Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome

Cédric Howald,[1,2,5] Andrea Tanzer,[3,5] Jacqueline Chrast,[1] Felix Kokocinski,[4] Thomas Derrien,[3] Nathalie Walters,[1] Jose M. Gonzalez,[4] Adam Frankish,[4] Bronwen L. Aken,[4] Thibaut Hourlier,[4] Jan-Hinnerk Vogel,[4] Simon White,[4] Stephen Searle,[4] Jennifer Harrow,[4] Tim J. Hubbard,[4] Roderic Guigó,[3,6] and Alexandre Reymond[1,6]

[1]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland; [2]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; [3]Centre de Regulacio Genomica, Grup de Recerca en Informatica Biomedica, E-08003 Barcelona, Spain; [4]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom

Within the ENCODE Consortium, GENCODE aimed to accurately annotate all protein-coding genes, pseudogenes, and noncoding transcribed loci in the human genome through manual curation and computational methods. Annotated transcript structures were assessed, and less well-supported loci were systematically, experimentally validated. Predicted exon–exon junctions were evaluated by RT-PCR amplification followed by highly multiplexed sequencing readout, a method we called RT-PCR-seq. Seventy-nine percent of all assessed junctions are confirmed by this evaluation procedure, demonstrating the high quality of the GENCODE gene set. RT-PCR-seq was also efficient to screen gene models predicted using the Human Body Map (HBM) RNA-seq data. We validated 73% of these predictions, thus confirming 1168 novel genes, mostly noncoding, which will further complement the GENCODE annotation. Our novel experimental validation pipeline is extremely sensitive, far more than unbiased transcriptome profiling through RNA sequencing, which is becoming the norm. For example, exon–exon junctions unique to GENCODE annotated transcripts are five times more likely to be corroborated with our targeted approach than with extensive large human transcriptome profiling. Data sets such as the HBM and ENCODE RNA-seq data fail sampling of low-expressed transcripts. Our RT-PCR-seq targeted approach also has the advantage of identifying novel exons of known genes, as we discovered unannotated exons in ~11% of assessed introns. We thus estimate that at least 18% of known loci have yet-unannotated exons. Our work demonstrates that the cataloging of all of the genic elements encoded in the human genome will necessitate a coordinated effort between unbiased and targeted approaches, like RNA-seq and RT-PCR-seq.

[Supplemental material is available for this article.]

The ENCODE (Encyclopedia of DNA Elements) Project aims to identify all functional elements encoded in the human genome and provide an annotated reference to the scientific community (The ENCODE Project Consortium 2004; Myers et al. 2011). These functional elements include genes, pseudogenes, transcripts isoforms, transcription start sites, and chromatin annotation, as well as long-range chromosomal interactions and methylation status (The ENCODE Project Consortium 2007, 2012).

Mammalian genomes are pervasively transcribed (i.e., the majority of their bases are incorporated in at least one primary transcript), and noncoding RNAs (ncRNAs) constitute the lion's share of this ubiquitous transcription (Kapranov et al. 2002; Okazaki et al. 2002; The ENCODE Project Consortium 2007; Brawand et al. 2011). Some transcripts even shatter the conventional gene structure by joining exons of different well-established coding loci (Parra et al. 2006; Denoeud et al. 2007; Djebali et al. 2012a). The biological importance of ncRNAs is the focus of heated debate (van Bakel et al. 2010; Clark et al. 2011) because they are generally lineage-specific and expressed at lower levels and more specifically than coding transcripts (Kowalczyk et al. 2012). Archetypes of their molecular function are, however, emerging (examples in Ulitsky et al. 2011; Cartault et al. 2012; for review, see Wang and Chang 2011). This unexpected complexity of transcriptomes and breadth of transcripts should be comprehensively annotated within a reference gene set upon which all other ENCODE Consortium analyses are built and that similar projects depending on an accurate description of gene elements in the human genome could use. GENCODE is the subproject of ENCODE whose goal it is to annotate all gene features accurately. It primarily relies on manual curation with moderate implementation of automated algorithms, merging manual HAVANA (The Human and Vertebrate Analysis and Annotation project) (Wilming et al. 2008) and automatic Ensembl annotation (Flicek et al. 2011). Gene and transcripts models are classified by two attributes: status and biotype (summarized in Methods). The "status" specifies the nature of the evidence exploited to define a model, thus indicating the level of confidence assigned to it, whereas the "biotype" denotes the biological class of genes and transcripts. Three different statuses are used—known, novel, and putative, while the number of biotypes is not restricted to accommodate new classes of genes that might be discovered (e.g., the recently characterized lincRNA biotype). The continuously evolving GENCODE

[5]These authors contributed equally to this work.
[6]Corresponding authors
E-mail roderic.guigo@crg.cat
E-mail alexandre.reymond@unil.ch

Article and supplemental material are at http://www.genome.org/cgi/doi/10.1101/gr.134478.111. Freely available online through the *Genome Research* Open Access option.

22:1698–1710 © 2012, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

gene set now contains 51,096 genes and 165,067 transcripts (GENCODE freeze version 8). This set, as well as the methods used for the annotation and the different statuses and biotypes, are described in a specific GENCODE companion manuscript (Harrow et al. 2012), while long noncoding (lnc), small RNAs, and pseudogenes annotated by the ENCODE project are detailed in Derrien et al. (2012), Djebali et al. (2012b), and Pei et al. (2012).

The quality, complexity, and depth reached by the GENCODE annotation were already demonstrated during the ENCODE pilot phase, which targeted only 1% of the human genome (The ENCODE Project Consortium 2004, 2007). While 84% of RefSeq and 76% of Ensembl exons exactly overlapped GENCODE exons, only 40% of GENCODE exons were contained within RefSeq or Ensembl (Harrow et al. 2006). Additionally, gene models predicted by automated algorithms (e.g., GeneID, SGP2, Genescan, Twinscan) (Burge and Karlin 1997; Parra et al. 2000, 2003; Flicek et al. 2003), which lay outside of the GENCODE annotation, could rarely be experimentally validated (3.4% success rate) (Guigo et al. 2003).

To assess the quality of the reference annotation for the ENCODE project reached by GENCODE freeze version 8, we systematically, experimentally evaluated the exon–intron structure of all transcripts rated as novel or putative with a novel procedure that combines traditional RT-PCR amplification with a highly multiplexed short read sequencing readout. In the present study, we describe this new approach and compare it with transcriptome profiling by RNA sequencing.

Results

RT-PCR-seq to validate genome annotations

During the pilot phase of GENCODE, gene models were experimentally validated by RT-PCR amplification, followed by gel separation and Sanger sequencing. The emergence of next-generation sequencers gave us the opportunity to replace some of these labor-intensive steps, abbreviate the process, and increase its accuracy. We implemented a novel experimental validation workflow called "RT-PCR-seq" that can be adapted to multiple scenarios in which targeted validation of splice sites is required. The specific implementation of the workflow in the framework of the GENCODE project included the following sequential steps (Fig. 1A):

(1) Lower-confidence GENCODE gene models flagged as "novel"—built using species-specific cDNAs or supported by cDNAs or proteins from another species—or "putative"—modeled using species-specific ESTs or ESTs from other species—are selected for experimental assessment.
(2) Pairs of primers are designed over suitable exon–exon junction(s) (Fig. 1B) (see Methods).
(3) These primer pairs are used to PCR-amplify individually the targeted exon–exon junctions from eight human poly(A)$^+$ cDNAs (brain, heart, kidney, liver, lung, spleen, skeletal muscle, and testis) (Fig. 1C).
(4) Aliquots of multiple amplimers (up to 4700) are subsequently pooled by tissue.
(5) Nonsonicated column-purified pools of amplimers are separated from primers on an agarose gel and used to generate sequencing libraries (see Methods).
(6) They are sequenced using a high-throughput sequencing platform (see Methods).
(7) Sequence tags are bioinformatically gauged, with criteria intimately related to the primer design protocol (see Methods and below) (Fig. 1C).
(8) Finally, validated data are submitted for display on the UCSC Genome Browser on the GENCODE track.

We first compared the novel RT-PCR-seq method with the traditional pipeline (PCR amplification, gel purification, and Sanger sequencing; described in Guigo et al. [2003] and used during the ENCODE pilot phase, Harrow et al. [2006]). A batch of 648 exon–exon junctions from known and novel GENCODE transcripts was experimentally assessed by both methods. We validated twice as many junctions with the new pipeline (308 [48%] compared with 156 [24%]). Conventional Sanger sequencing could only corroborate seven junctions not verified by RT-PCR-seq. These results indicated that the new procedure we established is not only highly scalable but also far more sensitive than classical approaches. Its success rate was further improved once we took advantage of the increase in read length provided by new sequencing chemistries (this trial was performed with 35-bp sequence reads). We raised it again by increasing the minimum amplicon size to 150 nt and strengthening self-priming parameters used to design the primer pairs (see below and Methods) reaching validation rates of 92% and 79% for known coding ($n = 158$) and noncoding transcripts ($n = 122$), respectively.

Secondly, we designed primers and selected for experimental validation by RT-PCR-seq a set of 10,162 different splice sites of 6831 "novel" or "putative" GENCODE genes representing 9213 distinct transcripts, as well as 486 ENCODE-predicted models (e.g., in V Gotea, H Petrykowska, L Elnitski, in prep.). Targeted splice sites can be divided into two classes: (1) exon–exon junctions where one primer could be placed within 75 nt of the junction ("Multi-span") (Fig. 1B), resulting in about half of the sequencing reads necessarily covering the junction (Fig. 2A); (2) junctions where this was unfeasible ("Multi") (Fig. 1B) and in which sequencing reads will generally not reach the splice site. We identified, however, a high number of sequencing reads crossing the targeted splice even for this category of amplimers (Fig. 2A). A third set of models is formed by the monoexonic transcripts ("Mono," all belonging to the set of ENCODE-predicted models) (Fig. 1B). Because models belonging to this category are sensitive to genomic DNA contaminations, they were assessed by amplification of cDNA in which a dNTP analog was incorporated as described in Washietl et al. (2007) (see Methods). Models belonging to each of these three categories were considered experimentally validated with different criteria as described in Methods and summarized in Figure 1C. We performed RT-PCR-seq for these 10,648 target loci and produced a total of 1,845,687,068 reads (Supplemental Table S1). Seventy-eight percent of them passed the quality threshold (mean *phred* quality score ≥ 23) and were mappable on the genome and/or the GENCODE transcriptome using Bowtie (Langmead et al. 2009). The overall validation rate across all tissues is extremely high, reaching between 73% and 87% for each biotype/status combination tested (Fig. 2B; Supplemental Tables S2, S3). Examples of validated "Multi-span" and monoexonic transcript models are presented in Supplemental Figure S1A,B. The transcriptome of testis showed the highest complexity (highest percentage of validated transcript models at 55%) (Fig. 2D) in accordance with previous reports and consolidating the view that chromatin is more relaxed in this tissue, leading to higher transcriptional activity (Denoeud et al. 2007). Each biotype/status combination is validated at comparable rates in the different investigated tissues with the exception of putative processed transcripts. Processed transcripts (putative and novel) are mainly identified in testis and are significantly more tissue-specific than other biotypes ($P = 1.52 \times 10^{-83}$, Fisher exact test)

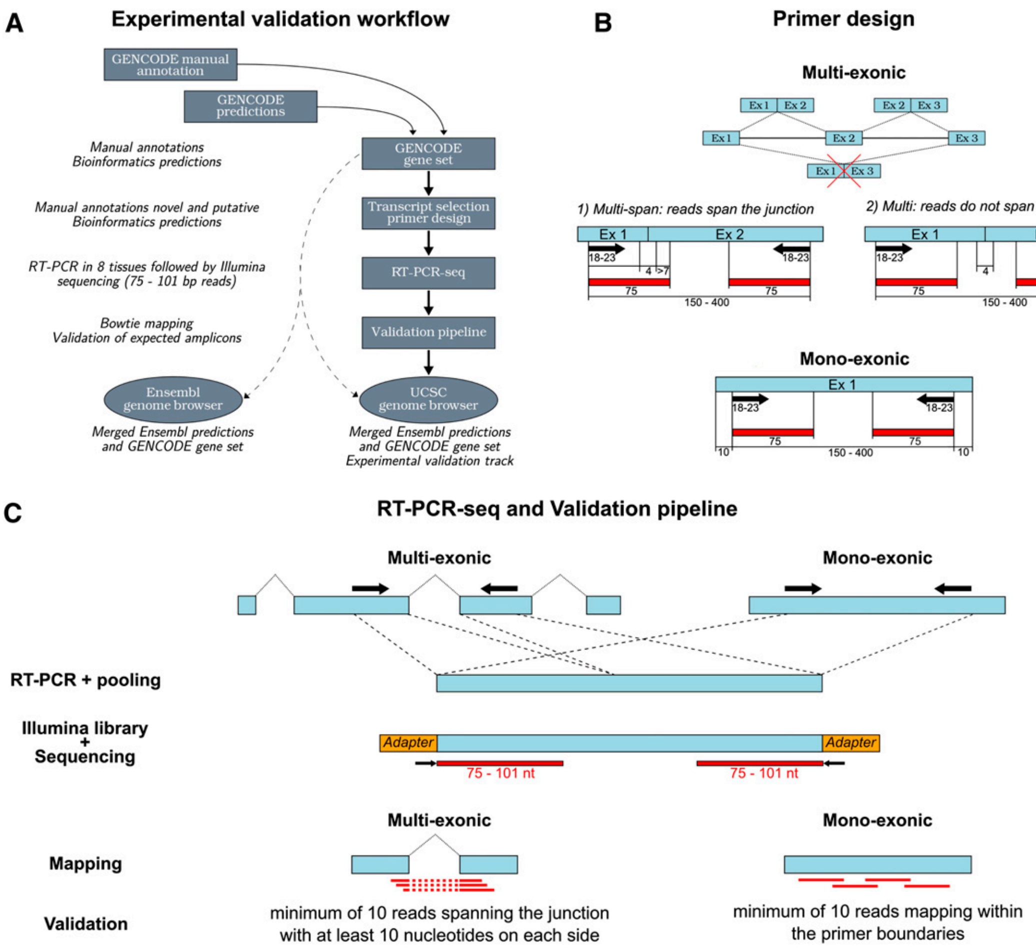
A
Experimental validation workflow
GENCODE manual annotation
GENCODE predictions
GENCODE gene set
Manual annotations
Bioinformatics predictions
Transcript selection primer design
Manual annotations novel and putative
Bioinformatics predictions
RT-PCR-seq
RT-PCR in 8 tissues followed by Illumina sequencing (75 - 101 bp reads)
Validation pipeline
Bowtie mapping
Validation of expected amplicons
Ensembl genome browser
UCSC genome browser
Merged Ensembl predictions and GENCODE gene set
Merged Ensembl predictions and GENCODE gene set
Experimental validation track
B
Primer design
Multi-exonic
Ex 1
Ex 2
Ex 3
1) Multi-span: reads span the junction
2) Multi: reads do not span the junction
18-23
4
>7
75
150 - 400
Mono-exonic
10
C
RT-PCR-seq and Validation pipeline
Multi-exonic
Mono-exonic
RT-PCR + pooling
Illumina library + Sequencing
Adapter
75 - 101 nt
Mapping
Validation
minimum of 10 reads spanning the junction with at least 10 nucleotides on each side
minimum of 10 reads mapping within the primer boundaries

Figure 1. RT-PCR-seq workflow. (*A*) Schematic workflow of the experimental validation of GENCODE transcript models with RT-PCR-seq. (*B*) Position of primers (black arrow) designed to validate gene models and corresponding sequencing reads (red rectangles) on targeted exons (blue rectangles); (Ex) exon. We experimentally assessed three categories of gene models: (1) spliced models in which one primer could be placed within 75 nt of a junction and that will result in about half of the sequencing reads covering the junction (''Multi-span''); (2) spliced models in which this was unfeasible (''Multi''); and (3) monoexonic genes (''Mono''). Numbers *below* exons, primers, sequencing reads, and amplimers indicate their respective sizes. Exon–exon junctions skipping an annotated cassette exon were not considered (see example barred with a red cross). (*C*) Schematic representation of the criteria used to experimentally validate GENCODE gene models (see Methods for details and panel *B* for color code).

(Fig. 2C). This is consistent with the hypothesis that a large fraction of GENCODE putative processed transcript loci correspond to long noncoding RNAs genes (lncRNAs), which are expressed at lower levels and more specifically than other genes (Derrien et al. 2012). Our ability to validate a large fraction of less well-supported GENCODE gene models further emphasizes the extremely high quality reached by the GENCODE gene set originating from the manual annotation involved.

RT-PCR-seq to identify novel transcript isoforms

We may have failed to validate some transcript models because they are expressed in a limited number of tissues, which were not targeted or because we only used a single condition for PCR amplification. Alternatively, the targeted splice-site prediction might be slightly shifted or wrongly annotated in regard to its real nature and position. Such a situation will impede validation by the RT-PCR-seq procedure because we are mapping reads with Bowtie (Langmead et al. 2009) against the expected amplicons without allowing any insertions or deletions. We better characterized unsubstantiated exon–exon junctions and found that in 146 cases (1.5% of assessed junctions; 6.9% of nonvalidated splice sites), the number of sequencing reads spanning the assessed exon–exon junction is simply below our validation threshold of 10 such reads (between one and nine reads; see Methods). Alternatively, we can surmise that in some cases, we are coamplifying multiple transcript isoforms with a single primer pair. In such circumstances, one or more of the isoforms might be outcompeted during amplification. To investigate how many unsubstantiated exon–exon junctions exhibit nontargeted splice sites in the amplified amplimers, we remapped all unmappable reads within the primer boundaries with GEM (http://sourceforge.net/apps/mediawiki/gemlibrary/index.php?title=The_GEM_library), a sequence aligner that allows, contrary to Tophat (Trapnell et al. 2009), split-mapping without prior prediction of genomic islands corresponding to exons. This is crucial because in RT-PCR-seq, one of the primers is intentionally designed very close to the assessed junction (Fig. 1B) (Methods). To quantify the fraction of PCR amplifications that allowed identification of novel isoforms, we implemented a specific scoring method (see Methods). We found 1119 (11%, n = 10,162) unannotated transcript models including 644 new internal exons

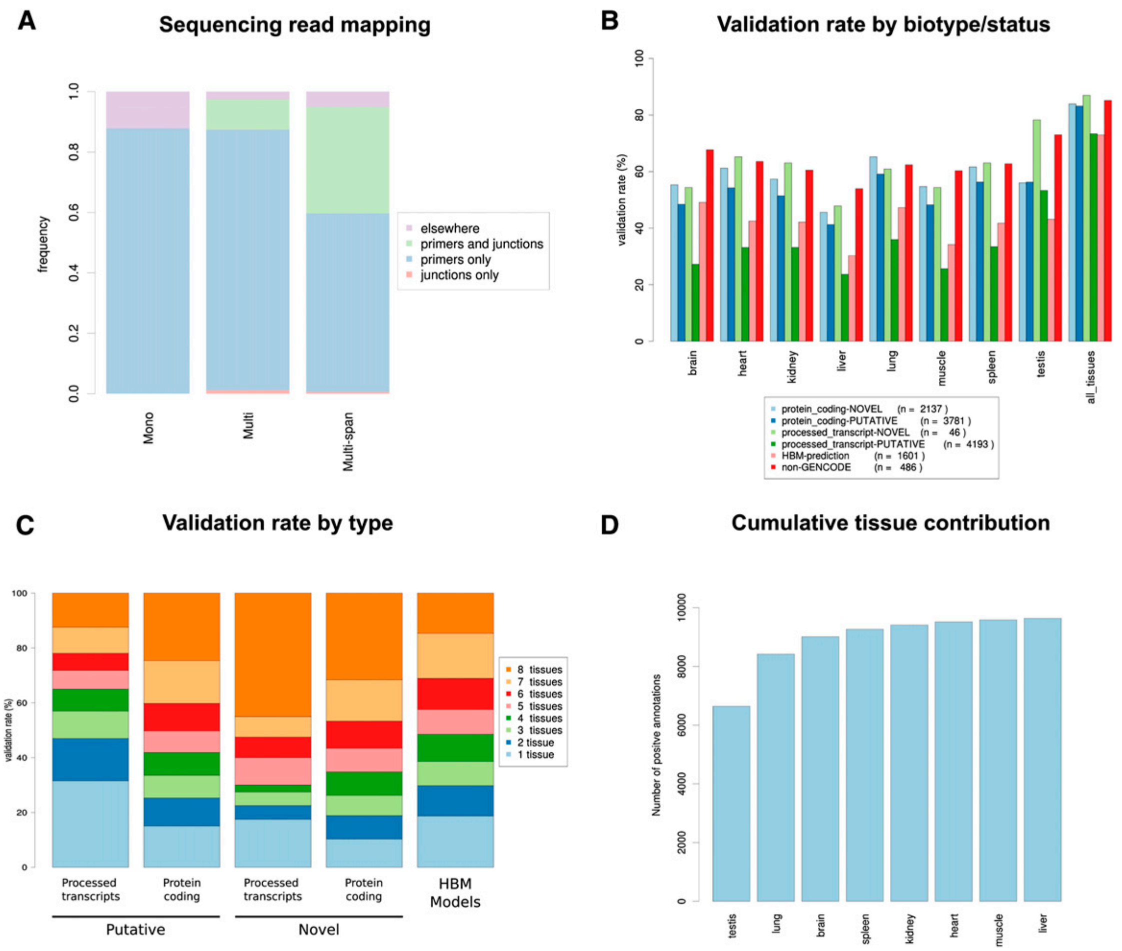

Figure 2. Validation rate of GENCODE and HBM gene models by RT-PCR-seq. (*A*) Mapping distribution of the sequencing reads obtained by RT-PCR-seq of gene models belonging to the "Mono," "Multi," and "Multi-span" categories (see text and Fig. 1B for details). (*B*) Validation rate of "novel" and "putative" GENCODE models (e.g., NOVEL processed transcripts) by biotype and status, of ENCODE Consortium predicted models ("non-GENCODE") and of models inferred from the Illumina Human Body Map RNA-seq effort (HBM-prediction) for each assessed tissue and all tissues together (*far right*). The corresponding statistics, in particular, results for biotypes/statuses with few tested models such as "NOVEL nonsense mediated decay" and "NOVEL retained intron," are presented in Supplemental Table S2. (*C*) Numbers of tissues in which validated GENCODE gene models (NOVEL protein coding, NOVEL processed transcripts, PUTATIVE processed transcripts, and PUTATIVE protein coding) and HBM models are detected. (*D*) Cumulative number of validated GENCODE gene models and HBM predictions in the eight assessed tissues.

and 568 novel splicing events within known exons (new exons that intersect not more than 80% of the length of a GENCODE annotated exon; see Methods; note that some assessed genomic intervals have both classes of novel exons). The vast majority of these novel exons (86%, 1046/1212) present canonical donor and acceptor sites. Forty-four percent and 83% of the internal new exons are overlapped over 90% of their length by sequencing reads of the deep transcriptome profiling of 16 different human tissues [4.99 billion individual sequence reads from adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testis, thyroid, and white blood cells; poly(A)$^+$ RNA] and 15 human cell lines [5.55 billion sequences from A549, AG04450, BJ, GM12878, H1-hESC, HMEC, HSMM, HUVEC, HeLa-S3, HepG2, K562, MCF7, NHLF, NHEK, and SK-N-SH; poly(A)$^+$ RNA] generated by the Illumina "Human Body Map" (HBM) project and ENCODE (Djebali et al. 2012b; The ENCODE Project Consortium 2012), confirming that they are new transcript models. Likewise, 45 novel internal exons within protein-coding loci are supported by newly released ESTs and/or cDNAs. Their splice sites (44 canonical, one noncanonical) are conserved across several species (A Nitsche, D Rose, M Fasold, PF Stadler, in prep.; see also http://splicemap.bioinf.uni-leipzig.de/). Thirty-four of these exons are incorporated in protein-coding transcripts, and the remaining 11 are integrated in processed transcripts. Some examples are presented in Figures 3 and 4 and Supplemental Figure S1C,D. A large fraction of these new exons is tissue-specific (69%), but some are detected ubiquitously (Figs. 3B, 4B) (e.g., 6% were identified in at least five tissues).

All internal canonically spliced new exons of protein-coding genes (n = 313) were subsequently manually annotated by the GENCODE pipeline. They are generally poorly conserved across vertebrates as shown by the distribution of their phastCons scores (Fig. 5A). They do not overlap repeats more than intergenic and intronic sequences as repeatedly shown for lineage specific exons (Supplemental Fig. S2; for review, see Keren et al. 2010). They can be subdivided into 70 new coding (22%), 173 new nonsense-mediated decay (NMD, 55%), 55 novel UTR (18%), and 15 new noncoding exons (5%) (Supplemental Table S4). Whereas it is

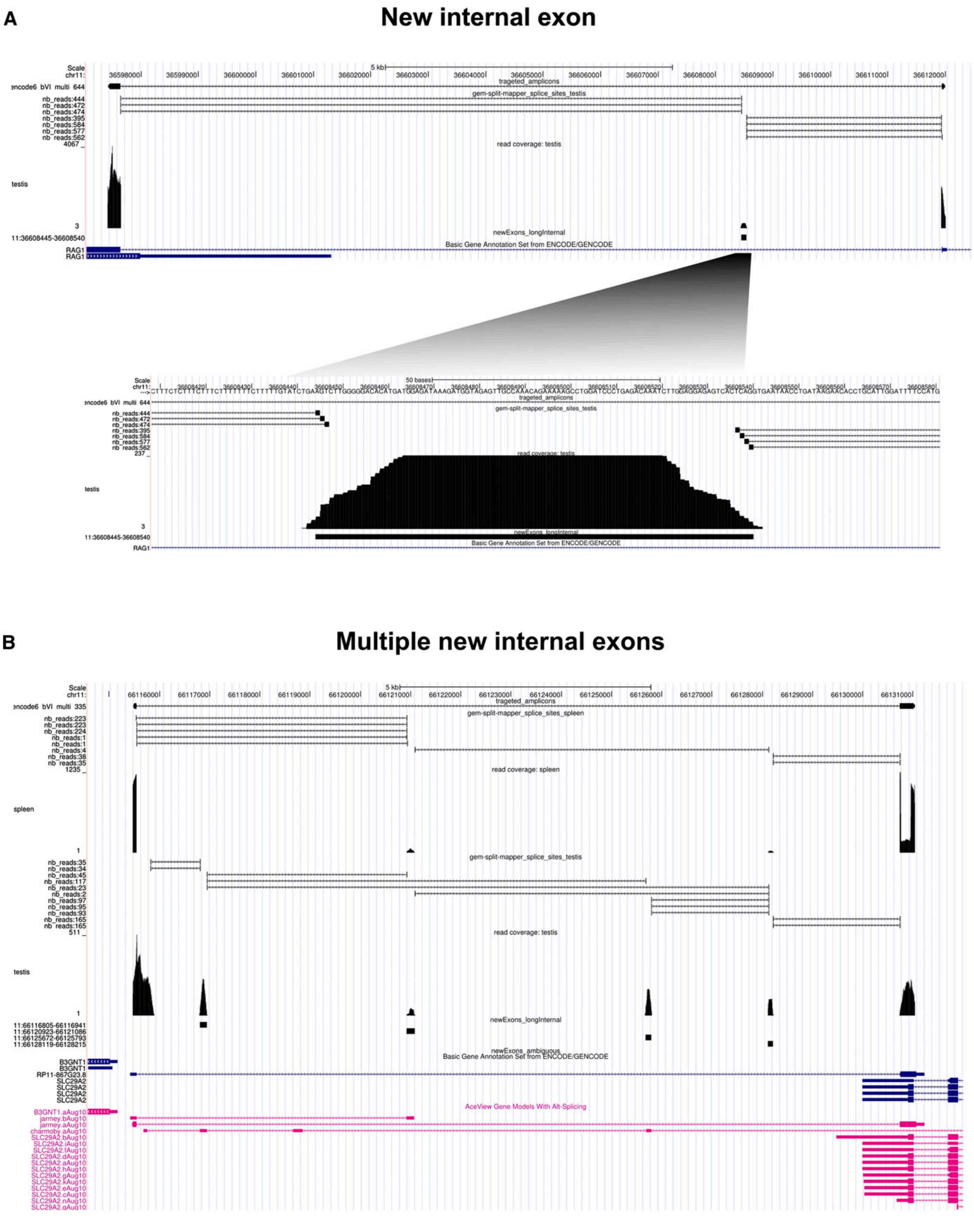

Figure 3. Examples of newly identified internal exons. The two panels show views from the UCSC Genome Browser. The tracks from *top* to *bottom* show the scale (black), coordinates (black), the tested GENCODE model (black boxes joined by thin black lines), the introns predicted using the GEM split-mapper, and unmapped reads from the indicated tissue (black ticks joined by thin black lines; number of split-reads for each predicted introns are indicated on the *left*; see Methods), the RT-PCR-seq sequence reads coverage from the indicated tissue (black; scale on the *left*), the newly exons identified (black boxes; coordinates are indicated on the *left*), the ENCODE/GENCODE annotated models (blue or green boxes [exons] joined by thin blue or green lines, respectively), and Aceview predictions using RNA-seq (magenta boxes [exons] joined by thin magenta lines) (Thierry-Mieg and Thierry-Mieg 2006). Identification of a novel internal exon in testis (*A*) and four novel internal exons and three novel transcript isoforms in spleen and testis (note that some, but not all four, novel exons are supported by RNA-seq) (*B*).

difficult to draw a general conclusion about the functionality of these new exons, some interesting cases could be pinpointed. For example, a new exon in the *BAD* gene interrupts one pro-apoptotic Pfam domain (Bcl-2_BAD, PF10514) but inserts another domain (GVQW, PF13900) commonly found in caspases, a family of proteins crucial to the apoptotic pathway (Fig. 6A). The two new NMD-inducing "poison" (Lareau et al. 2007) prone exons found in the *NR1H4* locus are highly specific to liver, possibly controlling expression of that gene in this tissue (Fig. 6B). Likewise, we identified two new mutually exclusive 5′-UTR *ECI2* exons (Fig. 6C). Some of the novel exons, especially within the coding, NMD and UTR categories, are evolutionarily conserved (see outliers in Fig. 5B). For example, we identified a new highly conserved exon within the *KIAA0528* gene (Fig. 6D). Its acceptor and donor sites are conserved back to medaka, while the encoded peptide is highly conserved back to the anolis lizard (Supplemental Fig. S3). We conclude that RT-PCR-seq can be used to further improve the current annotation and discover new gene structures.

RT-PCR-seq to substantiate RNA-seq predictions

Since we showed that GENCODE (or any other annotation for that matter) does not yet fully represent the complexity of the human transcriptome, we took advantage of the deep transcriptome profiling by HBM to uncover novel gene models. The 3.8 billion individual sequence reads were aligned on the human genome to predict alignment blocks (rough exon models), splice sites, and finally, novel gene models (see Methods for the Ensembl RNA-seq pipeline). At each locus, the transcript model with the greatest number of supporting reads is displayed on the Ensembl genome browser. Of them, 5918 do not overlap any loci depicted in GENCODE freeze version 7. Thus they potentially represent new noncoding RNA genes or alternatively unannotated 5′- or 3′-UTR portions of known genes, because the vast majority of these models were shown to have poor coding potential using comparative genomics and mass spectrometry (Lin et al. 2011; Harrow et al. 2012). We could design primers on splice-junctions of 1601 of those models to assess them experimentally by RT-PCR-seq. We validated 73% of the new HBM models outlined by the Ensembl predictions in an average of 4.5 tissues (Fig. 2B,C), de facto enriching the future complexity of the GENCODE annotation of noncoding RNAs genes by 1168 novel genes, a 3.7% increase. Because this rate of validation is close to the sensitivity of the RT-PCR-seq method with eight tissues for noncoding transcripts (79%; see above), we suggest that a large fraction of the nonvalidated HBM models might be bona fide transcripts rather than false-positive predictions. Our findings demonstrate the effectiveness of RNA-seq combined with RT-PCR-seq to uncover new genome features. These two technologies were simultaneously similarly paired to unravel expressed pseudogenes by the GENCODE Consortium (Pei et al. 2012).

Comparing RT-PCR-seq and RNA-seq

Then we compared the efficiency of RT-PCR-seq and transcriptome profiling with RNA-seq to identify and validate splice sites of lower confidence and poorly expressed gene models. We used the deep transcriptome profiling of 16 human tissues and 15 human cell lines achieved by HBM and ENCODE for these comparisons (see above). We distinguished three different classes of exon–exon junctions: (1) non-unique exon–exon junctions (i.e., common to at least two GENCODE gene isoforms and for whom we could design RT-PCR-seq primer pairs; $n = 135{,}617$); (2) specific unique exon–exon junctions (i.e., specific to a unique GENCODE gene isoform; $n = 52{,}994$); and (3) unique exon–exon junctions belonging to the lower-confidence novel or putative GENCODE models ($n = 8750$). Of the non-unique exon–exon junctions present in GENCODE freeze version 8, 63.6% (86,277; HBM) and 56.6% (76,791; ENCODE) are substantiated by at least two reads in at least one tissue (Fig. 7) (see Methods). Similarly, 34.4% (18,210; HBM) and 25.9% (13,717; ENCODE) of the specific unique exon–exon junction could be validated (Fig. 7). In contrast, only 16.0% (1397) and 8.7% (761) of low-confidence unique exon–exon junction models are found in the HBM and ENCODE RNA-seq data sets, respectively (Fig. 7) (Methods). These fractions should be compared with the 79% validation rate of the RT-PCR-seq method (8057 out of 10,162 tested junctions) (Fig. 7). The success rate reached by our novel targeted approach is therefore significantly higher than random sampling of RNA molecules ($P < 1 \times 10^{-321}$, Fisher exact test), which is impacted by a Poisson distribution leading to poor sampling of low-expressed transcripts.

RT-PCR-seq specificity and sensitivity

To assess the sensitivity of RT-PCR-seq, we compared its validation rate with the number of HBM sequencing reads crossing and thus validating the junction, a proxy of the abundance of this isoform. Extremely rarely transcribed junctions (i.e., splice junctions not identified in an average of 312 million of RNA-seq reads per tissue) were validated ~40% of the time (range 32%–49% depending on the tissue) by our RT-PCR-seq pipeline, while junctions supported by a single HBM read were validated at a rate of 90% (range 86%–94%) in each assessed tissue (Fig. 8A). This validation rate grew further to reach 100% for junctions overlapped by at least nine HBM reads (Fig. 8B). These results clearly demonstrate the extreme sensitivity of our targeted approach that combines PCR amplification and new sequencing technologies. Is this high sensitivity coming at a cost on specificity? To gauge the rate of false positives, we created random junctions by combining untargeted GENCODE7 exons from the same chromosome, encoded on the same strand, and respecting their 5′–3′ arrangement (see Methods). Not a single junction out of the 1,097,167 generated was validated, attesting that the RT-PCR-seq method exhibits simultaneously great specificity and high sensitivity.

Discussion

The sequencing of the human genome (Lander et al. 2001; Venter et al. 2001; The International Human Genome Sequencing Consortium 2004) and its complementation with whole-genome assessment of other mammals positioned at key evolutionary junctures of the mammalian kingdom—e.g., chimp (The Chimpanzee Sequencing and Analysis Consortium 2005), mouse (Waterston et al. 2002), cow (Elsik et al. 2009), opossum (Mikkelsen et al. 2007), and platypus (Warren et al. 2008)—have provided the raw material "to investigate biological phenomena in a comprehensive, unbiased, hypothesis-free manner" (Lander 2011). To fully exploit this information, the scientific community requires a reliable coding and noncoding gene catalog, the compilation of which was assigned, within ENCODE, to the GENCODE Consortium (Harrow et al. 2012). With more than 51,096 genes (20,026 coding and 31,070 noncoding; GENCODE version 8, March 2011), this manually curated annotation is richer than any previously available annotation (e.g., the UCSC Genome Browser) (Fujita et al. 2011). We demonstrate here that it is of extremely high quality, because even its lower-confidence gene and transcript models can be experimentally

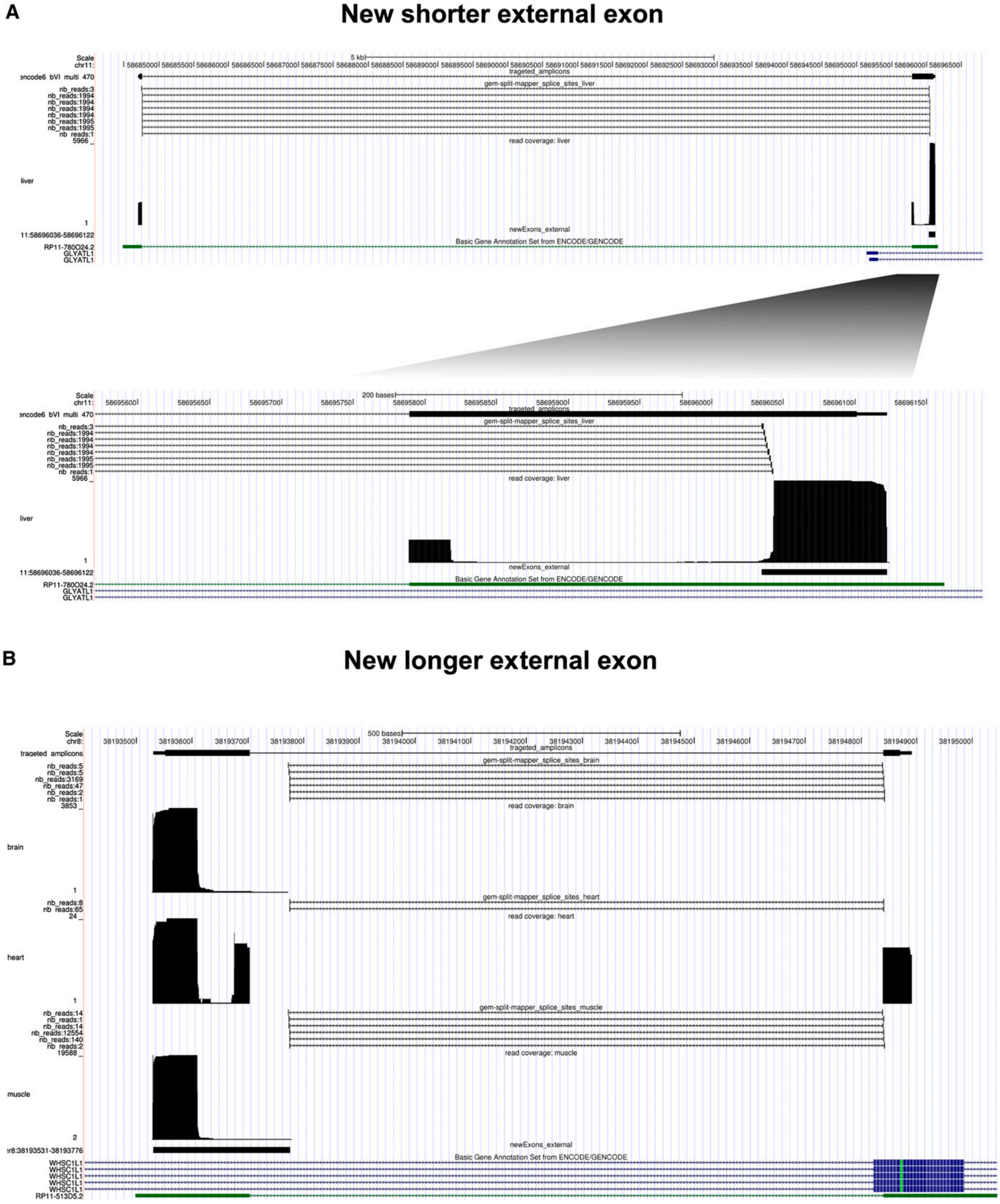

Figure 4. Examples of newly identified external exons. The different tracks are described in the legend of Figure 3. The first example is that of a new exon shorter than the annotated exon identified in liver (*A*), while the second is longer than the annotated exon and detected in three different tissues (brain, heart, and muscle) (*B*).

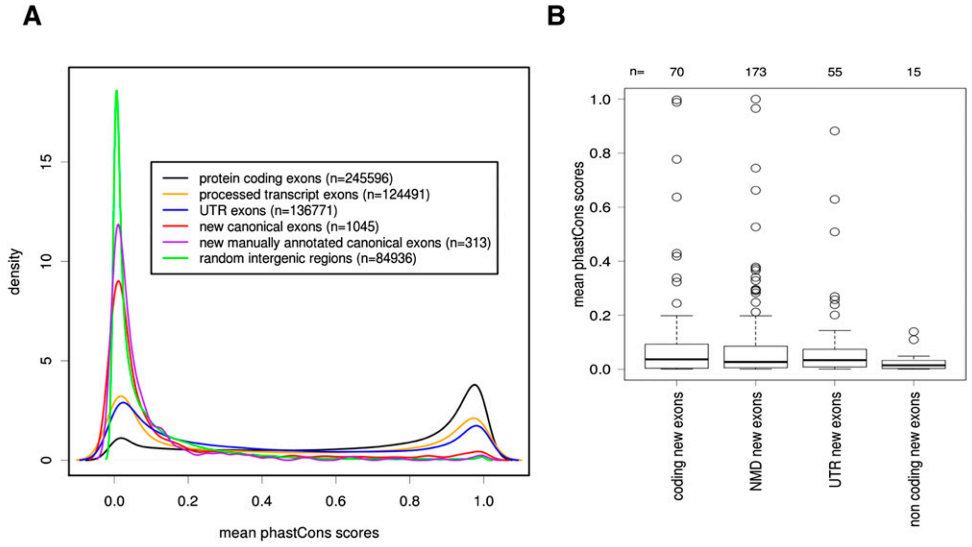

Figure 5. Conservation of newly identified exons. The conservation of the newly identified exons was assessed by comparing the distribution of phastCons scores of different genomic elements. The higher the phastCons score, the greater is the probability for a given element to be under negative selection. (*A*) Distribution of the mean phastCons scores of protein-coding exons (black line), noncoding (processed transcript) exons (yellow), UTR exons (blue), new canonical exons identified in this study (red), new canonical exons identified in this study, and mapping to protein-coding loci (purple) and random intergenic regions (green). (*B*) Boxplot distribution of the mean phastCons scores of new canonical exons identified in this study and mapping to protein-coding loci split per functional classes: protein coding, NMD (transcript subject to Nonsense Mediated Decay), UTR, and noncoding (see Methods).

validated using RT-PCR-seq (79.3% and 80.7% validation rate with more or less conservative criteria; see text for details). It does not, however, fully represent the complexity of the human transcriptome, since we identify novel internal exons in >11% of the genomic intervals we interrogated with RT-PCR-seq. Because GENCODE coding and long noncoding transcripts have an average of 4.3 and 2.2 exons, respectively (Harrow et al. 2012), we can conservatively estimate that ∼18% of the annotated genome loci have yet unrecognized exons. Large ongoing efforts to profile the human transcriptome by RNA-seq (e.g., HBM and ENCODE) confirm this assumption because they revealed a large number of transcribed islands that do not overlap GENCODE annotations (Djebali et al. 2012b). These isolated islands and archipelagos potentially represent novel exons and novel genes, respectively. Our RT-PCR-seq method has proven to be a powerful tool to evaluate such de novo transcript models derived from RNA-seq experiments. We validated 72% of the assessed model confirming that they are bona fide transcribed units missing from the current annotations. Together these observations support the notion that the human genome is pervasively transcribed as suggested by multiple authors (Kapranov et al. 2002; Denoeud et al. 2007; The ENCODE Project Consortium 2007; Djebali et al. 2008, 2012b; Clark et al. 2011). Further studies are warranted to identify the molecular roles of the new genes identified by RNA-seq (and validated by RT-PCR-seq) and the novel exons of known genes pinpointed by RT-PCR-seq. While we show that the majority of the novel yet-unannotated exons are not conserved, we found exons of coding transcripts that maintained their open reading frame (ORF) through evolution. Furthermore, when they are incorporated in coding transcripts, they modify the amino acid sequence and domain structure of the encoded protein about one-fifth of the time.

The complexity of the human transcriptome surpasses current RNA sequencing capabilities. Furthermore, lower-confidence and unique GENCODE annotated exon–exon junctions validated by RT-PCR-seq were often not validated by deep RNA-seq transcriptome profiling, which usually poorly samples low-expressed transcripts. Targeted approaches, such as RT-PCR-seq developed here or the recently described RNA CaptureSeq (Mercer et al. 2012), reach an exquisite sensitivity that exceeds that of RNA-seq. For example, we recovered and validated by RT-PCR-seq between 32% and 49% of extremely rarely transcribed exons (i.e., exons with splice junctions not represented within the HBM RNA-seq data set; see above for details). The major limitation of our method resides in the extremely stringent criteria we typically use to design primers, which often resulted in junctions that could not be tested. The partial relaxation of these designing standards allows us to notably increase the number of testable junctions. The only drawback is the possible coamplification of transcripts mapping elsewhere on the genome, which can be overcome by a deeper sequencing of the RT-PCR amplification pool. We are currently modifying the RT-PCR-seq method to allow quantification of RNA molecules in the future.

Our work underscores that targeted approaches will be needed, in coordination with unbiased approaches such as RNA-seq for the understanding and cataloging of all of the genic elements encoded in the human genome, one of the goals of the second decade following the sequencing of the human genome (Lander 2011).

Methods

GENCODE gene categories

The GENCODE annotation and the statuses and biotypes it uses are detailed in Harrow et al. (2012) and in the HAVANA annotation guidelines. The ones mentioned in this study are summarized below.

The "status" represents the types of evidence used to define one transcript; it reflects also the level of confidence that can be placed in a specific gene model. Three different statuses can be assigned to genes:

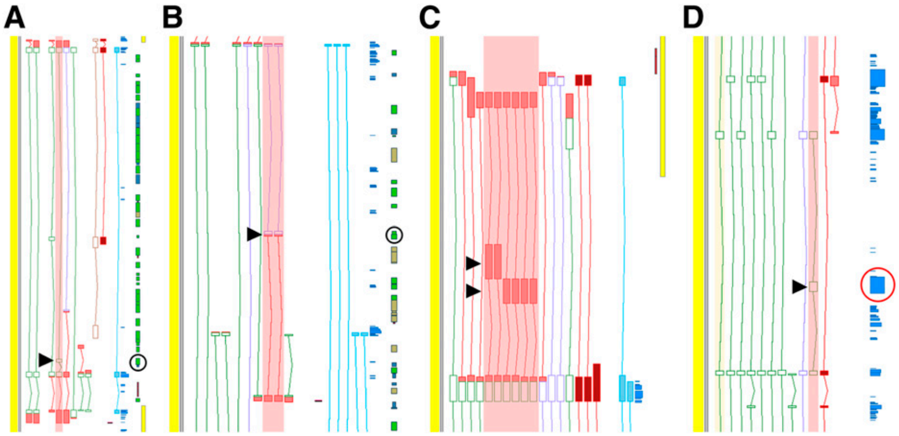

Figure 6. Function of newly identified internal exons of coding loci. Screenshots of gene annotation taken from the Zmap annotation interface. ORF exons of protein-coding models (open green boxes); UTR exons (filled red boxes); variants predicted to be subject to NMD have an ORF represented as open purple boxes. (*A–D*) Both pre-existing manual annotation and models incorporating model exons identified in this study (highlighted by red shading) are shown. (Black arrowheads) Novel exons; (black circles) overlapping repeat elements; (red circles) overlapping blocks of high cross-species conservation. (*A*) The BCL2-associated agonist of cell death (*BAD*) locus. A model incorporating the novel exon is predicted to encode an ORF that breaks the Pfam domain present in other protein-coding models at the locus (Bcl-2_BAD, PF10514) but introduces another domain (GVQW, PF13900) found in caspases, a family of proteins essential for apoptosis, the pathway regulated by the Bcl-2_BAD domain. The novel exon shows no cross-species conservation and overlaps a SINE. (*B*) The nuclear receptor subfamily 1, group H, member 4 (*NR1H4*) locus. The two highlighted variants that include the novel exon are both predicted to be subject to NMD and only differ at a small shift in their splice acceptors. Both transcripts are highly liver-specific, and the novel exon was not identified in any other tissue investigated; again there is no cross-species conservation and the novel exon overlaps a SINE. (*C*) The enoyl-CoA delta isomerase 2 (*ECI2*) locus. Here two novel exons were identified in the 5′-UTR region of the locus. Although there are only two novel exons, alternative splice donor and acceptor sites in flanking exons suggest many different intron combinations that expand the transcripts repertoire. Neither novel exon overlaps a repeat element or region of cross-species conservation. (*D*) The *KIAA0528* locus. A novel coding transcript incorporating a novel exon remains in-frame relative to other coding transcripts at the locus. The novel exon overlaps a region of exceptionally high conservation as indicated by a peak in the phastCons (44 mammals) track (in blue, circled). The alignment of this exon with other vertebrates' genomes is shown in Supplemental Figure S3.

- *Known*. Identical to known cDNAs or proteins from the same species and has an entry in species-specific model databases (HGNC or RefSeq for human).
- *Novel*. Identical or homologous to cDNAs from the same species, or proteins from all species.
- *Putative*. Identical or homologous to spliced ESTs from the same species.

Genes may have no associated *status* if this is not applicable, as for example, with the majority of pseudogenes.

The "biotype" indicates the biological significance of genes, as annotated in adherence to the HAVANA guidelines (http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/assets/guidelines.pdf).

- *Protein coding*. Contains an open reading frame (ORF).
- *Processed transcript*. Does not contain an ORF. In human, they are further subclassified into one of the following types: Non-coding, 3prime_overlapping_ncrna, Ambiguous_orf, Antisense, LincRNA, ncRNA_host, Retained_intron, Sense_intronic, Sense_overlapping, Processed transcript.

Primer design

The mapping location of the PCR primers is a crucial element of RT-PCR-seq because amplification products are directly sequenced without any fragmentation step (see below) resulting in a large fraction of sequencing reads corresponding to the 5′ and 3′ ends of the target regions. The "junction primer" is positioned within exon *x* not more than 65 bp away from the targeted junction to ensure that sequencing reads will cross the junctions with a minimum of 10 nt, while the second primer maps within exon $(x + 1)$ or $(x - 1)$. They are designed using Primer3 (Rozen and Skaletsky 2000) ("Multi-span" primers in Fig. 1B,C). We used parameters minimizing the formation of primer dimers and maximizing primers "stickiness" (Supplemental Table S5). Primer pairs are further filtered for mapping within repeat-regions and alternative priming within 30 kb in duplications and paralogous sequences (maximum of two tolerated mismatches). This stringent design procedure allowed designing assays in 182,907 out of 441,654 (41%) junctions. To increase the fraction of junctions we could possibly test, we then designed primer pairs further away from the junction ("Multi" primers in Fig. 1B,C). Primer pairs for monoexonic models were designed with the same parameters ("Mono" primers in Fig. 1B,C).

RT-PCR-seq

First-strand cDNA samples were prepared from eight commercial human poly(A)$^+$ RNAs (brain, heart, kidney, liver, lung, spleen, skeletal muscle, and testis; BD-Clontech) with the SuperScript III kit (Invitrogen). Amplifications were performed in a final volume of 12.5 μL with JumpStart REDTaq ReadyMix (Sigma-Aldrich) and a primer concentration of 0.4 μM in 384-well plates format on an automatized Evoware platform (TECAN) combined with a Tetrad2 thermocycler (Bio-Rad) that allows processing four plates in parallel. Because monoexonic amplification is sensitive to genomic DNA contaminations, monoexonic models were assessed by amplification of cDNA in which a dNTP analog was incorporated using the mRNA Selective PCR Kit (TAKARA) as described in Washietl et al. (2007).

Aliquots (two of 12.5 μL) of up to 4700 independent amplification reactions were pooled together respecting the tissue origin and purified with the QIAquick PCR Purification Kit (QIAGEN) following the manufacturer's instructions. Purified pools were directly used to generate sequencing libraries with the TruSeq DNA sample prep kit (Illumina) according to the manufacturer's recommendations, except for sonication fragmentation (expected amplicons size comprised between 150 and 400 bp) and exploitation of MetaPhor Agarose, an agarose specially made for separating small DNA fragments (Cambrex). TruSeq libraries were subsequently sequenced on a single Illumina Genome Analyzer IIx or multiplexed on Illumina HiSeq2000 lanes.

Validation of exon junctions by RT-PCR-seq

Sequence reads with *phred* quality score ≥23 were mapped with Bowtie version 0.12.5 (Langmead et al. 2009) against the genome, the predicted amplicons, and the transcriptome (all GENCODE

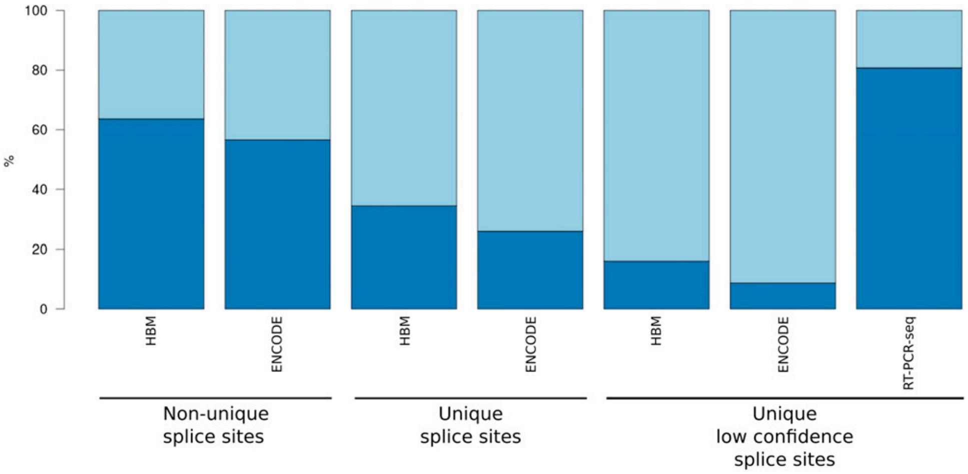

Figure 7. Comparison of validation rates. Validation rates (dark blue) of GENCODE non-unique splice junctions (i.e., common to more than one GENCODE transcript isoform; "common" junctions), GENCODE unique splice junctions (specific to a single GENCODE transcript isoform; "specific" junctions), and lower confidence GENCODE unique splice junctions (specific to a single novel or putative GENCODE transcript isoform; "specific and low expressed" junctions) by Illumina Human Body Map RNA-seq (HBM), ENCODE RNA-seq (ENCODE), and RT-PCR-seq are shown in bar plot format. Exon–exon junctions were considered substantiated by a RNA-seq data set if they were overlapped by at least two split-reads (Methods). The criteria used to validate a junction by RT-PCR-seq are detailed in the main text and schematized in Figure 1C. Note that both RNA-seq data sets fail to corroborate a substantial fraction of the rare and lowly expressed junctions when compared with RT-PCR-seq.

version 8 splice junctions), allowing up to two mismatches. Reads aligning at multiple positions were discarded unless their best hits (with the least mismatches) were unique. Multiexonic annotations ("Multi-span" and "Multi") (Fig. 1B,C) were considered experimentally validated if at least 10 reads cross the targeted splice-sites each with a minimum of 10 nt on both sides of the breakpoint. Monoexonic annotations (Mono) were corroborated if a minimum of 10 reads mapped anywhere within the primer boundaries (Fig. 1B,C). We generated output files containing information such as validation results, mapping results (BAM format), read coverage (bedGraph), and validating reads and expected amplicons (BED) with BEDtools and SAMtools (Li et al. 2009; Quinlan and Hall 2010), which can be loaded directly into web browsers such as the UCSC Genome Browser (Figs. 3, 4).

Split mapping and validation of novel exons

Sequencing reads are split-mapped using GEM (http://sourceforge.net/apps/mediawiki/gemlibrary). All possible exons resulting are compared with the GENCODE gene set, and only exons that intersect <80% of the length of a GENCODE annotated exon are considered. Potential novel exons are divided in three categories, and their confidence score (*S*) is computed independently.

External exons

Novel exons intersecting an exon in which a primer was designed were considered if S_{ext} was ≥1.5.

$$S_{ext} = E_{coverage} + (1 - (1/\log2(\text{SplitReads} + 1)))$$

- $E_{coverage}$: Exon coverage, proportion of the exon covered by reads mapping the genome or by reads split-mapping the targeted splice site.
- SplitReads: Number of split-mapped reads defining the splice site.
- The S_{ext} score has two components. The $E_{coverage}$ reflects how well the potential exon defined by GEM is covered by reads, while the second component approaches 1 as SplitReads increases.

Long internal exons long

Exons that do not contain any primed sequence were considered if $S_{longint}$ was ≥1.9.

$$S_{longint} = E_{accuracy} + (1 - (1/(\log2(\text{MinSplitReads} + 1))))$$

- $E_{accuracy}$: Exon accuracy, length of the intersection between the projected reads and the exon divided by the length of the union between projected reads and the exon.
- MinSplitReads: Internal exons are characterized by two novel splice sites. MinSplitReads is the number of reads spanning the less covered splice site.

Short internal exons

Exons shorter than the read length that do not contain any primed sequence were considered if $S_{shortint}$ was ≥20. Since the exon is shorter than the read length, no unspliced read will map the exon interval. The $S_{shortint}$ score corresponds to the number of reads spanning the less covered splice site.

Novel exons identified by RT-PCR-seq were considered validated if their coordinates intersected with those of at least one HBM or ENCODE RNA-seq data set sequence read.

Manual annotation of novel exons

New canonically spliced internal exons mapping within protein-coding loci were submitted to manual annotation by the GENCODE annotators. They were classified in four categories: (1) coding new exons that maintain the ORF between downstream and upstream exons. New exons can also be classified as coding if they disrupt the ORF but NMD is not triggered (i.e., novel exon in the most 3′ annotated intron and producing an alternative stop codon). (2) NMD

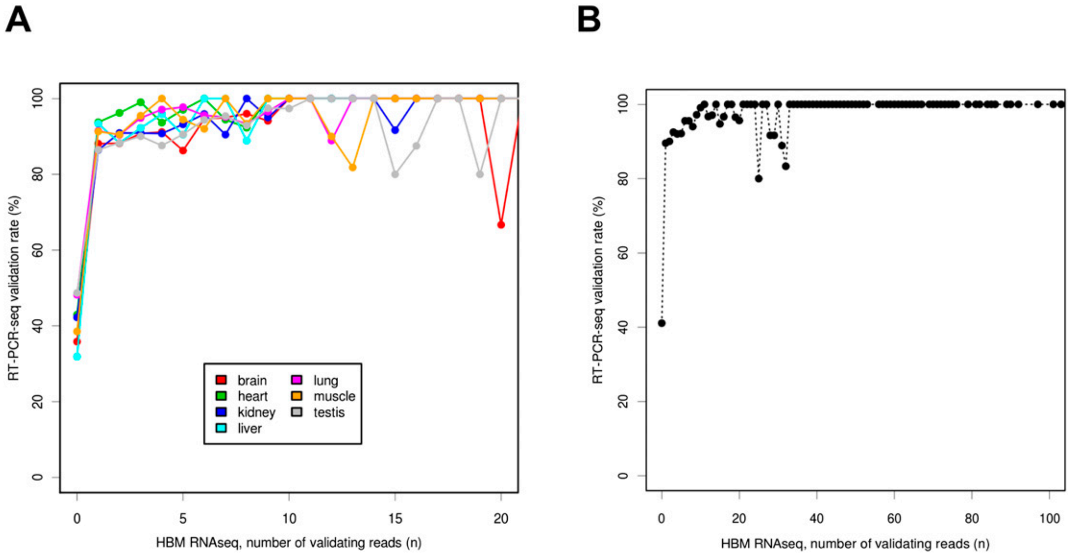

Figure 8. Sensitivity of the RT-PCR-seq method. The RT-PCR-seq validation rates of exon–exon junctions are shown as a function of the abundance of the targeted transcript isoforms. The latter was ascertained by counting the number of validating reads detected in HBM RNA-seq experiments. The results obtained for the seven tissues used in both HBM RNA-seq and GENCODE RT-PCR-seq were considered (brain, heart, lung, testis, liver, kidney, and skeletal muscle) either separately (*A*) or merged together (*B*).

new exons contain a stop codon further than 50 nt upstream of the splice donor (Zhang et al. 1998). (3) UTR new exons are incorporated in the 3′ or 5′ annotated UTR. (4) New exons not fulfilling any of these criteria are classified in the unspecified "noncoding" category.

Ensembl RNA-seq pipeline: Building gene models with RNA-seq data

The Illumina Human Body Map (HBM) from the 16 tissues were aligned against the human GRCh37 primary assembly for Ensembl release 59 using Exonerate (Slater and Birney 2005). The read alignments from all tissues were pooled and collapsed into alignment blocks that roughly corresponded to transcribed exons. Read pairing information was exploited to group exons into approximate transcript structures. For each tissue, sequence reads were realigned against these proto-transcripts with Exonerate to create a set of spliced alignments representing the introns. Proto-transcripts and spliced alignments were combined to create all possible combinations of transcript variants represented by the intron-supporting reads. At each locus, the putative variants were then filtered down, and one variant, representing the best-supported transcript model, was chosen for display on the Ensembl genome browser. These models can be visualized following the instructions deposited in http://www.ensembl.info/blog/2011/05/24/human-bodymap-2-0-data-from-illumina/.

Validation of exon junctions by RNA-seq

ENCODE and Illumina Human Body Map projects represent two of the most comprehensive and deepest RNA-seq paired-end data sets. Raw data sequences were mapped with GEM on the genome and transcriptome to build mate-pairs that use genomic reads, as well as reads that cross-splice junctions. We discarded reads mapping at multiple positions, overlapping mate-pairs, as well as mate-pairs mapping to different chromosomes, before intersecting the mate-pairs with the GENCODE gene annotation. Exon–exon junctions were validated by RNA-seq if (1) at least two mate reads from a tissue cross that junction by at least 10 nt; (2) these junction reads did not map to the genome, thus avoiding sequences that originate from retrotransposed pseudogenes; and (3) the other reads of the mate-pairs mapped to the same transcript. These very conservative criteria were applied to ensure that validated exon junctions were connected to exons of the annotated transcripts being assessed and did not originate from unannotated isoforms.

Sensitivity and specificity

The sensitivity was assessed by plotting the validation rate obtained by RT-PCR-seq against the number of validating reads (i.e., reads crossing the junction with at least 10 nt on each side) obtained by HBM RNA-seq. Only the seven tissues used in both HBM RNA-seq and GENCODE RT-PCR-seq were considered (brain, heart, lung, testis, liver, kidney, and skeletal muscle) (Fig. 8A,B). The specificity of the RT-PCR-seq method was estimated through our capacity of validating artificial splice sites generated by pairing GENCODE7 exons. To mimic bona fide splice sites, we only paired exons coming from the same chromosome, mapping on the same strand, and respecting the 5′–3′ order. A total of 40,739,244 testis reads from an experiment assessing the legitimacy of 1603 GENCODE models were mapped to 1,097,164 randomized exon pairs belonging to untargeted GENCODE loci. None of the random splice sites was validated with our standard validation parameters (minimum of 10 validating reads), and only two random splice sites were validated with less stringent parameters (two validating reads).

Data access

All Illumina sequencing reads used to validate exon junctions by RT-PCR-seq are deposited in the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/) under accession numbers ERP000774, ERP000781, ERP000972, and ERP001145. The Illumina Human Body Map (HBM) 50-bp paired-end and the 75-bp single Illumina sequence reads can be accessed at ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) accession: E-MTAB-513; ENA archive: ERP000546. The splice-sites targeted by RT-PCR-seq and their validations are listed within Supplemental Tables S1 and S3, respectively. The detailed manual annotation of the new canonically spliced exons can be found in Supplemental Table S4.

Acknowledgments

We thank the ENCODE Project Consortium and the Illumina "Human Body Map" project for making their data publicly available; and members of the Lausanne Genomic Technologies Facility and Angelika Merkel for technical help. This work was funded by National Human Genome Research Institute (NHGRI)/National Institutes of Health (NIH) grants to GENCODE (U54 HG004555) and Cold Spring Harbor Laboratories (U54 HG004557) subgroups of the ENCODE project. We acknowledge grants from the Swiss National Science Foundation to A.R., from the Spanish Ministry of Science (RD07/0067/0012, BIO2006-03380, and CSD2007-00050) to R.G., and from the Wellcome Trust (WT077198/Z/05/Z) to J.H., A.F., J.M.G., F.K., and T.J.H.

Author contributions: R.G., T.J.H., and A.R. designed the study. C.H. and A.T. performed the RT-PCR-seq experiments with the help of J.C., T.D., and N.W. F.K., A.T., J.M.G., and J.H. selected GENCODE models to be tested. A.F. assessed the coding potentials of novel exons, while B.L.A., T.H., J.H.V., S.W., and S.S. predicted gene models using HBM RNA-seq data. A.R. and C.H. wrote the manuscript with contributions from A.T. and R.G. T.J.H., R.G., and A.R. obtained the necessary financial support. All authors read and approved the final manuscript.

References

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478:** 343–348.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78–94.

Cartault F, Munier P, Benko E, Desguerre I, Hanein S, Boddaert N, Bandiera S, Vellayoudom J, Krejbich-Trotot P, Bintner M, et al. 2012. Mutation in a primate-conserved retrotransposon reveals a noncoding RNA as a mediator of infantile encephalopathy. *Proc Natl Acad Sci* **109:** 4980–4985.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.

Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, et al. 2011. The reality of pervasive transcription. *PLoS Biol* **9:** e1000625; discussion e1001102. doi: 10.1371/journal.pbio.1000625.

Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, et al. 2007. Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res* **17:** 746–759.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* (this issue). doi: 10.1101/gr.132159.111.

Djebali S, Kapranov P, Foissac S, Lagarde J, Reymond A, Ucla C, Wyss C, Drenkow J, Dumais E, Murray RR, et al. 2008. Efficient targeted transcript discovery via array-based normalization of RACE libraries. *Nat Methods* **5:** 629–635.

Djebali S, Lagarde J, Kapranov P, Lacroix V, Borel C, Mudge JM, Howald C, Foissac S, Ucla C, Chrast J, et al. 2012a. Evidence for transcript networks composed of chimeric RNAs in human cells. *PLoS ONE* **7:** e28213. doi: 10.1371/journal.pone.0028213.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi AM, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012b. Landscape of transcription in human cells. *Nature* (in press).

Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, Adelson DL, Eichler EE, Elnitski L, Guigo R, et al. 2009. The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science* **324:** 522–528.

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* (in press).

Flicek P, Keibler E, Hu P, Korf I, Brent MR. 2003. Leveraging the mouse genome for gene prediction in human: From whole-genome shotgun reads to a global synteny map. *Genome Res* **13:** 46–54.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2011. Ensembl 2011. *Nucleic Acids Res* **39:** D800–D806.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, et al. 2011. The UCSC Genome Browser database: Update 2011. *Nucleic Acids Res* **39:** D876–D882.

Guigo R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C, et al. 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc Natl Acad Sci* **100:** 1140–1145.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, et al. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol* **7:** S4.1–S4.9.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* (this issue). doi: 10.1101/gr.135350.111.

The International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916–919.

Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: Diversification, exon definition and function. *Nat Rev Genet* **11:** 345–355.

Kowalczyk MS, Higgs DR, Gingeras TR. 2012. Molecular biology: RNA discrimination. *Nature* **482:** 310–311.

Lander ES. 2011. Initial impact of the sequencing of the human genome. *Nature* **470:** 187–197.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446:** 926–929.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27:** i275–i282.

Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL. 2012. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* **30:** 99–104.

Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447:** 167–177.

Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, et al. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9:** e1001046. doi: 10.1371/journal.pbio.1001046.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.

Parra G, Blanco E, Guigo R. 2000. GeneID in *Drosophila*. *Genome Res* **10:** 511–515.

Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigo R. 2003. Comparative gene prediction in human and mouse. *Genome Res* **13:** 108–117.

Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, Antonarakis SE, Guigo R. 2006. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res* **16:** 37–44.

Pei S, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, et al. 2012. The GENCODE pseudogene resource. *Genome Biol* (in press).

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132:** 365–386.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31. doi: 10.1186/1471-2105-6-31.

Thierry-Mieg D, Thierry-Mieg J. 2006. AceView: A comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* **7:** S12.1–S12.14.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147:** 1537–1550.

van Bakel H, Nislow C, Blencowe BJ, Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol* **8:** e1000371. doi: 10.1371/journal.pbio.1000371.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Wang KC, Chang HY. 2011. Molecular mechanisms of long noncoding RNAs. *Mol Cell* **43:** 904–914.

Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grutzner F, Belov K, Miller W, Clarke L, Chinwalla AT, et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453:** 175–183.

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermuller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, et al. 2007. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* **17:** 852–864.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL. 2008. The vertebrate genome annotation (Vega) database. *Nucleic Acids Res* **36:** D753–D760.

Zhang J, Sun X, Qian Y, Maquat LE. 1998. Intron function in the nonsense-mediated decay of β-globin mRNA: Indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm. *RNA* **4:** 801–815.

Received November 7, 2011; accepted in revised form May 1, 2012.