells.** *Genome Res* 2007, **17:**1051-1060.
2. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E,

Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316:**1484-1488.

3. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, *et al.*: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309:**1559-1563.
4. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: **Identification of novel genes coding for small expressed RNAs.** *Science* 2001, **294:**853-858.
5. Pheasant M, Mattick JS: **Raising the estimate of functional human sequences.** *Genome Res* 2007, **17:**1245-1253.
6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
7. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
8. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:**931-945.
9. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin M, Kellis M, Lindblad-Toh K, Lander E: **Distinguishing protein-coding and noncoding genes in the human genome.** *Proc Natl Acad Sci USA* 2007, **104:** 19428-19433.
10. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M: **Large-scale copy number polymorphism in the human genome.** *Science* 2004, **305:**525-528.
11. Iafrate A, Feuk L, Rivera M, Listewnik M, Donahoe P, Qi Y, Scherer S, Lee C: **Detection of large-scale variation in the human genome.** *Nat Genet* 2004, **36:**949-951.
12. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, *et al.*: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453:**56-64.
13. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, Dike S, Wyss C, Henrichsen CN, Holroyd N, Dickson MC, Taylor R, Hance Z, Foissac S, Myers RM, Rogers J, Hubbard T, Harrow J, Guigó R, Gingeras TR, Antonarakis SE, Reymond A: **Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions.** *Genome Res* 2007, **17:**746-759.
14. Rozowsky JS, Newburger D, Sayward F, Wu J, Jordan G, Korbel JO, Nagalakshmi U, Yang J, Zheng D, Guigó R, Gingeras TR, Weissman S, Miller P, Snyder M, Gerstein MB: **The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci.** *Genome Res* 2007, **17:**732-745.
15. Fickett JW: **Recognition of protein coding regions in DNA sequences.** *Nucleic Acids Res* 1982, **10:**5303-5318.
16. Brent MR, Guigó R: **Recent advances in gene structure prediction.** *Curr Opin Struct Biol* 2004, **14:**264-272.
17. Mathé C, Sagot M-F, Schiex T, Rouzé P: **Current methods of gene prediction, their strengths and weaknesses.** *Nucleic Acids Res* 2002, **30:**4103-4117.
18. Jones S: **Prediction of genomic functional elements**. *Annu Rev Genomics Hum Genet* 2006, **7:**315-338.
19. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Gräf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kähäri A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, *et al.*: **Ensembl 2008.** *Nucleic Acids Res* 2008, **36(Database issue):** D707-D714.
20. Karolchik D, Hinrichs AS, Kent WJ: **The UCSC Genome Browser.** In *Curr Protocols Bioinf*, Chapter 1:Unit 1.4.
21. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2008, **36(Database issue):**D61-D65.
22. Maglott DR, Katz KS, Sicotte H, Pruitt KD: **NCBI's LocusLink and RefSeq.** *Nucleic Acids Res* 2000, **28:**126-128.
23. **Gnomon** [http://www.ncbi.nlm.nih.gov/projects/genome/guide/gnomon.shtml]
24. Wilming LG, Gilbert JGR, Howe K, Trevanion S, Hubbard T, Harrow JL: **The vertebrate genome annotation (Vega) database.** *Nucleic Acids Res* 2008, **36(Database issue):**D753-D760.
25. Searle S, Gilbert J, Iyer V, Clamp M: **The Otter annotation system.** *Genome Res* 2004, **14:**963-970.
26. **CCDS** [http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi]
27. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigó R: **GENCODE: producing a reference annotation for ENCODE.** *Genome Biol* 2006, **7(Suppl 1):**S4.1-9.
28. **GENCODE** [http://genome.imim.es/gencode]
29. ENCODE Project Consortium: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799-816.
30. ENCODE Project Consortium: **The ENCODE (ENCyclopedia Of DNA Elements) project.** *Science* 2004, **306:**636-640.
31. Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, Castelo R, Eyras E, Ucla C, Gingeras TR, Harrow J, Hubbard T, Lewis SE, Reese MG: **EGASP: the human ENCODE Genome Annotation Assessment Project.** *Genome Biol* 2006, **7(Suppl 1):**S2.1-31.
32. Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Bányai L, Patthy L: **Quality control of gene predictions.** In *Modern Genome Annotation.* Edited by The Biosapiens Network, Frishman D, Valencia A. Springer: 2008.
33. Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason PI, Albrecht M, Hegyi H, Giorgetti A, Raimondo D, Lagarde J, Laskowski RA, López G, Sadowski MI, Watson JD, Fariselli P, Rossi I, Nagy A, Kai W, Størling Z, Orsini M, Assenov Y, Blankenburg H, Huthmacher C, Ramírez F, Schlicker A, Denoeud F, Jones P, Kerrien S, *et al.*: **The implications of alternative splicing in the ENCODE protein complement.** *Proc Natl Acad Sci USA* 2007, **104:**5495-5500.
34. Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Banyai L, Patthy L: **Identification and correction of abnormal, incomplete and mispredicted proteins in public databases.** *BMC Bioinf* 2008, **9:**353.
35. **MisPred** [http://mispred.enzim.hu/]
36. **EPipe 1.0** [http://www.cbs.dtu.dk/services/EPipe-1.0]
37. Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, Liu ET, Ruan Y: **Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation.** *Nat Methods* 2005, **2:**105-111.
38. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SPA, Gingeras TR: **Large-scale transcriptional activity in chromosomes 21 and 22.** *Science* 2002, **296:**916-919.
39. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR: **Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays.** *Genome Res* 2005, **15:**987-997.
40. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science* 2008, **320:**1344-1349.
41. Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621-628.
42. Wilhelm B, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett C, Rogers J, Bahler J: **Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.** *Nature* 2008, **453:**1239-1243.
43. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.** *Science* 2008, **321:**956-960.

44. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*.** *Cell* 2008, **133:**523-536.
45. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods* 2008, **5:**613-619.
46. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, Sorek R: **Transcription-mediated gene fusion in the human genome.** *Genome Res* 2006, **16:**30-36.
47. Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, Antonarakis SE, Guigó R: **Tandem chimerism as a means to increase protein complexity in the human genome.** *Genome Res* 2006, **16:**37-44.
48. Vinckenbosch N, Dupanloup I, Kaessmann H: **Evolutionary fate of retroposed gene copies in the human genome.** *Proc Natl Acad Sci USA* 2006, **103:**3220-3225.
49. Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M, Ruan Y, Wei CL, Gingeras TR, Guigó R, Harrow J, Gerstein MB: **Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution.** *Genome Res* 2007, **17:**839-851.
50. Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H: **Emergence of young human genes after a burst of retroposition in primates.** *PLoS Biol* 2005, **3:**e357.
51. Werner A, Schmutzler G, Carlile M, Miles C, Peters H: **Expression profiling of antisense transcripts on DNA arrays.** *Physiol Genomics* 2007, **28:**294-300.
52. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.** *Genome Res* 2004, **14:**331-342.
53. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engström PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B, *et al.*: **Antisense transcription in the mammalian transcriptome.** *Science* 2005, **309:**1564-1566.
54. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde J, Drenkow J, Kapranov P, Gingeras TR, Guigó R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler PF: **Structured RNAs in the ENCODE selected regions of the human genome.** *Genome Res* 2007, **17:**852-864.
55. Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF: **Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome.** *Nat Biotechnol* 2005, **23:**1383-1390.
56. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, Surani MA, Sakaki Y, Sasaki H: **Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes.** *Nature* 2008, **453:**539-543.
57. Borel C, Gagnebin M, Gehrig C, Kriventseva EV, Zdobnov EM, Antonarakis SE: **Mapping of small RNAs in the human ENCODE regions.** *Am J Hum Genet* 2008, **82:**971-981.
58. Unneberg P, Claverie J-M: **Tentative mapping of transcription-induced interchromosomal interaction using chimeric EST and mRNA data.** *PLoS ONE* 2007, **2:**e254.
59. Eddy S: **What is a hidden Markov model?** *Nat Biotechnol* 2004, **22:**1315-1316.
60. Djebali S, Kapranov P, Foissac S, Lagarde J, Reymond A, Ucla C, Wyss C, Drenkow J, Dumais E, Murray RR, Lin C, Szeto D, Denoeud F, Calvo M, Frankish A, Harrow J, Makrythanasis P, Vidal M, Salehi-Ashtiani K, Antonarakis SE, Gingeras TR, Guigó R: **Efficient targeted transcript discovery via array-based normalization of RACE libraries.** *Nat Methods* 2008, **5:**629-635.