Universal alternative splicing of noncoding exons

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

[1] *Genomics and Epigenetics Division, Garvan Institute of Medical Research, NSW, Australia*

[2] *School of Biotechnology and Biomolecular Sciences, Faculty of Science, University of New South Wales, Sydney, Australia*

[3] *Australian Institute for Bioengineering and Nanotechnology, University of Queensland, QLD, Australia*

[4] *St Vincent's Clinical School, University of New South Wales, Sydney, Australia*

[5] *Pacific Biosciences, Menlo Park, CA, USA*

[6] *MRC Functional Genomics Unit, Department of Physiology and Genetics, University of Oxford, Oxford, UK*

[7] *Institute for Molecular Bioscience, University of Queensland, QLD, Australia*

* *These authors contributed equally.*

° *Joint corresponding authors: t.mercer@garvan.org.au, j.mattick@garvan.org.au*

The human genome encodes an unknown diversity of genes and isoforms. RNA Sequencing (RNA-Seq) suffers from an expression-dependent bias that impedes discovery of low-abundance transcripts and has prevented a complete census of human gene expression. Here 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 and identifies a fundamental distinction in the architecture of protein-coding and noncoding genes. Unlike their coding counterparts, and in contrast to the impression from more shallow surveys, noncoding exons undergo near-universal alternative splicing to produce an effectively inexhaustible variety of isoforms. Targeted RNA-Seq analysis of syntenic regions of the mouse genome indicates that few noncoding exons are shared between human and mouse. Despite this divergence, human alternative splicing profiles are recapitulated on Hsa21 in mouse nuclei, implying regulation by a local splicing code that is more strongly conserved than the noncoding isoforms themselves. We propose that noncoding exons are functionally modular, with combinatorial alternative splicing generating an enormous repertoire of potential regulatory RNAs and a rich transcriptional reservoir for adaptive gene evolution.

INTRODUCTION

The mammalian genome is transcribed into protein-coding and noncoding RNAs collectively termed the transcriptome (Djebali et al., 2012). The transcriptome is so large, diverse and dynamic that, even after a decade of investigation by RNA Sequencing (RNA-Seq; Wang et al., 2009), we are yet to reach its limits and 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 eroded by the recognition of pervasive transcription across the genome (Carninci et al., 2005; Clark et al., 2011) and interleaved alternative isoforms at individual loci (Kapranov et al., 2005; Mercer et al., 2011).

Deep RNA-Seq has unearthed 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). Similarly, many protein-coding genes express alternative isoforms that lack extended open reading frames (ORFs; Gonzàlez-Porta et al., 2013). Despite the growing list of functional noncoding RNAs (Quek et al., 2015) and the wide variety of roles they fulfill (Fatica and Bozzoni, 2014; Qureshi and Mehler, 2012; Satpathy and Chang, 2015), these unexpected findings have fuelled 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/or isoforms (Derrien et al., 2012; Hon et al., 2017; Iyer et al., 2015). 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 especially impeded the discovery and characterization of lncRNAs, since these are typically lowly (or precisely) expressed (Cabili et al., 2011; Liu et al., 2016; Mercer et al., 2008; Mukherjee et al., 2017). As a result, our understanding of lncRNA biology is relatively poor, and often informed by analyses of only those examples with sufficient expression to be accessible by RNA-Seq.

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). In addition to its demonstrated roles in the neurological phenotypes of trisomy 21, the Hsa21 gene *DYRK1A* is implicated in Alzheimer's and autism (Dang et al., 2017; Wegiel et al., 2011).

To measure transcription from Hsa21, but exclude the remainder of the genome, we used targeted RNA sequencing (RNA CaptureSeq; Mercer et al., 2014). Baits were tiled across the chromosome to capture its complete expression profile and transcripts were sequenced deeply on single-molecule and short-read platforms. This approach overcomes the expression-dependent bias of RNA-Seq to reveal the true dimensions of expressed gene populations encoded within a cross-section of the genome, and identifies a fundamental distinction in the architecture of protein-coding and noncoding RNAs.

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 sequenced reads are competitively sampled from a single pool, in which transcripts of varied abundance are

proportionally represented, lowly expressed transcripts commonly evade detection (Clark et al., 2011; Hardwick et al., 2016). Second; the accurate assembly of full-length isoforms from short (<150bp) sequencing reads is challenging, particularly when multiple alternative isoforms are transcribed from a single locus (Martin and Wang, 2011; Tilgner et al., 2015).

To overcome these challenges, we performed short-read (Illumina) and single-molecule (PacBio) RNA CaptureSeq (Mercer et al., 2014), targeting the complete expression profile of Hsa21. To do so, we generated biotin-labeled oligonucleotide baits tiling the entire non-repetitive Hsa21 sequence. These were used to capture full-length cDNA molecules, thereby restricting sequencing to transcripts expressed on Hsa21 (see **Supplementary Materials and Methods**).

To validate our approach, we analyzed Hsa21-enriched cDNA libraries from the human K562 cell-type by short-read sequencing. We achieved a median 223-fold enrichment in sequencing coverage and an expanded transcriptome assembly (2.1-fold increase in unique multi-exon transcripts), compared to parallel analysis by conventional RNA-Seq (**Figure S2**). The legitimacy of novel transcripts assembled from short-read CaptureSeq was confirmed by RT-PCR and Sanger sequencing (16/20 randomly selected examples; **Figure S3**; **Table S1**).

Next we analyzed Hsa21-enriched cDNA from human testis by deep single-molecule sequencing. 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 non-chimeric reads aligning to Hsa21, representing 910 Mb of usable transcript sequence concentrated in ~1.5% of the genome (**Figure S4A**). After filtering, we retrieved 101,478 full-length multi-exonic transcripts on Hsa21 (**Figure S4B**).

In addition, we performed short-read sequencing on Hsa21-enriched cDNA from testis, brain and kidney (**Figure S5A**). At least 100 million short-read alignments to Hsa21 were obtained for each tissue, meaning >7 billion alignments (per tissue) would be required to achieve equivalent coverage via traditional RNA-Seq (**Table S2**).

The majority (80.3%) of spliced short-read alignment junctions were concordantly mapped to an intron in our single-molecule Hsa21 transcriptome profile, with this being predictably higher for testis (84.7%), than brain (80.5%) or kidney (78.0%; **Figure S5B**). Likewise, the majority of unique internal exons in our single-molecule Hsa21 transcriptome were detected by short-reads, and detection rates were comparable to those of annotated exons in the Gencode (v19) catalog for both protein-coding (95.3% vs 97.5%) and noncoding exons (85.1% vs 87.8%; **Figure S5C**). Splice-junction concordance with more accurate short-reads suggests that the intron-exon architecture of single-molecule isoforms was, in general, correctly resolved.

Transcriptional landscape of chromosome 21

Hsa21 has frequently been used as a cross-sectional model of the genome (Cawley et al., 2004; Kampa et al., 2004), 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**). The combination of targeted single-molecule sequencing with saturating short-read coverage allows us to accurately resolve full-length isoforms on Hsa21 and perform quantitative analyses of their expression and splicing (Conesa et al., 2016).

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

At protein-coding gene loci on Hsa21, we identified 7,374 unique multi-exonic isoforms, of which 77% were novel, including many noncoding isoform variants (**Figure 1C**). These encoded up to 2,272 possible open reading frames (ORFs) that were not currently annotated, encompassing 1,046 novel coding exons and 1,433 novel ORF introns (**Figure 1C**; examples in **Figure S6**).

Although these are predicted ORFs only, and are not necessarily expressed as novel peptides, this is a considerable increase on current annotations (2.7-fold).

While this suggests that the protein-coding content of Hsa21 may be underestimated, the majority (68%) of novel isoforms to protein-coding genes were noncoding isoform variants or possessed novel untranslated regions (UTRs). Alternative splicing of UTRs (both 5' and 3') was common and often highly complex (examples in **Figure S7**). Single-molecule sequencing was particularly useful for resolving UTR variants, since it does not suffer from sequencing 'edge effects' that impact short-read transcript assembly (Martin and Wang, 2011). In total, protein-coding loci on Hsa21 encoded 6154 unique internal exons (29% novel) and 6056 unique canonical introns (51% novel; **Figure 1C**).

At noncoding loci, we identified 1,516 novel lncRNA isoforms across Hsa21, encompassing 1,348 unique internal exons (61% novel) and 2,589 unique canonical introns (81% novel; **Figure 1C**). Examples of rich isoform diversity at lncRNA loci were routinely resolved with targeted single-molecule sequencing (**Figure 1B**). We also generated more complete gene models for many known lncRNAs or incorporated multiple partial lncRNA annotations into single unified loci (**Figure S8A-C**). The splicing of lncRNAs with neighboring protein-coding genes to form long, non-annotated, extensively spliced UTRs was also routinely observed (**Figure S8D**).

By extrapolation, we estimate that the broader human genome encodes 383,000 unique isoforms to protein-coding genes that encompass 147,000 possible ORFs. This is 1.9-fold more isoforms than listed in the Gencode v26 catalog, and incorporates 1.4-fold more unique introns and 1.2-fold more unique internal exons. We predict noncoding gene loci express 98,000 multi-exonic isoforms (2.1-fold increase), incorporating 88,000 internal exons (1.2-fold) and 168,00 introns (1.7-fold). While we have profiled only three tissues, ensuring these are lower-bound estimates, it is clear that a large amount of transcriptional diversity remains unexplored.

Near-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 (see **Supplementary Materials and 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**; **Figure S9A**). This indicates that, although the majority of exons were discovered, noncoding isoforms were not exhaustively resolved, even with the enhanced sensitivity afforded by targeted sequencing.

Because each unique intron represents a different splicing event, this result suggests that alternative splicing generates a near-limitless diversity of isoforms at noncoding loci. To further assess alternative splicing, we calculated a Percent Splice Inclusion (PSI) score for each internal exon in our Hsa21 transcriptome (see **Supplementary Materials and Methods**). Unlike protein-coding exons (median PSI = 90.5%), almost all noncoding exons were alternatively spliced (median PSI = 55.5%; **Figure 2B; Figure S10A-D**). The abundances of protein-coding and noncoding introns, compared to exons, further supports this finding: that the relative difference between coding and noncoding introns (49.8-fold) is larger than for exons (19.2-fold; **Figure 2C**; **Figure S9B,C**) reflects the greater isoform diversity generated by enriched alternative splicing of noncoding RNAs.

The distinction in protein-coding and noncoding splicing diversity is illustrated by comparison of the protein-coding gene *SAMSN1* to its antisense lncRNA (*SAMSN1-AS1*) and a nearby intergenic lncRNA (*AJ006998.2*). Most single-molecule reads from *SAMSN1* match either of the gene's two annotated mRNA isoforms. By contrast, RNA CaptureSeq expanded the annotations for

SAMSN1-AS1 and *AJ006998.2*, with most single-molecule reads encoding unique alternative isoforms (**Figure 2D**).

We noted that near-universal alternative splicing was not limited to lncRNA exons but was similarly observed among noncoding exons within the UTRs of protein-coding genes (**Figure S10E**). For example, the gene *CHODL* exhibited rich alternative splicing in its 5'UTR that was not previously annotated (**Figure S11**). Canonical poly-pyrimidine enrichments were observed immediately upstream of splice acceptor sites at both lncRNA and UTR exons, and these motifs were indistinguishable from those found at protein-coding exons (**Figure S10F**). This indicates that noncoding exons are demarcated by *bona fide*, rather than cryptic splice sites.

Our observations are consistent with a recent study that showed, despite canonical *cis*-regulatory sequences, U2AF65 occupancy is relatively lower in lncRNAs than mRNAs, a feature independently correlated with heightened alternative over constitutive splicing (Melé et al., 2017). Moreover, a global reduction in splicing kinetics has been observed in lncRNAs (Mukherjee et al., 2017), relative to protein-coding genes, with this effect also apparent for UTR exons (Tilgner et al., 2012).

Supported by these data, 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 exons, allowing the spliceosome to explore the full range of noncoding exon combinations to generate an effectively inexhaustible noncoding isoform diversity.

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 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 at equivalent depth to human samples within matched mouse tissues (testis, brain and kidney), enabling comparison of the two transcriptomes at high-resolution (see **Supplementary Materials and Methods**).

As for the human chromosome, 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 to 88.4% on Hsa21 (**Figure 1D,3B**). In mouse, 25.1% of targeted bases were retained as mature exons, compared to 16.7% in human, with the remaining fraction (55.4% vs 71.7%) removed as introns (**Figure 1D,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 comparable in size to human (95%, based on unique internal exon count; **Figure 3B**).

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 to just 31% in mouse (**Figure 2B**). Reflecting this heightened splicing diversity, human protein-coding genes had more internal exons (20%), introns (30%), isoforms (24%) than their corresponding mouse orthologs (**Table S3**).

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 brain, we identified six novel internal exons in the human brain (in addition to all 13 currently annotated *DYRK1A* exons; **Figure 4A,B**). Extensive alternative splicing generated at least 11 novel *DYRK1A* isoforms, of which 10 comprise noncoding variants and one encodes a

novel ORF with a N-terminal modification to the DYRK1A protein (**Figure 4A,B**; interestingly, an analogous N-terminal modification regulates subcellular localization of the DYRK4 paralog (Papadopoulos et al., 2011)).

As was the case on Hsa21, the majority of novel exons discovered in the mouse syntenic regions were noncoding, ultimately outnumbering their protein-coding counterparts (1.1-fold; **Figure 3B**; **Figure S12A-B**), and these were similarly subject to near-universal alternative splicing (**Figure 2B**; **Figure S12C**). This indicates that the size and structure of the human and mouse transcriptomes are largely comparable, with each harboring large noncoding RNA populations that are diversified by prolific alternative splicing.

Despite this similarity, we found that individual lncRNAs were largely divergent between the two lineages (see **Supplementary Materials and Methods**). While 19% of splice-acceptor and 16% of splice-donor dinucleotides (AG/GT) were conserved between the human and mouse genomes, only ~2% of human lncRNA splice sites was expressed in any mouse tissue (**Figure 3C,D**). Despite being poorly conserved relative to their protein-coding counterparts, noncoding exons exhibited internal nucleotide conservation comparable to annotated DNase hypersensitive or transcription factor binding sites across vertebrate genomes (**Figure 3C,D**), with flanking splice sites showing a further, though relatively modest, conservation enrichment. This indicates that whilst similar in size and structure, the content of human and mouse noncoding RNA populations is 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 nuclei

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 (brain, kidney and testis; see **Supplementary Materials and 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 (**Figure 4A-D**). Globally, 87% of human-specific splice sites distinguishing human and mouse orthologs were also detected on Hsa21 in the Tc1 mouse (compared to 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 to 39% in mouse; **Figure S13B,C**). Furthermore, 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 primacy 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%; **Figure 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 principle component analysis or rank-correlation clustering (**Figure S14A-C**). 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 (**Figure S14A,B**). 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 nuclei 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, implying these are regulated by a code that is written into the local chromosome sequence, and can be correctly interpreted by the mouse spliceosome (this is consistent with previous reports; Barash et al., 2010; Barbosa-Morais et al., 2012). The splicing code is so highly conserved between human and mouse, that even human-specific exons and noncoding RNAs without orthologs in mouse are correctly spliced: that is, the elements mediating splicing are better conserved than the isoforms themselves. Therefore the splicing lexicon is deeply conserved, even whilst the elements it regulates undergo rapid turnover and re-deployment.

DISCUSSION

To overcome the expression-dependency of RNA-Seq, which impedes discovery and characterization of low abundance transcripts (Clark et al., 2011; Hardwick et al., 2016), we performed targeted RNA sequencing across human chromosome 21. The combination of single-molecule and short-read RNA CaptureSeq enabled accurate resolution of isoform diversity and analysis of alternative splicing at unprecedented depth.

We observed a fundamental, yet previously unappreciated, distinction between the architecture of protein-coding and noncoding gene content. Noncoding loci are, contrary to the impression from more shallow surveys (Cabili et al., 2011; Derrien et al., 2012), enriched for alternative splicing, with almost all noncoding exons being alternative. Therefore, while protein-coding genes are constrained by the requirement to maintain an ORF, no such constraint is imposed on noncoding RNAs, and noncoding exons are recombined with maximum flexibility.

This finding implies that noncoding exons are functionally modular. Operating as discrete cassettes with unique biochemical properties (probably defined by their individual secondary structures; Mercer and Mattick, 2013), noncoding exons are combinatorially assembled into a large variety of isoforms. We speculate that individual noncoding exons may form specific interactions with other biomolecules (proteins, RNAs and/or DNA-motifs), organizing these around the scaffold of a noncoding transcript. In this way, different noncoding isoforms could assemble different collections of binding partners to dynamically organize and regulate cellular processes.

The distinction between protein-coding and noncoding exons was also evident when comparing exons within ORFs to untranslated regions of the same gene (located in 5' or 3'UTRs, or specific to noncoding isoforms). This implies modularity in the functional architecture of UTR regions and also suggests a primary importance for the nonsense-mediated-decay (NMD) pathway in constraining protein-coding isoform diversity. Universal alternative splicing may represent the default mode of both coding and noncoding gene expression, with NMD subsequently responsible for eliminating promiscuous isoform diversity within protein-coding sequences.

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 sequencing 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 dynamic, and that there does not exist a definitive list of noncoding isoforms that can be feasibly catalogued.

The functional significance of noncoding RNAs remains the subject of debate among the genomics community. The generally low expression, weak sequence conservation and sheer number of unique lncRNAs has caused many to question their relevance (despite the growing list of functional examples; Quek et al., 2015). Regardless of its implications for lncRNA functionality, the near-universal alternative splicing of noncoding exons reported here certainly contributes to a reservoir of transcriptional diversity from which molecular innovations might evolve. *De novo* gene birth, an increasingly interesting facet of genome evolution (Kaessmann, 2010), is suspected to proceed via lncRNA precursors (Toll-Riera et al., 2009; Wu et al., 2011; Xie et al., 2012). Accordingly, the divergent evolution of lncRNAs observed between human and mouse lineages may reflect rapid adaptive gene evolution.

ACKNOWLEDGEMENTS

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 the Australian Research Council (Special Research Initiative in Stem Cell Science to L.K.N.). 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 strict accordance with the associated 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 bioinformatic 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. Correspondence should be addressed to T.R.M. (t.mercer@garvan.org.au).

CONFLICT OF INTEREST

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

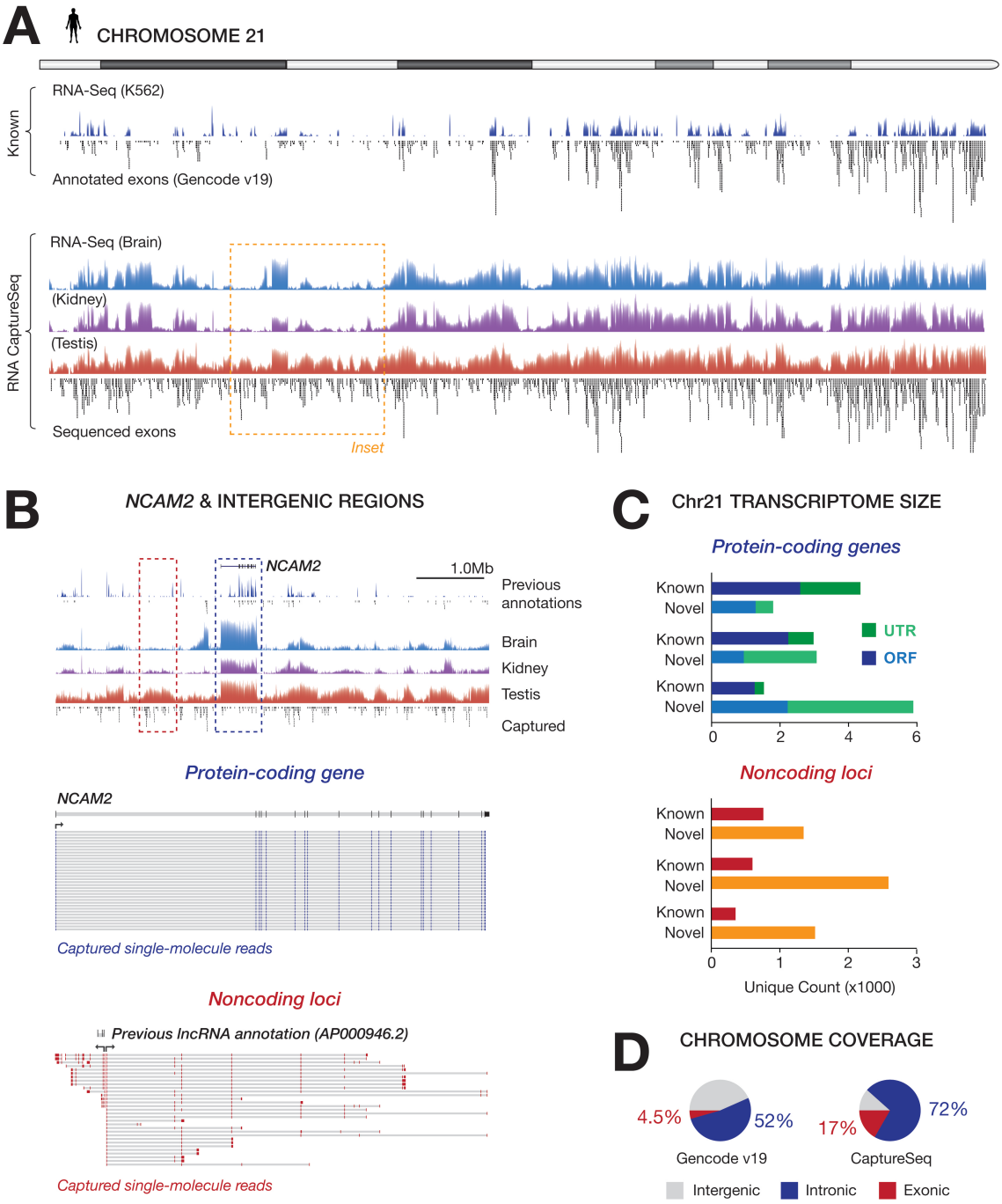

Figure 1. Transcriptional landscape of human chromosome 21. (**A**) Transcriptional activity recorded across the major arm of human chromosome 21 (Hsa21) by short-read and single-molecule RNA CaptureSeq. Normalized coverage (log scale) from short-read alignments is shown for brain (blue), kidney (purple) and testis (red), with coverage from traditional RNA-Seq (in K562 cells; navy) shown for comparison. Below coverage tracks 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. (**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 extends the gene model for the previously annotated lncRNA *AP000946.2* and resolves a novel lncRNA that is transcribed in the opposite direction. Both exhibit extensive alternative splicing, in contrast *NCAM2*, where the majority of reads represent non-redundant isoforms. (**C**) The number of unique internal exons, unique canonical introns and non-redundant 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, content belonging noncoding isoform or UTR variants (green) is distinguished from predicted ORFs (blue). (**D**) The proportion of captured bases on chromosome 21 that are exonic, intronic or silent, according to Gencode v19 (left) RNA CaptureSeq (right).

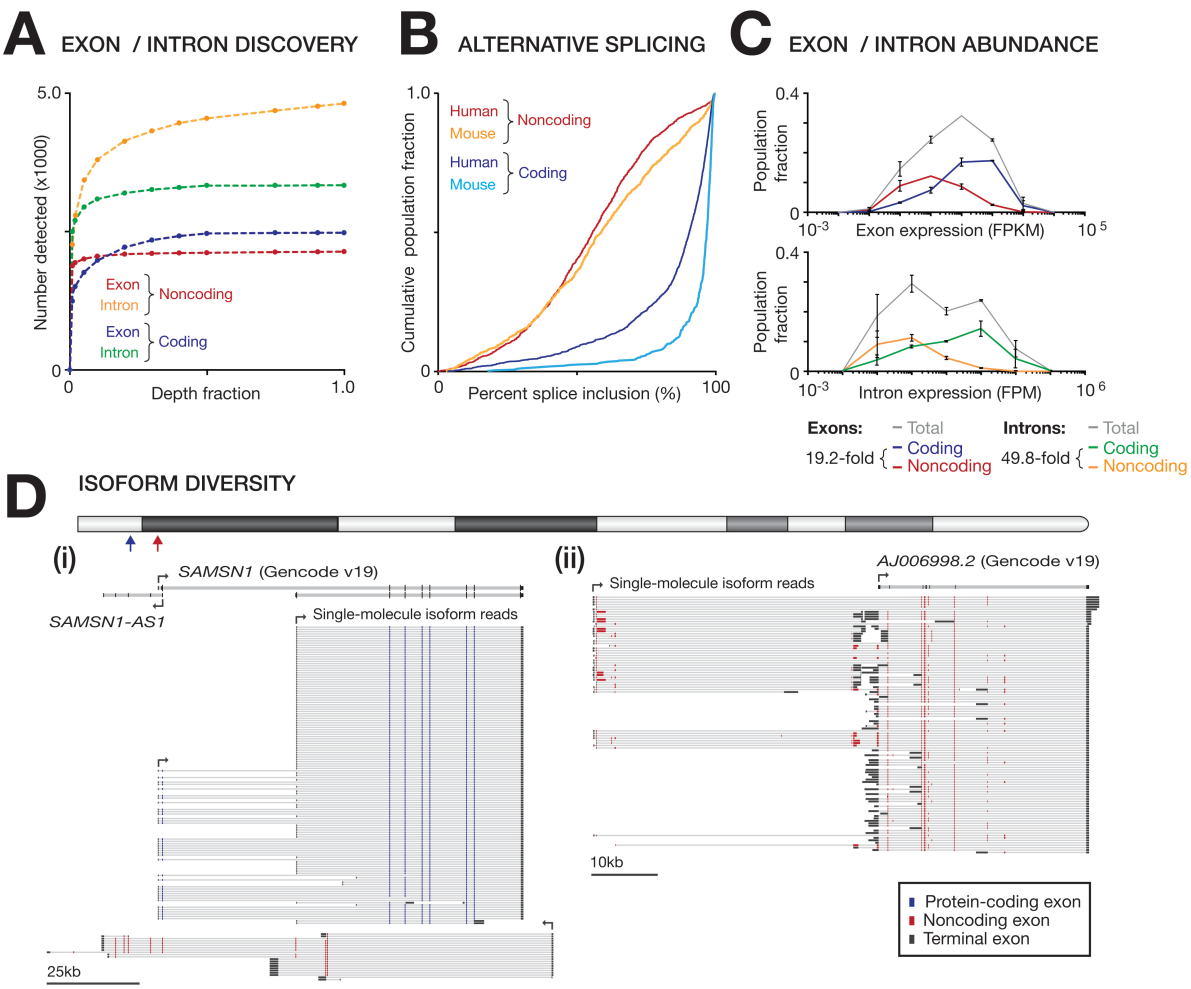

Figure 2. Near-universal alternative splicing of noncoding exons. (**A**) Discovery-saturation curves show the rate of detection for unique protein-coding/noncoding introns and unique internal exons (from the single-molecule Hsa21 transcriptome) relative to short-read sequencing depth. Depth fraction is relative to a combined pool of 300 million short-read alignments to Hsa21 from testis, brain and kidney (100 million each). (**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 indicate abundances of protein-coding/noncoding internal exons and introns in human testis (mean ± SD, n = 3). Grey line shows total exon/intron population. Median fold-difference between coding/noncoding populations is indicated 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*; (**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 most analyses.

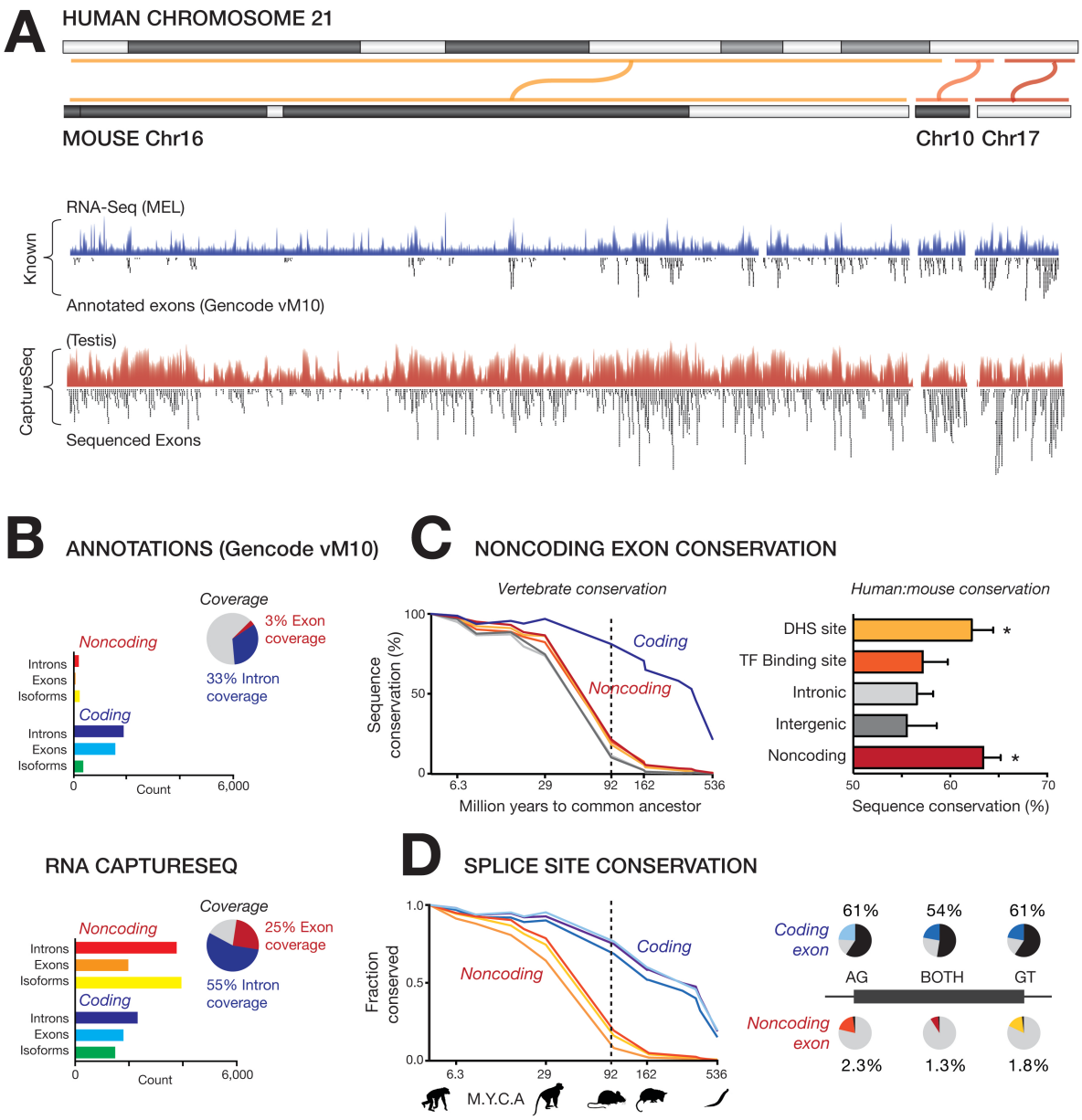

Figure 3. Transcriptional landscape of mouse syntenic regions and noncoding exon evolution. (**A**) Transcriptional activity recorded across regions of the mouse genome (Mmu16, Mmu17, Mmu10) syntenic to human chromosome 21 (Hsa21) by short-read RNA CaptureSeq. Normalized coverage (log scale) of short-read alignments is shown for testis (red), with traditional RNA-Seq (in MEL cells; navy) shown for comparison. Below coverage tracks are all unique internal exons (black) from transcripts in the Gencode vM10 reference catalog and those assembled from RNA CaptureSeq data (testis, brain, kidney combined). (**B**) Bar charts show the number of protein-coding/noncoding unique internal exons, unique canonical introns and unique transcript isoforms in Gencode vM10 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**; left) Rate of syntenic alignment for protein-coding/noncoding Hsa21 exons to other vertebrate genomes (ordered by evolutionary distance from human). DNase Hypersensitive (DHS) sites, Transcription Factor (TF) binding sites and randomly selected intergenic or intronic sequences are included for comparison. (**C**; right) Percentage sequence conservation between human and mouse sequences for features that could be aligned from the human (hg19) to mouse (mm10) genomes (mean ± SD, asterisks denote significant increases relative to intergenic and intronic sequences at $p < 0.0001$, unpaired t-test). (**D**) The rate of conservation for splice site dinucleotides (AG/GT/both) in other vertebrate genomes for protein-coding/noncoding exons on Hsa21. Pie charts indicate proportion of protein-coding/noncoding internal exons with splice site dinucleotides that are conserved in the mouse genome (red/blue) and the proportion for which equivalent splice sites were detected in mouse RNA CaptureSeq libraries (black).

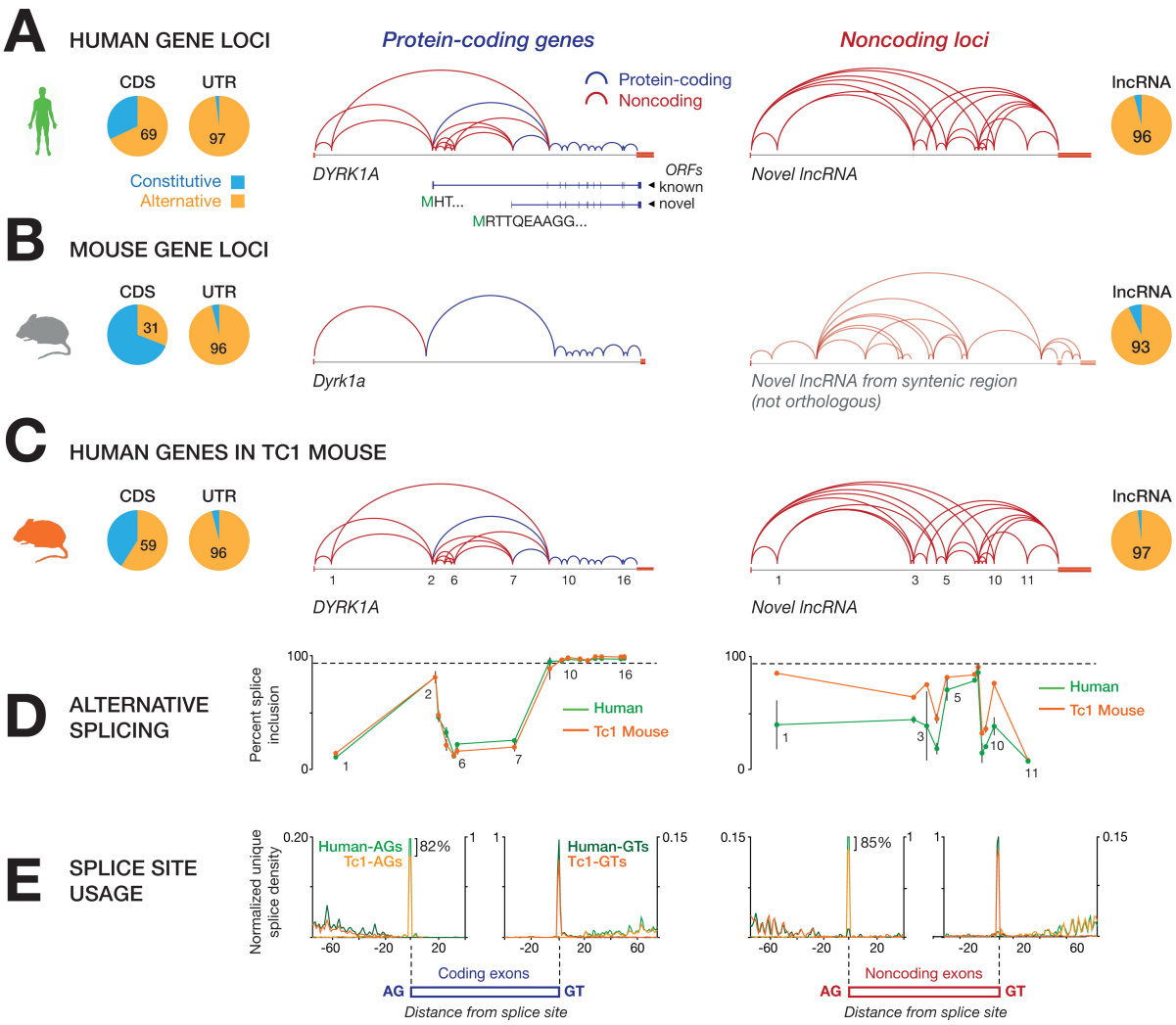

Figure 4. Splicing of human genes in mouse nuclei. (**A-C**) Exon-intron structures assembled for the protein-coding gene *DYRK1A* and a novel lncRNA locus from human-Hsa21 (**A**), mouse-syntenic regions (**B**) and on Hsa21 in the Tc1-mouse strain (**C**). Pie charts indicate the proportion of unique internal exons classified as constitutive (percentage splice inclusion; 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 illustrated 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 indicate 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.

REFERENCES

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., 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., 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., Rinn, J.L., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & Development 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.

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.

Dang, T., Duan, W.Y., Yu, B., Tong, D.L., Cheng, C., Zhang, Y.F., Wu, W., Ye, K., Zhang, W.X., Wu, M., et. al. 2017. Autism-associated Dyrk1a truncation mutants impair neuronal dendritic and spine growth and interfere with postnatal cortical development. Mol. Psychiatry.

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 Research 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.

Fatica, A., Bozzoni, I., 2014. Long non-coding RNAs: new players in cell differentiation and development. Nature Reviews Genetics 15, 7–21.

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

Hardwick, S.A., Chen, W.Y., Wong, T., Deveson, I.W., Blackburn, J., Andersen, S.B., Nielsen, L.K., Mattick, J.S., Mercer, T.R., 2016. Spliced synthetic genes as internal controls in RNA sequencing experiments. Nat Meth 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., Zhao, S. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet 47, 199–208.

Kaessmann, H., 2010. Origins, evolution, and phenotypic impact of new genes. Genome Research 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 Research 14, 331–342.

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

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.

Liu, S.J., Nowakowski, T.J., Pollen, A.A., Lui, J.H., Horlbeck, M.A., Attenello, F.J., He, D., Weissman, J.S., Kriegstein,

A.R., Diaz, A.A., Lim, D.A., 2016. Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biol. 17, 67.

Martin, J.A., Wang, Z., 2011. Next-generation transcriptome assembly. Nature Reviews Genetics 12, 671–682.

Melé, M., Mattioli, K., Mallard, W., Shechner, D.M., Gerhardinger, C., Rinn, J.L., 2017. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. Genome Research 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., Mattick, J.S., 2014. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. Nat Protoc 9, 989–1009.

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., Mattick, J.S., 2008. Specific expression of long noncoding RNAs in the mouse brain. Proc. Natl. Acad. Sci. U.S.A. 105, 716–721.

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

Mercer, T.R., Mattick, J.S., 2013. Structure and function of long noncoding RNAs in epigenetic regulation. Nature Structural & Molecular Biology 20, 300–307.

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

Mukherjee, N., Calviello, L., Hirsekorn, A., de Pretis, S., Pelizzola, M., Ohler, U., 2017. Integrative classification of human coding and noncoding genes through RNA metabolism profiles. Nature Structural & Molecular Biology 24, 86–96.

Papadopoulos, C., Arato, K., Lilienthal, E., Zerweck, J., Schutkowski, M., Chatain, N., Müller-Newen, G., Becker, W., la Luna, de, 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.

Quek, X.C., Thomson, D.W., Maag, J.L.V., Bartonicek, N., Signal, B., Clark, M.B., Gloss, B.S., Dinger, M.E., 2015. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Research 43, D168–73.

Qureshi, I.A., Mehler, M.F., 2012. Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. Nat. Rev. Neurosci. 13, 528–541.

Satpathy, A.T., Chang, H.Y., 2015. Long noncoding RNA in hematopoiesis and immunity. Immunity 42, 792–804.

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. CellReports 3, 2179–2190.

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

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., 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 Research 22, 1616–1625.

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., Alba, M.M., 2009. Origin of primate orphan genes: a comparative genomics approach. Molecular Biology and Evolution 26, 603–612.

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, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10, 57–63.

Wegiel, J., Gong, C.-X., Hwang, Y.-W., 2011. The role of DYRK1A in neurodegenerative diseases. FEBS J. 278, 236–245.

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

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

Xie, C., Zhang, Y.E., Chen, J.-Y., Liu, C.-J., Zhou, W.-Z., Li, Y., Zhang, M., Zhang, R., Wei, L., 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. Human Molecular Genetics 19, 2780–2791.

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.