




# High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing

Julien Lagarde<sup>1,2,7</sup>, Barbara Uszczynska-Ratajczak<sup>1,2,6,7</sup>, Silvia Carbonell<sup>3</sup>, Sílvia Pérez-Lluch<sup>1,2</sup> , Amaya Abad<sup>1,2</sup>, Carrie Davis<sup>4</sup>, Thomas R Gingeras<sup>4</sup>, Adam Frankish<sup>5</sup>, Jennifer Harrow<sup>5,6</sup>, Roderic Guigo<sup>1,2</sup>  & Rory Johnson<sup>1,2,6</sup> 

Accurate annotation of genes and their transcripts is a foundation of genomics, but currently no annotation technique combines throughput and accuracy. As a result, reference gene collections remain incomplete—many gene models are fragmentary, and thousands more remain uncataloged, particularly for long noncoding RNAs (lncRNAs). To accelerate lncRNA annotation, the GENCODE consortium has developed RNA Capture Long Seq (CLS), which combines targeted RNA capture with third-generation long-read sequencing. Here we present an experimental reannotation of the GENCODE intergenic lncRNA populations in matched human and mouse tissues that resulted in novel transcript models for 3,574 and 561 gene loci, respectively. CLS approximately doubled the annotated complexity of targeted loci, outperforming existing short-read techniques. Full-length transcript models produced by CLS enabled us to definitively characterize the genomic features of lncRNAs, including promoter and gene structure, and protein-coding potential. Thus, CLS removes a long-standing bottleneck in transcriptome annotation and generates manual-quality full-length transcript models at high-throughput scales.

lncRNAs represent a vast and relatively unexplored component of the mammalian genome. The assignment of lncRNA functions depends on the availability of high-quality transcriptome annotations. At present such annotations are still rudimentary: we have little idea of the total number of lncRNAs, and for those that have been identified, transcript structures remain largely incomplete.

Projects using diverse approaches have helped to increase both the number and size of available lncRNA annotations. Early gene sets, derived from a mixture of FANTOM cDNA sequencing efforts and public databases<sup>1,2</sup>, were joined by long intergenic noncoding RNA (lincRNA) sets discovered through chromatin signatures<sup>3</sup>. More recently, researchers have applied transcript-reconstruction software such as Cufflinks<sup>4</sup> to identify novel genes in short-read RNA-sequencing (RNA-seq) data sets<sup>5–9</sup>. However, the standard references for lncRNAs are currently the regularly updated manual annotations from GENCODE, which are based on the curation of cDNAs and expressed sequence tags by human annotators<sup>10,11</sup> and have been adopted by international genomics consortia<sup>12–15</sup>.

At present, annotation efforts face a necessary compromise between throughput and quality. Short-read-based transcriptome-reconstruction methods deliver large annotations with low financial and time investment, whereas manual annotation is slow and requires long-term funding. However, the quality of software-reconstructed annotations is often doubtful because of the inherent difficulty of reconstructing transcript structures from shorter sequence reads.

Such structures tend to be incomplete and often lack terminal exons or splice junctions between adjacent exons<sup>16</sup>. This particularly affects lncRNAs, whose low expression results in low read coverage<sup>11</sup>. The outcome is a growing divergence between large automated annotations of uncertain quality (e.g., 101,700 genes for NONCODE<sup>8</sup>) and the highly curated, ‘conservative’ GENCODE collection<sup>11</sup> (15,767 genes for version 25).

Annotation incompleteness takes two forms. First, genes may be entirely missing from an annotation; many genomic regions are suspected to transcribe RNA but contain no annotation, including ‘orphan’ small RNAs with presumed long precursors<sup>17</sup>, enhancers<sup>18</sup> and ultraconserved elements<sup>19,20</sup>. Second, annotated lncRNAs may represent partial gene structures. Start and end sites frequently lack independent supporting evidence<sup>11</sup>, and lncRNAs are shorter and have fewer exons than mRNAs<sup>7,11,21</sup>. Recently, a method of rapid amplification of cDNA ends followed by sequencing (RACE-seq) was developed to complete lncRNA annotations, albeit at relatively low throughput<sup>21</sup>.

One of the principal impediments to the annotation of lncRNAs is their low steady-state levels<sup>3,11</sup>. To overcome this, RNA capture sequencing (CaptureSeq)<sup>22</sup> is used to boost the concentration of low-abundance transcripts in cDNA libraries. Such studies depend on short-read sequencing and *in silico* transcript reconstruction<sup>22–24</sup>. Thus, although CaptureSeq achieves high throughput, its transcript structures lack the confidence required for inclusion in GENCODE.

<sup>1</sup>Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. <sup>2</sup>Universitat Pompeu Fabra (UPF), Barcelona, Spain. <sup>3</sup>R&D Department, Quantitative Genomic Medicine Laboratories (qGenomics), Barcelona, Spain. <sup>4</sup>Functional Genomics Group, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. <sup>5</sup>Wellcome Trust Sanger Institute, Hinxton, UK. <sup>6</sup>Present addresses: Centre of New Technologies, Warsaw, Poland (B.U.-R.); Illumina, Cambridge, UK (J.H.); Department of Clinical Research, University of Bern, Bern, Switzerland (R.J.). <sup>7</sup>These authors contributed equally to this work. Correspondence should be addressed to R.J. (rory.johnson@dbmr.unibe.ch) or R.G. (roderic.guigo@crp.cat).

Received 1 February; accepted 11 October; published online 6 November 2017; doi:10.1038/ng.3988

In this paper, we describe a new method, CLS, which couples targeted RNA capture with third-generation long-read cDNA sequencing. We used CLS to interrogate the GENCODE catalog of intergenic lncRNAs, together with thousands of suspected novel loci, in six human tissues and six mouse tissues. We demonstrate that CLS combines the throughput of CaptureSeq with high-confidence, complete transcript models from long-read sequencing, resulting in an advance in transcriptome annotation.

## RESULTS

### Application of CLS to complete lncRNA annotations

Our aim was to develop an experimental approach that could improve and extend reference transcript annotations while minimizing human intervention and avoiding *in silico* transcript assembly. We designed CLS, which couples targeted RNA capture to Pacific Biosciences (PacBio) third-generation long-read sequencing (Fig. 1a).

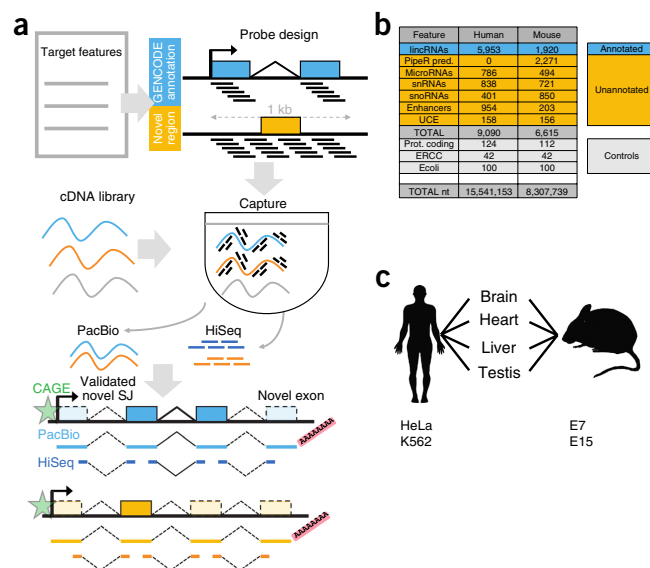
CLS can be used for two distinct objectives: to improve existing gene models, and to identify novel loci (Fig. 1a). Although in the present study we focused mainly on the former aim, we demonstrate that novel loci can be captured and sequenced. We created a comprehensive capture library targeting the set of intergenic GENCODE lncRNAs in human and mouse tissues. Annotations for humans are currently more complete than those for mice, and thus the annotations are different sizes (14,470 and 5,385 lncRNA genes in GENCODE releases 20 and M3, respectively). The GENCODE annotations probed in this study were principally multi-exonic transcripts based on polyadenylated (polyA+) cDNA/expressed sequence tag libraries, and thus were not likely to include ‘enhancer RNAs’<sup>10,25</sup>. To these we added tiled probes targeting loci that may produce lncRNAs: small RNA genes<sup>26</sup>, enhancers<sup>27</sup> and ultraconserved elements<sup>28</sup>. For mouse tissues we also added orthologous lncRNA predictions from PipeR<sup>29</sup>. We added numerous control probes, including a series that targeted half of the External RNA Controls Consortium (ERCC) synthetic spike-ins<sup>30</sup>. These sequences were targeted by capture libraries of temperature-matched and nonrepetitive oligonucleotide probes (Fig. 1b).

To access the maximal lncRNA diversity, we chose transcriptionally complex and biomedically relevant organs from mice and humans: whole brain, heart, liver and testis (Fig. 1c). We added two heavily studied human cell lines, HeLa and K562 (ref. 31), and two mouse embryonic time points (embryonic day 7 (E7) and E15).

We designed a protocol to capture full-length, oligo-dT-primed cDNAs (Online Methods). Barcoded, unfragmented cDNAs were pooled and captured. Preliminary qPCR analysis indicated enrichment for targeted regions (Supplementary Fig. 1a). PacBio sequencing tends to favor shorter templates in a mixture<sup>32</sup>. Therefore, we grouped pooled, captured cDNA into three size ranges (1–1.5 kb, 1.5–2.5 kb and >2.5 kb) (Supplementary Fig. 1b,c) and used it to construct sequencing libraries for PacBio single-molecule real-time (SMRT) sequencing technology<sup>33</sup>.

### CLS yields an enriched long-read transcriptome

We sequenced samples on 130 SMRT cells and obtained ~2 million reads in total for each species (Fig. 2a). We demultiplexed PacBio reads, or ‘reads of insert’ (ROIs), to retrieve their tissue of origin and mapped them to the genome. We observed high mapping rates (>99% in both cases), of which 86% and 88% were unique in human and mouse samples, respectively (Supplementary Fig. 2a). (Throughout the rest of the paper, all data are presented in the format “human/mouse.”) The use of short barcodes meant that for ~30% of reads, the tissue of origin could not be retrieved (Supplementary Fig. 2b). This could be remedied by the use of longer barcodes. Representation



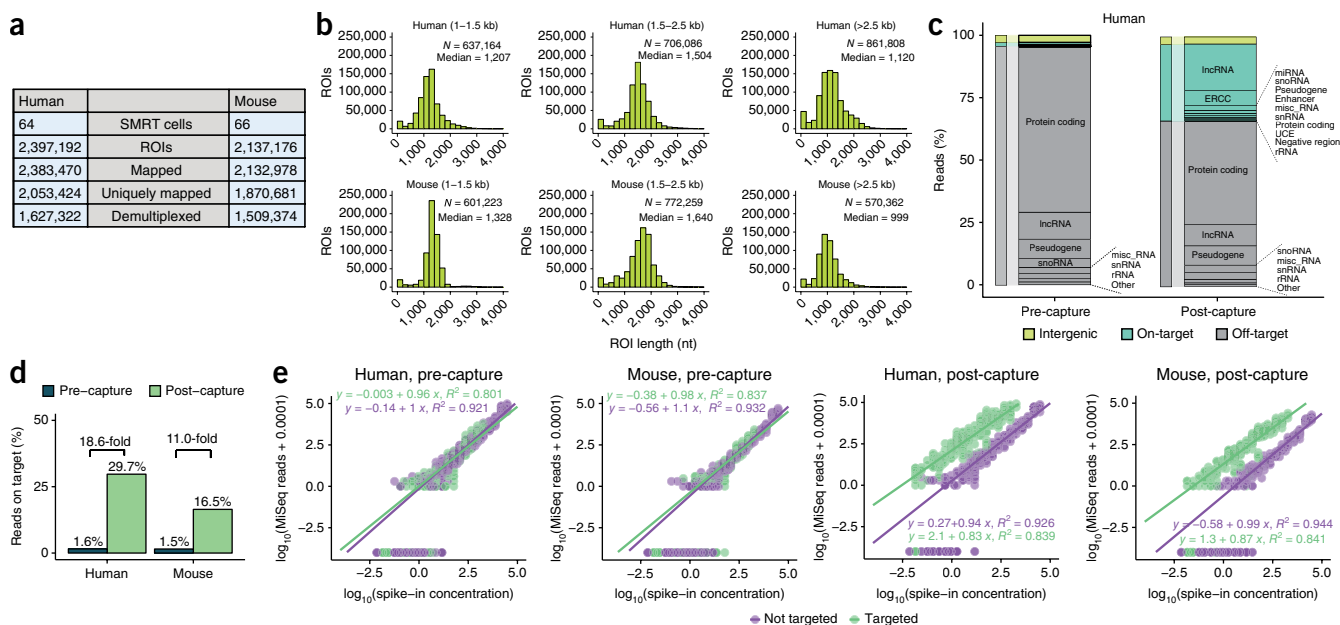
**Figure 1** Using the CLS approach to extend GENCODE lncRNA annotation. **(a)** The strategy for automated, high-quality transcriptome annotation. CLS can be used to complete existing annotations (blue) or to map novel transcript structures in suspected loci (gold). Capture oligonucleotides (black bars) are designed to tile across targeted regions. PacBio libraries are prepared for from the captured molecules. Illumina HiSeq short-read sequencing can be carried out for independent validation of predicted splice junctions (SJ). Predicted transcription start sites can be confirmed by CAGE clusters (green), and transcription termination sites by non-genomically encoded polyA+ sequences in PacBio reads (red). Rectangles with lighter shading and dashed outlines denote novel exons. **(b)** A summary of the human and mouse capture library designs. The numbers of individual gene loci probed are shown. PipeR pred., ortholog predictions in mouse genome of human lncRNAs made by PipeR<sup>29</sup>; snRNA, small nuclear RNA; snRNA, small nucleolar RNA; UCE, ultraconserved elements; Prot. coding, expression-matched, randomly selected protein-coding genes; ERCC, spike-in sequences; Ecoli, randomly selected *Escherichia coli* genomic regions (enhancers and UCes were probed on both strands, and these were counted separately). **(c)** Types of RNA samples used in the study.

was even across tissues, with the exception of testis (Supplementary Fig. 2d). ROIs had a median length of 1–1.5 kb (Fig. 2b), in agreement with previous reports<sup>32</sup> and exceeding the average lncRNA annotation of ~0.5 kb (ref. 11).

Capture performance is assessed on the basis of two factors: the ‘on-target’ rate—that is, the proportion of reads originating from probed regions—and enrichment, or the increase in the on-target rate after capture<sup>34</sup>. To estimate these, we sequenced pre- and post-capture libraries with MiSeq. CLS achieved on-target rates of 29.7%/16.5%, representing 19-fold/11-fold enrichment (Fig. 2c,d and Supplementary Fig. 2e). These rates are competitive with values for intergenic lncRNA capture from previous, short-read studies (Supplementary Fig. 2f,g). The majority of off-target signal arose from nontargeted, annotated protein-coding genes (Fig. 2c).

CLS on-target rates were similar to those from previous studies of fragmented cDNA<sup>35</sup> (Supplementary Fig. 2f,g), but lower than those observed with genomic DNA capture. Side-by-side comparisons showed that the capture of long cDNA fragments implies some loss in capture efficiency (Supplementary Fig. 2h,i), as has been observed by others<sup>24</sup>.

We used synthetic spike-in sequences at known concentrations to assess the sensitivity and quantitiveness of our method. We compared



**Figure 2** CLS yields an enriched, long-read transcriptome. **(a)** Sequencing statistics. **(b)** Length distributions of ROIs. Sequencing libraries were prepared from three size-selected cDNA fractions: 1–1.5 kb, 1.5–2.5 kb and >2.5 kb (**Supplementary Fig. 1b,c**). **(c)** A breakdown of sequenced reads by gene biotype, pre- and post-capture, for human samples (equivalent mouse data are presented in **Supplementary Fig. 2j**). The shading denotes the on/off-target status of the reads: green, reads from targeted features, including lncRNAs; gray, reads originating from annotated but not targeted features; yellow, reads from unannotated, nontargeted regions. The ERCC class comprised only those ERCC spike-ins that were probed. When a given read overlapped more than one targeted class of regions, it was counted in each of those classes separately. snoRNA, small nucleolar RNA; snRNA, small nuclear RNA; miRNA, microRNA; UCE, ultraconserved elements. **(d)** A summary of capture performance. The y-axis shows the percentage of all mapped ROIs originating from a targeted region ('on-target'). Enrichment was defined as the ratio of this value in post- versus pre-capture samples. Sequencing was done with MiSeq technology. **(e)** The response of read counts in captured cDNA to the input RNA concentration. Colored circles represent individual data points for 92 spiked-in synthetic ERCC RNA sequences; 42 were probed in the capture design (green), and the remaining 50 were not (violet). Green and purple lines represent linear fits to the corresponding data sets; the parameters are shown at the top of each plot. Given the log–log representation, a linear response of read counts to template concentration should yield an equation of type  $y = c + mx$ , where  $m$  is 1.

the relationship between sequence reads and starting concentration for the 42 probed and 50 nonprobed synthetic ERCC sequences in pre- and post-capture samples (**Fig. 2e**). We found that CLS was notably sensitive, extending detection sensitivity by two orders of magnitude, and was capable of detecting molecules at approximately  $5 \times 10^{-3}$  copies per cell (Online Methods). It was less quantitative than CaptureSeq<sup>24</sup>, particularly at higher concentrations where the slope fell below unity. This suggests saturation of probes by cDNA molecules during hybridization. A degree of noise, as inferred by the coefficient of determination ( $R^2$ ) between read counts and template concentration, was introduced by the capture process.

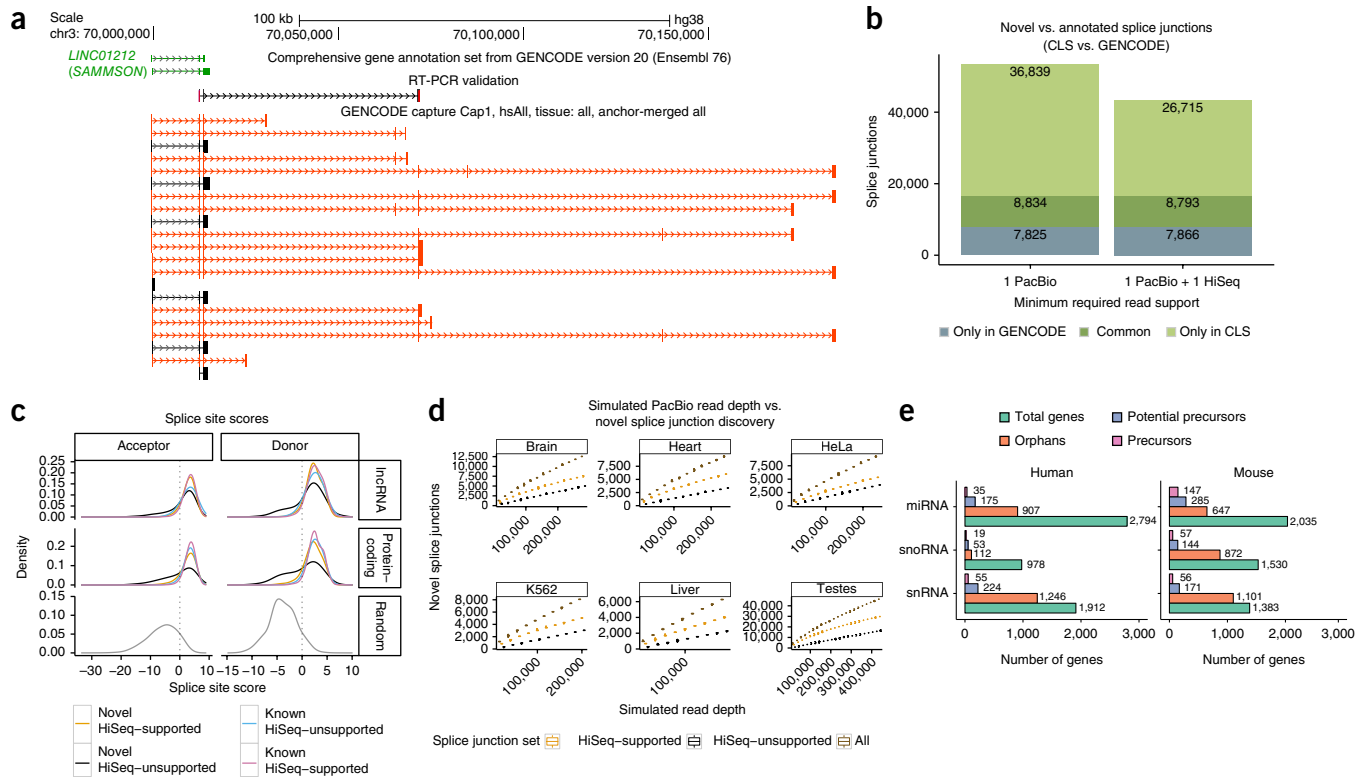
### CLS expands the complexity of known and novel lncRNAs

CLS uncovered a wealth of novel transcript structures in annotated lncRNA loci. In the *SAMMSON* oncogene<sup>36</sup> (*LINC01212*), we discovered previously unannotated exons, splice sites and transcription termination sites (**Fig. 3a**, **Supplementary Figs. 3–5**; examples validated by RT-PCR).

We quantified the amount of newly discovered complexity in targeted lncRNA loci. CLS detected 58%/45% of targeted lncRNA nucleotides and extended these annotations by 6.3/1.6 Mb (86%/64% increase compared with existing annotations) (**Supplementary Fig. 6a**). CLS discovered 45,673/11,038 distinct splice junctions, of which 36,839/8,847 were previously unidentified (**Fig. 3b**, **Supplementary Fig. 6b**). We noted 20,327 novel, high-confidence splice junctions in comparison with a deeper human splice junction reference catalog

composed of both GENCODE v20 and miTranscriptome<sup>7</sup> annotations (**Supplementary Fig. 6c**). For independent validation, and given the relatively high sequence insertion-deletion rate detected in PacBio reads (**Supplementary Fig. 2m**) (an analysis of sequencing error rates is presented in the Online Methods), we deep-sequenced captured cDNA with Illumina HiSeq at an average depth of 35 million/26 million paired-end reads per sample. Split reads from these data exactly matched 78%/75% of splice junctions from CLS. These 'high-confidence' splice junctions alone represent a 160%/111% increase over the existing, probed annotations (**Fig. 3b**, **Supplementary Fig. 6b**). The novel high-confidence lncRNA splice junctions were rather tissue specific, with the greatest numbers observed in testis (**Supplementary Fig. 6d**), and were also discovered across other classes of targeted and nontargeted loci (**Supplementary Fig. 6e**). We observed a greater frequency of intron-retention events in lncRNAs compared with that in protein-coding transcripts (**Supplementary Fig. 6f**).

To evaluate the biological significance of the novel lncRNA splice junctions, we computed their strength with standard position weight matrix models<sup>37</sup> (**Fig. 3c**, **Supplementary Fig. 7a**). High-confidence novel splice junctions from lncRNAs far exceeded the predicted strength of background splice-junction-like dinucleotides and were essentially indistinguishable from annotated splice junctions (**Fig. 3c**). Even unsupported novel splice junctions (**Fig. 3c**) tended to have high scores, although with low-scoring tails. Although they showed little evidence of sequence conservation according to standard measures (similar to lncRNA splice junctions in general; **Supplementary Fig. 7b**),



**Figure 3** Extending known lncRNA gene structures. **(a)** Novel transcript structures from the *SAMMSON* locus. Green, Gencode; black/red, known/novel CLS transcript models, respectively. An RT-PCR-amplified sequence is shown. **(b)** Splice junction discovery. The y-axis represents unique splice junctions for human samples (mouse data are presented in **Supplementary Fig. 6b**) within probed lncRNA loci. Left, all splice junctions; right, high-confidence, HiSeq-supported splice junctions. **Supplementary Figure 6c** shows a comparison to the MiTranscriptome catalog. **(c)** Splice junction motif strength. The plots show the distribution of predicted splice junction strength for splice site acceptors and donors in human samples (mouse data are presented in **Supplementary Fig. 7a**). Splice site strength was computed with GeneID<sup>37</sup>. Data are shown for nonredundant CLS splice junctions from targeted lncRNAs (top), protein-coding genes (middle), and randomly selected splice-site-like dinucleotides (bottom). **(d)** Splice junction discovery/saturation analysis in human samples. The plots show novel splice junctions discovered in simulations with increasing numbers of randomly sampled CLS ROIs. Splice junctions retrieved in each sample were stratified according to the level of support. Each individual box symbol in the box plots summarizes 50 samples. Equivalent mouse data are presented in **Supplementary Figure 8a**, and data for novel transcript model discovery are in **Supplementary Figure 8b**. **(e)** The identification of putative precursor transcripts of small RNA genes. Shown is the count of unique genes for each gene biotype. “Orphans” indicates genes with no annotated overlapping transcript in Gencode that were targeted in the capture library. “Potential precursors” are orphan RNAs residing in the intron of a novel CLS transcript model. “Precursors” reside in the exon of a novel transcript. snoRNA, small nucleolar RNA; snRNA, small nuclear RNA; miRNA, microRNA.

novel splice junctions showed weak but nonrandom evidence of selected function (**Supplementary Fig. 7c**).

We estimated how close these sequencing data were to saturation (i.e., to reaching a definitive annotation). We tested the rate of novel splice junction and transcript model discovery as a function of increasing depth of randomly sampled ROIs (**Fig. 3d**, **Supplementary Fig. 8a,b**). We observed a consistent increase in novelty with increasing depth for both low- and high-confidence splice junctions, up to that presented here. Similarly, no splice-junction-discovery saturation plateau was reached at increasing simulated HiSeq read depths (**Supplementary Fig. 8c**). Thus, considerable additional sequencing is required to complete existing lncRNA gene structures.

Beyond lncRNAs, CLS can be used to characterize other types of transcriptional units. As an illustration, we searched for precursors of small RNAs, whose annotation remains poor<sup>17</sup>. We probed 1-kb windows around all ‘orphan’ small RNAs (i.e., those with no annotated overlapping transcript). Note that although mature small nucleolar RNAs are nonpolyadenylated, they are processed from polyA<sup>+</sup> precursors<sup>38</sup>. We identified more than 100 likely primary transcripts, and hundreds more potential precursors that harbored small RNAs

within their introns (**Fig. 3e**). One interesting example was the cardiac-enriched hsa-miR-143, for which CLS identified a new RT-PCR-supported primary transcript belonging to the *CARMEN1* lncRNA gene (*CARMN*)<sup>39</sup> (**Supplementary Fig. 9**).

### Assembling a full-length lncRNA annotation

A unique benefit of the CLS approach is the ability to identify full-length transcript models with confident 5′ and 3′ termini. ROIs of oligo-dT-primed cDNAs carry a fragment of the poly(A) tail, which can identify the polyadenylation site with base-pair precision<sup>32</sup>. Using conservative filters, we found that 73%/64% of ROIs had identifiable polyadenylation sites (**Supplementary Table 1**) representing 16,961/12,894 novel sites compared with end positions of Gencode annotations. Known and novel polyadenylation sites were preceded by canonical polyadenylation motifs (**Supplementary Fig. 10a–d**). Similarly, the 5′ completeness of ROIs was confirmed by proximity to methyl-guanosine caps identified by cap analysis of gene expression (CAGE)<sup>15</sup> (**Supplementary Fig. 10e**). We used CAGE and polyadenylation sites to define the 5′ and 3′ completeness of all ROIs (**Fig. 4a**).



We developed a pipeline to merge ROIs into a nonredundant collection of transcript models. In contrast to previous approaches<sup>4</sup>, our ‘anchored merging’ method preserved confirmed internal transcription start sites (TSSs) and polyadenylation sites (Fig. 4b). Application of this method to captured ROIs resulted in a greater number of unique transcript models than would have been identified otherwise (Fig. 4c, Supplementary Fig. 11a). We identified 179,993/129,556 transcript models across all biotypes (Supplementary Table 2), 86%/87% of which displayed support of their entire intron chain by captured HiSeq split reads (Supplementary Table 3). In the well-studied *CCAT1* locus<sup>40</sup>, we identified novel full-length transcripts with 5′ and 3′ support (Fig. 4d). CLS here suggested that adjacent *CCAT1* and *CASC19* annotations are fragments of a single gene, a conclusion supported by RT-PCR (Fig. 4d).

Merged transcript models can be defined by their end support: full length (5′ and 3′ supported), 5′ only, 3′ only, or unsupported (Fig. 4b,e). We identified a total of 65,736/44,673 full-length transcript models (Fig. 4e, Supplementary Fig. 11b): 47,672 (73%)/37,244 (83%) arose from protein-coding genes, and 13,071 (20%)/5,329 (12%) from lncRNAs (Supplementary Table 2). An additional 3,742 (6%)/1,258 (3%) represented full-length models that spanned loci of different biotypes (Fig. 1b), usually including one protein-coding gene (‘multi-biotype’). Of the remaining noncoding full-length transcript models, 295/434 were novel, arising from unannotated gene loci. In total, 11,429/4,350 full-length structures arose from probed lncRNA loci, of which 8,494/3,168 (74%/73%) were novel (Supplementary Table 2). We identified at least one full-length transcript model for 19%/12% of the originally probed lncRNA annotations (Fig. 4f, Supplementary Fig. 11c). Independent evidence for gene promoters from DNase I hypersensitivity sites supported our 5′ identification strategy (Fig. 4g). Human lncRNAs with mouse orthologs had considerably more full-length transcript models, although the reverse was not observed (Supplementary Fig. 11d–g). This imbalance might be due to evolutionary factors (for example, the appearance of novel lncRNA isoform complexity during primate evolution) or technical biases; it is noteworthy that we had access to deeper CAGE data for humans than for mice (217,516 versus 129,465 TSSs), and that human lncRNA annotations were more complete than those for mice.

In addition to probed lncRNA loci, CLS also discovered several thousand novel transcript models that originated from unannotated regions and mapped to probed (Fig. 1b) or unprobed regions (Supplementary Fig. 11h,i). These transcript models tended to have lower detection rates (Supplementary Fig. 11j) consistent with low overall expression (Supplementary Fig. 11k) and lower rates of 5′ and 3′ support than probed lncRNAs, although a small number were full length (Fig. 4e, Supplementary Fig. 11b).

We next compared the performance of CLS to that of conventional, short-read CaptureSeq. We took advantage of our HiSeq analysis (212 million/156 million reads) of the same captured cDNAs to make a fair comparison between methods. Short-read methods depend on *in silico* transcriptome assembly; using PacBio reads as a reference, we found that the StringTie tool outperformed Cufflinks, which was used in previous CaptureSeq projects<sup>24,41</sup> (Supplementary Fig. 12a). Using intron chains to compare annotations, we found that CLS identified 69%/114% more novel transcript models than StringTie assembly (Fig. 4h, Supplementary Fig. 12b). CLS transcript models were more complete at 5′ and 3′ ends than StringTie assemblies were, and they were also more complete at the 3′ end compared with probed GENCODE annotations (Fig. 4i, Supplementary Fig. 12d–h). Thus, although StringTie transcript models are slightly longer (Fig. 4j, Supplementary Fig. 12c), they are far less likely to be

full length than CLS models are. This greater length might be attributable to the production of overly long 5′ extensions by StringTie, as suggested by the relatively high CAGE signal density downstream of StringTie TSSs (Supplementary Fig. 12g–h). CLS was more sensitive in the detection of repetitive regions and identified ~20% more repetitive nucleotides in human tissues (Supplementary Fig. 12i).

### Redefining lncRNA promoter and gene characteristics

With a full-length lncRNA catalog, we revisited the basic characteristics of lncRNA and protein-coding genes. lncRNA transcripts, as annotated, are substantially shorter and have fewer exons than mRNAs<sup>5,11</sup>. However, it has remained unresolved whether this is a genuine biological trend or simply the result of annotation incompleteness<sup>21</sup>. When we considered full-length transcript models from CLS, we found that the median lncRNA transcript length was 1,108/1,067 nucleotides, similar to that of mRNAs mapped according to the same criteria (1,240/1,320 nucleotides) (Fig. 5a, Supplementary Fig. 13a). This length difference of 11%/19% was statistically significant ( $P < 2 \times 10^{-16}$  for both human and mouse samples; two-sided Wilcoxon test). These measured lengths are still shorter than those of most annotated protein-coding transcripts (median of 1,543 nucleotides in GENCODE v20), but they are much longer than those of annotated lncRNAs (median of 668 nucleotides). There are two factors that preclude our making firm statements regarding the relative lengths of lncRNAs and mRNAs: the upper length limitation of PacBio reads (Fig. 2b), and the fact that our size-selection protocol selected against shorter transcripts. Nevertheless, we did not find evidence that lncRNAs are substantially shorter<sup>11</sup>. We expect that this issue will be definitively answered with future nanopore sequencing approaches.

In a previous study, we observed enrichment for two-exon genes in lncRNAs<sup>11</sup>. However, the results of the current study show that this was clearly an artifact arising from annotation incompleteness: the mean number of exons for lncRNAs in the full-length models was 4.27, compared with 6.69 for mRNAs (Fig. 5b, Supplementary Fig. 13b). This difference can be explained by lncRNAs’ longer exons, although they peak at approximately 150 bp, or one nucleosomal turn (Supplementary Fig. 13c).

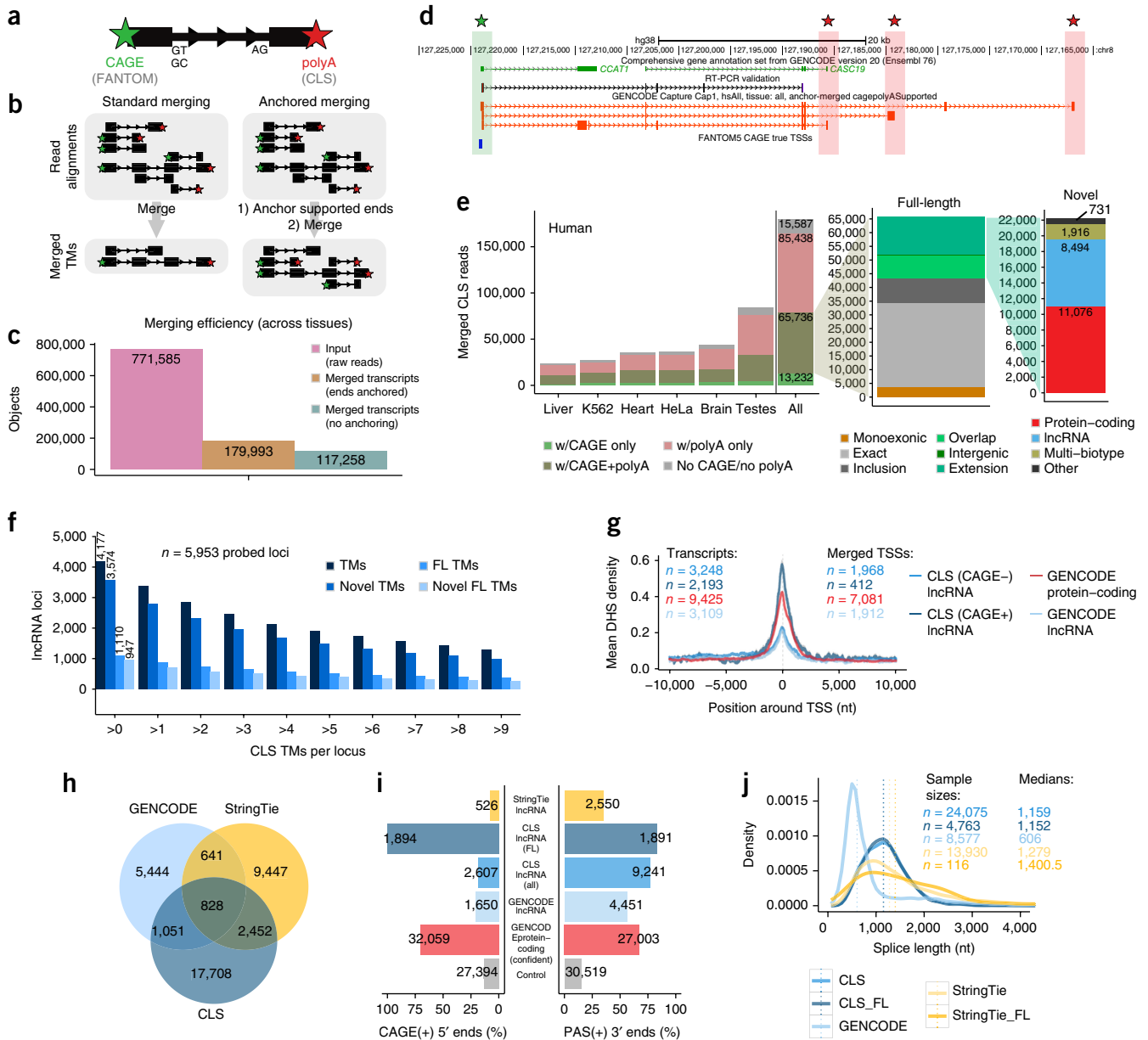
Improvements in TSS annotation are further demonstrated by the fact that full-length transcripts’ TSSs are, on average, closer to expected promoter features, including promoters and enhancers predicted by genome segmentations<sup>42</sup> and CpG islands, although not evolutionarily conserved elements or phenotypic genome-wide association study variants<sup>43</sup> (Fig. 5c). Accurate mapping of lncRNA promoters may provide new hypotheses for the mechanism by which such variants result in observed phenotypes. For example, improved 5′ annotation brings genome-wide association study SNP rs246185 closer to the TSS of RP11-65J2 (ENSG00000262454). Evidence for a functional link between the two is supported by the fact that rs246185 is an expression quantitative trait locus for RP11-65J2, which is expressed in heart and muscle<sup>44</sup> (Supplementary Fig. 13d,e).

The improved 5′ definition provided by CLS transcript models also allowed us to compare lncRNA and mRNA promoters. Recent studies based on the start positions of gene annotations have claimed that strong differences exist between lncRNA and mRNA promoters<sup>45,46</sup>. To make fair comparisons, we created an expression-matched set of mRNAs in HeLa and K562 cells, and removed bidirectional promoters. We compared these across a variety of data sets from ENCODE<sup>12</sup> (Supplementary Figs. 14 and 15).

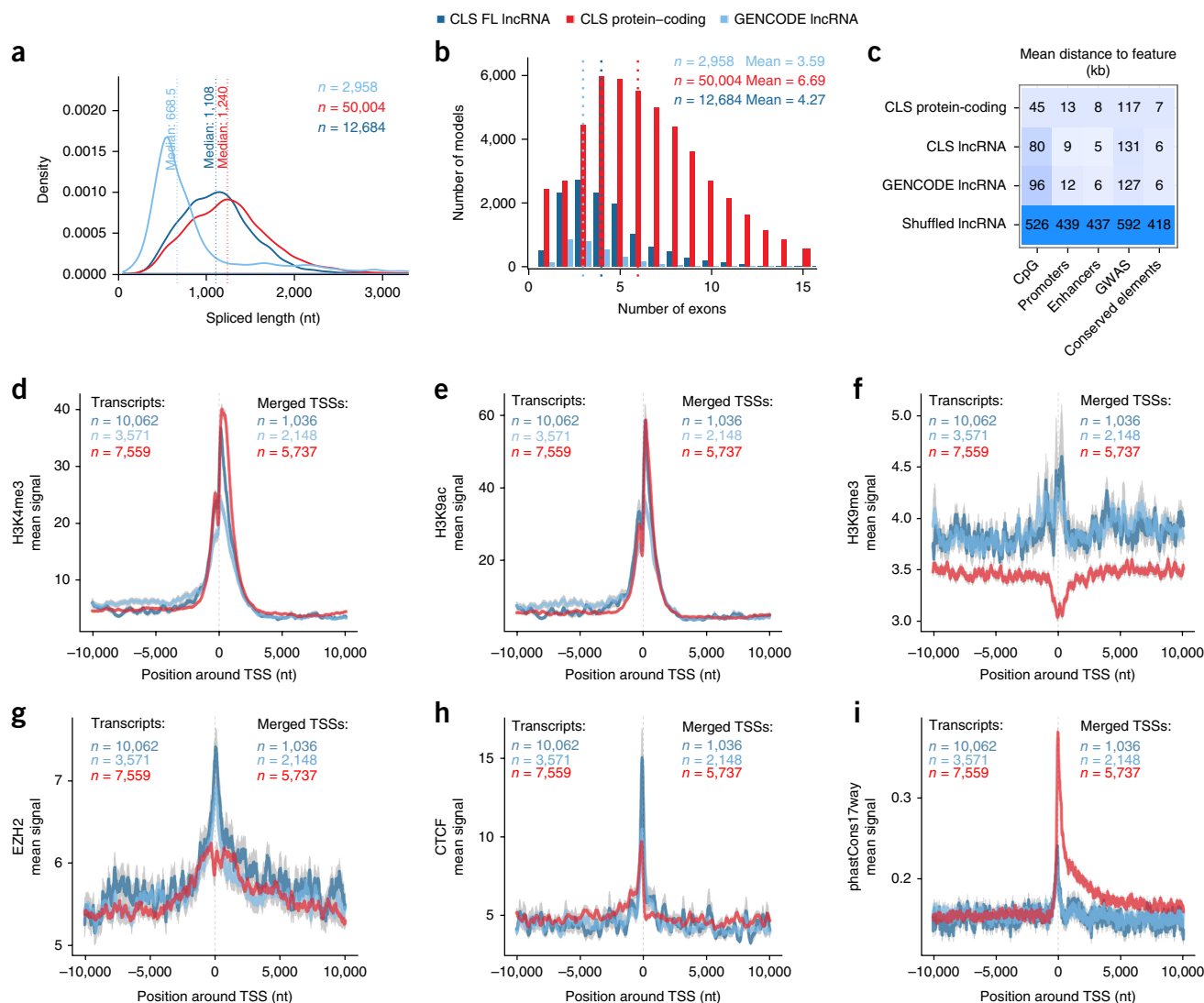
We observed a series of similar and divergent features of lncRNA and mRNA promoters. For example, activating promoter histone modifications such as H3K4me3 (Fig. 5d) and H3K9ac (Fig. 5e)

were essentially indistinguishable between full-length lncRNAs and protein-coding genes, which suggests that, when expression differences are accounted for, the active promoter architecture of lncRNAs

is not unique. The contrast between these findings and previous reports suggests that reliance on annotations alone in prior studies led to inaccurate promoter identification<sup>45,46</sup>.



**Figure 4** Full-length transcript annotation. **(a)** 5' and 3' termini of transcript models (TMs) inferred from CAGE clusters and poly(A) tails in ROIs, respectively. **(b)** In conventional ("standard") transcript merging, TSSs and polyadenylation sites overlapping other exons are lost. "Anchored" merging preserves such sites. **(c)** Anchored merging yields more distinct TMs. The data shown are for human. **(d)** Full-length TMs at the *CCAT1/CASC19* locus. Red, novel full-length TMs; green/red stars, CAGE/poly(A)-supported ends. An RT-PCR-amplified sequence is shown. **(e)** Anchored-merging TMs for human samples (mouse data are presented in **Supplementary Fig. 11b**). In all plots in this panel, the y-axis represents the number of unique TM counts. Left, all anchor-merged TMs, color-coded by end support. Middle, full-length TMs color-coded by novelty compared with GENCODE. Green, novel TMs (subcategories are described in the Online Methods). Right, novel full-length TMs color-coded by biotype. **(f)** Numbers of probed IncRNA loci mapped by CLS at increasing cutoffs for each category in human tissue (mouse data are presented in **Supplementary Fig. 11c**). FL, full-length. **(g)** DNase hypersensitivity site (DHS) coverage of TSSs in HeLa-S3 cells. The y-axis represents the mean DHS density per TSS. Data are plotted as mean values; gray fringes represent the s.e.m. "CAGE+" and "CAGE-" indicate CLS TMs with and without supported 5' ends, respectively. "GENCODE protein-coding" indicates TSSs of protein-coding genes. **(h)** A comparison of IncRNA transcript catalogs from GENCODE, CLS and StringTie within captured regions. The values shown are for human samples; mouse data are presented in **Supplementary Figure 12b–e**. **(i)** 5'/3' transcript completeness, estimated on the basis of CAGE and upstream polyadenylation signals (PAS), respectively. Shown is the proportion of transcript ends with such support (CAGE(+)/PAS(+)). The control was a random sample of internal exons. Data shown are for human tissue; mouse data are presented in **Supplementary Figure 12f**. **(j)** Splice length distributions of transcript catalogs. The dotted lines indicate the median values for the different groups. Data shown are for human samples; mouse data are presented in **Supplementary Figure 12c**.



**Figure 5** Properties of full-length lncRNA transcripts. **(a)** The mature, spliced transcript length of CLS full-length (FL) transcript models from targeted lncRNA loci, transcript models from the targeted and detected GENCODE lncRNA loci, and CLS full-length transcript models from protein-coding loci. **(b)** The number of exons per full-length transcript model from the same groups as in **a**. Dotted lines represent medians for the respective groups. **(c)** Distances from annotated TSSs to genomic features. Each cell shows the mean distance to the nearest neighboring feature for each TSS. TSS sets correspond to the classes from **a**. “Shuffled” indicates full-length lncRNA TSSs randomly placed throughout genome. **(d–i)** A comparison of promoter profiles across gene sets. The aggregate density of various features is shown across the TSSs of the indicated classes. Overlapping TSSs were merged within classes, and TSSs belonging to bidirectional promoters were discarded (Online Methods). The y-axis shows the mean signal per TSS for H3K4me3 **(d)**, H3K9ac **(e)**, H3K9me3 **(f)**, EZH2 **(g)**, CTCF **(h)** and conservation scores across 17 vertebrate species (phastCons17way) **(i)**; gray fringes represent the s.e.m. ChIP-seq experiments were carried out with HeLa cells (Online Methods). Dark blue, full-length lncRNA models from CLS; light blue, the GENCODE annotation models from which the CLS full-length lncRNA models were probed; red, a subset of protein-coding genes with similar expression in HeLa cells as the CLS lncRNAs.

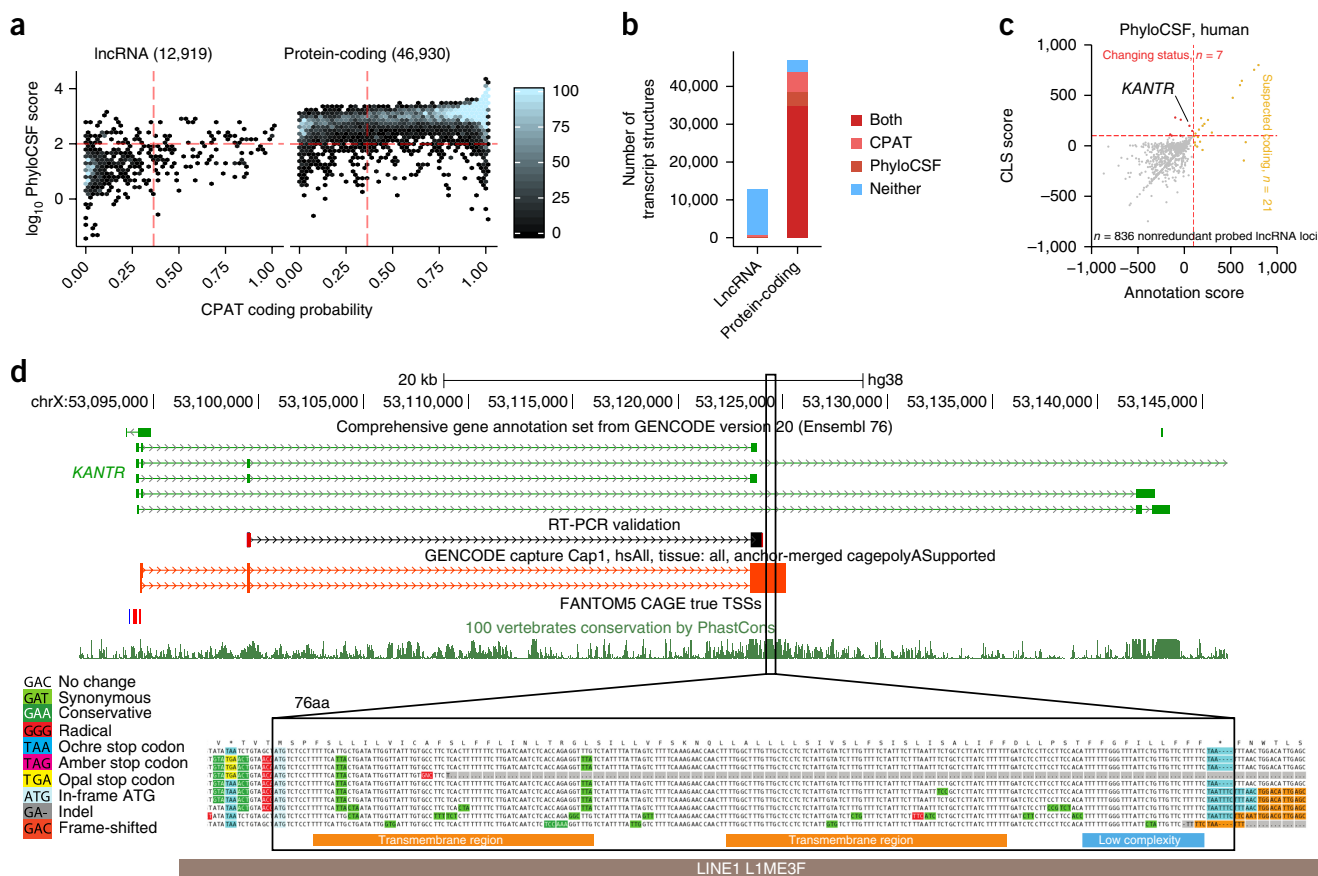
However, as observed previously, lncRNA promoters were distinguished by elevated levels of repressive chromatin marks such as H3K9me3 (**Fig. 5f**) and H3K27me3 (ref. 45) (**Supplementary Figs. 14 and 15**). This may have been a consequence of elevated recruitment to lncRNAs of the Polycomb repressive complex, as evidenced by its subunit Ezh2 (**Fig. 5g**). Promoters of lncRNAs were also distinguished by a localized peak of the insulator protein CTCF (**Fig. 5h**). Finally, there was a clear signal of evolutionary conservation at lncRNA promoters, although it was lower than that for protein-coding genes (**Fig. 5i**).

Two conclusions can be drawn. First, CLS-inferred TSSs have a greater density of expected promoter features compared with probed

annotations, thus demonstrating that CLS improves TSS annotation. Second, after adjustment for expression, lncRNAs have similar activating histone modifications, but distinct repressive modifications, compared with protein-coding genes.

#### Discovery of new potential open reading frames

A number of studies have suggested that lncRNA loci encode peptide sequences through unannotated open reading frames (ORFs)<sup>47,48</sup>. We searched for signals of protein-coding potential in full-length models by using two complementary methods based on evolutionary conservation and intrinsic sequence features<sup>49,50</sup> (**Fig. 6a**, Online Methods, **Supplementary Data Set 1**). This analysis revealed evidence for



**Figure 6** Protein-coding potential of full-length lncRNAs. **(a)** The predicted protein-coding potential of all full-length transcript models mapped to lncRNA (left) or protein-coding loci (right). Points represent full-length transcript models. The y-axis indicates the coding likelihood according to PhyloCSF, based on multiple genome alignments; the x-axis shows the likelihood calculated by CPAT, an alignment-free method. Red dashed lines indicate score thresholds, and values above those thresholds were considered representative of protein coding. Transcript models that mapped to multiple biotypes were not considered. **(b)** Numbers of classified transcript models from **a**. **(c)** The discovery of new protein-coding transcripts in full-length CLS reads with PhyloCSF. For each probed GENCODE gene annotation, we took the score of the best ORF across all transcripts (x-axis) and the score of the best ORF in the corresponding full-length CLS transcript models (y-axis). Yellow, loci from the GENCODE v20 annotation predicted to encode proteins; red, lncRNA loci where new ORFs were discovered as a result of CLS transcript models. **(d)** *KANTR*, an example of an annotated lncRNA locus where a novel protein-coding sequence was discovered. Top, the structure of the lncRNA and the associated ORF (highlighted region) falling within the range of novel full-length CLS transcripts (red). Note how this ORF lies outside the existing annotation (green) and overlaps a highly conserved region (see the PhastCons conservation track below). A sequence obtained by RT-PCR (black) is also shown. Bottom, conservative substitutions in the predicted 76-amino-acid ORF consistent with a functional peptide, generated by CodAlignView (“URLs”). High-confidence predicted SMART<sup>53</sup> domains are indicated by colored bars. This ORF lies within and antisense to an L1 transposable element (gray bar).

protein-coding potential in a small fraction of lncRNA full-length transcript models (109 of 1,271, or 8.6%), although a similar number of protein-coding full-length transcripts showed no evidence of protein coding (2,900 of 42,758, or 6.8%) (**Fig. 6b**).

CLS full-length models supported reclassification of protein-coding potential for five distinct gene loci (**Fig. 6c**, **Supplementary Fig. 16a**, **Supplementary Data Set 2**). A good example is the *KANTR* locus, where extension by CLS (supported by independent RT-PCR) identified a placental-mammal-conserved 76-amino-acid ORF with no detectable protein ortholog<sup>51</sup>. It is composed of two sequential transmembrane domains (**Fig. 6d**, **Supplementary Fig. 16e**) and derives from a LINE1 transposable element. Another case is *LINC01138*, linked to prostate cancer, for which a potential 42-amino-acid ORF was found in the extended transcript<sup>52</sup>. We could not find peptide evidence for translation of either ORF (Online Methods). Whole-cell expression, as well as cytoplasmic-to-nuclear distributions, also showed that the behavior of potentially protein-coding

lncRNAs was consistently more similar to that of annotated lncRNAs than to that of mRNAs (**Supplementary Fig. 16b–d**). Hence, CLS will be useful in improving biotype annotation of the small minority of lncRNAs that may encode proteins.

## DISCUSSION

We have introduced an annotation methodology that addresses the competing needs of quality and throughput. Capture long-read sequencing produces transcript models with quality approaching that of human annotators, yet with throughput similar to that of *in silico* transcriptome reconstruction. CLS improves upon existing assembly-based methods through not only confident exon connectivity but also (1) far higher rates of 5′ and 3′ completeness and (2) the carrying of encoded poly(A) tails.

CLS is also competitive in economic terms. Using conservative estimates with 2016 prices (\$2,460 for one lane of PE125bp HiSeq, and \$500 for one SMRT), and including the cost of sequencing alone,



we estimate that CLS yielded one novel, full-length lncRNA structure for every \$8 spent, compared with \$27 with conventional CaptureSeq. This difference is due to the greater rate of full-length transcript discovery by CLS.

Despite its advantages, CLS could still be optimized in several respects. First, the capture efficiency for long cDNAs can be improved by several-fold. Second, various technical factors limit the completeness of CLS transcript models, including sequencing reads that remain shorter than many transcripts, incomplete reverse transcription of the RNA template, and degradation of RNA molecules before reverse transcription. Resolution of these issues will be an important objective of future protocol improvements, and only after it has been achieved can we make definitive judgments about lncRNA transcript properties. In recent work separate from the current study, we further optimized the capture protocol, pushing on-target rates to around 35% (Online Methods and data not shown). However, the most dramatic gains in the cost-effectiveness and completeness of CLS will come from advances in sequencing technology. The latest nanopore cDNA sequencing promises to be ~150-fold less expensive per read than PacBio technology (0.01 versus 15 cents per read, respectively).

Full-length annotations have provided the most confident view so far of lncRNA gene properties. lncRNAs are more similar to mRNAs than previously thought in terms of splice length and exon count<sup>11</sup>. We noted a similar trend for promoters: when lncRNA promoters were accurately mapped by CLS and compared with expression-matched protein-coding genes, we found them to be surprisingly similar in terms of activating modifications. This suggests that previous studies that placed confidence in annotations of TSSs should be reassessed<sup>45,46</sup>. On the other hand, lncRNA promoters do have unique properties, including elevated levels of repressive histone modification, recruitment of Polycomb group proteins, and interaction with the insulator protein CTCF. To our knowledge, this is the first report to suggest a relationship between lncRNAs and insulator elements. Overall, these results suggest that lncRNA gene features per se are generally similar to those of mRNAs, after normalization for differences in expression. Finally, extended transcript models did not yield evidence for widespread protein-coding capacity encoded in lncRNAs.

Despite our success in mapping novel structures in annotated lncRNAs, we observed surprisingly low numbers of transcript models originating in the relatively fewer numbers of unannotated loci that we probed, including ultraconserved elements and developmental enhancers. This suggests that, at least in the tissue samples probed here, such elements do not give rise to substantial numbers of lncRNA-like, polyA+ transcripts.

In summary, by resolving a longstanding roadblock in lncRNA transcript annotation, the CLS approach promises to accelerate progress toward an eventual 'complete' mammalian transcriptome annotation. These updated lncRNA catalogs represent a valuable resource for the genomic and biomedical communities, and address fundamental issues of lncRNA biology.

**URLs.** CLS data portal, [https://public\\_docs.crg.es/rguigo/CLS/](https://public_docs.crg.es/rguigo/CLS/); pre-loaded CLS UCSC Genome Browser track hub, [http://genome-euro.ucsc.edu/cgi-bin/hgTracks?hubUrl=http://public\\_docs.crg.es/rguigo/CLS/data/trackHub/hub.txt](http://genome-euro.ucsc.edu/cgi-bin/hgTracks?hubUrl=http://public_docs.crg.es/rguigo/CLS/data/trackHub/hub.txt); CodAlign View, <https://data.broadinstitute.org/compbio1/cav.php>; ENCODE mycoplasma contamination guidelines, [https://www.encodeproject.org/documents/60b6b535-870f-436b-8943-a7e5787358eb/@@download/attachment/Cell\\_Culture\\_Guidelines.pdf](https://www.encodeproject.org/documents/60b6b535-870f-436b-8943-a7e5787358eb/@@download/attachment/Cell_Culture_Guidelines.pdf).

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGMENTS

We thank members of the Guigó laboratory for their valuable input and help with sample handling, data analysis and writing of the manuscript, including E. Palumbo, F. Reverter, A. Breschi, D. Pervouchine, C. Arnan and F. Camara. We thank L. Armengol (qGenomics) for advice on RNA capture, D. Garrido (CRG) for help with eQTL analysis, S. Bonnin (CRG) for help with data manipulation in R, and I. Jungreis (MIT) for advice on PhyloCSE. J. Wright and J. Choudhary (Sanger Institute) helped with the search for peptide hits to putative coding regions. S. Djebali (INRA, France) kindly made available the Compmerge utility. This work and its publication were supported by the National Human Genome Research Institute of the US National Institutes of Health (grants U41HG007234, U41HG007000 and U54HG007004) and the Wellcome Trust (grant WT098051 to R.G.). R.J. was supported by the Ramón y Cajal Subprogram of the Spanish Ministry of Economy and Competitiveness (grant RYC-2011-08851). Work in the laboratory of R.G. was supported by the National Human Genome Research Institute (awards U54HG0070, R01MH101814 and U41HG007234). This research was partly supported by NCCR RNA & Disease, funded by the Swiss National Science Foundation (to R.J.). We thank R. Garrido (CRG) for administrative support. We acknowledge support from the Spanish Ministry of Economy and Competitiveness, Centro de Excelencia Severo Ochoa 2013–2017 (SEV-2012-0208), and from the CERCA Programme, Generalitat de Catalunya.

## AUTHOR CONTRIBUTIONS

R.J., R.G., J.H., A.F., B.U.-R. and J.L. designed the experiment. S.C. generated cDNA libraries and performed the capture. C.D. and T.R.G. carried out PacBio sequencing of capture libraries. J.L. and B.U.-R. analyzed the data under the supervision of R.G. and R.J. R.J. wrote the manuscript, with contributions from J.L., B.U.-R. and R.G. S.P.-L. and A.A. performed the RT-PCR experiments.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

- Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
- Jia, H. *et al.* Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA* **16**, 1478–1487 (2010).
- Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
- Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
- Cabili, M.N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
- Hangauer, M.J., Vaughn, I.W. & McManus, M.T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet.* **9**, e1003569 (2013).
- Iyer, M.K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
- Zhao, Y. *et al.* NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* **44**, D203–D208 (2016).
- Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
- Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
- Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
- Bernstein, B.E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Chen, L. *et al.* Transcriptional diversity during lineage commitment of human blood progenitors. *Science* **345**, 1251033 (2014).
- Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

15. Forrest, A.R.R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
16. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
17. Georgakilas, G. *et al.* microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. *Nat. Commun.* **5**, 5700 (2014).
18. Ørom, U.A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46–58 (2010).
19. Ferdin, J. *et al.* HINCUTs in cancer: hypoxia-induced noncoding ultraconserved transcripts. *Cell Death Differ.* **20**, 1675–1687 (2013).
20. Calin, G.A. *et al.* Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* **12**, 215–229 (2007).
21. Lagarde, J. *et al.* Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq). *Nat. Commun.* **7**, 12339 (2016).
22. Mercer, T.R. *et al.* Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* **30**, 99–104 (2011).
23. Bussotti, G. *et al.* Improved definition of the mouse transcriptome via targeted RNA sequencing. *Genome Res.* **26**, 705–716 (2016).
24. Clark, M.B. *et al.* Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat. Methods* **12**, 339–342 (2015).
25. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
26. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014).
27. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L.A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007).
28. Dimitrieva, S. & Bucher, P. UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.* **41**, D101–D109 (2013).
29. Bussotti, G. *et al.* BlastR—fast and accurate database searches for non-coding RNAs. *Nucleic Acids Res.* **39**, 6886–6895 (2011).
30. Kralj, J.G. & Salit, M.L. Characterization of in vitro transcription amplification linearity and variability in the low copy number regime using External RNA Control Consortium (ERCC) spike-ins. *Anal. Bioanal. Chem.* **405**, 315–320 (2013).
31. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
32. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013).
33. Quail, M.A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).
34. Mercer, T.R. *et al.* Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* **9**, 989–1009 (2014).
35. García-García, G. *et al.* Assessment of the latest NGS enrichment capture methods in clinical context. *Sci. Rep.* **6**, 20948 (2016).
36. Leucci, E. *et al.* Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* **531**, 518–522 (2016).
37. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Curr. Protoc. Bioinformatics* **Chapter 4**, Unit 4.3 (2007).
38. Smith, C.M. & Steitz, J.A. Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5′-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol. Cell. Biol.* **18**, 6897–6909 (1998).
39. Ounzain, S. *et al.* CARMEN, a human super enhancer-associated long noncoding RNA controlling cardiac specification, differentiation and homeostasis. *J. Mol. Cell. Cardiol.* **89**, 98–112 (2015).
40. Nissan, A. *et al.* Colon cancer associated transcript-1: a novel RNA expressed in malignant and pre-malignant human tissues. *Int. J. Cancer* **130**, 1598–1606 (2012).
41. Perteza, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
42. Marques, A.C. *et al.* Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biol.* **14**, R131 (2013).
43. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
44. Arking, D.E. *et al.* Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nat. Genet.* **46**, 826–836 (2014).
45. Alam, T. *et al.* Promoter analysis reveals globally differential regulation of human long non-coding RNA and protein-coding genes. *PLoS One* **9**, e109443 (2014).
46. Melé, M. *et al.* Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res.* **27**, 27–37 (2017).
47. Mackowiak, S.D. *et al.* Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* **16**, 179 (2015).
48. Bazzini, A.A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33**, 981–993 (2014).
49. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74 (2013).
50. Lin, M.F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
51. Sauvageau, M. *et al.* Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife* **2**, e01749 (2013).
52. Wan, X. *et al.* Identification of androgen-responsive lincRNAs as diagnostic and prognostic markers for prostate cancer. *Oncotarget* **7**, 60503–60518 (2016).
53. Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.* **43**, D257–D260 (2015).

## ONLINE METHODS

**Library design.** *Design of human capture probes.* All designs were based on the GENCODE<sup>10</sup> version 20 annotation in human genome build hg38. For probe design, a target annotation was prepared in FASTA format and composed of sets of features. In each case, the entire set of features of each class was taken as a starting point, unless otherwise stated, and where necessary was lifted over to the hg38 assembly. Features that overlapped protein-coding gene loci were removed. Intergenic lncRNAs were extracted from the GENCODE v20 annotation and were taken as all those genes with no single transcript that overlapped or lay within 5 kb of any protein-coding gene. For small RNA loci, a 1-kb window centered on the small RNA was targeted.

At this stage, we quantified the expression of candidate regions with HBM/ENCODE RNA-seq data from appropriate human tissues and cell lines. We noticed that the top 20 most expressed features (mean expression across samples) produced approximately 71% of sequencing reads (**Supplementary Fig. 17**), and we removed these in order to favor rarer transcripts. A number of controls were added to the design. We included 100 protein-coding genes, with steady-state levels matched to the distribution of lncRNAs, and 100 random 1-kb genomic regions from the *Escherichia coli* genome. As additional negative controls, we included 100 intergenic regions of 1 kb each with no evidence from ENCODE ChromHMM for any transcriptional or regulatory activity<sup>54</sup>. Finally, out of the 92 ERCC sequences, we removed the top 8 most concentrated, and we selected half of the remainder ( $n = 42$ ) such that they evenly covered the concentration distribution. In total the design targeted 14,667 regions, which corresponded to ~15.5 Mb of human genome (hg38) and exons of 9,560 lincRNAs from 5,953 loci. The summary information for selected transcript targets in human is provided in **Supplementary Table 4**. Statistics on probed gene loci are presented in **Figure 1b**.

All targets were combined into a single FASTA file and submitted to Roche NimbleGen (Madison, WI) for probe design. The oligonucleotide probes were designed and synthesized as a SeqCap EZ Choice XL library according to the manufacturer's protocol. The oligonucleotide probes covered 86.6% of target regions directly, with an estimated 96.1% of target regions successfully targeted. Roche NimbleGen's policy prohibits the release of SeqCap's probe coordinates, but the design is available from the corresponding authors on request.

*Design of mouse capture probes.* Mouse library design was carried out essentially as for the human library, with some differences. All designs were based on the GENCODE version M3 annotation in genome build mm10. Candidate lncRNAs were filtered to remove those that overlapped any protein-coding gene within 5 kb. Homology-based predictions of mouse orthologs of human lncRNA were obtained via the PipeR pipeline<sup>29</sup>. As before, the top 20 most expressed lncRNAs, as estimated from ENCODE<sup>31</sup> RNA-seq data, were removed. The final design covered 8,708 regions, including 2,817 GENCODE vM3 lincRNA transcripts from 1,920 loci. The covered regions corresponded to 8.3 Mb. The summary information for selected transcript targets in mouse is provided in **Supplementary Table 5**. Statistics on probed gene loci are presented in **Figure 1b**.

Designed oligonucleotide probes covered 76.3% of target regions directly and 85.0% of target regions successfully targeted. Oligonucleotide probes were synthesized as a Roche NimbleGen SeqCap EZ Choice XL library. Roche NimbleGen's policy prohibits the release of SeqCap's probe coordinates, but the design is available from the corresponding authors on request.

**Sample preparation.** *RNA samples.* Commercial total RNA samples were obtained for four different adult human (Ambion AM6000) and mouse (Clontech 636644) tissues: heart, testes, liver and brain. We also obtained mouse E7 and E15 samples from the same panel. Human K562 and HeLa RNA was obtained directly from members of the ENCODE consortium<sup>31</sup>. Neither cell line used in this study is listed in the database of commonly misidentified cell lines maintained by ICLAC. Cell lines were not authenticated. Cell lines were tested for mycoplasma contamination as per ENCODE guidelines ("URLs"). The integrity of samples was tested by Bioanalyzer (Agilent), and all samples had values of 8.5 or higher. To 4  $\mu$ g of each RNA sample, we added 4  $\mu$ l of 1:100-diluted ERCC mix (Ambion 4456740) according to the manufacturer's protocol (**Supplementary Table 6**). Mixes 1 and 2 were assigned to samples as

described below. The samples containing ERCC controls were ribodepleted with Ribo-Zero (Epicentre; MRZE724), and successful rRNA removal was validated by Bioanalyzer.

*cDNA synthesis.* Full-length cDNA was synthesized via reverse-transcription of ribosome-depleted RNA samples with the SMARTer PCR cDNA synthesis kit (Clontech; 634926) and the Advantage 2 PCR kit (Clontech; 639206). Each cDNA was synthesized from 3.5  $\mu$ l of ribosome-depleted RNA according to the manufacturer's protocol, and two independent cDNA synthesis reactions were carried out for each sample. cDNA was primed with oligo(dT). The adaptors used in the cDNA library construction sequences (SMART IV oligonucleotide and CDS III/3' PCR primer) are listed in **Supplementary Data Set 3**.

All first-strand RNA obtained from the reaction was used for second-strand synthesis. We modified the synthesis cycling protocol from that specified by the manufacturer by increasing the extension time from 3 to 6 min to favor the synthesis of long strands. After protocol optimization, a total of 18 cycles was used to obtain the full-length cDNA libraries. The resulting cDNA was quantified with a NanoDrop ND-1000 full-spectrum spectrophotometer (Thermo Scientific). The library length and quality were also verified by Bioanalyzer.

**Capture.** *Library preparation.* cDNA samples were used to create barcoded, full-length libraries. The two aliquots of cDNA obtained in the preceding step were pooled, and 1  $\mu$ g was used for library preparation. One adenine was added to blunt cDNA 3' extremities, and Illumina Truseq adaptors were ligated. Different barcoded adaptor hexamer indexes were used to discriminate each sample (**Supplementary Table 7** and **Supplementary Data Set 3**). The overall structure of cDNA libraries is represented schematically in **Supplementary Figure 2c**.

The library was amplified for ten PCR cycles under standard Kapa Biosystems PCR conditions (low-throughput library prep; Kapa Biosystems, KK8232), except that the PCR extension step was increased to 3 min to allow long fragments to be fully amplified. The quality and length of libraries were checked with an Agilent 2100 Bioanalyzer. Library quantification was done with Qubit dsDNA BR assays (Thermo Fisher). For each cDNA sample, an additional Covaris-fragmented Illumina sequencing library was prepared for MiSeq and HiSeq sequencing according to standard protocols.

Standard Illumina 6-mer indexes were used for compatibility with blocking oligonucleotides in the SeqCap capture protocol (see below). We note that the use of these relatively short indexes led to the loss of information during later demultiplexing steps. Improving this issue through the use of standard 16-nt PacBio indexes should be a priority in future versions of CLS.

*Sample pooling.* Samples were pooled separately by species, such that all six human libraries were mixed at equimolar ratios, and similarly for mouse libraries. The final amount of each pool was 1  $\mu$ g.

*cDNA capture.* Human and mouse pools were dried and prepared for hybridization to NimbleGen SeqCap EZ Choice XL library capture probes according to the manufacturer's protocol (SeqCap EZ Library SR User's Guide Version 5.0). Hybridization was carried out for 72 h. A total of five separate parallel captures were performed for each species; four were used for subsequent PacBio sequencing, and the one remaining sample was used for Illumina sequencing.

Subsequent to the presented work, we managed to further optimize the efficiency of this capture process by implementing four changes to the described protocol:

1. Dry cDNA for resuspension before capture at 60 °C instead of 55 °C
2. Hybridization incubation time: 20 h instead of 72 h
3. For washing steps after capture, use a water bath instead of a dry bath
4. Blockers: additional blockers targeting the SMARTer adaptors used during library construction (sequences in **Supplementary Data Set 3**, "SMARTer\_blocker" and "SMARTer\_5p\_PCR\_blocker")

*Amplification and quality control of captured cDNA.* After hybridization, human and mouse pools were washed with m-280 streptavidin Dynabeads (Invitrogen 11205D) to eliminate nonspecific hybridization according to the recommendations in the Roche protocol. Human and mouse washed pools were PCR-amplified with Kapa HotStart ReadyMix 2X (Kapa Biosystems; KK1006). Two independent PCR reactions containing half of the washed



pool each were prepared to avoid PCR duplicates. Eighteen PCR cycles were performed, with an increased extension step of 3 min to allow long fragments to be fully amplified. The length of post-capture PacBio and Illumina libraries was verified by Bioanalyzer, and quantity was verified by Qubit.

**PacBio sequencing of captured cDNA.** *Pooling.* After quantification and quality control, the four post-capture libraries were pooled together by species to produce one unique human and one unique mouse pool. The 110  $\mu$ l of each sample were again quantified by Qubit dsDNA BR assay (Thermo Fisher), with 12.3  $\mu$ g for human and 9.57  $\mu$ g for mouse.

*Size selection.* Samples were subsequently size-selected with E-gel (Invitrogen) into three different ranges: 1,000–1,500 bp, 1,500–2,500 bp and >2,500 bp. We collected two shorter fractions of 200–500 bp and 500–1,000 bp, but after reviewing the preliminary sequencing data we decided not to scale them up because of the large number of reads in this size range obtained in the larger fractions. After size selection, each size fraction was dried and resuspended with 20  $\mu$ l of water and quantified by Qubit dsDNA BR assay (Thermo Fisher). These samples were then amplified again by PCR (four cycles) with Kapa HiFi HotStart (Kapa Biosystems) to reach the required amount for PacBio library preparation. The quality and length of obtained libraries were verified with Bioanalyzer and Qubit.

We checked the efficiency of size selection via analysis of spike-in sequences (Supplementary Fig. 1d). For each size-selected captured library, and for pre-capture libraries, we calculated the sequencing efficiency as a function of spike-in sequence length. Sequencing efficiency was defined for each spike-in sequence as follows: (number of reads)/(molar concentration  $\times$  sequence length  $\times$  total read count). This showed that, as expected, size selection boosted the sequencing of longer templates.

*PacBio library preparation.* Approximately 2  $\mu$ g of each of the size-fractionated and amplified DNAs was used for each of the human and mouse pools, for a total of 6 (3  $\times$  2) distinct samples. Sizes and concentrations were verified by Bioanalyzer. PacBio libraries were constructed for each sample with kit #100-250-100 (Pacific Biosciences) as per the manufacturer's protocol. Briefly, this involved polishing the PCR amplicon ends to 'blunt' them, ligating the SMRTbell adaptors, removing linear (nonligated) fragments of DNA, and carrying out AMPure bead purification followed by Bioanalyzer analysis to assess the size distribution and Qubit quantifications.

*PacBio sequencing and collection of post-capture data.* We ran each of the PacBio libraries on an initial SMRT cell to assess their respective performance and optimal sequencing concentration. Those that performed well were then scaled up to an additional 20 SMRT cells for deep data collection. The PacBio reagents and metrics used for each sample are listed in Supplementary Table 8. The sequencing was performed on a PacBio RSII instrument. Upon completion of the sequencing, SMRT cells from a given library were aggregated on SMRT Portal, and the PacBio post-processing method "RS\_ReadsOfInsert.1" was run on each aggregated sample to generate ROIs for downstream processing. This yielded a single FASTQ file per library.

**HiSeq sequencing of captured cDNA.** Post-capture Illumina cDNA libraries were sequenced on a HiSeq 2500 machine (2  $\times$  125 nt, v4, high-output mode). One sequencing lane was generated per species at a depth of ~212 million (human) or ~156 million (mouse) pairs of reads. Read pairs were demultiplexed with Illumina software. Note that these libraries were unstranded and Covaris-fragmented before capture.

**Demultiplexing of ROIs according to sample barcodes.** As previously mentioned, PacBio reads contained Illumina Truseq adaptors, universal (59 nt) and indexed (65 nt), that flanked targeted cDNAs (Supplementary Fig. 2c). To demultiplex samples (i.e., to determine the tissue of origin of each ROI), for each adaptor we selected its middle 26 nt. Each of the 26-mers derived from the indexed adaptors contained the hexamer barcode in the center. We used the GEM mapper<sup>55</sup> to demultiplex samples. PacBio reads were compiled into a FASTA file (one file per species) and indexed by GEM. Mapping the middle 26-mer of indexed adaptors to the PacBio read allowed us to assign it to its tissue of origin. The additional presence of the universal adaptor within ROIs was used to confirm the completeness of the insert. The GEM-based demultiplexing procedure allowed up to three mismatches ( $-m$  0.1) and

three indels ( $-e$  0.1) for accurate identification of the barcodes. The following non-default GEM parameters were used during the mapping step:  $-T$  3  $--max-big-indel-length$  0  $-s$  3  $-D$  4. We filtered out 'chimeric' ROIs (that is, reads arising from the concatenation of inserts during adaptor ligation) by removing those reads that contained more than one indexed or more than one universal TruSeq Illumina adaptor sequence.

Overall, we were able to demultiplex 1,627,322 and 1,509,374 ROIs in human and mouse samples, respectively (Fig. 2a, Supplementary Fig. 2b). As shown in Supplementary Figure 2d, only a minute fraction of human ROIs were assigned a mouse barcode (and vice versa), which highlights the high specificity of the demultiplexing procedure.

**Read-mapping.** All read-to-genome alignments were performed on genome assemblies GRCh38/hg38 (human) and GRCm38/mm10 (mouse). Mapping of ROIs from post-capture PacBio libraries to human and mouse genomes (in addition to sequences of 96 ERCC spike-in controls) was done with STAR<sup>56</sup> (v.2.4.0.1) compiled for long reads. For improved accuracy in splice junction mapping, a reference annotation was provided as a guide to the aligner. The reference annotation for human genes was built with the GENCODE v20 set and sequences of all other targeted regions. For mouse genes, exonic sequences of PipeR predictions along with sequences of all other additional targets were added to the reference annotation of GENCODE vM3. The following non-default parameters were used during the mapping step:  $--outFilterMultimapScoreRange$  20  $--outFilterScoreMinOverLread$  0  $--outFilterMatchNminOverLread$  0.5  $--outFilterMismatchNmax$  1000  $--winAnchorMultimapNmax$  200  $--seedSearchStartLmax$  50  $--seedPerReadNmax$  100000  $--seedPerWindowNmax$  100  $--alignTranscriptsPerReadNmax$  100000  $--alignTranscriptsPerWindowNmax$   $--genomeSAsparseD$  4  $--outsAMunmapped Within$   $--runThreadN$  6.

For analysis of MiSeq (pre-capture cDNA) and HiSeq (post-capture) data, FASTQ files were aligned to the human and mouse genomes (plus the sequences of 96 ERCC spike-in controls) with STAR<sup>56</sup> (v.2.4.0.1) compiled for short reads. The reference annotations described above were used to guide the mapper. To maximize the mapping rate, we aligned the mates of each pair of reads separately. The following non-default STAR parameters were specified:  $--outFilterMismatchNoverLmax$  0.04  $--alignIntronMin$  20  $--alignIntronMax$  1000000  $--alignMatesGapMax$  1000000  $--outsAMunmapped Within$   $--runThreadN$  6.

**Analysis of CLS performance and on-target enrichment.** *RNA-capture on-target enrichment.* We evaluated the overall RNA-capture performance by calculating an on-target rate in both MiSeq pre-capture and PacBio post-capture libraries. The on-target rate was defined as the ratio of the number of distinct ROIs mapping to targeted genomic regions (excluding ERCC RNA spike-in controls) to the total number of mapped ROIs. The number of reads overlapping targeted regions was calculated directly from the STAR BAM file with bedtools intersect<sup>57</sup>. Overlap was defined as  $\geq 1$  bp of intersection between the sequencing read and the exonic span of a feature on the same strand. The overall on-target fold enrichment was computed as the on-target rate in the post-capture library divided by the on-target rate in the pre-capture library.

We calculated enrichment separately by referencing two distinct sequencing data sets of post-capture cDNA: (a) the main PacBio reads, and (b) Illumina MiSeq of the same material. Figure 2d shows data for enrichments calculated with the latter data set: MiSeq post-capture versus MiSeq pre-capture. Equivalent enrichments for the former comparison (PacBio post-capture versus MiSeq pre-capture) were 16.6-fold/11.1-fold for human/mouse.

We compared CLS enrichments to values from a previous capture short-read sequencing (CSS) study<sup>24</sup>. We focused our analysis on the CSS tissues that were also assayed in CLS (human brain, heart, liver and testis), and computed on-target rates on lincRNAs more than 5 kb away from any protein-coding gene in both studies, based on GENCODE v20 and v19 for CLS and CSS, respectively. CSS pre-capture rates were estimated from pre-capture MiSeq libraries generated in the present work, and remapped to hg19/GENCODE 19. Across the four tissues studied, CLS outperformed CSS in terms of both on-target enrichment (in all samples) and post-capture on-target rate (in brain and testis only) (Supplementary Fig. 2f,g).



*Breakdown of sequencing reads by gene biotype.* Both human and mouse genomes, as well as ERCC spike-in sequences, were segmented into distinct classes of locus regions according to their gene biotype annotation and capture status (i.e., on-target versus off-target). The on- and off-target categories corresponded to standard, GENCODE-annotated gene biotypes (in simplified categories, as described in **Supplementary Note 1**, in addition to “Other,” which comprised mitochondrial genes), whereas the “Intergenic” class included all nontargeted and unannotated genome segments. Next, we calculated the proportion of pre- and post-capture MiSeq reads originating from each genome partition, using the read BAM files and the bedtools coverage utility<sup>57</sup>. Note that when a given read overlapped multiple regions of distinct biotype classes, it was counted in each of those classes separately. Secondary targets (i.e., genes that were not targeted per se but that overlapped targeted regions) were included in on-target biotype subclasses. The following additional hierarchical rules were applied in the assignment: the highest priority in the read classification was given to capture-targeted (“On-target”), then “Off-target”, and finally the “Intergenic” class; these three categories were mutually exclusive.

*Comparison of capture protocols and long cDNA capture efficiency.* We wished to compare the performance of the CLS protocol to that of other methods. We judged performance on the basis of (1) the percentage of reads in post-capture cDNA that originated from a targeted region (on-target rate), and (2) the enrichment, defined as the ratio of on-target rates in post/pre-capture cDNA. In all experiments, the off-the-shelf SeqCap RNA lncRNA enrichment kit (Roche) was used. Four distinct experiments were performed. For each one, the same aliquot of human kidney total RNA was used, and sequencing was done with Illumina MiSeq. The experiments were as follows:

1. Original CLS protocol (as used and described here), polyA-selected, unfragmented
2. Improved CLS protocol, polyA-selected, unfragmented
3. Improved CLS protocol, total RNA, unfragmented
4. Roche SeqCap RNA protocol, total RNA, fragmented

‘Improved’ CLS incorporated several adjustments designed to boost enrichment: the use of LoBind tubes, a drying step at 60 °C, a shorter incubation time, the use of Smarter blockers, and the use of a water bath at 47 °C for post-capture washes.

Findings are presented in **Supplementary Figure 2h,i** and together suggest that capture of long cDNAs yields lower on-target efficiency. Additional methods are included in **Supplementary Note 1**. Summary statistics on UMD-ROIs and double-bounded reads are presented in **Supplementary Table 9**. A comparison/integration of polyadenylation and splice junction strand inference approaches is presented in **Supplementary Table 10**. **Supplementary Table 11** shows the CAGE support of novel versus known PacBio TSSs. Details about TSS versus ChIP-seq and TSS conservation analyses are included in **Supplementary Tables 12 and 13**.

**Code availability.** All computer code used in this study is available from the corresponding authors upon request. Most programs have been deposited in GitHub as specified in “URLs.”

**Data availability.** Raw and processed data have been deposited in the Gene Expression Omnibus under accession [GSE93848](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE93848). RT-PCR validation sequences are available in **Supplementary Data Set 4**. Genome-aligned data were assembled into a public Track Hub, which can be loaded into the UCSC Genome Browser (see “URLs”). A **Life Sciences Reporting Summary** for this paper is available.

54. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
55. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).
56. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
57. Quinlan, A.R. BEDTools: the Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1–11.12.34 (2014).

## Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### ▶ Experimental design

#### 1. Sample size

Describe how sample size was determined.

One sample per tissue type

#### 2. Data exclusions

Describe any data exclusions.

None

#### 3. Replication

Describe whether the experimental findings were reliably reproduced.

N/A

#### 4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

N/A

#### 5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

N/A

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

#### 6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Confirmed  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The <u>exact sample size</u> ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)                                    |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement indicating how many times each experiment was replicated  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as an adjustment for multiple comparisons   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The test results (e.g. $P$ values) given as exact values whenever possible and with confidence intervals noted   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A clear description of statistics including <u>central tendency</u> (e.g. median, mean) and <u>variation</u> (e.g. standard deviation, interquartile range)  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clearly defined error bars  |

See the web collection on [statistics for biologists](#) for further resources and guidance.

## ► Software

Policy information about [availability of computer code](#)

### 7. Software

Describe the software used to analyze the data in this study.

R, perl, bedtools, STAR, StringTie, Cufflinks, liftOver, custom software available on GitHub, as specified in the "Code availability" section of the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* [guidance for providing algorithms and software for publication](#) provides further information on this topic.

## ► Materials and reagents

Policy information about [availability of materials](#)

### 8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No restrictions

### 9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A

### 10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

See Online Methods

b. Describe the method of cell line authentication used.

See Online Methods

c. Report whether the cell lines were tested for mycoplasma contamination.

See Online Methods

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

## ► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

### 11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

Mouse RNA samples were obtained from commercial sources.

Policy information about [studies involving human research participants](#)

### 12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Human RNA samples were obtained from commercial sources.