Downloaded from orbit.dtu.dk on: Oct 26, 2019

DTU Library

Universal Alternative Splicing of Noncoding Exons

Deveson, Ira W; Brunck, Marion E; Blackburn, James; Tseng, Elizabeth; Hon, Ting; Clark, Tyson A; Clark, Michael B; Crawford, Joanna; Dinger, Marcel E; Nielsen, Lars K

Published in:
Cell Systems

Link to article, DOI:
10.1016/j.cels.2017.12.005

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Deveson, I. W., Brunck, M. E., Blackburn, J., Tseng, E., Hon, T., Clark, T. A., ... Mercer, T. R. (2018). Universal Alternative Splicing of Noncoding Exons. *Cell Systems*, *6*(2), 153-155. https://doi.org/10.1016/j.cels.2017.12.005

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

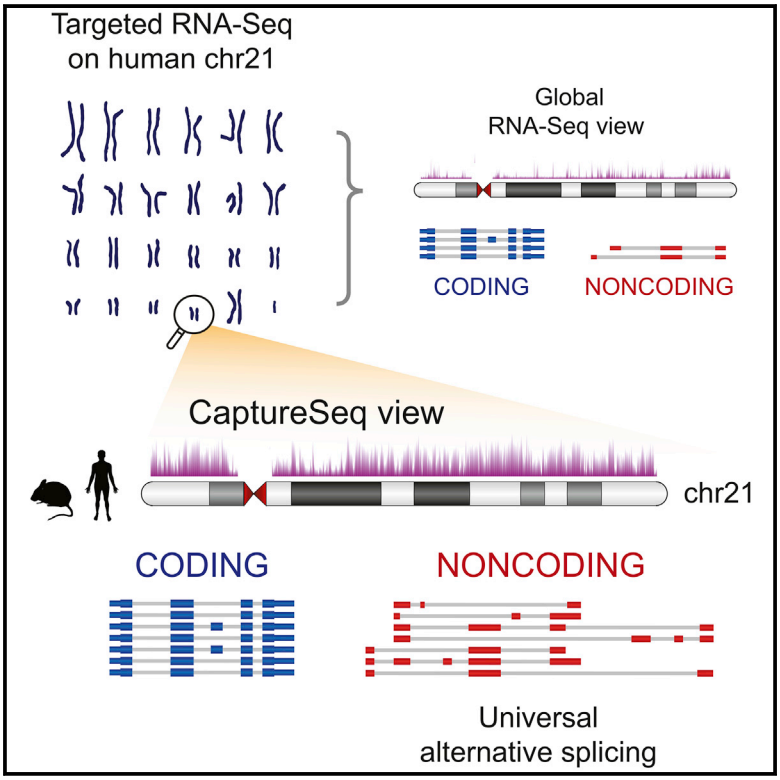
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Cell Systems

Report

Universal Alternative Splicing of Noncoding Exons

Graphical Abstract



Highlights

- Targeted short-read and single-molecule transcriptome survey of human chr21
- Unlike protein-coding exons, noncoding exons are universally alternatively spliced
- Mouse noncoding RNAs are similarly diverse but predominantly divergent
- Human splicing profiles (but not expression) are recapitulated on chr21 in mouse cell

Authors

Ira W. Deveson, Marion E. Brunck, James Blackburn, ..., Lars K. Nielsen, John S. Mattick, Tim R. Mercer

Correspondence

j.mattick@garvan.org.au (J.S.M.), t.mercer@garvan.org.au (T.R.M.)

In Brief

Our high-resolution analysis of human chr21 reveals a fundamental distinction in the architecture of protein-coding and noncoding gene content. Contrary to the impression from more shallow surveys, noncoding RNAs exhibit enriched splicing diversity, with noncoding exons (but not protein-coding counterparts) being near-universally alternatively spliced.

Deveson et al., 2018, Cell Systems 6, 245–255
February 28, 2018 Crown Copyright © 2017 Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2017.12.005

CellPress

Cell Systems
Report

CellPress

Universal Alternative Splicing of Noncoding Exons

Ira W. Deveson,[1,2,11] Marion E. Brunck,[3,4,11] James Blackburn,[1,5] Elizabeth Tseng,[6] Ting Hon,[6] Tyson A. Clark,[6] Michael B. Clark,[1,7] Joanna Crawford,[8] Marcel E. Dinger,[1,5] Lars K. Nielsen,[4,9] John S. Mattick,[1,2,5,*] and Tim R. Mercer[1,5,10,12,*]

[1]Garvan Institute of Medical Research, Sydney, NSW, Australia
[2]School of Biotechnology and Biomolecular Sciences, Faculty of Science, University of New South Wales, Sydney, NSW, Australia
[3]Centro de Biotecnologia FEMSA, Tecnologico de Monterrey, Campus Monterrey, Avenue Eugenio Garza Sada, Monterrey, NL, Mexico
[4]Australian Institute for Bioengineering and Nanotechnology, University of Queensland, Brisbane, QLD, Australia
[5]St Vincent's Clinical School, University of New South Wales, Sydney, NSW, Australia
[6]Pacific Biosciences, Menlo Park, CA, USA
[7]Department of Psychiatry, Warneford Hospital, University of Oxford, Oxford, UK
[8]Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, Australia
[9]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark
[10]Altius Institute for Biomedical Sciences, Seattle, WA, USA
[11]These authors contributed equally
[12]Lead Contact
*Correspondence: j.mattick@garvan.org.au (J.S.M.), t.mercer@garvan.org.au (T.R.M.)
https://doi.org/10.1016/j.cels.2017.12.005

SUMMARY

The human transcriptome is so large, diverse, and dynamic that, even after a decade of investigation by RNA sequencing (RNA-seq), we have yet to resolve its true dimensions. RNA-seq suffers from an expression-dependent bias that impedes characterization of low-abundance transcripts. We performed targeted single-molecule and short-read RNA-seq to survey the transcriptional landscape of a single human chromosome (Hsa21) at unprecedented resolution. Our analysis reaches the lower limits of the transcriptome, identifying a fundamental distinction between protein-coding and noncoding gene content: almost every noncoding exon undergoes alternative splicing, producing a seemingly limitless variety of isoforms. Analysis of syntenic regions of the mouse genome shows that few noncoding exons are shared between human and mouse, yet human splicing profiles are recapitulated on Hsa21 in mouse cells, indicative of regulation by a deeply conserved splicing code. We propose that noncoding exons are functionally modular, with alternative splicing generating an enormous repertoire of potential regulatory RNAs and a rich transcriptional reservoir for gene evolution.

INTRODUCTION

The genome is transcribed into a diverse range of protein-coding and noncoding RNAs collectively termed the transcriptome. The human transcriptome is so large and complex that, even after a decade of investigation by RNA sequencing (RNA-seq), we are yet to achieve a complete census of gene expression. Moreover, our view of a gene as a discrete entity, and of a single protein-coding gene as the functional unit of inheritance, has been undermined by the recognition of pervasive transcription across the genome and interleaved alternative isoforms at individual loci (Carninci et al., 2005; Clark et al., 2011; Djebali et al., 2012; Kapranov et al., 2005; Mercer et al., 2011; Sharon et al., 2013; Tilgner et al., 2015).

RNA-seq has revealed an abundance of small and large non-protein-coding RNAs that are antisense, intronic, or intergenic to protein-coding genes (Derrien et al., 2012; Hon et al., 2017; Iyer et al., 2015; You et al., 2017). Similarly, many protein-coding genes express alternative isoforms that lack extended open reading frames (ORFs; Gonzàlez-Porta et al., 2013). These findings have fueled one of the major debates of modern genetics: the functional relevance of noncoding RNA expression.

The initial sequencing of the human genome provided a catalog of around 20,000 protein-coding genes (Lander et al., 2001). However, at least as many long noncoding RNAs (lncRNAs) have since been identified, and new studies routinely discover novel genes and isoforms (Hon et al., 2017; Iyer et al., 2015; Sharon et al., 2013; Tilgner et al., 2015; You et al., 2017). This failure to achieve a comprehensive annotation of the transcriptome is partly due to the expression-dependent bias of RNA-seq, which limits the capacity of this technique to resolve low-abundance transcripts (Hardwick et al., 2016). This has impeded the discovery and characterization of lncRNAs, which are typically weakly expressed (Cabili et al., 2011; Derrien et al., 2012). As a result, our understanding of lncRNA biology has been largely informed by studying only those examples with sufficient expression for analysis by RNA-seq.

Another limitation of traditional RNA-seq is the reliance on computational assembly of full-length isoforms from short (~100–150 bp) sequencing reads. This is a difficult task, particularly when alternative splicing generates multiple partially redundant isoforms at an individual locus (Conesa et al., 2016). With the emergence of technologies for long-read sequencing it is now possible to read full-length isoforms as single

molecules, negating the challenges posed by transcript assembly. Leading studies have highlighted the utility of single-molecule techniques for resolving complex and precisely organized alternative splicing events (Sharon et al., 2013; Tilgner et al., 2015). However, depth remains a constraint, with rare transcripts often falling below the limits of sampling.

Here, we have attempted to reach the lower limits of the transcriptome by surveying gene expression from a single human chromosome at unprecedented resolution. Chromosome 21 (Hsa21) is the smallest human chromosome (48 Mb), is typical of the human genome in many features (e.g., gene content and repeat density; Figure S1), and has accordingly been used as a model system in transcriptomics (Cawley et al., 2004; Kampa et al., 2004). With trisomy of Hsa21 being the most common chromosomal aneuploidy in live-born children and the most frequent genetic cause of mental retardation, the gene content of this chromosome is also the subject of medical interest (Dierssen, 2012; Letourneau et al., 2014).

To resolve the complete expression profile of Hsa21 while excluding the remainder of the genome, we used targeted RNA-seq (CaptureSeq; Mercer et al., 2014). Complementary oligonucleotide probes were tiled across the chromosome to capture expressed transcripts that were then sequenced deeply on single-molecule and short-read platforms. This approach reduces the influence of the expression-dependent bias inherent to RNA-seq, allowing gene populations encoded within this cross-section of the genome to be observed at high resolution. This reveals a fundamental distinction in the architecture of protein-coding and noncoding RNA.

RESULTS

Targeted RNA-Seq Analysis of Human Chromosome 21

Two limitations of traditional RNA-seq have ensured that, despite considerable attention, the true dimensions of the human transcriptome remain unresolved. First, because sequencing reads are competitively sampled from a single pool in which transcripts of varied abundance are proportionally represented, weakly expressed transcripts commonly evade detection (Clark et al., 2011; Hardwick et al., 2016). Second, the accurate assembly of full-length isoforms from short sequencing reads is challenging, particularly when multiple alternative isoforms are transcribed from a single locus (Conesa et al., 2016; Tilgner et al., 2015).

To address these challenges, we performed single-molecule (PacBio RSII) and short-read (Illumina HiSeq) RNA CaptureSeq, targeting the complete expression profile of Hsa21. We generated biotin-labeled oligonucleotide probes tiling the entire nonrepetitive Hsa21 sequence. These were used to capture full-length cDNA molecules, thereby restricting sequencing to transcripts expressed on Hsa21 (STAR Methods).

To establish our approach, we first performed short-read sequencing on Hsa21-enriched cDNA libraries from the human K562 cell type and compared these with K562 samples analyzed in parallel by conventional RNA-seq (STAR Methods). Whereas just 1.3% of reads from RNA-seq libraries were uniquely aligned to Hsa21, this figure rose to 71.8% after capture, equating to a 54-fold coverage enrichment (Table S1). Analysis of ERCC (External RNA Controls Consortium) spike-in controls confirmed that RNA CaptureSeq accurately measured transcript abundances within the physiological range of gene expression (Figures S2A and S2B), as previously demonstrated (Clark et al., 2015; Mercer et al., 2014). The legitimacy of novel splice junctions identified by CaptureSeq was also verified by RT-PCR and Sanger sequencing (16 of 20 randomly selected examples; Figure S2C and Table S2).

Next, we analyzed Hsa21-enriched cDNA from human testis by deep single-molecule sequencing (STAR Methods). Testis exhibits a distinctly promiscuous transcriptional profile (Soumillon et al., 2013), marking this as an ideal tissue within which to conduct a broad survey of gene content encoded on Hsa21. We obtained 387,029 full-length nonchimeric reads aligning to Hsa21, representing 910 Mb of usable transcript sequence concentrated in ~1.5% of the genome. After filtering, we retrieved 101,478 full-length multi-exonic transcript reads on Hsa21 (Table S3; Figures S3A and S3B).

To reinforce single-molecule isoforms and, more importantly, enable quantitative analyses of expression and splicing, we also performed very deep short-read sequencing on Hsa21-enriched cDNA from testis, brain, and kidney (STAR Methods). An average 65-fold coverage enrichment, relative to conventional RNA-seq, was achieved, and at least 100 million reads uniquely aligning to Hsa21 were obtained for each tissue (Table S4). Strong correspondence between spliced short-read alignments and single-molecule transcripts in our Hsa21 transcriptome profile was observed, with 84.7% of alignment junctions being concordant with introns in single-molecule transcripts (Figures S3C and S3D). Because long PacBio reads provide reliable transcript scaffolds, while read counts for short Illumina reads accurately estimate the abundance of known isoforms, our combined CaptureSeq approach permits robust quantitative transcriptome analyses that do not rely on *de novo* transcript assembly.

Transcriptional Landscape of Human Chromosome 21

Hsa21 has frequently been used as a cross-sectional model for genomics because it is small (48 Mb; ~1.5% of the genome) and typical in terms of gene content, repeat density, and other features (Figure S1; Cawley et al., 2004; Kampa et al., 2004).

Our targeted analysis showed that essentially all (nonrepetitive) regions of Hsa21 encode spliced gene loci, greatly reducing intergenic regions (Figure 1A). This is best illustrated in two "gene deserts" that flank the *NCAM2* gene (extending 2.6 Mb upstream and 4.0 Mb downstream), which were largely devoid of transcript annotations. We discovered that these regions harbored numerous large, multi-exonic, and richly alternatively spliced lncRNAs (Figure 1B).

At protein-coding loci on Hsa21 we identified 7,310 unique multi-exonic isoforms, of which 77% were novel, with respect to current annotations (a combined transcriptome catalog of Gencode v19, MiTranscriptome v2, and FANTOM5; Figure 1C; Tables S5 and S6). Novel isoforms encoded up to 2,365 possible ORFs that were not currently annotated, encompassing 291 novel coding exons and 845 novel ORF introns (Figure 1C; for examples see Figure S4). Although these are predicted ORFs only, and the precise number retrieved is influenced by the choice of cutoff parameters (STAR Methods), this represents a considerable increase on existing annotations.

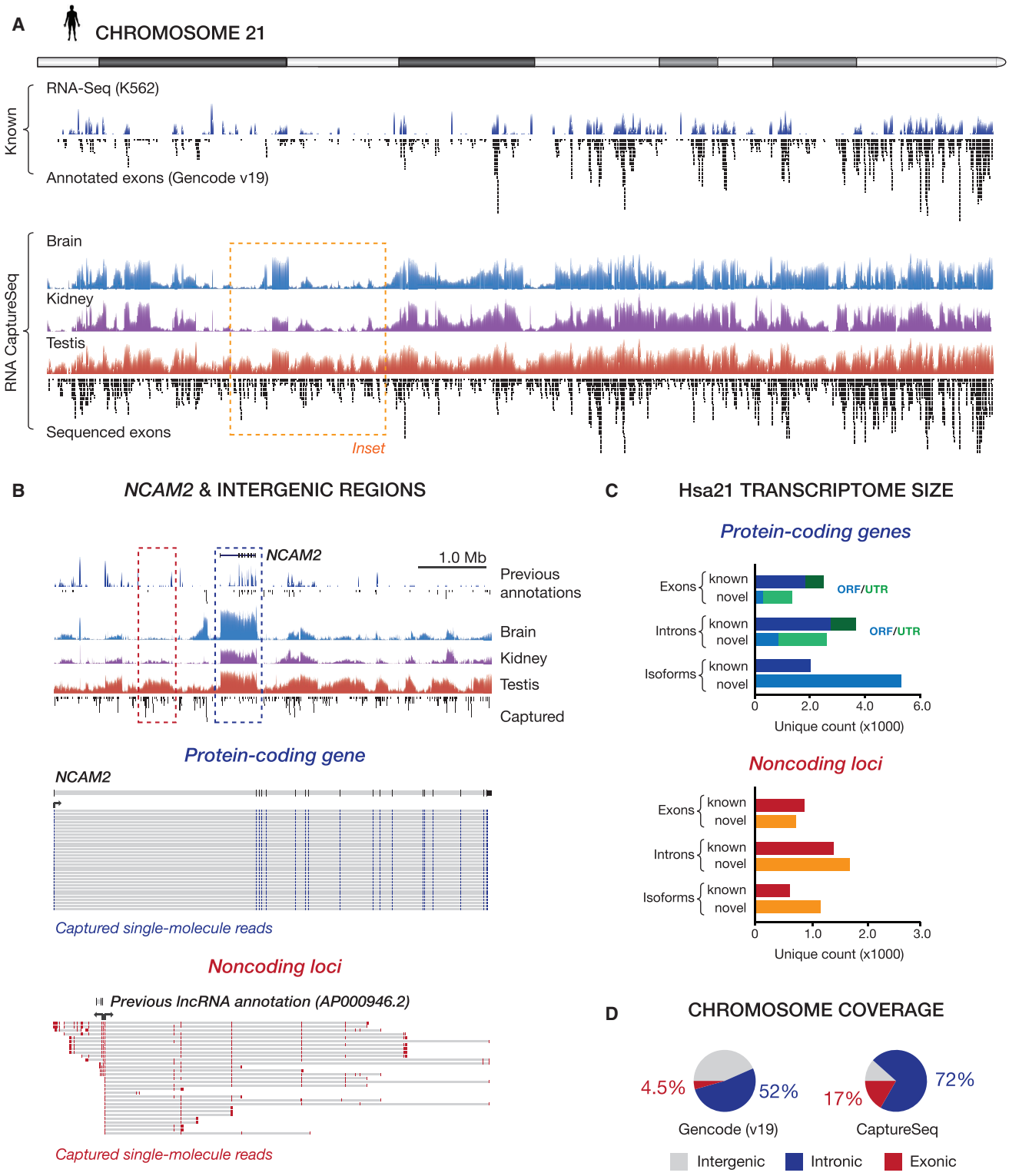


Figure 1. Transcriptional Landscape of Human Chromosome 21

(A) Transcriptional activity recorded across the major arm of human chromosome 21 (Hsa21) by short-read and single-molecule RNA CaptureSeq. Normalized coverage (log scale) from short-read alignments is shown for brain (blue), kidney (purple), and testis (red), with coverage from traditional RNA-seq (in K562 cells; navy) shown for comparison. The coverage tracks shown below are all unique internal exons (black) from transcripts resolved by single-molecule sequencing. Unique internal exons from the Gencode (v19) reference catalog are shown for comparison.

(legend continued on next page)

While this suggests that the protein-coding content of Hsa21 may be underestimated, most novel isoforms to protein-coding genes were noncoding isoform variants or possessed novel UTRs. Alternative splicing of UTRs (both 5′ and 3′) was common and often highly complex (for examples see Figure S5). Single-molecule sequencing was particularly useful for resolving UTR variants, since it does not suffer from sequencing ''edge effects'' that affect short-read transcript assembly (Martin and Wang, 2011). In total, protein-coding loci on Hsa21 encoded 3,931 unique internal exons (34% novel) and 6,365 unique canonical introns (43% novel; Figure 1C; Tables S5 and S6).

At noncoding loci, we identified 1,589 isoforms across Hsa21, encompassing 1,663 unique internal exons (45% novel) and 3,210 unique canonical introns (55% novel; Figure 1C; Tables S5 and S6). Examples of rich isoform diversity at lncRNA loci were routinely resolved with targeted single-molecule sequencing (Figures 1B and S6). In many instances, multiple partial lncRNA annotations were incorporated into single unified loci (Figures S6A–S6C), and the splicing of lncRNAs with neighboring protein-coding genes to form extensively spliced UTRs was also frequently observed (Figure S6D).

By extrapolating our Hsa21 annotations across the broader human genome, we can estimate the existence of 383,000 unique isoforms to protein-coding genes that encompass 147,000 possible ORFs. We predict noncoding gene loci to express 98,000 multi-exonic isoforms (2.1-fold increase by comparison with Gencode v26), incorporating 88,000 internal exons (1.2-fold) and 168,000 introns (1.7-fold). While we have profiled only three tissues, ensuring that these are lower-bound estimates, it is clear that a large amount of transcriptional diversity remains unexplored.

Universal Alternative Splicing of Noncoding Exons

To determine whether RNA CaptureSeq provided a complete profile of transcription on Hsa21, or whether further isoforms remain to be discovered, we generated discovery-saturation curves by incremental subsampling of short-read alignments (STAR Methods). The detection of protein-coding exons and introns approached saturation at a fraction of library depth, indicating that these were near comprehensively sampled (Figure 2A; note that terminal exons were not considered in this analysis). While the detection of noncoding exons also approached saturation, the discovery of noncoding introns (and consequently additional noncoding isoforms) continued progressively toward maximum sequencing depth (Figure 2A). This indicates that although the majority of exons were discovered, noncoding isoforms were not exhaustively resolved, even with the enhanced sensitivity afforded by RNA CaptureSeq.

Because each unique intron represents a different splicing event, this result suggests that alternative splicing generates a seemingly limitless diversity of noncoding isoform variants. To confirm this, we assessed the alternative splicing of internal exons according to percent splice inclusion (PSI) scores (STAR Methods). Unlike protein-coding exons (median PSI = 90.5%), almost all noncoding exons were alternatively spliced (median PSI = 55.5%; Figures 2B and S7). The relative abundances of coding and noncoding introns, compared with exons, further supports this finding; we found that the relative difference between coding and noncoding introns (49.8-fold) is larger than for exons (19.2-fold; Figure 2C), reflecting the greater isoform diversity generated by enriched alternative splicing of noncoding RNAs. We use the phrase universal alternative splicing to mean that nearly every noncoding exon is subject to alternative splicing.

The distinction between coding and noncoding splicing is illustrated by comparison of the protein-coding gene *SAMSN1* with its antisense lncRNA (*SAMSN1-AS1*) and a nearby intergenic lncRNA (*AJ006998.2*). While the majority of single-molecule *SAMSN1* transcripts correspond to one of just two mRNA isoforms for this gene, almost all *SAMSN1-AS1* and AJ006998.2 transcript molecules represent unique isoforms (Figure 2D). Universal alternative splicing was not limited to lncRNA exons but was similarly observed for untranslated exons at protein-coding loci (located in 5′ or 3′ UTRs, or specific to noncoding isoforms; Figures S8 and S9). For example, the 5′ UTR of the protein-coding gene *CHODL* exhibited extensive alternative splicing (Figure S10). Splice acceptor sites at lncRNA and UTR exons harbored canonical splicing elements that were indistinguishable from those found at protein-coding exons (Figure S8B), confirming that noncoding exons are demarcated by bona fide rather than cryptic splice sites.

Our analysis reveals a fundamental distinction between the organization of protein-coding and noncoding gene content. While the diversity of protein-coding isoforms is limited by the requirement to maintain an ORF, no such constraint is imposed on noncoding RNA, allowing the spliceosome to explore the full range of noncoding exon combinations to generate an effectively inexhaustible noncoding isoform diversity. It is worth noting here that this phenomenon is not limited to obscure noncoding RNAs encoded on Hsa21: using publicly available data (Cabili et al., 2011; Clark et al., 2015; Iyer et al., 2015) we examined four well-known functional lncRNAs—*XIST*, *HOTAIR*, *GOMAFU*, and *H19*—and revealed rich isoform diversity and near-universal alternative splicing at each locus (Figure S11).

Comparison of Transcriptional Landscapes in Human and Mouse

To establish whether the novel genes and isoforms unearthed on Hsa21 were conserved between human and mouse, and to determine whether noncoding exons are similarly enriched for

(B) Inset from (A): detail of intergenic regions flanking the protein-coding gene *NCAM2* (2.6 Mb upstream and 4.0 Mb downstream), in which multiple novel long noncoding RNAs (lncRNAs) were detected. Red and blue insets show single-molecule reads supporting *NCAM2* (blue) and two nearby lncRNAs (red). Single-molecule RNA CaptureSeq data extend the gene model for the previously annotated lncRNA AP000946.2 and resolve a novel lncRNA that is transcribed in the opposite direction. Both exhibit extensive alternative splicing, in contrast to *NCAM2*, where the majority of reads represents redundant isoforms.

(C) The number of unique internal exons, unique canonical introns, and nonredundant isoforms resolved by single-molecule RNA CaptureSeq. Content from protein-coding genes (upper) and noncoding loci (lower) are shown separately and, for protein-coding genes, exons/introns belonging to noncoding isoforms or UTR variants (green) are distinguished from predicted ORFs (blue).

(D) The proportion of captured bases on Hsa21 that are exonic, intronic, or silent, according to Gencode (v19; left) and RNA CaptureSeq (right).

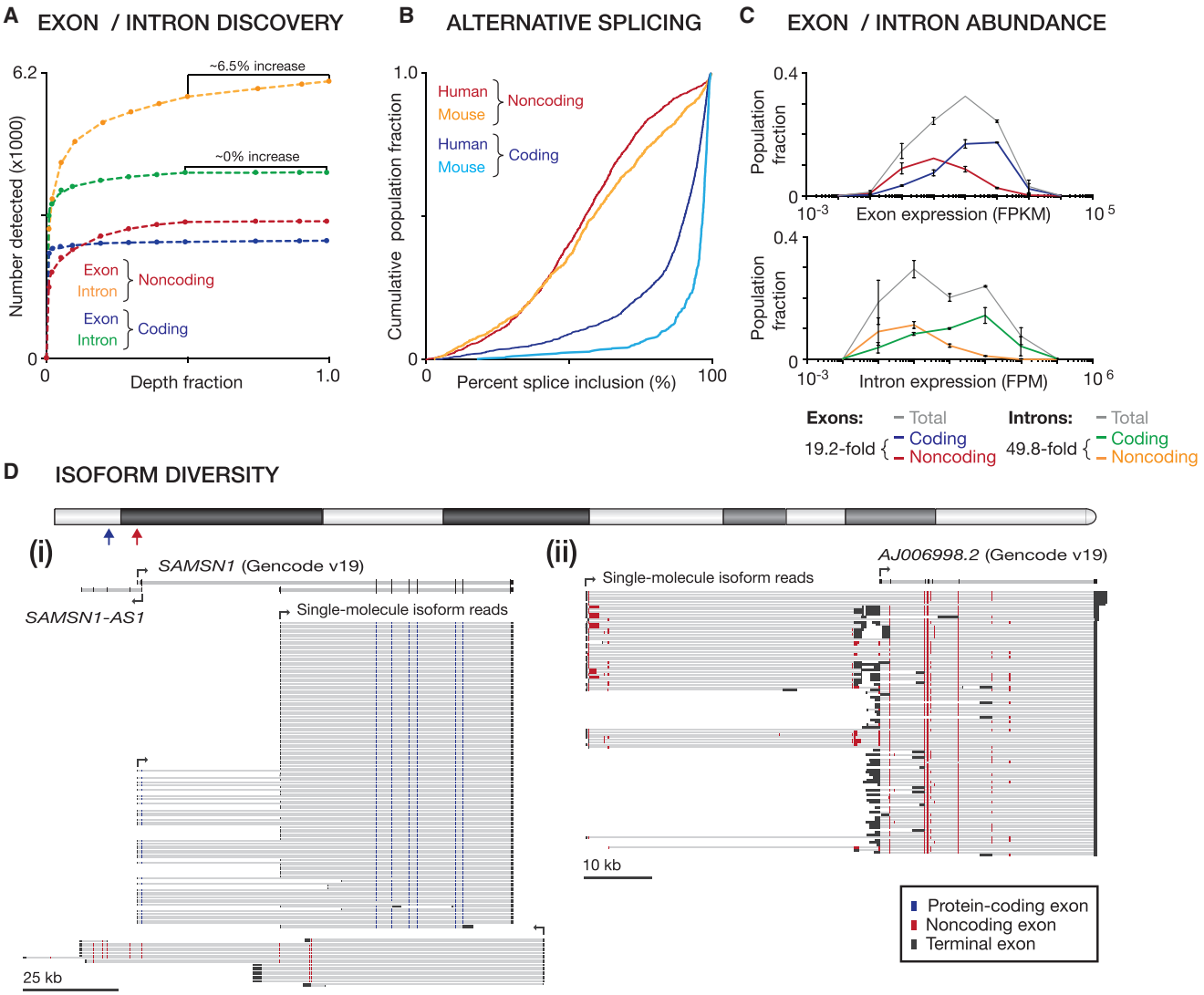


Figure 2. Universal Alternative Splicing of Noncoding Exons

(A) Discovery-saturation curves show the rate of detection for unique protein-coding/noncoding introns and unique internal exons (from the single-molecule Hsa21 transcriptome) relative to short-read sequencing depth. Depth fraction is relative to a combined pool of ~450 million short-read alignments to Hsa21 from testis, brain, and kidney.

(B) Cumulative frequency distributions show percent splice inclusion (PSI) scores for protein-coding/noncoding internal exons in human and mouse tissues.

(C) Binned frequency distributions show abundances of protein-coding/noncoding internal exons and introns in human testis (mean ± SD, n = 3). Gray line shows total exon/intron population. Median fold difference between coding/noncoding populations is shown below.

(D) Illustrative examples of isoforms resolved by single-molecule RNA CaptureSeq in human tissues. Annotated transcripts (Gencode v19) and mapped single-molecule isoform reads are shown at two loci: (i) the protein-coding gene *SAMSN1* and the noncoding antisense RNA *SAMSN1-AS1*; and (ii) the lncRNA AJ006998.2. Internal exons are identified as protein-coding (blue) or noncoding (red), which includes untranslated exons at protein-coding loci. Terminal exons (black) were excluded from analyses.

alternative splicing, we performed short-read RNA CaptureSeq across mouse genome regions syntenic to Hsa21 (located on mouse chromosomes 10, 16, and 17; Mouse Genome Sequencing Consortium et al., 2002). We obtained transcriptional profiles of equivalent depth to human samples within matched mouse tissues (Table S7), enabling comparison of the two transcriptomes at high resolution (STAR Methods).

As for Hsa21, the syntenic regions of the mouse genome were pervasively transcribed, largely eroding intergenic desert regions (Figure 3A). Spliced transcripts encompassed 80.5% of targeted bases in mouse, compared with 88.4% on Hsa21 (compare Figures 1D and 3B). In mouse, 25.1% of targeted bases were retained as mature exons, compared with 16.7% in human, with the remaining fraction (55.4% versus 71.7%) removed as introns (compare Figures 1D and 3B). The larger fraction of bases represented in mature exons in mouse (1.5-fold) likely reflects compaction of the mouse genome via accelerated genetic loss (Mouse Genome Sequencing Consortium et al., 2002; Vierstra et al., 2014; Yue et al., 2014) rather than higher gene content, with the mouse transcriptome assembly being somewhat

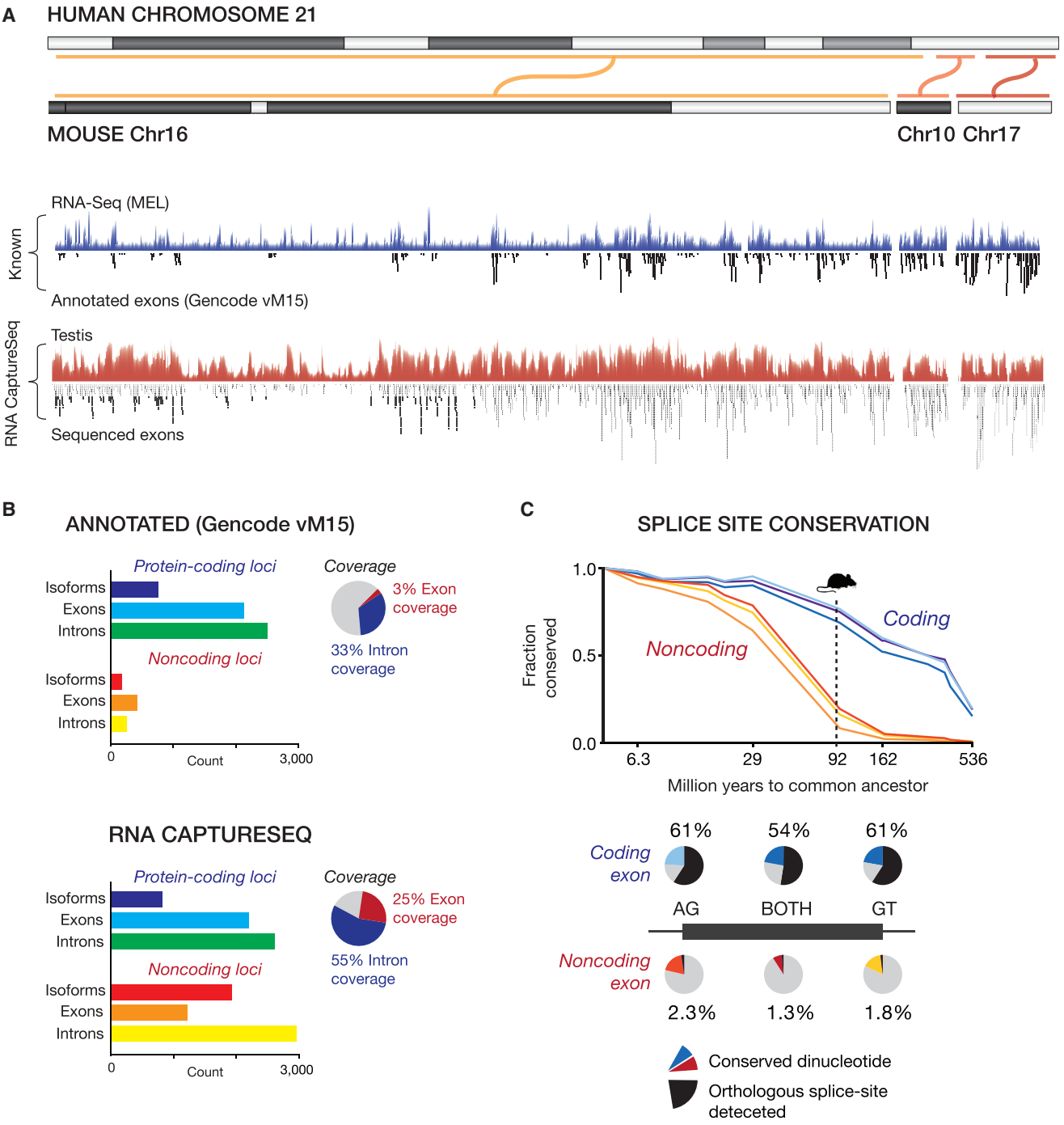


Figure 3. Transcriptional Landscape of Mouse Syntenic Regions and Noncoding Exon Evolution

(A) Transcriptional activity recorded across regions of the mouse genome (Mmu16, Mmu17, Mmu10) syntenic to human chromosome 21 (Hsa21) by short-read RNA CaptureSeq. Normalized coverage (log scale) of short-read alignments is shown for testis (red), with traditional RNA-seq (in MEL cells; navy) shown for comparison. Coverage tracks shown below are all unique internal exons (black) from transcripts in the Gencode (vM15) reference catalog and those assembled from RNA CaptureSeq data (testis, brain, and kidney combined).

(B) Bar charts show the number of protein-coding/noncoding unique internal exons, unique canonical introns, and unique transcript isoforms in Gencode (vM15) annotations (upper) or assembled from RNA CaptureSeq (lower; brain, kidney, and testis combined). Pie charts indicate the proportion of captured bases that are exonic, intronic, or silent.

(C) Upper: the fraction of coding/noncoding splice-site dinucleotides (AG/GT/both) detected on Hsa21 that are conserved in other vertebrate genomes (arranged by million years to common ancestor). Lower: pie charts indicate the proportion of Hsa21 exons with splice-site dinucleotides that are conserved in the mouse genome (red/blue) and the proportion for which an equivalent splice site could also be detected in mouse RNA CaptureSeq libraries (black).

smaller than human (77%, based on unique internal exon count; Table S8).

Although almost all protein-coding genes on Hsa21 have mouse orthologs, we observed a higher frequency of alternative splicing among human genes than their mouse counterparts: 69% of human protein-coding exons were classified as alternative (PSI < 95%) compared with just 31% in mouse (Figure 2B). As a result of this greater splicing diversity, human protein-coding genes had more internal exons (17%), introns (39%), and isoforms (65%) than their corresponding mouse orthologs (Table S8).

The *DYRK1A* gene, a leading candidate for autism and trisomy-21 phenotypes (Becker et al., 2014), provides an illustrative example of the increased splicing diversity distinguishing human genes from their mouse orthologs. While we found no novel exons or isoforms to the *Dyrk1A* gene in the mouse, we identified six novel internal exons in the human brain (in addition to all 13 currently annotated *DYRK1A* exons; Figures 4A and 4B). Extensive alternative splicing generated at least 11 novel *DYRK1A* isoforms, of which 10 comprise noncoding variants and one encodes a novel ORF with an N-terminal modification to the DYRK1A protein (Figures 4A and 4B; interestingly, an analogous N-terminal modification regulates subcellular localization of the DYRK4 paralog [Papadopoulos et al., 2011]).

The majority of novel exons discovered in the mouse syntenic regions were noncoding (Figure 3B), and these were similarly subject to near-universal alternative splicing (Figure 2B). This indicates that the size and structure of the human and mouse transcriptomes are largely comparable, with each harboring large noncoding RNA populations that exhibit prolific alternative splicing.

Despite this similarity, we found that individual lncRNAs were largely divergent between the two lineages (STAR Methods). Nineteen percent of lncRNA splice-acceptor and 16% of splice-donor dinucleotides (AG/GT) were conserved between the human and mouse genomes. However, a corresponding splice site was found in mouse for fewer than 2% of human sites, implying that noncoding exon orthologs are rare (Figure 3C). Although they were poorly conserved relative to their protein-coding counterparts, noncoding exons exhibited internal sequence constraint (base on Vertebrate PhlyoP scores) comparable with annotated DNase hypersensitive or transcription factor binding sites, with flanking splice sites showing a further, though relatively modest, conservation enrichment (Figures S12A and S12B). These data indicate that, while similar in size and structure, human and mouse noncoding RNA populations are largely distinct, echoing the reported divergence of regulatory elements between mouse and human genomes (Vierstra et al., 2014; Villar et al., 2015).

Human Chromosome 21 Expression and Splicing in Mouse Cells

The Tc1 mouse strain is a model for trisomy-21 that carries a stable copy of Hsa21 (Yu et al., 2010). The Tc1 mouse has also been used to compare the human and mouse transcriptomes, enabling the regulatory contributions of human *cis* elements and mouse *trans*-acting factors to be distinguished (Barbosa-Morais et al., 2012; Wilson et al., 2008). To investigate the regulation of transcriptome diversity, we performed short-read RNA CaptureSeq, targeting both Hsa21 and its syntenic mouse genome regions, in matched tissues of the Tc1 mouse (Table S9 and STAR Methods).

Most notably, the splicing profiles of genes encoded on Hsa21 were recapitulated in the Tc1 mouse as for human tissues, rather their mouse orthologs, where these were divergent. The *DYRK1A* gene again provides an illustrative example, with human-specific splice-site selection and quantitative exon usage faithfully recapitulated on Hsa21 in Tc1 samples (Figures 4A–4D). Globally, 87% of human-specific splice sites distinguishing human and mouse orthologs were also detected on Hsa21 in the Tc1 mouse (compared with 82% for shared splice sites; Figure S13A). Similarly, the alternative splicing frequency of human exons remained 2.1-fold higher than for mouse orthologs, and 88% of sites classified as alternative in human were also classified as alternative in Tc1 (compared with 39% in mouse; Figures S13B and S13C). When correlated according to the PSI profiles across all orthologous splice sites, we found that human, mouse, and Tc1 samples clustered according to chromosomal, rather than organismal, origin (Figure S14C). Together these data confirm the central importance of local *cis* sequence elements in defining exon boundaries and alternative exon inclusion.

The structure and splicing of lncRNAs encoded on Hsa21 was also precisely recapitulated in the Tc1 mouse, despite the absence of mouse orthologs. The majority (85%; Figure 4E) of human noncoding splice sites were also detected in Tc1 and noncoding exons were again near universally alternatively spliced (98%; Figures 4A–4C). Furthermore, quantitative splice-site usage and relative noncoding isoform abundance was maintained as for human tissues (Figure 4D), indicating that the local Hsa21 sequence is sufficient to establish splice-site position and regulate the proportional inclusion of noncoding exons by alternative splicing.

In contrast to splicing, we observed a global deregulation of expression in the Tc1 mouse, as assessed by principal component analysis or rank-correlation clustering (Figures S14A and S14B). This effect is best illustrated at intergenic regions flanking the *NCAM2* locus, where numerous lncRNA genes that are silenced in the human brain become deregulated, resulting in aberrant expression in the Tc1 mouse brain (Figure S15A). In fact, the expression of human protein-coding genes encoded on Hsa21 was more similar to expression of their mouse orthologs in corresponding mouse tissues than to the expression of the same human genes encoded on Hsa21 in the Tc1 mouse (Figures S14A and S14B). This deregulation was restricted to the human chromosome, with tissue-specific expression profiles still maintained across syntenic regions of the mouse genome in Tc1 mouse (Figure S15B).

This analysis appears to highlight a distinction in the evolution of expression and splicing regulation. The deregulation of human gene expression in mouse cells is consistent with the reported divergence of human and mouse regulatory elements, including enhancers and transcription factor binding sites (Vierstra et al., 2014; Villar et al., 2015). In contrast, splicing profiles were largely recapitulated on Hsa21 in the Tc1 mouse, as has been observed previously for protein-coding exons (Barbosa-Morais et al., 2012). This implies that splicing is largely regulated by *cis* elements in the local chromosome sequence (Barash et al., 2010)

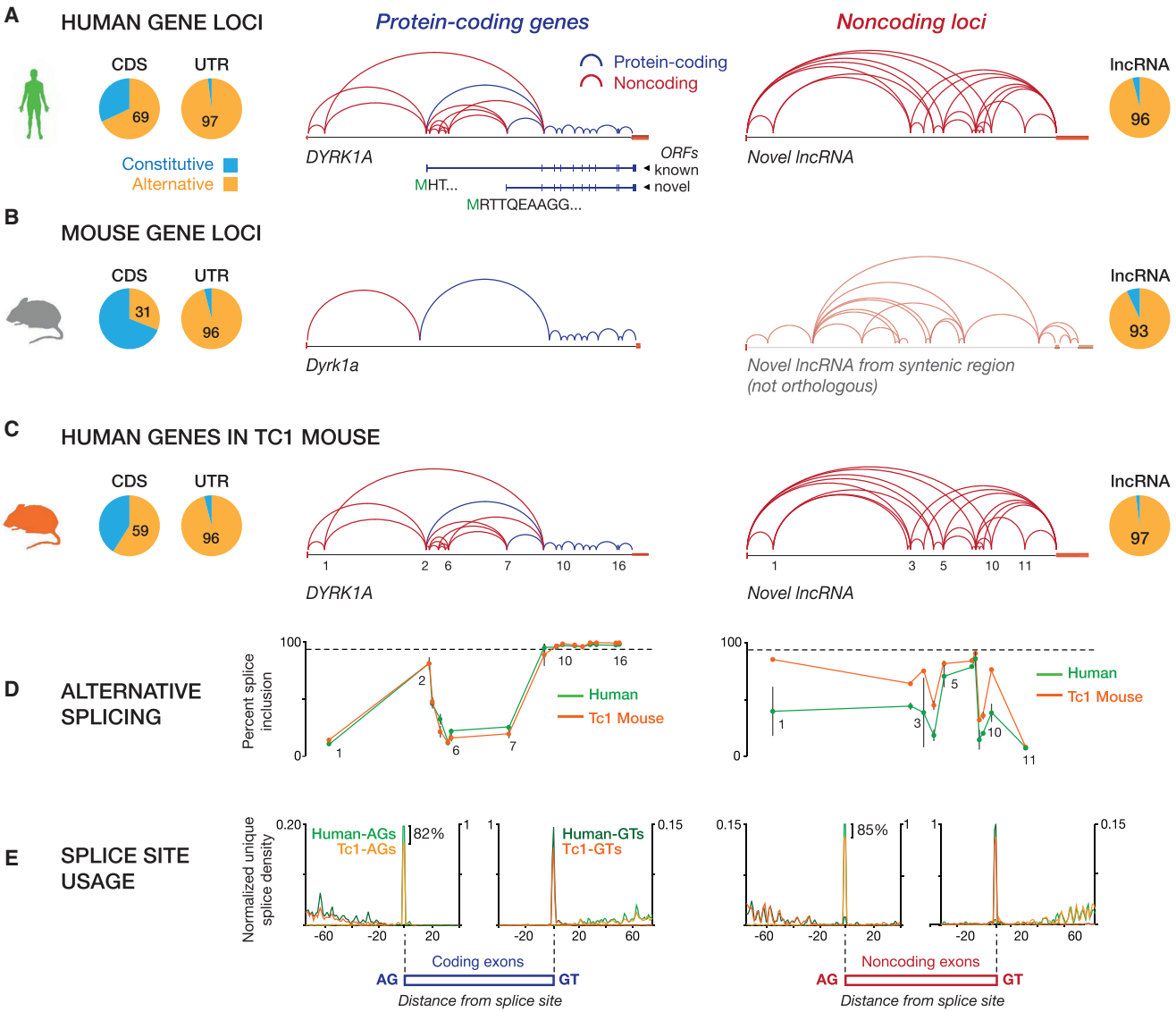


Figure 4. Splicing of Human Genes in the Tc1 Mouse

(A–C) Exon-intron structures assembled for the protein-coding gene *DYRK1A* and a novel lncRNA locus from human Hsa21 (A), mouse syntenic regions (B), and on Hsa21 in the Tc1 mouse strain (C). Pie charts indicate the proportion of unique internal exons classified as constitutive (PSI > 95) or alternative, for protein-coding exons (CDS) and untranslated exons at coding loci (UTR) and lncRNA exons.

(D) Plots show relative isoform abundance of *DYRK1A* and novel lncRNA loci in human and Tc1 mouse, as indicated by PSI values for each internal exon (aligned to exons in gene models above). PSI values shown for *DYRK1A* are measured from human brain libraries (mean ± SD, n = 2) and lncRNA from testis libraries (n = 3).

(E) Density plots show global concordance of unique splice junction selection between human Hsa21 (mean ± SD, n = 2) and Tc1-Hsa21 libraries (n = 3) for human exons. Density values are normalized relative to human Hsa21 libraries.

and can be correctly interpreted by the mouse spliceosome. Our findings imply that the splicing code is so highly conserved that human-specific exons and noncoding RNAs without orthologs in mouse are correctly spliced. This demonstrates deep conservation of the lexicon that governs splicing, even while the isoforms produced undergo rapid diversification and turnover.

DISCUSSION

To overcome the expression-dependent bias in RNA-seq, which impedes discovery and characterization of low-abundance transcripts (Hardwick et al., 2016), we performed targeted RNA-seq across human chromosome 21. The combination of single-molecule and short-read RNA CaptureSeq enabled accurate resolution of isoform diversity and quantitative analysis of alternative splicing at unprecedented depth.

Noncoding loci are, contrary to the impression from more shallow surveys (Cabili et al., 2011; Derrien et al., 2012), enriched for alternative splicing, with noncoding exons being near universally classified as alternative. This finding is consistent with previous reports of lower splicing efficiency and U2AF65 occupancy among lncRNAs than mRNAs, features

independently correlated with heightened alternative over constitutive splicing (Melé et al., 2017; Mukherjee et al., 2017; Tilgner et al., 2012).

Therefore, while protein-coding genes are constrained by the requirement to maintain an ORF, it appears that no similar constraint is imposed on noncoding RNA. This suggests that noncoding exons are functionally modular, operating as discrete cassettes that are recombined with maximum flexibility. One can envision a scenario where individual noncoding exons interact independently with other biomolecules (proteins, RNAs, and/or DNA-motifs), organizing these around the scaffold of a noncoding transcript. In this way, alternative isoforms could assemble different collections of binding partners to dynamically regulate cellular processes. The distinction between protein-coding and noncoding RNA was also evident when comparing exons within ORFs with untranslated regions of coding loci (located in 5′ or 3′ UTRs, or specific to noncoding isoforms), implying similar modularity in the functional architecture of untranslated regions.

Low expression is often cited as evidence against the functional relevance of novel transcripts, such as the wide variety of rare noncoding isoforms identified in our survey. However, while weakly expressed genes/isoforms are unlikely to fulfill structural or metabolic functions, there are precedents for these fulfilling regulatory roles. For example, Marinov et al. (2014) found the mRNAs of expressed transcription factors to be present, on average, at just ~3 copies per cell within a homogeneous cell line (GM12878). Single-cell studies routinely reveal rare cell types within human tissues on the basis of their unique gene expression profiles. These may be represented by just a few cells within a community of hundreds or thousands of cells (Grun et al., 2015; La Manno et al., 2016; Muraro et al., 2016). Given that mRNAs encoding regulatory molecules such as transcription factors are expressed at only a few copies per cell, the regulatory factors that define the identity of rare cell types are expected to be present at very low frequencies in whole-tissue transcriptomes. It would be premature, therefore, to discount the relevance of any transcript on the basis of low abundance alone.

Moreover, while not every rare isoform is necessarily important, and the promiscuous splicing we observed might simply reflect a lack of selective pressure, noncoding RNAs collectively form a large reservoir of transcriptional diversity from which molecular innovations might evolve and new genes may be born (Kaessmann, 2010; Toll-Riera et al., 2009; Wu et al., 2011; Xie et al., 2012). The use of alternative splicing to generate noncoding transcriptional diversity and, thereby, drive gene evolution is consistent with the divergence of noncoding exons reported here. By contrast, the splicing code that governs such transcriptional diversity remains closely conserved between human and mouse.

Despite concerted efforts over the past decade, we are yet to achieve a complete census of human gene expression. Even our use of targeted single-molecule RNA-seq was insufficient to resolve the full complement of noncoding isoforms encoded on Hsa21. Instead, we found a seemingly limitless diversity of noncoding isoforms. Given the range of combinatorial possibilities, we suggest that the noncoding RNA population may be inherently plastic, and that there does not exist a finite list of noncoding isoforms that can be feasibly cataloged.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human Samples
  - Mouse Samples
  - K562 Cells
- METHOD DETAILS
  - Total RNA Extraction
  - Capture Enrichment
  - PacBio SMRTbell Library Preparation and Single-Molecule Sequencing
  - Short-Read (Illumina) Sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Validation of Hsa21 RNA CaptureSeq in K562 Cells
  - Single-Molecule Hsa21 Transcriptome Survey
  - Short-Read Hsa21 RNA CaptureSeq in Human Tissues
  - Discovery Saturation Curves
  - Percent Splice Inclusion (PSI) Values
  - Comparison of Human and Mouse Transcriptomes
  - Conservation Analyses
  - Analysis of Hsa21 Splicing and Expression in the Tc1 Mouse
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes 15 figures and 10 tables and can be found with this article online at https://doi.org/10.1016/j.cels.2017.12.005.

ACKNOWLEDGMENTS

The authors acknowledge the following funding sources: an Australian National Health and Medical Research Council (NHMRC) Project grant (APP1062106 to T.R.M.), NHMRC Australia Fellowship (631668 to J.S.M.), an NHMRC Early Career Fellowship (APP1072662 to M.B.C.), an EMBO Long Term Fellowship (ALTF 864-2013 to M.B.C.), the Australian Research Council (Special Research Initiative in Stem Cell Science to L.K.N.), a Cancer Institute NSW Early Career Fellowship (2018/ECF013 to I.W.D.), and the generous support of the Paramor family (to T.R.M.). The contents of the published material are solely the responsibility of the administering institution, a participating institution, or individual authors and do not reflect the views of NHMRC or ARC. The authors thank the ENCODE consortium for the provision of data; data were employed in accordance with the data-release policy.

AUTHOR CONTRIBUTIONS

T.R.M. and M.E.B. conceived the project and designed experiments, with advice from J.S.M., L.K.N., and M.E.D. M.E.B., M.B.C., J.C., and J.B. performed capture enrichment and library preparations for short-read sequencing. M.E.B. and J.B. performed PCR validations or assembled transcripts. J.B. and T.H. performed long-read sequencing, overseen by T.A.C. I.W.D. and E.T. performed bioinformatics analyses. I.W.D. and T.R.M. prepared the manuscript with support from M.E.B., J.B., M.B.C., L.K.N., and J.S.M. T.R.M., M.E.D., L.K.N., and J.S.M. provided funding.

DECLARATION OF INTERESTS

T.R.M. was a recipient of a Roche Discovery Agreement (2014). M.B.C. has received research support from Roche/Nimblegen for unrelated research projects. E.T., T.H., and T.A.C. are employees of Pacific Biosciences.

Received: May 13, 2017
Revised: October 18, 2017
Accepted: December 8, 2017
Published: January 24, 2018

REFERENCES

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53–59.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science *338*, 1587–1593.

Becker, W., Soppa, U., and Tejedor, F.J. (2014). DYRK1A: a potential drug target for multiple Down syndrome neuropathologies. CNS Neurol. Disord. Drug Targets *13*, 26–33.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., et al. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell *116*, 499–509.

Clark, M.B., Amaral, P.P., Schlesinger, F.J., Dinger, M.E., Taft, R.J., Rinn, J.L., Ponting, C.P., Stadler, P.F., Morris, K.V., Morillon, A., et al. (2011). The reality of pervasive transcription. PLoS Biol. *9*, e1000625.

Clark, M.B., Mercer, T.R., Bussotti, G., Leonardi, T., Haynes, K.R., Crawford, J., Brunck, M.E., Cao, K.-A.L., Thomas, G.P., Chen, W.Y., et al. (2015). Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. Nat. Methods *12*, 339–342.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., et al. (2016). A survey of best practices for RNA-seq data analysis. Genome Biol. *17*, 13.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Dierssen, M. (2012). Down syndrome: the brain in trisomic mode. Nat. Rev. Neurosci. *13*, 844–858.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. *14*, R70.

Grun, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature *525*, 251–255.

Hardwick, S.A., Chen, W.Y., Wong, T., Deveson, I.W., Blackburn, J., Andersen, S.B., Nielsen, L.K., Mattick, J.S., and Mercer, T.R. (2016). Spliced synthetic genes as internal controls in RNA sequencing experiments. Nat. Methods *13*, 792–798.

Hon, C.C., Ramilowski, J.A., Harshbarger, J., Bertin, N., Rackham, O.J.L., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T.M., Severin, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5′ ends. Nature *543*, 199–204.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., and Zhao, S. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. *47*, 199–208.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome Res. *21*, 1543–1551.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res. *20*, 1313–1326.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., et al. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res. *14*, 331–342.

Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S., and Gingeras, T.R. (2005). Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res. *15*, 987–997.

La Manno, G., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L.E., Stott, S.R.W., Toledo, E.M., Villaescusa, J.C., et al. (2016). Molecular diversity of midbrain development in mouse, human, and stem cells. Cell *167*, 566–580.e19.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., 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.

Letourneau, A., Santoni, F.A., Bonilla, X., Sailani, M.R., Gonzalez, D., Kind, J., Chevalier, C., Thurman, R., Sandstrom, R.S., Hibaoui, Y., et al. (2014). Domains of genome-wide gene expression dysregulation in Down's syndrome. Nature *508*, 345–350.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map (SAM) format and SAMtools. Bioinformatics *25*, 2078–2079.

Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. *24*, 496–510.

Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. Nat. Rev. Genet. *12*, 671–682.

Melé, M., Mattioli, K., Mallard, W., Shechner, D.M., Gerhardinger, C., and Rinn, J.L. (2017). Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. Genome Res. *27*, 27–37.

Mercer, T.R., Clark, M.B., Crawford, J., Brunck, M.E., Gerhardt, D.J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. Nat. Protoc. *9*, 989–1009.

Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A., Mattick, J.S., and Rinn, J.L. (2011). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat. Biotechnol. *30*, 99–104.

Muraro, M.J., Dharmadhikari, G., Grun, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M.A., Carlotti, F., de Koning, E.J.P., and van Oudenaarden, A. (2016). A single-cell transcriptome atlas of the human pancreas. Cell Syst. *3*, 385–394.e3.

Mukherjee, N., Calviello, L., Hirsekorn, A., de Pretis, S., Pelizzola, M., and Ohler, U. (2017). Integrative classification of human coding and noncoding genes through RNA metabolism profiles. Nat. Struc. Mol. Biol. *24*, 86–96.

Papadopoulos, C., Arato, K., Lilienthal, E., Zerweck, J., Schutkowski, M., Chatain, N., Müller-Newen, G., Becker, W., and de la Luna, S. (2011). Splice variants of the dual specificity tyrosine phosphorylation-regulated kinase 4 (DYRK4) differ in their subcellular localization and catalytic activity. J. Biol. Chem. *286*, 5494–5505.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. *31*, 1009–1014.

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., et al. (2013). Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep. *3*, 2179–2190.

Tilgner, H., Jahanbani, F., Blauwkamp, T., Moshrefi, A., Jaeger, E., Chen, F., Harel, I., Bustamante, C.D., Rasmussen, M., and Snyder, M.P. (2015). Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nat. Biotechnol. *33*, 736–742.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. *22*, 1616–1625.

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., and Albá, M.M. (2009). Origin of primate orphan genes: a comparative genomics approach. Mol. Biol. Evol. *26*, 603–612.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R.S., Stehling-Sun, S., Sabo, P.J., Byron, R., Humbert, R., et al. (2014). Mouse regulatory DNA landscapes reveal global principles of *cis*-regulatory evolution. Science *346*, 1007–1012.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554–566.

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P., and Li, W. (2013). CPAT: coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucl. Acids Res. *41*, e74.

Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., 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.

Wilson, M.D., Barbosa-Morais, N.L., Schmidt, D., Conboy, C.M., Vanes, L., Tybulewicz, V.L.J., Fisher, E.M.C., Tavaré, S., and Odom, D.T. (2008). Species-specific transcription in mice carrying human chromosome 21. Science *322*, 434–438.

Wu, D.D., Irwin, D.M., and Zhang, Y.P. (2011). De novo origin of human protein-coding genes. PLoS Genet. *7*, e1002379.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics *21*, 1859–1875.

Xie, C., Zhang, Y.E., Chen, J.Y., Liu, C.J., Zhou, W.Z., Li, Y., Zhang, M., Zhang, R., Wei, L., and Li, C.Y. (2012). Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. PLoS Genet. *8*, e1002942.

Yu, T., Li, Z., Jia, Z., Clapcote, S.J., Liu, C., Li, S., Asrar, S., Pao, A., Chen, R., Fan, N., et al. (2010). A mouse model of Down syndrome trisomic for all human chromosome 21 syntenic regions. Hum. Mol. Genet. *19*, 2780–2791.

You, B.-H., Yoon, S.-H., and Nam, J.-W. (2017). High-confidence coding and noncoding transcriptome maps. Genome Res. *27*, 1050–1062.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature *515*, 355–364.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Human RNA Survey Panel: total RNA from healthy adult tissue (brain, kidney, and testis)	Ambion	AM6000
Critical Commercial Assays		
SeqCap EZ Enrichment Kit (custom probe design)	Roche-Nimblegen	06266312001
TruSeq Stranded mRNA Library Preparation	Illumina	RS-122-2101
Iso-Seq SMRTBell Template Preparation	Pacific Biosciences	http://www.pacb.com/blog/intro-to-iso-seq-method-full-leng/
Deposited Data		
Raw and analyzed data	This paper	GEO:GSE99637
Human reference genome (hg19)	UCSC Browser	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/
Mouse reference genome (mm10)	UCSC Browser	http://hgdownload.soe.ucsc.edu/goldenPath/mm10/
Gencode transcriptome annotations (v19 and vM15)	Gencode Project	https://www.gencodegenes.org/
MiTranscriptome annotation	Iyer et al., 2015	http://mitranscriptome.org/
FANTOM5 transcriptome annotation	Hon et al., 2017	http://fantom.gsc.riken.jp/5/
Experimental Models: Cell Lines		
K562 cells	ATCC	https://www.atcc.org/products/all/CCL-243.aspx
Experimental Models: Organisms/Strains		
WT mouse (C57BL/6J x 129S8/SvEv)	Jackson Laboratory, Maine, USA	-
TC1 mouse strain (Yu et al., 2010)	Jackson Laboratory, Maine, USA	-
Oligonucleotides		
Custom primers for splice-junction validation	IDT	Table S2
Software and Algorithms		
STAR (2.4.2a)	Dobin et al., 2013	https://github.com/alexdobin/STAR
Samtools (1.5)	Li et al., 2009	http://samtools.sourceforge.net/
Bedtools (2.25.0)	Quinlan lab, University of Utah	http://bedtools.readthedocs.io/en/latest/index.html
TrimGalore (0.4.1)	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/
StringTie (1.3.3b)	Pertea et al., 2015	https://ccb.jhu.edu/software/stringtie/
Cuffcompare/Cuffmerge (2.2.1)	Trapnell et al., 2012	http://cole-trapnell-lab.github.io/cufflinks/
GMAP	Wu and Watanabe, 2005	http://research-pub.gene.com/gmap/
LiftOver	UCSC Browser	http://genome.ucsc.edu/cgi-bin/hgLiftOver
CPAT (1.2.2)	Wang et al., 2013	http://lilab.research.bcm.edu/cpat/
RSEM (1.2.30)	Li and Dewey, 2011	https://deweylab.github.io/RSEM/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Tim R. Mercer (t.mercer@garvan.org.au).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Samples

Total RNA samples from healthy adult tissues were acquired from a commercial vendor (Ambion Human RNA Survey Panel). Ambion certifies that all of human-derived materials have been prepared from tissue obtained with consent from a fully informed donor or a member of the donor's family. Two replicate samples for human brain and kidney, and three replicates for testis were analyzed.

Mouse Samples

All animals were handled, housed, and used in the experiments in accordance with protocols approved by the Animal Ethics Committee of The University of Queensland, Brisbane, Australia. Three male mice (C57BL/6J x 129S8/SvEv) and 3 Tc1 female breeders were imported from the Jackson Laboratory, Maine, USA. 10 x F1 pups produced from paired mating were toed for identification and genotyping. Genotyping was performed on gDNA extracted from toe tissues using primers specific to human and mouse *JAM2* genes. Three Tc1 F1 males and 3 WT littermates were sacrificed between 6 and 9 weeks of age, and testis, kidney and brain tissues were immediately harvested in TRIzol reagent (Ambion), snap-frozen on dry ice and stored at -80°C until processing.

K562 Cells

The human lymphoblast cell line, K562, was obtained from the American Type Culture Centre (ATCC). Cells were not independently verified or tested for mycoplasma. Cells were cultured according to Coriell Institute's growth protocols and standards. Briefly, K562 cells were cultured in RPMI 1640 medium (Gibco) supplemented with 10% FBS at 37°C under 5% CO_2. Cells (passage 9 or 10) were grown to ~80% confluence before RNA extraction.

METHOD DETAILS

Total RNA Extraction

Mouse Samples

Tissue samples (brain, kidney and testis), harvested in trizol and snap-frozen (see above), were defrosted on ice and 125 mg MicroBeads (MoBio) were added to sample tubes. Tissues and cells were disrupted by 2 cycles of 45 sec at 6500 rpm on a tissue homogenizer (Precellys) connected to a cooling system (Cryolis) keeping the samples temperature < 4°C during the cell lysis procedure. Bead-free cell lysates were transferred to a new 1.5 mL tube and RNA was immediately harvested, as above. Brain RNA was extracted a second time in TRIzol due to high lipid contamination present after the first extraction.

K562 Cells

Total RNA was extracted in 3mL TRIzol reagent (Ambion). Extraction of the aqueous phase was performed using chlorophorm and RNA precipitation was achieved with isopropanol. RNA pellets were washed in ice-cold 75% EtOH and allowed to dry before suspension in DEPC-treated H_2O.

Capture Enrichment

Oligonucleotide probes targeting the entire non-repetitive portion of Hsa21 (hg19) and its syntenic regions in the mouse (mm10) genome, obtained using the UCSC Genome Browser liftOver tool, were designed and synthesized by Roche/NimbleGen. In total, the array targeted 25.3 Mb encompassing 0.87% of the human genome and 20.3 Mb encompassing 0.77% of the mouse genome.

Total RNA samples were assessed for potential gDNA contamination by PCR and accordingly treated with Turbo DNAse (Ambion). Samples were spiked with ERCC RNA spike-in controls (Jiang et al., 2011) to a final 1% concentration and rRNA-depleted using the Ribo-Zero rRNA removal magnetic kit (Epicentre). cDNA libraries were prepared from rRNA-depleted samples using the Illumina TruSeq stranded mRNA low-template kit. Pre-capture LMPCR amplified libraries were purified using AMPure XP beads (Beckman Coulter Genomics), and successful library preparation was validated using a DNA 1000 kit (Agilent) on a Bioanalyser. Average library sizes were ~280-310 bp.

Capture hybridization was performed in 96-well plates, with hybridization probes incubated at 47°C for 64-77 h. Post-capture LMPCR amplification was performed for 17 cycles. Amplified post-capture libraries were purified using AMPure XP beads, and validated using a DNA 1000 kit on a Bioanalyzer. Samples that retained unincorporated primers were cleaned a second time using a QIAquick PCR purification kit (Qiagen), and successful primer removal was validated on a Bioanalyzer. For a detailed protocol see (Mercer et al., 2014).

PacBio SMRTbell Library Preparation and Single-Molecule Sequencing

A total of 4.5 μg of captured full-length cDNA was subjected to size fractionation using the Sage Science BluePippin system into four size bins (0-2kb, 1-3kb, 3-6kb and 4-10kb). Eluted size fractions were subsequently re-amplified and purified with AMPure PB beads. Size distributions of the fractions were checked for quality on a 2100 BioAnalyzer (Agilent).

Approximately 10 μg of purified amplicon was taken into Iso-Seq SMRTBell library preparation (https://pacbio.secure.force.com/SamplePrep). Separate SMRTbell libraries were generated for each of the four size bins (Table S10). Two of the SMRTBell libraries

(3-6kb and 4-10kb) were size-selected again using the Sage Science BluePippen system to remove trace amounts of small inserts. A total of 24 SMRT Cells (6 cells for each of the size-selected SMRTBell library) were sequenced on the PacBio *RS II* platform using P6-C4 chemistry with 3 to 4 hour acquisition time.

Short-Read (Illumina) Sequencing

Short-read sequencing libraries were prepared with the Illumina TruSeq stranded mRNA low-template kit. Libraries were sequenced at the Garvan Institute (Sydney, NSW, Australia) on the Illumina HiSeq 2000 or Illumina HiSeq 2500 platforms, generating 2 x 101bp and 2 x 125bp paired-end sequencing, respectively.

QUANTIFICATION AND STATISTICAL ANALYSIS

Validation of Hsa21 RNA CaptureSeq in K562 Cells

To establish the Hsa21 CaptureSeq approach we first carried out short-read sequencing (Illumina HiSeq 2000) on Hsa21-enriched libraries from the human K562 cell-type (n = 3) and compared these to K562 samples analyzed in parallel by conventional RNA-Seq (n = 3; Table S1).

Sequencing libraries were trimmed using TrimGalore (https://www.bioinformatics.babraham.ac.uk-/index.html; default parameters), then aligned to a combined index of the hg19 reference genome and ERCC spike-in sequences. Alignment was performed by STAR (Dobin et al., 2013) with the following custom parameters:

–twopassMode Basic –outFilterIntronMotifs RemoveNoncanonical –alignIntronMin 20 –alignIntronMax 500000 \

On-target alignment rates were calculated by determining the number of uniquely mapped reads (MapQ=255) within Hsa21, as a fraction of all uniquely mapped reads across hg19. The average fold-increase in on-target rate for CaptureSeq samples, relative to non-captured samples, provides an estimate of the coverage enrichment achieved by capture (Table S1).

Relative abundances (in FPKM) of ERCC RNA spike-ins were determined using RSEM (Li and Dewey, 2011), allowing the accuracy of transcript quantification to be assessed in captured/non-captured samples (Figures S2A and S2B). Progressively decreasing enrichment at very high ERCC concentrations was observed due to saturation of CaptureSeq oligonucleotide baits, but is unlikely to affect transcripts within the physiological range of gene expression (Clark et al., 2015; Mercer et al., 2014).

Single-Molecule Hsa21 Transcriptome Survey

To avoid potential artifacts associated with *de novo* transcript assembly, we analyzed Hsa21-enriched cDNA from human testis by deep single-molecule PacBio sequencing. Testis was selected because it provides a broad survey of gene-content encoded on Hsa21 (Soumillon et al., 2013).

The PacBio Iso-Seq "classify" protocol was used to generate full-length, non-chimeric (FLNC) reads (https://github.com/PacificBiosciences/cDNA_primer). Briefly, for each sequencing ZMW, a circular consensus sequence (CCS) was generated. Each CCS sequence was identified as full-length, non-chimeric if (1) both the 5' and 3' cDNA primer and the polyA tail preceding the 3' primer was identified at the two ends; and (2) the sequence does not contain cDNA primers or polyA tails in the middle of the sequence (indication of library artifacts). Because FLNC reads are supported by the presence of a polyA tail, they are considered to represent mature transcript isoforms. Resolution of both sequencing primers ensures that FLNC reads have been sequenced in completeness. However, the possibility of degradation due to the use of template switching without 5' cap capture means that some FLNC reads may not be biologically full-length.

FLNC reads were mapped to the human reference genome (hg19) using GMAP (Wu and Watanabe, 2005) and reads that mapped to Hsa21 with maximum mapping confidence (MapQ=40) were retained. In total, 387,029 reads aligned to Hsa21, constituting 910 Mb of usable transcript sequence concentrated in ~1.5% of the genome (Figures S3A and S3B; Tables S3 and 10).

To retain only high-quality multi-exonic single-molecule transcripts, we next discarded: (1) transcripts with < 3 exons; (2) transcripts that contained one or more non-canonical introns (based on AG/GT splice sites); (3) any transcript that was entirely contained as a partial fragment of a longer transcript or an annotated transcript (based on internal exon-intron-exon chain); (4) any transcript with the following CuffCompare classifier: e, i, o, p, r, s (Trapnell et al., 2012). After filtering, we retrieved 101,478 full-length multi-exonic transcripts (Figure S3B; Table S3). To our knowledge, this represents the deepest single-molecule transcriptome survey of a large genome region ever performed.

Transcripts were next assessed for overlap with known protein-coding genes. Transcripts that shared exonic overlap (on the same strand) with a known gene but not an identical internal exon-intron-exon chain were considered to represent novel isoforms of known genes. Redundant transcripts (sharing identical internal exon-intron-exon chains) were collapsed.

These were classified as protein-coding or noncoding via the Coding Potential Assessment Tool (Wang et al., 2013; longest ORF $\geq$ 100 codons, coding potential score $\geq$ 0.99). For coding transcripts we used TransDecoder to predict ORFs. Redundant ORFs and fragments fully contained within longer ORFs were discarded. ORFs were overlapped with full-length transcripts to distinguish coding exons from noncoding exons belonging to coding loci, including spliced UTR exons and exons incorporated exclusively into noncoding isoform variants at coding loci.

Transcripts that shared no exonic overlap with a known protein-coding gene (Gencode v19) and lacking a predicted ORF (CPAT; longest ORF < 100 codons, coding potential score < 0.99) were classified as lncRNA transcripts. Redundant transcripts (sharing identical internal exon-intron-exon chains) were collapsed.

Short-Read Hsa21 RNA CaptureSeq in Human Tissues

To obtain quantitative information about transcripts in our single-molecule Hsa21 profile, short-read sequencing (Illumina HiSeq 2500) was performed on duplicate Hsa21-enriched libraries from brain, kidney and testis. Sequencing libraries were trimmed and aligned as described above (see above). On-target alignments rates were calculated as above (see 3.1) and reads that were uniquely mapped (MapQ=255) within Hsa21 were retained for further analysis. At least 100 million alignments to Hsa21 were obtained for each tissue (Table S4).

Because *de novo* transcript assembly from short sequencing reads may generate incorrect transcript models, especially for low-abundance isoforms (Hardwick et al., 2016), we did not perform transcript assembly. For all analyses of human tissues, the single-molecule PacBio transcriptome described above (see above) was used as a reference, with spliced-short read alignments used to quantitatively evaluate expression and splicing of PacBio transcripts.

The majority of spliced short-read alignment junctions were concordantly mapped to an intron in our single-molecule Hsa21 transcriptome profile (Figure S3B). Likewise, the majority of unique internal exons in our single-molecule Hsa21 transcriptome were correctly detected (both boundaries specified by $\geq$ 3 spliced short-read alignment termini) in at least one tissue (Figure S3C). This concordance with more accurate short-reads (Conesa et al., 2016) suggests that single-molecule isoforms were, in general, correctly aligned to Hsa21, and their internal exon-intron-exon architecture was correctly resolved.

Discovery Saturation Curves

To generate discovery-saturation curves, short-read alignments from brain, kidney and testis were combined, then incrementally subsampled from a maximum of ~450 million alignments. We assessed the detection of introns and exons within our single-molecule Hsa21 transcriptome by short-read alignments at each depth-increment. The analysis was limited to unique canonical introns and unique internal exons (as opposed to terminal exons) since these should be precisely demarcated by short-read alignment junctions at either end. An internal exon from our transcriptome was considered to be detected within a given library if both boundaries were specified by $\geq$ 3 spliced short-reads. An intron from our transcriptome was considered to be detected within a given library if it was spanned by $\geq$ 3 spliced short-read alignment junctions, mapping exactly to its 5' and 3' ends.

Percent Splice Inclusion (PSI) Values

To further assess alternative splicing, we calculated a Percent Splice Inclusion (PSI) score for each internal exon in our Hsa21 transcriptome, as has been done previously (Barbosa-Morais et al., 2012; Tilgner et al., 2015). For each exon, we counted spliced reads that used its 5' splice site ($\mathbf{In_1}$) and 3' splice site ($\mathbf{In_2}$) and spliced reads that spanned/skipped the exon (**Out**). The PSI value for that exon was then calculated as:

$$PSI = \frac{(\mathbf{In1} + \mathbf{In2})/2}{\mathbf{Out} + (\mathbf{In1} + \mathbf{In2})/2} \times 100$$

We confirmed that the occurrence of lower inclusion frequencies among noncoding exons (median PSI = 55.5%) than protein-coding exons (median PSI = 90.5%; Figure 2B) was not simply caused their lower gene expression, since exon-PSI and overall gene expression metrics were independent from one and other (Figure S7). Applying a threshold of PSI > 95% to distinguish constitutive from alternatively spliced exons, as has been done previously (Barbosa-Morais et al., 2012; Tilgner et al., 2015), almost all noncoding exons (97%) would be classified as alternative. We note that the use of a single, essentially arbitrary, PSI threshold is simplistic. However, this difference was indicative of a clear, population-wide depletion of PSI scores among noncoding exons, relative to their protein-coding counterparts, with a clear difference observable at any chosen threshold (Figure 2B). The use of PacBio reads as a scaffold for calculating PSI scores ensures we are working with accurate transcripts (free of potential artifacts from transcript assembly). However, PacBio reads are less suitable for quantitative analyses because (i) size fractionation and heavy PCR amplification during the PacBio library preparation may distort transcript abundances within the population and (ii) saturating read-depth is required for accurate analyses of splicing. Therefore we consider our combined approach to be the most reliable way to assess alternative splicing. PSI scores were also calculated independently using only spliced PacBio long-read alignments, rather than spliced short-reads. This analysis produced similar results (Figure S9).

Comparison of Human and Mouse Transcriptomes

To perform a fair comparison of transcriptome dimensions between human and mouse profiles, we generated *de novo* transcriptome assemblies for each, at matched depth, rather than using the single-molecule human transcriptome (which would skew the comparison, since long-read sequencing was not performed on mouse tissues). While we note that *de novo* assembly may produce some spurious transcript models, the comparison of transcriptome dimensions between human and mouse is fair, since biases should affect both assemblies equally. Additionally, because assemblies were generated without the support of reference gene catalogs, they are not affected by discrepancies in the quality/completeness of human/mouse genome annotation.

After aligning reads to hg19 or mm10 with STAR and retaining only on-target, uniquely mapped reads (see above), libraries were subsampled to achieve a maximum matched depth of ~100 million alignments per tissue (Tables S4 and S7). Assemblies were generated from these matched libraries using StringTie (Pertea et al., 2015). Assemblies from replicates for each tissue were merged using Cuffmerge (Trapnell et al., 2012). Full-length assembled transcript sequences were filtered and classified as above (see above).

Conservation Analyses

To assess genomic-level sequence constraint we calculated average per-base PhyloP scores from 100 Vertebrate MultiZ alignments, obtained via the UCSC Genome Browser, within noncoding exons. Intergenic and intronic sequences of equivalent size were selected at random and used to provide a measure of neutral/background conservation. In addition to noncoding exons, we examined noncoding splice site dinucleotides (AG/GT), which were more strongly conserved than internal exon sequences, as well as annotated transcription factor binding sites and DHS sites on Hsa21 (Figure S12).

While PhyloP scores provide a useful measure of constraint on genomic sequence (inferred by comparison to other vertebrate genomes), our targeted transcriptome survey of syntenic human-mouse genome regions, allowed us to evaluate whether conserved exonic splice site nucleotides are actually transcribed and/or used as splice sites in both lineages. To do so, we lifted splice-site dinucleotides (AG/GT) demarcating human noncoding exons to the mouse genome using the UCSC liftOver tool. If a human splice-site dinucleotide was found to be conserved in the mouse genome, we tested to see if an equivalent junction could be found among mouse CaptureSeq alignments. While ~15% of human noncoding splice site nucleotides are conserved in mouse at the genomic level, a corresponding splice site was found in mouse for fewer than 2% of human sites.

Analysis of Hsa21 Splicing and Expression in the Tc1 Mouse

The Tc1 mouse strain is a model for trisomy-21 that carries a stable copy of Hsa21 (Yu et al., 2010). We applied short-read RNA CaptureSeq (Illumina HiSeq 2500) to provide saturating coverage of both the Hsa21 transcriptome and syntenic mouse chromosome regions in Tc1 mouse tissues (brain, kidney and testis; Table S9). To assess alternative splicing we calculated PSI scores (see above) for protein-coding and noncoding exons expressed on Hsa21 or the mouse syntenic regions in Tc1 tissues. We compared PSI values at Tc1 exons to their equivalent exons detected in human or mouse, and between orthologous exons from the two species (Figure S13). To perform global comparisons of alternative splicing profiles among orthologous human/mouse exons, we clustered human/mouse/Tc1 mouse tissues according to their PSI scores using either principle component analysis or rank-correlation clustering (Figure S14). We did the same for gene expression, using FPKM values for orthologous genes rather than PSI scores.

DATA AND SOFTWARE AVAILABILITY

Raw sequencing data and a combined transcriptome annotation are available via the NCBI Gene Expression Omnibus (GEO) under the following accession: GSE99637.