UNIVERSITAT DE BARCELONA

FACULTAT DE BIOLOGIA

DEPARTAMENT DE GENÈTICA

Programa de Doctorat en Genètica

Genomic Characterization of Human Long Noncoding RNAs

Memòria presentada per

Julien Lagarde

per a optar al grau de Doctor per la

Universitat de Barcelona

Treball realitzat al Centre de Regulació Genòmica (CRG)

Doctorand

Julien Lagarde

Director

Roderic Guigó

Centre de Regulació Genòmica

Tutor

Josep F. Abril

Universitat de Barcelona

Barcelona, Setembre de 2019

Acknowledgments

I was fortunate enough to enjoy my years of PhD. These have been filled with thoroughly rewarding intellectual experiences, and I am thankful to many people for making it possible. First and foremost, I would like to express my profound gratitude to my Director Roderic Guigó, who encouraged me early on to pursue a Doctorate, provided an ideal environment to carry it out, secured funding and guided this project. Most of the research presented here was co-supervised by my dear friend Rory Johnson, to whom I am also deeply grateful. Roderic and Rory's insight, intellectual sharpness, passion and tireless dedication to Science, humane qualities and good humor are and will always be an inspiration for me.

Besides Roderic and Rory, I also thank Barbara Uszczynska-Ratajczak and Sílvia Carbonell, for their crucial contributions to the success of this study. I feel proud of what we achieved together and I will keep very fond memories of this happy and exciting period of my life.

This work also benefitted from the scientific contributions of Jennifer Harrow, Adam Frankish, Jonathan Mudge, José Manuel González (European Bioinformatics Institute, UK), Javier Santoyo-López (Roslin Institute, Edinburgh, UK), Tom Gingeras and Sara Goodwin (Cold Spring Harbor Laboratory, USA), Irwin Jungreis (Massachusetts Institute of Technology, USA), Jyoti Choudhary and James Wright (Institute for Cancer Research, UK), and Lluis Armengol (qGenomics, Spain).

At the Centre de Regulació Genòmica, the following (likely non-exhaustive) list of outstanding people also helped, in random order: Sarah Djebali, Sílvia Pérez-Lluch, Amaya Abad, Emilio Palumbo, Dmitri Pervouchine, Ferran Reverter, Alessandra Breschi, Carme Arnan, Diego Garrido, Anna Vlasova, Beatrice Borsari, Joao Curado, Ramil Nurtdinov, Francisco Camara, Marina Ruiz-Romero, Valentin Wucher and, last but not least, our all-around phenomenal administrative assistant Romina Garrido.

Josep Abril, Rosa Maria Baruquer, Pere Martinez, Marta Pascual and Montse Corominas helped me through the somewhat arcane administrative procedures of the Universitat de Barcelona. To Pep, in particular, thanks for the tutoring and the geeky wisdom you are always so eager to communicate.

My thanks to the following people for useful feedback on various parts of the Thesis manuscript: Sarah Bonnin, Sílvia Carbonell, Sílvia Pérez-Lluch, Rory Johnson, Josep Abril and Roderic Guigó.

Among the Guigó lab and CRG alumni, my friends France Denoeud, Thomas Derrien, Vincent Lacroix, Pedro Ferreira, Sylvain Foissac, Christoforos Nikolaou and Filipe Pinto Teixeira Sousa contributed to making me a better scientist and human being.

Thanks to my love, soulmate, best friend Sarah Bonnin for her constant support, in R and otherwise, but mainly otherwise. You make my world a better place.

Pour finir, merci à ma famille et surtout à mes parents, pour leur amour, l'éducation qu'ils m'ont prodiguée et leur soutien sans faille. Ce studieux mois d'août 2019 à Langogne restera parmi mes plus beaux souvenirs avec vous. Merci pour ces moments, et pour avoir su y entretenir une atmosphère si propice au travail intellectuel.

Abstract

The human genome contains an astonishingly large fraction of noncoding DNA, which is pervasively transcribed into thousands of long noncoding RNAs (lncRNAs) – long transcripts with no discernible protein-coding potential. However, little is known about lncRNAs' biological functions, and their genome annotations show evident signs of inadequacy: existing gene models are sketchy, and many lncRNAs remain uncatalogued. This annotation incompleteness hampers lncRNA functional characterization, notably by failing to accurately describe gene boundaries. To address this issue, the present work aims to advance towards a complete and accurate annotation of lncRNA genes in the human genome. Using a high-throughput, targeted long-read transcriptome sequencing methodology, this study uncovers thousands of novel lncRNAs, approximately doubling the annotated transcript complexity within targeted loci. The method presented vastly outperforms competing techniques in accuracy, and precisely maps many previously unknown, strongly supported lncRNA transcript boundaries. This augmented catalog provides the most definitive view of the genomic properties of lncRNAs to date, while contributing a robust foundation for future lncRNA functional characterization.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ARM - Anchored Read Merging
bp - basepair
CAGE - Cap Analysis of Gene Expression
CCS - Circular Consensus Sequence
cDNA - complementary DNA
CDS - Coding Sequence
CLS - Capture Long-read Sequencing
CRISPR - Clustered Regularly Interspaced Short Palindromic Repeats
CRISPRa - CRISPR activation
CRISPRi - CRISPR interference
CTCF - CCCTC-binding Factor
DHS - DNase I Hypersensitive Site
DNA - Deoxyribonucleic Acid
dRNA - direct RNA
EC - Evolutionary Conservation
EGASP - ENCODE Genome Annotation Assessment Project
ENCODE - ENCyclopedia Of DNA Elements
eRNA - enhancer RNA
EST - Expressed Sequence Tag
GRM - Greedy Read Merging
GWAS - Genome-Wide Association Study
H3K27me3 - trimethylation of histone H3 at lysine 27
H3K36me3 - trimethylation of histone H3 at lysine 36
H3K4me1 - monomethylation of histone H3 at lysine 4
H3K4me3 - trimethylation of histone H3 at lysine 4
H3K9ac - acetylation of histone H3 at lysine 9
H3K9me3 - trimethylation of histone H3 at lysine 9
lincRNA - long intervening (sometimes intergenic) noncoding RNA
lncRNA - long noncoding RNA
mRNA - messenger RNA (protein-coding)
ncRNA - noncoding RNA
NET-CAGE - Native Elongating Transcript-Cap Analysis of Gene Expression

NGS - Next-Generation Sequencing
nt - nucleotide
ONT - Oxford Nanopore Technologies
ORF - Open Reading Frame
PCR - Polymerase Chain Reaction
Pol II - RNA Polymerase II
PRC2 - Polycomb Repressive Complex 2
RACE - Rapid Amplification of cDNA Ends
RNA - Ribonucleic Acid
RPKM - Reads Per Kilobase of exon per Million mapped reads
rRNA - ribosomal RNA
RT - Reverse Transcriptase
RT-PCR - Reverse Transcriptase PCR
sgRNA - single-guide RNA
SGS - Second-Generation Sequencing
SLR-Seq - Synthetic Long-read Sequencing
SMRT - Single-Molecule Real-Time
snoRNA - small nucleolar RNA
SNP - Single-Nucleotide Polymorphism
snRNA - small nuclear RNA
TE - Transposable Element
TGS - Third-Generation Sequencing
TM - Transcript Model
tRNA - transfer RNA
TSS - Transcription Start Site
UCE - Ultra-Conserved Element
UMI - Unique Molecular Identifier
UTR - Untranslated Region

Introduction

I

From Mendel's garden to whole-genome sequencing: a brief historical perspective

A genome encodes the information necessary for its host organism's development and physiological operation over its entire lifetime – namely, its phenotype. Since Mendel's breeding experiments established the concept of discrete inheritable units[1] – later coined *genes* by Johannsen[2] – in the second half of the nineteenth century, relating genes and phenotypes has been the *raison d'être* of Genetics.

Mendel and his contemporaries were unaware of the cytological – let alone molecular – basis of heredity. Mendel's laws of heredity went largely unnoticed or misunderstood by the scientific community during almost four decades after the Moravian monk published his seminal work[3]. Soon after their rediscovery, Sutton[4,5] and Boveri[6] independently suggested the chromosomes to be the carriers of genetic information, thereby introducing the *chromosome theory* of heredity. Studying grasshoppers' germ cell division, Sutton directly made the connection between his observations and Mendel's theoretical framework. He aptly concludes his 1902 paper with the following statement:

> *"[...] the association of paternal and maternal chromosomes in pairs and their subsequent separation during the reducing division [...] may constitute the physical basis of the Mendelian law of heredity."*[4]

Originally met with controversy, the chromosome theory was later validated successively by Carothers[7] and Morgan[8]. Naturally, assigning such a crucial genetic role to chromosomes drove scientists towards studying them in more depth. Sticking to a strict etymological definition of *genomics* (the study of genomes), one might argue that it marked the birth of this discipline.

Early efforts to map genetic markers onto the genome came to fruition in the 1910s. Sturtevant and Morgan, by investigating patterns of gene co-segregation over

generations of *Drosophila melanogaster* populations, were able to produce the first genetic map (Figure 1)[8,9]. Their so-called *linkage maps* relied on the assumption that the closer physically genes are on a chromosome, the lower their chance of being separated from each other during meiotic crossover, and hence, the higher their probability of being inherited together. Morgan's team could thus estimate the relative distances that separate genes on *Drosophila* chromosomes. The mere ability to build such consistent maps was highly significant: it meant that genes were *linearly ordered* along chromosomes.

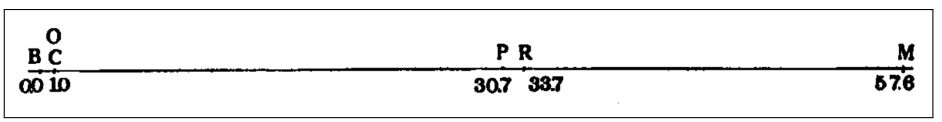


Figure 1: Sturtevant's genetic map of six *Drosophila* sex-linked genes on their chromosome. Top row (traits, modern nomenclature between parentheses when applicable): B, "Black body" ("Yellow body"); C, "Eye color" ("White eyes"); O, "Eosin eyes" ("White eyes"); P, "Vermilion eyes"; R, "Miniature wings"; M, "Rudimentary wings". Bottom row: distance. The unit of distance is taken as a portion of the chromosome of such length that, on average, one cross-over will occur in it out of every 100 gametes formed. That is, percent of cross-overs is used as an index of distance. Reproduced from[9].

Decades more were necessary to progress from Morgan and Sturtevant's painstakingly assembled linkage maps to nucleotide-resolved, annotated genome sequences. DNA's basic chemistry – a polymer of adenine (A), cytosine (C), guanine (G) and thymine (T) – had been described long before[10]; yet, its biological role was still obscure. DNA finally came to prominence when it was proposed by Avery and colleagues as the likely biomolecular basis of heredity[11]. The hypothesis was confirmed in 1952 with the Hershey-Chase experiment[12], and shortly after by Watson and Crick[13]. After detailing their double helix model and its associated A-T G-C base complementarity rules, the British-American pair note:

> *"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."*[13]

As a direct connection between chromosomes, DNA and heredity was finally established, resolving DNA sequences became imperative. The concept of a *genetic program* governing an organism's development, as formulated by Jacob, Monod[14] and Mayr[15] in 1961, added momentum to this enterprise. Watson and Crick's discovery

also paved the way for first-generation DNA sequencing methods, which were developed in the late seventies by Maxam, Gilbert[16] and Sanger[17,18]. Those were quickly applied to decoding the genomes of simple organisms, and subsequent gradual improvements and democratization of sequencing and molecular biology techniques enabled the tackling of increasingly complex genomes (see summary timeline in Table 1; for a more complete history of DNA sequencing methods, see Heather and Chain[19]).

Organism	Domain	Haploid genome size	Date finished	Remarks	References
MS2	Virus (RNA)	3.5 kb	1976	First genome	[20]
ϕX174	Virus (DNA)	5.4 kb	1977	First DNA genome	[21]
H. influenzae	Prokaryote	1,8 Mb	1995	First genome of a free-living organism	[22]
S. cerevisiae	Eukaryote	12 Mb	1996	First genome of a eukaryote	[23]
C. elegans	Eukaryote	97 Mb	1998	First genome of a multicellular organism	[24]
D. melanogaster	Eukaryote	120 Mb	2000		[25]
H. sapiens	Eukaryote	3 Gb	2001		[26,27]
M. musculus	Eukaryote	2,5 Gb	2002		[28]

Table 1: Timeline of some landmark whole-genome sequencing projects.

Yet, a bare sequence of As, Ts, Gs and Cs is next to useless in the context of biological research; in other words, a genome's sequence (namely, its *data*) is only as valuable as its annotation (*i.e.*, its associated *metadata*). Moving to an interpretable genome sequence requires to locate genes and other landmarks on the genome, assigning them biological functions and describing their regulation, to name a few. These are some of the most important challenges at the heart of genome annotation.

II

Gene annotation in the genomic era

If Sturtevant and Morgan can be credited with producing the first genome maps, a fundamental distinction needs to be made between those and present-day, high-resolution genome annotations. The two proceed from opposite approaches: because the early *Drosophila* geneticists were unaware of the biochemical nature of the genetic material, they mapped *phenotypic traits* to the genome. In contrast, the *modus operandi* of modern genomics consists in first *defining a genotype*, then investigate how phenotypic traits emerge from it – a process called *reverse genetics*. To put it another way, it is concerned with relating each nucleotide in a genome to its biological function. This procedure is slow: today, in the so-called genomic era, there remain thousands of identified genomic features whose function is unknown, even in extremely well-studied organisms (see *e.g.* Wood *et al.*[29]).

Recent large-scale genomic studies have also prompted the expansion and redefinition of the term "gene": it is no longer merely an abstract unit of heredity, as defined by classical genetics. Rather, a gene is nowadays generally defined as a genomic region encoding a set of overlapping functional products, usually RNA transcripts – a more practical, but still imperfect working definition for modern genomics[30].

The process of genome annotation can be roughly subdivided into three consecutive layers[31], as summarized in Figure 2:

1. ***Nucleotide-level annotation:*** This first step consists in precisely locating genomic landmarks – genes, transcripts, repeats, promoters, enhancers, etc. Gene finding is one of the most critical components of nucleotide-level annotation. Gene finding in prokaryotic genomes is relatively straightforward, due to their extreme compactness: in the bacterium *Escherichia coli*, for example, the average distance between adjacent genes is only 118 base pairs[32]. To summarize, once a bacterial genome is assembled, finding protein-coding regions relies, for the most part, on finding open reading frames (ORFs) longer than a certain threshold, *in silico*. Because prokaryotic noncoding RNA (ncRNA) genes are

usually very well conserved evolutionarily, they are also readily identifiable. In eukaryotes, the low gene density, combined with the presence of introns and alternative splicing, significantly hampers the task of finding both coding and ncRNA genes (see Section II.1).

2. ***Product-level annotation:*** When a genomic feature is expressed into a product in the cell, such as an RNA and/or a polypeptide, the identification of this product is carried out at this point. This is usually achieved through sequence homology – searching the sequence in question against a database of known protein or RNA molecules, under the assumption that sequence conservation implies function conservation.

3. ***Process-level annotation:*** At this final stage, gene products are assigned a function; that is, they are integrated into the broader framework of the cell and organism's physiology. Automated *in silico* methods aimed at predicting gene function exist, mainly for coding sequences[33,34]. However, in the absence of clues gathered by these computational tools, this step involves more laborious molecular biology techniques. Broadly speaking, these generally consist in disrupting the expression or sequence content of the gene of interest, followed by the study of the phenotypic consequences of the perturbation. Such methods include mutagenesis, gene knock-out (*i.e.*, inactivation of a gene at the DNA level), knock-in (*i.e.*, insertion or replacement of a gene in the genome) and knock-down (*i.e.*, gene silencing at the RNA level).

Naturally, this basic annotation workflow needs to be adapted to non-standard cases. A very low level of sequence conservation, for example, might make product-level annotation of a given gene unworkable. This step may be consequently skipped altogether, and one may need to proceed directly to process-level – that is, functional – annotation. Similarly, functional annotation can in principle be directly inferred from product-level annotation if the product in question shows homology to an already well-characterized biomolecule – *i.e.,* an RNA or a protein – from another organism. In fact, it is estimated that 98% of the entries in Gene Ontology – the most widely used gene function annotation database – are inferred using computational means only[35]. On the other hand, nucleotide-level annotation and gene finding, the focus of the present work, arguably constitute the indispensable groundwork of any functional genomic study.

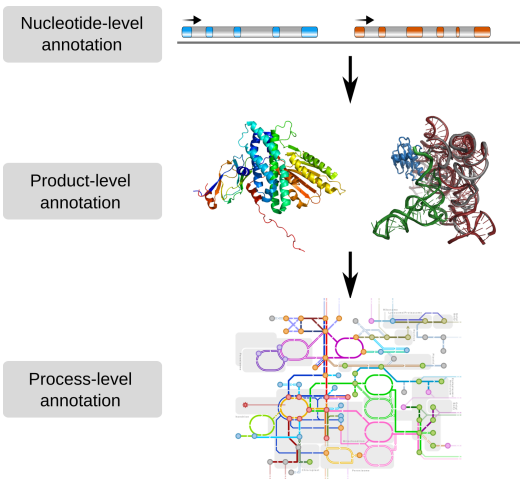


Figure 2: The three layers of genome annotation. See text for details (nomenclature and figure inspired by[31])

II.1. Gene finding in eukaryotic genomes

Gene finding is at the core of nucleotide-level genome annotation – henceforth simply referred to as *annotation*. Given the size and complexity of eukaryotic genomes, gene annotation techniques always incorporate an important computational component, and can be roughly subdivided into *de novo* (*i.e.*, using solely genome sequences as input) and *evidence-based* (*i.e.*, using both genomic and external, non-genomic evidence) methods.

II.1.1. *De novo* gene prediction

As is the case with bacterial genomes, a draft annotation of eukaryotic genomes can be obtained with computational methods using intrinsic genomic evidence only. A subtype of these *de novo* methods, labeled *ab initio*, consists in only using the information contained in the genome sequence of interest. From a practical standpoint, this is by far the most straightforward approach for the annotator. However, *ab initio* gene finding is considerably more challenging – and as a result, less accurate – in eukaryotes than in bacteria or viruses, notably because of the typically more complex structure of eukaryotic genes (Figure 3). In vertebrates, for instance, genes are usually composed of short exons, interspersed with much longer introns. Human genes

are an apt illustration of this tendency, where exons are on average 249 basepairs (bp)-long, while the intron average length is 6,450 bp. Moreover, recent estimates point to less than 5% of the ∼3 billion bp that compose the human genome being exonic. When only considering the protein-coding portion of exons, this percentage falls even lower (1.1%) [1].

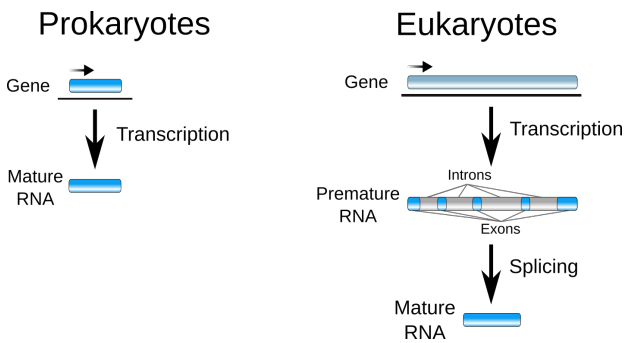


Figure 3: Gene structure and RNA maturation in prokaryotes and eukaryotes. In prokaryotes, the RNA is considered mature straight after transcription. In eukaryotes, both exons (blue boxes) and introns (gray boxes) are first transcribed, then introns are excised from the premature RNA during the splicing process.

Since a gene's biological product is encoded exclusively in its exons, gene annotation implies not only mapping genes, but also precisely delineating their exon/intron structure. It should then appear clearly that, from a computational gene finding perspective, the signal (exons) to search for is deeply buried in noise (intronic and intergenic sequences). The situation is further complicated by alternative splicing, a pervasive phenomenon in vertebrates by which a gene is processed into multiple, distinct transcripts[37,38] (Figure 4).

II.1.1.1. Coding gene prediction

Even protein-coding gene finding, a trivial undertaking in prokaryotes, is strongly impeded by the presence of introns and the resulting lack of ORF contiguity in eukaryotic genomes. Therefore, to obtain best results, *ab initio* eukaryotic gene predictors must rely on sophisticated mathematical models – such as hidden Markov models, conditional random fields and machine learning techniques – to detect splic-

[1] Statistics calculated from the GENCODE Human reference annotation (https://www.gencodegenes.org)[36], version 21 (release: 2014).

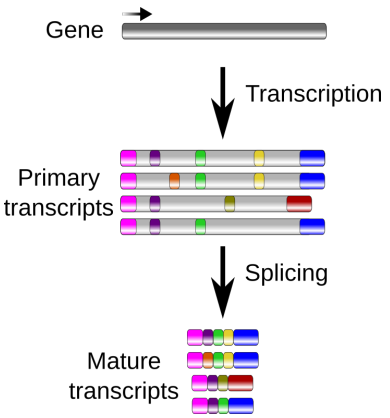


Figure 4: Alternative splicing and transcript diversity. In eukaryotes, one gene can give rise to multiple, distinct mature RNA products through alternative splicing. The gene (dark gray box) is transcribed into a primary RNA transcript composed of putative introns (light gray boxes) and exons (colored boxes). Introns can be differentially spliced out to produce a variety of alternative, mature transcript isoforms.

ing signals and sequence signatures characteristic of coding regions (see Brent[39] for a review). In practice, *ab initio* programs – examples of which include GeneID[40] or GENSCAN[41] – tend to identify individual coding exons reasonably well. Nevertheless, they usually fall short of accurately chaining these into full-length ORFs, as demonstrated by the EGASP (ENCODE Genome Annotation Assessment Project) experiment[42].

There are, in addition, at least two other important facets of gene finding where *ab initio* algorithms perform particularly poorly: alternative splicing and untranslated region (UTR) prediction. Because alternative splicing is regulated not only by *cis* sequence signals but also by *trans*-acting factors highly contingent on the cellular context, it is still near-impossible to predict from genomic sequence alone, despite recent efforts to decipher a so-called "splicing code"[43,44]. Therefore, prediction programs have no other option than to simply ignore alternative isoforms, and output only what they consider to be the most likely chain of coding exons. On the other hand, UTRs – transcribed regions upstream and downstream of the ORF in coding genes, see Figure 5 – which are major players in post-transcriptional gene regulation[45,46], are less constrained and notoriously difficult to model at the sequence level, compared to coding regions. As a result, UTR computational predictions are wildly inaccurate, to

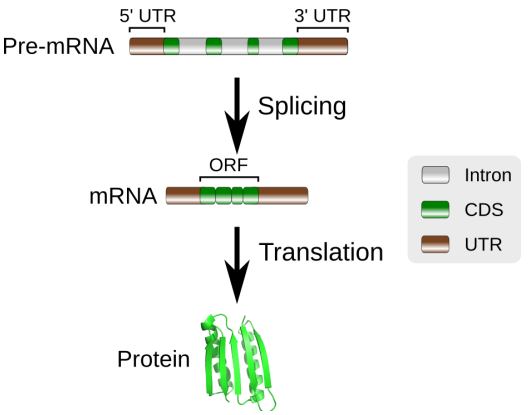


Figure 5: Protein-coding gene structure in eukaryotes. The pre-mRNA contains exons (colored boxes) as well as introns. The ORF, which specifies the protein's aminoacid sequence through blocks of coding sequences (CDS), is flanked by untranslated regions (UTRs) on both sides.

the extent that *ab initio* software generally do not even attempt to report them[39,42,47]. UTRs can be of considerable length (Table 2), and the inability to correctly predict them prevents the identification of potentially important post-transcriptional regulatory elements frequently contained in their sequences, including binding motifs, microRNA response elements and RNA secondary structures[45,48].

Organism	**5′ UTRs**				**3′ UTRs**			
	N	Median	Mean	Maximum	*N*	Median	Mean	Maximum
H. sapiens	86,836	92	145	5,545	59,519	416	956	32,873
M. musculus	54,850	92	139	8,962	40,430	491	950	39,400

Table 2: Transcript UTR length in *H. sapiens* and *M. musculus*. The number (*N*), median, mean and maximum length (in nucleotides) of 5′ and 3′ UTRs are indicated for human and mouse, as observed in the GENCODE annotation. See Supplementary Methods for further details.

II.1.1.2. Noncoding RNA gene prediction

It emerges from the above that *ab initio* approaches generally fail at providing an accurate picture of protein-coding regions in a genome. To make matters worse, their

performance appears even poorer with respect to noncoding RNA genes, with the exception of a few well-characterized RNA species like ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small nucleolar RNAs (snoRNAs) and microRNAs. From a computational standpoint, most ncRNAs share a lot of similarities with the UTRs of coding transcripts, leading to comparable difficulties in predicting them (see Section II.1.1.1). Besides the absence of protein-coding potential, these include lower evolutionary conservation and lack of good mathematical models to detect them, for example. Furthermore, the absence of an open reading frame makes it difficult to assess the completeness of a noncoding transcript model. In other words, the "grammar" of RNA is much less well understood than that of protein-coding sequences, and this substantially impacts ncRNA prediction.

Genome-wide *ab initio* ncRNA prediction approaches generally attempt to detect thermodynamically stable structures[49]. This is because although RNAs can act as linear sequences, they also often fold into secondary and tertiary structure to perform their function. These methods, however, dramatically lack in accuracy, as random RNA sequences tend to fold into stable structures just as well as functional RNAs[50]. In addition, some ncRNAs, particularly long ones[51], tend to be largely unstructured[52].

II.1.1.3. Using genomic homology to improve gene prediction: evolution to the rescue

One way to enhance the performance of *de novo* prediction tools in both coding and non-coding regions is to exploit comparative genomics approaches. Usually, this involves the alignment of the target genome to one or more informant genomes from related species. Gene finding software such as TWINSCAN[53], SGP2[54] and CONTRAST[55] leverage such homology information to better model coding regions, leading to limited improvements in sensitivity and precision[42].

NcRNA predictors incorporating evolutionary models also exist. Those operate by searching genomic sequences for evolutionarily conserved secondary structures, by comparative analysis of covariation in homologous sequence alignments. A plethora of such *de novo* structured RNA predictors have been developed, including EvoFold[56], RNAz[57] and CMFinder[58]. These algorithms, however, suffer from a worryingly high false positive rate[59,60]. In addition, those tend to predict small RNAs and short-range structured regions rather than long, full-length transcripts[60], which limits their usefulness in practice.

II.1.2. Evidence-based gene annotation

Gene annotation consists in finding precisely which parts of the genome are expressed as mature RNA. It follows naturally that sequencing RNA products and mapping them back to their original genome location should constitute the gold standard of gene annotation. These methods are in principle capable of resolving alternative isoforms as well as non-coding RNA features, two important aspects where purely computational pipelines noticeably fail (see Section II.1.1). In fact, such empirical transcriptome-based approaches have been shown to almost always greatly outperform *de novo* predictors[42].

II.1.2.1. Methods for transcriptome sequencing

Overview. Because RNA is an inherently unstable molecule, and since most sequencing chemistries are only suited for DNA, transcriptome sequencing studies generally start with the conversion of RNA to double-stranded, complementary DNA (cDNA) (Figure 6). cDNA has the additional advantage of being amplifiable *in vitro* by Polymerase Chain Reaction (PCR), and/or *in vivo* by cloning it into bacterial vectors. On the other hand, one should note that cDNA synthesis can introduce important experimental biases, especially during first-strand synthesis. Two of those can severely affect downstream gene annotations: Reverse Transcriptase (RT) template switching, and spurious internal polyA priming. The first occurs when the elongating RT-cDNA complex dissociates from the RNA template, and subsequently reanneals to a sequence similar to the original one, but at a different template location, in *cis* or *trans*. In both cases, RT template switching produces experimental artifacts – false introns for the former, false chimeric transcripts for the latter – that can lead to erroneous annotations[61,62]. The second RT-related pitfall stems from the hybridization of the oligo-deoxythymidine (oligo-dT) primer to an internal stretch of As on the RNA template, instead of the intended polyA tail. This results in the first-strand synthesis starting from within the RNA template instead of its 3′ end. Consequently, a cDNA molecule truncated upstream of its true 3′ end is produced[63]. These two experimental biases can be flagged by annotation pipelines, as their sequence substrates – direct short repeats around intron boundaries for the former, genome-encoded polyA runs for the latter – can be detected reasonably easily in downstream bioinformatic analyses.

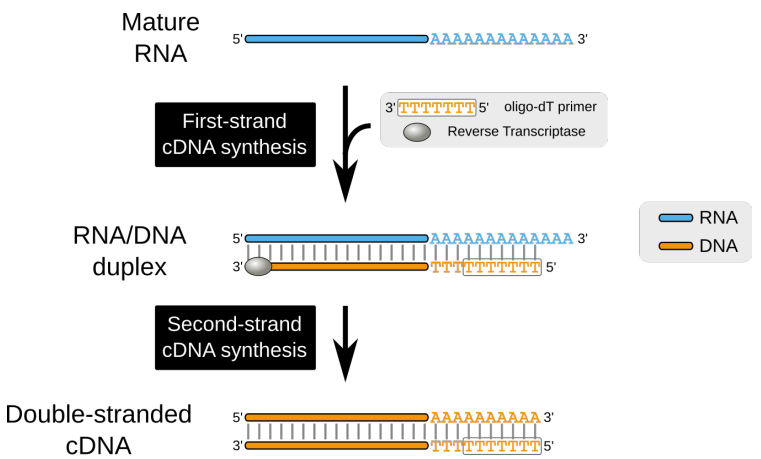


Figure 6: cDNA synthesis in transcriptomics studies (simplified). The method takes advantage of the polyadenylated (polyA) tail present at the 3′ end of most eukaryotic mature RNAs. First-strand cDNA synthesis is initiated at the 3′ end of the transcript by the reverse transcriptase, thanks to a complementary oligo-dT primer (in orange fonts, framed) which anneals at random on the RNA's polyA tail. The final product is a double-stranded cDNA whose sequence is identical (or complementary, depending on its strand) to the original RNA sequence. Note that non-oligo-dT RT priming methods exist, such as those using gene-specific oligonucleotides (see Section II.1.3.2) or random hexamers, but those do not produce full-length cDNAs.

Sanger-based approaches. Before the advent of high-throughput sequencing techniques, transcript sequencing required laborious and costly laboratory procedures involving cDNA vector cloning and Sanger sequencing of random bacterial clones[64]. Since the Sanger method requires a known sequence to prime onto, the sequencing reaction usually starts on either end of the vector. For this reason, reads tend to cover preferentially the 5′ and 3′ ends of the cDNA insert. Moreover, because of the length limitations of the Sanger method, which rarely yields more than 1,000 bp per read, sequencing cDNA clones results in relatively short (*ca.* 500-800 bp), single-pass reads – although with a very high (>99.999%) per-base accuracy[65]. Owing to the far greater length of most mammalian transcripts (Figure 7), these so-called Expressed Sequence Tags (ESTs)[66,67] cover cloned cDNA sequences only partially. Obtaining full-length cDNA sequences hence required low-throughput approaches, such as iterative "primer walking" along the cDNA. Despite these difficulties, initiatives like the Mammalian Gene Collection[68], FANTOM[69] and others[70] successfully

relied on such techniques to build relatively large mammalian gene catalogs. As of today, such early cDNA and EST databases still constitute the bulk of the evidence underlying RefSeq and GENCODE[36,71], the two most widely used mammalian gene annotation resources.

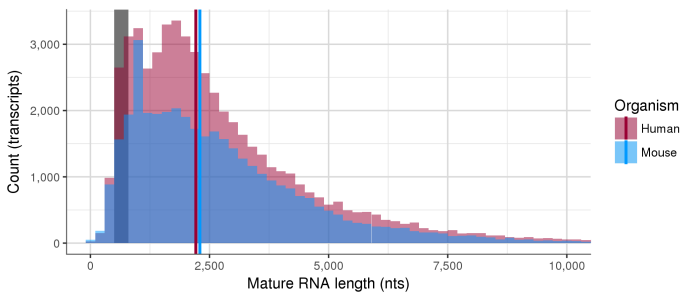


Figure 7: Transcript length distribution in mammalians vs typical Sanger sequencing read length. Histogram of transcript lengths in Human (magenta) and Mouse (blue) annotated genes. X-axis: mature RNA length (that is, excluding introns) in nucleotides (nts). Respective median lengths (Human: 2,213 nts, Mouse: 2,298 nts) are depicted as vertical lines for both species, using the same color scheme. The approximate range of read length output by Sanger sequencing (500-800 nts) is indicated by a gray rectangle. The X-axis is cut at 10,000 nts for clarity. See Supplementary Methods for further details.

Second-Generation Sequencing (SGS) methods. Sometimes still unfortunately called *Next*-Generation Sequencing (NGS) methods more than a decade after their breakthrough, these techniques enabled a major step towards routine large-scale transcriptome sequencing, starting in the mid-2000s. In addition to their much higher yield and lower cost per sequenced base, those offer other important technical advantages over Sanger methods, including the elimination of the cumbersome cDNA cloning step. As a result, SGS RNA sequencing (RNA-Seq) quickly supplanted Sanger-based cDNA sequencing as the method of choice for transcriptome studies (Figure 8).

The first commercialized SGS technology was the 454 pyrosequencing platform[72], which yielded a maximum of one million reads per run, and reads of up to 1,000bp-long. Because of its high cost and relatively low throughtput compared to other SGS platforms, however, the 454 instrument was not commercially successful and therefore discontinued by Roche in 2016. Nowadays, SGS is synonym with

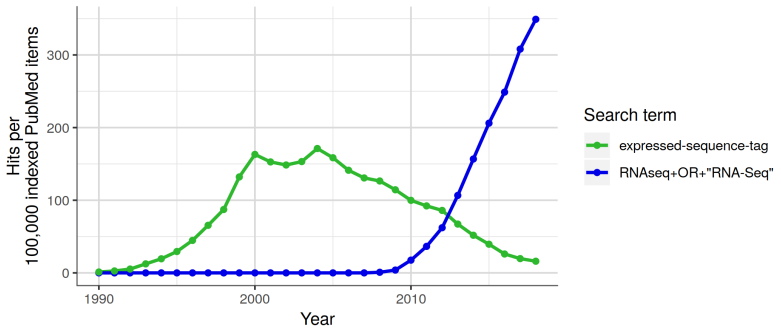


Figure 8: Publication trends of transcriptome sequencing methods over time. The number of PubMed hits per 100,000 indexed papers per year is represented for two search terms associated with transcriptome sequencing techniques (ESTs: green, RNA-Seq: blue) between 1990 and 2018.

Solexa/Illumina sequencing[73], its most popular implementation.

Despite its name, SGS RNA-Seq does not operate directly on RNA, but rather on cDNA molecules. SGS platforms are able to output hundreds of millions of high-quality (∼0.1% error rate for the HiSeq platform as of 2016[74]) sequencing reads per run, but those are typically in the range of 30 to 150 bases in length, *i.e.*, shorter than Sanger sequencing reads. Consequently, short-read RNA-Seq does not provide full-length cDNA sequences, and transcripts need to be physically fragmented prior to sequencing in order to achieve uniform coverage[75].

Re-assembling full-length transcript sequences from SGS short reads *in silico* is possible, in principle. Examples of such algorithms abound, Scripture[76], Cufflinks[77] and StringTie[78] being some of the most widely used. These programs, however, face the problem of assigning each short read to its originating transcript, which becomes near-insoluble in case of extensive transcript overlap and alternative splicing. For this reason, computational transcript reconstruction techniques end up lacking greatly in accuracy: when compared to a ground-truth transcript annotation, their sensitivity and precision generally do not exceed 25% and 50%, respectively, with particularly poor results around the 5′ and 3′ transcript boundaries[78,79].

Third Generation Sequencing (TGS), long-read technologies. These latest methods enable significant improvements compared to SGS in terms of read length, at the cost of lower throughput and higher error rate. For the first time, they bring

the possibility to describe long RNA transcripts from 5′ to 3′ end in a single step and without fragmentation, circumventing the need for computational read assembly. Such platforms have been made available principally by two manufacturers: Pacific Biosciences (PacBio) and, more recently, Oxford Nanopore Technologies (ONT). Both methods are capable of generating multi-kilobase-long reads and can currently yield up to a few million reads per run, while relying on radically different principles.

PacBio's Single-Molecule Real-Time (SMRT) technology consists in the incorporation of labeled nucleotides by a DNA polymerase along individual DNA molecules, which is recorded by a laser and camera system in microwells. Each nucleotide is read with a rather high ∼13% error rate, on average[74]. However, the ligation of hairpin adapters (called "SMRTbells", Figure 9) to each end of cDNA molecules allows the DNA polymerase to circle around its substrate and thus, read the same sequence repeatedly on both strands. Building a consensus sequence out of multiple such passes on the same cDNA molecule is possible, thanks to the stochastic nature of SMRT sequencing errors. This results in a so-called Circular Consensus Sequence (CCS), which exhibits error rates of less than 1% – still an order of magnitude higher than Illumina HiSeq. PacBio's cDNA sequencing application, IsoSeq, has been successfully applied to transcriptome sequencing since 2013[80].

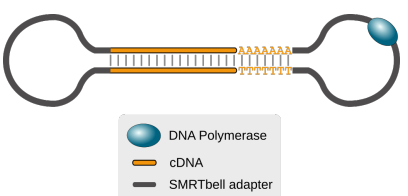


Figure 9: Schematic structure of SMRTbell cDNA libraries used in PacBio sequencing. Ligated SMRTbell adapters circularize the cDNA molecule, which enables the DNA polymerase to perform multiple passes along it, and hence correct sequencing errors.

In contrast, ONT devices rely on multiple electro-sensitive protein pores fixed on a membrane, through which single-stranded cDNA molecules pass. During this process, the pore's electrical current is modulated as a function of the DNA sequence currently traversing the pore. This current is measured in real time by the instrument, and the resulting signal converted to a nucleotide sequence[81]. ONT runs can produce astonishingly long reads – the current record being more than 2 megabases, for genomic DNA[82]. However, its error rate is similar to that of single-pass PacBio reads, but contrary to the SMRT technology, ONT multi-pass sequence consensus building has yet to become mainstream, despite recent developments in this direction[83], which

strongly handicaps nanopore technologies for transcriptome sequencing, despite its very low cost[74,84].

Due to physical constraints, longer DNA molecules tend to load into microwells or nanopores with more difficulty than shorter ones. One drawback of both TGS long-read technologies is therefore their strong bias towards shorter DNA molecules. This phenomenon can be mitigated by separating cDNA libraries into distinct size fractions and sequencing these separately – a laborious and inefficient process, though[85].

Synthetic Long-Read Sequencing (SLR-Seq). At the frontier between Illumina and long-read sequencing lies SLR-Seq, in which cDNAs are diluted and separated into partitions, either through microtiter wells or emulsions. Each cDNA partition is then fragmented and barcoded using Unique Molecular Identifiers (UMIs), followed by short-read Illumina sequencing. Thanks to the presence of UMIs, short reads can be reassembled computationally into full-length isoforms with high confidence. In theory, SLR-Seq thus promises the depth and high sequence accuracy of Illumina sequencing, combined with unambiguous transcript reconstruction, which is lacking in standard short-read RNA-Seq (see 'Second-Generation Sequencing (SGS) methods', page 14)[86–88]. Nevertheless, while SLR-Seq isoform assembly does gain in accuracy thanks to the presence of unique barcodes, it does not fully guarantee against artifactual reconstruction – in cases where several abundant, similar isoforms end up in the same partition, for instance.

Direct RNA (dRNA) sequencing. The ability to directly sequence RNA without prior conversion to cDNA is a recent and promising innovation. During recent years, two dRNA sequencing technologies were made available to the research community: the Helicos/HeliScope single molecule fluorescent sequencing platform[89], and more recently, nanopore sequencing. As of today, the latter – and only surviving – method, commercialized by ONT, is beginning to appear applicable to large-scale transcriptome sequencing projects[90]. While eliminating the biases and experimental artifacts associated with cDNA synthesis mentioned before (see 'Overview', page 12), ONT dRNA sequencing also allows the detection of modified, non-canonical bases in RNA, clearing the way to the elucidation of the "epitranscriptome"[91,92]. Another advantage of dRNA sequencing compared to cDNA-based approaches is that it reads a transcript's *native* polyA tail – rather than the product of random oligo-dT priming on it, see Figure 6 – which brings the ability to measure its actual length. Because of the unamplifiable nature of RNA, dRNA often require prohibitive amounts of input material, however. Thus, important impracticalities remain with current dRNA pro-

tocols, beyond the aforementioned high error rate plaguing nanopore sequencing in general.

II.1.2.2. Building empirical gene and transcript catalogs

Mapping transcript evidence to the genome. Determining where a sequenced RNA originates on the genome can be achieved via sequence similarity search between the transcript and the genome sequences. Except in cases of genomic sequence duplication such as paralogy and pseudogenization, this mapping is usually unambiguous, provided the transcript sequence is long enough. Historical local alignment algorithms such as BLAST (Basic Local Alignment Search Tool)[93,94] are not adapted to transcript-to-genome alignments, because they do not model splice sites properly and consider introns as mere alignment gaps, leading to gross inaccuracies around exon/exon junctions[95]. The growth of EST and cDNA sequence databases in the late 1990s and early 2000s created a demand for specialized spliced aligners, such as EST_GENOME[96], BLAT (BLAST-Like Alignment Tool)[97] and GMAP[98]. Most of these tools would later struggle to cope with the enormous data volumes produced by SGS methods, which prompted the development of another generation of short-read, ultra-fast spliced aligners (*e.g.* GEM[99], STAR[100] and TopHat[77]).

Nevertheless, automatic mapping of transcript evidence to the genome is not sufficient to build accurate gene and transcript models. When building their high-confidence gene catalogs, annotation groups such as RefSeq[71] and GENCODE[36] incorporate a manual curation step, during which experts review all relevant features of an empirical model – alignment quality, splice sites, ORF contiguity *etc.* – before including it in their catalog. While extremely laborious and costly, this strategy has proven useful, as manual annotations tend to display a much higher rate of experimental validation than fully automated ones, and thus are considered far more reliable[42,101]. Due to their notorious unreliability (see 'Second-Generation Sequencing (SGS) methods', page 14), transcript models reconstructed *in silico* from SGS short reads are not included in reference resources such as GENCODE[36] – some are present in RefSeq, though[71].

When compared to first- and second-generation sequencing methods, recent TGS approaches hold the potential for high-throughput, full-length genome annotation at a comparatively low cost (see previous section). Their much higher yield, however, makes the systematic manual curation of TGS-based gene models exceedingly time-consuming. On the other hand, TGS's higher base error rate demands stricter quality filtering – and possibly sequence correction – for resulting gene models to

reach acceptable levels of confidence without human curation.

Gene annotation resources. RefSeq[71] and GENCODE[36] are the two most widely-used gene annotation catalogs. These gene sets consist of both automatically-generated and manually-reviewed gene models, separated into tiers. However, GENCODE contains a much higher proportion of curated gene models compared to RefSeq, at least within coding regions (93.4% and 45%, respectively[102]). While both resources present their own pros and cons, it is generally admitted that GENCODE provides a more complete representation of the human and mouse gene sets[102–104]. In addition, and in contrast to RefSeq, GENCODE incorporates both discovery- and validation-oriented experimental components to its annotation pipeline, using both transcriptomics and proteomics techniques[101,105–107]. As a result, GENCODE is considered the reference gene annotation in human and mouse, and has been adopted by various large international consortia, including ENCODE[108–111], GTEx[112,113] and the International Cancer Genome Consortium (ICGC)[114].

How many genes in a mammalian genome? The number of genes in the human genome has been a matter of debate for more than half a century[115]. During the few years preceding the completion of the human genome, most estimates fell in the range of 50,000 to 100,000 protein-coding genes[116,117]. To the surprise of many, the complete genome sequence revealed a mere 31,000 such genes[27], and the tally has fallen steadily since then. Nowadays, both RefSeq and GENCODE agree on a gene count of ∼20,000 coding genes in the human genome, and a number of long noncoding genes between 15,000 and 18,000[118]. Gene counts in the mouse genome are fairly dissimilar to that of human (Figure 10). However, it is still unclear how much of this difference is explained by biology rather than by the more mature status of the human genome annotation.

There is uncertainty as to how these gene counts will evolve in the future. Since coding signatures can nowadays be identified with relatively high sensitivity on the genome (see Section II.1.1.1), one could reasonably speculate that the protein-coding gene catalog is next to complete, and recent deep sequencing surveys indeed seem to support this notion (see Section III.3.2.2). A recent study, however, challenged this consensus by claiming the discovery of hundreds of novel human coding genes through ultra-deep, short-read-based trancript assembly[119], but later proved extremely controversial[118,120].

Because of the difficulties in identifying them in genomic sequences (see Section II.1.1.2), the situation is less definitive for noncoding genes. Thus, there is reason to

believe that the corresponding tally will substantially increase as methods of detection improve (see Sections II.1.3 and III.3).

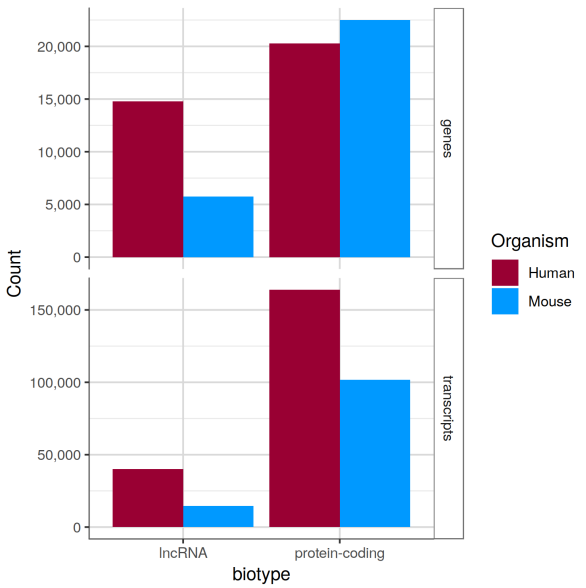


Figure 10: Basic GENCODE annotation statistics in the human and mouse genomes. Shown are the gene (top) and transcript (bottom) counts for the long noncoding RNA (lncRNA, left) and protein-coding (right) biotypes. Other biotypes are omitted for simplicity. Biotypes correspond to broad functional classes of genes. See Supplementary Methods for further details.

II.1.3. Elucidating the hidden transcriptome

If expression-based methods can be considered, as a whole, far better suited for genome annotation than *de novo* gene finding methods, technical factors severely limit their sensitivity: sequencing techniques, which transcriptome studies rely upon, are inherently limited in their read output, and there is strong evidence for the existence of a hidden layer of unannotated transcripts, a deep transcriptome lying under the radar of current experimental techniques.

The statistical representation of a given transcript in a cDNA library directly relates to its RNA copy number in the cell. Moreover, distinct transcripts can be expressed at wildly different levels *within* a cell – from one copy to hundreds of thou-

sands of copies per cell[121]. The problem is further aggravated by the fact that RNA-Seq libraries are usually constructed from entire tissues and bulk cell populations, which has the effect of masking out possibly massive gene expression level differences *across* the cells under study. Rare or transient cell types such as circulating tumour cells, early embryonic cells and adult stem cells, are difficult to isolate from tissue extracts and may contain distinctive RNA populations, for example[122].

Since sequencing technologies always incorporate a random sampling process, this leads to lowly-expressed genes being vastly underrepresented or even absent from standard cDNA sequencing libraries. In 2009, Blencowe *et al.* suggested that ∼700 million short Illumina reads would be needed to correctly detect 95% of the transcripts present in an RNA sample[123]. More recently, the SEQC/MAQC-III Consortium reported the sustained discovery of unannotated introns at depths greater than 2 billion Illumina HiSeq reads, using discovery - saturation curves in a variety of human samples[124]. This indicates that even ultra-deep sequencing – which is far above the capacity of current TGS long-read technologies – is not sufficient to detect all transcripts present in a sample. For this reason, a few laboratory techniques (detailed below) have been developed to mitigate this relative shallowness.

II.1.3.1. Normalization and subtraction of cDNA libraries

Since over 90% of cellular RNAs are ribosomal[121], one of the first steps of cDNA library preparation usually consists in the subtraction of such RNA species, which are largely irrelevant for genome annotation. rRNA depletion is relatively easily achieved by sequence-specific hybridization techniques, and/or polyA selection[125] – as rRNAs are devoid of polyA tails. The latter has the further advantage of also subtracting tRNAs, another overwhelmingly abundant and uninformative RNA species.

However, large differences in concentrations remain even within the polyadenylated transcript population, which leads to the sampling issues mentioned above. Hence, normalization methods, aimed at equalizing cDNA abundances in libraries, have been developed over the years[126–128]. Most of them are based on hybridization kinetics of nucleic acids. Briefly, double-stranded cDNAs are denatured, and then subjected to rehybridization; during this process, abundant molecules tend to reanneal more effectively than rare ones. The mixture is then subjected to digestion by a duplex-specific nuclease, which preferentially targets abundant, double-stranded DNA. This procedure results in sharp decreases in the representation levels of abundant protein-coding genes, and a 3- to 4-fold enrichment for rarer ones[126]. This procedure is still used nowadays, and was recently applied successfully to TGS (PacBio

IsoSeq) library preparation in chicken[129]. However, such untargeted methods are quite inefficient and, by definition, highly unspecific. In addition, concerns have emerged recently as to the possible length biases such procedures may introduce[130].

II.1.3.2. Targeted RT-PCR/RACE

RT-PCR (Reverse Transcriptase PCR) and RACE (Rapid Amplification of cDNA ends)[131] are two standard molecular biology methods based on PCR transcript amplification using specific oligonucleotide primer sequences (Figure 11). RT-PCR relies on two converging transcript-specific primers; thus, an RT-PCR product consists exclusively in the sequence of the targeted transcript that is comprised between the two oligonucleotides. RACE, on the other hand, is based on a single gene-specific primer, paired with a universal – non gene-specific – primer located at the opposite end of the transcript. As a result, RACE can be performed in both 5′ and 3′ directions of the targeted transcript, and is therefore well-suited for precisely delimiting transcript boundaries. Since RACE relies on only one specific oligonucleotide, it is more prone to non-targeted amplification[132,133]. Using a sequential nested primer design can reduce the presence of such unwanted products, however[131].

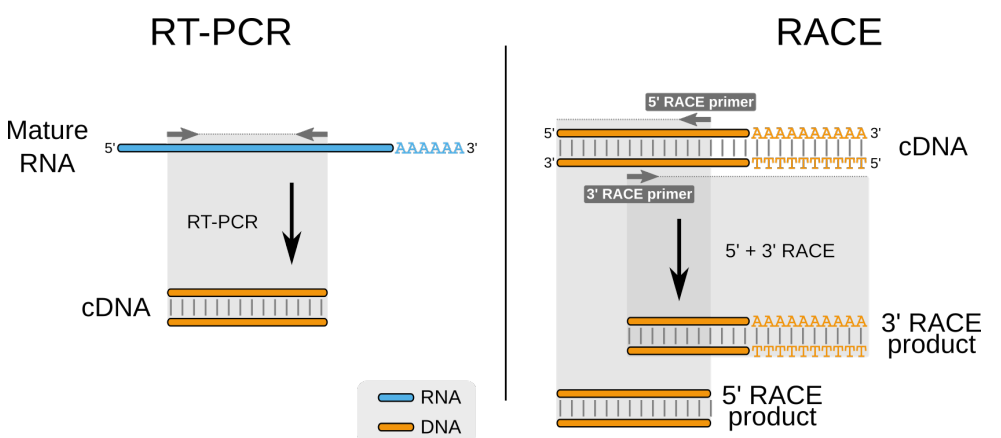


Figure 11: Schematic comparison of the RT-PCR (left) and RACE (right) techniques. The regions amplified by each primer (or pair of primers, depicted as small dark grey arrows) are highlighted in grey.

While both RT-PCR and RACE have been heavily used for some time in low-throughput experiments (typically to quantify transcripts, confirm individual exon/intron structures or transcript ends), recent examples of medium- to large-scale, often discovery-oriented applications have been developed. In 2003,

Guigó *et al.*, guided by *de novo* gene predictions, used RT-PCR coupled with Sanger sequencing to identify more than a thousand novel human genes[134]. More recently, RT-PCR was also applied by the GENCODE consortium to the high-throughput validation of annotated splice junctions using Illumina sequencing (RT-PCR-Seq)[106].

Examples of high-throughput applications of RACE exist, but mainly in the context of hybridization array experiments – which are notoriously noisy and low-resolution – rather than sequencing[105,135]. The idea of coupling RACE with high-throughput sequencing (RACE-Seq) was suggested by Olivarius *et al.* in 2009[136]. The authors applied the procedure to only a few protein-coding genes, however, while combining it with short-read sequencing and the disadvantages it supposes (see 'Second-Generation Sequencing (SGS) methods', page 14).

Despite their high sensitivity and remarkable efficiency, both techniques suffer from one important drawback. By design, they essentially cannot provide complete transcript sequences: RT-PCR products lack both ends of the targeted transcripts, while 5′ and 3′ RACE products are deprived of transcripts' 3′ and 5′ ends, respectively.

II.1.3.3. Targeted cDNA capture

The application of hybridization methods to enrich for specific sequences has been successfully used for some time[137], and is routinely employed in the context of exome sequencing – the targeted sequencing of genome exonic sequences[138,139]. Current capture technologies involve the design of biotinylated oligonucleotide probes tiling sequences of interest. In targeted transcriptome sequencing surveys, probed cDNAs hybridize preferentially to the oligonucleotides, which then bind to streptavidin-linked magnetic beads. The hybridized cDNA products are thereby isolated from undesired cDNA fragments and subsequently eluted (Figure 12). Thus, the resulting cDNA capture library consists in highly enriched targeted molecules that can be subsequently sequenced using high-throughput methods (CaptureSeq)[140].

Nowadays, custom capture designs targeting several megabases of arbitrary sequence can be routinely manufactured. This has enabled large-scale transcript discovery studies focusing on cancer-related genes[141] or low-abundance long non-coding RNAs (lncRNAs)[142,143], for example. However, none of these surveys employed long-read sequencing on capture libraries. Therefore, authors had to rely on short-read-based transcript assembly methods, which inevitably casts doubts on their resulting transcript models (see 'Second-Generation Sequencing (SGS) methods', page 14).

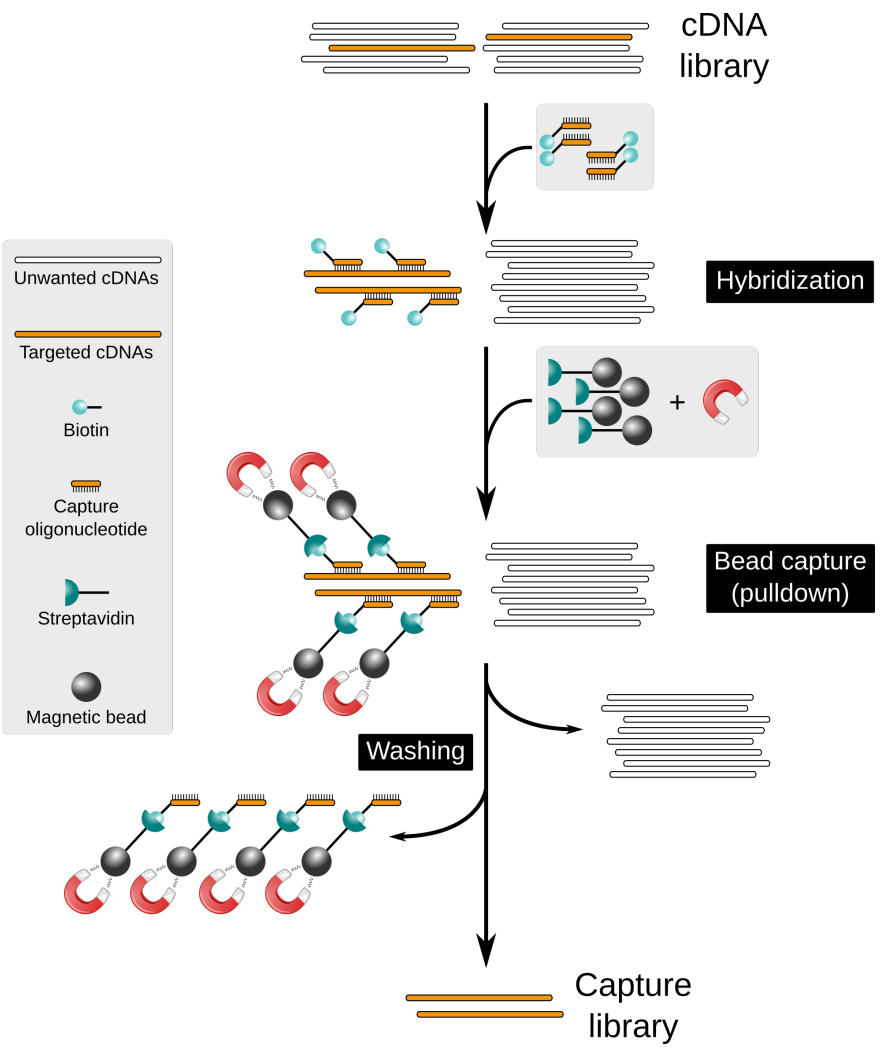


Figure 12: Simplified flowchart of a solution-based cDNA capture protocol (see text for details).

Although much more costly than the cDNA normalization methods mentioned before (see Section II.1.3.1), CaptureSeq achieves considerably higher, more specific target enrichment, while retaining relative RNA abundance information[141,143]. This technique thus paves the way for the efficient interrogation of virtually any potentially transcribed region of the genome, in the context of gene annotation as well as gene quantification.

II.1.4. Gene annotations in their genomic and cellular context

The human and mouse genomes have been the subject of intense scrutiny, particularly since the complete assembly of their respective sequences. International consortia like ENCODE[108–111], NIH Roadmap Epigenomics[144,145], GTEx[112,113] and FANTOM[146–148] have generated thousands of genome-wide assays, whose results are mapped onto these two reference sequences. Those constitute an invaluable resource for integrating gene annotations in their genomic context and facilitate both their validation and biological characterization.

Hallmarks of transcription initiation such as CAGE (Cap Analysis of Gene Expression) clusters[146], DNase I hypersensitive sites (DHS)[149], transcription factor binding sites[150] and promoter-associated chromatin marks – *e.g.* trimethylation of histone H3 at lysine 4 (H3K4me3) –[145,151] can be leveraged to assign levels of confidence to annotated gene TSSs. Similarly, genomic overlaps between transcriptional units and marks of monomethylation of histone H3 at lysine 4 (H3K4me1) are suggestive of enhancer-associated RNAs (see Section III.2.1).

Existing RNA-Seq data can be employed in at least two ways. First, short read mappings split over exon-exon junction can be used as substantiating evidence for dubious intron models[152]. Second, expression profiling of transcript models using RNA-Seq in tissues – healthy or diseased – and subcellular compartments can also provide clues as to their spatial localization and function[153,154].

Finally, more than 70,000 trait- and disease-associated human genetic variants, identified through genome-wide association studies (GWAS)[155], can be mapped onto gene models and shed light on their phenotypic function, role in disease or expression regulation[156–159].

In summary, there is a tight methodological interplay between genome-wide studies and basic gene annotations. While the former quite obviously need the latter as a foundation, it is also true that gene annotations benefit from the superimposed multiple layers of "-omics" evidence.

III

The noncoding genome

One of the distinguishing features of eukaryotic genomes is their large size and low protein-coding content. As previously mentioned, only 1.1% of the human genome bases are currently annotated as coding, and this fraction is not likely to increase substantially upon deeper inspection (see Section II.1 and 'How many genes in a mammalian genome?', page 19). The question then arises as to the composition of the remainder of the genome, and more importantly, what biological function it serves – if any.

III.1. Eukaryotic genome complexity and noncoding DNA: an evolutionary perspective

III.1.1. The *C-value* paradox

Genome size varies considerably across all domains of life (Figure 13, *X*-axis). There is an overall correlation between organism complexity (or at least an intuitive notion we may have of it) and genome size from bacteria to multicellular eukaryotes. This tendency, however, drops within the eukaryotic kingdom, where there is a puzzling discrepancy between genome size (the so-called "*C-value*") and organismal complexity. If DNA contains the entire genetic program needed for an organism to develop, why do complex organisms such as mammals not accordingly possess the largest genomes? Why does *Triticum aestivum* – the common wheat – carry a genome an order of magnitude larger than *Homo sapiens*'? Similarly, it seems utterly counter-intuitive that the genome of *Trichomonas vaginalis*, a unicellular protist, is larger than that of *Drosophila melanogaster*, a complex animal. Putting aside the rather subjective notion of organism complexity, even closely related species can bear genomes of markedly different sizes. Within the *Sorghum* plant genus, for example, haploid

genome sizes vary from 600 Mb to more than 5 Gb[160]. These are some of the apparent inconsistencies that have collectively been grouped under the term "*C-value* paradox" since the 1950s[161,162].

Ohno was one of the first to propose an explanation for this enigma in 1972: he argued that a substantial fraction of eukaryotic genomes bears little to no adaptive advantage, while being tolerated by the host organism. This fraction, which he dubbed "junk DNA", is therefore free to grow over generations, as long as this genomic expansion does not hamper the organism's fitness. As his paper's concluding remark suggests, he believed the bulk of this non-functional genomic DNA consisted in pseudogenes:

> *"Triumphs as well as failures of nature's past experiments appear to be contained in our genome."*[163]

This "junk", however, was later found to be mostly composed of transposable elements (TEs) instead – mobile, "selfish" DNA capable of self-replication[164,165]. It is now established that millions of copies of TEs – or relics thereof – account for an astoundingly large fraction of eukaryotic genomes, including 37.5% of the mouse genome[28], 45% of the human genome[27] and up to 85% of plant genomes[166]. In accordance with the "selfish DNA expansion" hypothesis, a clear correlation can be observed in eukaryotes between genome size and proportion of repetitive elements – of which 75% are identified as TEs[167]. In addition, an estimate of only ∼3.5% of noncoding sequences are highly conserved across mammals[28,168], while less than 1% are conserved across distant vertebrates[169], arousing further suspicion as to their overall functionality.

III.1.2. The *G-value* paradox

The number of protein-coding genes encoded per genome (the "*G-value*") follows a trend similar to that of genome size across the tree of life (Figure 13, *Y*-axis): a global increase with organism complexity that breaks off within the eukaryotic kingdom, particularly in land plants and animals – the *G-value* paradox[173,174].

The *G-value*, as one would expect, positively correlates with genome size. Yet, this trend is also lost when focusing on the larger plant and animal genomes: overall, the dynamic range of *G-values* across kingdoms is also much smaller than that of the *C-value* – approximately ∼2.5 *vs* ∼4.5 orders of magnitude, respectively (Figure 13; note that some hybrid plant species such as *T. aestivum* contain multiple,

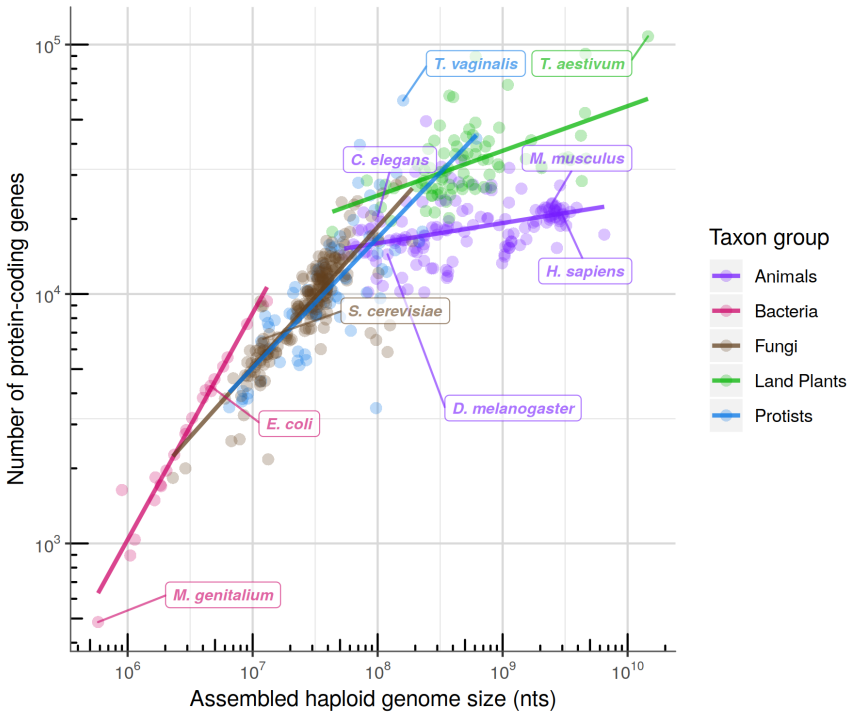


Figure 13: Genome size *vs* coding gene content in free-living bacteria and eukaryotes (log-log plot). Each dot corresponds to an organism whose genome has been assembled and annotated at least as a draft version (note that the list of organisms shown is broadly representative, but not exhaustive). Linear regression lines for each taxon group are depicted using their respective colors. A few genomes of interest are labelled with the corresponding species name. Data compiled from[167,170–172].

partly redundant "sub-genomes", which artificially inflates their genome size and gene number[175]). This observation supports the idea, mentioned in the previous section, of a widespread expansion of noncoding DNA during plant and animal evolution.

A few non-mutually exclusive theories have been put forward to solve the *G-value* paradox (reviewed by Hahn and Wray[174]). Alternative splicing, for instance, has been proposed to contribute to phenotypic innovation by increasing transcript diversity without new gene acquisitions. There is indeed evidence that rates of alternative splicing are correlated with phenotypic complexity[38,176,177]. Furthermore,

analyses suggest that alternative splicing and gene duplication – a major source of functional innovation in evolution[178,179] – might constitute interchangeable evolutionary mechanisms to increase transcript diversity[180–182].

Other authors argue that the key to biological complexity lies in the noncoding part of the genome[183]. In support of this view, the genomes of vertebrates – arguably the most complex organisms on earth – show the highest noncoding content of all life forms[183]. While certainly attractive, this idea must however be reconciled with the fact that the vast majority of noncoding nucleotides in eukaryotic genomes is not evolutionarily constrained[184].

Lastly, a very real possibility exists that the *G-value* enigma is erroneously deemed paradoxical due to overly simplistic premises. While it may certainly not be incorrect to expect a relationship between genome and organism complexity, a gene count constitutes a very crude measure of the information content of a genome, and we may not currently possess the necessary tools to accurately measure the latter.

III.2. Functional noncoding DNA: separating the tare from the wheat

There is strong evidence that a large proportion of the human genome is nonfunctional, junk DNA (see Section III.1.1). Based on evolutionary conservation and mutational load arguments, and in line with Ohno's work[163], some authors propose an upper limit on its functional fraction of 20 to 25%, under the *selected effect* definition of "function" – namely, an evolutionarily selected one[185–187]. Even under this conservative model, this means that a substantial portion of the noncoding genome is likely functional.

III.2.1. Known examples of functional noncoding DNA sequences

Many cases of noncoding, yet functional genome elements are well-known. Some noncoding RNA species, such as snRNAs (small nuclear RNAs), snoRNAs, rRNAs, microRNAs, tRNAs and some lncRNAs (see Section III.3) are extensively characterized. Most of these are annotated automatically with reasonable accuracy in mammalian genomes (see Section II.1.1.2), and occupy a tiny fraction of the human

genome (Table 3).

Category	# genes	# genomic nts covered (kb)	% of genome sequence covered
miRNA	3,837	339	0.011%
rRNA	549	63	0.002%
snoRNA	978	115	0.004%
snRNA	1,912	211	0.007%
tRNA	649	49	0.002%
Total	7,925	776	0.025%

Table 3: Annotated ncRNA genes in the human genome. Statistics (based on the GENCODE annotation) are reported only for the major types of ncRNAs, and omitting lncRNAs, which are covered extensively in Section III.3. See Supplementary Methods for further technical details.

Other functionally important noncoding elements include *cis*-regulatory sequences that control gene expression (*e.g.*, promoters, enhancers, insulators and silencers), origins of replications, telomeres and centromeres. In its most conservative assessment, the ENCODE consortium estimates that a minimum of 8.5% of the human genome is involved in *cis* gene regulation, based on ChIP-Seq and DNase footprinting assays in a limited number of cell lines[110].

Among those regulatory sequences, enhancers constitute a particularly interesting case. Those *cis*-acting sequences activate gene transcription through long-range interactions with the promoter of their target genes. Enhancers can be located equally upstream or downstream of their target promoter, sometimes hundreds of thousands of nucleotides away[188,189]. Intriguingly, some active enhancers have been shown to be transcribed into so-called eRNAs (enhancer RNAs)[190,191] – non-polyadenylated, usually unspliced noncoding RNAs expressed at low levels[148]. eRNAs have long been suspected of being mere by-products of enhancer activation. However, evidence has recently emerged that they may in fact play a crucial role in this process, by actually triggering the molecular cascade leading to enhancer activation[192].

A large fraction of *cis*-regulatory sequences, enhancers included, are not constrained evolutionarily[110]. At first glance, this may suggest that these are the product of noisy, non-specific DNA-protein binding. However, many well-characterized orthologous *cis*-regulatory binding sites show high rates of sequence divergence – to the point of becoming unrecognizable at the sequence level, even among closely related species[193–196]. This, incidentally, indicates that a lack of sequence conservation does not necessarily imply a lack of functionality in noncoding genomic features.

III.2.2. Is all the rest "junk"?

Putting aside known ncRNAs and regulatory sequences, several lines of evidence support the idea that the remainder of the noncoding genome may contain yet-to-be-discovered, possibly transcribed functional elements, even in apparently poorly constrained regions.

Lineage-specific noncoding genetic variants. Genome-wide association studies (GWAS) aim at identifying loci that associate with risk for complex diseases and phenotypic traits. A substantial fraction of the more than 70,000 human variants identified using this method are well-replicated, highlighting the robustness of the inferred associations[155]. A strong link between a locus variant and a phenotypic trait or disease obviously indicates a function for the locus in question, although GWAS does not provide any mechanistic information as to this function. Surprisingly, the vast majority (∼93%) of trait- and disease-associated loci lie within non-coding regions, of which thousands are located in the intergenic space, far away from any annotated gene[197]. Recent studies show that a large proportion of noncoding GWAS SNPs (Single Nucleotide Polymorphisms) fall within or near DNase hypersensitive sites, suggesting an important role in gene regulation[110,197,198]. Furthermore, many such intergenic, noncoding variant sites have been shown to be under human-specific purifying selection[199,200]. Again, this suggests that cross-species evolutionary conservation provides only incomplete insights into the functionality of noncoding elements.

Ultra-conserved elements. At the other end of the evolutionary conservation spectrum, ultra-conserved elements (UCEs) are genome segments of 200 to 700 bases that are 100% identical in mouse, rat and human, and highly conserved all the way to dog and chicken[201]. Of the 481 UCEs identified in the human genome, only 93 fall within protein-coding exons, and 140 lie in gene deserts, more than 10kb away from the nearest annotated gene. Such extreme levels of evolutionary conservation are indicative of essential biological functions. As a result, these enigmatic noncoding UCEs have been studied extensively since their discovery. Results indicate that many UCEs function as enhancers with important, subtle roles in vertebrate development[202–205]. As for any enhancer, these regions may also be transcribed into eRNAs (see Section III.2.1). Other studies have established an association between a few transcribed UCEs, outside of enhancer regions, and cancer[206,207]. Still, a large fraction of UCEs remains uncharacterized.

Structured ncRNAs in unannotated regions. *De novo* computational methods predict millions of structured ncRNAs in the intergenic space – although these approaches generate a high number of false positives (see Section II.1.1.2). For example, using covariation information, Smith *et al.* predict more than four million evolutionarily constrained RNA structures, covering 13.6% of the human genome in total[208]. The authors estimate their false discovery rate to be in the range of 5 to 22%, although this might be an underestimation[209]. Even under this assumption, experimentally probing the most confidently predicted regions from such datasets – using RNA capture, for example (see Section II.1.3.3) – could help shed light on the biology of the noncoding genome.

Junk DNA: a platform for evolutionary innovation? There are several documented cases of exaptation – the process by which a trait's function shifts during evolution – of selfish DNA. Parasitic elements such as TEs, in particular, have been shown to contribute to the evolution of vertebrate gene regulatory networks[210,211], in line with Barbara McClintock's original idea of TEs acting as "controlling elements"[212]. For example, in a human-mouse comparative study, Sundaram *et al.* found thousands of orthologous, TE-derived transcription factor binding sites with strong evidence of purifying selection[213]. Other studies have highlighted the possible role of TEs in shaping the evolution of splice sites[214,215], promoters[216,217], enhancers[218,219], lncRNAs[220,221] and microRNAs[222]. Finally, a 2011 survey comparing the genomes of 29 mammals found more than 280,000 conserved noncoding elements derived from TEs in the human genome, covering a total of $\sim$7 Mb[223]. Overall, this suggests that genomic domestication of selfish DNA is a significant phenomenon that may drive functional innovation in eukaryotes.

It is important to note, however, that those examples probably constitute exceptions rather than the rule[224], and that most TEs in vertebrate genomes should be presumed functionless unless proven otherwise. One should also refrain from the temptation of trying to fit junk DNA into a teleological framework. TEs and other junk elements can indeed be co-opted, however they are not actively preserved in genomes as "evolvable" material, as some authors insinuate (see *e.g.* Jain[225] and Barroso in Ecker *et al.*[226]). In the words of Sydney Brenner,

> *"There is a strong and widely held belief that all organisms are perfect and that everything within them is there for a function. Believers ascribe to the Darwinian natural selection process a fastidious prescience that it cannot possibly have and some go so far as to think that patently useless features of existing organisms are there as investments for the future."*[227]

III.3. Long noncoding RNAs: the last frontier of gene annotation

III.3.1. Pervasive transcription in mammalian genomes

The evidence for widespread, "intergenic" transcription in mammalian genomes remained anecdotal (see *e.g.* Ashe *et al.*[228]) until the 2000s, when technology improvements enabled deep, genome-wide transcriptome surveys. Large-scale cDNA sequencing projects[147,229], RNA tiling array hybridization studies[109,230–232], genome-wide maps of active promoters[110,233] and finally, RNA-Seq[153] revealed a profusion of such unannotated, mainly noncoding transcribed elements in both human and mouse. Noncoding transcription is not confined to the intergenic space, and as much as 11% of the human transcriptome derives from ncRNA transcripts that are interleaved with protein-coding ones[234].

In a landmark study using ultra-deep RNA-Seq in 15 human cell lines, the ENCODE consortium reports that overall, about half of the human genome is covered by mature, polyadenylated transcripts. Those cover a quarter of the intergenic space, and this fraction is mainly accounted for by long noncoding RNAs. Not surprisingly, detected lncRNAs, including intergenic ones, are expressed at a much lower level than protein-coding genes: about 80% of them are present in ENCODE samples in one or fewer copies per cell, compared with 25% for protein-coding transcripts[153]. However, low copy numbers measured in such bulk experiments do not necessarily reflect consistently low abundance in all cells, since RNA-Seq averages expression values across the cell population under study, as detailed in Section II.1.3.

Transcription does not imply function, however, and the biological significance of pervasive transcription is controversial[235]. Transcription initiation by RNA polymerase II (Pol II) – the enzyme synthetizing the large majority of intergenic and protein-coding transcripts – is known to be a leaky, stochastic process[236–238]. Consistent with their overall low expression levels and evolutionary conservation[59], a significant fraction of intergenic transcripts may be spurious products of such noisy transcription, and destined to rapid degradation by the cell machinery.

III.3.2. LncRNAs: a heterogeneous gene class

LncRNAs are long (>200 nts), usually capped and polyadenylated transcripts with no discernible protein-coding potential[239]. They constitute a broad category of transcripts, present in a wide range of eukaryotic taxa, from fungi to plants and animals[240,241].

III.3.2.1. Early lncRNA discoveries

The characterization of the first mammalian lncRNAs actually predates the genomic era and the discovery of widespread intergenic transcription. *H19*, the first discovered eukaryotic lncRNA[242], is a spliced, polyadenylated 2.3kb-long RNA product, relatively conserved across mammals (Figure S3)[243]. *H19* directly acts as an RNA molecule and is involved in the control of growth and cell proliferation during early mammal embryonic development[244]. Shortly after the discovery of *H19* in the early 1990s, it was found that *Xist* ("X-inactive specific transcript"), another lncRNA, was the main effector of chromosome X inactivation in female mammals[245–247]. *Xist*, an extraordinarily long (∼17 kb) spliced RNA, acts by "coating" one of the X chromosomes, which ultimately leads to the transcriptional silencing of the entire chromosome[248]. Interestingly, *Xist* does not exhibit a particularly high level of conservation: its average sequence similarity among mammals does not exceed that of the UTRs of mRNAs (Figure S10)[249,250].

III.3.2.2. LncRNA biology in the genomic era: map first, ask questions later

At a time when the main focus of attention was protein-coding genes, the examples of *H19* and *Xist* brought lncRNAs into the spotlight, and stimulated a global undertaking to identify and characterize noncoding genes. Early cDNA sequencing efforts, notably from the FANTOM initiative, uncovered thousands of such RNAs covering both strands of the human and mouse genomes[70,147,251]. Despite this explosion in the number of annotated mammalian lncRNAs, only a handful had been clearly characterized functionally by the end of the 2000s[51]. Besides *Xist* and *H19*, examples included noncoding genes involved in the regulation of nuclear import (*Nron*[252]), *trans*-acting gene regulation (*HOTAIR*[253], Figure S4, *MALAT1*[254], Figure S6), and genetic imprinting (*Air*[255,256], Figure S1) (see Section III.3.2.4).

Detecting lincRNAs via chromatin signatures. In 2009, Guttman *et al.* tackled the problem of lncRNA detection from a radically original angle. Instead of deep transcriptome sequencing, they opted for chromatin signatures to identify transcribed intergenic features in mouse[51]. The team first observed that genes actively transcribed by Pol II were associated with the H3K4me3 mark at their promoter and H3K36me3 (trimethylation of histone H3 at lysine 36) along their transcribed region[257]. They then took advantage of these so-called "K4-K36 domain" signatures to detect long intervening (*i.e.* intergenic) noncoding RNAs (lincRNAs, or standalone lncRNAs, see Figure 14) genome-wide. Using this method, the authors could identify more than a thousand novel lincRNA regions – with many bearing signs of purifying selection – and were able to reveal important biological functions for a few of them[51,258]. Nonetheless, however sensitive in pinpointing transcribed genomic regions this epigenomic approach is, it does not directly provide full-length transcript models *per se*, in contrast to transcriptomic techniques.

SGS-based lncRNA annotation. More studies followed Guttman's results, this time based on non-targeted, short-read transcriptome sequencing. Each added its own worth of novel mammalian lncRNAs to previous gene catalogs, while discovering little to no novel protein-coding genes[259–262], supporting the idea that the coding gene catalog is saturating. Others employed the targeted CaptureSeq methodology (see Section II.1.3.3) to dive even deeper into the noncoding transcriptome, uncovering lncRNAs with estimated concentrations as low as ∼0.0006 transcripts per cell[142,143,263]. Importantly, all of the aforementioned large-scale lncRNA sequencing surveys employed short-read sequencing methods, whose pitfalls have been previously discussed (see 'Second-Generation Sequencing (SGS) methods', page 14).

LncRNA annotation resources. LncRNA annotations are compiled into generic, curated gene sets (*e.g.* GENCODE[36], RefSeq[71]) or into databases containing software-reconstructed transcript catalogs, often specifically lncRNA-oriented (*e.g.* BIGTranscriptome[264], NONCODE[265], MiTranscriptome[262], FANTOM CAT[266]). These resources vary greatly in size: GENCODE and RefSeq contain ∼15,000 lncRNA locus entries (Figure 10), while NONCODE, the currently most comprehensive catalog (which integrates many of the aforementioned collections), reports close to 100,000 of them. Merits and weaknesses of different annotation approaches have been discussed previously (see Section II.1.2.2). Thanks to extensive manual curation, and by excluding SGS-assembled transcripts from its sources, GENCODE arguably provides the most conservative, highest-quality reference catalog for lncRNAs[267,268].

However, this comes at the cost of low sensitivity[269]: GENCODE's cautious choice of data sources implies that most of its underlying empirical evidence currently derives from low-depth, Sanger-based EST and cDNA surveys. Importantly, a systematic comparison of lncRNA gene collections has yet to be performed.

III.3.2.3. "Omics" of lncRNAs

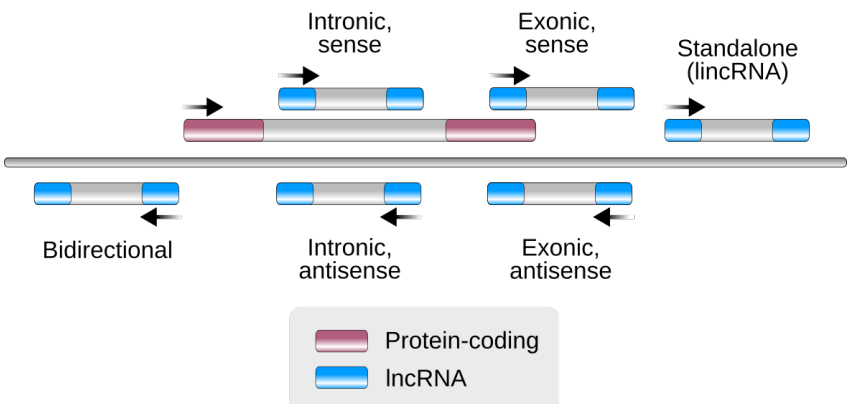


Figure 14: Positional classification of long noncoding RNAs. LncRNAs are categorized based on their genomic location with respect to annotated protein-coding genes. Colored and light gray boxes represent exons and introns, respectively.

In the absence of a clearly defined biological function, most lncRNAs are classified according to their genomic position with respect to protein-coding genes (Figure 14). LncRNAs interleaved with protein-coding genes are further subdivided based on the features they cover (intronic or exonic) on the one hand, and their genomic strand relative to the coding gene (sense or antisense) on the other. Bidirectional lncRNAs share a divergent promoter with a protein-coding gene. Finally, standalone lncRNAs (lincRNAs) lie in the intergenic space, sometimes far away from any other annotated gene. In general, belonging to one of these categories does not augur the biological function of the lncRNA in question. Anecdotal evidence suggests, however, that some antisense lncRNAs tend to regulate the expression of their "host" coding gene in *cis*[270] (see also Section III.3.2.4).

Detailed analyses of recent gene catalogs have revealed various genomic peculiarities apparently distinguishing lncRNA transcripts from their coding counterparts. Notably, mature lncRNAs typically have fewer exons[260,268,271] and are shorter than

protein-coding transcripts[268,271,272]. Other prominent lncRNA idiosyncrasies are listed in the next paragraphs.

Expression and RNA processing landscape. In general, lncRNAs display low expression levels, with a few notable exceptions (see examples in Section III.3.2.4). Together, they typically account for less than 3% of the RNA mass in tissue samples[273], with median abundances at least an order of magnitude lower than that of protein-coding genes[260,268,271,272] (see also Section III.3.1). Recent data suggests that these low steady-state levels result from reduced rates of transcription rather than RNA instability[274]. LncRNAs are also expressed in a highly tissue-specific fashion, predominantly in brain and testis cell types[260,268]. LncRNAs seem to partly share biogenesis pathways with mRNAs, including Pol II transcription, 5′ capping, 3′ polyadenylation and splicing[239]. Interestingly, lncRNAs exhibit slightly less efficient splicing than coding transcripts[275], although they use exon splicing signals of similar strength[268]. Many well-studied lncRNAs function in the cell nucleus (see Section III.3.2.4), and large-scale analyses seem to confirm that overall, this compartment is indeed enriched in lncRNAs compared to coding transcripts[268]. This enrichment may indicate localized functions in the nucleus, however it is also consistent with a rapid turnover and degradation of spurious lncRNA transcripts after their synthesis.

Evolutionary genomics of lncRNAs. In order to avoid the confounding influence of overlapping protein-coding genes, evolutionary analyses of lncRNA sequences usually focus on standalone lincRNAs. As a gene class, lincRNAs are much less evolutionarily conserved than protein-coding genes at the primary sequence level. Deep comparative surveys of lncRNA repertoires also indicate that a large majority of lincRNA genes are lineage- or species-specific and undergo rapid turnover[276]. At least 20% of human lincRNAs show hominid-specific expression[277]. Further, sequence divergence among vertebrates is so extreme that only 6% of zebrafish lincRNAs are alignable to human and mouse lincRNAs at the primary sequence level[272]. Between human and mouse, only 12% of lincRNAs are conserved[260]. Even functional lincRNAs with identified orthologs, such as *NEAT1*, sometimes show no evidence of evolutionary constraint over large stretches of their exonic sequence (Figure S7)[267,278]. However, evidence has been found of both secondary structure[279] and positional conservation[272] for at least a few cases of lincRNAs that are not conserved at the primary level. Within functional lncRNAs, some correlation can be found between type and level of evolutionary conservation on the one hand, and mode of action on the other. For example, lncRNAs exhibiting strong synteny

in combination with low sequence conservation tend to act in *cis* on nearby genes[267] (see Section III.3.2.4).

LncRNAs are enriched in transposable elements: an estimated 41% of lncRNA nucleotides are covered by TEs, and altogether, more than 80% of annotated lncRNAs contain at least one TE, or relic thereof[280]. While this observation is consistent with "junk", spurious transcription, a few cases of TE exaptation into functional elements have been documented (see Section III.2.2), including for lncRNAs. Examples include the mammalian *ANRIL* gene, which domesticated multiple TEs into exons in the primate lineage[281], *Xist*[282] and *AS-Uchl1*[270]. It has been proposed that TEs occasionally play critical roles in lncRNA evolutionary innovation by providing novel functional domains[220,221].

Genomic environment and annotation quality. If relating genes to their genomic surroundings can shed light into various aspects of their biology (see Section II.1.4), it can also be used to assess the quality and completeness of annotated gene models. An in-depth analysis of the GENCODE catalog revealed that lncRNAs are globally depleted in hallmarks of transcription initiation (*e.g.*, CAGE clusters[146]) and termination (*e.g.*, paired-end ditags[283], polydadenylation signals[284]) when compared to coding genes[268]. Other teams have made similar observations, ascribing these differences to globally differential transcription regulation of lncRNAs and protein-coding genes[274,285]. A more parsimonious explanation attributes these properties to annotation artifacts rather than real biology, however. It is indeed plausible that the lack of experimental support for the 5′ and 3′ ends of lncRNAs, together with their short length and low number of exons, is simply due to the incompleteness of available lncRNA gene models. This latter hypothesis is consistent with the low expression levels of lncRNAs and the resulting difficulties in sampling them from RNA extracts in a non-truncated full-length form (see Section II.1.3).

Are lncRNAs translated? Finally, there is evidence that at least some genes, initially labelled as lncRNAs, are in fact translated into proteins[286,287]. Those typically encode small peptides (< 100 aminoacids) that are difficult to detect for annotation pipelines because of their short length. The introduction of high-throughput sequencing of ribosome-protected fragments (Ribo-Seq) revealed unexpected associations between lncRNAs and ribosomes, suggesting that the former undergo translation[288]. Although ribosome association does not necessarily imply translation into a functional protein product, polypeptide production could be demonstrated in at least a few cases[289,290].

III.3.2.4. Known functional lncRNAs

Functional characterization, to this day, still struggles to keep pace with the ever-growing catalog of lncRNAs: of the tens of thousands of currently annotated lncRNA genes, less than 500 have robustly assigned biological functions[291]. Based on our current understanding, functional lncRNAs can be subdivided into those that perform regulatory functions in *cis*, and those that act in *trans*[292].

Cis-acting lncRNAs. *Xist* (see Section III.3.2.1) is the most famous and well-characterized example of a *cis*-acting lncRNA. *Air* (also known as *AIRN*, Figure S1) is an lncRNA that silences its target gene, *Igf2r*, through antisense transcription[255,256]. Intriguingly, *Air*'s regulatory activity is sequence-independent, and stems solely from the process of its transcription[293]. More cases of such *cis*-regulatory lncRNAs that carry out their function through the act of transcription, rather than via their RNA product, have been recently described, such as *Upperhand*[294] (also known as *HAND2-AS1*, Figure S9), *Blustr* (also known as *Gm13261*, Figure S2) in mouse[295], and *Linc-p21*[296]. These examples may represent a common mechanism of sequence-independent transcriptional regulation. As one could expect from such a non-sequence-specific mode of action, in most cases the level of conservation of these lncRNAs is low[293,295] (see Supplementary Figures).

Trans-acting lncRNAs. LncRNAs can act away from their site of transcription, *i.e.*, in *trans*. *HOTAIR* (Figure S4), for example, is thought to function as a scaffold that recruits chromatin-modifying complexes to the distant *HOXD* locus[253,297]. In mouse, *lincRNA-EPS* (also known as *Ttc39aos1*, Figure S5) has been shown to down-regulate in *trans* various immune response genes, by interacting directly with their promoter sequences in macrophages[298,299]. Remarkably, *lincRNA-EPS* is expressed at relatively low levels (∼11 copies per cell[299]) and harbors only very limited sequence conservation with its presumed human ortholog[298].

LncRNAs can also play structural roles in the functional organization of the nucleus's architecture, such as *MALAT1* (Figure S6), an abundant, highly conserved noncoding gene, or *NEAT1* (Figure S7). Both these RNAs are associated with active chromatin, but in distinct, functionally important sub-nuclear bodies – nuclear speckles for the former, paraspeckles for the latter. Evidence suggests that they contribute to the localization of their respective sub-nuclear complexes close to actively transcribed genes[254,300–305].

Other functions of *trans*-acting lncRNAs include the regulation of the activity of

other proteins or RNAs in the cell. *NORAD* (Figure S8), owing to its high affinity with *PUMILIO* proteins, acts as an inhibitor of their activity through a molecular decoy mechanism[306,307]. *CDR1as* (also known as *ciRS-7*) also functions as a molecular "sponge", but with microRNAs instead of proteins[308,309].

III.3.2.5. Navigating the vast *terra incognita* of uncharacterized lncRNAs.

Searching for functional clues in lncRNAs. Although definitely demonstrating non-functionality is impossible, a large proportion of lncRNAs are likely to be spurious products of transcriptional noise, as discussed in Section III.3.1. Understandably, sequence conservation is often employed to infer functionality – or lack thereof – or prioritize functional characterization of newly discovered lncRNA gene models. Primary sequence conservation is an extremely *specific* predictor of function – one would reasonably assume a conserved element to be functional. However, as the numerous aforementioned examples of experimentally validated lncRNAs show, predictions solely based on sequence conservation suffer from a substantial rate of false negatives – *i.e.*, non-conserved, albeit functional lncRNAs (Figure 15).

Similarly, low expression levels are often considered evidence of the bogus nature of lncRNAs. While reasonable on the surface, this argument is mostly based on RNA concentrations in "bulk" experiments, and ignores possibly localized modes of action, as exposed in Sections II.1.3 and III.3.1. Furthermore, examples of low-abundance, yet functional lncRNAs are starting to emerge, including *lincRNA-EPS* (see Section III.3.2.4) and *VELUCT*, a recently described regulatory lncRNA present at less than one copy per cell[310].

Finally, some characteristics of noncoding RNAs, such as proper post-transcriptional processing (*e.g.*, splicing, 5′ capping and polyadenylation) and tissue-specific expression, are often insistently taken as indicators of function (see *e.g.* Mercer *et al.*[311], Dinger *et al.*[312]). This is likely a misinterpretation, however, as all these observations are perfectly compatible with the intrinsically noisy nature of biochemical processes[239]. It is also tempting to view lncRNA's combination of tissue-specific expression patterns and volatile sequence evolution as suggestive of lineage-specific, delicately orchestrated gene regulation events during development. This model is supported by some recent experimental data[313,314], but may be considered anecdotal within the large, heterogeneous universe of lncRNAs until further evidence is gathered.

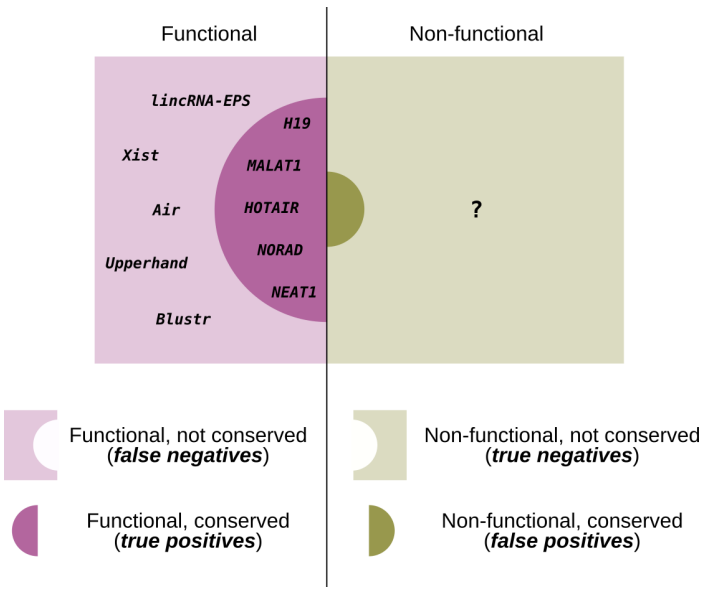


Figure 15: Evolutionary conservation of primary sequence as a predictor of lncRNA function. LncRNAs can be subdivided into functional (left, purple) and non-functional (right, green) categories. Evolutionary conservation (EC) of RNA sequences is often used to infer their biological significance – namely, their function. EC will predict true positives (conserved, functional elements), false negatives (non-conserved, functional elements) as well as true negatives (non-functional, non-conserved sequences) and false positives (conserved, non-functional sequences). It is generally admitted that EC should generate little to no false positives: it is highly *specific*. However, experimental data shows that EC sometimes fails to predict functional lncRNAs, namely, it lacks in *sensitivity*. A few examples of functional lncRNAs are reported in the corresponding plot areas (see text and Supplementary Figures for details and bibliographic references). Note that neither the area sizes, nor the number of examples represented are exactly proportional to the real size of the corresponding sets.

High-quality annotations as a foundation of lncRNA experimental characterization. Gene function can be studied using reverse genetics perturbation techniques (see Section II). The CRISPR-Cas9 method (Clustered Regularly Interspaced Short Palindromic Repeats Cas9) coupled with individual single-guide RNAs (sgRNAs) efficiently produces loss-of-function mutations in protein-coding genes[315,316]. Because noncoding sequences are much less constrained, such point

mutations are unlikely to reliably inactivate lncRNAs, however. An additional, prominent difficulty with lncRNA functional studies lies in the capacity to distinguish between effects due to *cis*-acting DNA elements *vs* RNA-level modes of action when attempting to explain a phenotype. As a result, lncRNA functional screens often employ combinations of several disruption techniques, including promoter or gene knockouts using pairs of sgRNAs[317,318], genomic insertion of premature polyadenylation signals[295], CRISPRi (CRISPR interference, *i.e.*, CRISPR-mediated transcription inhibition), CRISPRa (CRISPR-mediated activation of transcription)[319], and RNA interference[320].

These experimental methods hold the promise of genome-scale, high-throughput lncRNA functional surveys. They also have one important feature in common, which is that they rely on accurate, exhaustive gene annotations. As presented above, there is ample evidence, however, that mammalian lncRNA gene sets lack in such qualities. SGS-based resources, while extremely sensitive, suffer from patent inaccuracies due to the difficulties in reassembling transcript models from short read data (see 'Second-Generation Sequencing (SGS) methods', page 14 and Section III.3.2.2). In comparison, reference catalogs such as GENCODE, while lacking in depth, arguably provide a more precise view of the mammalian transcriptome – although reference transcript models still exhibit doubtful characteristics, particularly at their boundaries (see Section III.3.2.3). In other words, annotation incompleteness stems from a currently inevitable trade-off between depth and quality, which consequently hampers the full deployment of lncRNA functional genomic studies.

Objectives

The main objective of the present Thesis Project is to produce a more accurate long noncoding RNA genome annotation than is currently available in human. Providing such a complete and reliable resource is critical to the understanding of this gene class's biology. Using targeted transcriptome sequencing methods, the objectives of this Thesis Project are:

1. To develop high-throughput, high-quality lncRNA annotation methods based on current long-read sequencing technologies, reducing manual intervention to a minimum;

2. To improve the accuracy of the currently annotated lncRNA transcript models in the reference GENCODE resource, and to advance towards a comprehensive catalog of human lncRNA gene loci;

3. In light of this enhanced annotation, to re-evaluate the genomic properties of lncRNAs reported in past studies;

4. To systematically compare the strengths and weaknesses of publicly available lncRNA annotation resources, including the improved GENCODE catalog generated in this Thesis Project.

Impact and authorship report of the publications

All three peer-reviewed articles included in this thesis dissertation have been published in high-impact journals. Julien Lagarde is the co-first author of articles I and II, and second author of article III, a review. The specific contributions of Julien Lagarde to each publication are indicated in the following sections, together with the 5-year impact factor of the journal, as reported by the Nature Publishing Group's 2018 journal metrics[1]. Individual author contributions are also available within each published article.

[1] https://www.nature.com/nature-research/about/journal-metrics

I

Extension of lncRNAs with RACE-Seq

Lagarde J, Uszczynska-Ratajczak B, Santoyo-Lopez J, Gonzalez JM, Tapanari E, Mudge JM, Steward CA, Wilming L, Tanzer A, Howald C, Chrast J, Vela-Boza A, Rueda A, Lopez-Domingo FJ, Dopazo J, Reymond A, Guigó R, Harrow J. *Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq)*. ***Nature Communications*** 2016 Aug 17; 7:12339.

- **URL:** https://doi.org/10.1038/ncomms12339
- **5-year impact factor (2018): 13.811**
- **Author's contribution:** Julien Lagarde made major contributions to the design of the experiment and the writing of the manuscript, and analyzed the data.

II

High-throughput annotation of lncRNAs with CLS

Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R, Johnson R. *High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing*. ***Nature Genetics*** 2017 Dec; 49(12):1731-1740.

- **URL:** https://doi.org/10.1038/ng.3988
- **5-year impact factor (2018): 31.077**
- **Author's contribution:** Julien Lagarde made major contributions to the design of the project, designed the CLS bioinformatics workflow, analyzed the data, and contributed to the writing of the manuscript. He also wrote most of the original code for this project (see Appendix, 'Relevant software written by the author').

III

Towards a complete map of human lncRNAs

Uszczynska-Ratajczak B, **Lagarde J**, Frankish A, Guigó R, Johnson R. *Towards a complete map of the human long non-coding RNA transcriptome*. ***Nature Reviews Genetics*** 2018 Sep; 19(9):535-548.

- **URL:** https://doi.org/10.1038/s41576-018-0017-y
- **5-year impact factor (2018): 42.812**
- **Author's contribution:** Julien Lagarde participated in the analysis of the data, and contributed original code (notably the `buildLoci` software, see Appendix, 'Relevant software written by the author').

The director, Roderic Guigó

Publications

I

Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq)

Lagarde J, Uszczynska-Ratajczak B, Santoyo-Lopez J, Gonzalez JM, Tapanari E, Mudge JM, Steward CA, Wilming L, Tanzer A, Howald C, Chrast J, Vela-Boza A, Rueda A, Lopez-Domingo FJ, Dopazo J, Reymond A, Guigó R, Harrow J. *Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq).* ***Nature Communications*** 2016 Aug 17; 7:12339.

URL: https://doi.org/10.1038/ncomms12339

Abstract:

Long non-coding RNAs (lncRNAs) constitute a large, yet mostly uncharacterized fraction of the mammalian transcriptome. Such characterization requires a comprehensive, high-quality annotation of their gene structure and boundaries, which is currently lacking. Here we describe RACE-Seq, an experimental workflow designed to address this based on RACE (rapid amplification of cDNA ends) and long-read RNA sequencing. We apply RACE-Seq to 398 human lncRNA genes in seven tissues, leading to the discovery of 2,556 on-target, novel transcripts. About 60% of the targeted loci are extended in either 5′ or 3′, often reaching genomic hallmarks of gene boundaries. Analysis of the novel transcripts suggests that lncRNAs are as long, have as many exons and undergo as much alternative splicing as protein-coding genes, contrary to current assumptions. Overall, we show that RACE-Seq is an effective tool to annotate an organism's deep transcriptome, and compares favourably to other targeted sequencing techniques.

I.1. Main article

nature COMMUNICATIONS

ARTICLE

Received 20 Apr 2016 | Accepted 23 Jun 2016 | Published 17 Aug 2016

DOI: 10.1038/ncomms12339 OPEN

Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq)

Julien Lagarde[1,2,*], Barbara Uszczynska-Ratajczak[1,2,*], Javier Santoyo-Lopez[3,†], Jose Manuel Gonzalez[4], Electra Tapanari[4,†], Jonathan M. Mudge[4], Charles A. Steward[4], Laurens Wilming[4], Andrea Tanzer[1,2,†], Cédric Howald[5,†], Jacqueline Chrast[5], Alicia Vela-Boza[3,6], Antonio Rueda[3], Francisco J. Lopez-Domingo[3], Joaquin Dopazo[3,7,8], Alexandre Reymond[5], Roderic Guigó[1,2] & Jennifer Harrow[4]

Long non-coding RNAs (lncRNAs) constitute a large, yet mostly uncharacterized fraction of the mammalian transcriptome. Such characterization requires a comprehensive, high-quality annotation of their gene structure and boundaries, which is currently lacking. Here we describe RACE-Seq, an experimental workflow designed to address this based on RACE (rapid amplification of cDNA ends) and long-read RNA sequencing. We apply RACE-Seq to 398 human lncRNA genes in seven tissues, leading to the discovery of 2,556 on-target, novel transcripts. About 60% of the targeted loci are extended in either 5′ or 3′, often reaching genomic hallmarks of gene boundaries. Analysis of the novel transcripts suggests that lncRNAs are as long, have as many exons and undergo as much alternative splicing as protein-coding genes, contrary to current assumptions. Overall, we show that RACE-Seq is an effective tool to annotate an organism's deep transcriptome, and compares favourably to other targeted sequencing techniques.

[1] Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Dr Aiguader 88, 08003 Barcelona, Spain. [2] Universitat Pompeu Fabra (UPF), Barcelona, Spain. [3] Genomics and Bioinformatics Platform of Andalusia (GBPA), 41092 Seville, Spain. [4] Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1HH, UK. [5] Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. [6] Roche Diagnostics, 08174 Sant Cugat Del Vallès, Barcelona, Spain. [7] Computational Genomics Department, Centro de Investigación Príncipe Felipe, 46012 Valencia, Spain. [8] Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe, 46012 Valencia, Spain. * These authors contributed equally to this work. † Present addresses: Edinburgh Genomics, The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, Edinburgh EH25 9RG, UK (J.S.-L.); European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK (E.T.); Department of Theoretical Chemistry, University of Vienna, Waehringerstrasse 17, 1090 Vienna, Austria (A.T.); Division of Genetic Medicine, Geneva University Hospitals, Geneva, Switzerland (C.H.). Correspondence and requests for materials should be addressed to R.G. (email: roderic.guigo@crg.cat) or to J.H. (email: jla1@sanger.ac.uk).

The mammalian transcriptome is composed of a complex mixture of protein-coding and non-protein-coding RNA molecules. Increasing interest has been brought to bear on the latter, most notably on long non-coding RNAs (lncRNAs). A small but growing number of lncRNAs has been reported to play diverse roles in biological and pathological processes[1,2]; however, the vast majority still awaits functional characterization. Such characterization depends on accurate and comprehensive annotation of the complete repertoire of lncRNA transcript structures. This has been the focus of considerable efforts in recent years[3–7]. Arguably the most refined and widely used lncRNA annotation is the catalogue of 15,000 human lncRNA loci published by GENCODE, alongside the Encyclopedia of DNA Elements (ENCODE) data release in 2012 (ref. 8). Various consortia, including the 1,000 genomes project[9], The Cancer Genome Atlas (TCGA)[10] and The Genotype-Tissue Expression (GTEx) Consortium[11] use GENCODE as their reference annotation.

LncRNA gene annotations remain incomplete and methods to define them continue to evolve. In contrast to protein-coding genes, lncRNA gene annotations tend to have poorly defined boundaries, as judged by their lack of characteristic hallmarks of transcription initiation and termination[8]. While computational methods can provide some guidance[12], accurate gene annotation requires the use of high-confidence transcriptomic evidence, such as sequencing of full-length cDNA[13]. Until a few years ago, only low-depth techniques, such as Sanger sequencing of expressed sequence tags (ESTs)[14], were used. Recent advances in high-throughput cDNA sequencing technology, that is, RNA-seq[15,16], have provided deep sampling of the human transcriptome[17]. When used in the context of gene annotation, however, these techniques still exhibit limitations due to the necessary compromise between read length and sequencing depth. Long-read sequencing (for example, Roche 454, Pacific Biosciences) can in principle provide close to full-length transcript sequences, but at low depth. Short-read RNA-Seq experiments (for example, Illumina Hi-Seq) routinely produce hundreds of millions of reads. However, such reads are far shorter than a typical mRNA or lncRNA transcript, which severely hampers accurate full-length isoform assembly[18]. In summary, current non-targeted, conventional cDNA sequencing methods are ineffective for reading the full dynamic range of transcript expression in the cell. This means that low-expressed transcripts, that is, the majority of lncRNAs, suffer from incomplete annotations.

Technical methods are being developed to address the problem of low-abundance transcript annotation. Recently, a high-throughput sequencing method called CaptureSeq was used for lncRNA characterization, in conjunction with Illumina short-read sequencing. It achieves targeted transcript enrichment by the hybridization of cDNA (derived from cellular RNA) to bead-linked oligonucleotide probes that are tiled and complementary to exons[19,20]. RNA CaptureSeq proved to be effective for the discovery of novel lowly expressed transcripts and allows for their quantification and assembly. However, this procedure has not been designed to specifically address the proper definition of 5′ and 3′ transcript ends, and as a result other methods are required for the precise experimental annotation of gene boundaries.

To improve the annotation of the boundaries of low-expressed genes, we coupled the widely used RACE technique (rapid amplification of cDNA ends[21]) to high-throughput sequencing—'RACE-Seq'. In RACE-Seq, we carry out RACE with primers designed in targeted loci with the aim of producing cDNA sequences that reach the transcript termini. RACE products are then subjected to high-throughput long-read sequencing (for example, Roche 454). We here apply RACE-Seq to a selection of 398 lncRNA loci from the reference GENCODE v7 catalogue[7], most of them low-expressed and lacking typical gene boundary hallmarks. We discover 2,556 novel, manually curated rare isoforms. Two thirds of those extend their parent locus beyond their previously annotated boundaries, often reaching marks of transcription initiation and termination, such as CAGE tags and poly-adenylation sites. We found that both the sensitivity and specificity of RACE-Seq are greatly enhanced by the use of a second, nested set of priming oligonucleotides. Overall, we show that RACE-Seq is a highly efficient method, well-suited for both novel isoform discovery and gene boundary characterization.

Results

RACE-Seq general strategy and proof-of-concept. The outline of the RACE-Seq procedure is depicted in Fig. 1. For each locus in a given set of annotated genes, 5′ and/or 3′ RACE primers are designed *in silico* along the transcript sequences so that the resulting RACE product has a suitable size for the long-read sequencing platform in use (see Methods). To limit off-target RACE amplification, it is beneficial to ignore primers exhibiting substantial sequence identity with any transcribed region in the genome other than their intended target (>80% identity in our test case). To increase further RACE specificity, a second 'nested' primer, placed as close as possible, downstream of the first one, can be designed using the same selection criteria as before. RACE reactions are then carried out in RNA extracted from the cellular samples of interest. Finally, RACE products are subsequently sequenced using a high-throughput long-read sequencing platform, and resulting reads are aligned and assembled into spliced transcripts on the genome.

As a proof-of-concept, we targeted 398 distinct lncRNA loci from the GENCODE v7 annotation[7,8], and performed RACE-Seq on a set of cDNA libraries from 7 human tissues (brain, heart, kidney, liver, lung, spleen and testis) known to cover a large fraction of the annotated human transcriptome[22]. We subdivided our set of target lncRNAs in two subsets, depending on whether their annotated 5′ end was supported by CAGE tags (as identified by the FANTOM project[3] ($N=180$), or not ($N=218$) (see Methods)). The RACE cDNA mixtures were then sequenced using the Roche 454 FLX+ platform. Reads obtained have an average length of ~600 bp. Sequenced reads were mapped to the genome using a combination of BLAT[23] and GMAP[24], and the resulting alignments were manually curated and incorporated into the GENCODE human gene set.

We obtained a first batch of RACE-Seq (referred to as 'standard RACE' below) using standard, non-nested RACE primers in each of the 398 targets. We then performed nested RACE on aliquots of the standard RACE reactions so as to improve the assay's sensitivity and specificity. In total, adding an extra pilot set of standard RACE libraries, we sequenced ~22 million reads in 40 RACE libraries (Supplementary Figs 1 and 2). We obtained at least one alignable RACE product for 94% of the 398 targeted loci, and discovered at least 1 novel, manually curated isoform for 343 of them.

Novel gene boundaries. With RACE-Seq, out of the 398 targeted lncRNAs, we extended 176 and 193 loci further in 5′ and 3′, respectively, and 131 in both directions (Fig. 2a, left panel). In total, the boundaries of 238 loci (60%) were expanded in either direction. These genomic extensions were accounted for by 752 and 848 distinct 5′ and 3′ RACE products, respectively (Fig. 2a, right panel). Eighty two novel transcripts extended their parent locus in both 5′ and 3′.

RACE-Seq was particularly successful in extending CAGE-unsupported loci: the median/mean genomic length of 5′ extensions were +21/−8,479 and −376/−14,440 (negative

NATURE COMMUNICATIONS | DOI: 10.1038/ncomms12339 ARTICLE

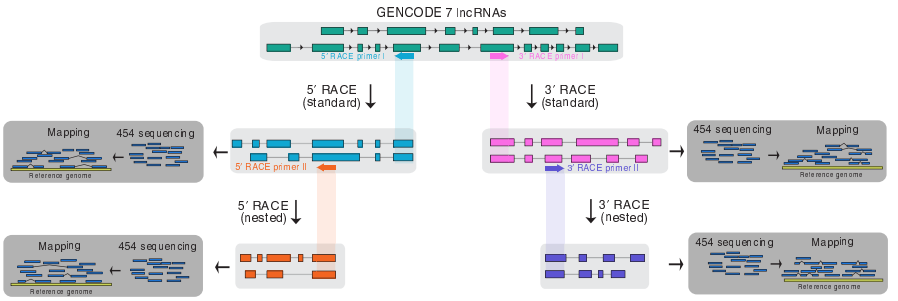


Figure 1 | Schematic overview of RACE-seq. Standard 5′ and 3′ RACE primers are designed to target exons of a gene and produce primary RACE products, which undergo a second round of RACE reactions using nested 5′ or 3′ RACE primers. Both standard and nested 5′ and 3′ RACE products are subjected to long-read sequencing, followed by mapping to the reference genome.

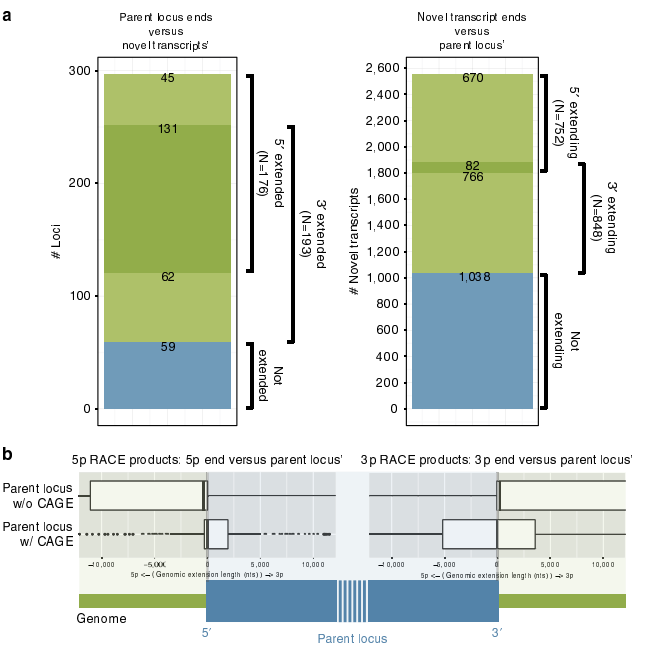


Figure 2 | Locus extension and novel transcript boundaries. (**a**) Venn diagrams depicting the proportion of loci (left panel) and transcripts (right panel) extended in 5′ and/or 3′ direction. (**b**) Novel boundaries for CAGE-supported (bottom box-plot) and CAGE-unsupported loci (top box-plot). A schematic depiction of a target locus is provided below the plots. The viewing range of the box plots is reduced (−10,000, 10,000 nucleotides) for clarity.

values represent novel transcription start sites (TSSs) upstream of the annotated locus), respectively, for CAGE-supported and unsupported loci (Fig. 2b, and example in Fig. 5b). Surprisingly, we observed a similar phenomenon at the 3′ end of targeted loci: the mean/median genomic length of 3′ extensions amounted to −15/−526 and +225/+8,518 (positive values correspond to novel transcription termination sites (TTSs) downstream of the annotated locus'), respectively, for CAGE-supported and unsupported loci. We speculate that this observation is due to the pre-RACE-Seq GENCODE set being mostly based on oligo-dT-primed ESTs, which tend to cover preferentially the 3′ end of transcripts. As a consequence of this bias, a transcript model that is complete at its 5′ end (that is, CAGE-supported) is also likely to be complete at its 3′ end, which is consistent with our results.

Table 1 | Comparison of various TTS data sets with Merck PolyA-Seq peaks.

Data set	Total #TTS	#TTS close to a polyA-Seq tag (±100 nts)	% TTS close to a polyA-Seq tag (±100 nts)
Targets (pre-RACE)	535	83	16%
Targets updated (post-RACE)	1,027	99	10%
Protein coding	17,940	7,019	39%
lncRNAs	12,556	2,223	18%

lncRNA, long non-coding RNA; RACE, rapid amplification of cDNA ends; TTS, transcription termination site.
Statistics are also reported for the full sets of GENCODE-annotated protein-coding genes and lncRNAs for reference.

Table 2 | Comparison of pre- and post-RACE TTSs data sets with polyA peaks called using our RACE-Seq data.

Data set	Total #TTS	#TTS close to a RACE-Seq inferred polyA-Seq tag (±100 nts)	%TTS close to a RACE-Seq inferred polyA-Seq tag (±100 nts)
Targets (pre-RACE)	535	206	39%
Targets updated (post-RACE)	1,027	321	31%

RACE, rapid amplification of cDNA ends; TTS, transcription termination site.

In addition, we observed that when novel TSSs were discovered in CAGE-supported loci, they were much more likely to be supported by another CAGE peak than CAGE-unsupported loci (74% versus 56% CAGE support, see Supplementary Fig. 3). Overall, the experiment uncovered 873 non-redundant TSSs, of which 615 were previously unknown—including 252 (41%) that were CAGE-supported (Supplementary Table 1).

We also assessed the accuracy of the newly annotated TTSs by comparing them with experimentally established poly-adenylation (polyA) sites (Merck PolyA-Seq data sets[25]). The overall proportion of TTS within 100 nucleotides of a PolyA-Seq tag slightly decreased from 16% pre-RACE to 10% post-RACE (Table 1). Yet, the raw count of TTSs supported by PolyA-Seq was improved after RACE-Seq, albeit very marginally (from 83 to 99 PolyA-Seq-supported TTSs). In addition, we identified polyA sites ourselves by searching for non-templated polyA/T tails in partially mapped 3′ RACE-Seq reads. Using this method, we were able to precisely map 1,212 distinct polyA sites near our targets, and compared those with the 3′ ends of our transcript set. We observed a much higher number of TTSs in the near vicinity (± 100 nucleotides) of these sites (206 and 321 pre-RACE and post-RACE TTSs, respectively) (Table 2). This indicates that the low Merck PolyA-Seq coverage of our TTSs is probably due to the limited depth and tissue coverage of PolyA-Seq compared with our RACE-Seq data.

On-target enrichment and sensitivity of RACE-Seq. Since (1) RACE operates with only one internal oligonucleotide primer, and (2) our targeted genes are very lowly expressed ones, we expected this experiment to yield a high number of off-target products. We found that, on average across all samples, 94% of uniquely mapped sequencing reads overlapped GENCODE v7 genic regions (Supplementary Fig. 4), indicating insignificant genomic contamination of our cDNA libraries. The vast majority of reads arose from annotated genic regions, and 3.9% of them, on average, fell within the targeted locus boundaries when using standard RACE (Fig. 3a and Supplementary Table 2). This corresponds to a 3.1-fold enrichment of reads originating from transcripts compared with untargeted sequencing (as estimated using GTEx RNA-Seq data in matched tissues, see Methods). In contrast, nested RACE yielded an average of 36.4% on-target reads across all tissues (that is, a 9.5-fold increase in specificity compared with standard RACE, and 29.2-fold over expected in untargeted sequencing), allowing much deeper sequencing of the target loci. Similarly, the number of non-targeted loci producing reads decreased 2.2-fold when using nested RACE (on average, 5,025 amplified non-targeted loci in standard RACE, versus 2,332 in nested RACE, see Supplementary Fig. 5).

The total number of loci successfully amplified by RACE-Seq was 374 (94%, regardless of RACE direction), 326 (82%) and 341 (86%) for 3′ RACE and 5′ RACE, respectively (Supplementary Table 3). When further assessing the sensitivity of RACE-Seq, we noticed the benefits of nested RACE over standard RACE again. Overall, 12.5% (351 versus 312) more targets could be amplified in nested RACE-Seq than in standard RACE-Seq. The majority of positive loci (289, that is, 73% of the total) were detected in both nested and standard RACE, and only 23 of them were positive in standard RACE only (Supplementary Fig. 6).

In each individual tissue, nested RACE-Seq always outperformed standard RACE-Seq (Fig. 3b). The median number of positive targets was 49 (12%) across all standard RACE-Seq experiments, and 130 (33%) across the nested ones. The difference in sensitivity between nested and standard RACE samples was particularly remarkable in the kidney 3′ RACE samples (36% versus 8% success rate, respectively), and less noticeable in more transcriptionally complex tissues, such as testis (3′ RACE; 58 versus 50% success rate, respectively). We attribute the nested sets' sensitivity improvements to its better specificity, which, by limiting the number of off-target reads, leads to a deeper sampling of the targeted transcripts. Taken globally, these results indicate that, as one could expect, nested RACE-Seq is far more informative that standard RACE-Seq, that is, surpasses it in both sensitivity and specificity terms.

Novel isoforms in targeted regions and tissue origin. We extended 176 lncRNA loci at the 5′ end and 193 loci at the 3′ end out of the total of 398 loci targeted for extension from GENCODE v7 (Fig. 2a). After extension, re-annotation and loci merging where necessary, the total number of lncRNA loci was reduced to 343. One putative lncRNA locus (OTTHUMG00000009351), when extended, was found to bear coding potential as it was extended to overlap the *LRRC7* (Leucine-rich repeat containing) coding locus, thus its biotype was changed to protein-coding. About 57 transcripts were merged into existing protein-coding loci (see example in Supplementary Fig. 7, where a putative lncRNA is re-annotated to be part of the

NATURE COMMUNICATIONS | DOI: 10.1038/ncomms12339 ARTICLE

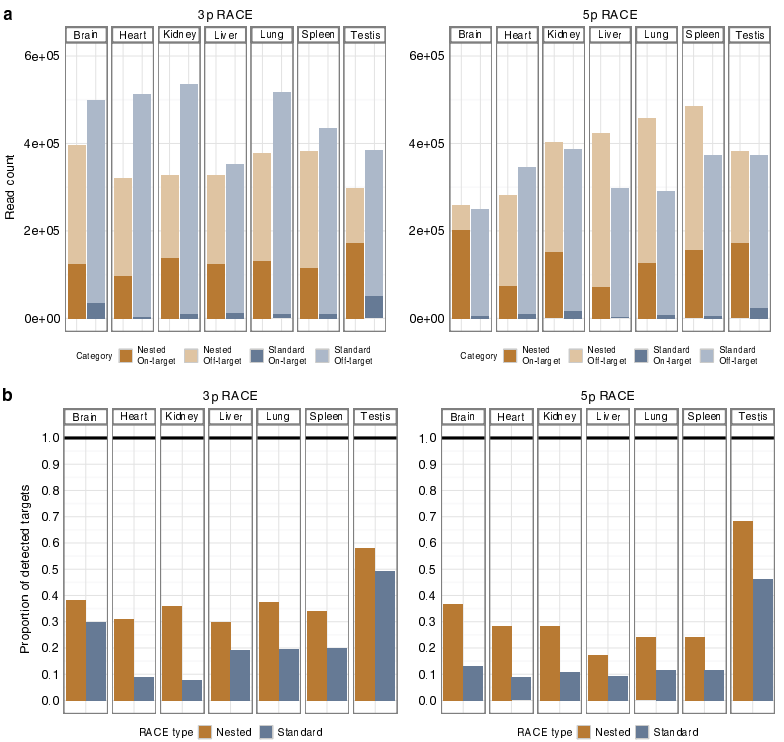


Figure 3 | On-target RACE enrichment and RACE-Seq specificity. (**a**) Number of RACE-Seq reads falling into exonic regions of targeted genes (dark shades) and outside of them (light shades), after using standard (blue) and nested (orange) 5′ and 3′ RACE. (**b**) Proportion of targets detected by standard (blue) and nested (orange) 5′ and 3′ RACE-Seq.

Table 3 | Table summarizing basic annotation statistics before and after RACE-Seq.

Data set	#loci	#transcripts	#transcripts per locus	#exons (all)	#exons (unique)	#exons per transcript
Targets (pre-RACE)	398	597	1.5	1,889	1,695	3.2
Targets updated (post-RACE)	343	2,556	7.5	10,139	5,326 (4,626)	4.0

RACE, rapid amplification of cDNA ends.
Unique exons are those having distinct coordinates on the genome. The number of previously unannotated unique exons is indicated between parentheses in the penultimate column.

PIGL (phosphatidylinositol glycan anchor biosynthesis, class L) locus using RACE-Seq read data). The number of alternatively spliced variants generated by the 5′ and 3′ RACE increased by >4 fold from 597 to 2,556 (Table 3), and the median length of the transcripts slightly increased from 623 to 704, although not significantly ($P = 0.7$, Wilcoxon rank sum test with continuity correction) (Fig. 4a). It should be mentioned that RACE, by design, does not produce full-length, TSS-to-TTS transcripts. This is because RACE products, by definition, start at their originating primer's position along the targeted transcript. Therefore, we speculate that the length of post-RACE transcripts is heavily underestimated.

The average number of transcripts per locus increased from 1.5 (597/398) pre-RACE to 7.5 (2,556/343) post-RACE (Table 3 and Fig. 4b). The total number of splice junctions increased from 1,093 pre-RACE to 3,085 post-RACE (Table 4). One lncRNA, *PCB1-AS1* (OTTHUMG00000153728), antisense to *PCBP1*, had the highest number of alternatively spliced transcripts, increasing from 40 transcripts pre-RACE to 170. The function of this lncRNA is currently unknown, however, the *PCBP1* protein is known to act as a translational coactivator[26] and mediate the degradation of mitochondrial antiviral signals. Interestingly, *PCB1-AS1* was already highlighted by Derrien *et al.*[8] as the most alternatively spliced lncRNA gene in the GENCODE v7

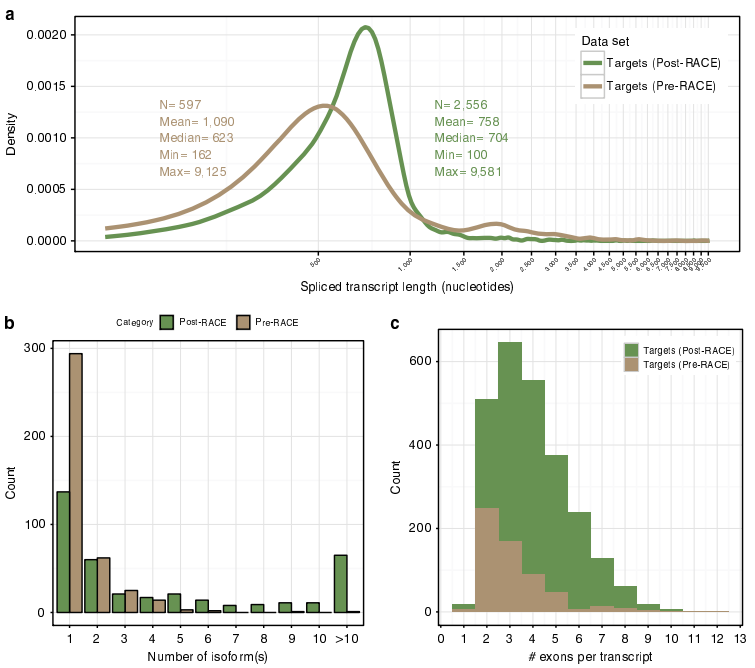


Figure 4 | New isoform discovery and annotation. (**a**) Length distribution of spliced transcripts (logarithmic scale) for pre- (brown) and post-RACE-Seq (green) targets. (**b**) Distribution of the number of alternatively spliced isoforms per pre- (brown) and post-RACE-Seq (green) targeted gene locus. (**c**) Exon count distribution in pre- (brown) and post-RACE-Seq (green) transcripts.

Table 4 | Proportion of annotated splice junctions in pre- and post-RACE-Seq targets supported by short-read ENCODE or GTEx RNA-Seq data.

Data set	Total #unique splice junctions	#supported by ENCODE or GTEx RNA-Seq	%supported by ENCODE or GTEx RNA-Seq
Targets (pre-RACE)	1,093	771	71%
Targets updated (post-RACE)	3,085*	975	31%
Protein coding	82,627	74,090	90%
lncRNAs	24,133	16,937	67%

lncRNA, long non-coding RNA; RACE, rapid amplification of cDNA ends.
Both data sets are derived from conventional, unbiased sequencing experiments. (* represents novel introns only).

catalogue. Figure 5a shows a common occurrence in the annotation where two separate lncRNA loci have been extended to produce one larger new locus (*LINC01246*) with over 50 new spliced transcripts.

The majority, 63% ($N = 1{,}618$), of the 2,556 RACE-Seq derived transcripts were from testis and 20% ($N = 516$) from brain (Supplementary Fig. 8). The rest of the tissues (heart, kidney, liver, lung and spleen) contributed ~13% of novel transcripts. Many genes that appeared to be extensively alternatively spliced (>25 transcripts) such as *TEX1*, *LINC0069* and *LAMTOR5-AS*, are detected in all 7 tissues examined.

In total, 4,626 novel exons were discovered, bringing the average number of exons per transcript from 3.2 pre-RACE-Seq to 4.0 post-RACE-Seq. Derrien *et al.*[8] made the striking observation that lncRNAs have a very strong bias towards two-exon structures and exhibit less alternatively spliced isoforms per locus compared with protein-coding genes. Our results suggest that these are artifacts arising from inaccurate annotation of lncRNA transcript structures, since the biases towards both two-exon transcripts and isoform-poor genes disappear in the post-RACE-Seq transcripts (Fig. 4b,c).

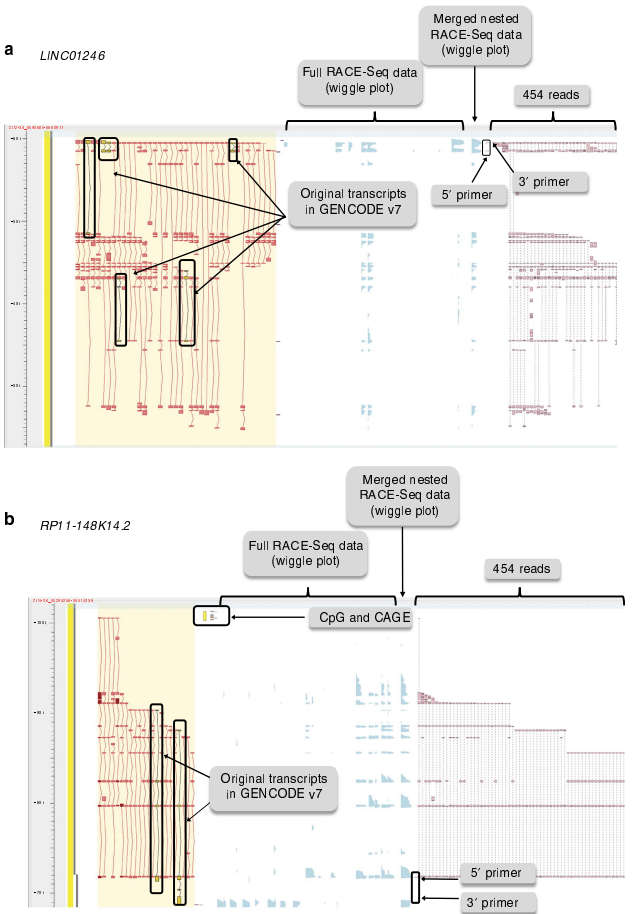


Figure 5 | Locus examples. (**a**) Two separate loci were merged into one larger locus (LINC01246). This example illustrates the large number of alternative splicing events found using the RACE-Seq approach. The red filled transcripts (far left) indicate the manual annotation models built from the 454 reads (far right in pink), as visualized in the ZMap browser (http://www.sanger.ac.uk/science/tools/zmap). (**b**) RACE-Seq reads (far right, in pink) establishes the Transcriptional start site (TSS) of an existing incomplete lincRNA, by extending the 5′ end of the gene to a CpG island (yellow box) and is also supported by FANTOM5 CAGE data (small pink boxes).

Comparison with other transcriptome sequencing methods. To further evaluate RACE-Seq, we compared its performance with other transcriptome sequencing methods. First, we used non-targeted, conventional RNA sequencing data generated by the GTEx[11] and ENCODE[6] consortia. We analysed the GTEx pilot data freeze, which consists of RNA-seq data collected from 1,641 samples from 175 human individuals, representing up to 43 tissues per individual (29 solid organ tissues, 11 brain regions, whole blood and 2 cell lines). GTEx RNA-Seq samples were sequenced to an average 80 million of pair-end Illumina reads (2×76 bp) per sample. The ENCODE data set is smaller (55 human cell lines and 104 samples), but on the other hand much more deeply sequenced (200–250 million pair-end reads (2×100 or 2×76 per sample)). We found that 71% of pre-RACE-Seq splice junctions from the targeted loci were supported by short-read Illumina ENCODE or GTEx RNA-Seq data. This proportion dropped to 31% when looking only into novel splice junctions found in transcripts discovered through RACE-Seq (Table 4). This result strongly suggests that the coverage of weakly expressed novel transcripts in non-targeted, conventional RNA-Seq experiments is shallow. It is also reflected by the proportion of overall support of splice junctions for

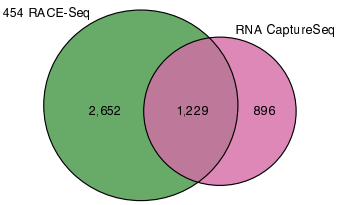


Figure 6 | RACE-Seq performance compared with CaptureSeq. Venn diagram indicating the number of annotated and unannotated on-target (± 5 kb) splice junctions discovered by RNA CaptureSeq and RACE-seq. Only the top 25% splice junctions with canonical splice sites ranked by read coverage were included in this analysis (see Methods).

protein-coding and lncRNA genes from GENCODE v7 by ENCODE or GTEx RNA-Seq data. Almost all splice junctions from annotated protein-coding loci show short-read support, while it is the case for <70% of lncRNAs.

To fully assess the performance of RACE-Seq, we compared our results with another targeted RNA sequencing method, capture sequencing (RNA CaptureSeq). RNA CaptureSeq enhances coverage of weakly expressed transcripts by focusing sequencing on genes of interest, thus enabling deeper sampling of low-abundance isoforms[19,20]. We analysed a subset of seven matching tissues from RNA CaptureSeq data set generated for lncRNAs profiling across 20 different human tissues by Clark *et al.*[20]. The support rate for both pre- and post-RACE transcripts is much higher compared with conventional RNA-Seq experiments, with 83% and 60% of splice junctions supported by RNA CaptureSeq data, respectively (Supplementary Table 4). To investigate whether RACE-Seq provides deeper interrogation of transcriptional events, we compared the set of splice junctions produced by each method within boundaries (± 5 kb) of 366 loci targeted by both studies. To compensate for differences in sequencing depth, we considered only the top quartile of canonical splice junctions, as ranked by read coverage, in each data set. RNA CaptureSeq enabled detection of 2,125 splice junctions, while RACE-Seq of 3,881 (83% more). Moreover, 1,229 splice junctions were supported by both methods, which constituted 60% of the total number of splice junctions seen in RNA CaptureSeq and roughly 30% from RACE-Seq (Fig. 6). Both techniques produce splice junctions uniformly distributed across targeted loci, including 5′ and 3′ ends (Supplementary Fig. 9). It is important to stress that the isoform discovery rate of both methods is expected to be negatively correlated with the number of targeted genes (16,453 in the study by Clark *et al.*[20], 398 in the present one), owing to the limited sequencing depth they rely on. These differences are not fully accounted for in our analysis, and may therefore favour our method over CaptureSeq in this comparison.

In addition, we used the read data by Clark *et al.*[20] from equivalent tissues to build CaptureSeq Cufflinks[27] transcript models overlapping our target genes, and mapped their corresponding 5′ ends. We derived a total of 343 non-redundant TSSs from this set, of which only 70 (20%, including 37 supported by FANTOM5 CAGE data) were previously unknown according to GENCODE (Supplementary Table 1). When compared with the output of RACE-Seq (873 TSSs, including 615 novel ones, see section above), this highlights the superiority of this latter technique at uncovering novel TSSs in comparison with CaptureSeq.

Discussion

Increased resolution in available technologies to monitor cellular transcriptomes have recently unveiled a plethora of RNA species beyond mRNAs. Among them, some lncRNAs have been shown to play important roles in cell function[28,29]. LncRNAs have characteristic tissue specificity and low-expression levels, which makes them challenging to annotate. While mRNAs, as well as some small RNA families, exhibit sequence and/or structural constraints that can be employed by computational methods to facilitate their identification and annotation, such constraints are mostly absent among lncRNAs[30]. There is indeed strong evidence that the exonic structure and the transcript termini of lncRNAs are not as well-annotated as those of protein-coding genes. For instance, only 15% of them have ENCODE CAGE data support at their 5′ end compared with 55% of protein-coding loci, according to a 2012 study[8].

Here we introduced the RACE-Seq methodology and used it to enhance lncRNA annotation. The idea of combining RACE with high-throughput sequencing was previously described by Olivarius *et al.*[31]. However, this study presented only 5′ RACE analysis of 17 protein-coding genes and compared single short-read Illumina sequencing with Sanger sequencing, and thus did not fully explore the high-throughput potential of this approach. In contrast, we tested the approach on a set of almost 400 human lncRNAs in 7 tissues, with both 5′ and 3′ RACE, and uncovered many previously unannotated transcripts. We increased the number of transcripts per lncRNA locus from 1.5 to 7.5 (see Table 3), and extended the 5′ and/or 3′ boundaries of the loci in 60% of the cases. The CAGE coverage of TSSs within the targeted genes increased by 28% at the end of the experiment—from 180 to (180 + 50 =) 230 CAGE-supported loci (Supplementary Fig. 3). Particularly useful was the usage of nested RACE-Seq, which lead to a 2.2-fold reduction in the number of detected off-target loci (Supplementary Fig. 5) compared with standard RACE-Seq.

While RACE-Seq leads to the identification of many novel transcripts, still only about 50% of the transcripts are on average full-length in a given locus. This could be improved by replacing the 454 technology, which has an average read length of 600 bp, with a longer-read sequencing technology, such as PacBio or Nanopore[32]. The sensitivity of RACE-Seq coupled with longer reads will facilitate automatic assembly of individual transcripts, which has proved problematic and inaccurate when using shorter reads[18], and it will lead to improved annotations. Indeed, we used the very large collection of short-read RNA-Seq samples from multiple tissues compiled by the GTEx project[11], and found that only 31% of the targeted lncRNA splice junctions could be detected in this data set. This highlights that conventional, unbiased short-read RNA-Seq suffers from a limited sampling capacity given the large dynamic range of transcript abundances within the cell.

To alleviate, in part, the poor sensitivity of unbiased methods, strategies that target specific genomic regions have already been developed. Notably CaptureSeq[19,20] uses oligonucleotide capture to perform short RNA-Seq in RNA populations enriched for selected loci. Still, we found that 40% of the RACE-Seq splice junctions are not observed in the output of CaptureSeq, and that although only 7 tissues were used in RACE-Seq this resulted in the discovery of a larger number of transcripts (6.6 on average per locus) than CaptureSeq (3.6 transcripts per locus on average) even when 20 tissues were employed. This highlights the importance of using longer reads and shows that RACE-Seq is very efficient to target specific gene classes, such as lncRNAs to uncover deep transcriptional complexity. Moreover, RACE-Seq has the advantage over CaptureSeq that it solves much more

accurately, and with more sensitivity the 5′ and 3′ end of transcripts.

The vast transcriptional complexity uncovered through RACE-Seq could reflect functional non-coding RNAs[33], or alternatively may be a result of experimental artifacts, and transcriptional noise[34]. The fact that many of the extensively alternatively spliced (>25 transcripts) loci such as *TEX1*, *LINC0069* and *LAMTOR5-AS* show expression in all 7 tissues examined indicates that this complexity is not due to experimental artifacts, and could instead be an indication of the vast functional potential of non-coding RNAs[33]. However, the amount of biological 'noise' that exists within the transcriptome remains a source of much debate[34,35], and other methods will be required to rigorously establish the functionality of these transcripts.

Recently, Tilgner *et al.*[36] used a new sequencing method, Synthetic long-read RNA sequencing (SLR-RNA-Seq) in which small pools of full-length cDNAs are fragmented and sequenced using small-read-sequencing, and then re-assembled. Since Tilgner *et al.*[36] also examined the transcriptome of human brain tissue, we examined the data to investigate if any of the targeted lncRNA loci were detected by SLR-RNA-Seq. Around 37% ($N=149$) of our targeted lncRNA genes were covered by reads, however, only 9% or RACE-Seq splice junctions were detected by SLR-RNA-Seq reads (see Supplementary Fig. 10 and Supplementary Table 5). This alternative method of deep short-read sequencing, combined with targeted, nested RACE-Seq, could potentially provide a cheaper alternative to more expensive longer-read sequencing.

Methods

Target selection and primer design. The experiment was designed using a fully automatized pipeline, which contents are available on request. Illumina HBM (Human Body Map 2.0) RNA-Seq data was used and RPKMs (reads per kilobase of exon per million mapped reads) for all GENCODE v7 lncRNAs were calculated. We selected lncRNAs that were expressed in at least one HBM experiment with an RPKM >5 and that were lacking CAGE/PET support in ENCODE cell line experiments[6,17]. The spliced RNA sequences for the top 398 lncRNAs, ranked by mean RPKM across cell lines, were extracted and used as input for primer design. At the time of the experimental design, CAGE data on matched tissues were not available, therefore we had to rely on ENCODE CAGE experiments, performed on various cell lines, all quite distinct from our set of tissues. On the public release of matched tissue CAGE data from the FANTOM5 consortium[3], we re-calculated CAGE support of the 398 RACE-Seq-targeted loci. We found that, in fact, 180 of them had at least 1 CAGE tag in their vicinity (±50 nucleotides, on the same strand) in at least 1 of the matched FANTOM5 tissues.

Non-specific regions within the candidate sequences were masked to avoid off-target RACE products. These regions were established by aligning candidate sequences against all GENCODE v7 transcript sequences using the BLAST program[37]. Regions having >80% sequence similarity to any GENCODE v7 transcript from a distinct locus were hard-masked. Only stranded overlap was considered. We then generated for each candidate transcript, all possible 5′ and 3′ RACE primers using primer3 with the following parameters: PRIMER_INTERNAL_OPT_SIZE = 25, PRIMER_INTERNAL_MIN_SIZE = 23, PRIMER_INTERNAL_MAX_SIZE = 27, PRIMER_INTERNAL_OPT_TM = 70.0, PRIMER_INTERNAL_MIN_TM = 68.0, PRIMER_INTERNAL_MAX_TM = 72.0, PRIMER_INTERNAL_MIN_GC = 50, PRIMER_INTERNAL_MAX_GC = 70, PRIMER_INTERNAL_OPT_GC_PERCENT = 60.

In total, we could design 3′ and 5′ RACE for all 398 targets in standard RACE. 361 and 367 nested primers could be designed for 3′/5′ RACE, respectively.

The full list of RACE primer sequences, together with their corresponding transcript targets and mean RPKM, is provided as a tab-separated file in the Supplementary Data section.

RACE reactions. Nested and non-nested 5′ and 3′ RACE products were obtained using the Clontech SMART RACE cDNA Amplification kit and the Advantage 2 Proofreading Polymerase PCR kit (Clontech Laboratories, Mountain View, CA, USA, catalogue no. 634914) according to the manufacturer's instructions. PolyA + RNA from a panel of seven human tissues was used (brain, heart, kidney, liver, lung, spleen and testis), all from Clontech Laboratories. RACE- and nested RACE-specific primers were synthesized by Life Technologies Europe BV and were diluted to a final concentration of 200 nM. Each RACE reaction was performed in an independent well on a 384 well plate, and PCRs were done using liquid-handling robots.

Double-stranded cDNA synthesis, adaptor ligations to the synthesized cDNA and RACE reactions were performed according to the manufacturers' instructions. Nested RACEs were performed with 0.5 µl of the initial RACEs in a final volume of 12.5 µl. The cycling parameters were: RACE 5 × (94 °C 30″, 70 °C 30″, 72 °C 3′), 5 × (94 °C 30″, 68 °C 30″, 72 °C 3′), 20 × (94 °C 30″, 66 °C 30″, 72 °C 3′); nested RACE 25 × (94 °C 30″, 68 °C 30″, 72 °C 3′). We then pooled by tissue, 2 µl of all nested RACE reactions, and pools were purified using Qiaquick PCR purification kit (Qiagen, CA, USA) before proceeding with 454+ library preparation.

GS-FLX 454 + library preparation and sequencing. cDNA RACE samples were analysed on a DNA 7,500 Chip (2,100 Bioanalyzer, Agilent Technologies Inc, Santa Clara, CA, USA) to assess fragment size and sample integrity. For samples with a mean fragment size of ≥2 Kbp, 1 µg of material is subjected to nebulization and then used to prepare a rapid ligation (RL) genomic shotgun library using the Rapid Library Preparation Method Manual (GS FLX+ Series-XL+, May 2011, Roche 454 Life Sciences). For RACE samples with a mean fragment size smaller than 2 Kbp the nebulization step is avoided, starting library preparation directly with 800 ng of material. A modification was introduced in the small fragment removal of this library preparation, to allow only for the removal of fragments under 400 bp instead of fragments under 650 bp. Then, the quality of the RL libraries was assessed by running an aliquot of the library in a High Sensitivity Chip (2,100 Bioanalyzer). Library quantification was performed generating a RL standard curve and using a 96-well Plate fluorometer, according with the manufacturers instructions, and using the Rapid Library Quantitation Calculator (www.454.com/my454).

Samples were titrated using the emPCR amplification Method Manual Lib-L SV (GS FLX+ Series-XL+, May 2011) to know the optimal point of copies per bead (cpb) needed to obtain a 10% enriched beads. Then, a large volume emulsion PCR was performed using the emPCR amplification Method Manual Lib-L LV (GS FLX+ Series-XL+, May 2011).

Sequencing was performed at the Genomic and Bioinformatics Platform of Andalusia (GBPA) using half 454-pyrosequencing plate per sample using a Roche 454 GS FLX+ instrument, and GS FLX+ reagents (Roche 454 Life Sciences). After the sequencing was finished, sequencing images were analysed using the Shotgun-pipeline to generate SFF files.

Read pre-processing and mapping. FASTQ files were extracted from SFF files by the program sffextract (http://bioinf.comav.upv.es/sff_extract/index.html). Cutadapt was used to remove adaptors, and reads shorter than 100 nts were filtered out (5′ RACE Adapter: 5′-CTAATACGACTCACTATAGGGCAAGCAGTGGTATCAACGCAGAGTACT-3′, 3′ RACE Adapter: 5′-CTAATACGACTCACTATAGGGCAAGCAGTGGTATCAACGCAGAGTACGCGGG)-3′.

Low quality nucleotides in 3′ end were hard-trimmed by calculating the mean quality of the last three nucleotides and removing bases progressively until reaching a mean quality >20 (Sanger scale).

Two different approaches where used to map the reads to the reference genome:

a) The Inchworm wrapper program[38] was used to run *BLAT*[23] (v35) and generate SAM files, by setting minimum per cent identity at 95% and considering just the best single hit per read reported by BLAT. Intron prediction by Inchworm is based on the presence of splice consensus sites in the ends of the gaps (gap size> 20).

b) Using the GMAP program[24] (version 31 March 2013 with all parameters set to default except `--min-identity=0.95 --force-xs-dir -B 5 -t 5 -f samse $file --min-intronlength=30 --split-output`). Only unique mappings were considered in subsequent analyses.

Both BLAT and GMAP mappings were performed against the hg19 (GRCh37) assembly of the human genome.

Coverage and on-target enrichment calculation. Mapped reads were compared with annotated regions using the BEDtools suite[39] v2.17.0. Reads were considered on-target when they overlapped exonic regions of the targeted transcripts. We estimated the expected read coverage of transcripts in a typical non-targeted RNA-Seq experiment using GTEx[11] data in matched tissues (SRA accessions: SRR1403958, SRR1340617, SRR1314940, SRR1080294, SRR809807, SRR1069539 and SRR1458955). We then calculated the on-target enrichment achieved by RACE-Seq using the following formula:

$$\text{Enrichment} = R/E$$

Where

R = proportion (across tissues) of mapped RACE-Seq reads on-target

E = proportion (across tissues) of mapped GTEx RNA-Seq reads on-target. The results of the comparison between GTEx read coverage and both standard and nested RACE-Seq on the 398 targets are summarized in Supplementary Table 2.

Transcripts manual annotation. Manual annotation was performed according to GENCODE standards[7]. Briefly, imported BAM files (merged outputs of GMAP and BLAT) representing the aligned RNA-seq reads were displayed in our in-house annotation tool, ZMAP (http://www.sanger.ac.uk/science/tools/zmap). Splice sites and alignments for all non-redundant novel intron combinations and exon extensions were evaluated manually and, when confirmed, used to create new or

extend existing transcript models. While the target loci were lncRNAs, where warranted by the RNA-seq data, biotypes were modified (for example, from non-coding to coding if RNA-seq read joins lncRNA target to a coding gene).

Characterization of novel transcript boundaries. In this part of the analysis, we split all post-RACE-Seq transcript models into 5′ and 3′ RACE products. We reasoned that 5′ RACE product models may be anchored at their originating primer location at the 3′ end, hence obfuscating the global analysis of genuine transcript 3′ ends (and likewise for 3′ RACE products versus 5′ ends). We did so by assigning the most probable originating RACE direction (5′ or 3′) to each post-RACE GENCODE transcript model: a transcript extending an annotated locus further in 5′, and/or whose 3′ end started within 50 bps of a 5′ RACE primer (and on the opposite strand) was labelled a 5′ RACE product. Similarly, one extending an annotated locus further in 3′, and/or whose 5′ end started within 50 bps of a 3′ RACE primer (and on the same strand) was labelled a 3′ RACE product.

This resulted in 1,427 and 1,420 transcript objects likely produced by 5′ RACE and 3′ RACE that were used for TSS and TTS analysis, respectively. About 291 transcripts were shared between those 2 sets (that is, likely complete structures from 5′ to 3′). Overall, we could assign a probable RACE direction to 2,556 transcripts out of 2,641, that is, 97%.

Transcription start sites. All annotated 5′ RACE-Seq TSSs were clustered (that is, all TSSs on the same strand and <51 bps away were merged into one). About 615 out of these 873 clustered TSSs were considered novel, that is, they lied farther than 100 bps from any of the targeted GENCODE 7 transcript's TSSs. We compared TSSs against merged, tissue-matched CAGE data from the FANTOM5 consortium, and considered them CAGE-supported if a CAGE tag could be found on the same strand, within 50 bps on either side. Of the 615 novel TSSs, 252 (41%) were found to be CAGE supported.

TSSs uncovered with RACE-Seq were compared with the TSSs of their originating locus. Any RACE-Seq TSS upstream of the 5′-most TSS of its original, GENCODE 7-annotated locus, was labelled as 'extending', and the corresponding locus as '5′-extended'.

Transcription termination sites. All annotated 3′ RACE-Seq TTSs were clustered (that is, all TTSs on the same strand and <151 bps away were merged into one). The clustering distance was chosen longer than for TSSs because of the 'leakier' nature of transcription termination compared with transcription initiation. We compared TTSs against merged data from available matched tissues (brain, kidney, liver, muscle, testis) from the Merck PolyA-Seq set[25], as downloaded from the UCSC genome browser. We considered a TTS polyA-Seq-supported if a PolyA-Seq tag could be found on the same strand, within 100 bps on either side.

We also inferred polyA sites from RACE-Seq data. To do so, we selected mapped 3′ RACE reads and searched for characteristic non-templated stretches of >20 Ts or As (allowing for 10% mismatches) at their ends. About 1,212 non-redundant polyA sites were mapped, and compared with RACE-Seq TTSs in the same manner as for Merck PolyA-Seq sites.

TTSs uncovered with RACE-Seq were compared with the TTSs of their originating locus. Any RACE-Seq TTS downstream of the 3′-most TTS of its original, GENCODE 7-annotated locus, was labelled as 'extending', and the corresponding locus as '3′-extended'.

Conventional unbiased short-read RNA-seq. Integrative Pipeline for Splicing Analyses (IPSA, unpublished, https://github.com/pervouchine/ipsa) was employed to locate splice junctions from 1,641 GTEx, 104 ENCODE and 38 454-RACE-Seq bam files, respectively. Alignments for GTEx and ENCODE data sets were produced by each consortium using their respective official processing pipelines. IPSA was run with the default parameters except `–entropy 3`. The analysis of splice junction support was done by comparing the two lists of splice junctions: one consisting of GTEx or ENCODE splice junctions produced by IPSA, and a second containing splice junctions derived from GTF files of pre- or post-RACE transcripts.

RNA CaptureSeq. RNA CaptureSeq FASTQ files for seven tissues matched with RACE-seq experiment were downloaded from BioProject (PRJNA261251). The sequences were aligned to the reference human genome (GRCh37/hg19) using STAR[40] v 2.4.0, according to the instructions specified by Clark *et al.*[20]. Both standard and nested 454-RACE-seq sequences were also re-mapped using STAR v 2.4.0. The following non-default STAR parameters were applied: `--outSAMunmapped Within --alignSJDBoverhangMin 1 --outFilterType BySJout`. Again, IPSA (with the same parameters as those mentioned above) was run to produce the lists of annotated and unannotated splice junctions for both data sets. Support rate of splice junctions annotated by GENCODE for pre- and post-RACE loci (only those targeted by both studies) by RNA CaptureSeq and 454-RACE-Seq was investigated by intersecting those with the set of splice junctions detected by IPSA for each data set. Next, annotation-free splice junction analysis was performed to further compare RNA CaptureSeq and 454-RACE-Seq. This analysis was done by selecting annotated and unannotated splice junctions with canonical splice sites, located within targeted loci boundaries (± 5kb) from the IPSA output. Next, splice junctions were ranked by read coverage and only those supported by top quartile read count were further analysed.

We assessed the TSSs discovered by CaptureSeq by re-building Cufflinks[27] transcript models from Clark *et al.*'s[20] provided BAM files on brain, heart, kidney, liver, lung, spleen and testis (GEO accession GSE61474). Reads falling on and in the vicinity (± 500 bps) of our GENCODE 7 targeted genes were selected and fed to Cufflinks v.2.2.1 with all options set to default except '`--library-type fr-firststrand`', '`-u`' and using GENCODE 7 as a guide ('`-g`'). We derived TSSs from Cufflinks' output after selecting those 1,156 transcript models that overlapped RACE-Seq-targeted exons, had the 'full_read_support' GFF attribute set to 'yes' and an FPKM value >0 in any of the 7 tissues. The resulting 343 non-redundant TSSs were then processed the same way as RACE-Seqs (see section above and results in Supplementary Table 1).

Synthetic long-read sequencing. The sequences for 11 human brain samples produced using SLR-seq were downloaded from Sequence Read Archive (SRP049776) and aligned to the reference genome (GRCh37) using GMAP and parameters specified by Tilgner *et al.*[36]. An in-house developed pipeline was applied to produce the list of splice junctions from genomic alignments.

Data availability. All computer code is available from the authors upon request. Sequence data have been deposited in the European Nucleotide Archive (ENA) under accession number ERP012249. All curated novel isoforms were incorporated into the human GENCODE set (version 22 onwards). In addition, a data portal, including a UCSC track hub, is available at http://public-docs.crg.es/rguigo/Papers/2016_lagarde-uszczynska_RACE-Seq/.

References

1. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154,** 26–46 (2013).
2. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long non-coding RNAs: past, present, and future. *Genetics* **193,** 651–669 (2013).
3. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507,** 462–470 (2014).
4. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).
5. Stamatoyannopoulos, J. A. *et al.* An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.* **13,** 418 (2012).
6. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
7. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
8. Derrien, T. *et al.* The GENCODE v7 catalogue of human long non-coding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22,** 1775–1789 (2012).
9. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
10. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).
11. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
12. Harrow, J. *et al.* Identifying protein-coding genes in genomic sequences. *Genome Biol.* **10,** 201 (2009).
13. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13,** 329–342 (2012).
14. Adams, M. D., Soares, M. B., Kerlavage, A. R., Fields, C. & Venter, J. C. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat. Genet.* **4,** 373–380 (1993).
15. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).
16. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).
17. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
18. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10,** 1177–1184 (2013).
19. Mercer, T. R. *et al.* Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* **9,** 989–1009 (2014).
20. Clark, M. B. *et al.* Quantitative gene profiling of long non-coding RNAs with targeted RNA sequencing. *Nat. Methods* **12,** 339–342 (2015).
21. Yeku, O. & Frohman, M. A. Rapid amplification of cDNA ends (RACE). *Methods Mol. Biol.* **703,** 107–122 (2011).
22. Howald, C. *et al.* Combining RT-PCR-seq and RNA-seq to catalogue all genic elements encoded in the human genome. *Genome Res.* **22,** 1698–1710 (2012).
23. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).

24. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21,** 1859–1875 (2005).
25. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22,** 1173–1183 (2012).
26. Zhou, X., You, F., Chen, H. & Jiang, Z. Poly(C)-binding protein 1 (PCBP1) mediates housekeeping degradation of mitochondrial antiviral signaling (MAVS). *Cell Res.* **22,** 717–727 (2012).
27. 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).
28. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309,** 1559–1563 (2005).
29. Mattick, J. S. & Rinn, J. L. Discovery and annotation of long non-coding RNAs. *Nat. Struct. Mol. Biol.* **22,** 5–7 (2015).
30. Gorodkin, J. & Hofacker, I. L. From structure prediction to genomic screens for novel non-coding RNAs. *PLoS Comput. Biol.* **7,** e1002100 (2011).
31. Olivarius, S., Plessy, C. & Carninci, P. High-throughput verification of transcriptional starting sites by Deep-RACE. *Biotechniques* **46,** 130–132 (2009).
32. Bolisetty, M. T., Rajadinakaran, G. & Graveley, B. R. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol.* **16,** 204 (2015).
33. Clark, M. B. *et al.* The reality of pervasive transcription. *PLoS Biol.* **9,** e1000625 (2011).
34. van Bakel, H., Nislow, C., Blencowe, B. J. & Hughes, T. R. Most 'dark matter' transcripts are associated with known genes. *PLoS Biol.* **8,** e1000371 (2010).
35. Mudge, J. M., Frankish, A. & Harrow, J. Functional transcriptomics in the post-ENCODE era. *Genome Res.* **23,** 1961–1973 (2013).
36. Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* **33,** 736–742 (2015).
37. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
38. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–652 (2011).
39. Quinlan, A. R. BEDTools: the swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47,** 11.12.1–11.12.34 (2014).
40. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

Acknowledgements

This work and publication were supported by the National Human Genome Research Institute of the National Institutes of Health (grant numbers U41HG007234, U41HG007000 and U54HG007004) and the Wellcome Trust (grant number WT098051). Work in laboratory of R.G. was supported by Awards Number U54HG0070, R01MH101814 and U41HG007234 from the National Human Genome Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We acknowledge support of the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013–2017', SEV-2012-0208 and grant BIO2011-26205. We thank members of the Guigó laboratory for their valuable input when analysing data and writing the manuscript, in particular Rory Johnson, Dmitri Pervouchine and Sarah Djebali; Romina Garrido (CRG) for administrative assistance, and Roche for providing library preparation and sequencing reagents for the nested RACE experiments.

Author contributions

R.G., J.H., J.D., J.S.-L., A.T., A.Re. and J.L. designed the experiment. A.Ru., F.J.L-D, A.T., J.D., J.S.-L., B.U.-R., J.M.G., E.T., J.M.M., C.S., L.W. and J.L. analysed the data. C.H. and J.C. performed the RACE amplification. A.V.-B. performed the 454 sequencing of the RACE products. J.L., B.U.-R., J.H. and R.G. wrote the manuscript.

Additional information

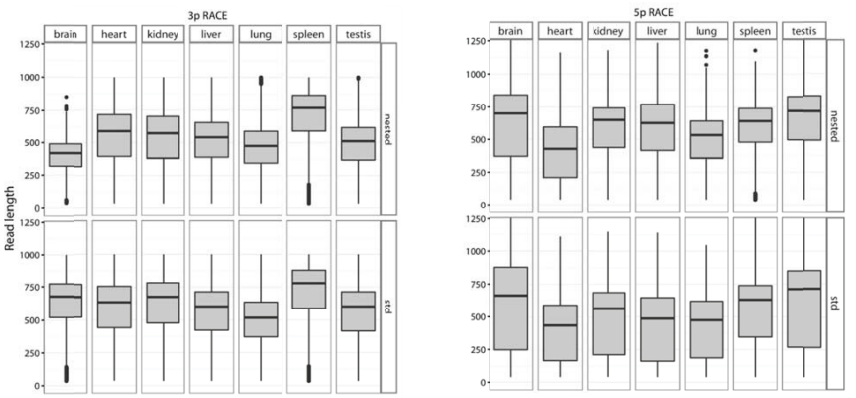
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

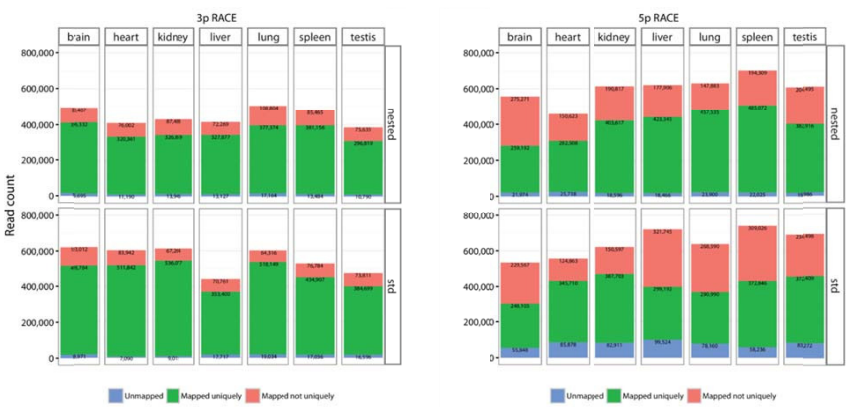
Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Lagarde, J. *et al.* Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq). *Nat. Commun.* 7:12339 doi: 10.1038/ncomms12339 (2016).

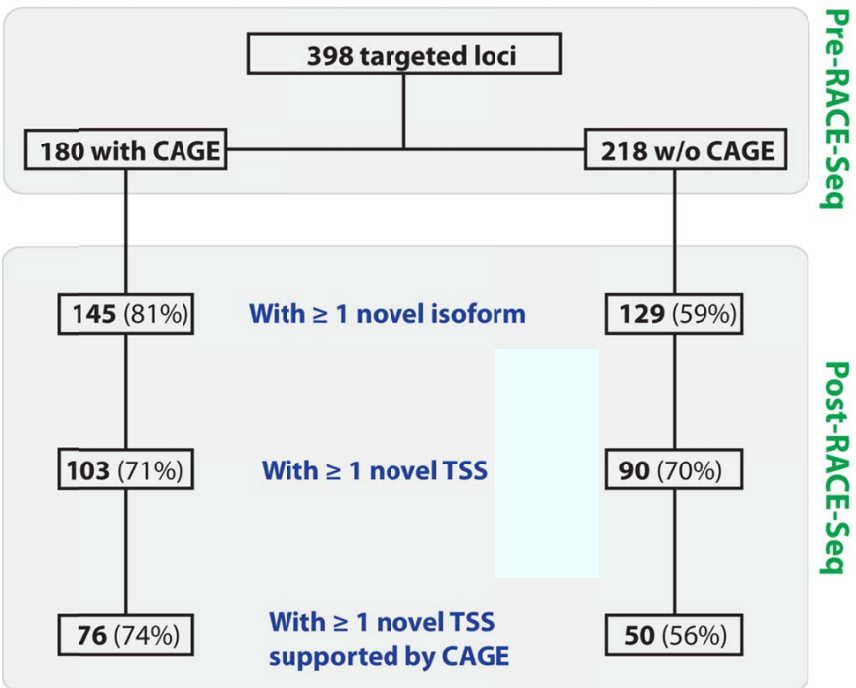
I.2. Supplementary information



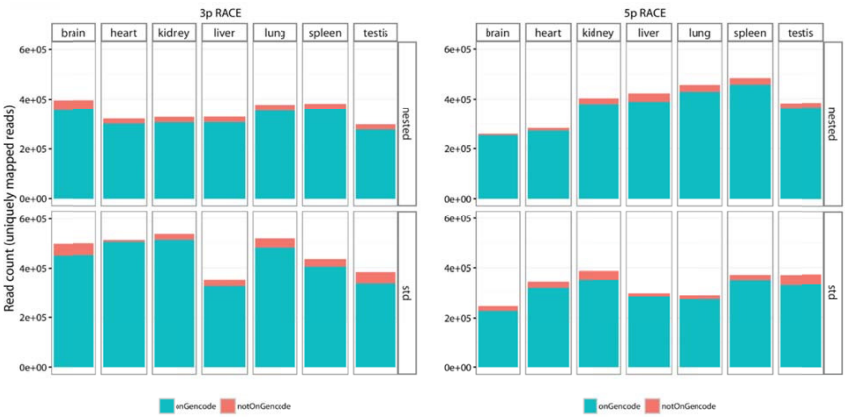
Supplementary Figure 1: Box-plots of read length distributions across seven tissues targeted by 5' and 3' standard ("std") and nested RACE-Seq.



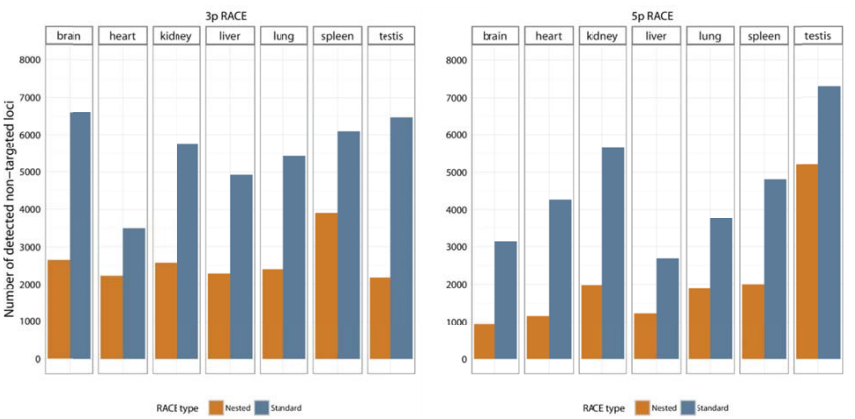
Supplementary Figure 2: Genome mapping statistics. Bar plot showing the number of 454 reads that were unmapped (blue) mapped uniquely (green) and multiple times (coral) to the reference human genome (GRCh37).



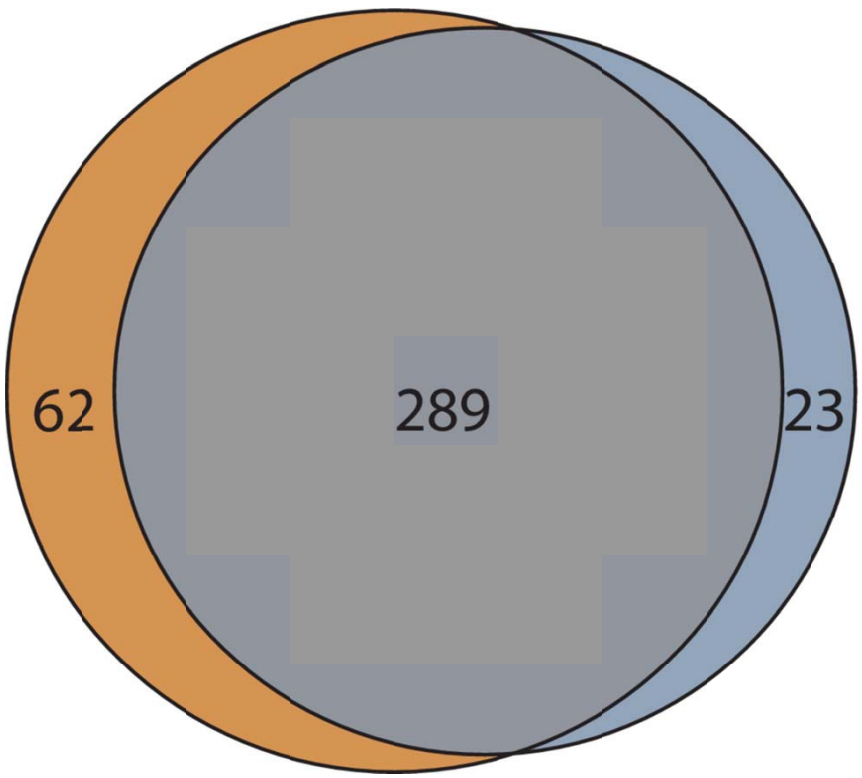
Supplementary Figure 3: Flowchart explaining the CAGE enrichment analysis and summarized results.



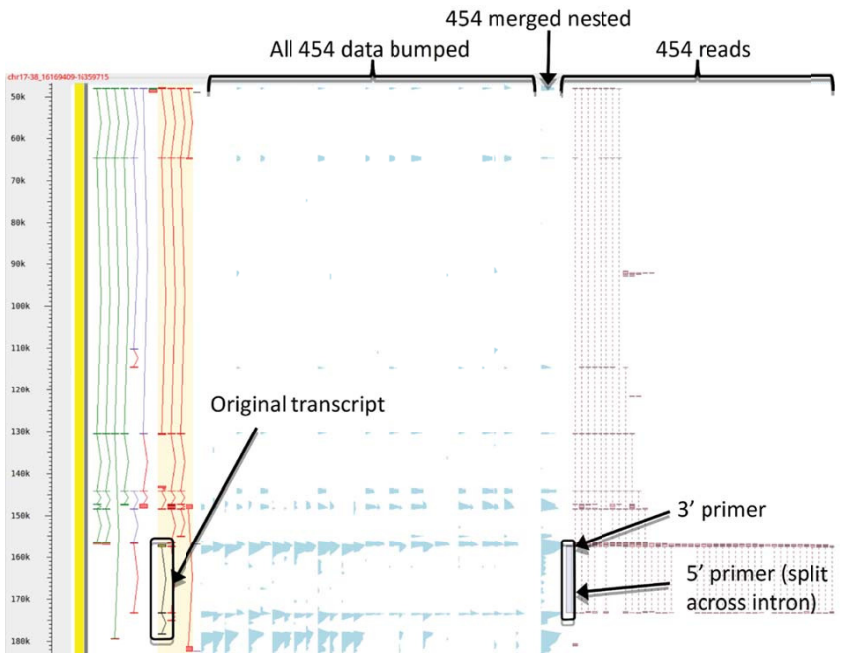
Supplementary Figure 4: Read mapping statistics. Number of uniquely mapped 454 reads overlapping GENCODE-annotated loci (green), vs. in intergenic regions (coral).



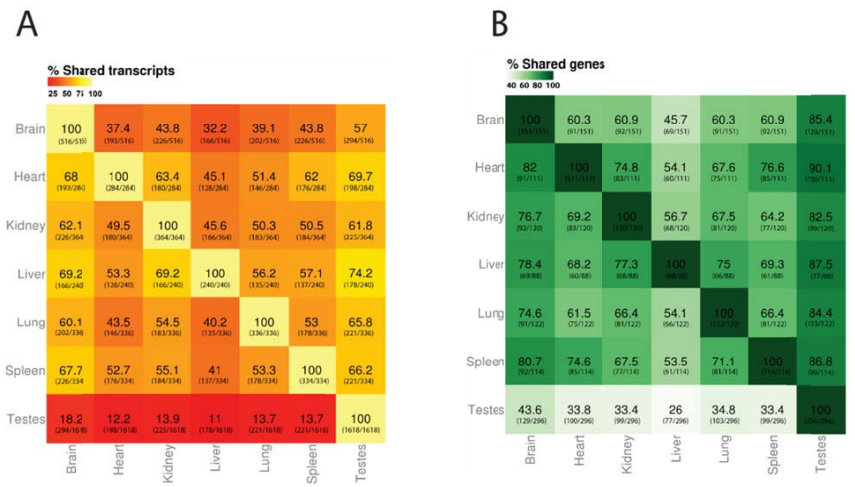
Supplementary Figure 5: Number of amplified non-targeted loci in nested and standard RACE-Seq in the seven tissues assayed.



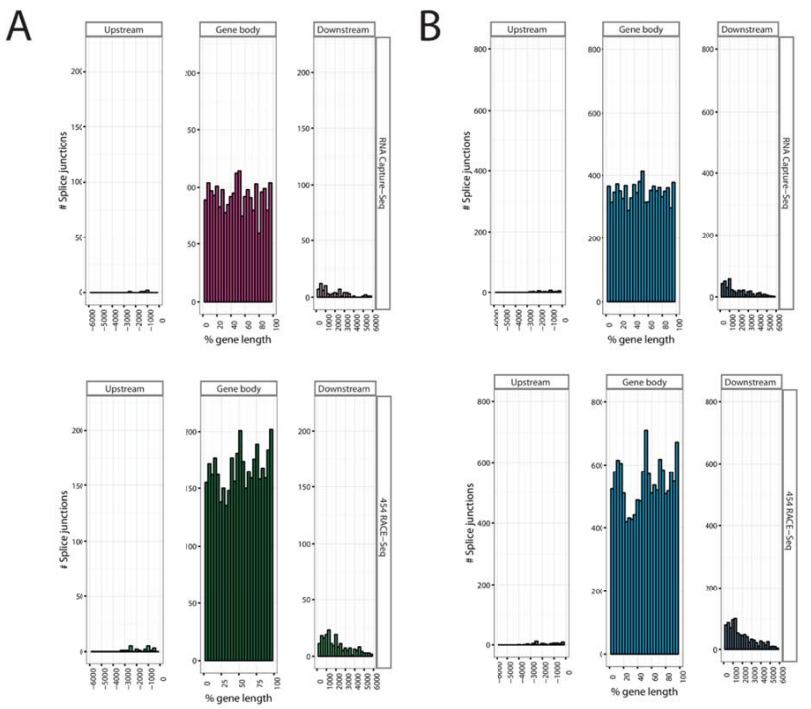
Supplementary Figure 6: Venn diagram comparing the sets of targeted loci that could be amplified in standard, primary RACE (blue) and nested RACE (orange)



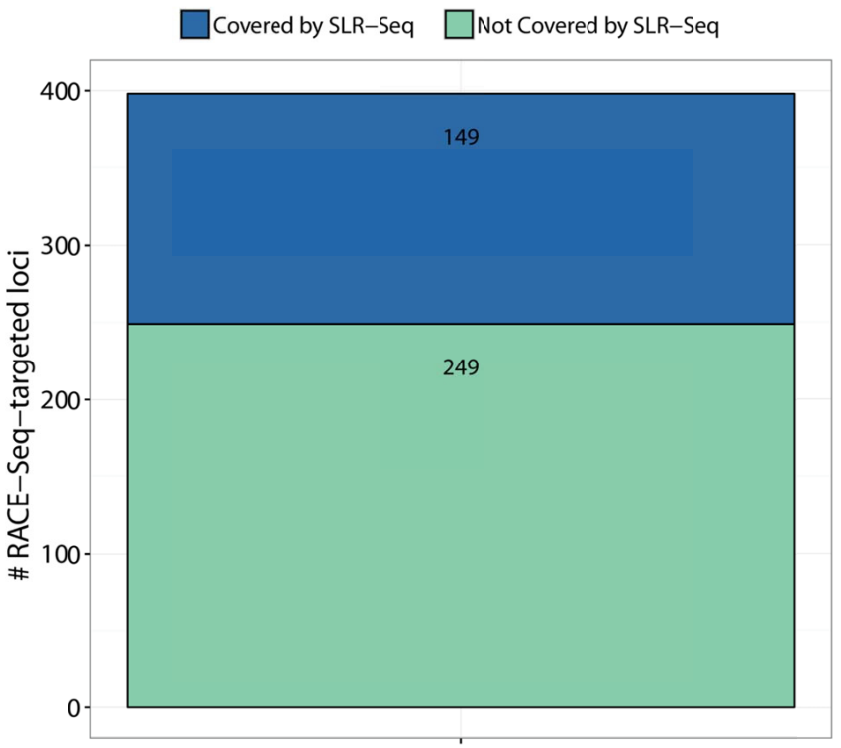
Supplementary Figure 7: Genome browser screenshot of the PIGL locus. The original locus (highlighted, bottom left) was subsumed into the coding locus PIGL. The RNA-seq signal plot (light blue, all 454 data bumped) shows the density of reads mapping to exons. The RACE



Supplementary Figure 8: Detection of lncRNAs in targeted tissues. Heat map showing the number of detected lncRNAs in each tissue and their proportion shared across other targeted tissues at the (A) transcript level (B) gene level. Transcripts were reconstructed directly from the read alignments, using transcript structure compatibility as merging criterion. A given locus was considered as expressed in a given tissue if at least one of its transcript was successfully reconstructed.



Supplementary Figure 9: Distribution of detected splice junctions by RACE-Seq (bottom panel) and CaptureSeq (top panel) along targeted genes. Bar plots showing the location of splice junctions within targeted loci boundaries (+/- 5kb). (a) All annotated and unannotated splice junctions within targeted loci boundaries (+/- 5kb). ***(A)*** *All annotated and unannotated splice junctions detected in RNA CaptureSeq and RACE-Seq data sets.* ***(B)*** *The top 25% of annotated and unannotated canonical splice junctions ranked by read coverage.*



Supplementary Figure 10: Detection of pre- and post-RACE-Seq targets by SLR-Seq. Number of loci covered (blue) and not covered (green) by SLR-Seq reads.

Dataset	Total # TSS (clustered)	# Clustered TSS +/- 50 bp from CAGE tag	% Clustered TSS +/- 50 bp from CAGE tag
Targets (pre-RACE)	527	241	46%
Targets (post-RACE, all)	873	415	48%
Targets (post-RACE, novel)	615	252	41%
CaptureSeq transcript models (all)	343	203	59%
CaptureSeq transcript models (novel)	70	37	53%

Supplementary Table 1: Table summarizing TSS discovery and CAGE coverage statistics in both RACE-Seq and Clark et al.*'s CaptureSeq.*

Tissue	Pre RACE-Seq (GTEx data)			Post RACE-Seq (standard)				Post RACE-Seq (nested)			
	Total # mapped reads	# reads within targeted transcripts	% reads within targeted transcripts	Total # mapped reads	# reads within targeted transcripts	% reads within targeted transcripts	On-target fold enrichment	Total # mapped reads	# reads within targeted transcripts	% reads within targeted transcripts	On-target fold enrichment
brain	37,162,929	330,959	0.9%	746,889	41,019	5.5%	6.2	655,524	328,666	50.1%	56.3
heart	35,835,990	564,311	1.6%	857,552	14,475	1.7%	1.1	602,869	171,668	28.5%	18.1
kidney	45,384,859	544,064	1.2%	923,780	28,104	3.0%	2.5	730,476	291,291	39.9%	33.3
liver	39,432,611	282,230	0.7%	652,592	18,345	2.8%	3.9	751,222	199,511	26.6%	37.1
lung	48,531,622	792,644	1.6%	809,139	18,894	2.3%	1.4	834,909	256,859	30.8%	18.8
spleen	33,051,498	475,424	1.4%	807,753	16,775	2.1%	1.4	866,228	272,386	31.4%	21.9
testis	43,888,261	545,528	1.2%	757,108	76,250	10.1%	8.1	679,735	345,709	50.9%	40.9
TOTAL	**283,287,770**	**3,535,160**	**1.2%**	**5,554,813**	**213,862**	**3.9%**	**3.1**	**5,120,963**	**1,866,090**	**36.4%**	**29.2**

Supplementary Table 2: On-target read enrichment statistics. 5' and 3' RACE datasets were merged in each tissue.

# Targets	RACE direction	RACE type	# Targets successfully RACE'd	% Targets successfully RACE'd
398	5'	Standard	248	62%
		Nested	314	79%
		Standard + Nested	341	86%
	3'	Standard	255	64%
		Nested	293	73%
		Standard + Nested	326	82%

Supplementary Table 3: Number and proportion of successfully RACE-amplified targets.

Dataset	Total # unique splice junctions	# Supported by RNA CaptureSeq	% Supported by RNA CaptureSeq
Targets (pre-RACE)	1,093	903	83%
Targets updated (post-RACE)	3,664	2,211	60%

Supplementary Table 4: Detection of pre- and post-RACE-Seq targets by RNA CaptureSeq. Proportion of annotated splice junctions in pre- and post-RACE-Seq targets supported by RNA CaptureSeq.

Dataset	Total # unique splice junctions	# Supported by 454-RACE-Seq	% Supported by 454-RACE-Seq	# Supported by SLR-Seq	% Supported by SLR-Seq
Targets (pre-RACE)	1,093	817	74.8%	226	20.68%
Targets updated (post-RACE)	3,664	3,277	89.4%	281	9.11%

Supplementary Table 5: Comparison of splice junction support by 454 RACE-seq and SLR-seq.

II

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

Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R, Johnson R. *High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing*. ***Nature Genetics*** 2017 Dec; 49(12):1731-1740.

URL: https://doi.org/10.1038/ng.3988

Abstract:

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.

II.1. Main article

ARTICLES

nature genetics

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

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

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[1,2], were joined by long intergenic noncoding RNA (linc RNA) sets discovered through chromatin signatures[3]. More recently, researchers have applied transcript-reconstruction software such as Cufflinks[4] to identify novel genes in short-read RNA-sequencing (RNA-seq) data sets[5–9]. 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[10,11] and have been adopted by international genomics consortia[12–15].

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[16]. This particularly affects lncRNAs, whose low expression results in low read coverage[11]. The outcome is a growing divergence between large automated annotations of uncertain quality (e.g., 101,700 genes for NONCODE[8]) and the highly curated, 'conservative' GENCODE collection[11] (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[17], enhancers[18] and ultraconserved elements[19,20]. Second, annotated lncRNAs may represent partial gene structures. Start and end sites frequently lack independent supporting evidence[11], and lncRNAs are shorter and have fewer exons than mRNAs[7,11,21]. 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[21].

One of the principal impediments to the annotation of lncRNAs is their low steady-state levels[3,11]. To overcome this, RNA capture sequencing (CaptureSeq)[22] 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[22–24]. Thus, although CaptureSeq achieves high throughput, its transcript structures lack the confidence required for inclusion in GENCODE.

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [3]R&D Department, Quantitative Genomic Medicine Laboratories (qGenomics), Barcelona, Spain. [4]Functional Genomics Group, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [5]Wellcome Trust Sanger Institute, Hinxton, UK. [6]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.). [7]These authors contributed equally to this work. Correspondence should be addressed to R.J. (rory.johnson@dbmr.unibe.ch) or R.G. (roderic.guigo@crg.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'[10,25]. To these we added tiled probes targeting loci that may produce lncRNAs: small RNA genes[26], enhancers[27] and ultraconserved elements[28]. For mouse tissues we also added orthologous lncRNA predictions from PipeR[29]. We added numerous control probes, including a series that targeted half of the External RNA Controls Consortium (ERCC) synthetic spike-ins[30]. 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[32]. 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[33].

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

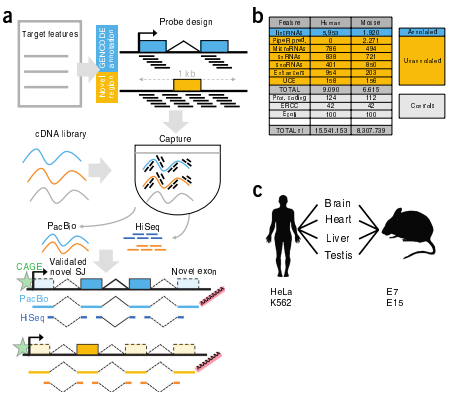


Figure 1 Using the CLS approach to extend GENCODE lncRNA annotation. (**a**) The strategy for automated, high-quality transcriptome annotation. CLS can be used to complete existing annotations (blue) or to map novel transcript structures in suspected loci (gold). Capture oligonucleotides (black bars) 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[29]; snRNA, small nuclear RNA; snoRNA, 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[32] 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[34]. 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[35] (**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[24].

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

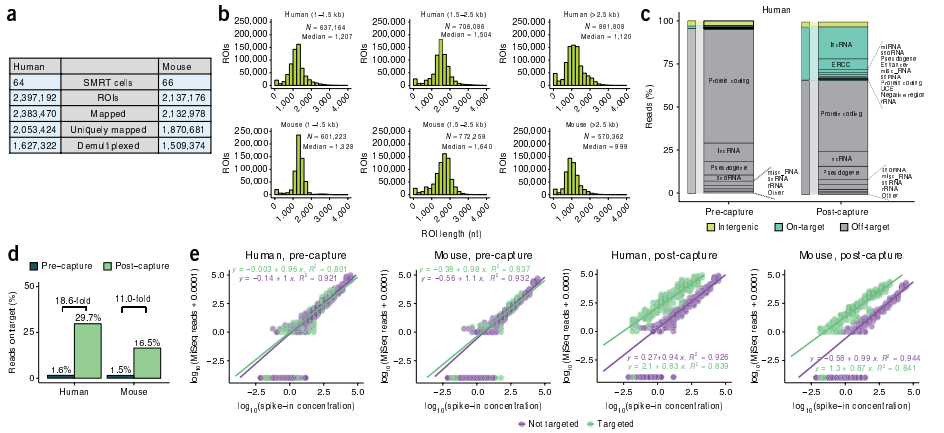


Figure 2 CLS yields an enriched, long-read transcriptome. (**a**) Sequencing statistics. (**b**) Length distributions of ROIs. Sequencing libraries were prepared from three size-selected cDNA fractions: 1–1.5 kb, 1.5–2.5 kb and >2.5 kb (**Supplementary Fig. 1b,c**). (**c**) A breakdown of sequenced reads 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 concentrate 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 5×10^{-3} copies per cell (Online Methods). It was less quantitative than CaptureSeq[24], 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[36] (*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[7] 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[37] (**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**),

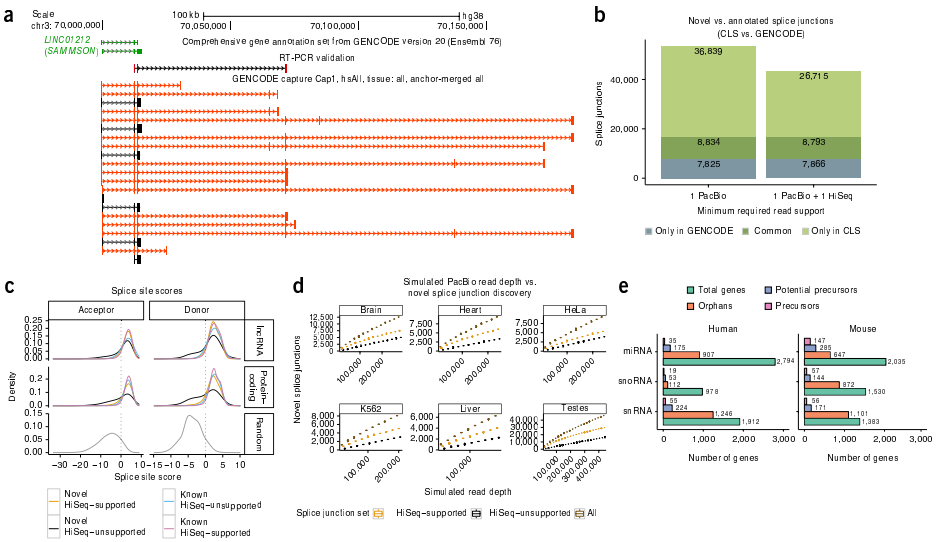


Figure 3 Extending known lncRNA gene structures. (**a**) Novel transcript structures from the *SAMMSON* locus. Green, GENCODE; black/red, known/novel CLS transcript models, respectively. An RT-PCR-amplified sequence is shown. (**b**) Splice junction discovery. The *y*-axis represents unique splice junctions 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[37]. 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[17]. 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+ precursors[38]. 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*)[39] (**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[32]. 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)[15] (**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[4], 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[40], 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[24,41] (**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[5,11]. However, it has remained unresolved whether this is a genuine biological trend or simply the result of annotation incompleteness[21]. 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[11]. 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[11]. 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[42] and CpG islands, although not evolutionarily conserved elements or phenotypic genome-wide association study variants[43] (**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[44] (**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[45,46]. 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[12] (**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[45,46].

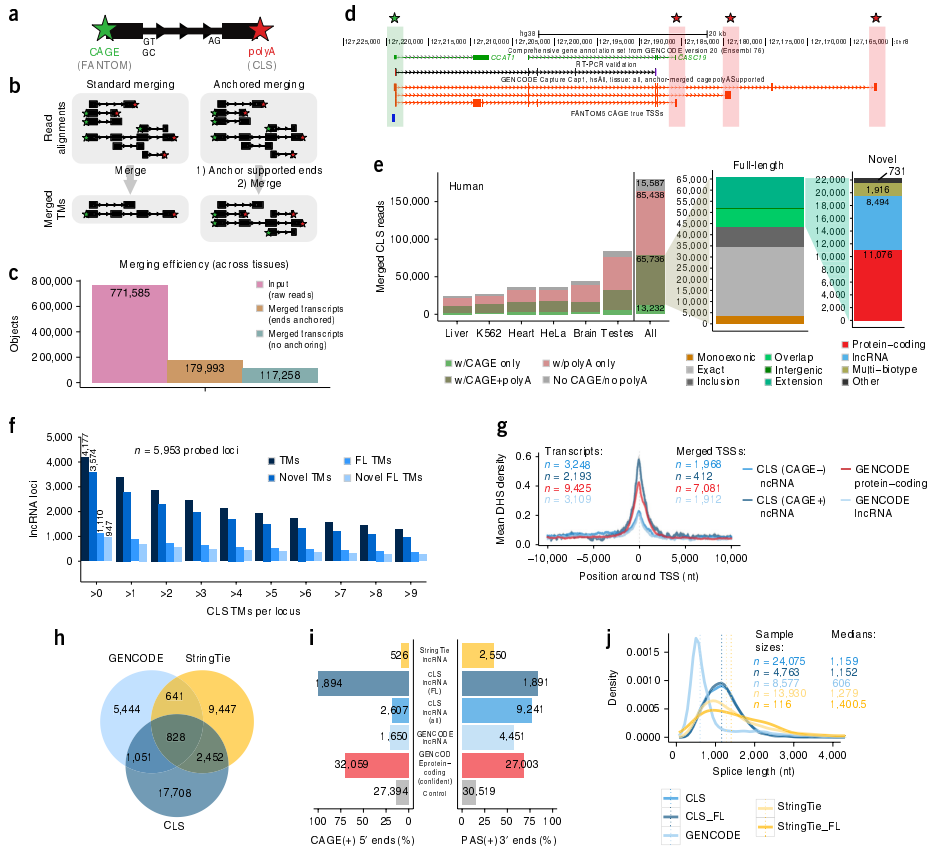


Figure 4 Full-length transcript annotation. (**a**) 5′ and 3′ termini of transcript models (TMs) inferred from CAGE clusters and poly(A) tails in ROIs, respectively. (**b**) In conventional ("standard") transcript merging, TSSs and polyadenylation sites overlapping other exons are lost. 'Anchored' merging 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 lncRNA 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 lncRNA 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**.

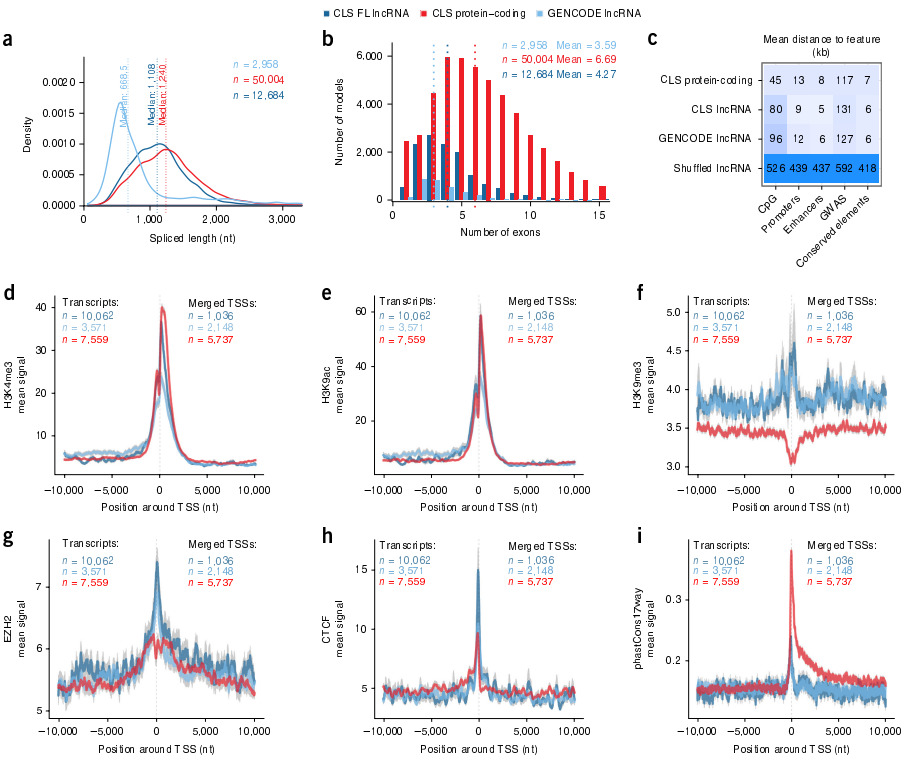


Figure 5 Properties of full-length lncRNA transcripts. (**a**) The mature, spliced transcript length of CLS full-length (FL) transcript models from targeted lncRNA loci, transcript models from the targeted and detected GENCODE lncRNA loci, and CLS full-length transcript models from protein-coding loci. (**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)[47,48]. 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[49,50] (**Fig. 6a**, Online Methods, **Supplementary Data Set 1**). This analysis revealed evidence for

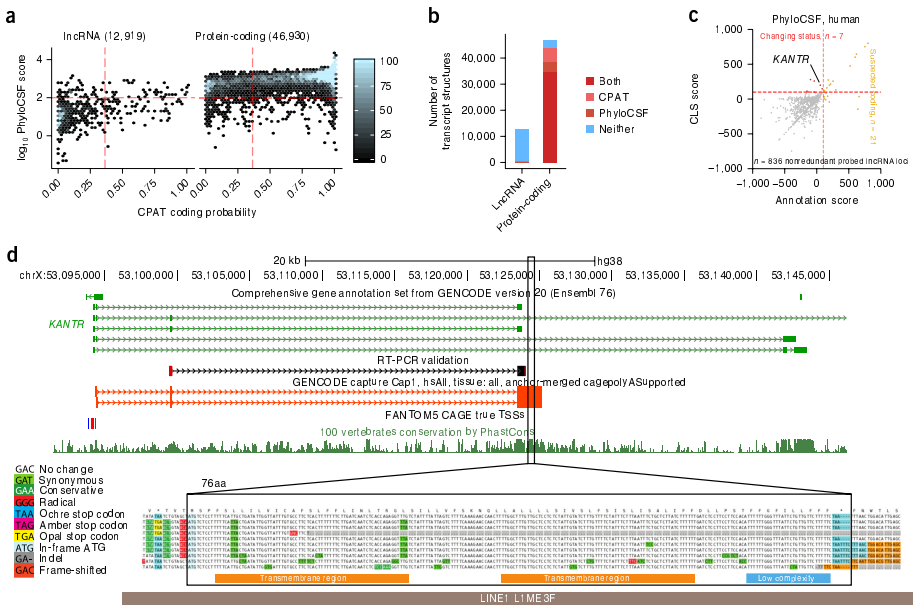


Figure 6 Protein-coding potential of full-length lncRNAs. (**a**) The predicted protein-coding potential of all full-length transcript models mapped to lncRNA (left) or protein-coding loci (right). Points represent full-length transcript models. The *y*-axis indicates the coding likelihood according 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[53] 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[51]. 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[52]. 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[11]. 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[45,46]. 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/; preloaded 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; CodAlignView, 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.

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 PhyloCSF. 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.

1. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
2. Jia, H. *et al.* Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA* **16**, 1478–1487 (2010).
3. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
4. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
5. 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).
6. 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).
7. Iyer, M.K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
8. Zhao, Y. *et al.* NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* **44**, D203–D208 (2016).
9. 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).
10. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
11. 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).
12. Bernstein, B.E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
13. Chen, L. *et al.* Transcriptional diversity during lineage commitment of human blood progenitors. *Science* **345**, 1251033 (2014).
14. 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. Pertea, 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 lncRNAs 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[10] 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[54]. 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[29]. As before, the top 20 most expressed lncRNAs, as estimated from ENCODE[31] 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[31]. 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 µg of each RNA sample, we added 4 µl 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 synthetized 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 synthetized from 3.5 µl 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 µg 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 µg.

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 µl of each sample were again quantified by Qubit dsDNA BR assay (Thermo Fisher), with 12.3 µg for human and 9.57 µg 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 µl 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 × sequence length × total read count). This showed that, as expected, size selection boosted the sequencing of longer templates.

PacBio library preparation. Approximately 2 µg 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 × 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 × 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[55] 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[56] (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[56] (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[57]. Overlap was defined as ≥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[24]. 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**).

doi:10.1038/ng.3988

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[57]. 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. 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).

nature research

Corresponding author(s): Rory Johnson / Roderic Guigo

☐ Initial submission ☐ Revised version ☒ Final submission

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	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	☐	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	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)
☒	☐	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Nature Genetics: doi:10.1038/ng.3988

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.

Nature Genetics: doi:10.1038/ng.3988

II.2. Supplementary information

Supplementary Note

From Lagarde *et al.*, **High-throughput annotation of full-length long noncoding RNAs with Capture Long-Read Sequencing** (DOI: https://doi.org/10.1101/105064)

Contents

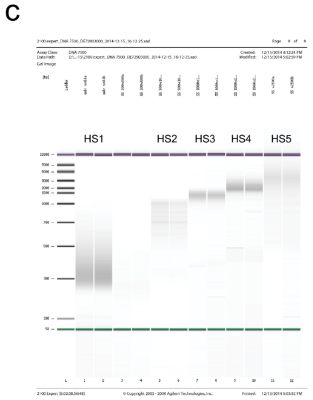
Supplementary Figures

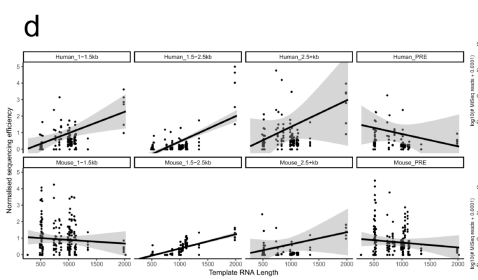
Note

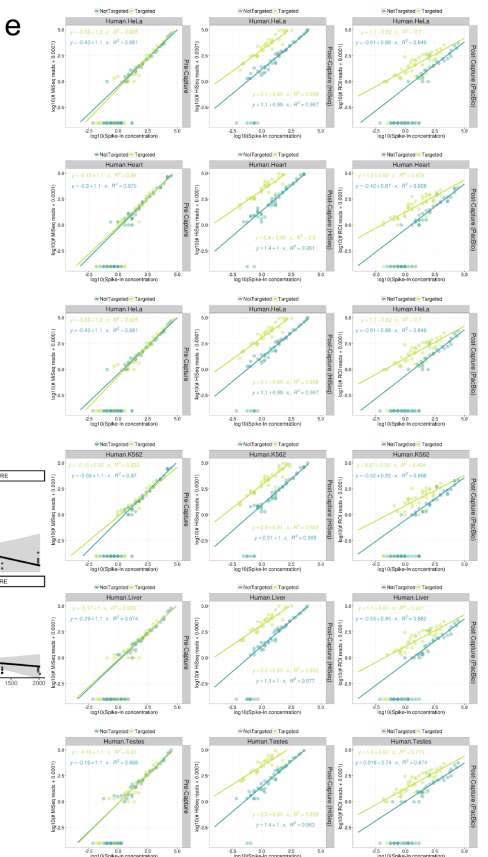
Each supplementary figure is followed by its title and caption on the following page. Links and page numbers in the Table of Contents refer to figure captions.

b

	Range (nt)	Human
1	200-500	HS1
2	500-1000	HS2
3	1000-1500	HS3
4	1500-2500	HS4
5	>2500	HS5







Supplementary Figure 1

Supplementary Figure 1: RNA Capture enrichment, PacBio cDNA Size Fractionation and Sequencing

(a) qPCR validation of enrichment

Quantitative PCR was performed to assess capture performance. Templates were pooled cDNA, before and after capture. Separate amplifications were performed on cDNA prepared for MiSeq (fragmented) or PacBio (full-length). Primers were designed to a selection of target sequences: GAPDH/ GNBL21/ Actb/ Gusb are housekeeping mRNAs; Spike-in 1&2 were targeted in the capture library; Spike-in 4 was present but not targeted. Lnc 1-4 and mR 1-4 refer to randomly selected, targeted lncRNAs and mRNAs respectively. Note that spike-ins are common to human and mouse experiments. y-axis shows the value of (Ct-POST - Ct-PRE), where Ct refers to the PCR threshold cycle. PCRs were carried out in technical triplicate and the mean is shown. Also shown are error bars denoting the standard deviation.

(b) Size fractionation of captured cDNA

cDNAs were size selected into five ranges. The last three were selected for subsequent sequencing.

(c) Agarose gel electrophoresis of size-selected post-capture cDNA samples

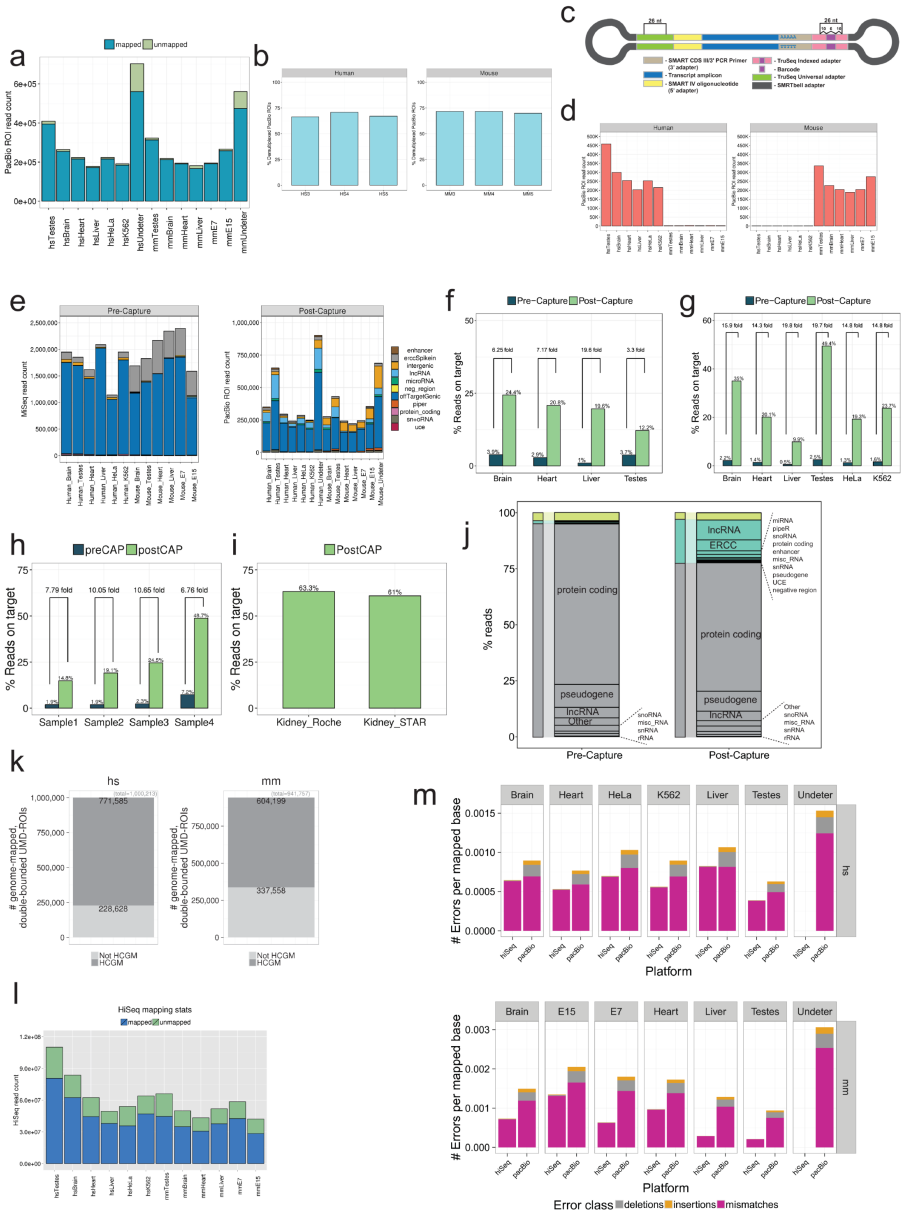
(d) Template length-dependence of PacBio sequencing

Each panel shows data from a different sequencing library. Top row: human; bottom row: mouse. The first three panels of each row show post-capture PacBio data for indicated size-selected fractions. The fourth panels of each row show similar data for pre-capture MiSeq data, which did not undergo size selection. Every point represents one of the synthetic ERCC spike-in RNA sequences added to samples prior to library preparation. x-axes show the length of these sequences. y-axes show the normalised sequencing efficiency: the sequencing reads per molecule, normalised to length and sequencing depth. Details may be found in the Methods. Lines show the best linear fit, and shading indicates the 95% confidence interval.

(e) Spike-in detection curves for individual human tissues

Data are analogous to Figure 2e, but broken down by tissues (rows). First column: pre-capture samples, with MiSeq sequencing; Second column: post-capture samples, with HiSeq sequencing; Third column: post-capture samples, with PacBio sequencing. Note the log scales for each axis. Each point represents one of 92 spiked-in synthetic ERCC RNA sequences. 42 were probed in the capture design (light green), while the remaining 50 were not (dark green). Lines represent linear fits to each dataset, whose parameters are shown above. Given the log-log representation, a linear response of read counts to template concentrate should yield an equation of type $y = c + mx$, where m is 1.

Nature Genetics: doi:10.1038/ng.3988



Supplementary Figure 2

Supplementary Figure 2: Sequencing library structure and statistics

(a) Read mapping statistics

Shown are the numbers of reads, broken down by originating sample, which could be mapped to the genome. "Undeter" refers to reads from which the barcode could not be confidently identified. "hs": human; "mm": mouse.

(b) ROI demultiplexing efficiency

The y-axis indicates the fraction of reads in each pooled sample whose sample of origin could be inferred based on hexamer barcodes. x-axis columns indicate the three size-selected fractions of each species.

(c) Schematic structure of PacBio reads and library adapters

Indexed adapters carry a unique 6-nt barcode, specific to the originating sample. The adapter and barcode sequences are available in Supplementary Data 3.

(d) Demultiplexed ROIs by sample of origin

Undetermined reads are not shown. "hs": human; "mm": mouse.

(e) Capture enrichment by tissue

Pre-Capture data was generated using MiSeq reads of pooled cDNA prior to capture, while PostCapture data represents PacBio ROI reads. "Undeter" refers to reads from which the barcode could not be confidently identified, and hence from an undetermined sample. Colours refer to the biotype of the feature to which the reads map. These feature classes are composed of targeted features (Figure 1b) or off-target features (either genic in dark blue, or intergenic in orange). Most off-target genic features are protein-coding genes. Notice the increase in representation of targeted features (mainly lncRNAs, light blue) in Post-capture compared to Pre-capture samples.

(f-g) Capture performance in individual tissues for Capture Short-Seq (CSS, data from Clark *et al.*, 2015) (f) and CLS (g)

The y-axis shows the percent of all mapped ROIs originating from targeted regions. Enrichment is defined as the ratio of this value in Post- and Pre-capture samples. Note that pre-capture rates in (f) were estimated using pre-capture MiSeq libraries generated in the present study.

(h-i) Comparing capture protocols shows that long cDNA targets yield lower capture efficiency

Sample1: Original CLS protocol (as used and described here), PolyA-selected, unfragmented. Sample2: Improved CLS protocol (see Methods), PolyA-selected, unfragmented. Sample 3: Improved CLS protocol, Total RNA, unfragmented. Sample 4: Roche SeqCap RNA protocol, Total RNA, fragmented. **(h)** Performance statistics for the four captures. Compare the on-target rates for Post-capture ("postCAP") material in Sample2/Sample4 and Sample3/Sample4. **(i)** Performance statistics for in-house data provided by Roche, for SeqCap capture of fragmented kidney cDNA.

(j) Breakdown of sequenced reads by gene biotype, pre- (left) and post-capture (right), for mouse

Colours denote the on/off-target status of the genomic region from which the reads originate, namely: Grey: reads originating from annotated but not targeted features; green: reads from targeted features, including lncRNAs; yellow: reads from unannotated, non-targeted regions. The ERCC class comprises only those ERCC spike-ins that were probed in this experiment. Note that when a given read overlapped more than one targeted class of regions, it was counted in each of these classes separately. Equivalent human data are found in Figure 2c.

(k) Results of the selection of High-Confidence Genome Mappings (HCGMs)

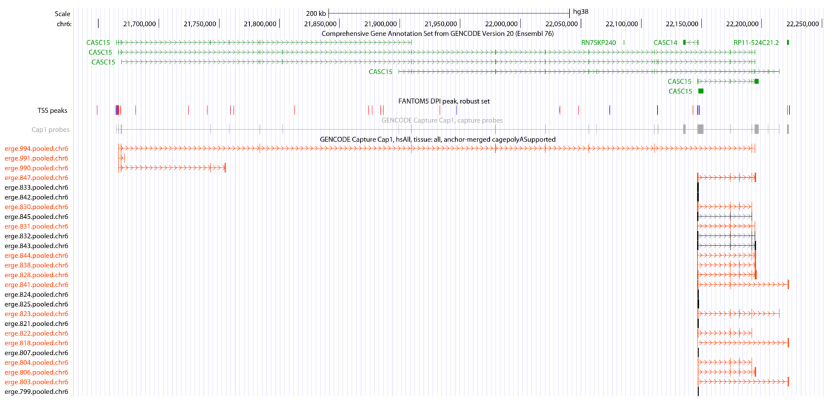
The number of double-bounded ROIs with and without HCGM (see definition in Methods) is represented in human ("hs", left panel) and mouse ("mm", right panel). The total number of double-bounded ROIs is reported at the top of each bar, in light grey.

(l) Post-Capture HiSeq read mapping statistics

(m) Sequencing error rates in human (top) and mouse (bottom) samples

The rate of sequencing error per sample and sequencing platform is represented on the y-axis. Sequencing errors are subdivided into mismatches (magenta), deletions (grey) and insertions (orange) with respect to the genome reference. The top of each bar corresponds to the global error rate in each library. "Undeter": undetermined reads, *i.e.*, non-demultiplexed (not available for HiSeq libraries).

CASC15 / ENSG00000272168



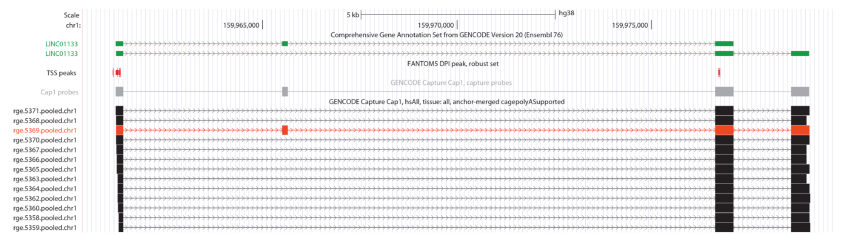
GAS5 / ENSG00000234741

Supplementary Figure 3

Supplementary Figure 3: Examples of known lncRNAs with changes in their annotated structures

Nature Genetics: doi:10.1038/ng.3988

LINC01133 / ENSG00000224259



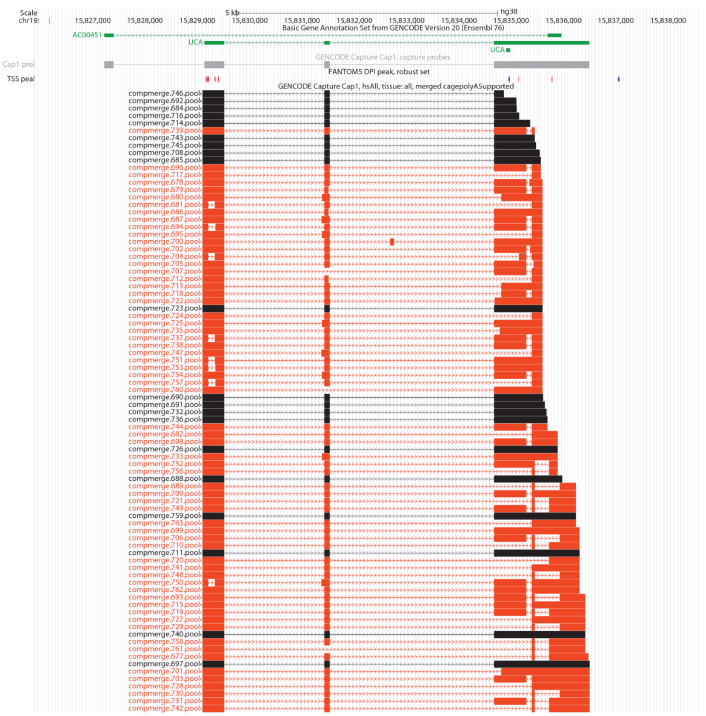
Supplementary Figure 4

Nature Genetics: doi:10.1038/ng.3988

Supplementary Figure 4: Examples of known lncRNAs with almost no change in their annotated structures

Nature Genetics: doi:10.1038/ng.3988

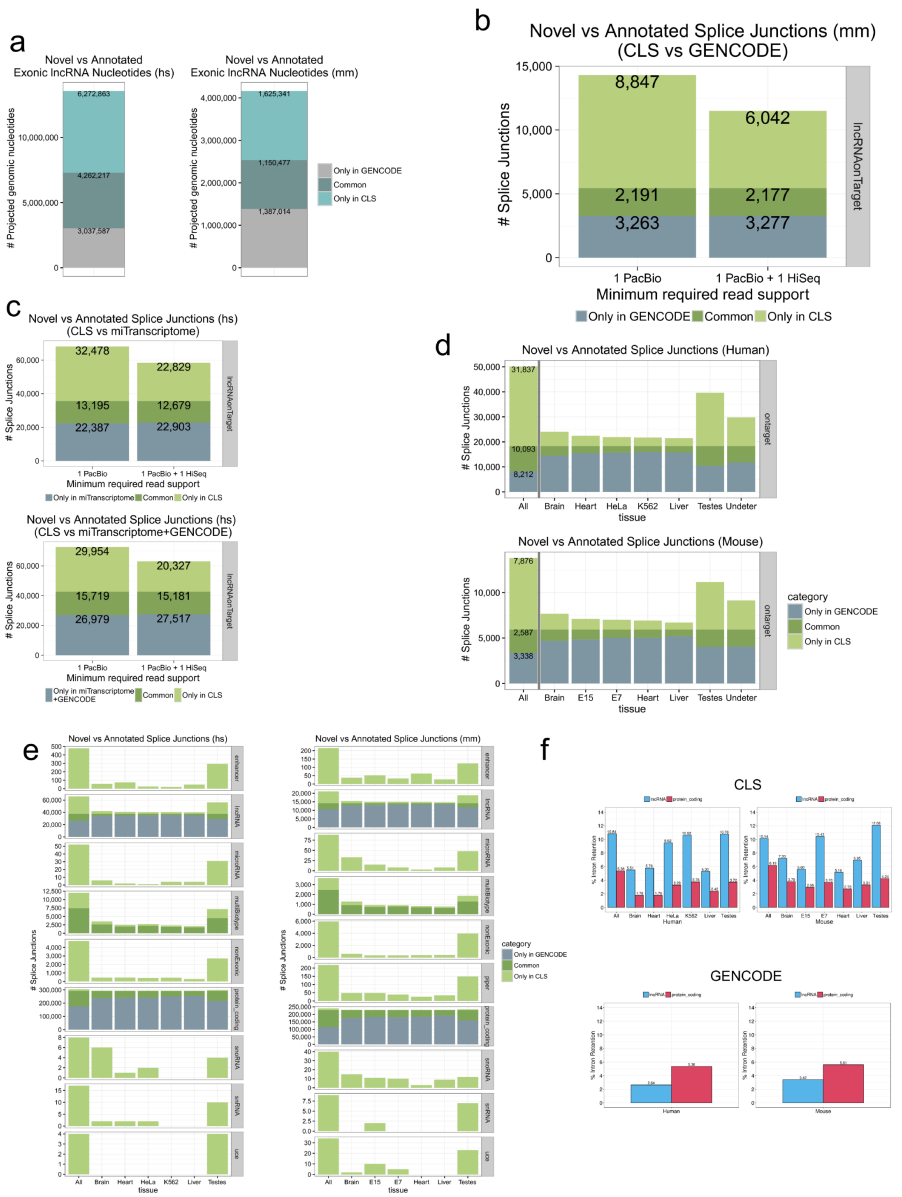
UCA1 / ENSG00000214049



Supplementary Figure 5

Nature Genetics: doi:10.1038/ng.3988

Supplementary Figure 5: Examples of known lncRNAs with changes in their annotated structures (II)



Supplementary Figure 6

Nature Genetics: doi:10.1038/ng.3988

Supplementary Figure 6: Exon and Intron Discovery by CLS

(a) Novel exonic bases discovered by CLS

Figures show the number of nucleotides that (i) are annotated in the targeted GENCODE lncRNA annotation but not detected ("Only in GENCODE"), (ii) are annotated and detected by CLS ("Common") or (iii) detected nucleotides that are not present in GENCODE and hence novel ("Only in CLS"). Left: human; Right: mouse. Note that nucleotide counts are from collapsed (merged annotations) and hence are non-redundant. Data on novel nucleotides only refer to ROIs that map to targeted lncRNA loci.

(b) Discovery of splice junctions (SJs) in targeted lncRNAs for mouse

GENCODE v.M3 is used as a reference. The *y*-axis denotes counts of unique SJs. Only "on-target" junctions originating from probed lncRNA loci are considered. Grey represents annotated SJs that are not detected. Dark green represents annotated SJs that are detected by CLS. Light green represent novel SJs that are identified by CLS but not present in the annotation. The left column represents all SJs, and the right column represents only high-confidence SJs, supported by at least one split-read from Illumina short read sequencing. Equivalent human data is in Figure 3b.

(c) Discovery of splice junctions (SJs) in targeted lncRNAs for human (comparison with *miTranscriptome*)

Novel splice junction discovery with respect to *miTranscriptome* (top panel) and the union of GENCODE and *miTranscriptome* (bottom panel) SJ sets. The figure layout and color legend is analogous to (b).

(d) Novel splice junctions by tissue in targeted loci

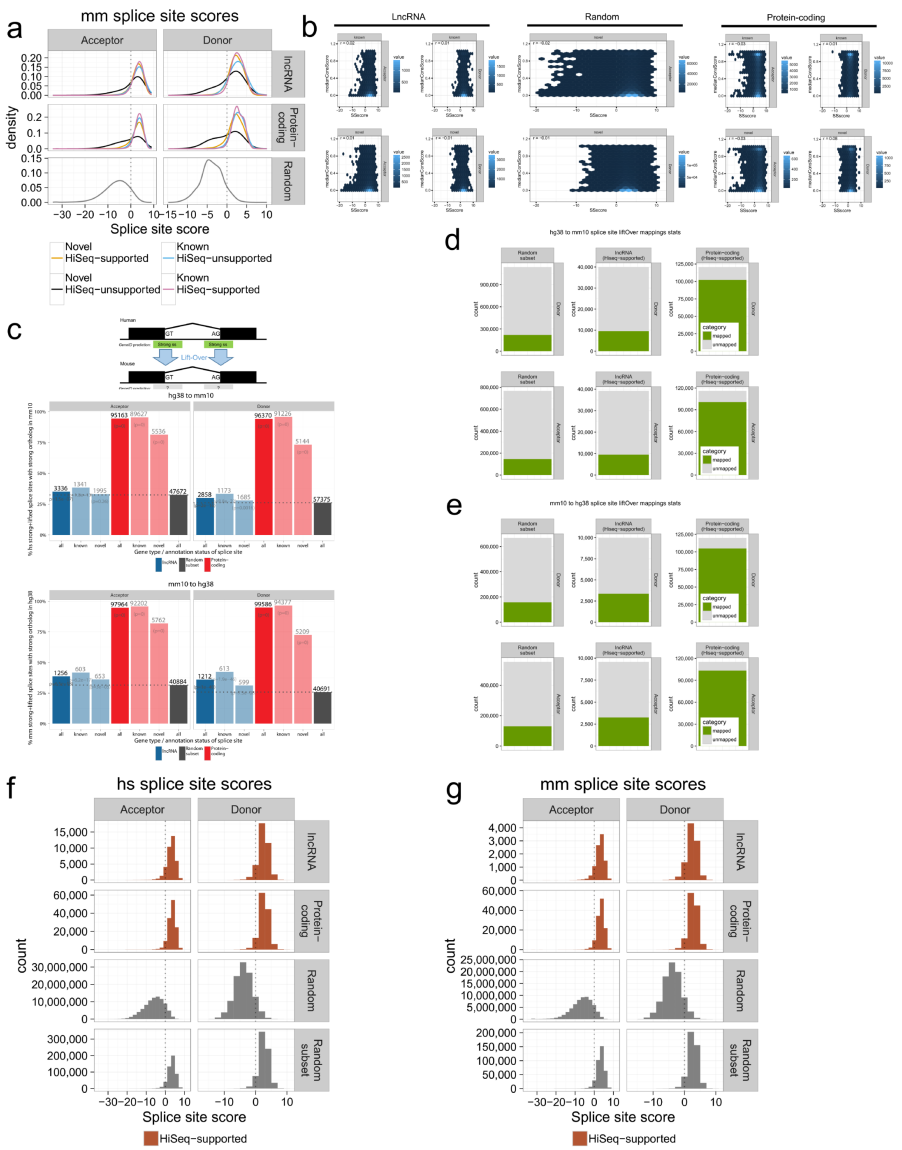
Figures display the number of splice sites discovered by CLS and compared to GENCODE annotations, broken down by tissue. Only high-confidence, HiSeq short read-supported CLS junctions are considered.

(e) Splice junction discovery statistics by tissue and biotype

Figures display the number of splice junctions discovered by CLS and compared to GENCODE annotations in human (left panel) and mouse (right panel), broken down by tissue and ROI biotype. Only high-confidence, HiSeq short read-supported CLS junctions are considered.

(f) Analysis of intron retention (IR) rates

Top panel: Proportion of transcripts with at least one retained intron in lncRNA and protein-coding CLS transcripts in human (left) and mouse (right). Bottom panel: Proportion of GENCODE lncRNA and protein-coding transcripts with at least one retained intron in human and mouse samples. Red indicates IR rate in lncRNAs, while blue indicates IR rate in protein coding transcripts.



Supplementary Figure 7

Supplementary Figure 7: Splice site (SS) motif and evolutionary analysis

(a) Splice site motif quality in mouse

Panels plot the distribution of predicted SJ strength, for acceptors (left) and donors (right). Splice site strength was computed using position weight matrices from *geneid*. Data are shown for non-redundant SJs from CLS transcript models from targeted lncRNAs (top), protein-coding genes (middle), or background distribution sampled from randomly-selected AG (acceptor-like) and GT (donor-like) dinucleotides (bottom). Analogous human data can be found in Figure 3c.

(b) Evolutionary conservation of known and novel splice sites

Panels show the distribution of splice sites (broken down by donor and acceptor sites) as a function of base-level nucleotide conservation ("medianConsScore", as calculated by PhastCons 100 vertebrate alignments) (y-axis) and predicted splice site strength ("SSscore", as determined by the *geneid* software) (x-axis).

(c) Evolutionary conservation of splice sites

The figures show the rate of conservation of different classes of splice sites (SSs). Conservation is defined by having a high-strength predicted SS at the orthologous site in the other genome. Orthologous regions were obtained from whole-genome alignments. Percentages only relate to those SSs for which an alignment exists (see (d) and (e)), HiSeq-supported, and deemed "strong" (*i.e.*, with a positive *geneid* score) in both the original and target genomes. Upper panel: conservation of human SSs in mouse; Lower panel: conservation of mouse SSs in human. Dark shades: all sites; Light shades: known/novel subsets of SSs. Background sites (referred to as "Random subset") are nearby putative SSs with no evidence of splicing, but with similar *geneid* scores, see (f) and (g). The actual SS counts are specified above each bar. Statistical significance for each set of SSs was estimated using Chi-square test of conserved/non-conserved sites, compared to background sites, and the obtained p-values are reported on each bar.

(d) human (hg38) to mouse (mm10) liftOver mapping statistics

Depicted in green are the number of hg38 strong SSs of each category for which an orthologous site could be found in mm10 using whole-genome alignments.

(e) mouse (mm10) to human (hg38) liftOver mapping statistics

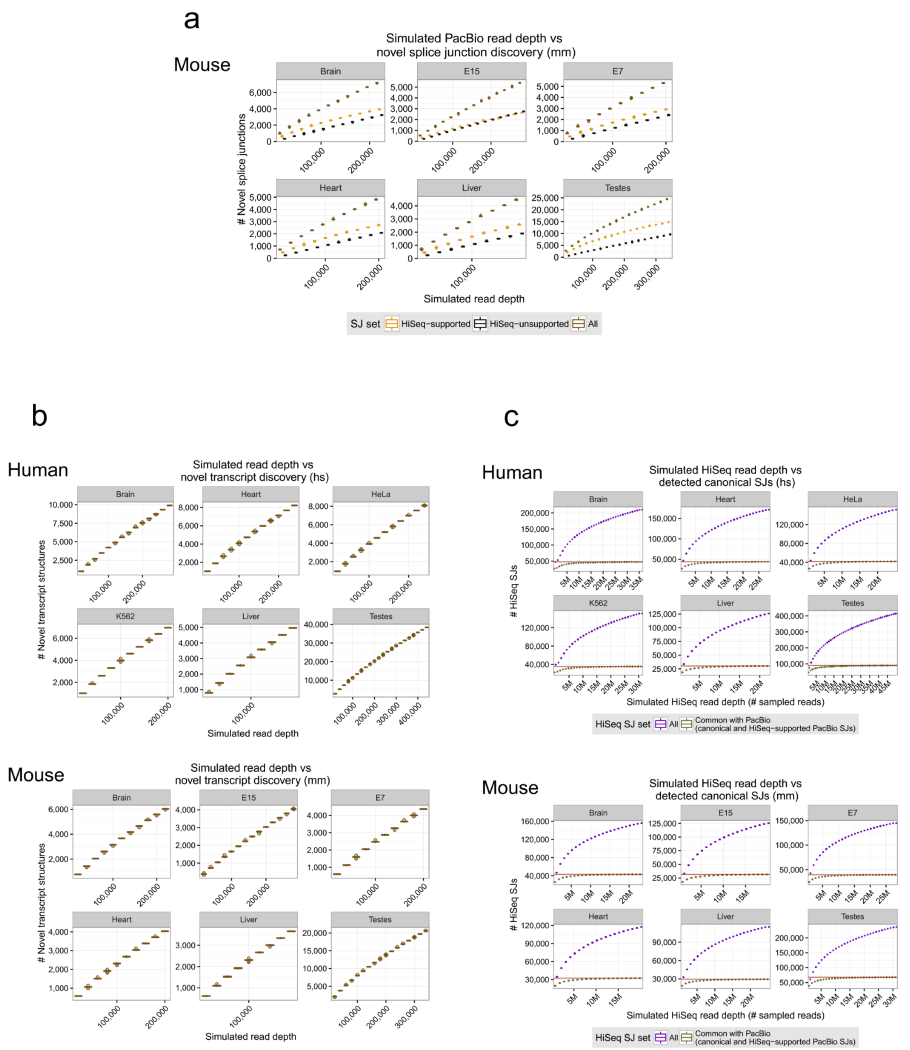
Depicted in green are the number of mm10 strong SSs of each category for which an orthologous site could be found in hg38 using whole-genome alignments.

(f) *geneid* score distribution of human splice sites used in the conservation analysis

The "Random subset" category corresponds to splice sites sampled from the "Random" set, such that its overall score distribution mimics that of lncRNA and protein-coding sites, depicted in the two upper panels.

(g) *geneid* score distribution of mouse splice sites used in the conservation analysis

Legend: see (f).



Supplementary Figure 8

Supplementary Figure 8: Discovery/Saturation analysis

(a) Novel splice junction discovery as a function of sequencing depth in mouse

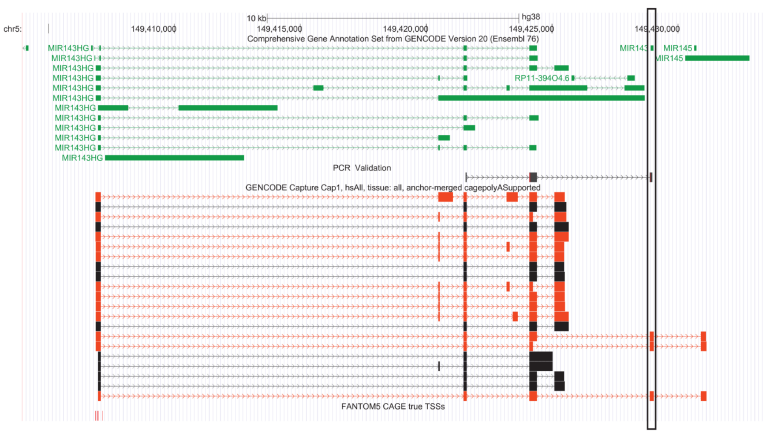
Each panel represents the number of novel splice junctions (SJs) discovered (y-axis) in a simulated analysis where increasing numbers of mapped ROIs (x-axis) were randomly sampled from the experiment. The SJs retrieved at each read depth were further stratified by level of sequencing support (Dark brown: all PacBio SJs; Orange: HiSeq-supported PacBio SJs; Black: HiSeq-unsupported PacBio SJs). Each randomization was repeated fifty times, and a boxplot summarizes the results at each simulated depth. The highest y value represents the actual number of novel SJs discovered. Analogous data for human is to be found in Figure 3d.

(b) Novel transcript discovery simulations for human (upper section) and mouse (lower section)

Each panel represents the number of novel transcript models (TMs) discovered (y-axis) in simulated analysis where increasing numbers of mapped ROIs (x-axis) were randomly sampled from the experiment. The randomizations were repeated a hundred times, and a boxplot summarizes the results at each simulated depth. The highest y value represents the actual number of novel TMs discovered.

(c) Splice junction discovery simulations for human (upper section) and mouse (lower section) using captured HiSeq reads

Each panel represents the number of splice junctions (SJs) discovered (y-axis) in simulated analysis where increasing numbers of HiSeq reads (x-axis) were randomly sampled from the experiment. The randomizations were repeated five times, and a boxplot summarizes the results at each simulated depth (purple: all HiSeq-derived SJs; brown: HiSeq-derived SJs also detected in PacBio matched samples). The highest y value represents the actual number of novel SJs discovered in each sample using HiSeq. The horizontal red line marks the number of HiSeq-supported PacBio SJs detected in the corresponding sample.

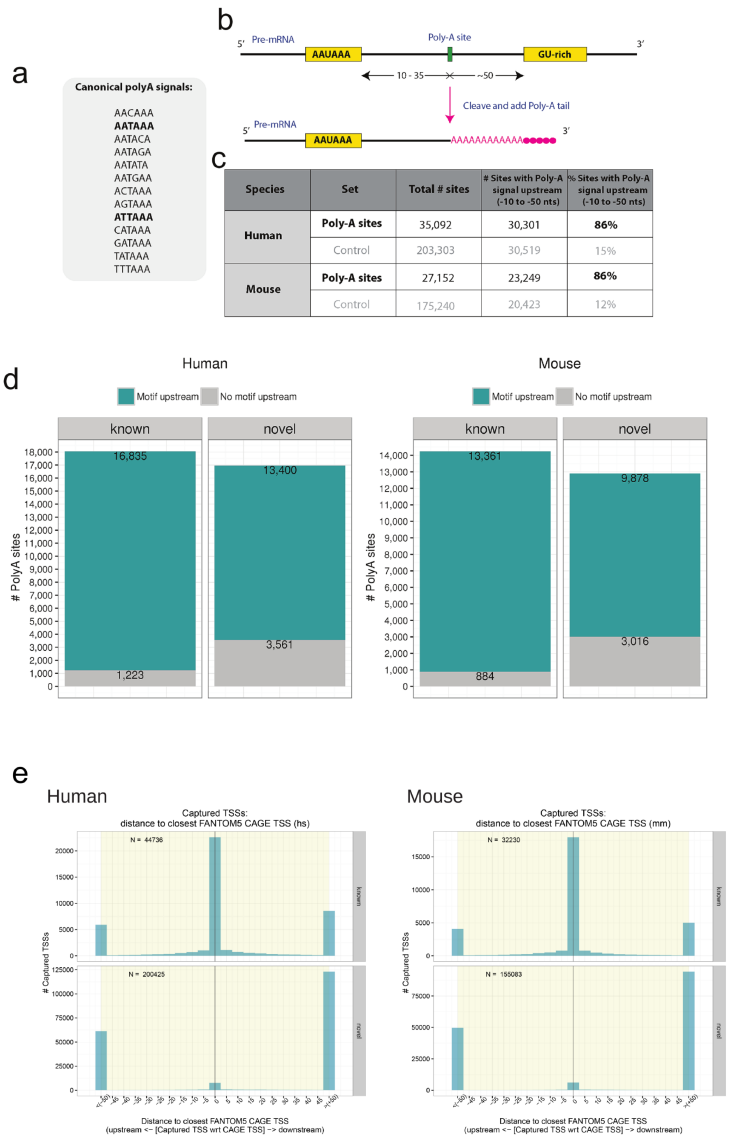


Supplementary Figure 9

Nature Genetics: doi:10.1038/ng.3988

Supplementary Figure 9: Evidence that CARMEN1 isoforms are precursors of hsa-mir-143

Shown is the CARMEN1 locus (chr5:149,402,925-149,452,858, hg38). GENCODE v20 annotation is green, capture probe targets in grey, full length CLS transcript models in black (known) and red (novel). Also visible are tracks for CAGE peaks from FANTOM and polyA sites from this study. Note the existence of novel isoforms directly overlapping on the same strand the mature hsa-mir-143 (boxed), for which no precursor annotation exists in GENCODE. Also shown is the sequence obtained by RT-PCR and Sanger sequencing (black).



Supplementary Figure 10

Supplementary Figure 10: Analysis of transcript ends

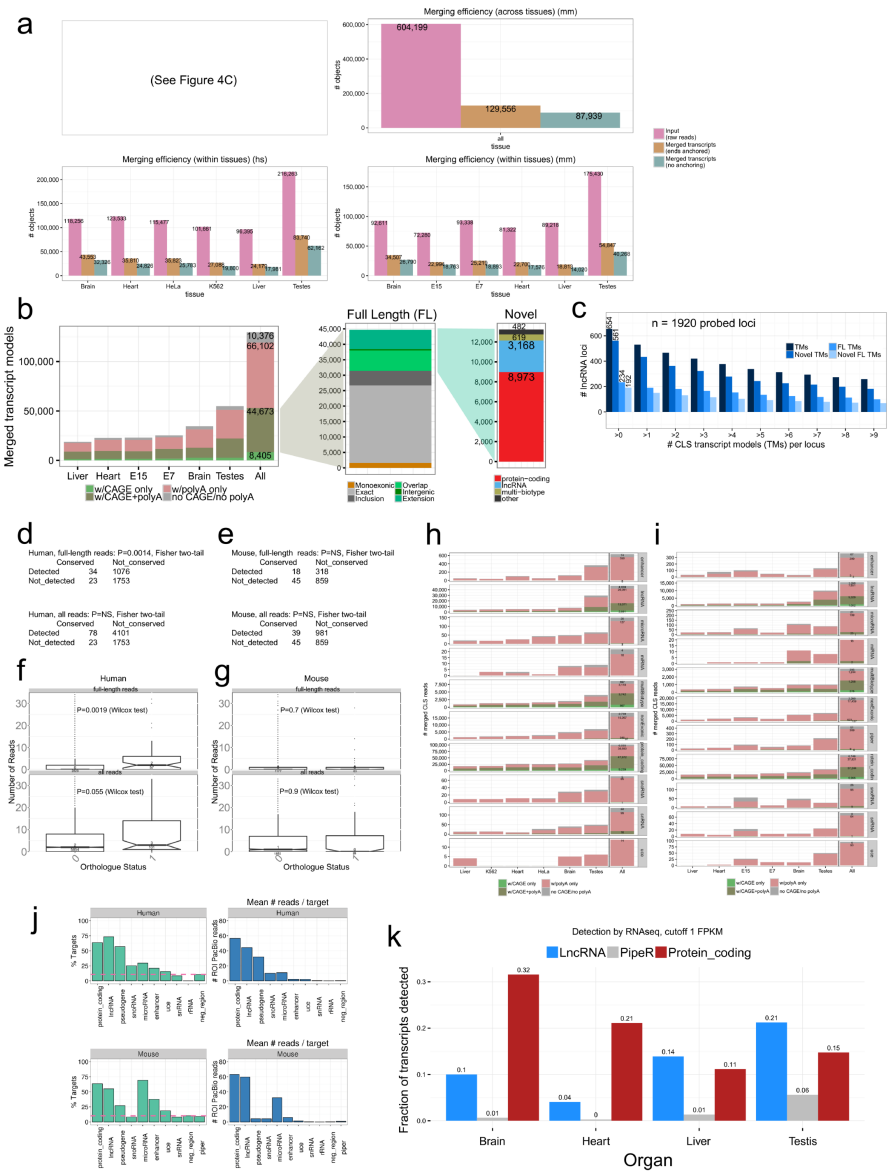
(a-d) Analysis of polyadenylation motifs around known and novel transcript 3′ ends

(a) list of the polyA motif considered here to be canonical. **(b)** Overview of the pre-mRNA termination and polyadenylation process. PolyA tails are generally added between 10 and 35 nt downstream of the polyA motif. **(c)** The rate at which CLS-discovered 3′ ends contain a canonical polyA signal. Control sites were generated by selecting the middle of non-terminal captured exons more than 100nt distal from the nearest captured polyA site. **(d)** Comparison of polyA motif frequency between known and novel 3′ ends.

(e) Captured TSS distance to closest CAGE cluster

Left: human; right: mouse; top: known TSSs; bottom: novel TSSs (w.r.t. GENCODE). Each plot is a histogram of the distance of the 5′ end of the start of a transcript model annotation to the FANTOM5 CAGE TSS. To the left are cases where captured TSS is upstream of the nearest CAGE TSS. The two extreme bins ("<(-50)" and ">(+50)") contain all cases where the closest CAGE cluster lies more than 50 bases away. The population size of both sets is reported in the top left corner of each plot.

Nature Genetics: doi:10.1038/ng.3988



Supplementary Figure 11

Supplementary Figure 11: Transcript merging, end support and detection in CLS

(a) Performance of anchored transcript model merging compared to a conventional approach

Charts indicate the number of unique transcript models created in each case (human on left, mouse on right). Upper panels show all reads, lower panel show reads broken down by sample origin.

(b) Anchor-merged transcript models identified by CLS in mouse

The y-axis of each panel shows unique transcript model (TM) counts. Left panel: All merged TMs, coloured by end support. Middle panel: Full length (FL) TMs, broken down by novelty with respect to existing GENCODE annotations. Green areas are novel and multi-exonic: "overlap" intersect an annotation on the same strand, but do not respect all its splice junctions; "intergenic" overlap no annotation on the same strand; "extension" respect all of an annotation′s splice junctions, and add novel ones. Right panel: Novel FL TMs, coloured by their biotype. "Other" refers to transcripts not mapping to any GENCODE protein-coding or lncRNA annotation. Note that the majority of "multi-biotype" models link a protein-coding gene to another locus. Equivalent data for mouse are found in Figure 4e.

(c) Probed lncRNA loci vs CLS transcript isoforms in mouse

The total numbers of probed lncRNA loci giving rise to CLS transcript models (TMs), novel TMs, full-length CLS TMs (FL TMs) and novel FL TMs in mouse, at increasing minimum cutoffs for each category.

(d-g) Detection rates of lncRNAs with evolutionary orthologues

(d-e): Contingency tables show the numbers of detected gene annotations in each category: "Detected" is defined as having one or more mapping reads, either of any type (upper row) or only full-length (lower row). None were significant by Fisher′s test (two-tailed). **(f-g)**: Boxplots show the same data as above, but broken down by the numbers of reads per gene. Numbers above boxes show the median (upper number) and number of data points (lower number). Orthologue status "0" and "1" indicate lncRNAs without / with identified orthologues, respectively. Further details may be found in the Methods.

(h-i) Transcript completeness by biotype and tissue source, in human (h) and mouse (i)

Figures show the number of unique merged transcript structures. Transcripts are coloured by 5′ or 3′ completeness.

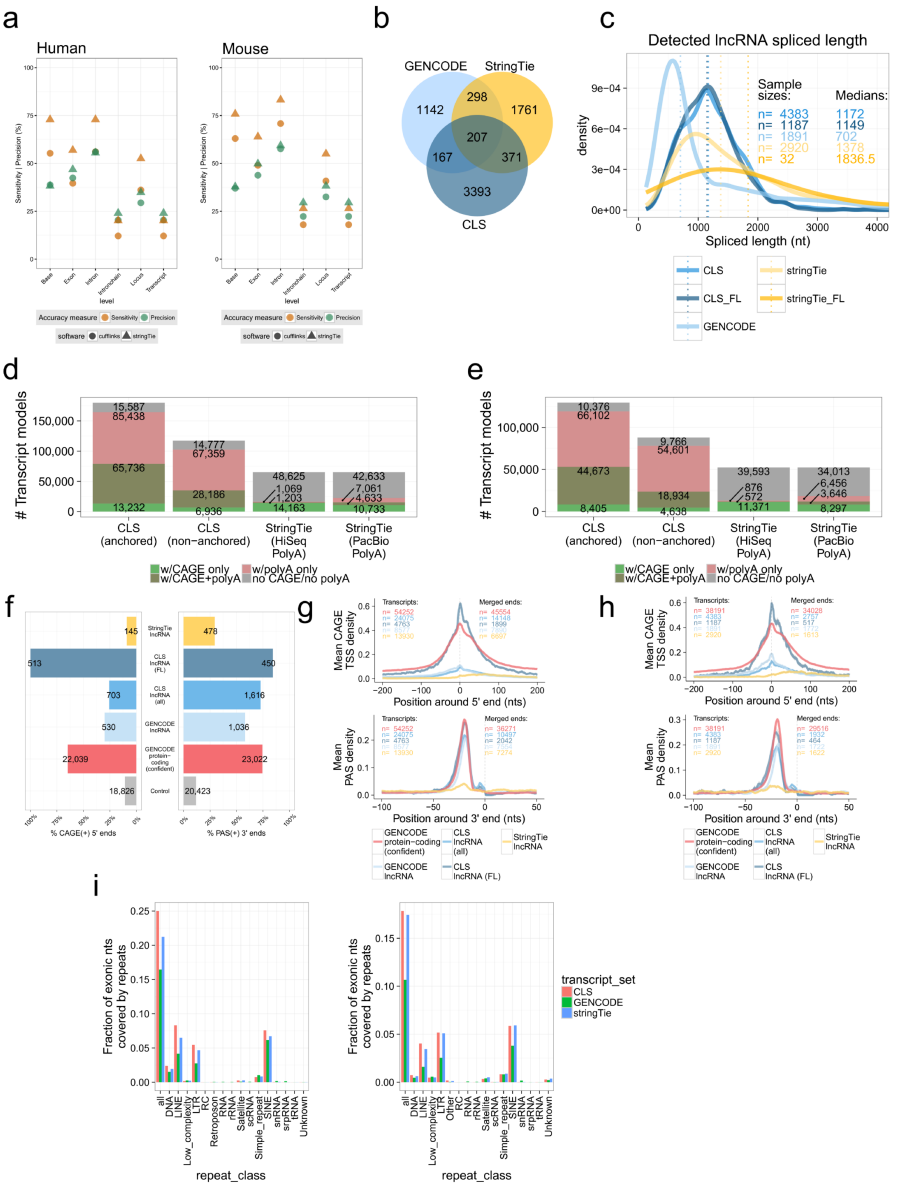
(j) Validation rates across target categories

Left panels show the percentage of probed targets detected by at least one ROI in human (top) and mouse (bottom). The rate of detection of negative regions is indicated with a pink dashed line. Right panels show the average number of ROIs detected per target class.

(k) Expression of PipeR lncRNA predictions in mouse tissues

Shown is the fraction of detected transcript models in each class, as measured by HiSeq in pre-captured samples and using a detection cut-off of >1 FPKM. Numbers of analysed transcripts: lncRNA - 8170, PipeR - 2469, protein-coding - 77,499.

Nature Genetics: doi:10.1038/ng.3988



Supplementary Figure 12

Nature Genetics: doi:10.1038/ng.3988

Supplementary Figure 12: Comparison of CLS with short-read transcript reconstruction methods

(a) Benchmarking of the *StringTie* and *Cufflinks* transcript assembly methods using PacBio evidence as a reference

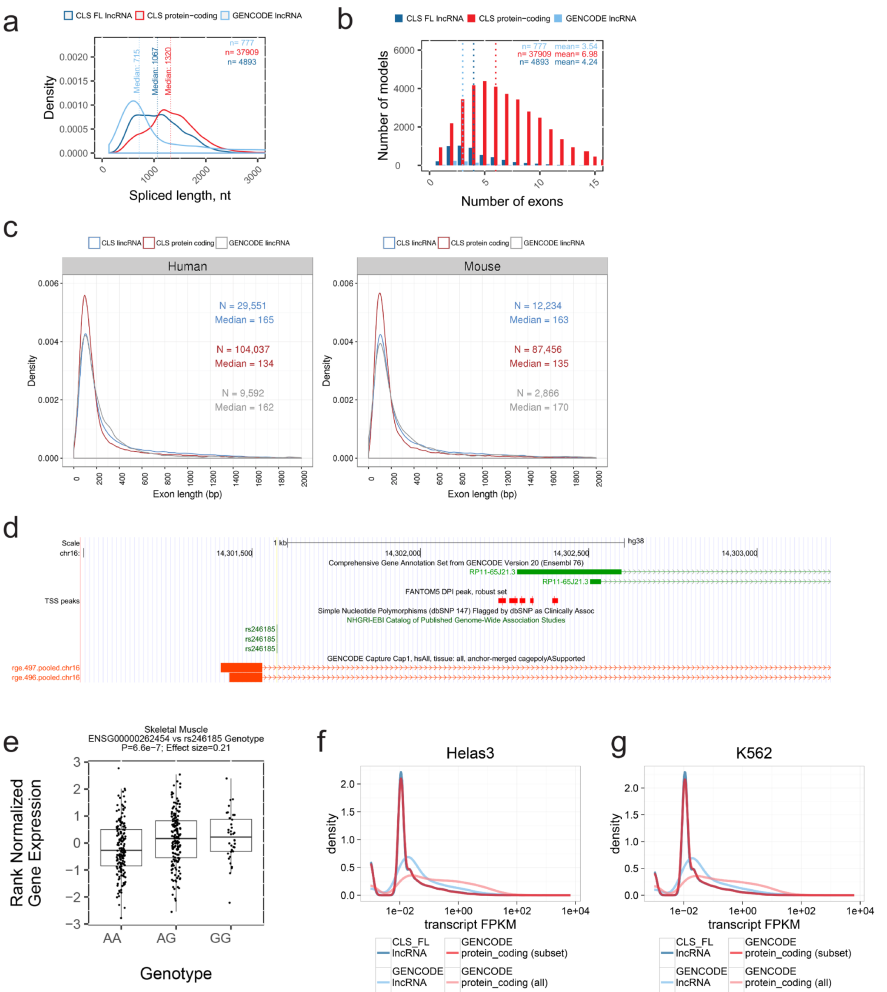
The y-axis displays the sensitivity/precision of each method in retrieving the indicated transcriptome elements (x-axis), as defined by CLS TMs.

(b-h) Comparison of *StringTie* and CLS transcript models (TMs)

Data shown in (b), (c) and (f) are for mouse; equivalent human data may be found in Figure 4. **(b)** Comparison of the numbers of unique transcript models present in each collection. Shared transcripts are defined by having identical intron chains. **(c)** Spliced length distributions of indicated non-redundant transcript catalogues. "FL" indicates the subset of transcripts from each catalogue that has 5′ support from CAGE, and 3′ support from PacBio-identified polyA sites. The median spliced length of each population is indicated by a vertical dotted line. **(d)** and **(e)**: *StringTie* models 5′ and 3′ completeness in human and mouse, respectively, compared to CLS merged models. "HiSeq PolyA" / "PacBio PolyA": comparison with polyA sites called using captured HiSeq / PacBio reads, respectively (see Methods). **(f)** Comparing completeness of transcript annotations: 5′ and 3′ completeness as estimated by CAGE overlap and upstream polyadenylation signal (PAS), with respect to 5′ (left) and 3′ ends (right), respectively. Neighbouring transcript ends were merged within each individual set (maximum distance: +/- 5nts on the same strand). "*GENCODE lncRNA*": subset of probed GENCODE lncRNAs detected by CLS or *StringTie*. "*GENCODE protein-coding (confident)*": 5′/3′ boundaries of high-confidence GENCODE protein-coding transcripts (see Methods). Control sites represent a random sample of internal exons′ middle coordinate. Represented is the proportion of transcript ends with CAGE or PAS support in each set (mouse). CAGE(+) 5′ ends are those TSSs having a CAGE cluster within a +/-50 bases window around them. Similarly, PAS(+) 3′ ends correspond to 3′ ends falling 10 to 50 bases downstream of a PAS motif. Note that full length (FL) CLS models have, by definition, a CAGE signal at 5′ end, and thus have 100% 5′ completeness. Corresponding counts of CAGE- or PAS-supported features are indicated on each bar. **(g-h)** CAGE TSS (top panel) and PAS (bottom panel) density aggregate plots, in human (g) and mouse (h). The mean density of CAGE TSSs and PAS (AATAAA and ATTAAA motifs) over each genomic position around various sets of transcript ends is represented. CAGE TSSs and PAS were required to overlap tested genomic regions on the same strand. Sample sizes (number of transcripts and number of merged ends after clustering) are indicated within each graph. Grey fringes represent the standard error of the mean. Transcript ends were merged as in (f), except for 3′ ends, for which a maximum clustering distance of 50 nucleotides was applied.

(i) Genome repeat coverage in CLS, *StringTie* and GENCODE exons

Shown is the fraction of exonic nucleotides covering genome repeats of various classes in each set of transcripts (left: human; right: mouse).



Supplementary Figure 13

Supplementary Figure 13: Full-length lncRNA transcripts: properties and genomic environment

(a-b) Comparison of lncRNA and mRNA transcript structure in mouse

(a) The mature, spliced transcript length of: CLS full-length transcript models from targeted lncRNA loci (dark blue); transcript models from the targeted and detected GENCODE lncRNA loci (light blue); CLS full-length transcript models from protein-coding loci (red). **(b)** The numbers of exons per full length transcript model, from the same groups as in (a). Dotted lines represent medians.

(c) Exon length distributions

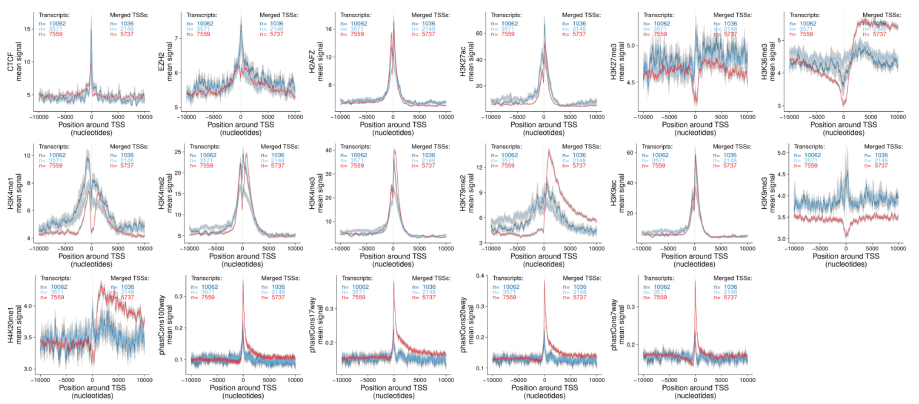
The distribution of exon lengths of: CLS full-length transcript models from targeted lncRNA loci (blue); transcript models from the targeted and detected GENCODE lncRNA loci (grey); CLS full-length transcript models from protein-coding loci (red). Left: human; right: mouse.

(d-e) Example of an expression QTL at lncRNA RP11-65J2

(d) The RP11-65J21.3 (ENSG00000262454) locus, showing phenotype-associated SNP rs246185. Existing GENCODE *v*20 annotation is shown in green, novel full-length transcript models in red. **(e)** Expression of ENSG00000262454 in muscle of GTEx individuals, broken down by genotype of rs246185. eQTL analysis was obtained from the GTEx Portal (http://www.gtexportal.org/home/).

(f-g) Creating an expression-matched set of protein-coding genes

Panels (f) and (g) show the distribution of whole-cell RNA levels for indicated transcript sets in HeLa and K562 cells, respectively. Note the log scale of the x-axis. Data are shown for CLS full-length lncRNA transcript models (dark blue), as well as the original GENCODE annotations to which they map (light blue). Also shown are data for all protein-coding genes (light red). From the latter, a subset was sampled with a similar expression distribution as the CLS lncRNAs (dark red).



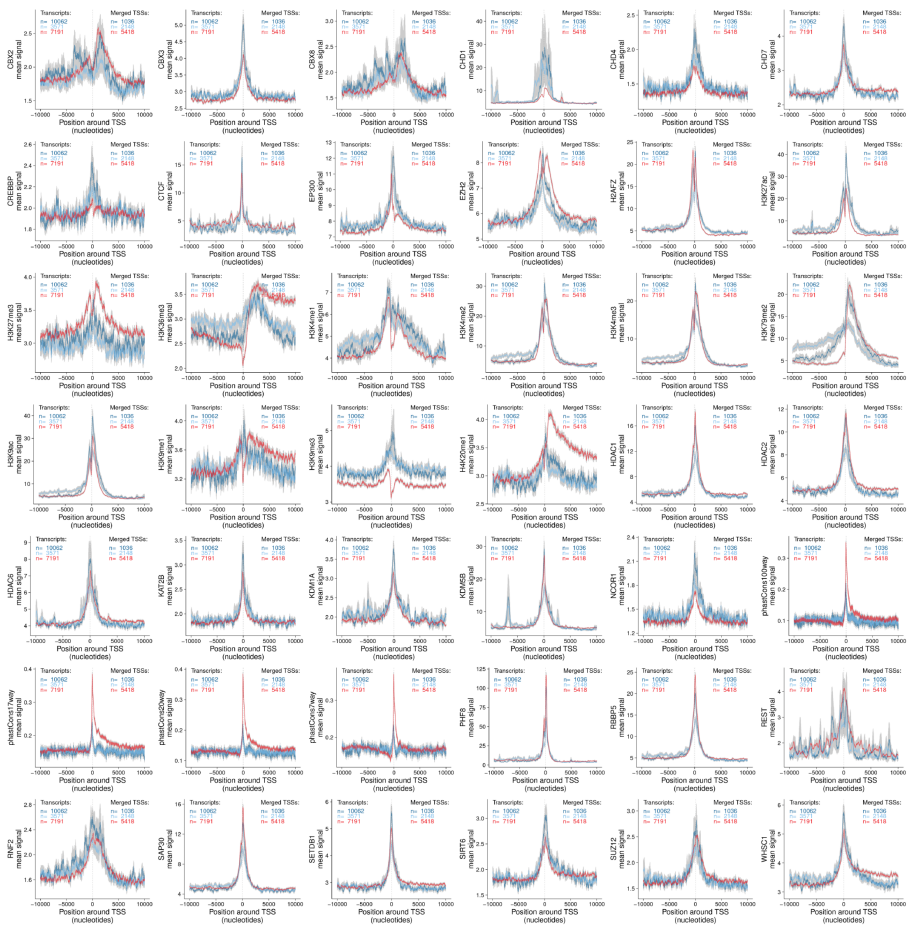
Supplementary Figure 14

Nature Genetics: doi:10.1038/ng.3988

Supplementary Figure 14: Characteristics of "standalone" promoters in HeLa

Montage of all signal density plots produced in HeLa across sets of "standalone" (*i.e.*, non-bidirectional) TSSs. The aggregate density of various features is shown across the TSS of indicated gene classes. Note that overlapping TSSs were merged within classes. The *y*-axis denotes the mean signal per TSS, and grey fringes represent the standard error of the mean. Gene sets are: Dark blue, full length lncRNA models from CLS; Light blue, the GENCODE annotation models from which the latter were derived; Red, a subset of protein-coding genes with similar expression in HeLa as the CLS lncRNAs.

Nature Genetics: doi:10.1038/ng.3988

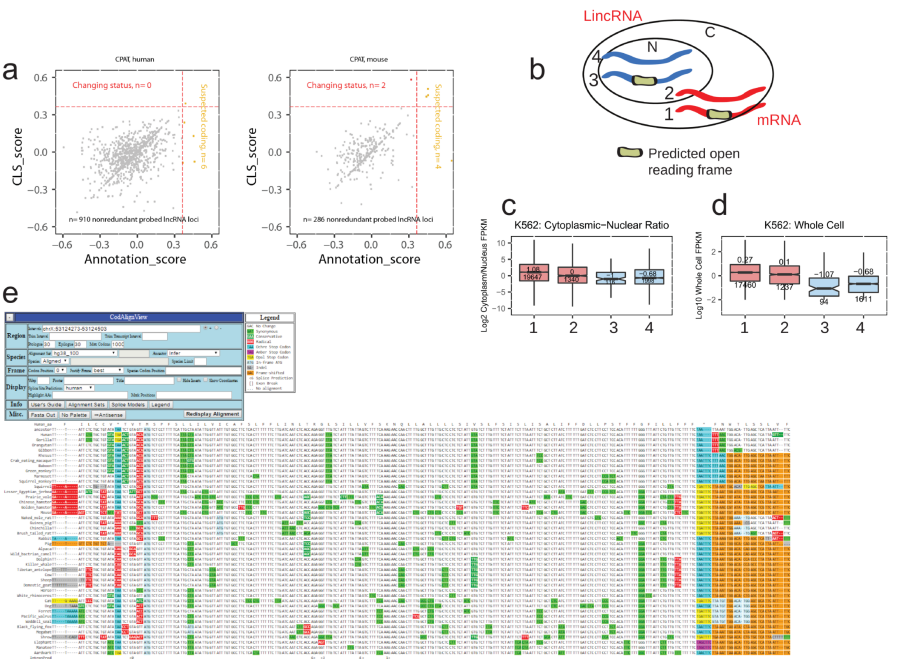


Supplementary Figure 15

Supplementary Figure 15: Characteristics of "standalone" promoters in K562

Montage of all signal density plots produced in K562 across sets of "standalone" (*i.e.*, non-bidirectional) TSSs. The aggregate density of various features is shown across the TSS of indicated gene classes. Note that overlapping TSSs were merged within classes. The *y*-axis denotes the mean signal per TSS, and grey fringes represent the standard error of the mean. Gene sets are: Dark blue, full length lncRNA models from CLS; Light blue, the GENCODE annotation models from which the latter were derived; Red, a subset of protein-coding genes with similar expression in K562 as the CLS lncRNAs.

Nature Genetics: doi:10.1038/ng.3988



Supplementary Figure 16

Nature Genetics: doi:10.1038/ng.3988

Supplementary Figure 16: Analysis of protein-coding potential and subcellular localization

(a) Changes in protein-coding status due to long read extension

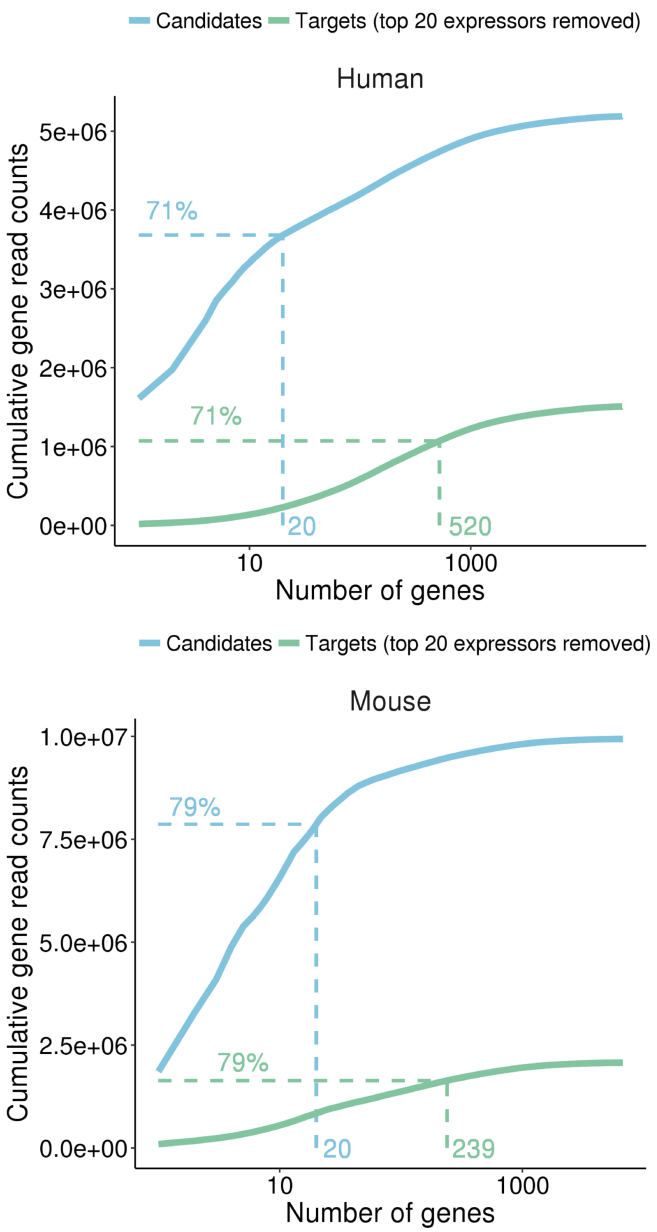
Changes in CPAT-predicted protein-coding potential in lncRNAs due to extension by CLS. Each point represents a probed and detected lncRNA gene. For each gene, the highest-scoring associated transcript model is used. The x-axis denotes the CPAT score of original GENCODE annotation, and the y-axis the score of associated full-length read models from CLS. Red lines indicate the prediction threshold dividing coding and non-coding. In yellow are shown gene loci that may be protein-coding, prior to CLS. In red are shown gene loci whose status changes following CLS.

(b-d) Expression and localisation properties of full-length transcript models in K562 cells, broken down by annotated and predicted coding potential

(b) Schematic of subcellular localisation of annotated lncRNAs (blue) and mRNAs (red). Indicated are identified ORFs in these transcripts in beige colour. **(c)** Subcellular localisation of transcripts in K562 cells. Localisation (y-axis) is estimated from RNAseq data by the log2 ratio of cytoplasmic RPKM / nucleus RPKM. Inside each box are displayed the median value (above) and the number of transcript models considered (below). Samples are numbered as in (b). **(d)** Similar to (c), but showing whole-cell expression values. Note that here, ORFs are defined to be present if predicted by either PhyloCSF or CPAT.

(e) Detailed view of KANTR short ORF

This corresponds to region chrX:53124273-53124488 in the hg38_100 alignment set, using CodAlignView (https://data.broadinstitute.org/compbio1/cav.php).



Supplementary Figure 17

Nature Genetics: doi:10.1038/ng.3988

Supplementary Figure 17: Removing high-expressed genes that dominate sequencing

Graphs show the cumulative number of sequencing reads originating from ranked lists of GENCODE lncRNA genes before (blue, "Candidates") and after (green, "Targets") removing the 20 most highly expressed genes, emphasizing the high fraction of all reads originating from the top 20 genes (note the logarithmic scale on the *x*-axis). The remaining gene models ("Targets") were used for capture probe design. Blue dashed lines represent the percentage of reads accounted for by the top 20 genes in the "Candidates" set. Green dashed lines depict the number of genes accounting for that percentage of reads (*i.e.*, 71% in human, 79% in mouse) in the "Targets" set. These plots were produced using matched, public Illumina short-read RNAseq data corresponding to the organs studied here. Upper panel: human; lower panel: mouse.

Supplementary Tables

Species	Sample	Total # uniquely mapped double-bounded reads	# polyA reads	% polyA reads	# on-genome polyA site clusters (+/-5nts), min 2 reads
Human	**Brain**	170,012	106,505	63%	9,607
	Heart	153,214	115,973	76%	8,502
	HeLa	150,196	109,023	73%	8,211
	K562	128,994	98,758	77%	6,097
	Liver	118,868	94,739	80%	5,786
	Testes	278,929	206,457	74%	16,850
	Total	**1,000,213**	**731,455**	**73%**	**35,092**
Mouse	**Brain**	150,371	85,679	57%	7,903
	E15	185,837	69,564	37%	5,271
	E7	131,314	93,459	71%	6,419
	Heart	117,908	79,320	67%	5,608
	Liver	123,941	96,774	78%	4,867
	Testes	232,386	176,318	76%	12,852
	Total	**941,757**	**601,114**	**64%**	**27,152**

Supplementary Table 1

Supplementary Table 1: Statistics on polyA site identification

Statistics on polyA site identification in double-bounded, genome-mapped reads (*i.e.* excluding ERCC spike-ins).

a	b	c	d	e	f	g	h
Species	**Biotype**	**# merged transcript models**	**# merged FL transcript models**	**% merged transcript models that are FL (d/c)**	**# novel, FL transcript models**	**% FL transcript models that are novel (f/d)**	**# loci affected by novel, FL transcript models**
Human	enhancer	634	1	0%	1	100%	1
	lncRNA	47,002 (42,463)	13,071 (11,429)	28%	8,494	65%	1,220 (812)
	microRNA	172	2	1%	2	100%	1
	misc_RNA	19	5	26%	3	60%	1
	Mt_rRNA	28	18	64%	0	0%	0
	Mt_tRNA	4	3	75%	1	33%	1
	multiBiotype	8,616	3,742	43%	1,916	51%	1,027
	neg_region	45	3	7%	3	100%	1
	nonExonic	18,751	548	3%	287	52%	0
	prot_coding	102,156	47,672	47%	11,076	23%	4,294
	pseudogene	2,344	655	28%	429	66%	103
	snoRNA	71	0	0%	0	N/A	0
	snRNA	137	16	12%	5	31%	2
	uce	14	0	0%	0	N/A	0
	Total	**179,993**	**65,736**	**37%**	**22,217**	**34%**	**N/A**
Mouse	enhancer	364	5	1%	5	100%	1
	lncRNA	15,580 (13,130)	5,329 (4,350)	34%	3,168	59%	448 (249)
	microRNA	266	27	10%	9	33%	6
	misc_RNA	15	0	0%	0	N/A	0
	Mt_rRNA	42	0	0%	0	N/A	0
	Mt_tRNA	7	2	29%	0	0%	0
	multiBiotype	3,075	1,258	41%	619	49%	337
	neg_region	37	0	0%	0	N/A	0
	nonExonic	20,469	623	3%	419	67%	0
	piper	433	9	2%	8	89%	3
	prot_coding	88,177	37,244	42%	8,973	24%	3,608
	pseudogene	791	167	21%	40	24%	19
	snoRNA	131	9	7%	1	11%	1
	snRNA	73	0	0%	0	N/A	0
	uce	96	0	0%	0	N/A	0
	Total	**129,556**	**44,673**	**35%**	**13,242**	**30%**	**N/A**

Supplementary Table 2

Supplementary Table 2: Breakdown of captured transcripts by gene biotype and novelty

Numbers refer to transcript models merged across all tissue samples. Counts corresponding to lncRNA probed regions are reported between parentheses where appropriate. The number of annotated loci originating these transcript models is indicated in the rightmost column.

Species	Merging method	End support level	Total # TMs	# HiSeq-supported TMs	% HiSeq-supported TMs
Human	anchored	any	179,993	155,617	86%
		CAGE+polyA	65,736	60,046	91%
		CAGE	78,968	71,363	90%
		polyA	151,174	134,537	89%
	standard	any	117,258	94,163	80%
		CAGE+polyA	28,186	23,070	82%
		CAGE	35,122	28,494	81%
		polyA	95,545	80,135	84%
Mouse	anchored	any	129,556	113,186	87%
		CAGE+polyA	44,673	40,693	91%
		CAGE	53,078	47,924	90%
		polyA	110,775	99,362	90%
	standard	any	87,939	72,413	82%
		CAGE+polyA	18,934	15,261	81%
		CAGE	23,572	19,027	81%
		polyA	73,535	62,757	85%

Supplementary Table 3

Supplementary Table 3: HiSeq support of merged CLS transcript models

Numbers refer to transcript models (TMs) merged across all tissue samples using the "anchored" and "standard" (*i.e.*, non-anchored) methods. HiSeq-supported TMs refer to those TMs whose entire set of introns are supported by at least one split short read in the captured HiSeq libraries. These transcript models are referred to as "HiSeq-supported TMs" elsewhere in the paper. "CAGE+polyA" end support level corresponds to full-length TMs. "Any" end support level refers to all merged TMs, including the ones without CAGE/polyA end support.

Feature	Source	Number of targeted transcripts	Comments
lncRNAs (intergenic)	GENCODE v20	9,560	
microRNA	mirBase v20	785	Tiled 1kb
snoRNA	GENCODE v20	401	Tiled 1kb
snRNA	GENCODE v20	838	Tiled 1kb
VISTA enhancers	http://enhancer.lbl.gov/	1,908	
Ultraconserved elements	UCNEbase	316	Any UCE less than 500 bp long were removed.
Protein-coding	GENCODE v20	100	Expression matched to lncRNAs
E. coli (random genomic)		100	Identical in human and mouse libraries
ERCC sequences (selected)	https://www.thermofisher.com/order/catalog/product/4456740	42	Identical in human and mouse libraries

Supplementary Table 4

Supplementary Table 4: Target regions for capture library design (human)

Feature	Source	Number of targeted transcripts	Comments
lncRNAs (intergenic)	GENCODE vM3	2,817	
Orthologues of human lncRNAs	PipeR	2,469	
microRNA	mirBase v20	494	Tiled 1kb
snoRNA	GENCODE vM3	850	Tiled 1kb
snRNA	GENCODE vM3	721	Tiled 1kb
VISTA enhancers	http://enhancer.lbl.gov/	406	
Ultraconserved elements	UCNEbase	312	
Protein-coding (expression matched)	GENCODE vM3	100	
E. coli (random genomic)		100	Identical in human and mouse libraries
ERCC sequences (selected)	https://www.thermofisher.com/order/catalog/product/4456740	42	Identical in human and mouse libraries

Supplementary Table 5

Supplementary Table 5: Target regions for capture library design (mouse)

Species	Sample	ERCC mix
Human	**Heart**	Mix1
	Testes	Mix2
	Liver	Mix2
	Brain	Mix1
	HeLa	Mix2
	K562	Mix1
Mouse	**Heart**	Mix1
	Testes	Mix2
	Liver	Mix1
	Brain	Mix2
	E7	Mix1
	E15	Mix2

Supplementary Table 6

Supplementary Table 6: ERCC spike-in mixes used per library

Species	Sample type	Illumina index ID	Index Sequence
Human	**Heart**	AD020	GTGGCC
	Testes	AD021	GTTTCG
	Liver	AD022	CGTACG
	Brain	AD023	GAGTGG
	HeLa	AD025	ACTGAT
	K562	AD027	ATTCCT
Mouse	**Heart**	AD013	AGTCAA
	Testes	AD014	AGTTCC
	Liver	AD015	ATGTCA
	Brain	AD016	CCGTCC
	E7	AD018	GTCCGC
	E15	AD019	GTGAAA

Supplementary Table 7

Supplementary Table 7: Index / barcode sequences

See Supplementary Data 3 for full adapter sequences.

Label	Size Range	# SMRT-cells	Loading Concentration (pM)	Loading Method	Read Bases of Insert	Mean Read Length of Insert	Mean Read Quality of Insert	Mean Number of Passes
MM_1	240 - 646	1	500	diff	3,479,177	329	98.9%	44
MM_2	438 - 3400	1	500	diff	13,436,069	594	99.0%	30
MM_3	896 - 5931	21	250	Mag	784,093,824	1304	99.0%	15
MM_4	672 - 6841	21	400	Mag	1,205,620,473	1561	99.1%	13
MM_5	500 - 5000	21	25	diff	51,313,641	986	99.0%	18
HS_1	253 - 698	1	500	diff	10,808,941	332	98.9%	45
HS_2	388 - 3262	1	500	diff	12,585,406	606	98.9%	28
HS_3	503 - 12138	21	25	Mag	64,413,386	1087	98.8%	17
HS_4	551 - 11636	21	25/35	Mag	1,049,441,562	1486	99.3%	13
HS_5	558 - 5000	21	40	Mag	989,112,311	1147	99.2%	16

Supplementary Table 8

Supplementary Table 8: Summary of PacBio sequencing

MM: mouse; HS: human.

Properties common to all samples/fractions:

- PacBio Kit: #100-259-100
- Polymerase used: P6/C4 (except HS_4: P5/C4 and HS_5: P4/C3)
- Movie length: 4h
- Post-run analysis: RS_ReadsOfInsert.1
- Files generated: FASTQ

	a	b	c	d	e
Dataset	**# Mapped reads**	**# Uniquely mapped reads (UMD-ROIs)**	**# double-bounded UMD-ROIs**	**# genome-mapped, double-bounded UMD-ROIs**	**% double-bounded UMD-ROIs (c/b)**
Hs Brain	274,732	265,170	200,767	170,012	76%
Hs Heart	232,699	224,910	176,357	153,214	78%
Hs HeLa	233,303	220,340	169,847	150,196	77%
Hs K562	198,890	185,684	140,307	128,994	76%
Hs Liver	185,797	180,379	136,492	118,868	76%
Hs Testes	423,727	405,447	306,445	278,929	76%
Hs - total	**1,549,148**	**1,481,930**	**1,130,215**	**1,000,213**	**76%**
Mm Brain	231,189	219,521	168,322	150,371	77%
Mm E15	280,718	266,837	201,578	185,837	76%
Mm E7	208,421	194,473	146,325	131,314	75%
Mm Heart	207,954	193,872	133,684	117,908	69%
Mm Liver	181,337	174,843	137,294	123,941	79%
Mm Testes	342,414	329,350	252,582	232,386	77%
Mm - total	**1,452,033**	**1,378,896**	**1,039,785**	**941,757**	**75%**

Supplementary Table 9

Supplementary Table 9: Summary statistics on UMD-ROIs and double-bounded reads

Hs: Human; Mm: Mouse.

UMD-ROIs: Uniquely Mapped and Demultiplexed ROIs

Undetermined (*i.e.*, non-demultiplexed) reads are not reported, as they do not bear a recognizable index sequence, by definition. Genome-mapped reads refer to reads not mapped to ERCC spike-in sequences.

	a	b	c	d	e	f	g	h	i
Species	**Total # uniquely mapped reads**	**# reads stranded by at least one method**	**# reads stranded by both methods**	**# reads with same strand inferred by both methods**	**# reads stranded by polyA method only**	**# reads stranded by SJ method only**	**% reads stranded by at least one method (b/a)**	**% reads stranded by both methods (c/a)**	**% reads with same strand inferred by both methods (d/c)**
Human	2,053,424	1,446,986	566,109	564,258	168,398	712,479	70.5%	27.6%	99.7%
Mouse	1,870,681	1,255,423	491,493	490,110	111,877	652,053	67.1%	26.3%	99.7%

Supplementary Table 10

Supplementary Table 10: Comparison/integration of polyA and SJ strand inference approaches

"Undetermined" (*i.e.*, non-demultiplexed) ROIs are included in the total.

Species	TSS type	# TSSs	# CAGE-supported TSSs	% CAGE-supported TSSs
Human	novel	200,425	16,305	8.1%
	known	44,736	30,352	67.9%
Mouse	novel	155,083	11,255	7.3%
	known	32,230	23,195	72.0%

Supplementary Table 11

Supplementary Table 11: CAGE support of novel vs known PacBio TSSs

A TSS is considered supported if a FANTOM "true" TSS is found within 50 bases around it on the same genomic strand (see Methods).

Cell line	ChIP-Seq antibody target	ENCODE portal file accession
Helas3	CTCF	ENCFF000BAN
Helas3	EZH2	ENCFF000BAV
Helas3	H2AFZ	ENCFF000BAZ
Helas3	H3K27ac	ENCFF000BBR
Helas3	H3K27me3	ENCFF000BBX
Helas3	H3K36me3	ENCFF000BCD
Helas3	H3K4me1	ENCFF000BBF
Helas3	H3K4me2	ENCFF000BCJ
Helas3	H3K4me3	ENCFF000BCP
Helas3	H3K79me2	ENCFF000BCV
Helas3	H3K9ac	ENCFF000BDB
Helas3	H3K9me3	ENCFF000BBL
Helas3	H4K20me1	ENCFF000BDI
K562	CBX2	ENCFF000BVA
K562	CBX3	ENCFF000BVE
K562	CBX8	ENCFF000BVI
K562	CHD1	ENCFF000BVO
K562	CHD4	ENCFF000BVT
K562	CHD7	ENCFF000BVW
K562	CREBBP	ENCFF000BUW
K562	CTCF	ENCFF000BWF
K562	EP300	ENCFF000CAL
K562	EZH2	ENCFF000BWL
K562	H2AFZ	ENCFF000BWT
K562	H3K27ac	ENCFF000BWY
K562	H3K27me3	ENCFF000BXD
K562	H3K36me3	ENCFF000BXJ
K562	H3K4me1	ENCFF000BXQ
K562	H3K4me2	ENCFF000BXV
K562	H3K4me3	ENCFF000BYB
K562	H3K79me2	ENCFF000BYH
K562	H3K9ac	ENCFF000BYN
K562	H3K9me1	ENCFF000BYR
K562	H3K9me3	ENCFF000BYX
K562	H4K20me1	ENCFF000BYZ
K562	HDAC1	ENCFF000BZF
K562	HDAC2	ENCFF000BZL
K562	HDAC6	ENCFF000BZR
K562	KAT2B	ENCFF000CAO
K562	KDM1A	ENCFF000BZV
K562	KDM5B	ENCFF000CBA
K562	NCOR1	ENCFF000BZZ
K562	PHF8	ENCFF000CAT
K562	RBBP5	ENCFF000CBL
K562	REST	ENCFF000CBP
K562	RNF2	ENCFF000CBQ
K562	SAP30	ENCFF000CBV
K562	SETDB1	ENCFF000CBY
K562	SIRT6	ENCFF000CCC
K562	SUZ12	ENCFF000CCH
K562	WHSC1	ENCFF000CAD

Supplementary Table 12

Supplementary Table 12: Datasets used in the TSS vs ChIP-Seq analysis

All files are of "signal" type, in bigWig format, and were obtained from the official ENCODE portal (https://www.encodeproject.org).

All corresponding experiments were performed in Bradley Bernstein's lab at the Broad Institute.

Dataset name	Description	Population size	
		# Transcripts	# Merged TSSs
CLS_FL lncRNA	Merged, full-length captured human transcripts	**10062**	**1036**
GENCODE lncRNA	GENCODE v.20 transcript models of simplified biotype "lncRNA", overlapping exons of CLS_FL lncRNAs on the same genomic strand (obtained using bedtools intersect -split -s -u -a GENCODE_lncRNAs -b CLS_FL_lncRNA)	**3571**	**2148**
GENCODE protein-coding	Subset of GENCODE v.20 transcript models, of transcript_type "protein_coding", matched to CLS_FL lncRNAs for transcript expression in K562 and HeLaS3 cell lines, with CAGE-supported TSSs (i.e., +/- 20 bases from a FANTOM "true" TSS on the same strand)	7,559 (HeLaS3) 7,191 (K562)	5,737 (HeLaS3) 5,418 (K562)

Supplementary Table 13

Supplementary Table 13: Transcript collections used in the TSS vs ChIP-Seq and TSS conservation analyses

Supplementary Methods

Abbreviations

- **FL**: full length
- **HCGM**: High-Confidence Genome Mapping
- **ROI**: read of insert, *i.e.* PacBio read
- **SJ**: splice junction
- **SS**: splice site
- **TM**: transcript model
- **TSS**: Transcription Start Site
- **UMD-ROI**: Uniquely Mapped and Demultiplexed ROI

Post-processing of ROI alignments

Selection of uniquely mapped ROIs

Demultiplexed ROIs mapped uniquely on the genome were selected from the BAM files using the bamflag utility (https://github.com/pervouchine/bamflag) with the "*-m2 -u*" options. This procedure resulted in a set of 1,481,930 (human) and 1,378,896 (mouse) reads, referred to as UMD-ROIs (Uniquely Mapped and Demultiplexed ROIs) hereafter.

Identification of "double-bounded" ROIs

We defined a set of double-bounded reads, namely, UMD-ROIs bounded by a Universal Adapter at one end, and an Indexed Adapter at the other (See schema in Supplementary Figure 2c). We reasoned that such reads should contain the entire cDNA sequence inserted between the two library adapters, and therefore be enriched in fully sequenced inserts.

Globally, about three quarters of uniquely mapped reads were found to be double-bounded both in human (1,130,215 reads, *i.e.* 76%, of which 1,000,213 on-genome) and mouse (1,039,785 reads, *i.e.* 75%, of which 941,757 on-genome). More detailed statistics on double-bounded UMD-ROIs are provided in Table 9.

Identification of poly-adenylated ROIs, on-genome polyA sites and signals

PolyA site calling

We identified poly-adenylated UMD-ROIs and on-genome polyA sites using the *samToPolyA* utility (https://github.com/julienlag/samToPolyA), developed in-house, with the following options: *minClipped=20*, *minAcontent=0.9*, *minUpMisPrimeAlength=10*. That is, we searched for read alignments where a genome match was immediately followed by a final stretch of more than 20 unaligned As or Ts (ignoring adapter sequences, and allowing up to 10% of non-A/non-T nucleotides over the total length of the tail), resulting in a set of potential poly-adenylated reads and on-genome polyA sites. Hits immediately preceded by an upstream A-rich genomic sequence ($>$ 10bp, with $\leq$ 1 non-A bp) were discarded, in order to avoid erroneously calling polyA sites from internally RT-primed cDNAs.

Using this conservative procedure, 731,455 (73%) and 601,114 (64%) reads were found to be poly-adenylated in human and mouse, respectively. Resulting on-genome polyA sites were subsequently merged into clusters using the *bedtools merge* utility [1], using a maximum clustering distance of 5 bases ("*-d 5*"), and forcing strandedness ("*-s*"). Only on-genome polyA site clusters supported by a minimum of 2 reads were kept for further analysis. In total, 35,092 (human) / 27,152 (mouse) non-redundant polyA sites were identified with this procedure. Table 1 summarizes the results of the polyA calling pipeline.

Proximity of polyA signals

We scanned the immediate 5' proximity of our polyA sites for the presence of poly-adenylation signals mentioned in Lopez *et al.*, 2006 [2] (Supplementary Figure 10a). Specifically, we extracted the [-50, -10] sequence window upstream of each non-redundant polyA site, and checked if at least one of those motifs was present in it. We performed the same operation with a collection of negative sites. This latter set was obtained by extracting the middle coordinate of each of our non-terminal PacBio captured exons, distal (+/- 100 bases) to any identified 3' end in our data, and subsequently merging them ("*bedtools merge -d 5 -s*").

Using this method, we established that globally, 86% of observed polyA sites were preceded by a polyA signal in both human and mouse, compared to 12/15% for negative sites, respectively (Supplementary Figure 10c). The same analysis was performed separately on "known" (*i.e.*, sites falling within +/- 50 bases of a GENCODE-annotated 3' end on the same strand) and "novel" (*i.e.*, sites falling more than 50 bases away from of a GENCODE-annotated 3' end on the same strand) polyA sites. We found that although novel sites were slightly more depleted in polyA signals when compared to known ones, they were overall far above the 12/15% random expectation (Supplementary Figure 10d).

ROI genomic strand inference

As PacBio SMRT cDNA sequencing is not directional, we inferred the genome strand of all (including non-demultiplexed) 2,053,424 (human) / 1,870,681 (mouse) uniquely mapped ROIs using the following two methods, in parallel.

"PolyA" approach

We used the *samToPolyA* utility (https://github.com/julienlag/samToPolyA, see PolyA site calling) to assign a genomic strand to poly-adenylated ROIs. Reads where a polyA tail was detected at their 3' end were assigned a '+' genomic strand, whereas reads with a polyT tail at their 5' end were deduced to originate from the '-' strand.

"Splice Junction" (SJ) approach

We extracted part of the SJ sequences (*i.e.* the first and last two nucleotides of each intron) of all ROI unique spliced mappings. We identified, when possible, canonical SJ motifs (GT and AG at the donor and acceptor site, respectively) in each intron of this dataset, and assigned it a genomic strand accordingly: '+' (plus) for GT/AG introns, and '-' (minus) for CT/AC (*i.e.*, the reverse-complement of GT/AG) introns. Each spliced ROI was then assigned a genomic strand based on the inferred strand of the majority of its constituting introns.

Integration of the polyA and SJ approaches

When an ROI could be assigned a genomic strand with both approaches, we found that the agreement between the two methods was 99.7%. Overall, 1,446,986 (70.5%, human) / 1,255,423 (67.1%, mouse) ROIs could be stranded (*i.e.*, assigned a genomic strand) based on at least one method (See Table 10). In rare cases of conflict, priority was given to the strand information obtained *via* the polyA method over the SJ one.

ROI-to-locus/biotype assignment

We assigned each mapped and stranded ROI an originating annotated locus by comparing PacBio mappings to the reference Gencode annotations, Gencode v.20 (human) and v.M3 (mouse), using the *bedtools intersect* program [1] with the following options: *-split* (ignore introns, *i.e.* only exonic overlaps were considered) *-s* (force strandedness) *-wao* (output overlapping entries from both files), only on exon records in both datasets.

Based on this data, we could then assign a unique annotation biotype to each ROI, based on overlapping GENCODE annotations, where available. In most cases, we used the original GENCODE gene_type attribute for this purpose. To simplify, however, ROIs overlapping loci of the following GENCODE gene types were tagged "*lncRNA*", though: "*antisense*", "*lincRNA*", "*processed_transcript*", "*sense_intronic*" and "*sense_overlapping*". When falling outside of GENCODE exonic regions, biotypes were attributed according to the type of capture probe the ROIs overlapped (*e.g.* enhancer, UCE, PipeR, *etc.*). As a last resort, ROIs falling outside of any GENCODE-annotated exon or probed element were tagged "*nonExonic*". When an ROI overlapped exons of multiple biotypes, it was flagged as "*multiBiotype*".

Construction of a HCGM set (High-Confidence ROI Genome Mappings)

We built a collection of High-Confidence ROI Genome Mappings from 1,000,213 (human) and 941,757 (mouse) genome-mapped, double-bounded UMD-ROIs. HCGMs were defined as follows:

- If spliced, read mappings can be composed only of canonical splice junctions (GT or GC as donor site, AG as acceptor site) over their entire mapped length,
- If unspliced, reads need to bear a detectable polyA tail, using the procedure explained in PolyA site calling.

Using these criteria, we identified a set of 771,585 (*i.e.*, 77% of genome-mapped, double-bounded UMD-ROIs) and 604,199 (*i.e.*, 64%) HCGMs in human and mouse, respectively (see Supplementary Figure 2k).

Sequencing error rate estimation

We evaluated the sequencing error rate of the CLS PacBio and HiSeq sequencing output with *qualimap BAMQC* (version 2.2.1)[3]. This software relies on SAM's NM and MD optional attributes for error rate calculations, therefore we re-mapped our reads as detailed in the Online Methods, adding the "*–outSAMattributes NM MD*" option to *STAR*'s [4] parameters.

Qualimap bamqc was run with default options on these BAM files, and the following information was extracted from each library's *genome_results.txt* reports: number of mapped bases, number of mismatches, number of insertions, and number of deletions. We then computed the mismatch, insertion and deletion rates per mapped base in each library (Supplementary Figure 2m). The global error rate was calculated as the sum of mismatches, insertions and deletions, divided by the total number of mapped bases. We observed that PacBio libraries had a ~ 2.1 times higher global error rate than HiSeq ones (1.37x10-3 *vs* 6.5x10-4 errors per mapped base on average, across all human and mouse samples). Both HiSeq and PacBio global error rates were mainly accounted for by sequence mismatches. As expected, non-demultiplexed PacBio reads were enriched in sequencing errors, which might explain why their sample barcode could not be identified in the first place.

Strikingly, PacBio reads were characterized by a much higher rate of insertions (7.5x10-5 vs 4.7x10-6 per mapped base on average, *i.e.* 16 times higher) and deletions (2x10-4 vs 1x10-5 per mapped base on average, *i.e.* 20 times higher) than their HiSeq counterparts. The relatively high rate of PacBio deletion errors casts some doubt on introns detected with this technology, and highlights the need for their systematic confirmation by HiSeq, which was performed in this study (see Extraction of Splice Junctions and Splice Sites, HiSeq support and novelty assessment below).

It should be noted that this analysis considers any sequence difference between RNA-Seq reads and the reference genome as a sequencing error, and therefore does not account for genuine, non-artefactual genome sequence vari-

ation. Thus, the error rates reported here may be slightly over-estimated for both HiSeq and PacBio reads.

Read merging and creation of a full-length lncRNA catalog

Read redundancy was reduced by merging transcript structures with compatible intron chains using the *compmerge* program (https://github.com/sdjebali/Compmerge). We used an original strategy, named "*anchored merging*", which consists in preventing reads with high-confidence boundaries - in our case, supported by a FANTOM true TSS at their 5' end (see Identification of high-confidence Transcription Start Sites using CAGE data below) and/or a captured, PacBio-encoded polyA site at their 3' end - from being merged into another longer read, regardless of their intron chain structure (see Figure 4b). The goal of this extra anchoring step is to preserve all transcript structures with high-confidence TSSs/3' ends, including those falling within exonic regions, which would be lost otherwise.

We anchored polyA- and CAGE-supported HCGMs before merging them using the *anchorTranscriptsEnds* software utility (https://github.com/julienlag/anchorTranscriptsEnds). First, we adjusted all high-confidence 5'/3' ends into clusters. That is, we merged close and overlapping sites using the *bedtools merge* utility, with a maximum clustering distance of 5 bases ("*-d 5*"), and forcing strandedness ("*-s*"). Each individual 5'/3' end belonging to a cluster was assigned its start/end coordinate, respectively - meaning that terminal exons were sometimes extended by a few nucleotides when necessary. In doing so, we ensured that within a cluster, all sites aligned at the exact same position. We subsequently added an "anchor" to all high-confidence, adjusted sites. This step consisted in attaching an artificial, biologically implausible chain of exons (*i.e.*, four 1 nucleotide-long exons, separated by 3 nucleotide-long introns) to each transcript model, upstream or downstream of its high-confidence 5' or 3' end, respectively. These false exons served as anchors to supported start and termination sites during the merging step, and were discarded immediately afterwards.

For comparison, we also performed a standard, "*non-anchored*" merging of HCGMs in parallel. The results of both strategies, across and within our interrogated tissues, are summarized in Supplementary Figure 11a. Following this merging step, we assigned a parent gene as well as a biotype to all merged transcript models (TMs), using the procedure described in ROI-to-locus/biotype assignment.

The end support - *i.e.*, by CAGE true TSS at the 5' end, and poly-adenylation at the 3' end - of each anchor-merged TMs was then deduced from the properties of its constituting ROIs, obtained from the procedures detailed in PolyA site calling and Identification of high-confidence Transcription Start Sites using CAGE data. Accordingly, the full-length set of TMs (referred to as "CLS_FL") consists only of models bounded by such high-confidence 5' and 3' ends. In addition, all their splice junctions are canonical, as they constitute a subset of HCGMs. The results of the read merging and selection of full-length transcript structures are detailed in Table 2, columns a-e.

The end support of transcript models merged using the standard (*i.e.* non-anchored) procedure was deduced not from their constituting ROIs', but rather, from the on-genome comparison of their end coordinates to CAGE TSSs and captured polyA sites (obtained with the methods described in PolyA site calling). 5'/3' ends were considered supported if they laid less than 20/5 bases away from a CAGE TSS / polyA site, respectively, and on the same genomic strand. The results of this comparison are summarized in Supplementary Figures 12c-d (second bar from the left).

Identification of high-confidence Transcription Start Sites using CAGE data

We used CAGE (Cap Analysis of Gene Expression) data produced by the FANTOM consortium [5] to single out high-confidence Transcription Start Sites (TSSs) in our mapped data. To do so, we compared the 5' ends of our HCGMs to the CAGE TSSs identified as "true" TSSs by FANTOM's TSS classifier (http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/TSS_classifier/TSSpredictionREADME.pdf) across FANTOM-interrogated tissues. The CAGE TSS files were downloaded from http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/TSS_classifier/ and lifted to hg38 and mm10 using the *liftOver* command-line tool (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/).

Captured TSSs were considered high-confidence (*i.e.*, CAGE-supported) if a FANTOM "true" TSS was found within a window of +/- 20 bases around it, on the same genomic strand (using *bedtools closest* with options "*-s -D b -t first -a <HCGM TSSs> -b <FANTOM true TSSs>*").

In addition, we analyzed the CAGE coverage of "known" (*i.e.*, sites falling within +/- 50 bases of a GENCODE-annotated TSS on the same strand) and "novel" (*i.e.*, sites falling more than 50 bases away from of a GENCODE-annotated TSS on the same strand) PacBio TSSs separately. To do so, for each non-redundant TSS (obtained using *bedtools merge -n -s -d 5*) of the two populations, we computed the distance to the closest FANTOM "true" TSS (using *bedtools closest* with options "*-s -D b -t first -a <HCGM TSSs> -b <FANTOM true TSSs>*").

We observed that novel PacBio TSSs far outnumber known ones in both species (200,425 vs 44,736 in human, 155,083 vs 32,230 in mouse, respectively, see Supplementary Figure 10e). While the CAGE coverage of known sites was higher, thousands of novel TSSs found a CAGE cluster in their close vicinity (+/- 50 bases on the same genomic strand, see Table 11).

Splice Junction analysis

Extraction of Splice Junctions and Splice Sites, HiSeq support and novelty assessment

PacBio Splice Junctions (SJs) were gathered from HCGMs (see Construction of a HCGM set (High-Confidence ROI Genome Mappings)), and as such, they were all canonical (GT|GC / AG). They were assigned a biotype based on that of their originating reads (see ROI-to-locus/biotype assignment). The *IPSA* suite [6] (*Integrative Pipeline for Splicing Analyses*, https://github.com/pervouchine/ipsa-full) was employed to extract SJs and their read counts from *STAR* [4] alignments of Illumina HiSeq data. *IPSA* was run with the default parameters. GENCODE versions 20 and M3 were used as a reference for human and mouse, respectively. All operations were performed on a non-redundant set of distinct SJs, which were uniquely identified by their chromosome, start/end coordinates, and genomic strand. A PacBio SJ was defined as HiSeq-supported if the exact same intron was also observed in the post-capture HiSeq data. HiSeq SJ support was also computed at the level of entire merged CLS transcript models (TMs). Overall, 86.5 % (human) / 87% (mouse) of TMs displayed HiSeq support of their complete intron chain (Table 3). This rate amounted to 91% when considering full-length TMs only.

We proceeded similarly when comparing PacBio SJs to GENCODE-annotated introns: they were flagged as "*known*" when an exact equivalent was found in the comprehensive GENCODE set, and "*novel*" otherwise. The "*known/novel*" status of each SJ was also propagated to its constituting donor and acceptor splice sites (SSs).

A comparison of captured SJs to the human *miTranscriptome* catalog [7] was also performed. We downloaded the GTF data (version 2) from http://mitranscriptome.org/download/mitranscriptome.gtf.tar.gz, converted it to BED,

and mapped its original GRCh37 (hg19) coordinates to GRCh38 (hg38) using *liftOver* (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/). At this stage, 377,382 of 384,066 transcripts (98.3%) were successfully lifted over. A biotype was then assigned to these transcripts, following the procedure described in ROI-to-locus/biotype assignment, and we kept only those models of biotype *lncRNA* that overlapped a CLS-probed genomic region (N=32,502). 35,582 unique, canonical (GT|GC / AG) SJs were subsequently extracted and compared to human CLS (Supplementary Figure 6c, top panel). The same analysis was performed on the union of GENCODE 20 and miTranscriptome SJs (N= 42,698) within probed lncRNA regions (Supplementary Figure 6c, bottom panel).

Analysis of splicing motifs

We analyzed PacBio donor and acceptor splice sites (SS) separately. We employed the *geneid* software [8] (version 1.4, with options *-a -d -G -P <parameter file>*) to score individual sites using Position Weight Matrices computed on annotated human genes (parameter file available at https://public_docs.crg.es/rguigo/CLS/data/human.param.Feb_22_2006_GC). The score calculated by *geneid* for a given site *S* corresponds to the log-likelihood ratio of *S* in an actual SS *vs.* *S* in a false SS. We built control ("*Random*") sets of splice sites separately for donor and acceptor sites. To do so, we selected all putative splice sites (GT and GC for donors, and AG for acceptors) within genomic regions overlapped by introns or exons of HCGMs. We then filtered out any site observed as spliced in GENCODE or our PacBio SJs, and scored the remainder with *geneid*, as explained above.

Human-mouse evolutionary conservation of splice sites

The conservation of HiSeq-supported PacBio splice sites between human and mouse was analyzed by mapping "strong" SSs (namely, SSs with positive *geneid* scores, see Analysis of splicing motifs) from one species to the other using *liftOver* (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/) with "*reciprocal best*" alignment chains (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/vsMm10/reciprocalBest/hg38.mm10.rbest.chain.gz and http://hgdownload.cse.ucsc.edu/goldenPath/mm10/vsHg38/reciprocalBest/mm10.hg38.rbest.chain.gz for human and mouse, respectively). Supplementary Figure 7d-e summarizes the results of the SS *liftOver* step.

A subset of the "*Random*" collection of splice sites (described in Analysis of splicing motifs, and matched for *geneid* scores) was used as a control. This sample was produced by using the *matchDistribution* tool (https://github.com/julienlag/matchDistribution, commit version "*a72706a*", 500 sequential passes, 500 bins, with default options) with the *geneid* score distribution of "*Random*" sites as the "subject" set, and the *geneid* scores of the union of protein-coding and lncRNA sites as the "target" distribution. The result was a sample of "*Random*", unspliced sites matching the splicing strength of both lncRNA and protein-coding SSs in our data (see Supplementary Figures 7f-g).

After mapping high-strength SSs from those three collections from one species' genome to the other, we counted the number of orthologous sites in the destination genome that were also scored positively by *geneid*, as explained in Analysis of splicing motifs. We observed that although much weaker than that of protein-coding sites, the conservation of lncRNA SS strength was overall significantly above background (Chi-square test of conserved/non-conserved sites compared to "*Random*" sites) (see Supplementary Figure 7c).

Intron retention

Intron retention (IR) rates were calculated using *bedtools intersect* [1] on the CLS and GENCODE transcript sets. Note that CLS transcript models whose entire set of introns was not HiSeq-supported (see Extraction of Splice Junctions and Splice Sites, HiSeq support and novelty assessment) were ignored in this analysis. An intron was considered retained if its boundaries were fully contained within at least one individual exon of the same transcript set (*bedtools* option: "*-f1*") and on the same strand (*bedtools* option: "*-s*"). The IR rates

reported in Supplementary Figure 6f were calculated as the proportion of transcripts with a least one intron retained.

Identification of novel transcript structures

We used the *comptr* program (https://github.com/sdjebali/Comptr) to compare the intron chains of our merged TMs (the "*assessed*" set, obtained as described in Read merging and creation of a full-length lncRNA catalog) to the comprehensive set of GENCODE 20 and M3 transcripts (the "*reference*" set). We considered novel only the transcripts categorized by *comptr* as "*Extension*" (*i.e.*, there is a reference transcript with all its introns equal to the assessed transcript but the assessed transcript has additional introns), "*Intergenic_or_antisense*" (*i.e.*, the assessed transcript is stranded and spliced but does not overlap any reference transcript on the same genomic strand) and "*Overlap*" (*i.e.*, there is a reference transcript overlapped by the assessed transcript on the same strand). The results of the full-length TM structure comparison are summarized in Figure 4e-f, and detailed in Table 2, which also reports the number of annotated loci giving rise to novel FL structures.

Simulated read depth versus discovery rate

In order to evaluate the completeness of our post-capture annotation - *i.e.*, how close it is to saturation - we calculated the number of novel SJs and novel transcript structures discovered at increasing ROI sequencing depths in each tissue sample. We randomly sampled ROIs from unfiltered BAM files (that is, including unmapped ROIs) at increasing depths, in increments of 20,000 reads until the total of available reads was reached in each tissue sample. A combination of samtools [9] and standard GNU Linux utilities (*head*, *shuf*) was employed for that purpose.

We then counted the number of novel individual SJs (procedure described in Extraction of Splice Junctions and Splice Sites, HiSeq support and novelty assessment), or novel intron chains (see Identification of novel transcript structures) the sampled ROIs gave rise to. Each randomization at a given read depth was repeated 50/100 times for individual SJ and full intron chain simulations, respectively. When assessing the novelty of individual introns, we stratified the SJs generated at each read depth by level of sequencing support (all PacBio junctions, PacBio junctions with HiSeq support, and PacBio junctions without HiSeq support). Results of the simulations are presented in Figure 3d, as well as Supplementary Figure 8a-b. Results of the simulations using HiSeq short-read sequencing of captured cDNA are reported in Supplementary Figure 8c.

Analysis of protein-coding potential

The set of full-length transcript models were used as input for the programs CPAT [10] and PhyloCSF [11]. CPAT uses intrinsic sequence properties to predict coding potential. PhyloCSF, in contrast, uses evolutionary signatures of selection on coding sequences.

CPAT was run according to creator's protocol [10]. Hexamer tables and logit models were created using the *Human_ORF.fa* and *Human_NONCODE.fa* files and used for both human and mouse analyses. We used the cutoff value of 0.364 to distinguish coding from noncoding transcripts.

PhyloCSF was run on spliced alignments of transcripts using a custom pipeline. GTF format annotations were used to extract multiple alignment file (MAF, obtained from the UCSC Genome Browser for hg38) format blocks, which were stitched together to recreate the multiple alignment of the processed transcript. This was converted to a FASTA file and input to PhyloCSF. The parameter file used was "*29 mammals*". The settings used were "*–dna –aa –frame=3*

–removeRefGaps –orf=ATGStop –minCodons=20". A score threshold of 100 was used to define protein-coding transcripts.

Analysis of cytoplasmic/nuclear localization

PolyA RNA sequencing data from whole cell, nucleus and cytoplasm of ten human cell lines was obtained from the ENCODE portal (https://www.encodeproject.org/) [12, 13]. An annotation in GTF format was constructed by combining the annotation of merged transcript models with GENCODE v20. The *GRAPE* pipeline [14] (https://github.com/guigolab/grape-nf) was used to quantify these models, using *STAR* [4] (v.2.4.0j) with RSEM [15] (v.1.2.21) for quantification. The following non-default parameter was specified for *STAR*: "*–outFilterMismatchNmax 4*". *RSEM* was used in transcriptome mode. Next, for every full length CLS model, the log2 ratio was calculated of cytoplasmic to nuclear RPKM values. Any transcript model with a zero value in either compartment was discarded.

Evaluation of Illumina-based transcript reconstruction methods in matched samples

Global assessment of reconstruction software accuracy

We comprehensively assessed the accuracy of the short-read transcript reconstruction algorithms *StringTie* [16] and *Cufflinks* [17] using CLS transcript models (TMs) in matched samples as a gold standard. Capture HiSeq reads were mapped to the corresponding reference genome (hg38 and mm10) using *STAR*, as described in the Online Methods, adding the "*–outSAMattributes XS*" parameter, in order to comply with *StringTie* and *Cufflinks* requirements when used with unstranded reads. Reads mapping to ERCC spike-in sequences were discarded.

StringTie (v1.3.3) was run with default parameters except "*-p6*" (*i.e.*, 6 CPU threads), and *Cufflinks* (v2.2.1) with the "*–multi-read-correct*" and "*-p6*" options. Running on human data, *Cufflinks* hung for more than 10 days on a single region (chr12:65981298-65981423 in hg38), and thus had to be restarted with the offending region masked.

StringTie and *Cufflinks*'s respective outputs were then compared to the full set of 94,163 (human) / 72,413 (mouse) HiSeq-supported, standard-merged CLS TMs (see Read merging and creation of a full-length lncRNA catalog and Extraction of Splice Junctions and Splice Sites, HiSeq support and novelty assessment) as the reference annotation file. We obtained the corresponding sensitivity and precision measures using *gffcompare* (v0.9.9c) (https://github.com/gpertea/gffcompare), run with options "*-N*" and "*-M*" (*i.e.*, ignore single-exon transcripts and transfrags in both reference and test sets). While fair at the "base" and "intron" levels, sensitivity and precision were particularly poor at the "intron chain" and "transcript" levels for both programs (Supplementary Figure 12a). *StringTie* substantially and consistently outperformed *Cufflinks* at all accuracy levels, and we therefore decided to further analyze only *StringTie* models for the sake of simplicity.

StringTie produced a total of 94,082 (human) and 171,439 (mouse) distinct transcripts, merged across all assayed tissues in each organism. Of those, 65,060 (human, *i.e.* 69%) and 52,412 (mouse, *i.e.* 31%) could be assigned a genomic strand by the program (all unstranded models were single-exon transfrags, and were ignored in the rest of the analysis).

Following the procedure used for CLS models (see ROI-to-locus/biotype assignment), we found that 13,930 (human) and 2,920 (mouse) stranded *StringTie* models originated from probed lncRNA genomic regions. We then extracted intron chains from these models and performed a 3-way comparison with CLS and GENCODE transcript models falling within targeted lncRNA regions (Figure 4h and Supplementary Figure 12b). Spliced length statistics of these *StringTie*

TMs are presented in Figure 4i (human) and Supplementary Figure 12c (mouse), side-by-side with the CLS TM set.

End support of CLS and *StringTie*-reconstructed transcripts

We analyzed the end support of stranded *StringTie* models and compared it to CLS TMs. Since polyA tails are not preserved in Illumina-based reconstructed transcripts, we independently called polyA sites by applying the method described in PolyA site calling to uniquely mapped capture HiSeq reads.

Requiring a minimum of 2 reads supporting a given site, we could detect only 2,572 (human) / 2,278 (mouse) distinct polyA sites, that is, 14 (human) / 12 (mouse) times less than when using PacBio ROIs in matched samples with the same parameters (Table 1). We attribute this lack of sensitivity to the well-documented relative depletion of HiSeq reads towards the ends of transcripts [18].

Failure to properly resolve polyA sites using HiSeq data led us to evaluate *StringTie* models' 3' end completeness with CLS-called polyA sites instead. The full-length status assessment of *StringTie*-reconstructed transcripts was, as a result, performed exactly as for standard (*i.e.*, non-anchored) -merged CLS TMs (see Read merging and creation of a full-length lncRNA catalog), for a fair comparison. Only 4,633 (*i.e.* 7%, human) / 3,646 (*i.e.* 7%, mouse) *StringTie* transcripts were considered full-length using our criteria, a much lower rate than the one observed in the standard-merged CLS TM set (28,186, *i.e.* 24% in human, and 18,934, *i.e.* 22% in mouse, see Supplementary Figure 12d-e for a side-by-side comparison). The fraction of full-length *StringTie* TMs decreased immensely within probed lncRNA regions (116/13,930, *i.e.* 0.8% in human, and 32/2,920, *i.e.* 1.1% in mouse).

The 5' and 3' completeness of *StringTie* and CLS TMs were further analyzed and compared with the following datasets: GENCODE lncRNAs (5' and 3' ends from annotated lncRNAs originating *StringTie* or CLS TMs), GENCODE protein-coding transcripts (a confident set of protein-coding transcripts, not tagged *mRNA_end_NF* nor *mRNA_start_NF* in the original GENCODE GTF files), and a control set of sites (middle coordinate of internal exons, as described in Proximity of polyA signals). All sets of sites were individually clustered to reduce redundancy ("*bedtools merge -d 5 -s*").

We then assessed the proximity of each TSS to FANTOM5 CAGE true TSSs (as described in Identification of high-confidence Transcription Start Sites using CAGE data), and of each 3' end to canonical polyA signal motifs (PAS, as described in Proximity of polyA signals) using a combination of *bedtools slop* and *bedtools intersect* [1]. We considered a TSS CAGE-supported ("CAGE(+)") if a CAGE cluster could be found +/-50 bases around them, on the same strand. Similarly, "PAS(+)" 3' ends are those CLS polyA sites falling 10 to 50 bases downstream of a PAS motif (Figure 4i and Supplementary Figure 12f).

In addition, we present aggregate plots of PAS and CAGE TSSs around various sets of transcript ends in Supplementary Figure 12g-h.

Genome repeat coverage

Repeat elements in both mm10 and hg38 were downloaded from the UCSC Table Browser (`https://genome.ucsc.edu/cgi-bin/hgTables`, *RepeatMasker* track) in tabular format. Repeat features were split into different classes according to their *repClass* attribute, converted to BED format and projected on the genome using *bedtools merge* [1]. Next, repeat elements were compared to projected exons from the CLS, GENCODE and *StringTie* TM sets of all biotypes, using *bedtools coverage*. Only stranded models were considered in the *StringTie* TM set. The fraction of exonic nucleotides covering genome repeats of various classes in each set of TMs is reported in Supplementary Figure 12i.

Estimating capture sensitivity using spike-ins

Inspection of sensitivity curves for spike-ins in individual tissues (Supplementary Figure 1e) shows a detection threshold around 5.6x10-2 attomol (-1.25 in log10 units) for captured molecules. In 4 µg of a 1:100 dilution of spike-in RNA that was added to 4 µg of each RNA sample, this threshold is equivalent to 1344 molecules. We assume here that the total RNA content of a single cell is 5 pg [19], making this threshold equate to 7x10-3 molecules per cell. Non-captured sequences' detection threshold lies approximately 30-fold higher (1.5 log10 units), or 0.21 molecules per cell.

TSS overlap analysis

Coordinates of indicated features were downloaded and mapped to hg38 using *liftOver* (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/) where appropriate. CpG islands were downloaded from the UCSC Table Browser (hg38) (https://genome.ucsc.edu/cgi-bin/hgTables).

Promoter and enhancer maps were downloaded as all files corresponding to 15-state *ChromHMM* predictions brain, heart and liver from Epigenome Roadmap [20]. These coordinates were merged separately by promoter (states: 1, 2, 3, 4, 14) / enhancer predictions (states: 7, 8, 9, 10, 11, 15). GWAS SNPs (gwas_catalog_v1.0-associations_e84_r2016-05-08.tsv) were obtained from the GWAS Catalog [21]. Conserved elements, obtained using *PhastCons* [22] 46-way primate alignments, were downloaded from the UCSC Genome Browser (hg38).

Comparison of human TSSs with DNase-Seq (DHS), ChIP-Seq and conservation tracks

Input datasets

We compared various sets of human TSSs to ENCODE ChIP-Seq (in cell lines K562 and HeLa, see "signal" BigWig file list in Table 12), DNase-Seq (HeLa-S3 DNase Hypersensitive Sites hotspots, downloaded from https://www.encodeproject.org/files/ENCFF968ECA/) and conservation (*phastCons* [22] scores downloaded from http://hgdownload-test.cse.ucsc.edu/goldenPath/hg38/) tracks. At the time of the study, the relevant ENCODE ChIP-Seq signal files were only available on human assembly hg19, therefore we mapped our TSSs to this genome version using *liftOver* (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/) before proceeding with the ChIP-Seq comparison.

The three collections of transcripts used as input were "*CLS_FL lncRNA*" (described in Read merging and creation of a full-length lncRNA catalog), "*GENCODE lncRNA*" (*i.e.*, GENCODE-annotated lncRNAs detected by CLS), and "*GENCODE protein-coding*" (defined in Transcript expression matching of the GENCODE protein-coding set below). The TSS sets analyzed consisted in a "*standalone*" version, where transcripts originating from bi-directional promoters were filtered out. These were generated by removing all transcripts whose TSS fell within 1,000 bases upstream of any GENCODE or captured FL TSS on the opposite genomic strand. Given the fuzzy nature of ChIP-Seq and DNase-Seq peaks, we merged TSSs within each set using a rather large maximum clustering distance of 200 bases ("*bedtools merge -s -d 200*" [1]).

The basic characteristics of each TSS dataset used in the ChIP-Seq and conservation analyses are reported in Table 13.

Transcript expression matching of the GENCODE protein-coding set

We selected a subset of protein-coding transcripts with expression similar to that of CLS_FL lncRNAs in K562 and HeLa cells. First, we merged the

CLS_FL transcript models with GENCODE v.20, then quantified the resulting transcripts in K562 and HeLaS3 ENCODE polyA+ whole-cell RNA-Seq experiments *ENCSR000CPH* and *ENCSR000CPR* (downloaded from the ENCODE portal, https://www.encodeproject.org), respectively.

The transcript quantifications were computed using *GRAPE* [14] (https://github.com/guigolab/grape-nf, commit version "*bcaa6688b9*", bundling *STAR* [4] version 2.4.0j and *RSEM* [15] version 1.2.21) running under the *NextFlow* framework [23] (version 0.17.3, https://www.nextflow.io/).

We then extracted the posterior mean estimate FPKM ("*pme_FPKM*") values of CLS_FL lncRNA and GENCODE protein-coding transcripts from *RSEM* output. The subset of expression-matched GENCODE protein-coding transcripts was obtained using the *matchDistribution* software tool (https://github.com/julienlag/matchDistribution, commit version "*a72706a*", 10 sequential passes, 500 bins, with option "*–transform=log10*") with GENCODE protein-coding *pme_FPKM*s as the "*subject*" set, and CLS_FL lncRNA as the "*target*" distribution, separately on K562 and HeLaS3.

The results of the expression matching are presented in Supplementary Figure 13f-g and Table 13.

Aggregate plots of signal density surrounding TSSs

We employed *bwtool* [24] (https://github.com/CRG-Barcelona/bwtool) to produce aggregate plots of ChIP-Seq read density and conservation scores on the aforementioned TSS collections, using the following command: "*bwtool agg -long-form -header -expanded -firstbase 10000:10000 <TSS set> <signal BigWig file> output.txt*".

The mean signal and standard error of the mean were extracted from *bwtool*'s output and plotted as a function of the nucleotide position around TSSs (Supplementary Figures 14 and 15).

Comparison of TSSs and DNase Hypersensitive Sites (DHS) in HeLa cells

HeLa-S3 DHS peak ("hotspot") BED files were first converted to the BigWig format using "*bedtools genomecov -bga*" [1] followed by *bedGraphToBigWig* (http://hgdownload.soe.ucsc.edu/admin/exe/).

The resulting BigWig files were subsequently compared only to the "*raw*" merged TSSs of transcripts detected in HeLa by CLS, using "*bwtool agg -long-form -header -expanded -firstbase 10000:10000*" to obtain the mean DHS hotspot density at each base surrounding TSSs. The "*GENCODE protein-coding*" set was "expression-matched" in HeLa-S3, as described in Transcript expression matching of the GENCODE protein-coding set. The "*GENCODE lncRNA*" set consisted in GENCODE v.20 lncRNA transcripts detected by CLS in HeLa-S3.

Testing predicted peptides

Using a published proteogenomics workflow [25], we searched the Human Proteome Project (C-HPP) [26] database of testis peptides for those matching predicted ORFs, but found no hits at a threshold of 0.01 PEP.

Identifying lncRNA orthologues

We defined orthology using *MultiZ* sequence alignments [27]. Taking the entire genomic span of GENCODE lncRNA gene annotations, we created human-to-mouse and mouse-to-human orthology mappings using *liftOver* (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/) with chain files from the UCSC Genome Browser (hg38 -> mm10, mm10 -> hg38). *liftOver* was run with "*-minMatch=0.8*" (minimum fraction of nucleotides mapped), and all other options set to default.

Orthology was then defined using strand-specific intersection, requiring a minimum of 5% of the genomic span of both elements to overlap. Orthologue lists were now defined in two ways: "*Reciprocal*" requires reciprocal mapping in both directions (131 pairs where both have an ID from GENCODE 20 / M3); "*Union*" is the union of both directions mappings, without requiring reciprocal hits (293). Note that these numbers refer to the entire lncRNA annotation. The subset of these reciprocal pairs was then obtained that map to the probed lncRNA annotation in either species. We proceeded with the larger *Union* set to boost statistical significance, being 101/84 orthologous and probed genes, for human and mouse respectively. This set we defined as "conserved" having an orthologue in the other species. Amongst these were cases such as *SNHG17* and *MALAT1*.

We asked whether conserved lncRNAs were more likely to be detected (having >0 mapping reads, using either all reads or only full-length reads). We also compared the number of reads between conserved and non-conserved lncRNAs. Both analyses were performed separately in each species, and presented in Supplementary Figure 11d-g.

RT-PCR experimental validation of CLS transcript models

500 ng of total RNA from HeLa, brain and testis samples (the same we used for the capture assays) were used for retrotranscription. Retrotranscription was performed with ReverseAid retrotranscriptase (Thermo Scientific), using both oligo-dT and random hexamers as primers, following the manufacturer's protocol.

For CARMN and KANTR, testis and brain cDNAs, respectively, were amplified for 40 cycles at 56°C annealing. For CASC19, HeLa cDNA was amplified for 40 cycles at 56°C annealing, gel purified and amplified for 40 more cycles to enrich for specific bands. The SAMMSON transcript was amplified from testis, for 40 cycles at 54°C annealing with Expand polymerase.

PCRs were performed with KOD DNA Polymerase (Novagen) using primers to be found in Supplementary Data 3. The amplicons were sequenced using Sanger sequencing and are available in Supplementary Data 4.

Bibliography

1. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics* **47,** 11.12.1–11.12.34. ISSN: 1934-340X (2014).
2. Lopez, F, Granjeaud, S, Ara, T, Ghattas, B & Gautheret, D. The disparate nature of "intergenic" polyadenylation sites. *RNA* **12,** 1794–1801 (2006).
3. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32,** 292–294. ISSN: 1367-4803 (2016).
4. Dobin, A *et al.* STAR: ultrafast universal RNA-seq aligner. eng. *Bioinformatics* **29,** 15–21 (2013).
5. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507,** 462–70. ISSN: 1476-4687 (2014).
6. Pervouchine, D. D., Knowles, D. G. & Guigó, R. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29,** 273–274. ISSN: 13674803 (2013).
7. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47,** 199–208. ISSN: 1061-4036 (2015).
8. Blanco, E., Parra, G. & Guigó, R. in *Curr. Protoc. Bioinforma.* Unit 4.3 (John Wiley and Sons, Inc., Hoboken, NJ, USA, 2007). doi:10.1002/0471250953.bi0403s18.
9. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079. ISSN: 1367-4803 (2009).
10. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41,** e74–e74. ISSN: 1362-4962 (2013).
11. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27,** i275–i282. ISSN: 1367-4803 (2011).
12. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–8. ISSN: 1476-4687 (2012).
13. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 1476-4687 (2012).
14. Knowles, D. G., Roder, M., Merkel, A. & Guigo, R. Grape RNA-Seq analysis pipeline environment. *Bioinformatics* **29,** 614–621. ISSN: 1367-4803 (2013).
15. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323. ISSN: 1471-2105 (2011).
16. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33,** 290–295. ISSN: 1087-0156 (2015).
17. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. eng. *Nat Protoc* **7,** 562–578. ISSN: 1754-2189 (2012).
18. Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.* **32,** 915–925. ISSN: 1087-0156 (2014).

19. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21,** 1160–1167. ISSN: 1088-9051 (2011).

20. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330. ISSN: 0028-0836 (2015).

21. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901. ISSN: 0305-1048 (2017).

22. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050. ISSN: 1088-9051 (2005).

23. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35,** 316–319. ISSN: 1546-1696 (2017).

24. Pohl, A. & Beato, M. bwtool: a tool for bigWig files. *Bioinformatics* **30,** 1618–9. ISSN: 1367-4811 (2014).

25. Wright, J. C. *et al.* Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat. Commun.* **7,** 11778. ISSN: 2041-1723 (2016).

26. Zhang, Y. *et al.* Tissue-Based Proteogenomics Reveals that Human Testis Endows Plentiful Missing Proteins. *J. Proteome Res.* **14,** 3583–94. ISSN: 1535-3907 (2015).

27. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–15. ISSN: 1088-9051 (2004).

III

Towards a complete map of the human long non-coding RNA transcriptome

Uszczynska-Ratajczak B, **Lagarde J**, Frankish A, Guigó R, Johnson R. *Towards a complete map of the human long non-coding RNA transcriptome*. ***Nature Reviews Genetics*** 2018 Sep; 19(9):535-548.

URL: https://doi.org/10.1038/s41576-018-0017-y

Abstract:

Gene maps, or annotations, enable us to navigate the functional landscape of our genome. They are a resource upon which virtually all studies depend, from single-gene to genome-wide scales and from basic molecular biology to medical genetics. Yet present-day annotations suffer from trade-offs between quality and size, with serious but often unappreciated consequences for downstream studies. This is particularly true for long non-coding RNAs (lncRNAs), which are poorly characterized compared to protein-coding genes. Long-read sequencing technologies promise to improve current annotations, paving the way towards a complete annotation of lncRNAs expressed throughout a human lifetime.

III.1. Main article

REVIEWS

NON-CODING RNA

Towards a complete map of the human long non-coding RNA transcriptome

*Barbara Uszczynska-Ratajczak[1], Julien Lagarde[2,3], Adam Frankish[4], Roderic Guigó[2,3] and Rory Johnson[5,6]**

Abstract | Gene maps, or annotations, enable us to navigate the functional landscape of our genome. They are a resource upon which virtually all studies depend, from single-gene to genome-wide scales and from basic molecular biology to medical genetics. Yet present-day annotations suffer from trade-offs between quality and size, with serious but often unappreciated consequences for downstream studies. This is particularly true for long non-coding RNAs (lncRNAs), which are poorly characterized compared to protein-coding genes. Long-read sequencing technologies promise to improve current annotations, paving the way towards a complete annotation of lncRNAs expressed throughout a human lifetime.

Long non-coding RNAs (lncRNAs). RNA transcripts ≥200 nucleotides long that do not encode any identifiable peptide product.

A fundamental goal of biology is to understand how the instructions to create and maintain an organism are encoded in its DNA sequence. From worm to man, the genomes of different species house remarkably similar numbers of protein-coding genes[1], prompting the notion that many aspects of complex organisms arise from non-protein-coding regions. These non-coding regions comprise a rich diversity of regulatory and functional units, amongst the most numerous of which are loci encoding long non-coding RNAs (lncRNAs)[2]. Next-generation sequencing has identified tens of thousands of lncRNA loci, from single-celled eukaryotes to humans[3]. The sequences of lncRNAs are under purifying evolutionary selection[4,5], and a substantial fraction yield clear phenotypic effects in both in vitro and in vivo loss of function studies[6–10]. Growing numbers of lncRNAs have been linked to human diseases[11]. However, their functionality remains contentious[12], and the number of experimentally characterized or disease-associated lncRNAs lies in the hundreds, or ≤1% of identified loci[13].

Closing this gulf between mapped and experimentally validated lncRNAs has prompted functional studies of growing scope. These studies have depended on the development of the fundamental resource of annotations, which describe the genomic locations, sequences and exon structure of lncRNA transcripts. As the basis of microarray designs, early lncRNA annotations enabled researchers to perform the first generation of functional genomics studies, implicating lncRNAs in processes as diverse as embryonic stem cell pluripotency[14], reprogramming[15], tumour suppression[16], neuronal differentiation[17] and cardiac differentiation[18]. More recently, large-scale functional screens based on the CRISPR–Cas system have been applied to hundreds or thousands of lncRNAs in a single experiment[19].

Several different annotations exist for the human genome (TABLE 1), each with advantages and drawbacks that might not be immediately evident. They are based on two principal strategies of automated and manual annotation. Automated annotation typically employs transcriptome assembly approaches that are rapid and inexpensive but produce incomplete and inaccurate annotations. Manual annotation yields high-quality catalogues but at slow rates and requiring substantial long-term economic support. Both approaches suffer from a variety of deficiencies that are important for end users to understand.

Recent technical developments promise to revolutionize annotation methods. Third-generation sequencing technologies are capable of reading entire RNA or cDNA molecules. Combined with methods to capture desired transcripts, third-generation sequencing promises to extend and improve existing lncRNA annotations rapidly and cost-effectively. These advances make it feasible to envisage the eventual complete annotation of the genome, whereby the entirety of biologically relevant genes, transcripts and exons is catalogued in all cell types throughout the human lifespan. A key subsidiary aim will be to define what threshold constitutes biological relevance and hence whether expression (or other) thresholds should be used for inclusion in final annotations[20].

This Review has two main objectives. The first is to provide an overview of the current state of lncRNA annotations: how they are created, how good they are, best practice in their use, and the development of quantitative standards by which they might be evaluated and compared. The second is to discuss how emerging

[1]*Centre of New Technologies, University of Warsaw, Warsaw, Poland.*
[2]*Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain.*
[3]*Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain.*
[4]*European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.*
[5]*Department of Medical Oncology, Inselspital, University Hospital and University of Bern, Bern, Switzerland.*
[6]*Department of Biomedical Research (DBMR), University of Bern, Bern, Switzerland.*
* *e-mail: rory.johnson@dbmr.unibe.ch*

https://doi.org/10.1038/s41576-018-0001-7-y

Table 1 | **lncRNA annotations**

Name (version)	Reported size (gene loci)	Methods[a]	Comments	Completeness	Comprehensiveness[b]	Exhaustiveness[c]
NONCODE (v5)	96,308	Integration of other databases	The most comprehensive resource	8.9%	67,276	2.3
MiTranscriptome (v2)	63,615	Assembly from short reads	Mainly cancer samples	4.4%	45,088	4.4
FANTOM CAT (v1)	27,919	Assembly, other annotations and CAGE evidence	Mapped 5′ ends using CAGE tags	15.8%	27,278	3.3
RefSeq (GCF_000001405.37_GRCh38.p11)	15,791	Manual (based on cDNA) and automated annotation (based on RNA-seq data)	The oldest annotation	11.0%	14,889	1.9
GENCODE (v27)	15,778	Manual annotation based on cDNA, ESTs and high-quality long-read data	Used by most consortia and integrated with Ensembl	13.5%	15,063	1.9
BIGTranscriptome (v1)	14,158	Assembly, with CAGE and 3P-seq evidence	Full-length transcripts	27.7%	12,632	2.1
GENCODE+	13,434	Union of GENCODE (v20) and CLS lncRNAs with anchor-merged CLS transcript models	Extension of GENCODE by CLS	24.0%	13,434	3.3
CLS FL	807	lncRNAs from GENCODE+ with CAGE and poly(A) evidence	Full-length transcripts	71.7%	807	5.5
Protein-coding[d]	19,502	GENCODE confident protein-coding transcripts	Not tagged *mRNA_end_NF* nor *mRNA_start_NF* in the original GENCODE v27 GTF file	53.8%	18,995	2.9

All numbers correct as of the end of 2017. MiTranscriptome, Functional Annotation of the Mammalian genome (FANTOM) cap analysis of gene expression (CAGE)-associated transcriptome (CAT) and BIGTranscriptome long non-coding RNA (lncRNA) catalogues were lifted over to the Genome Reference Consortium Human Build 38 (GRCh38) genome assembly. 3P-seq, poly(A)-position profiling by sequencing; CLS, capture long-read sequencing; EST, expressed sequence tag; RNA-seq, RNA sequencing. [a]Assembly in the Methods column refers to transcriptome assembly using short reads from RNA-seq. [b]Comprehensiveness is the total number of gene loci boundaries defined using buildLoci. To compare gene sets in a consistent way, the assembly patches were excluded, and the gene loci boundaries were redefined using buildLoci, which explains discrepancies between gene numbers presented here and those reported in original publications. [c]Exhaustiveness is the average number of isoforms per gene locus. Figures for completeness, comprehensiveness and exhaustiveness as presented in FIG. 3 are shown here. [d]A set of protein-coding transcripts was used as a reference.

technologies will have an impact on these annotations and may alter our understanding of what constitutes the human lncRNA transcriptome. Although we focus mainly on human studies, the following discussions are of relevance to other model and non-model organisms. Of note, the lncRNAs discussed here are almost exclusively those of the polyadenylated (polyA+) fraction, owing to the fact that most transcriptomic surveys have been performed on conventional, oligo-dT-primed cDNA. The universe of polyA− lncRNAs remains largely unexplored and may hold many functional molecules[21].

lncRNA annotations: a research foundation

Structure of lncRNA annotations and biotypes. Annotations, whether of protein-coding or lncRNA-encoding genes, are hierarchical: they are composed of gene loci, each of which is composed of one or more partially overlapping transcripts, themselves composed of one or more exons (FIG. 1a). In the absence of a clear understanding of their sequence–structure–function relationship, lncRNAs have tended to be classified by their genomic organization, in other words, the relationship of their encoding locus to the nearest protein-coding gene (FIG. 1b). In the context of genome annotation, this can be used as a biotype label. The principal dichotomy of genomic organization is genic versus intergenic, or lncRNAs that overlap or do not overlap a protein-coding gene, respectively. The latter are also referred to as long intergenic non-coding RNAs (lincRNAs). Genic lncRNAs may be subdivided by the precise nature of their overlap with the protein-coding gene, and there is some evidence for distinct functions and features between these classes[22]. By numbers, lncRNAs tend to be approximately equally divided into genic and intergenic classes.

Why are lncRNAs difficult to annotate? lncRNA annotations lag considerably behind those of protein-coding genes, for reasons that go beyond their more recent discovery. There are at least three factors that make lncRNA annotation challenging. First, lncRNAs are relatively lowly expressed, meaning that their transcripts will be weakly sampled in any unbiased transcriptomic data, including expressed sequence tags (ESTs), RNA sequencing (RNA-seq) and cap analysis of gene expression

Annotation
Catalogue of gene loci comprising detailed and hierarchical information on their genomic coordinates and that of their constituent transcript isoforms and exons, all of which are assigned unique and stable identifiers.

Transcriptome assembly
The use of bioinformatic algorithms to reconstruct gene and transcript models based on short sequence reads.

Manual annotation
The creation of gene and transcript models by human annotators based on RNA and protein evidence and according to defined protocols.

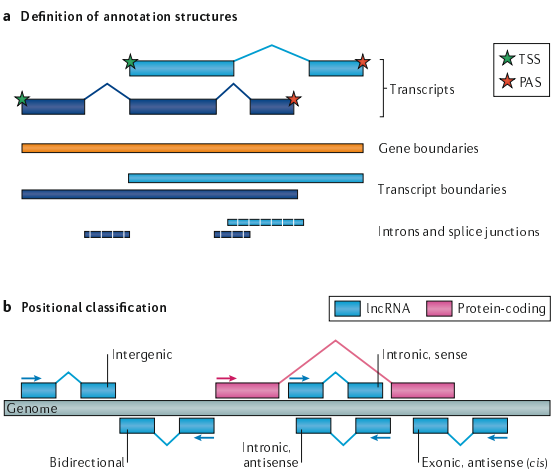


Fig. 1 | **Basic concepts of lncRNA annotations. a** | The principal structures of a long non-coding RNA (lncRNA) to be annotated. Annotations are hierarchical: they are composed of gene loci, each of which is composed of one or more partially overlapping transcripts, themselves composed of one or more exons (blue rectangles). **b** | Positional classification of lncRNAs with respect to the nearest protein-coding gene. Genic lncRNAs overlap a protein-coding gene locus, whereas intergenic lncRNAs, also known as long intergenic non-coding RNAs (lincRNAs), do not. Transcripts that overlap a protein-coding gene on the opposite strand are identified as antisense. PAS, polyadenylation site; TSS, transcription start site.

(CAGE) data[2,23]. Second, our understanding of the lncRNA sequence–function relationship is poor (BOX 1). Thus, in contrast to the information-rich, readily identifiable open reading frame (ORF) of protein-coding genes, sequence features or functional elements cannot presently be used to identify novel lncRNAs. Third, lncRNAs tend to be weakly conserved during evolution[24,25], making it challenging to identify their orthologues or paralogues by sequence similarity. Consequently, lncRNA annotation relies almost entirely on physical transcriptomic evidence.

The importance of accurate annotations. The fundamental nature of lncRNA annotations means that uncertainties or inaccuracies can have a profound impact on downstream projects. For example, during studies on the developing bat wing, researchers used microarrays to identify what seemed to be an intergenic lncRNA upstream of the gene encoding the developmental factor Meis2 (REF. [26]). However, careful analysis revealed that the cDNA sequence upon which the annotation had been based was most likely an internally primed fragment of the *Meis2* 5′ untranslated region (UTR)[26]. Similarly, an annotated lncRNA whose orthologue was knocked out in mouse, *Kantr*, was identified through an analysis of full-length transcript models from long-read sequencing to be a protein-coding transcript with an ORF in a previously unannotated exon[9,27]. Finally, 1/2-sbsRNA AF087999, which has been proposed to regulate mRNAs in *trans* through Staufen binding, lies within the 3′ UTR of the *RBM4* gene. There is little evidence supporting AF087999 as an independent transcript, leaving it unresolved whether it is a standalone lncRNA or a misannotated UTR fragment[28].

Amongst the most frequent use of lncRNA annotations is as a reference for quantifying and identifying differentially expressed genes and transcripts in RNA-seq experiments. Quantifier programs, such as RSEM[29] or Kallisto[30], take annotation files as an input together with mapped RNA-seq reads and attempt to estimate abundances of lncRNA transcripts. This is a challenging problem, particularly for lowly-expressed transcripts[31]. Inaccuracies or omissions in lncRNA annotations will propagate to transcript abundance estimates. For example, an excessively long 3′ exon annotation will lead to artificially low expression estimates, given that measures such as fragments per kilobase per million mapped (FPKM) are scaled to the annotated length of transcripts[32].

Accurate estimates of lncRNA transcription start sites (TSSs) are of particular importance for studies of lncRNA promoters or CRISPR–Cas screens, which depend on targeting Cas9 molecules to gene promoters[6,19]. Such studies should only examine transcripts with confident 5′ ends, which may be achieved by using independent evidence such as CAGE data to exclude unvalidated TSSs[27,33–35].

Biomedical applications for lncRNA annotations are of growing importance. The recent availability of cancer genomes has enabled searches for driver lncRNAs, whose mutations are positively selected for during tumorigenesis[36,37]. Predictions are critically dependent on lncRNA annotation quality. Similarly, diagnostic screening and genome-wide association studies (GWAS) depend on making accurate inferences of the functional impact of trait-associated mutations[38]. Such mutations are often assumed to be regulatory when they fall outside exonic regions. Truncated lncRNA annotations could therefore lead to the misinterpretation of mutations that actually fall inside a lncRNA exon and act through the mature lncRNA transcript, for example, by modulating a microRNA response element, as in the case of *lnc-LAMC2-1:1* (REF. [39]). Finally, the identification of lncRNA biomarkers, such as *PCA3* for the detection of prostate tumours[40], uses RNA-seq quantified against lncRNA annotations. In cases where the analysis output is a diagnosis, annotation quality can thus have a direct impact on patient outcomes.

Additional examples of the diverse uses for lncRNA annotations include evolutionary phylogenies[24], analysis of splicing regulation and conservation[41], identification of small ORFs (sORFs)[42], lncRNA-specific gene properties[25] and RNA modifications[43]. Finally, the success of the nascent field of lncRNA functional domain prediction will depend in large part on the availability of comprehensive and complete lncRNA annotations (BOX 1).

The ecosystem of annotations

Thanks to ongoing efforts over the past two decades (BOX 2), a range of lncRNA annotation resources obtained by different methods are presently available. Contemporary annotation efforts are principally based

Biotype
An annotation label referring to the genomic classification, processing or other characteristics of a locus or transcript intended to provide insights into biological function.

Expressed sequence tags
(ESTs). An early transcriptomic method in which short fragments of transcribed regions, often from 5′ or 3′ ends, are identified through sequencing of cDNA.

Cap analysis of gene expression
(CAGE). A cap-trapping and sequencing method that is considered a gold standard for mapping RNA 5′ ends.

Transcript models
Abstract descriptions of a transcription event, defining the genomic location of the start point, the end point and splice junctions.

Box 1 | **Beyond gene annotation: mapping functions and domains**

In tandem with complete gene annotations, an additional objective is to predict and label molecular, biological and disease functions of long non-coding RNAs (lncRNAs). This aim is held back by our poor understanding of the sequence–function relationship of lncRNAs, in contrast to protein-coding genes whose function can usually be predicted from primary sequence alone[52]. Here, we discuss a selection of promising methods to predict the functions and functional domains of lncRNAs. It will be interesting in the future to see such information integrated with annotation databases, LNCipedia and LncRNAWiki being the only resources thus far to do this[61,62].

Gene-level functional annotation

Strategies to predict lncRNA functions have traditionally involved reassigning functional labels from protein-coding mRNAs to lncRNAs based on expression patterns. Tissue profiles of lncRNAs and mRNAs are determined from RNA sequencing (RNA-seq) or microarray data and then used to create mixed gene clusters by correlation. Significantly enriched functional labels attached to mRNAs in each cluster, such as Gene Ontology, Kyoto Encyclopedia of Genes and Genomes (KEGG) or disease association terms[103–105], are assigned to any lncRNAs in the same cluster. This widely used approach is often referred to as guilt by association[23]. However, it assumes that expression patterns hold information on molecular functionality. Algorithms of growing sophistication, often integrating additional data, are being applied to this problem[106–108]. A lack of gold standard data means that it is difficult to assess the power of such techniques, although new databases may help resolve this[11,109,110].

The expression of lncRNAs within the cell, or subcellular localization, may hold more useful clues for molecular functions. RNA-seq-based maps of lncRNA levels in compartments of the cell, including nucleus, cytoplasm and other organelles, can be used to create maps of localization[111–113]. These data are then used to classify lncRNAs by their localization according to defined cut-offs[114]. Although this approach does not make specific functional predictions, it can provide broad pointers: for example, nuclear-enriched transcripts may regulate transcription, while cytoplasmic transcripts are more likely to play post-transcriptional roles. Localization data may also be used to search for domains or motifs that promote lncRNA trafficking to specific cellular sites[115–117].

Mapping lncRNA functional elements

The prevailing view is that lncRNAs, similar to proteins, are modular and composed of separable 'functional elements' (REFS[117–119]). Convincing evidence is available for a limited number of cases[117,118,120–124], but the global annotation of elements would be a powerful basis for predicting lncRNA functions.

Elements can be predicted by a variety of methods. Evolutionary conservation of RNA structure is a statistically rigorous way of finding putative functional elements[88]. Protein-binding data are useful in identifying molecular interactors and their binding sites, although they have the drawback that sensitivity depends on expression, which is usually low for lncRNAs[125]. Maps of inferred or experimentally identified microRNA sites may point to post-transcriptional regulatory roles, such as for competitive endogenous RNAs (ceRNAs)[125,126]. lncRNAs may interact with genomic DNA through the formation of triplex structures that can be predicted bioinformatically[127]. Other studies have attempted to map functional sites through transposable elements[89,115,116,128].

either on automated transcriptome assemblies from short reads or on manual annotation of existing cDNA and EST libraries (FIG. 2). Recent years have seen considerable efforts in consolidating lncRNA collections, with attention shifting from quantity to quality and a premium placed on 5′ and 3′ completeness. In this section, we review presently available annotations, grouped by method.

Annotations based on transcriptome assembly using short reads. Short-read RNA-seq experiments produce hundreds of millions of reads, providing a deep sampling of even large mammalian transcriptomes. These reads can be used to annotate transcripts from known and novel genes, both coding and non-coding. However, the fact that reads are much shorter than typical mRNAs and lncRNAs means that they must be bioinformatically assembled to infer the structure of the underlying transcript (FIG. 2a). Despite drawbacks inherent in this approach (discussed below), RNA-seq has facilitated the creation of large lncRNA catalogues.

The MiTranscriptome annotation combines 6,503 data sets, heavily weighted to 27 cancer types, to automatically annotate 58,648 lncRNA genes using a two-stage assembly strategy[44]. At the time of its creation, 54% of loci were not present in any other available resource.

Several studies are taking steps to improve the completeness of annotations. The Functional Annotation of the Mammalian genome (FANTOM) CAGE-associated transcriptome (CAT) meta-assembly combines both published sources and in-house short-read assemblies[45]. What sets this collection apart is its use of CAGE tags, which mark transcript TSSs, to identify 5′-complete transcript models. The resulting 27,919 gene loci are more complete at the 5′ end compared with other annotations, as judged by independent evidence, such as histone 3 lysine 4 trimethylation (H3K4me3) and DNase I hypersensitivity sites (DHSs)[45]. One drawback of CAGE is that, similar to other RNA-dependent methods, its signal scales with expression[46]; hence, lowly-expressed transcripts are more weakly represented.

The BIGTranscriptome catalogue comprises transcripts that are complete at both the 5′ and 3′ ends[47]. It employs a new method, CAFE, which is capable of inferring strands of unstranded RNA-seq reads. Consequently, CAFE overcomes strand ambiguity, which particularly affects genic transcript models generated from unstranded data sets, such as those from the Human Body Map (HBM) or the Genotype-Tissue Expression (GTEx) project[48]. CAGE and poly(A)-position profiling by sequencing (3P-seq) were used to assess 5′-end and 3′-end completeness, respectively[45,49]. Combining 169 RNA-seq data sets, BIGTranscriptome comprises 1,725 novel full-length lncRNA loci.

Annotations based on manual curation. Gene annotation remains one of the few high-throughput scientific activities where humans still outperform computers. In manual annotation, a team of human annotators systematically assembles transcriptomic and genomic evidence into gene models according to defined protocols. By inspection of high-quality transcript evidence, principally from ESTs and cDNA databases, annotators can create fairly confident annotations, free from many of the artefacts inherent in automated approaches (FIG. 2b).

The most widely used manual annotation is GENCODE[2,50], which stands out thanks to its extensive experimental validation and integration into the Ensembl annotation set[2]. Whereas the main GENCODE protein-coding gene annotation is created by merging the output from two pipelines, one manual and one automated, the lncRNA annotation is almost entirely manual. Individual transcript models are annotated and grouped together on the basis of genomic overlap of exons and splice sites into gene loci.

Unsurprisingly, manual annotation is much slower than automated approaches. Nevertheless, GENCODE annotations, released at 6-month intervals, have grown rapidly since 2012 (BOX 3; FIG. 3a). Moreover, single-exon

Fragments per kilobase per million mapped (FPKM). One of the principal units of RNA abundance in the context of RNA sequencing experiments, defined as the number of sequenced fragments per kilobase of annotation per million mapped fragments.

models and transcript models supported by transcriptomic data from long-read sequencing are now being introduced (discussed below).

Newly created transcript models are assessed for protein-coding potential (BOX 4) and whether they are likely to be functional or pseudogenic. Where there is no evidence of coding potential from mass spectrometry data, orthologues or paralogues in reference databases such as UniProt[51], structural or functional protein domains identified by Pfam[52] or conservation data such as PhyloCSF[53], a locus is defined as non-coding. lncRNAs from the literature are assessed with equal stringency. Although much of the annotation of lncRNAs was completed during first-pass manual annotation across the whole human genome, targeted (re)annotation of missing or truncated lncRNAs is now underway.

All new transcripts and genes are assigned stable identifiers on their creation. All updates to annotation are captured in an increment to the version of the gene and transcript identifier (that is, "ENSGXXXXXXX.2"). For example, when extension or trimming of a transcript is undertaken in light of new data or where new data emerges to strongly support changing the biotype of a locus (BOX 3), updates will be made and a version increment applied.

Owing to the quality deriving from its manual annotation, regularly updated versions, long-term support, well-defined and consistent source data, identifier stability and integration into Ensembl[50], GENCODE has been adopted by most large-scale genomics projects, including the Encyclopedia of DNA Elements (ENCODE)[54] (for which it was originally created), GTEx project[48], International Cancer Genome Consortium (ICGC)[55], Blueprint[56], Epigenome Roadmap[57] and FANTOM[45]. The use of stable Ensembl identifiers simplifies the integration of data across projects and releases. However, the inherent weakness of GENCODE is its relatively small size: 15,778 genes in human (version 27) and 11,975 in mouse (version M15). Of note, the mouse annotation project was commenced later, accounting for the difference in size with human.

Another manual gene annotation resource, Reference Sequence (RefSeq), was created and is maintained by the National Center for Biotechnology Information (NCBI) and covers multiple species, including human[58]. Consisting of a mixture of manual and automated annotations, RefSeq is created using a variety of evidence, including cDNAs, ESTs and RNA-seq. Entries carry unique and stable identifiers and are associated with metadata summarizing their annotation history. Of relevance in this context are non-coding RNA annotations with accessions 'NR_' and 'XR_', which refer to manually curated models (NR) and products of an automated pipeline based on Illumina data (XR), respectively. Thus, the RefSeq annotation process is similar to GENCODE, with the exception of usage of RNA-seq. Along with GENCODE, RefSeq is one of the most widely used lncRNA annotations[59].

Integrative annotations. A number of other lncRNA collections are worthy of note. NONCODE has, since 2005, integrated annotations from a mixture of manual literature searches and other annotations[3]. The latest version, NONCODE (version 5), is to our knowledge the single largest present collection, describing 96,308 lncRNA gene loci in human alone (as of November 2017). It also has data for 15 species other than human and mouse.

RNACentral is a large-scale resource of non-coding RNA sequences, integrating various other databases, which lists 116,292 lncRNA sequences at the time of writing[60]. It is based on sequences, rather than annotations, making the total number of lncRNA loci unclear.

Finally, LNCipedia and LncRNAWiki stand out in their usefulness for integrating functional data. LNCipedia holds a database of 48,028 carefully filtered lncRNA genes from a range of sources[61]. Users may access information on peptide mapping, coding potential, RNA folding and microRNA recognition. Similarly, LncRNAWiki holds a variety of useful information, including disease association and putative small peptides, and is an invaluable resource of manually curated functional information for hundreds of lncRNAs[62].

Box 2 | The evolution of lncRNA collections

The first hint at the volume of long non-coding RNAs (lncRNAs) populating our genome came from genomic microarray technology. Starting in 2002, tiled microarrays with increasing density and genomic span revealed extensive transcription outside of then-known gene loci[129]. However, the exact sequence and hence protein-coding potential, of those transcripts could not be resolved with this technology. The sequences of these unannotated transcripts were first resolved by massive cDNA sequencing undertaken by the Functional Annotation of the Mammalian genome (FANTOM) consortium[130,131]. The consortium used a combination of cap analysis of gene expression (CAGE), which can identify transcription start sites (TSSs) by sequencing the 3′ end of cDNAs (that is, the 5′ end of RNAs), and ditag sequencing (also known as paired-end tag sequencing), which is capable of identifying both TSSs and polyadenylation sites. Approximately one-third of cDNAs did not contain identifiable protein-coding sequences; in other words, they were lncRNAs. This data set facilitated the first studies demonstrating purifying evolutionary selection on lncRNAs as a population, implying that at least a subset is functional rather than "transcriptional noise"[4].

lncRNA genes were also identified indirectly through their patterns of histone modifications[23]. Reasoning that lncRNA genes may carry similar combinations of histone 3 lysine 4 trimethylation (H3K4Me3) and histone 3 lysine 36 trimethylation (H3K36Me3) modifications — known markers of active protein-coding genes — researchers identified approximately 1,000 long intergenic non-coding RNAs (lincRNAs) in human and mouse[23,132]. These lincRNA genes exhibited low steady-state expression levels compared with mRNAs, now known to be a general property of lncRNAs.

Growing volumes of publicly available cDNA sequences opened the way to accurate lncRNA annotations, similar to those for protein-coding genes. The first catalogue of 5,446 human lncRNA loci was generated largely on the basis of cDNAs filtered by an open reading frame (ORF) prediction tool and a pipeline based on the protein basic local alignment search tool (BLASTP)[133].

The advent of RNA sequencing (RNA-seq) democratized lncRNA annotation. Using only a sequencer and off-the-shelf computational tools, any laboratory was able to identify thousands of lncRNA loci in their favourite cell type. A central requirement for this approach is transcriptome assembly, whereby computational algorithms are used to reconstruct the underlying transcript structures responsible for observed RNA-seq reads[97] (FIG. 2a). Reference-based methods that make use of read-to-genome alignments to infer transcript structures tend to be more accurate than de novo methods[98]. Foremost amongst reference-based assemblers are Cufflinks[97] and, more recently, StringTie[99]. In the first attempt to apply RNA-seq to lncRNA annotation, Cabili et al.[134] assembled RNA sequences from a variety of human tissues to yield a total of 4,662 lncRNA loci. This study discovered another fundamental property of lncRNAs: high tissue-specificity and cell type-specificity.

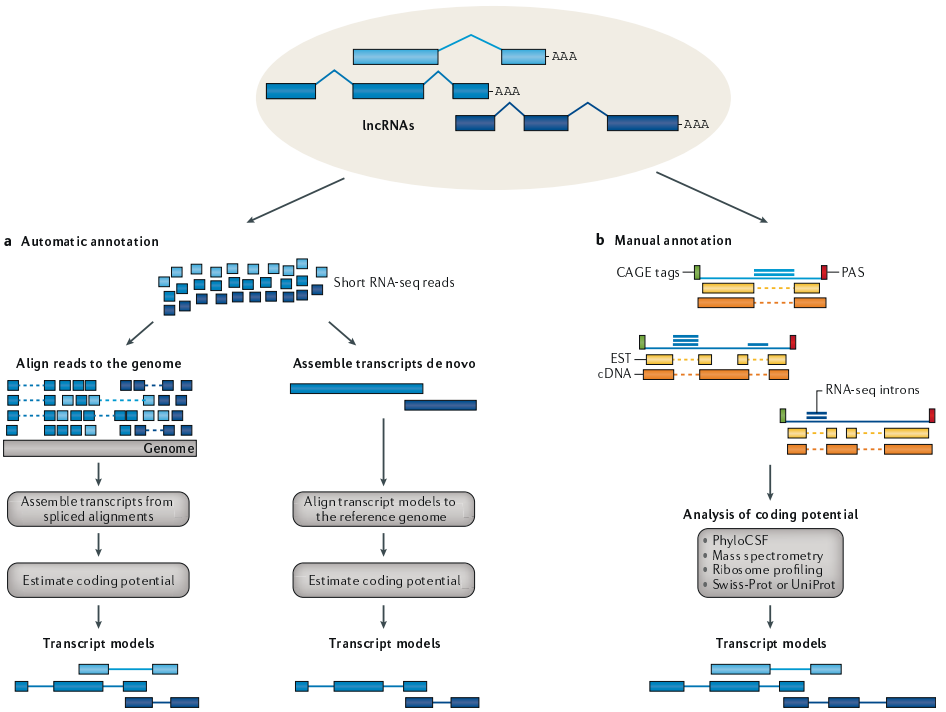


Fig. 2 | **Annotation strategies for lncRNAs. a** | Automatic annotation based on RNA sequencing (RNA-seq) may follow two distinct strategies that differ in how the genome reference is used. The align-then-assemble strategy (left) aligns reads to the reference genome to reveal possible splicing events and then assembles reads into transcript models. The assemble-then-align strategy (right) builds transcript models de novo, directly from the RNA-seq reads, and then aligns them to the reference genome to determine their exon–intron structure. De novo transcriptome assembly has more explorative potential than alignment-based assembly but tends to have worse performance[68]. **b** | In manual annotation, human annotators employ various sources of data to build transcript models. Expressed sequence tags (ESTs) and cDNA form the primary evidence for transcript models and are often supplemented with RNA-seq reads to validate introns, cap analysis of gene expression (CAGE) clusters to identify 5′ ends[45] and poly(A)-position profiling by sequencing (3P-seq) to identify polyadenylation sites (PASs)[47]. A key step in the annotation process is to assess the protein-coding potential of transcripts, usually on the basis of a combination of methods. lncRNA, long non-coding RNA.

How good are lncRNA annotations?

Overlap between annotations. Annotations tend to have low overlap (FIG. 3b). For example, the two largest annotations, MiTranscriptome and NONCODE, have just 27.7% and 45.5% of genes in common, respectively. Not surprisingly, NONCODE encompasses more than 97% of GENCODE, which it incorporates. What is perhaps unexpected is the poor overlap that is observed between the two manual annotations, GENCODE and RefSeq (34.6% and 44%, respectively). Overall, the low overlap points to much scope for merging of annotations to improve comprehensiveness.

Quality metrics for annotations. An ideal annotation would be a record of every locus expressed at any point in time from the genome of a given species. An important requirement for future lncRNA mapping projects is the development of standards for assessing quality that go beyond anecdotal examples. For the present discussion, we make the following definitions of annotation quality: (a) comprehensiveness — the fraction of all gene loci that are included; (b) exhaustiveness — the fraction of all transcripts from each locus that are known; (c) completeness — the fraction of transcript models that cover the entire length, from start to end, of the physical RNA molecule. Obviously, comprehensiveness

Box 3 | **Using GENCODE lncRNA annotations**

Availability

The GENCODE annotation of long non-coding RNAs (lncRNAs) is released in alignment with versioned Ensembl updates. Mouse releases are prefixed M; the most recent human release is GENCODE v27 and for mouse vM15. Full GENCODE annotations are available in GTF and GFF3 formats from the Ensembl and University of California Santa Cruz genome browsers and from gencodegenes.org. Separate files containing only lncRNA transcripts are also available. The GENCODE site also houses a full archive of previous releases and their statistics.

Biotypes

A full description of all GENCODE lncRNA biotypes is presented in the HAVANA annotation guidelines[135]. More recently, biotypes relating to other genomic features have been added and are being populated; for example, a 'bidirectional-promoter lncRNA' describes a locus where a lncRNA lies on the opposite strand to a protein-coding gene and there is evidence — for example, from cap analysis of gene expression (CAGE) data — that their transcription start sites lie within a window of 200 bp.

Biotype labels should be employed with caution as they tend to exhibit considerable inertia. As they are defined with reference to nearby protein-coding gene structures, any changes in those structures can lead to a change in the biotype. If users require up-to-date biotype information, it is recommended to regenerate them, for example, by using the lncrna.annotator script or the classifier module within FEELnc[136].

GENCODE Comprehensive versus GENCODE Basic

GENCODE Comprehensive comprises the entire annotation of transcript models. As lncRNA annotations become increasingly complex, a need arises for a simplified annotation: GENCODE Basic. GENCODE Basic contains at least one transcript for every gene locus, ensuring full gene representation. For protein-coding loci, all coding transcripts with full-length coding DNA sequence (that is, ATG to stop codon) are included in the Basic set. For complex lncRNA loci, the Basic set is generated by including the minimal set of transcripts that capture >80% of the splice sites.

and exhaustiveness are impossible to define, as we do not know the total number of lncRNA genes or transcripts. Nevertheless, we can at least compare proxies for these metrics between annotations (TABLE 1) to get a comparative picture. By contrast, a minimum bound can be placed on completeness, owing to the availability of independent evidence for transcript 5′ and 3′ boundaries.

Based on these three metrics, we have compared the discussed lncRNA annotations (FIG. 3c). Most striking is the general anti-correlation between comprehensiveness and completeness. In other words, there is a trade-off between quality and size: smaller annotations tend to have higher completeness (although this remains low in absolute terms) and vice versa. Amongst the smaller annotations, BIGTranscriptome is the leader in terms of completeness, although with low numbers of annotated transcripts per gene. The two manual annotations, GENCODE and RefSeq, have comparable profiles. For the larger annotations, MiTranscriptome has just 4.4% of complete (full-length) transcript models (FIG. 3c), which is most likely the result of its dependence on transcriptome assembly. NONCODE beats MiTranscriptome in size and completeness but with lower exhaustiveness. FANTOM CAT represents a compromise between completeness and comprehensiveness. Of note, we find substantially lower 5′ completeness than originally reported[45], which is due to the use of more stringent CAGE cut-off thresholds: only robust CAGE clusters (FANTOM5 phase 1/2 robust (n = 201,802)) were considered, and FANTOM5 phase 2 unfiltered CAGE clusters (n = 4,218,430) were discarded owing to their seemingly high background rate.

Data are also displayed for protein-coding genes as a reference, with the assumption that their annotation is of the highest quality. The protein-coding gene annotation should be comprehensive, as not many are expected to remain undiscovered[63]. We also included a recently generated set of full-length lncRNA transcript models produced using capture long-read sequencing (CLS) technology (discussed below)[27]. These models display high completeness, in part because their 5′ ends were defined using the same CAGE data as used here for evaluation. Incorporation of CLS models into GENCODE resulted in an improved annotation, GENCODE+ (TABLE 1), with dramatically higher completeness. It is noteworthy that GENCODE+ has a slightly reduced gene count as a result of unifying artefactually separate gene models in existing annotations.

One important caveat of this analysis is that CAGE clusters used for 5′-end definition are expression-dependent and only available for a defined set of tissues. This likely accounts, at least in part, for the fact that protein-coding genes have apparent 5′ completeness <100% (FIG. 3d) and will also underestimate completeness of lower expressed lncRNAs. However, it is also possible that some protein-coding gene annotations remain incomplete.

The use of proxies for comprehensiveness (numbers of loci) and exhaustiveness (transcripts per gene) makes the key assumption that no false-positive annotations exist. This assumption is probably incorrect and will affect some annotations more than others. In particular, assembly-based collections may hold substantial numbers of false-positive transcripts. Inspection of splice junctions supports the idea that certain annotations, particularly NONCODE, suffer from high rates of false-positive structures (FIG. 3e).

Overall, this analysis illustrates the strengths and weaknesses of contemporary annotations. It highlights the great scope for improving lncRNA annotations, first, by increasing their completeness to levels observed for protein-coding genes and, second, by improving their comprehensiveness by merging diverse available resources.

Sources of incompleteness. The lack of completeness in existing lncRNA annotations may be traced to several historical and technical factors. cDNA molecules often tend to be 5′ truncated, owing to a combination of RNA degradation and the tendency of reverse transcriptase molecules to disengage before reaching the 5′ end of the template RNA, often as a result of RNA secondary structures[64]. In short-read RNA-seq, a range of processes create non-uniformity in read coverage, particularly at the 5′ and 3′ ends[65]. Together, these factors introduce a tendency for short-read assemblies and cDNA libraries, upon which most annotations are based, to be 5′ and 3′ incomplete[34,35,45,66].

More generally, the assembly of transcriptomes from short reads is inherently challenging. Assembly

REVIEWS

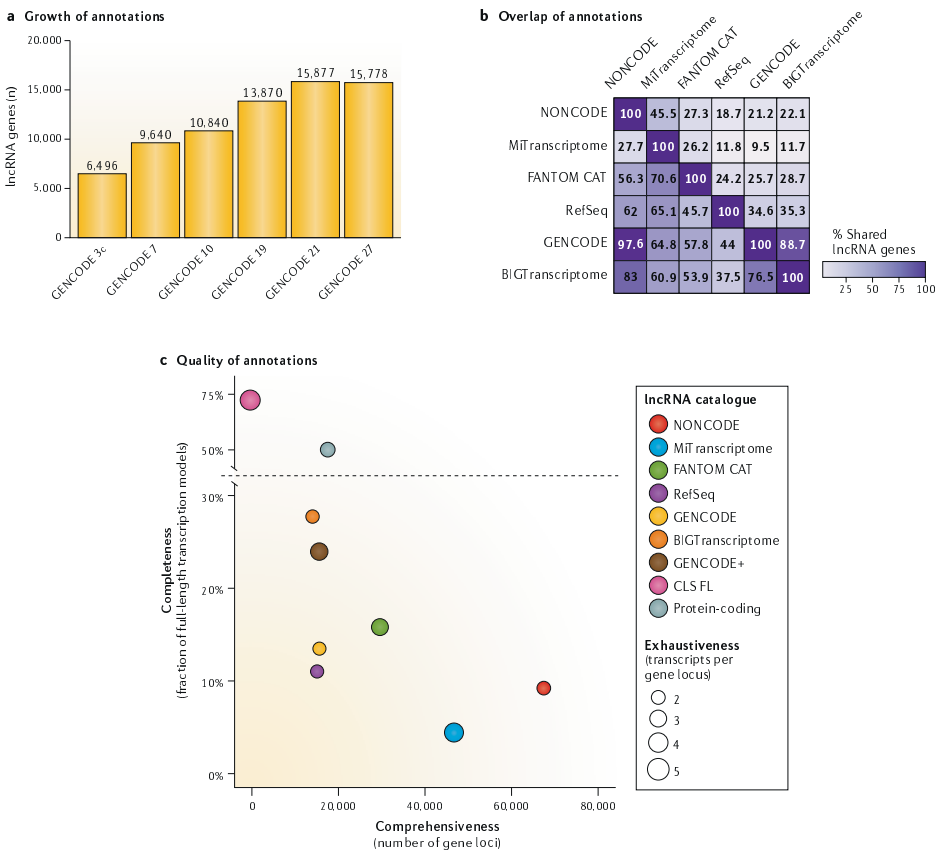


Fig. 3 | **Comparison of leading lncRNA annotations. a** | Growth of GENCODE long non-coding RNA (lncRNA) collection overtime, in terms of gene loci. Only reference releases are included. **b** | Overlap between annotations at the gene level, based on a medium-stringency definition. Values represent the percentage of gene loci in the annotation of each row that overlap the annotation in each column. Overlap is defined as at least 60% of the span of the shorter gene on the same strand. Only genes with at least one multiexonic transcript were included. See TABLE 1 for details. **c** | Comparison of quality metrics between annotations. x-axis: comprehensiveness, or the total number of gene loci; y-axis: completeness, or percentage of transcript structures whose start is supported by a robust phase 1/2 Functional Annotation of the Mammalian genome (FANTOM) cap analysis of gene expression (CAGE) cluster (n = 201,802) within ±50 bases and whose end contains a canonical polyadenylation motif[154] within a window of 10–50 bp upstream. Circle diameters reflect exhaustiveness, or mean number of transcripts per gene. GENCODE+is the union of GENCODE version 20 with non-anchor-merged capture long-read sequencing (CLS) transcript models. Protein-coding is a set of confident GENCODE protein-coding transcripts as described in REF.[27]. **d** | As for part **c**, but separately for 5′ and 3′ completeness. **e** | The distribution of predicted splice junction strength for splice site acceptors and donors in each lncRNA catalogue, as calculated by the GeneID software[155]. The plots show non-redundant splice sites from lncRNA annotations sets (top), confident GENCODE protein-coding transcripts (middle), and 500,000 randomly selected GC|GT donors+AG acceptors with no evidence of splicing in any of the annotation sets under study (bottom). For each non-canonical splice site not scored by GeneID, a random score between −30 and −20 was assigned.

programs, including the widely used Cufflinks[67], have high error rates. Whereas exons are identified with reasonable sensitivity, their assembly into correct transcripts is particularly difficult[68]. Simulations across a range of assembly programs demonstrate a mean sensitivity of only 41% in assembling expressed genes, dropping to 21% at the transcript level[68]. The majority of such transcript models lacks one or more exons. Assemblies are sensitive to gene expression levels and coverage uniformity[68], which has a particular impact on lowly-expressed lncRNAs. However, even when controlling for expression, transcriptome assemblies are less sensitive for lncRNAs compared with mRNAs for unknown reasons[68]. More recent assemblers such as StringTie and Scallop run far faster than Cufflinks and have demonstrably better sensitivity and specificity, but resultant assemblies remain far from ideal[69,70]. In a study using StringTie to assemble synthetic spliced RNAs, it was found that for the correct assembly of >50% of its nucleotides, a transcript must be expressed at a level equivalent to 23 FPKMs — far in excess of the average lncRNA or even mRNA[66]. These issues will result in low comprehensiveness, exhaustiveness and completeness of annotations based on transcriptome assemblies.

Another issue that probably has an impact on comprehensiveness is of historical nature: the material used for the generation of cDNA libraries has been biased towards adult tissues, tumour samples and cell lines[2,44]. Thus, modern annotations may omit much of the wealth of lncRNAs likely to be expressed during embryogenesis, development and childhood[48]. Similarly, certain lncRNAs may only be expressed in rare subpopulations of cells within a tissue or even cell culture[71] and thus are likely to be missed owing to the low apparent expression in bulk cell samples.

In summary, present annotations are likely to fall short in all the quality metrics described, leaving thousands of gene loci, transcripts and exonic nucleotides unmapped.

Emerging technologies in lncRNA annotation

Advances in key technologies for targeting and sequencing lncRNA transcripts promise to directly address the two principal challenges facing lncRNA annotations of low target abundance and incomplete transcript models.

Long-read sequencing technologies. Pacific Biosciences' (PacBio) technology employs zero-mode waveguides to sequence single circularized DNA or cDNA molecules. Around 40,000 reads are produced by each lane, with an average ~1.5 kb length[27,72]. This length is several-fold longer than the average exon, meaning that the exon connectivity of complete or almost complete transcript structures can be resolved. A recent study in human showed that, for transcripts up to ~1.5–2.0 kb, the majority of reads yields full-length transcript structures, falling short on average 47 nt and 6 nt from the annotated 5′ and 3′ sites, respectively[72].

Raw PacBio sequencing reads tend to have relatively high error rates[72]. To mitigate this issue, consensus reads of insert (ROIs) are assembled from multiple passes of the same circular template molecule. Resulting per-base sequencing errors are moderate, approximately twofold greater than for Illumina and with a tendency for nucleotide deletions[72]. At this rate, the majority of reads can be mapped with high confidence[73]. Despite its advantages, widespread adoption of PacBio is hindered by its cost and low throughput. Given the low representation of lncRNAs within the cellular transcriptome, pure PacBio sequencing would be an inefficient method to map lncRNA loci[72]. Perhaps its greatest drawback is its sequencing preference for short templates in a mixture. This limitation creates the need to size-select cDNA libraries, introducing a length-dependent bias in the sequenced transcripts[27,72].

Nanopore-based technologies read single molecules in real time[74]; nucleic acid molecules are translocated at a controlled rate through a membrane-bound protein nanopore. Changes to electric currents through nanopores are used to infer the identity of each nucleotide. This technology has reached the mainstream market with Oxford Nanopore's MinION technology[74], which is capable of returning ~5 million reads per flow cell, at a cost of ~€500.

Nanopore has a range of advantages over other sequencing approaches. By dispensing with the amplification or enzymatic modification of target molecules, important sources of bias are avoided. cDNA molecules can be directly sequenced with minimal preparation[75], and it may even be feasible to identify chemically modified bases[76]. A recent report describes direct sequencing of RNA (as opposed to cDNA) from a variety of samples, with read lengths of up to 7.5 kb and sequencing accuracy of 80%[77]. Reads are free from biases regarding template length or GC content, which affect other technologies[78]. Most importantly in the present context, nanopore sequencing yields reads of lengths that are virtually unlimited and that far exceed known lncRNAs and mRNAs[78]. These beneficial properties of throughput, read length and cost make nanopore technology highly appealing in the context of gene annotation.

Box 4 | Are lncRNAs really non-coding?

The extent to which protein-coding capacity is a qualitative (binary) or quantitative (gradual) property of RNAs has long been debated[137]. Recently, functional small peptides have been identified in transcripts previously annotated as long non-coding (lncRNAs)[42,138]. More broadly, ribosome profiling[139,140] and bioinformatic[141] studies have claimed that a large proportion of annotated lncRNAs encode proteins.

However, these findings are not yet conclusive. Ribosome interaction itself is suggestive, but not direct, evidence of coding potential[142,143]. For bioinformatic identification, a large fraction of purported, novel coding transcripts are likely to be false positives, arising from inadequate statistical approaches that do not correctly account for technical and biological noise[144–148]. For example, of the ~350 best-supported novel open reading frames (ORFs) proposed by Mackowiak et al.[141] and manually reviewed by GENCODE, only 35 could be verified (A.F., unpublished observation). Together with the presently low number of cases for which peptides have directly been observed, this observation means that it may be premature to suppose that most lncRNAs are translated into functional peptides.

This is not to say that annotations should not be rigorously screened to flag "transcripts of unknown coding potential" (REF.[136]). A variety of tools exist to predict protein-coding regions in RNA sequences, which may be classified amongst those using intrinsic sequence properties (for example, Coding-Potential Assessment Tool (CPAT)[149]), similarity to known proteins (for example, Coding Potential Calculator (CPC)[148]) and evolutionary signatures of protein evolution (for example, PhyloCSF[53]). The latter tool is considered to have the greatest sensitivity, particularly for short peptides[136,149] but identifies only evolutionarily conserved peptides and is computationally intensive. More direct evidence comes from mass spectrometry, although low sensitivity and the short length of potential peptides complicates their identification[94,150–152], and care must be taken to correctly estimate false-positive predictions[153].

Most annotation pipelines integrate one or several of these approaches[136]. For GENCODE, in addition to comparing putative ORFs within lncRNAs to entries in reference protein databases, such as UniProt and Pfam, all lncRNAs are routinely tested using both PhyloCSF and dedicated proteogenomics filtering. Manual re-examination can lead to reclassification of dubious lncRNAs. However, this is fairly infrequent: a stringent proteogenomics workflow to reprocess >52 million spectra revealed more than 1,400 putative novel protein-coding genes, but only 16 were confirmed following detailed reanalysis and just 8 fell in annotated lncRNA loci[94].

RNA capture sequencing. lncRNAs tend to be expressed approximately one order of magnitude lower than mRNA and represent about 1–2% of polyA+ RNA in a cell[27,72,79,80]. This creates a considerable hurdle for annotation, because lncRNA molecules are simply less likely to be sampled at a given depth of sequencing. One solution is to boost their apparent concentration in a cDNA library using oligonucleotide capture in a technique known as RNA CaptureSeq[81,82]. Custom libraries of tiled complementary oligonucleotide probes are used to enrich a population of desired targets in solution. This approach boosts the representation of lncRNA sequences to >25%, improving sequencing coverage by tens of fold compared with a conventional, uncaptured sample[81]. To date, this approach has been used successfully in human[83] and mouse[84] tissues.

RNA CaptureSeq demonstrates powerfully increased sensitivity compared with conventional, unbiased sequencing methods, typically discovering novel transcripts and gene loci expressed at far less than one copy per cell[83]. Often, adjacent and erroneously separate annotations are merged, or annotated loci are extended with new exonic sequence[84]. However, previous studies have largely relied on short-read Illumina sequencing coupled to Cufflinks transcriptome assembly[81,83,84]. Consequently, resulting annotations suffer from the same uncertainties and weaknesses as discussed above and have low 5′ and 3′ coverage[66,68].

The dependence of RNA CaptureSeq on short reads has recently been overcome by coupling it to PacBio technology in a method termed CLS[27,85]. By using long reads, CLS avoids the issues associated with short reads and transcript assembly, enabling the identification of full-length transcript models. By integrating CAGE data and fragments of poly(A) tails contained in PacBio reads, CLS can assess the completeness of transcript models at 5′ and 3′ ends, respectively (FIG. 4). The use of template-switching reverse transcriptase technology to generate almost full-length cDNAs can boost 5′ completeness further[84]. Short reads sequenced from the same samples can be used to assess the accuracy of splice site predictions[86]. As such, CLS marries the enhanced sequencing coverage provided by capture to transcript model confidence afforded by long reads. In the first report of this method, 2 million reads each in human and mouse across a panel of tissues and cells yielded novel full-length transcript models from 947 previously annotated human lncRNA loci[27]. Although the annotation complexity of the probed regions was approximately doubled, there was no sign of saturation of splice junctions, indicating that much more sequencing depth will be required to establish definitive gene structures in detected loci. A similar conclusion was reached in a study using essentially the same strategy to survey the transcriptional landscape of chromosome 21 in human testis[85].

Although the speed and cost of the CLS approach make it a substantial step towards more comprehensive lncRNA annotations, it suffers from some weaknesses that must first be resolved. The lengths of full-length transcript models are limited by PacBio reads, leaving many targeted transcripts incomplete. Also, sequencing depths are insufficient to saturate targeted loci. The incorporation of nanopore sequencing technology in the CLS workflow should help to overcome these barriers.

Towards complete lncRNA annotations

With the tools of long-read sequencing and RNA capture in hand, we may now envisage an eventual complete lncRNA annotation: maps of the entire universe of lncRNAs expressed throughout the lifetime of an organism, beginning with *Homo sapiens*.

A roadmap. The most obvious route to complete annotation lies in the systematic application of CLS coupled to nanopore sequencing (FIG. 4). Capture library designs would have two main components. First, in order to complete existing annotations, the entire catalogue of known lncRNAs would be targeted[27]. Second, in order to map unknown lncRNAs, suspected loci lying outside of annotated exons would be probed. These would come from two main sources: first, loci with a high confidence for lncRNA production, such as physical evidence from RNA-seq-derived assemblies and introns[44,87], and second, regions with more speculative evidence, such as predicted lncRNA orthologues from other species[24], bioinformatic predictions[88], GWAS regions[89] or small RNAs with presumed long precursors[90].

Oligonucleotide capture
A method for enriching cDNA libraries with sequences of interest using solution-phase hybridization to tiled, labelled oligonucleotide probes.

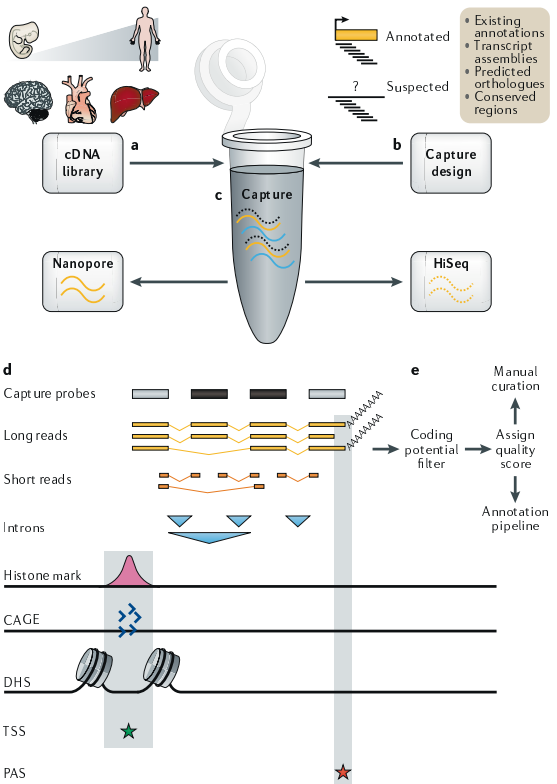


Fig. 4 | **Integrating capture and long-read sequencing with annotation pipelines.** **a** | Full-length cDNA libraries are prepared from a variety of tissues across the human lifespan. **b** | Target annotations are prepared from a variety of known and suspected long non-coding RNA (lncRNA) loci and used to design capture probes (black bars). **c** | Solution-phase oligonucleotide capture is performed, and enriched cDNA libraries are sequenced by long-read nanopore and short-read Illumina technologies. **d** | The resulting long reads are collapsed to produce non-redundant transcript models. The completeness and accuracy of these models are assessed using various evidence: introns (blue triangles) by short reads; transcription start site (TSS; green star) by promoter histone modifications, cap analysis of gene expression (CAGE) clusters and DNase I hypersensitivity sites (DHSs); and polyadenylation site (PAS; red star) by long-read-encoded poly(A) tails. **e** | With this information, transcript models are graded for completeness, checked for protein-coding potential and passed to annotators for either direct incorporation into annotation pipelines (for complete models) or further manual curation (incomplete models).

Capture libraries would be used to probe diverse organ and tissue panels across the human lifespan from embryos to aged adults, thus going beyond the adult organ panels and tumours that tend to dominate present data sets[91,92]. Given that organs are complex mixtures of common and rare cell types, it will also be beneficial to probe purified cell populations[93]. Technology permitting, this may eventually be extended to sampling single-cell transcriptomes of rare types that would be missed in bulk preparations[71]. Finally, the majority of transcriptome studies to date have been performed on individuals of European ancestry, making future sampling across different human populations a priority.

Such an ambitious project would entail considerable logistical and economic challenges. As recognized by ENCODE[54], a practical first step would be to focus on complete collections of lncRNA in defined cell types or organs. These might entail complex organs or cell lines of particular scientific or biomedical relevance, such as ENCODE cell lines[80].

Captured cDNA libraries would be sequenced using nanopore technology, up to a rationally chosen depth, defined below. The accuracy of the 5′ end, the 3′ end and splice junctions would be validated using independent data sets, in addition to bioinformatic and experimental screening for protein-coding capacity[53,94]. With this level of quality, resulting transcript models can be added to existing annotations with low levels of scrutiny by human annotators, minimizing the delay between sequencing and public availability.

How do we know when to stop?. A number of considerations must guide decisions regarding resource allocation in annotation projects. First, we must take care to focus efforts on collecting lncRNAs of biological relevance. Unfortunately, we remain far from having reliable methods for distinguishing functional lncRNAs from transcriptional noise. Although imposing a minimum expression threshold is an obvious path, the discovery of apparently functional lncRNAs with expression of <<1 copy per cell[20] would argue against imposing a hard expression cut-off. Nevertheless, to maximize usefulness in downstream applications such as RNA-seq, it is sometimes helpful to eliminate unnecessary complexity arising from growing numbers of transcript isoforms. This has prompted the creation of simplified annotations such as GENCODE's Basic annotation (BOX 3).

A question of singular importance to the design of annotation projects is: is the lncRNA population finite, and if so, how many transcripts and loci does it comprise? Or conversely, is an effort at complete annotation doomed by the fact that the transcriptome is infinite, owing to pervasive transcription or unlimited combinatorial splicing[83]? Certainly, after a decade of research, we are little closer to assigning an upper bound to the first question. Recent CLS studies finished sequencing before saturating even already known lncRNA loci[27], while a recent study claims that lncRNA genes explore astronomical numbers of available splicing combinations[85]. Furthermore, present upper estimates of lncRNA numbers are biased towards adult cell types, raising the possibility of existence of untold numbers of developmentally regulated lncRNAs.

A further source of complexity could be 'personal' transcriptomes — lncRNAs that are unique to individuals or populations[95,96]. Such transcripts might arise from individual-specific genomic regions that are not

represented in the reference or else shared genomic regions that are active in certain individuals thanks to processes such as transposon insertion[97,98] or *trans*-acting factors[99]. Even if the size of every individual personal transcriptome is small, summed across the entire population it could be enormous. Efforts to map personal genomes and transcriptomes are underway with the ENCODE Tissue Expression (EnTEx) project amongst others[100]. Personal lncRNAs, if they exist, may explain individual-specific phenotypes and features and could be of crucial importance to personalized medicine.

However, there is evidence supporting the finiteness of the lncRNA transcriptome. Simulations performed on relatively shallow CLS sequences from an admittedly limited range of tissues exhibited a decreasing rate of discovery with depth[27], indicating that lncRNA transcript complexity tends towards an asymptote. Deveson et al. seem to have exhaustively mapped all exons on chromosome 21 in testis[85]. Similarly, in analyses of nearly the entire volume of public RNA-seq data, the number of splice sites almost reached a plateau[87]. Finally, a more focused study in B cells also found evidence for an upper threshold in lncRNA isoform diversity[101]. Therefore, although lncRNA transcripts are highly complex and challenging to exhaustively map, a full map of at least their exons and splice sites is tractable.

Nevertheless, in any large-scale annotation project involving third-generation sequencing at depth, it will be imperative to periodically monitor the rate of novel transcript discovery in each tissue sample as a function of sequencing depth. This will indicate when transcriptome complexity has been saturated and hence when sequencing resources should be reallocated to other samples.

Conclusions and perspective

lncRNA annotations are a fundamental resource for basic research and also have growing importance for practical applications such as personalized medicine[102]. Although it has been argued, quite reasonably, that many lncRNAs may represent non-functional noise, the growing number of clearly documented counter-examples suggests that at least a substantial fraction of transcripts is functional in the strictest sense of enhancing organismal fitness.

The rapidly growing volume of the annotated lncRNA transcriptome will bring benefits but also new challenges, particularly in making this information available in a way that maximizes usefulness without sacrificing genuine biological complexity.

At present, lncRNA annotations lag far behind those for protein-coding genes, to an extent not often appreciated by individual researchers. However, there is now an opportunity to create complete annotations, at least in certain well-defined cell types. This will not only open new vistas into the molecular biology of the cell, disease mechanisms and diagnostics, but also enable us to answer fundamental questions about the functionality of lncRNAs.

Published online 23 May 2018

1. Liu, G., Mattick, J. & Taft, R. J. A meta-analysis of the genomic and transcriptomic composition of complex life. *Cell Cycle* **12**, 2061–2072 (2013).
2. 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).
3. Fang, S. et al. NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* **46**, D308–D314 (2018).
This study presents the latest instalment of the long-running NONCODE annotation, which was amongst the first ncRNA annotations and currently represents the most extensive collection.
4. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* **17**, 556–565 (2007).
This study initially demonstrated that lncRNA exons and promoters are under purifying evolutionary selection and hence provided strong evidence that, as a gene class, they are functional.
5. Pegueroles, C. & Gabaldón, T. Secondary structure impacts patterns of selection in human lncRNAs. *BMC Biol.* **14**, 60 (2016).
6. Zhu, S. et al. Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR–Cas9 library. *Nat. Biotechnol.* **34**, 1279–1286 (2016).
7. Wen, K. et al. Critical roles of long noncoding RNAs in *Drosophila* spermatogenesis. *Genome Res.* **26**, 1233–1244 (2016).
8. Li, L. & Chang, H. Y. Physiological roles of long noncoding RNAs: insight from knockout mice. *Trends Cell Biol.* **24**, 594–602 (2014).
9. Sauvageau, M. et al. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife* **2**, e01749 (2013).
10. Ip, J. Y. et al. Gomafu lncRNA knockout mice exhibit mild hyperactivity with enhanced responsiveness to the psychostimulant methamphetamine. *Sci. Rep.* **6**, 27204 (2016).
11. Chen, G. et al. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* **41**, D983–D986 (2013).
12. Amândio, A. R., Necsulea, A., Joye, E., Mascrez, B. & Duboule, D. Hotair is dispensible for mouse development. *PLoS Genet.* **12**, e1006232 (2016).
13. Quek, X. C. et al. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.* **43**, D168–D173 (2015).
For many years, this publication was the reference resource for manually curated, experimentally validated functional lncRNAs.
14. Sheik Mohamed, J., Gaughwin, P. M., Lim, B., Robson, P. & Lipovich, L. Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. *RNA* **16**, 324–337 (2010).
15. Loewer, S. et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.* **42**, 1113–1117 (2010).
16. Huarte, M. et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409–419 (2010).
17. Ng, S.-Y., Johnson, R. & Stanton, L. W. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J.* **31**, 522–533 (2012).
18. 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).
19. Liu, S. J. et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, aah7111 (2017).
This paper provides a map of hundreds of proliferation-altering lncRNAs across seven human cell lines, representing an invaluable resource of functional genes.
20. Seiler, J. et al. The lncRNA VELUCT strongly regulates viability of lung cancer cells despite its extremely low abundance. *Nucleic Acids Res.* **45**, 5458–5469 (2017).
This study presents an intriguing example of an extremely lowly expressed lncRNA that yields a reproducible cellular phenotype after knockdown, thereby challenging the notion that expression cut-off thresholds can be used to discriminate functional lncRNAs.
21. Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L.-L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.* **12**, R16 (2011).
22. Carrieri, C. et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* **491**, 454–457 (2012).
23. Guttman, M. et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
24. Herzog, H. et al. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* **11**, 1110–1122 (2015).
25. Haerty, W. & Ponting, C. P. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* **21**, 320–332 (2015).
26. Mason, M. K. et al. Retinoic acid-independent expression of Meis2 during autopod patterning in the developing bat and mouse limb. *Evodevo* **6**, 6 (2015).
27. Lagarde, J. et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* **49**, 1731–1740 (2017).
This study describes the method of CLS for mapping full-length transcript models in human and mouse samples.
28. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature* **470**, 284–288 (2011).
29. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
30. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
31. Kanitz, A. et al. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.* **16**, 150 (2015).

32. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
33. 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).
34. 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).
35. Melé, M. et al. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res.* **27**, 27–37 (2017).
36. Lanzós, A. et al. Discovery of cancer driver long noncoding RNAs across 1112 tumour genomes: new candidates and distinguishing features. *Sci. Rep.* **7**, 41544 (2017).
37. Juul, M. et al. Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate. *eLife* **6**, e21778 (2017).
38. Tan, J. Y. et al. *cis* -acting complex-trait-associated lincRNA expression correlates with modulation of chromosomal architecture. *Cell Rep.* **18**, 2280–2288 (2017).
39. Gong, J. et al. A functional polymorphism in lnc-LAMC2-1:1 confers risk of colorectal cancer by affecting miRNA binding. *Carcinogenesis* **37**, 443–451 (2016).
40. de Kok, J. B. et al. DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. *Cancer Res.* **62**, 2695–2698 (2002).
41. Tilgner, H. et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* **22**, 1616–1625 (2012).
42. Anderson, D. M. et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* **160**, 595–606 (2015).
43. Zhou, K. I. et al. N^6-methyladenosine modification in a long noncoding RNA hairpin predisposes its conformation to protein binding. *J. Mol. Biol.* **428**, 822–833 (2016).
44. Iyer, M. K. et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
This publication describes MiTranscriptome, the largest annotation to date based on transcriptome assembly using thousands of tumour RNA-seq samples.
45. Hon, C.-C. et al. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543**, 199–204 (2017).
46. Carninci, P. et al. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37**, 327–336 (1996).
47. You, B.-H., Yoon, S.-H. & Nam, J.-W. High-confidence coding and noncoding transcriptome maps. *Genome Res.* **27**, 1050–1062 (2017).
This study first attempted the automated annotation of full-length transcripts using CAGE and 3P-seq data.
48. Mele, M. et al. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
49. Jan, C. H., Friedman, R. C., Ruby, J. G. & Bartel, D. P. Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. *Nature* **469**, 97–101 (2011).
50. Harrow, J. et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
This report represents the reference publication for the GENCODE annotation of protein-coding and non-coding genes.
51. Apweiler, R. et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**, 115D–119 (2004).
52. Sonnhammer, E., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* **26**, 320–322 (1998).
53. 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).
54. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
55. Hudson (Chairperson), T. J. et al. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
56. Adams, D. et al. BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.* **30**, 224–226 (2012).
57. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
58. Pruitt, K. D. et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–D763 (2014).
59. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
60. The RNAcentral Consortium. RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Res.* **45**, D128–D134 (2017).
61. Volders, P.-J. et al. An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.* **43**, D174–D180 (2015).
62. Ma, L. et al. LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs. *Nucleic Acids Res.* **43**, D187–D192 (2015).
63. Ezkurdia, I. et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.* **23**, 5866–5878 (2014).
64. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. & Siebert, P. D. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* **30**, 892–897 (2001).
65. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38**, e131–e131 (2010).
66. Hardwick, S. A. et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods* **13**, 792–798 (2016).
A groundbreaking study using artificial spliced RNAs from a simulated genome as a gold standard by which to evaluate the sensitivity and specificity of transcriptome assembly methods.
67. 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).
68. Steijger, T. et al. Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
A key resource benchmarking the ability of a range of transcriptome assembly tools to recall annotated exons and transcripts, highlighting their overall poor performance.
69. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
70. Shao, M. & Kingsford, C. Scallop enables accurate assembly of transcripts through phasing-preserving graph decomposition. Preprint at *bioRxiv*, 123612 (2017).
71. Liu, S. J. et al. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol.* **17**, 67 (2016).
72. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013).
An early detailed view of human transcriptome sequencing using PacBio long-read technology, which established benchmarks for error rates, read lengths and sensitivity in detecting known and novel transcripts.
73. Weirather, J. L. et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6**, 100 (2017).
74. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).
75. Byrne, A. et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8**, 16027 (2017).
76. Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R. & Akeson, M. Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing. Preprint at *bioRxiv*, 132274 (2017).
77. Garalde, D. R. et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
An early glimpse of unlimited-length direct RNA-seq using nanopore technology.
78. Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D. & Ragoussis, J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci. Rep.* **6**, 31602 (2016).
79. Housman, G. & Ulitsky, I. Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. *Biochim. Biophys. Acta* **1859**, 31–40 (2016).
80. Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
81. Mercer, T. R. et al. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* **9**, 989–1009 (2014).
82. Mercer, T. R. et al. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* **30**, 99–104 (2012).
Description of the RNA CaptureSeq method, identifying novel isoforms of deeply-studied protein-coding and lncRNA genes.
83. Clark, M. B. et al. Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat. Methods* **12**, 339–342 (2015).
84. Bussotti, G. et al. Improved definition of the mouse transcriptome via targeted RNA sequencing. *Genome Res.* **26**, 705–716 (2016).
85. Deveson, I. W. et al. Universal alternative splicing of noncoding exons. *Cell Syst.* **6**, 245–255.e5 (2018).
86. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl Acad. Sci. USA* **111**, 9869–9874 (2014).
87. Nellore, A. et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol.* **17**, 266 (2016).
Describes intropolis, a large-scale data set of splice junctions from essentially all short-read RNA-seq experiments to date, which suggests that the number of splice junctions in the human genome can be exhaustively mapped.
88. Seemann, S. E. et al. The identification and functional annotation of RNA structures conserved in vertebrates. *Genome Res.* **27**, 1371–1383 (2017).
A rigorous data set of evolutionarily conserved structures in lncRNA exons, sure to be of value in future efforts to map their functional elements.
89. Bartonicek, N. et al. Intergenic disease-associated regions are abundant in novel transcripts. *Genome Biol.* **18**, 241 (2017).
90. Saini, H. K., Griffiths-Jones, S. & Enright, A. J. Genomic analysis of human microRNA transcripts. *Proc. Natl Acad. Sci. USA* **104**, 17719–17724 (2007).
91. Jaffe, A. E. et al. Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nat. Neurosci.* **18**, 154–161 (2014).
92. Gerrard, D. T. et al. An integrative transcriptomic atlas of organogenesis in human embryos. *eLife* **5**, e15565 (2016).
93. Ahn, R. S. et al. Transcriptional landscape of epithelial and immune cell populations revealed through FACS-seq of healthy human skin. *Sci. Rep.* **7**, 1343 (2017).
94. Wright, J. C. et al. Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat. Commun.* **7**, 11778 (2016).
A description of how large-scale peptidomic data sets can be used at controlled false-discovery rates to identify misidentified protein-coding transcripts amongst lncRNA annotations.
95. Gonzalez-Porta, M., Calvo, M., Sammeth, M. & Guigo, R. Estimation of alternative splicing variability in human populations. *Genome Res.* **22**, 528–538 (2012).
96. Kornienko, A. E. et al. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol.* **17**, 14 (2016).
97. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* **13**, R107 (2012).
98. Kapusta, A. et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* **9**, e1003470 (2013).

99. Kasowski, M. et al. Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).
100. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
101. Sen, R., Doose, G. & Stadler, P. Rare splice variants in long non-coding RNAs. *Non-Coding RNA* **3**, 23 (2017).
102. Nguyen, Q. & Carninci, P. Expression specificity of disease-associated lncRNAs: toward personalized medicine. *Curr. Top. Microbiol. Immunol.* **394**, 237–258 (2016).
103. Kanehisa, M. et al. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205 (2014).
104. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
105. Kibbe, W. A. et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* **43**, D1071–D1078 (2015).
106. Yu, G. et al. BRWLDA: bi-random walks for predicting lncRNA-disease associations. *Oncotarget* **8**, 60429–60446 (2017).
107. Zhang, J., Zhang, Z., Wang, Z., Liu, Y. & Deng, L. Ontological function annotation of long non-coding RNAs through hierarchical multi-label classification. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btx833 (2017).
108. Guo, X. et al. Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucleic Acids Res.* **41**, e35 (2013).
109. Ning, S. et al. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.* **44**, D980–D985 (2016).
110. Carlevaro-Fita, J. et al. Unique genomic features and deeply-conserved functions of long non-coding RNAs in the Cancer LncRNA Census (CLC). Preprint at *bioRxiv*, 152769 (2017).
111. Kaewsapsak, P., Shechner, D. M., Mallard, W., Rinn, J. L. & Ting, A. Y. Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *eLife* **6**, e29224 (2017).
112. Mas-Ponte, D. et al. LncATLAS database for subcellular localisation of long noncoding RNAs. *RNA* **23**, 1080–1087 (2017).
113. Benoit Bouvrette, L. P. et al. CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in *Drosophila* and human cells. *RNA* **24**, 98–113 (2018).
114. Cabili, M. N. et al. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* **16**, 20 (2015).
115. Lubelsky, Y. & Ulitsky, I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. Preprint at *bioRxiv*, 189746 (2017).
116. Carlevaro-Fita, J., Das, M., Polidori, T., Navarro, C. & Johnson, R. Ancient exapted transposable elements promote nuclear enrichment of long noncoding RNAs. Preprint at *bioRxiv*, 189753 (2017).
117. Zhang, B. et al. A novel RNA motif mediates the strict nuclear localization of a long noncoding RNA. *Mol. Cell. Biol.* **34**, 2318–2329 (2014).
118. Marín-Béjar, O. et al. The human lncRNA LINC-PINT inhibits tumor cell invasion through a highly conserved sequence element. *Genome Biol.* **18**, 202 (2017).
119. Guttman, M. & Rinn, J. L. Modular regulatory principles of large non-coding RNAs. *Nature* **482**, 339–346 (2012).
120. Smola, M. J. et al. SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proc. Natl Acad. Sci. USA* **113**, 10322–10327 (2016).
121. Fang, R., Moss, W. N., Rutenberg-Schoenberg, M. & Simon, M. D. Probing Xist RNA structure in cells using Targeted Structure-Seq. *PLoS Genet.* **11**, e1005668 (2015).
122. Hawkes, E. J. et al. COOLAIR antisense RNAs form evolutionarily conserved elaborate secondary structures. *Cell Rep.* **16**, 3087–3096 (2016).
123. Xue, Z. et al. A G-rich motif in the lncRNA Braveheart interacts with a zinc-finger transcription factor to specify the cardiovascular lineage. *Mol. Cell* **64**, 37–50 (2016).
124. Lee, S. et al. Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. *Cell* **164**, 69–80 (2016).
125. Li, J.-H., Liu, S., Zhou, H., Qu, L.-H. & Yang, J.-H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **42**, D92–D97 (2014).
126. Paraskevopoulou, M. D. et al. DIANA-LncBasev2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Res.* **44**, D231–D238 (2016).
127. Buske, F. A., Bauer, D. C., Mattick, J. S. & Bailey, T. L. Triplex-Inspector: an analysis tool for triplex-mediated targeting of genomic loci. *Bioinformatics* **29**, 1895–1897 (2013).
128. Kelley, D. R., Hendrickson, D. G., Tenen, D. & Rinn, J. L. Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biol.* **15**, 537 (2014).
129. Kapranov, P. et al. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
130. Carninci, P. et al. The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
131. Carninci, P. et al. Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* **13**, 1273–1289 (2003).
132. Khalil, A. M. et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA* **106**, 11667–11672 (2009).
133. Jia, H. et al. Genome-wide computational identification and manual annotation of human long noncoding. *RNA genes. RNA* **16**, 1478–1487 (2010).
134. Cabili, M. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
135. [No authors listed.] HAVANA Annotation Guidelines, Version 2.4. *Wellcome Sanger Institute* ftp://ftp.sanger.ac.uk/pub/project/havana/Guidelines/Guidelines_March_2016.pdf (2016).
136. Wucher, V. et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.* **45**, gkw1306 (2017).
137. Dinger, M. E., Pang, K. C., Mercer, T. R. & Mattick, J. S. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.* **4**, e1000176 (2008).
138. Huang, J.-Z. et al. A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer growth. *Mol. Cell* **68**, 171–184.e6 (2017).
139. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802 (2011).
140. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *eLife* **3**, e03523 (2014).
141. Mackowiak, S. D. et al. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* **16**, 179 (2015).
142. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**, 240–251 (2013).
143. Carlevaro-Fita, J., Rahim, A., Guigó, R., Vardy, L. A. & Johnson, R. Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. *RNA* **22**, 867–882 (2016).
144. Banfai, B. et al. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22**, 1646–1657 (2012).
145. Verheggen, K. et al. Noncoding after all: biases in proteomics data do not explain observed absence of lncRNA translation products. *J. Proteome Res.* **16**, 2508–2515 (2017).
One of several studies that carefully examines proteomic evidence for productive translation of lncRNAs.
146. Bruford, E. A., Lane, L. & Harrow, J. Devising a consensus framework for validation of novel human coding loci. *J. Proteome Res.* **14**, 4945–4948 (2015).
147. Wang, L. et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74 (2013).
148. Kong, L. et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **35**, W345–W349 (2007).
A pioneering bioinformatic tool for the discrimination of protein-coding and non-coding transcripts, in this case using an alignment-free sequence-feature and homology strategy.
149. Nelson, B. R. et al. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* **351**, 271–275 (2016).
150. Ma, J. et al. Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J. Proteome Res.* **13**, 1757–1765 (2014).
151. Gibb, E. A. et al. Activation of an endogenous retrovirus-associated long non-coding RNA in human adenocarcinoma. *Genome Med.* **7**, 22 (2015).
152. Gascoigne, D. K. et al. Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. *Bioinformatics* **28**, 3042–3050 (2012).
153. Ezkurdia, I. et al. The potential clinical impact of the release of two drafts of the human proteome. *Expert Rev. Proteom.* **12**, 579–593 (2015).
154. Lopez, F., Granjeaud, S., Ara, T., Ghattas, B. & Gautheret, D. The disparate nature of "intergenic" polyadenylation sites. *RNA* **12**, 1794–1801 (2006).
155. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Curr. Protoc. Bioinformatics* https://doi.org/10.1002/0471250953.bi0403s18 (2007).

Acknowledgements
R.J. acknowledges the support of the Swiss National Science Foundation through the National Centres for Competence in Research (NCCR) 'RNA & Disease' and the Medical Faculty of the University Hospital and University of Bern. The authors thank J. Carlevaro-Fita (University of Bern) for help with data analysis and J. Harrow (Illumina), J. Mudge (European Bioinformatics Institute), P. Flicek (European Bioinformatics Institute) and I. Jungreis (Massachusetts Institute of Technology) for fruitful discussions and feedback. A.F. is supported by the Wellcome Trust (WT098051 and WT108749/Z/15/Z), the National Human Genome Research Institute (NHGRI) (U41HG007234, 2U41HG007234) and the European Molecular Biology Laboratory. Work described in this publication was 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.). Work in the laboratory of R.G. was supported by the National Human Genome Research Institute (awards U54HG007000, R01MH101814 and U41HG007234), the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017' and CERCA Programme/Generalitat de Catalunya. The authors thank the following individuals for administrative support: R. Carrido (Centre for Genomic Regulation) and S. Roesselet and D. Re (both at the University of Bern).

Author contributions
B.U.-R. and R.J. researched data for the article. B.U.-R., A.F. and R.J. wrote the article. All authors provided substantial contributions to discussions of the content and reviewed and/or edited the manuscript before submission.

Competing interests
The authors declare no competing interests.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reviewer information
Nature Reviews Genetics thanks M. Dinger, I. Ulitsky and the other, anonymous reviewer(s) for their contribution to the peer review of this work.

RELATED LINKS
buildLoci: https://github.com/julienlag/buildLoci
GENCODE: www.gencodegenes.org
lncrna.annotator: https://github.com/gold4ab/shared_scripts/tree/master/lncRNA.annotator
UniProt: http://www.uniprot.org/
Pfam: http://pfam.xfam.org/

Discussion

I

Shedding light on the deep transcriptome with RACE-Seq and Capture Long-read Sequencing

The present work introduces two highly sensitive techniques for the detection and annotation of long noncoding RNAs: RACE-Seq (Publication I[321]) and Capture Long-read Sequencing (CLS, Publication II[322]). Both methodologies combine targeted transcriptome enrichment to medium- (RACE-Seq) and long- (CLS) read sequencing technologies, in order to improve the detection and delineation of lncRNA transcripts in mammalians. The choice of sequencing technologies – Roche GS-FLX 454+ for RACE-Seq, PacBio RSII for CLS – was dictated by their output read length and commercial availability when the projects started. Although Publication II is based on experiments and analyses performed in both human and mouse, this Discussion focuses primarily on the results obtained in human.

I.1. The RACE-Seq proof-of-concept

RACE (Rapid Amplification of cDNA Ends) is an established molecular biology method. It has been used for many years in a low-throughput context to precisely map individual transcript termini. While the idea of multiplexing RACE products for large-scale sequencing is not novel[136], the present work does it on an unprecedented scale – 398 targeted GENCODE v7 human lncRNA loci, in both 5′ and 3′ directions – while combining it with a longer-read sequencing technology. We were able to quantify the advantage of adding a sequential nested step to the RACE protocol: across all 7 human tissues assayed, nested RACE provided a mean 9.5-fold increase in target specificity compared to standard, single-oligonucleotide RACE. RACE-Seq coupled to 454 sequencing (yielding a median read length of ∼600 nts) led to the discovery of 2,556 novel transcripts overlapping targeted regions. As a result, 238 genes (60%) saw their boundaries extended in either 5′ or 3′ – 131 (33%) in both directions – after manual curation of the mapped 454 data. By stratifying the targeted

loci into 5′ CAGE-supported and CAGE-unsupported categories (see Introduction, Section II.1.4), our study shows that genes belonging to the former class are much less likely to be extended by RACE, confirming the robustness of CAGE clusters as predictors of transcription start sites. Within targeted loci, RACE-Seq discovers a wealth of previously unannotated features, including 615 novel TSSs – of which 252 (41%) are supported by CAGE – while bringing up the number of annotated isoforms per locus ∼5-fold overall. After re-annotation of the 398 loci, their median mature transcript length increased slightly (623 *vs* 704 nts), although not significantly.

Taken together, results from Publication I indicate that (1), GENCODE v7 gene models do not provide an entirely faithful representation of human lncRNA transcripts, and (2), RACE-Seq enables a better definition of their 5′ and 3′ termini, and reveals the extent of alternative splicing within these loci. Nonetheless, this study, because it was designed as a proof-of-concept, suffers from a few biases that prevent the extrapolation of its results to the entire human lncRNA catalog. First, the chosen 398 target loci are relatively highly expressed, with a median RPKM (Reads Per Kilobase of exon per Million mapped reads) of 8.3, compared to <1 RPKM across the whole lncRNA population, as reported by Derrien *et al.*[268]. Second, the known technical shortcomings of RACE (see Introduction, Section II.1.3.2), together with the limited read length of 454 sequencing, raise doubts as to the full-length nature of the RACE-Seq product sequences as a whole. Consequently, we speculate that the length and end-completeness of the resulting lncRNA transcript models are substantially underestimated. Coupling RACE-Seq with more modern, longer-read sequencing platforms such as PacBio and Oxford Nanopore has the potential to reduce the length bias likely introduced by the 454 platform. Lastly, while probably amenable to a larger set of targets, the RACE-Seq protocol still involves laborious steps that prevent this technique from reaching the throughput of CLS (see below), and is not realistically applicable to the large-scale interrogation of the unannotated genomic space.

I.2. High-throughput empirical lncRNA annotation with CLS

The CLS approach, presented in Publication II, addresses many of the limitations of RACE-Seq and other targeted transcriptome sequencing methodologies. It combines for the first time RNA capture with third-generation, long-read sequencing. Our study applies RNA capture to 9,060 human and 6,615 mouse features, totaling

∼15.5 and ∼8.3 megabases of probed regions, respectively. Included in the capture designs were 5,953 (human) and 1,920 (mouse) annotated lincRNA genes, representing 41% and 36% of the entire GENCODE lincRNA catalog, respectively. The remainder of the capture probes targeted various other types of noncoding genomic regions, from UCEs to predicted enhancers and *de novo* computational predictions. Capture libraries were enriched from cDNA libraries generated in four species-matched tissues, two murine whole embryos and two human immortalized cell lines. Aiming to reduce the length bias introduced by long-read platforms (see Introduction, Section II.1.2.1), the capture libraries were separated into three distinct size fractions, and subsequently sequenced on the Pacific Biosciences RSII platform, yielding about two million long circular consensus reads per species. The capture step allowed a substantial 19- and 11-fold on-target read enrichment compared to non-captured samples in human and mouse, respectively.

CLS reads were subjected to a fully automatized, high-throughput genome annotation pipeline. The CLS bioinformatics workflow involves various sequential steps including cDNA-to-genome mapping, read merging into transcript models (TMs), as well as a series of stringent filters aimed at enhancing the quality of the resulting annotation without manual curation (Figure 16). The 5′ and 3′ completeness of each TM was confirmed using FANTOM CAGE data[146] and DHS peaks[110] as well as polyA tails (encoded in the CLS long reads) and proximity to polyadenylation signals[284], respectively.

In human, CLS produced a collection of 179,993 non-redundant TMs overlapping genes of various GENCODE biotypes, including lincRNAs and protein-coding genes. From the 65,736 TMs that were considered full-length, 8,494 consisted in novel lncRNA structures arising from 947 probed loci, often revealing previously unannotated locus boundaries. CLS also uncovered thousands of novel transcripts within the intergenic space, including ∼18,000 in non-exonic regions, >600 overlapping enhancer predictions and >7,000 bridging regions of distinct biotypes – often involving protein-coding exons. Within detected lincRNA regions, the number of transcripts with confident TSSs increased by 58% compared to GENCODE (2,607 *vs* 1,650 CAGE-supported TMs, respectively).

A random sample of 240 CLS TMs, stratified by level of confidence (*i.e.*, intron HiSeq support, 5′ CAGE and 3′ polyA support) were assessed by GENCODE annotation curators, post-publication. TMs supported by at least one of the three kinds of evidence had a validation rate of >96%, while only 62% of unsupported transcript models passed manual inspection [1]. Although performed on a small sample, this

[1] Jose-Manuel González, Jonathan Mudge and Adam Frankish, European Bioinformatics Institute, per-

clearly highlights the robustness of our data processing pipeline and indicates that it approaches the quality of manual annotation.

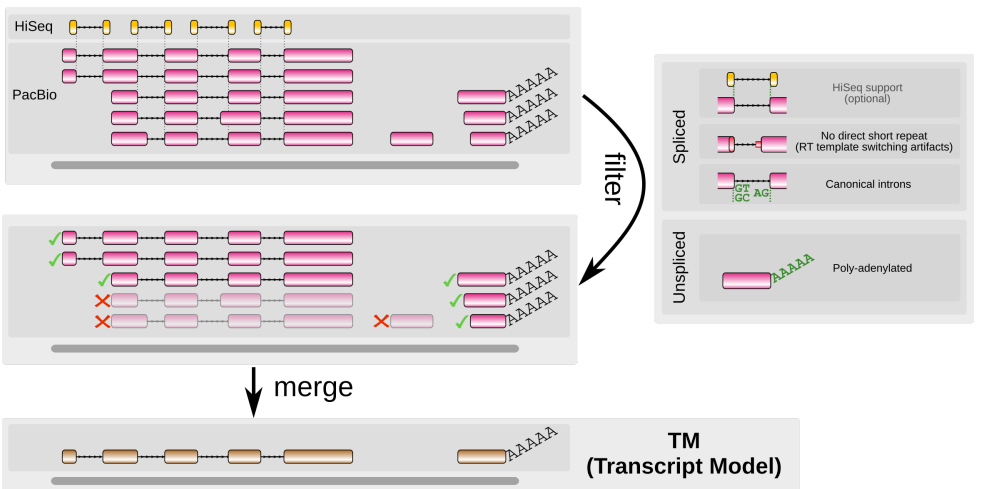


Figure 16: PacBio read processing in CLS. Reads from cDNA capture libraries – including both long PacBio CCS (pink) and short Illumina Hiseq reads (orange) – are mapped to the genome. Each PacBio read mapping subsequently undergoes a series of filters designed to guarantee that the downstream annotation is free of artifacts possibly arising from cDNA synthesis and/or low sequencing quality (see Introduction, Section II.1.2.1). Spliced PacBio reads are required to be devoid of non-canonical introns. Support of individual splice junctions by HiSeq split reads can be included as an additional requirement. Spliced reads presenting dubious introns containing direct short repeats are eliminated, as they may represent spurious products of RT template switching during cDNA synthesis. Unspliced reads, on the other hand, must be polyadenylated to pass validation, in order to remove genomic DNA contamination artifacts. Reads that passed validation are then merged into non-redundant transcript models (brown).

CLS introduces an original read merging methodology. This so-called "anchored" approach (or ARM, for "Anchored Read Merging") aims to preserve all transcripts with strongly supported internal TSSs and polyadenylation sites (see Figure 4b of Publication II), and as a result yields ∼54% more TMs than conventional, "Greedy" Read Merging (GRM) methods (179,993 *vs* 117,258 TMs across all CLS samples, re-

sonal communication.

spectively). An example from a recent study demonstrates the relevance of this approach. Using NET-CAGE (Native Elongating Transcript-CAGE), Hirabayashi and colleagues were able to clearly detect two alternative promoters in the *ZNHIT1* human locus, separated by ∼500 bp. NET-CAGE data shows that RNA products of the upstream ("minor") promoter degrade much more rapidly than those emerging from the downstream ("major") promoter. The authors proceeded to leverage the CLS ARM annotation generated in the present work to establish that the two promoters give rise to two distinct transcript isoforms, differing only by the length of their 5′ UTR[323]. As shown in Figure 17, the isoform generated from the major promoter would have disappeared from the CLS annotation had we used a GRM procedure. This testifies to the relevance of the anchored merging approach for downstream applications, as highlighted here with this study of RNA stability mediated by 5′ UTR length.

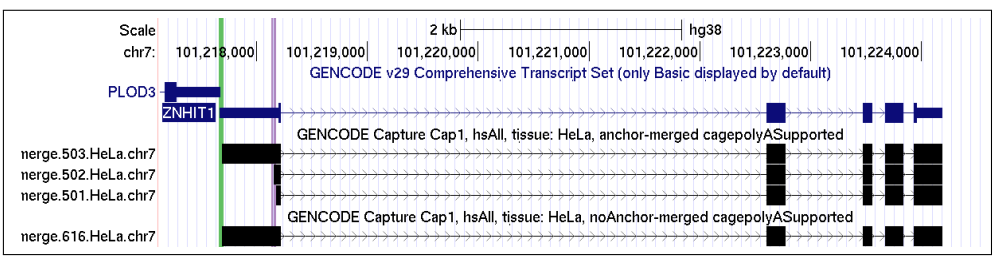


Figure 17: Anchored (ARM) *vs* greedy (GRM) read merging in the *ZNHIT1* human protein-coding locus. Using NET-CAGE in HeLa cells, Hirabayashi *et al.* identify two distinct promoters (depicted as vertical bars) in this gene: a "major" one (purple) and a "minor" one (green), corresponding to distinct short and long transcript isoforms, respectively[323]. GENCODE v29 annotates only the long transcript isoform (top track, blue) in this region, as it relies on the greedy approach. CLS yields full-length reads overlapping this (untargeted) locus in HeLa cells (bottom tracks, black). However, a GRM procedure masks the short isoform, by incorporating it into the longer one ("*noAnchor-merged*" CLS track). In contrast, the ARM transcript set ("*anchor-merged*" CLS track) preserves both the long and short isoforms in the final annotation set.

Importantly, third-generation long-read sequencing allowed us to bypass *in silico* transcript reconstruction, leading to more accurate TMs. We were able to quantify that advantage by comparing CLS-derived TMs to transcripts reassembled by the gold-standard StringTie software[78] using matched, deep Illumina HiSeq capture li-

braries as input. Within targeted lincRNA regions, CLS TMs showed spectacularly superior quality when compared to StringTie models. Notably, while 4,763 (∼20%) CLS TMs were deemed full-length, only 116 (∼0.8%) StringTie models exhibited such characteristic. Our results suggest that this is due to StringTie's tendency to over-extend transcript models beyond their true 5′ and 3′ boundaries. Overall, this analysis unequivocally confirms and quantifies, for the first time, the clear superiority of TGS over SGS methods in the context of noncoding gene annotation.

In summary, Publication II demonstrates the value of RNA capture coupled with third-generation long-read PacBio sequencing for annotating the deep transcriptome. CLS matches the quality of manual annotation for a much lower cost, with the throughput of SGS-based annotation. While RACE-Seq and CLS are here exploited in the context of lncRNA annotation, applications to any low-abundance transcript species are possible. For example, both techniques could be used to probe and delineate the full isoform complement of subsets of protein-coding genes, or adapted to a wide range of organisms other than human and mouse. In contrast to RACE-Seq, which relies on single oligonucleotides as probing units, CLS can efficiently target large genomic regions. It can also be easily scaled up: the latest commercialized capture kits allow custom designs of up to 100 megabases of tiled genome sequence [2].

In future implementations, the CLS experimental workflow could benefit from recent technological improvements. The completeness of cDNAs can be enhanced through the use of 5′ cap enrichment techniques such as CapTrap – the technique pioneered at the RIKEN Institute to generate CAGE libraries. Indeed, preliminary results indicate that CapTrap does boost the completeness rate of CLS transcript models almost two-fold (data not shown). Employing alternative, non-oligo-dT cDNA sequencing techniques may also uncover a profusion of non-polyadenylated transcripts, and could help improve the low rate of enhancer RNA detection observed in the present study. However, such libraries might contain a large proportion of immature, partially spliced RNAs of little interest to genome annotation.

TGS technologies have also evolved enormously since this project was carried out. Nanopore technologies offer a democratic, cheaper, but more error-prone alternative to PacBio sequencing. PacBio, on the other hand, has made substantial improvements to its product line. Its latest-generation Sequel II platform promises longer reads without size selection, base accuracy matching that of Illumina HiSeq, and yields of up to four million reads per run (instead of ∼30,000 with the RS II instrument) for a

[2] https://sequencing.roche.com/en/products-solutions/by-category/target-enrichment/hybridization/seqcap-rna-choice.html (September 2019).

much-reduced per-read cost.

II

An updated view of lncRNA genes

The proper genomic and transcriptomic characterization of lncRNA genes is crucial to the understanding of their biological roles and regulation. The present work contributes a much-needed, exceptionally detailed – although still preliminary – resource to investigate the landscape of lncRNA genome biology. Our full-length CLS transcript catalog provides the most confident view, to date, of the transcriptional and post-transcriptional regulation of those genes.

Full-length CLS data shows that lncRNA mature transcripts are likely much longer than previously assumed (median length: 668 nts in GENCODE v21 *vs* 1,108 nts in CLS), but slightly shorter than protein-coding transcripts (media length: 1,240 nts observed in CLS). A definitive statement as to the length of lncRNAs based on CLS would be premature, though, as the read length biases associated with the PacBio RSII platform may not provide a fully representative picture of the lncRNA transcriptome.

Like Derrien *et al.*'s analysis[268], CLS confirms that lncRNA splice sites are of similar strength compared to those of protein-coding genes. The lncRNAs detected by CLS are, to a large extent, polyadenylated via canonical sequence signals. Since it relies on oligo-dT-based cDNA synthesis in its current form, CLS is unfortunately agnostic to non-polyA transcripts, including the majority of enhancer RNAs (see Introduction, Section III.2.1). Hence, by design, CLS still provides an incomplete picture of long noncoding RNA biogenesis.

A fundamental question is whether expanding the GENCODE lncRNA catalog, as CLS does, reveals potential, hitherto unannotated open reading frames. We ad-

dressed this by searching *in silico* for signatures of coding sequences within the CLS transcript collection. To this end, we employed two software tools that use distinct, orthogonal approaches, CPAT (Coding Potential Assessment Tool, which uses an alignment-free logistic regression model)[324] and PhyloCSF (a comparative genomics method)[325]. Results from both methods globally concur, and indicate that only a handful of CLS lncRNAs may code for proteins. It is however possible that none of these methods are sensitive enough to properly detect very small ORFs sometimes translated from lncRNAs (see Introduction, Section III.3.2.4). A recently published approach, based on relating codon frequencies in lncRNA sequences to tRNA abundances in the cell[326] may help further clarify the issue of the coding potential of lncRNAs.

While our work confirms the overall low evolutionary conservation of lncRNA exons reported previously, we detect clear signs of evolutionary constraint at some of their TSSs, even when eliminating bi-directional promoters shared with protein-coding genes from the data. This immediately suggests a regulatory role for the lncRNAs under these promoters' control, possibly through "act-of-transcription" *cis* mechanisms (see Introduction, Section III.3.2.4).

The CLS high-confidence TSS collection enabled an extensive characterization of lncRNA promoters' epigenomic environment. Using ENCODE ChIP-Seq data in matched human cell lines (HeLa and K562), and controlling for gene expression levels, we show that lncRNA promoters share similar levels of active chromatin marks H3K4me3 and H3K9ac (acetylation of histone 3 at lysine 9) as those of protein-coding genes, contradicting previous reports[274,285]. This strongly suggests that the observed differences in these latter studies were at least partly confounded by the 5′ incompleteness of the underlying annotation.

We also report features distinguishing lncRNA from protein-coding promoters. Intriguingly, our analysis reveals that in HeLa, lncRNA promoters show more elevated levels of CTCF (CCCTC-binding factor) binding, a multitasking DNA-binding protein involved in transcriptional regulation, gene insulation and three-dimensional genome organization[327]. In addition, and as noted elsewhere[274,285], lncRNA promoters generally exhibit higher levels of marks associated with repressed chromatin: in HeLa cells, we detect enriched signals of the H3K9me3 (trimethylation of histone 3 at lysine 9) and H3K27me3 (trimethylation of histone 3 at lysine 27) modifications around lncRNA promoters – consistent with the concomitantly elevated levels of the EZH2 catalytic subunit of the Polycomb repressive complex 2 (PRC2) that deposits H3K27me3 marks[328,329]. The observed coexistence of high levels of some of these marks specifically in lncRNAs – while they are generally inversely corre-

lated in protein-coding gene promoters – seems inconsistent, however. For example, PRC2 has been shown to be inhibited in *cis* by H3K4me3 marks[330], while H3K9ac and H3K9me3, since they affect the same aminoacid residue, are mutually exclusive on the same histone molecule – although not on the same nucleosome. It is still unclear whether these apparently contradictory observations stem from averaging ChIP-Seq signals across heterogeneous cells and/or across different lncRNA gene sub-populations, or if they reflect a more biologically relevant mechanism.

III

How far are we from lncRNA annotation completeness?

The present work substantially expands existing long noncoding RNA annotations, and constitutes a significant step towards a complete, high-confidence map of lncRNAs in the human genome. There is ample evidence, presented in this work and elsewhere, that this map is nowhere near completion, however.

Publication II presents discovery - saturation curves which indicate that each step towards deeper sequencing leads to the discovery of more transcripts and splice junctions, in all interrogated tissues. Importantly, this phenomenon is also observed for the highest-confidence CLS transcript set and in both captured HiSeq and PacBio reads, suggesting that this absence of saturation is accounted for by genuine transcripts. This observation has been made elsewhere, often with much deeper datasets[124,331], and suggests that the noncoding transcriptome is still not fully sampled, even in targeted transcriptomics studies.

Recently, Deveson *et al.* used CaptureSeq, coupled with both short- (HiSeq) and long-read (PacBio RSII) sequencing, to probe the entire human chromosome 21, in-

cluding both lncRNAs and protein-coding genes. They report what they call "universal alternative splicing" – *i.e.*, near-infinite combinations of exons into transcript structures – within lncRNAs, but not in protein-coding genes, where the number of observed combinations of exons rapidly saturates[331]. Under this model, reminiscent of Borges' *Library of Babel* story [3], noncoding genes constitute a virtually limitless reservoir of distinct transcript structures. It remains to be determined whether universal alternative splicing of noncoding exons is a biologically relevant phenomenon, or the result of relaxed constraints on largely functionless RNAs. Regardless, it makes the complete cataloguing of the lncRNA transcriptome a daunting, almost Sisyphean task (see also the Commentary on this study in Appendix, Additional relevant publication, page 213).

Publication III[332] introduces the concepts of *completeness* (*i.e.*, fraction of full-length transcripts), *exhaustiveness* (*i.e.*, number of isoforms per locus) and *comprehensiveness* (*i.e.*, total number of loci) to characterize an annotation set (Figure 18) [4]. We use these metrics to compare various public lncRNA resources – NONCODE[265], MiTranscriptome[262], FANTOM CAT[266], RefSeq[71], GENCODE v27[36], BIGTranscriptome[264] and the CLS annotation generated in the present work, automatically merged into GENCODE ("GENCODE+"). Comprehensiveness and exhaustiveness are impossible to measure, since the total number of lncRNA isoforms and loci is unknown. Nevertheless, a rough, relative estimate of their values can be obtained for each catalog, by comparing it to its counterparts. To ensure a fair comparison of their comprehensiveness, and since the definition of a gene locus varies depending on the resource, we uniformly re-merged all transcripts from each dataset into loci using an original piece of software (see '`buildLoci`' in Appendix, Section IV).

What stands out from our analyses, as presented in Publication III, is that the competing annotation sets show little overlap. Even GENCODE and RefSeq, the two most widely used reference catalogs, have less than 50% of lncRNA genes in common. Unsurprisingly, NONCODE, which integrates most of the other resources, is the most comprehensive, with 67,276 gene loci. Across datasets, we observe a negative correlation between comprehensiveness and completeness. For example, the two most comprehensive catalogs, NONCODE and MiTranscriptome, show extremely large (>91%) proportions of incomplete transcripts, consistent with their being mainly composed of software-reconstructed gene models. As judged by the difference between GENCODE and GENCODE+, CLS substantially improves GEN-

[3] Jorge Luis Borges, *Ficciones*, 1944.

[4] Note that outside of this section, these three terms are employed in their more general sense.

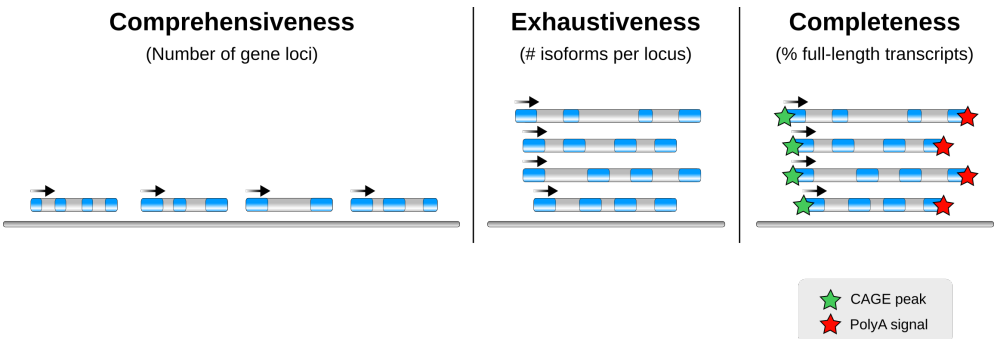


Figure 18: Comprehensiveness, exhaustiveness and completeness of gene annotation catalogs. In this analysis, transcript models are deemed full-length if they are bounded by FANTOM CAGE peaks[146] and polyadenylation signals[284] at their 5′ and 3′ ends, respectively.

CODE lncRNA transcripts' completeness (bringing it from 13.5% to 24%) and exhaustiveness (from 1.9 to 3.3 isoforms per locus).

With regard to annotation completeness, remarkably, none of the lncRNA catalogs under scrutiny come close to matching the gold standard of GENCODE protein-coding genes (53.8%). Recent improvements in cDNA library preparation and targeted TGS methods (see Section I.2) will likely make this target achievable for the GENCODE lncRNA catalog in the near future. In terms of number of annotated loci, GENCODE still lags far behind deep collections such as MiTranscriptome and NONCODE. Enhancing GENCODE's comprehensiveness could be accomplished by applying CLS to the union of all genomic bases covered by the aforementioned lncRNA resources. It is noteworthy that these resources, like CLS, are mostly based on oligo-dT cDNA synthesis, and therefore ignore the non-polyadenylated transcriptome, some of which may be functional[333].

Other sources of annotation incompleteness – in the general sense of the word – are harder to tackle. Subtle fluctuations in gene expression during an organism's lifetime may render the detection of transiently expressed transcripts difficult. Recent work has demonstrated the bursty nature of transcription in several organisms, including mammalians[334–336]. The interplay between bursty RNA production and degradation can result in temporary, isolated peaks of gene expression[337]. As a consequence, since all RNA libraries are essentially a snapshot of the cells' RNA steady-state levels, one can speculate that detecting all transcripts, expressed during development or under certain physiological conditions, would require a near-infinite number of biological samplings across developmental time points and conditions. This is,

of course, largely unfeasible by current – or near-future – standards. Nonetheless, these issues are beginning to be addressed[313,314], and there is hope that the development of single-cell RNA-Seq techniques[122,338–341] will enable a better representation of an entire organism's transcriptome across cell types and conditions.

Conclusion

The main conclusions of the present work are the following:

1. We introduce the Capture Long-read Sequencing (CLS) approach, a high-throughput methodology designed to annotate long noncoding RNAs (lncRNAs) in mammalian genomes. CLS couples RNA capture and third-generation, long-read sequencing for the first time. We demonstrate that this method matches manual annotation in quality, while bypassing the costly and time-consuming bottleneck of human curation. Our data also shows and quantifies the clear superiority of CLS over short-read-based transcript assembly methods.

2. Through the application of CLS – and to a lesser extent, RACE-Seq – the present study substantially improves the quality of the GENCODE reference long noncoding RNA annotation in the human genome. The CLS methodology reveals a wealth of novel long noncoding RNA structures, and produces a much-improved definition of their transcript boundaries. CLS thus provides a significantly more robust foundation for the functional characterization of human long noncoding RNAs.

3. CLS enables a confident re-assessment of important lncRNA gene properties:

- Annotation expansion does not reveal hitherto undetected coding potential in lncRNAs. Thus, overall, lncRNAs are indeed noncoding.
- Mature lncRNA transcripts are likely much longer than previously anticipated, and their length approaches that of protein-coding transcripts.
- The chromatin environment of lncRNA promoters is similar in terms of activating marks, but is enriched in repressive modifications compared to that of protein-coding gene promoters.

4. Compared to other public human lncRNA collections, the CLS-augmented GENCODE lncRNA catalog provides a highly competitive compromise between accuracy, exhaustiveness (number of annotated transcript isoforms per locus) and completeness (proportion of full-length transcript models). However, presumably because CLS focused mainly on already annotated regions in the present work, GENCODE+ still lacks in exhaustiveness (number of annotated gene loci).

Bibliography

1. Mendel, G. Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn* **IV,** 3–47 (1865).
2. Johannsen, W. Elemente der exakten Erblichkeitslehre. *Jena, G. Fischer* (1909).
3. Henig, R. M. *The Monk in the Garden: The Lost and Found Genius of Gregor Mendel, the Father of Genetics* (Boston: Houghton Mifflin, 2000).
4. Sutton, W. On the morphology of the chromosome group in Brachystola magna. *Biological Bulletin* **4,** 24–39 (1902).
5. Sutton, W. The chromosomes in heredity. *Biological Bulletin* **4,** 231–251 (1903).
6. Boveri, T. Ergebnisse über die Konstitution der chromatischen Substanz des Zelkerns. *Jena, G. Fischer* (1904).
7. Carothers, E. E. The mendelian ratio in relation to certain orthopteran chromosomes. *Journal of Morphology* **24,** 487–511 (1913).
8. Morgan, T. H., Sturtevant, A. H., Muller, H. & Bridges, C. *The mechanism of Mendelian heredity* (New York : H. Holt and company, 1915).
9. Sturtevant, A. H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology* **14,** 43–59 (1913).
10. Kossel, A. Weitere Beiträge zur Chemie des Zellkerns. *Zeitschrift für Physiologische Chemie* **10,** 148–264 (1886).
11. Avery, O. T., MacLeod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *The Journal of Experimental Medicine* **79,** 137–158 (Feb. 1, 1944).
12. Hershey, A. D. & Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology* **36,** 39–56 (Sept. 20, 1952).
13. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171,** 737 (Apr. 1953).
14. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3,** 318–356 (June 1961).
15. Mayr, E. Cause and Effect in Biology: Kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science* **134,** 1501–1506 (Nov. 10, 1961).
16. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74,** 560–564 (Feb. 1977).
17. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94,** 441–448 (May 25, 1975).
18. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74,** 5463–5467 (Dec. 1977).
19. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107,** 1–8 (Jan. 2016).
20. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260,** 500–507 (Apr. 8, 1976).
21. Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265,** 687–695 (Feb. 24, 1977).
22. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269,** 496–512 (July 28, 1995).
23. Goffeau, A. *et al.* Life with 6000 genes. *Science (New York, N.Y.)* **274,** 546, 563–567 (Oct. 25, 1996).
24. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science (New York, N.Y.)* **282,** 2012–2018 (Dec. 11, 1998).
25. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science (New York, N.Y.)* **287,** 2185–2195 (Mar. 24, 2000).
26. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291,** 1304–1351 (Feb. 16, 2001).
27. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (Feb. 15, 2001).
28. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (Dec. 5, 2002).
29. Wood, V. *et al.* Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biology* **9,** 180241 (Feb. 28, 2019).
30. Gerstein, M. B. *et al.* What is a gene, post-ENCODE? History and updated definition. *Genome Research* **17,** 669–681 (June 2007).
31. Stein, L. Genome annotation: from sequence to biology. *Nature Reviews Genetics* **2,** 493–503 (July 2001).
32. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science (New York, N.Y.)* **277,** 1453–1462 (Sept. 5, 1997).

33. Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O. & Ofran, Y. Automatic prediction of protein function. *Cellular and molecular life sciences: CMLS* **60,** 2637–2650 (Dec. 2003).

34. Gabaldón, T. & Huynen, M. A. Prediction of protein function and pathways in the genome era. *Cellular and molecular life sciences: CMLS* **61,** 930–944 (Apr. 2004).

35. Du Plessis, L., Škunca, N. & Dessimoz, C. The what, where, how and why of gene ontology—a primer for bioinformaticians. *Briefings in Bioinformatics* **12,** 723–735 (Nov. 2011).

36. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47,** D766–D773 (D1 Jan. 8, 2019).

37. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (Nov. 2008).

38. Kim, E., Magen, A. & Ast, G. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research* **35,** 125–131 (Jan. 2007).

39. Brent, M. R. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nature Reviews. Genetics* **9,** 62–73 (Jan. 2008).

40. Blanco, E., Parra, G. & Guigo, R. Using geneid to identify genes. *Current Protocols in Bioinformatics* **Chapter 4,** Unit 4.3 (June 2007).

41. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268,** 78–94 (Apr. 25, 1997).

42. Guigó, R. *et al.* EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biology* **7,** S2 (Suppl 1 2006).

43. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465,** 53–59 (May 2010).

44. Baralle, M. & Baralle, F. E. The splicing code. *Biosystems. Code Biology* **164,** 39–48 (Feb. 1, 2018).

45. Mayr, C. Regulation by 3'-Untranslated Regions. *Annual Review of Genetics* **51,** 171–194 (2017).

46. Leppek, K., Das, R. & Barna, M. Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nature Reviews Molecular Cell Biology* **19,** 158–174 (Mar. 2018).

47. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* **13,** 329–342 (May 2012).

48. Mignone, F., Gissi, C., Liuni, S. & Pesole, G. Untranslated regions of mRNAs. *Genome Biology* **3,** reviews0004.1–reviews0004.10 (2002).

49. Le, S. V., Chen, J. H., Currey, K. M. & Maizel, J. V. A program for predicting significant RNA secondary structures. *Computer applications in the biosciences: CABIOS* **4,** 153–159 (Mar. 1988).

50. Rivas, E. & Eddy, S. R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics (Oxford, England)* **16,** 583–605 (July 2000).

51. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–227 (Mar. 2009).

52. Gorodkin, J. & Hofacker, I. L. From structure prediction to genomic screens for novel non-coding RNAs. *PLoS computational biology* **7,** e1002100 (Aug. 2011).

53. Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics (Oxford, England)* **17 Suppl 1,** S140–148 (2001).

54. Parra, G. *et al.* Comparative Gene Prediction in Human and Mouse. *Genome Research* **13,** 108–117 (Jan. 1, 2003).

55. Gross, S. S., Do, C. B., Sirota, M. & Batzoglou, S. CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biology* **8,** R269 (2007).

56. Pedersen, J. S. *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2,** e33 (2006).

57. Washietl, S., Hofacker, I. L. & Stadler, P. F. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* **102,** 2454–2459 (2005).

58. Yao, Z., Weinberg, Z. & Ruzzo, W. L. CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics (Oxford, England)* **22,** 445–452 (Feb. 15, 2006).

59. Babak, T., Blencowe, B. J. & Hughes, T. R. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC bioinformatics* **8,** 33 (Jan. 30, 2007).

60. Meyer, I. M. A practical guide to the art of RNA gene prediction. *Briefings in Bioinformatics* **8,** 396–414 (Nov. 1, 2007).

61. Roy, S. W. & Irimia, M. When good transcripts go bad: artifactual RT-PCR 'splicing' and genome analysis. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* **30,** 601–605 (June 2008).

62. Cocquet, J., Chong, A., Zhang, G. & Veitia, R. A. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88,** 127–131 (July 2006).

63. Nordström, K. J. V. *et al.* Critical evaluation of the FANTOM3 non-coding RNA transcripts. *Genomics* **94,** 169–176 (Sept. 2009).

64. Nagaraj, S. H., Gasser, R. B. & Ranganathan, S. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics* **8,** 6–21 (Jan. 1, 2007).

65. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature Biotechnology* **26,** 1135–1145 (Oct. 2008).

66. Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science (New York, N.Y.)* **252,** 1651–1656 (June 21, 1991).

67. Hillier, L. D. *et al*. Generation and analysis of 280,000 human expressed sequence tags. *Genome Research* **6,** 807–828 (Sept. 1, 1996).

68. Gerhard, D. S. *et al.* The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Research* **14,** 2121–2127 (Oct. 2004).

69. Kawai, J. *et al.* Functional annotation of a full-length mouse cDNA collection. *Nature* **409,** 685–690 (Feb. 8, 2001).

70. Ota, T. *et al.* Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nature Genetics* **36,** 40–45 (Jan. 2004).

71. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44,** D733–745 (D1 Jan. 4, 2016).

72. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376 (Sept. 2005).

73. Bennett, S. T., Barnes, C., Cox, A., Davies, L. & Brown, C. Toward the $1000 human genome. *Pharmacogenomics* **6,** 373–382 (July 1, 2005).

74. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews. Genetics* **17,** 333–351 (2016).

75. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5,** 621–628 (2008).

76. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28,** 503–510 (2010).

77. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7,** 562–578 (Mar. 2012).

78. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33,** 290–295 (Feb. 2015).

79. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods* **10** (2013).

80. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nature biotechnology* **31,** 1009–1014 (Nov. 2013).

81. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* **4,** 265–270 (Apr. 2009).

82. Payne, A., Holmes, N., Rakyan, V. & Loose, M. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv,* 312256 (May 3, 2018).

83. Volden, R. *et al.* Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proceedings of the National Academy of Sciences* **115,** 9726–9731 (Sept. 25, 2018).

84. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6,** 100 (2017).

85. Borgström, E., Lundin, S. & Lundeberg, J. Large scale library generation for high throughput sequencing. *PloS One* **6,** e19119 (Apr. 27, 2011).

86. Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nature Biotechnology* **33,** 736–42 (May 2015).

87. Tilgner, H. *et al.* Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. *Genome Research* **28,** 231–242 (2018).

88. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics,* 1 (July 24, 2019).

89. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461,** 814–818 (Oct. 2009).

90. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods* **15,** 201–206 (2018).

91. Liu, H. *et al.* Accurate detection of m6A RNA modifications in native RNA sequences. *bioRxiv,* 525741 (Jan. 21, 2019).

92. Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R. & Akeson, M. Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PloS One* **14,** e0216709 (2019).

93. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215,** 403–10 (Oct. 1990).

94. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25,** 3389–3402 (Sept. 1, 1997).

95. Brent, M. R. Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Research* **15,** 1777–1786 (Dec. 1, 2005).

96. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Computer applications in the biosciences: CABIOS* **13,** 477–478 (Aug. 1997).

97. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome research* **12,** 656–64 (Apr. 2002).

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

99. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature methods* **9,** 1185–8 (Dec. 2012).

100. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

101. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome biology* **7 Suppl 1** (2006).

102. Frankish, A. *et al.* Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics* **16,** S2 (June 18, 2015).

103. Zhao, S. & Zhang, B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics* **16,** 97 (Feb. 18, 2015).

104. Wu, P.-Y., Phan, J. H. & Wang, M. D. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. *BMC Bioinformatics* **14,** S8 (Nov. 4, 2013).

105. Denoeud, F. *et al.* Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Research* **17** (2007).

106. Howald, C. *et al.* Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome research* **22,** 1698–710 (Sept. 2012).

107. Wright, J. C. *et al.* Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nature communications* **7,** 11778 (June 2016).

108. Feingold, E. *et al.* The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* **306** (2004).

109. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447** (2007).

110. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** (2012).

111. Stamatoyannopoulos, J. *et al.* An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology* **13** (2012).

112. GTEx Consortium, T. G. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* **45,** 580–5 (June 2013).

113. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348,** 648–660 (May 2015).

114. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (Apr. 2010).

115. Vogel, F. A Preliminary Estimate of the Number of Human Genes. *Nature* **201,** 847–847 (Feb. 1964).

116. Fields, C., Adams, M. D., White, O. & Venter, J. C. How many genes in the human genome? *Nature Genetics* **7,** 345–346 (July 1994).

117. Antequera, F. & Bird, A. Predicting the total number of human genes. *Nature Genetics* **8,** 114–114 (Oct. 1994).

118. Willyard, C. New human gene tally reignites debate. *Nature* **558,** 354–355 (June 19, 2018).

119. Pertea, M. *et al.* CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology* **19,** 208 (Nov. 28, 2018).

120. Jungreis, I. *et al.* Nearly all new protein-coding predictions in the CHESS database are not protein-coding. *bioRxiv,* 360602 (July 2, 2018).

121. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews. Genetics* **15,** 121–132 (Feb. 2014).

122. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews. Genetics* **14,** 618–630 (Sept. 2013).

123. Blencowe, B. J., Ahmad, S. & Lee, L. J. Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes & Development* **23,** 1379–1386 (June 15, 2009).

124. Su, Z. *et al.* A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology* **32,** 903–914 (Aug. 2014).

125. Kingston, R. E. Preparation of poly(A)+ RNA. *Current Protocols in Molecular Biology* **Chapter 4,** Unit4.5 (May 2001).

126. Zhulidov, P. A. *et al.* Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Research* **32,** e37 (2004).

127. Carninci, P. *et al.* Normalization and Subtraction of Cap-Trapper-Selected cDNAs to Prepare Full-Length cDNA Libraries for Rapid Discovery of New Genes. *Genome Research* **10,** 1617–1630 (Oct. 2000).

128. Bogdanov, E. A. *et al.* Normalizing cDNA libraries. *Current Protocols in Molecular Biology* **Chapter 5,** Unit 5.12.1–27 (Apr. 2010).

129. Kuo, R. I. *et al.* Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC genomics* **18,** 323 (2017).

130. Hoang, N. V., Furtado, A., Perlo, V., Botha, F. C. & Henry, R. J. The Impact of cDNA Normalization on Long-Read Sequencing of a Complex Transcriptome. *Frontiers in Genetics* **10** (July 23, 2019).

131. Yeku, O. & Frohman, M. A. Rapid amplification of cDNA ends (RACE). *Methods in molecular biology (Clifton, N.J.)* **703,** 107–22 (Jan. 2011).

132. Schaefer, B. C. Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends. *Analytical Biochemistry* **227,** 255–273 (May 20, 1995).

133. Bower, N. I. & Johnston, I. A. Targeted rapid amplification of cDNA ends (T-RACE)—an improved RACE reaction through degradation of non-target sequences. *Nucleic Acids Research* **38,** e194 (Nov. 2010).

134. Guigo, R. *et al.* Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 1140–1145 (Feb. 4, 2003).

135. Djebali, S. *et al.* Efficient targeted transcript discovery via array-based normalization of RACE libraries. *Nature Methods* **5** (2008).

136. Olivarius, S., Plessy, C. & Carninci, P. High-throughput verification of transcriptional starting sites by Deep-RACE. *BioTechniques* **46,** 130–2 (Feb. 2009).

137. Yehle, C. O. *et al.* A solution hybridization assay for ribosomal RNA from bacteria using biotinylated DNA probes and enzyme-labeled antibody to DNA:RNA. *Molecular and Cellular Probes* **1,** 177–193 (June 1987).

138. Bashiardes, S. *et al.* Direct genomic selection. *Nature Methods* **2,** 63–69 (Jan. 2005).

139. Teer, J. K. & Mullikin, J. C. Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics* **19,** R145–151 (R2 Oct. 15, 2010).

140. Mercer, T. R. *et al.* Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nature protocols* **9,** 989–1009 (May 2014).

141. Levin, J. Z. *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biology* **10,** R115 (Oct. 16, 2009).

142. Mercer, T. R. *et al.* Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nature biotechnology* **30,** 99–104 (Jan. 2012).

143. Clark, M. B. *et al.* Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nature Methods* **12,** 339–42 (Mar. 2015).

144. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology* **28,** 1045–1048 (Oct. 2010).

145. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (Feb. 2015).

146. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* A promoter-level mammalian expression atlas. *Nature* **507,** 462–470 (Mar. 27, 2014).

147. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.)* **309,** 1559–63 (Sept. 2005).

148. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–61 (Mar. 2014).

149. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (Sept. 6, 2012).

150. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489,** 83–90 (Sept. 2012).

151. Ernst, J. *et al.* Systematic analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (May 5, 2011).

152. Nellore, A. *et al.* Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome biology* **17,** 266 (2016).

153. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489** (2012).

154. Carlevaro-Fita, J. & Johnson, R. Global Positioning System: Understanding Long Noncoding RNAs through Subcellular Localization. *Molecular Cell* **73,** 869–883 (2019).

155. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* **45,** D896–D901 (D1 Jan. 2017).

156. Mele, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348,** 660–665 (2015).

157. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550,** 204–213 (2017).

158. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics* **50,** 956–967 (July 2018).

159. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550,** 239–243 (Oct. 2017).

160. Price, H. J. *et al.* Genome evolution in the genus Sorghum (Poaceae). *Annals of Botany* **95,** 219–227 (Jan. 2005).

161. Mirsky, A. E. & Ris, H. The desoxyribonucleic acid content of animal cells and its evolutionary significance. *The Journal of General Physiology* **34,** 451–462 (Mar. 20, 1951).

162. Thomas, C. A. The genetic organization of chromosomes. *Annual Review of Genetics* **5,** 237–256 (1971).

163. Ohno, S. So much "junk" DNA in our genome. *Brookhaven Symposia in Biology* **23,** 366–370 (1972).

164. Orgel, L. E. & Crick, F. H. Selfish DNA: the ultimate parasite. *Nature* **284,** 604–607 (Apr. 17, 1980).

165. Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284,** 601–603 (Apr. 17, 1980).

166. Lee, S.-I. & Kim, N.-S. Transposable Elements and Genome Size Variations in Plants. *Genomics & Informatics* **12,** 87–97 (Sept. 2014).

167. Elliott, T. A. & Gregory, T. R. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370** (Sept. 26, 2015).

168. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15,** 1034–1050 (Aug. 2005).

169. Thomas, J. W. *et al.* Comparative analyses of multispecies sequences from targeted genomic regions. *Nature* **424,** 788–793 (Aug. 14, 2003).

170. Schneiker, S. *et al.* Complete genome sequence of the myxobacterium *Sorangium cellulosum*. *Nature Biotechnology* **25,** 1281–1289 (Nov. 2007).

171. Konstantinidis, K. T. & Tiedje, J. M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proceedings of the National Academy of Sciences* **101,** 3160–3165 (Mar. 2, 2004).

172. International Wheat Genome Sequencing Consortium (IWGSC) *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science (New York, N.Y.)* **361** (2018).

173. Claverie, J. M. What if there are only 30,000 human genes? *Science (New York, N.Y.)* **291,** 1255–1257 (Feb. 16, 2001).

174. Hahn, M. W. & Wray, G. A. The g-value paradox. *Evolution & Development* **4,** 73–75 (2002).

175. International Wheat Genome Sequencing Consortium (IWGSC). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science (New York, N.Y.)* **345,** 1251788 (July 18, 2014).

176. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463,** 457–463 (Jan. 28, 2010).

177. Xing, Y. & Lee, C. Relating alternative splicing to proteome complexity and genome evolution. *Advances in Experimental Medicine and Biology* **623,** 36–49 (2007).

178. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science (New York, N.Y.)* **290,** 1151–1155 (Nov. 10, 2000).

179. Holland, P. W. H., Marlétaz, F., Maeso, I., Dunwell, T. L. & Paps, J. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **372** (2017).

180. Talavera, D., Vogel, C., Orozco, M., Teichmann, S. A. & de la Cruz, X. The (in)dependence of alternative splicing and gene duplication. *PLoS computational biology* **3,** e33 (Mar. 2, 2007).

181. Kopelman, N. M., Lancet, D. & Yanai, I. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nature Genetics* **37,** 588–589 (June 2005).

182. Su, Z., Wang, J., Yu, J., Huang, X. & Gu, X. Evolution of alternative splicing after gene duplication. *Genome Research* **16,** 182–189 (Feb. 2006).

183. Taft, R. J., Pheasant, M. & Mattick, J. S. The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* **29,** 288–299 (Mar. 2007).

184. Rands, C. M., Meader, S., Ponting, C. P. & Lunter, G. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genetics* **10** (July 24, 2014).

185. Comings, D. E. The structure and function of chromatin. *Advances in Human Genetics* **3,** 237–431 (1972).

186. Doolittle, W. F., Brunet, T. D., Linquist, S. & Gregory, T. R. Distinguishing between "Function" and "Effect" in Genome Biology. *Genome Biology and Evolution* **6,** 1234–1237 (May 9, 2014).

187. Graur, D. An Upper Limit on the Functional Fraction of the Human Genome. *Genome Biology and Evolution* **9,** 1880–1885 (July 1, 2017).

188. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nature Reviews. Genetics* **14,** 288–295 (2013).

189. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507,** 371–375 (Mar. 20, 2014).

190. De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS biology* **8,** e1000384 (May 11, 2010).

191. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).

192. Bose, D. A. *et al.* RNA Binding to CBP Stimulates Histone Acetylation and Transcription. *Cell* **168,** 135–149.e22 (Jan. 12, 2017).

193. Xie, D. *et al.* Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Research* **20,** 804–815 (June 2010).

194. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (New York, N.Y.)* **328,** 1036–1040 (May 21, 2010).

195. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics* **42,** 631–634 (July 2010).

196. Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nature Genetics* **42,** 806–810 (Sept. 2010).

197. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).

198. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Research* **22,** 1748–1759 (Sept. 2012).

199. Ward, L. D. & Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science (New York, N.Y.)* **337,** 1675–1678 (Sept. 28, 2012).

200. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science (New York, N.Y.)* **342,** 1235587 (Oct. 4, 2013).

201. Bejerano, G. *et al.* Ultraconserved Elements in the Human Genome. *Science* **304,** 1321–1325 (May 28, 2004).

202. Woolfe, A. *et al.* Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLOS Biology* **3,** e7 (Nov. 11, 2004).

203. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444,** 499–502 (Nov. 2006).

204. Ahituv, N. *et al.* Deletion of Ultraconserved Elements Yields Viable Mice. *PLOS Biology* **5,** e234 (Sept. 4, 2007).

205. Dickel, D. E. *et al.* Ultraconserved Enhancers Are Required for Normal Development. *Cell* **172,** 491–499.e15 (2018).

206. McCole, R. B., Fonseka, C. Y., Koren, A. & Wu, C.-T. Abnormal dosage of ultraconserved elements is highly disfavored in healthy cells but not cancer cells. *PLoS genetics* **10,** e1004646 (Oct. 2014).

207. Braconi, C. *et al.* Expression and functional role of a transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 786–791 (Jan. 11, 2011).

208. Smith, M. A., Gesell, T., Stadler, P. F. & Mattick, J. S. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res* **41,** 8220–8236 (2013).

209. Eddy, S. R. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annual Review of Biophysics* **43,** 433–456 (2014).

210. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nature Reviews. Genetics* **9,** 397–405 (May 2008).

211. Bourque, G. Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr Opin Genet Dev* **19,** 607–612 (2009).

212. Mcclintock, B. Controlling elements and the gene. *Cold Spring Harbor Symposia on Quantitative Biology* **21,** 197–216 (1956).

213. Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Research* **24,** 1963–1976 (Dec. 2014).

214. Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G. The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science (New York, N.Y.)* **300,** 1288–1291 (May 23, 2003).

215. Sela, N. *et al.* Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biol* **8,** R127 (2007).

216. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41,** 563–571 (2009).

217. Huda, A., Bowen, N. J., Conley, A. B. & Jordan, I. K. Epigenetic regulation of transposable element derived human gene promoters. *Gene* **475,** 39–48 (2011).

218. Pi, W. *et al.* The LTR enhancer of ERV-9 human endogenous retrovirus is active in oocytes and progenitor cells in transgenic zebrafish and humans. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 805–810 (Jan. 20, 2004).

219. Ling, J. *et al.* The solitary long terminal repeats of ERV-9 endogenous retrovirus are conserved during primate evolution and possess enhancer activities in embryonic and hematopoietic cells. *Journal of Virology* **76,** 2410–2423 (Mar. 2002).

220. Johnson, R. & Guigó, R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA (New York, N.Y.)* **20,** 959–76 (July 2014).

221. Carlevaro-Fita, J., Das, M., Polidori, T., Navarro, C. & Johnson, R. Ancient exapted transposable elements promote nuclear enrichment of long noncoding RNAs. *bioRxiv,* 189753 (Oct. 2017).

222. Roberts, J. T., Cardin, S. E. & Borchert, G. M. Burgeoning evidence indicates that microRNAs were initially formed from transposable element sequences. *Mobile Genetic Elements* **4,** e29255 (2014).

223. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).

224. De Souza, F. S., Franchini, L. F. & Rubinstein, M. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? *Mol Biol Evol* **30,** 1239–1251 (2013).

225. Jain, H. K. Incidental DNA. *Nature* **288,** 647–648 (Dec. 25, 1980).

226. Ecker, J. R. *et al.* Genomics: ENCODE explained. *Nature* **489,** 52–55 (Sept. 6, 2012).

227. Brenner, n. Refuge of spandrels. *Current biology: CB* **8,** R669 (Sept. 24, 1998).

228. Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N. J. Intergenic transcription and transinduction of the human beta-globin locus. *Genes & Development* **11,** 2494–2509 (Oct. 1, 1997).

229. Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309,** 1564–1566 (2005).

230. Kapranov, P. *et al.* Large-Scale Transcriptional Activity in Chromosomes 21 and 22. *Science* **296,** 916–919 (May 2002).

231. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science (New York, N.Y.)* **308,** 1149–1154 (May 20, 2005).

232. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science (New York, N.Y.)* **316,** 1484–1488 (June 8, 2007).

233. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436,** 876–880 (Aug. 11, 2005).

234. He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. *Science (New York, N.Y.)* **322,** 1855–1857 (Dec. 19, 2008).

235. Palazzo, A. F. & Lee, E. S. Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics* **6,** 2 (2015).

236. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature structural & molecular biology* **14,** 103–5 (Feb. 2007).

237. Cheung, V. *et al.* Chromatin- and transcription-related factors repress transcription from within coding regions throughout the Saccharomyces cerevisiae genome. *PLoS biology* **6,** e277 (Nov. 11, 2008).

238. White, M. A., Myers, C. A., Corbo, J. C. & Cohen, B. A. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 11952–11957 (July 16, 2013).

239. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154,** 26–46 (July 2013).

240. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long noncoding RNAs: past, present, and future. *Genetics* **193,** 651–69 (Mar. 2013).

241. Wang, H.-L. V. & Chekanova, J. A. Long Noncoding RNAs in Plants. *Advances in Experimental Medicine and Biology* **1008,** 133–154 (2017).

242. Jarroux, J., Morillon, A. & Pinskaya, M. History, Discovery, and Classification of lncRNAs. *Advances in Experimental Medicine and Biology* **1008,** 1–46 (2017).

243. Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. *Molecular and cellular biology* **10,** 28–36 (Jan. 1990).

244. Gabory, A. *et al.* H19 acts as a trans regulator of the imprinted gene network controlling growth in mice. *Development (Cambridge, England)* **136,** 3413–3421 (Oct. 2009).

245. Brown, C. J. *et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349,** 38–44 (1991).

246. Brown, C. J. *et al.* The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71,** 527–542 (Oct. 30, 1992).

247. Brockdorff, N. *et al.* The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71,** 515–26 (Oct. 1992).

248. Wutz, A. & Jaenisch, R. A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. *Molecular Cell* **5,** 695–705 (Apr. 2000).

249. Nesterova, T. B. *et al.* Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Research* **11,** 833–849 (May 2001).

250. Chureau, C. *et al.* Comparative Sequence Analysis of the X-Inactivation Center Region in Mouse, Human, and Bovine. *Genome Research* **12,** 894–908 (June 2002).

251. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420,** 563–573 (Dec. 5, 2002).

252. Willingham, A. T. *et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science (New York, N.Y.)* **309,** 1570–1573 (Sept. 2, 2005).

253. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129,** 1311–1323 (2007).

254. Ji, P. *et al.* MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22,** 8031–8041 (2003).

255. Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415,** 810–813 (Feb. 14, 2002).

256. Sotomaru, Y. *et al.* Unregulated expression of the imprinted genes H19 and Igf2r in mouse uniparental fetuses. *The Journal of Biological Chemistry* **277,** 12474–12478 (Apr. 5, 2002).

257. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (Aug. 2, 2007).

258. Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142,** 409–419 (Aug. 2010).

259. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28,** 511–515 (May 2010).

260. Cabili, M. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes and Development* **25,** 1915–1927 (Sept. 2011).

261. 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 Genetics* **9** (ed Rinn, J. L.) e1003569 (June 2013).

262. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics* **47,** 199–208 (Jan. 2015).

263. Bussotti, G. *et al.* Improved definition of the mouse transcriptome via targeted RNA sequencing. *Genome Research* **26,** 705–716 (May 2016).

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

265. Fang, S. *et al.* NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Research* (Nov. 2017).

266. Hon, C.-C. *et al.* An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* (Mar. 2017).

267. Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nature Reviews Genetics* **17,** 601–614 (Oct. 2016).

268. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research* **22,** 1775–1789 (Sept. 2012).

269. Morillon, A. & Gautheret, D. Bridging the gap between reference and real transcriptomes. *Genome Biology* **20,** 112 (June 3, 2019).

270. Carrieri, C. *et al.* Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* **491,** 454–7 (Nov. 2012).

271. Pauli, A. *et al.* Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Research* **22,** 577–591 (Mar. 2012).

272. Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147,** 1537–1550 (2011).

273. Van Bakel, H., Nislow, C., Blencowe, B. J. & Hughes, T. R. Most "dark matter" transcripts are associated with known genes. *PLoS biology* **8,** e1000371 (May 18, 2010).

274. Melé, M. *et al.* Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome research* (Dec. 2016).

275. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* **22,** 1616–1625 (2012).

276. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505,** 635–40 (Jan. 2014).

277. Washietl, S., Kellis, M. & Garber, M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* (2014).

278. Hezroni, H. *et al.* Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Reports* **11,** 1110–1122 (May 2015).

279. Torarinsson, E., Sawera, M., Havgaard, J. H., Fredholm, M. & Gorodkin, J. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Research* **16,** 885–889 (July 2006).

280. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology* **13,** R107 (Nov. 26, 2012).

281. He, S., Gu, W., Li, Y. & Zhu, H. ANRIL/CDKN2B-AS shows two-stage clade-specific evolution and becomes conserved after transposon insertions in simians. *BMC Evol Biol* **13,** 247 (2013).

282. Elisaphenko, E. A. *et al.* A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. *PLoS One* **3,** e2521 (2008).

283. Ng, P., Wei, C.-L. & Ruan, Y. Paired-end diTagging for transcriptome and genome analysis. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **Chapter 21,** Unit 21.12 (July 2007).

284. Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome Research* **10,** 1001–1010 (July 2000).

285. Alam, T. *et al.* Promoter Analysis Reveals Globally Differential Regulation of Human Long Non-Coding RNA and Protein-Coding Genes. *PLoS ONE* **9** (ed Mantovani, R.) e109443 (Oct. 2014).

286. Matsumoto, A. *et al.* mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* **541,** 228–232 (2017).

287. Anderson, D. M. *et al.* A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* **160,** 595–606 (Feb. 2015).

288. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).

289. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet* **15,** 205–213 (2014).

290. Choi, S.-W., Kim, H.-W. & Nam, J.-W. The small peptide world in long noncoding RNAs. *Briefings in Bioinformatics* (June 29, 2018).

291. Quek, X. C. *et al.* lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic acids research* **43,** D168–73 (Database issue Jan. 2015).

292. Kopp, F. & Mendell, J. T. Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell* **172,** 393–407 (Jan. 2018).

293. Latos, P. A. *et al.* Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science (New York, N.Y.)* **338,** 1469–1472 (Dec. 14, 2012).

294. Anderson, K. M. *et al.* Transcription of the noncoding RNA upperhand controls Hand2 expression and heart development. *Nature* **539,** 433–436 (2016).

295. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539,** 452–455 (Oct. 2016).

296. Groff, A. F. *et al.* In Vivo Characterization of Linc-p21 Reveals Functional cis-Regulatory DNA Elements. *Cell Rep* **16,** 2178–2186 (2016).

297. Tsai, M. C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329,** 689–693 (2010).

298. Hu, W., Yuan, B., Flygare, J. & Lodish, H. F. Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. *Genes & Development* **25,** 2573–2578 (Dec. 15, 2011).

299. Atianand, M. K. *et al.* A Long Noncoding RNA lincRNA-EPS Acts as a Transcriptional Brake to Restrain Inflammation. *Cell* **165,** 1672–1685 (June 16, 2016).

300. Sunwoo, H. *et al.* MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Research* **19,** 347–359 (Mar. 2009).

301. Clemson, C. M. *et al.* An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Molecular Cell* **33,** 717–726 (Mar. 27, 2009).

302. Tripathi, V. *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular Cell* **39,** 925–938 (Sept. 24, 2010).

303. Bernard, D. *et al.* A long nuclear-retained noncoding RNA regulates synaptogenesis by modulating gene expression. *The EMBO journal* **29,** 3082–3093 (Sept. 15, 2010).

304. Hutchinson, J. N. *et al.* A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC genomics* **8,** 39 (Feb. 1, 2007).

305. West, J. A. *et al.* The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Molecular Cell* **55,** 791–802 (Sept. 4, 2014).

306. Tichon, A. *et al.* A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. *Nature Communications* **7,** 12209 (2016).

307. Lee, S. *et al.* Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* **164,** 69–80 (Jan. 14, 2016).

308. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495,** 333–338 (Mar. 21, 2013).

309. Hansen, T. B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* **495,** 384–388 (Mar. 21, 2013).

310. Seiler, J. *et al.* The lncRNA VELUCT strongly regulates viability of lung cancer cells despite its extremely low abundance. *Nucleic Acids Research* **45,** 5458–5469 (May 19, 2017).

311. Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F. & Mattick, J. S. Specific expression of long noncoding RNAs in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 716–721 (Jan. 15, 2008).

312. Dinger, M. E., Amaral, P. P., Mercer, T. R. & Mattick, J. S. Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. *Briefings in Functional Genomics & Proteomics* **8,** 407–423 (Nov. 2009).

313. Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571,** 510–514 (2019).

314. Field, A. R. *et al.* Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes. *Stem Cell Reports* **12,** 245–257 (Jan. 10, 2019).

315. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)* **339,** 819–823 (Feb. 15, 2013).

316. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science (New York, N.Y.)* **339,** 823–826 (Feb. 15, 2013).

317. Zhu, S. *et al.* Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR–Cas9 library. *Nature Biotechnology* **34,** 1279–1286 (Oct. 2016).

318. Aparicio-Prat, E. *et al.* DECKO: Single-oligo, dual-CRISPR deletion of genomic elements including long non-coding RNAs. *BMC Genomics* **16,** 846 (2015).

319. Dominguez, A. A., Lim, W. A. & Qi, L. S. Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation. *Nature reviews. Molecular cell biology* **17,** 5–15 (Jan. 2016).

320. Mohr, S. E., Smith, J. A., Shamu, C. E., Neumüller, R. A. & Perrimon, N. RNAi screening comes of age: improved techniques and complementary approaches. *Nature Reviews. Molecular Cell Biology* **15,** 591–600 (Sept. 2014).

321. Lagarde, J. *et al.* Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq). *Nature Communications* **7** (2016).

322. Lagarde, J. *et al.* High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nature Genetics* **49,** 1731–1740 (Nov. 2017).

323. Hirabayashi, S. *et al.* NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nature Genetics* **51,** 1369–1379 (Sept. 2019).

324. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research* **41,** e74–e74 (Apr. 2013).

325. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27,** i275–i282 (July 2011).

326. Brümmer, A., Dreos, R., Marques, A. C. & Bergmann, S. LincRNA sequences are biased to counteract their translation. *bioRxiv*, 737890 (Aug. 16, 2019).

327. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137,** 1194–1211 (June 26, 2009).

328. Cao, R. & Zhang, Y. The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. *Current Opinion in Genetics & Development* **14,** 155–164 (Apr. 2004).

329. Ku, M. *et al.* Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS genetics* **4,** e1000242 (Oct. 2008).

330. Schmitges, F. W. *et al.* Histone methylation by PRC2 is inhibited by active chromatin marks. *Molecular Cell* **42,** 330–341 (May 6, 2011).

331. Deveson, I. W. *et al.* Universal Alternative Splicing of Noncoding Exons. *Cell Systems* **6,** 245–255.e5 (Feb. 2018).

332. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nature Reviews Genetics* (2018).

333. Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L.-L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biology* **12,** R16 (2011).

334. Pedraza, J. M. & Paulsson, J. Effects of molecular memory and bursting on fluctuations in gene expression. *Science (New York, N.Y.)* **319,** 339–343 (Jan. 18, 2008).

335. Dar, R. D. *et al.* Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 17454–17459 (Oct. 23, 2012).

336. Bahar Halpern, K. *et al.* Bursty gene expression in the intact mammalian liver. *Molecular Cell* **58,** 147–156 (Apr. 2, 2015).

337. Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature Biotechnology* **29,** 436–442 (May 2011).

338. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6,** 377–382 (May 2009).

339. Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods* **11,** 22–24 (Jan. 2014).

340. Cabili, M. N. *et al.* Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome biology* **16,** 20 (Jan. 2015).

341. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13,** 599–604 (2018).

342. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Research* **12,** 996–1006 (June 1, 2002).

343. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20,** 110–121 (2010).

344. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England)* (May 16, 2018).

Appendix

I

Supplementary Methods

- Table 2, Figure 7, Figure 10, Table 3: statistics derived from the GENCODE Human and/or Mouse reference annotations[36], versions 21 and M4, respectively (release: 2014)[1].
- Table 2, Figure 7: only protein-coding transcripts with confidently annotated ends (*i.e.*, not tagged 'mRNA_end_NF' or 'mRNA_end_NF') are considered.
- Figure 10: statistics based on gene biotypes only, which are simplified as follows:
  - **lncRNA**: "antisense", "non_coding", "bidirectional_promoter_lncrna", "macro_lncRNA", "lincRNA", "processed_transcript", "sense_intronic", "sense_overlapping"
  - **protein-coding**: "protein_coding", "IG_*", "TR_*"

[1] https://www.gencodegenes.org

II

Supplementary Figures

The following gallery of figures represent UCSC genome browser[2] plots[342] of a few lncRNA loci of interest, as annotated by GENCODE in human (hg38 assembly, GENCODE version 29) and mouse (mm10 assembly, GENCODE version M20)[36]. Exons are represented as solid blue boxes, with CDSs (when present) thicker. Introns are depicted as thin arrowed lines, with the direction of the arrows indicating the genomic strand of transcription. The bottom green track represent sequence conservation scores across vertebrates, as calculated by the PhastCons software based on whole-genome alignments[343]. The PhastCons score value can be interpreted as the probability that a given sequence is conserved according to the underlying phylogenetic model, from zero (not conserved, plot baseline) to one (highly conserved, indicated with a horizontal line).

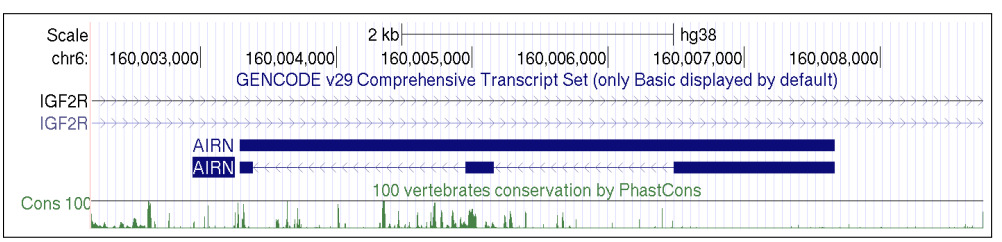


Figure S1: The *Air* (a.k.a. *AIRN*) human locus.

[2] http://genome.ucsc.edu/

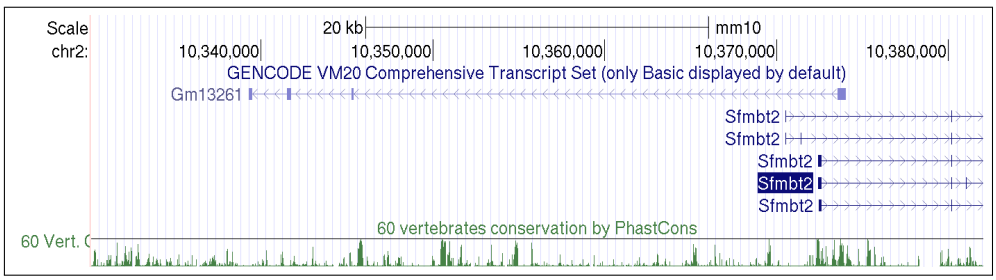


Figure S2: The *Blustr* (a.k.a. *Gm13261*) mouse locus. Note that the GENCODE transcript model slightly differs from the one reported in[295].

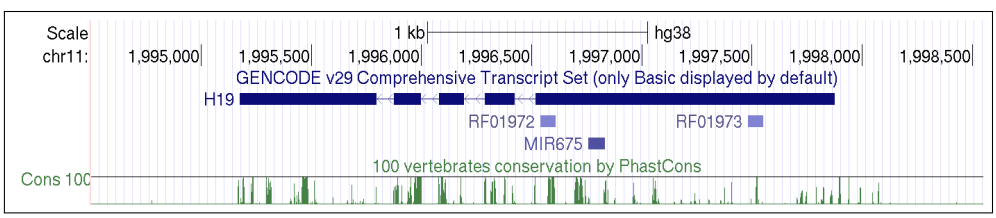


Figure S3: The *H19* human locus.

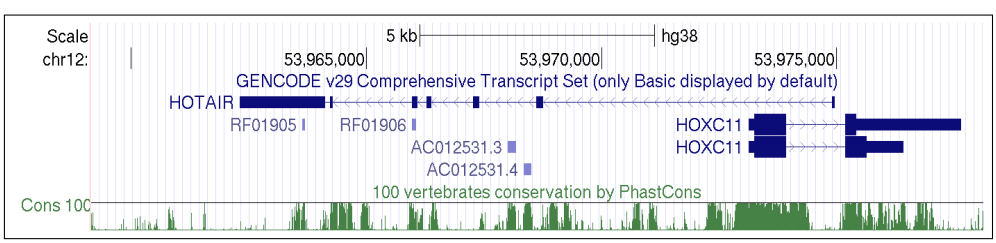


Figure S4: The *HOTAIR* human locus.

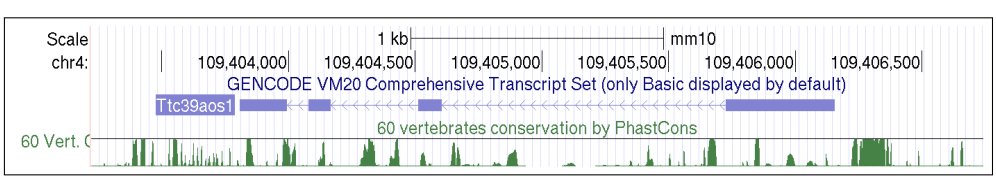


Figure S5: The *lincRNA-EPS* (a.k.a. *Ttc39aos1*) mouse locus.

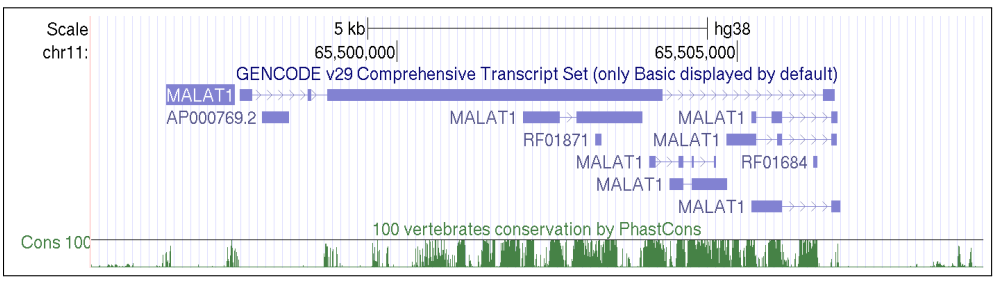


Figure S6: The *MALAT1* human locus.

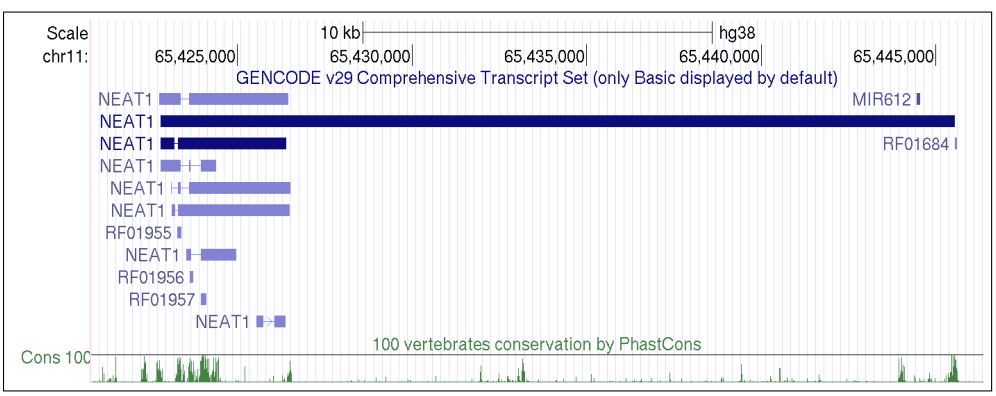


Figure S7: The *NEAT1* human locus.

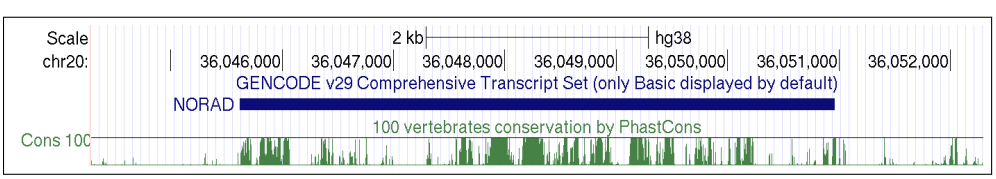


Figure S8: The *NORAD* human locus.

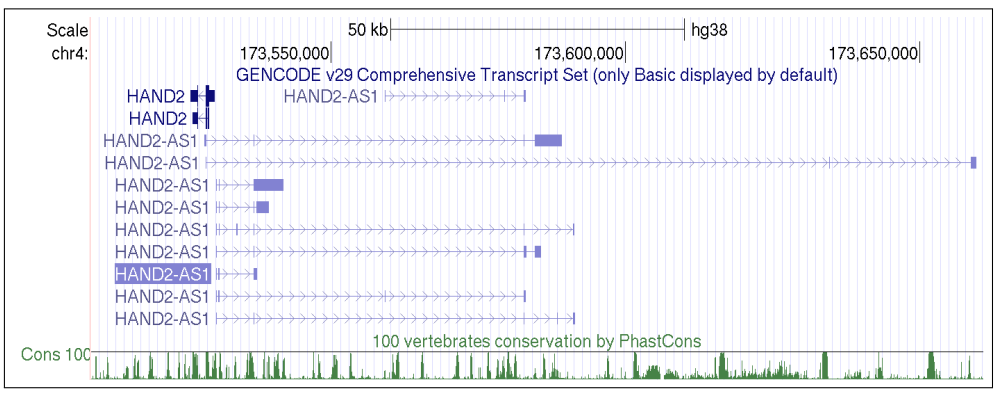


Figure S9: The *Upperhand* (a.k.a. *HANDS2-AS1*) human locus.

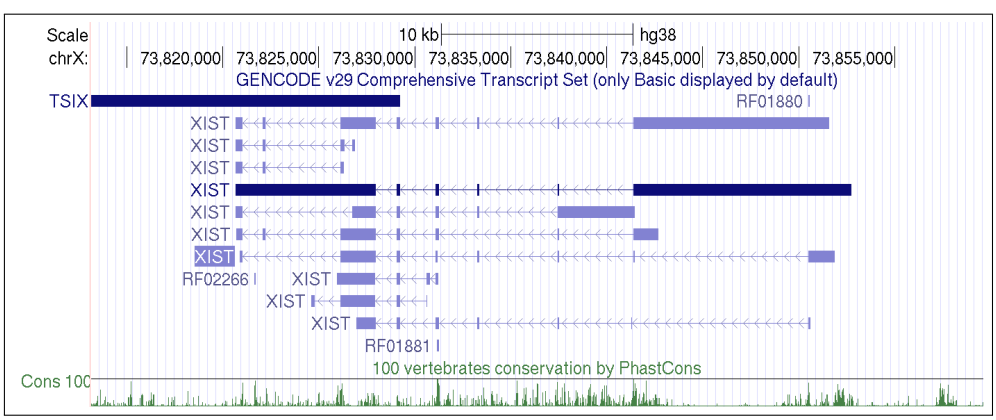


Figure S10: The *Xist* human locus.

III

Additional relevant publication

Lagarde J, Johnson R. *Capturing a long look at our genetic library*. ***Cell Systems*** 2018 Feb; 6(2):153-155.

URL: https://doi.org/10.1016/j.cels.2018.02.003

This article is an invited commentary on the Deveson *et al.* study entitled "*Universal Alternative Splicing of Noncoding Exons*", published in the same journal issue[331]. It was not peer-reviewed.

Abstract:

Long-read sequencing, coupled to cDNA capture, provides an unrivaled view of the transcriptome of chromosome 21, revealing surprises about the splicing of long noncoding RNAs.

Cell Systems
Previews

Capturing a Long Look at Our Genetic Library

Julien Lagarde[1,2] and Rory Johnson[3,4,*]
[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
[2]Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain
[3]Department of Medical Oncology, Inselspital, University Hospital and University of Bern, 3010 Bern, Switzerland
[4]Department of Biomedical Research (DBMR), University of Bern, 3008 Bern, Switzerland
*Correspondence: rory.johnson@dbmr.unibe.ch
https://doi.org/10.1016/j.cels.2018.02.003

Long-read sequencing, coupled to cDNA capture, provides an unrivaled view of the transcriptome of chromosome 21, revealing surprises about the splicing of long noncoding RNAs.

In his story "The Library of Babel," Jorge Luis Borges imagines a library of books with every conceivable permutation of letters. Every story told, or to be told, is found there. Similarly, our genome contains every gene and transcript, coding and noncoding, to be expressed during the human lifetime. But our catalog of this genetic library remains unsatisfactory—our books miss entire chapters, and many are completely unaccounted for.

In this issue of *Cell Systems*, Mercer, Mattick, and colleagues mark an important step in overcoming this by reporting a deep survey of the transcriptome of long noncoding RNAs (lncRNAs) and mRNAs on human chromosome 21 (Chr21) (Deveson et al., 2018). This is made possible by coupling two powerful techniques: cDNA capture and sequencing on a Pacific Biosciences (PacBio) instrument. This work joins several recent studies harnessing the unrivaled power of third-generation single-molecule sequencing to accurately survey the transcriptome (Sharon et al., 2013; Lagarde et al., 2017). This technology frees us from our dependence on short-read technology and opens fundamental questions about lncRNA biology.

Until now, researchers have relied heavily on assembly of short reads from Illumina-based RNA sequencing (RNA-seq) experiments to map lncRNAs. Programs such as Cufflinks have enabled labs to create catalogs in their favorite cell type or organism (Trapnell et al., 2010). But accurately assembling the exons of long transcripts from much shorter reads is a daunting algorithmic challenge. Sensitivity is low (entire genes are missed), false positives are frequent (i.e., nonexistent transcripts are assembled), and almost all transcripts fall short of the true 5′ and 3′ ends (Lagarde et al., 2017; Steijger et al., 2013). This is particularly acute for lncRNAs due to their low expression and sparse read coverage. Nevertheless, the canon of lncRNA knowledge rests largely upon these catalogs.

Deveson et al. and others have realized that long-read technology can overcome these issues. It can confidently report exon connectivity, while 3′ ends are identified by encoded polyA tails, and 5′ are also frequently reached (Lagarde et al., 2017; Sharon et al., 2013). However, the low sequencing depth of PacBio (~50,000 reads per lane versus ~300 million for Illumina), coupled to lncRNAs' low expression, introduces a new challenge (Sharon et al., 2013).

To address this challenge, Deveson et al. coupled long-read sequencing to cDNA capture. The latter method, pioneered by the authors themselves, focuses sequencing firepower onto known or suspected RNA-producing loci – particularly valuable for low-expressed lncRNAs (Mercer et al., 2014). Using oligonucleotide capture, cDNA libraries are first enriched for regions of interest—here, the entire human Chr21, representing 1.5% of the genome (Deveson et al., 2018). Captured cDNA, of which >70% originates from Chr21, is then sequenced by both long- and short-read technologies. In this way, the length of PacBio is harnessed to map exon connectivity, while deeper short reads can accurately quantify expression and splicing. A similar approach was recently employed by the GENCODE consortium to improve annotation of known lncRNAs (Lagarde et al., 2017).

By coupling third-generation sequencing to cDNA capture, the authors produce one of the deepest ever transcriptional maps of a human chromosome. They report a dataset of 387,029 reads from K562 cells and testis, revealing altogether 1,589 lncRNA transcript models. Approximately half of identified exons are novel. Simulation experiments suggest that, at least in the tissue panel sampled, the number of cataloged exons is approaching saturation.

These confident transcript models enable us to revisit old questions about lncRNAs, as well as formulate completely new ones. For example, we can ask to what extent lncRNA genes or products differ from protein-coding genes—the answer seems to be surprisingly little (Lagarde et al., 2017). In terms of mature length, or promoter chromatin, previously observed differences don't hold up to scrutiny afforded by full-length structures. In addition are myriad instances where incorrect gene annotations are extended to their full length or when separate fragmentary annotations are united to form a single, correct gene model (Lagarde et al., 2017; Deveson et al., 2018).

Deveson et al. extend these observations to splicing. It is known that lncRNA splicing exhibits some distinct properties compared to coding genes: it is less efficient (Tilgner et al., 2012), and their splice sites are less conserved (Nitsche et al., 2015). But now, Deveson et al. have identified another potentially more interesting hallmark of lncRNA: high rate of alternative splicing. Comparing the "percent spliced in" (PSI), a measure of frequency with

CellPress

Cell Systems
Previews

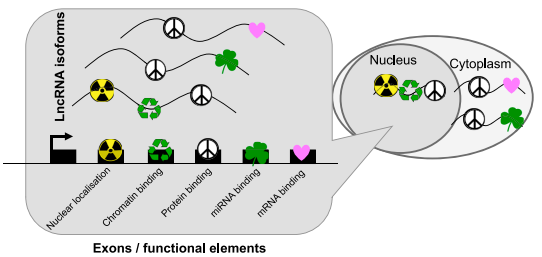


Figure 1. Generating Long Noncoding RNA Diversity
Deveson et al. propose that extensive alternative splicing may generate long noncoding RNAs (lncRNAs) with diverse functions through the differential inclusion of modular elements mediating nuclear localization or chromatin-, protein-, miRNA-, or mRNA-binding.

which exons are included in transcripts, they show that lncRNA exons tend to be alternative far more than those of coding genes. In other words, a given lncRNA transcript tends to choose just a subset of available exons. The authors name this "universal alternative splicing."

Universal alternative splicing, if validated, could have profound implications for our understanding of lncRNA functions. LncRNAs are thought to be modular: composed of combinations of functional elements and analogous to protein domains (Guttman and Rinn, 2012). There is growing interest in identifying such elements, but so far, relatively few are known (Marín-Béjar et al., 2017). The differential inclusion of such elements through exon splicing could be a mechanism of producing lncRNAs with diverse functions (Figure 1). Indeed, without the constraint to maintain an open reading frame, lncRNAs could exploit this mechanism more freely than mRNAs.

Of course, as is often the case for lncRNAs, one can interpret most observations equally as evidence for function or the lack of function. Does widespread alternative splicing reflect modularity, or simply relaxed constraint? This joins other features of lncRNAs—such as tissue-specific expression, nuclear localization, and lower evolutionary conservation—as features that can be interpreted in polar opposite ways. While Deveson et al. articulate an attractive argument for a "functionalist" interpretation, we would argue that one should adopt non-functionality as a null hypothesis to be falsified. The authors also showed that low PSI is a property of other transcribed noncoding sequences, namely untranslated regions, suggesting that splicing constraint is relaxed when an open reading frame is not present. Nevertheless, natural selection tends to make good use of available biological variation. It is entirely likely that, once better maps of lncRNA elements become available, compelling examples of alternatively spliced isoforms with demonstrably different functions will be uncovered.

On a more practical level, if lncRNA splicing really is as complex as suggested, it will have quite serious ramifications for how we annotate these genes. Is there value to users in individually annotating a vast assembly of splice variants? Or will we have to find more economical and abstract ways of annotating the splicing structure of a lncRNA?

Either way, long-read sequencing will likely lead to rapid improvements of lncRNA (and even protein-coding) gene annotations in coming years. Researchers will no doubt have to revisit more long-held assumptions about lncRNAs and re-quantify old short-read RNA-seq datasets using these new annotations. Differential gene expression studies can be carried out to find new targets in old data.

PacBio technology still suffers from several drawbacks that limit its usefulness in mapping lncRNAs. These include its high cost, low throughput, and cDNA read lengths that still do not exceed ~3 kb (Lagarde et al., 2017; Sharon et al., 2013). On the horizon is a technology that promises to resolve all these issues: direct RNA-seq by nanopore. The MinION from Oxford Nanopore Technologies offers direct RNA-seq with essentially no length limit (Garalde et al., 2018). Now, the race is on to apply this approach to lncRNAs.

By containing every possible book, Borges library held vastly more nonsense books than meaningful ones, including every possible error-containing version of any real book. The challenge for us now is to understand whether this applies to splicing of lncRNAs. Are they Borgian nonsense produced in an absence of selective pressure? Or a powerful mechanism for generating functional diversity through combinatorics?

ACKNOWLEDGMENTS

R.J. is supported by the Swiss National Science foundation through the National Centres for Competence in Research (NCCR) "RNA & Disease" and by the Medical Faculty of the University of Bern. J.L. is supported by the National Human Genome Research Institute of the US National Institutes of Health (grant U41HG007234).

REFERENCES

Deveson, I.W., Brunc, M.E., Blackburn, J., Tseng, E., Hon, T., Clark, T.A., Clark, M.B., Crawford, J., Dinger, M.E., Nielsen, L.K., et al. (2018). Universal alternative splicing of noncoding exons. Cell Syst. *6*, this issue, 245–255.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods. Published online January 15, 2018. https://doi.org/10.1038/nmeth.4577.

Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large non-coding RNAs. Nature *482*, 339–346.

Lagarde, J., Uszczynska-Ratajczak, B., Carbonell, S., Pérez-Lluch, S., Abad, A., Davis, C., Gingeras, T.R., Frankish, A., Harrow, J., Guigo, R., and Johnson, R. (2017). High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nat. Genet. *49*, 1731–1740.

Marín-Béjar, O., Mas, A.M., González, J., Martinez, D., Athie, A., Morales, X., Galduroz, M., Raimondi, I., Grossi, E., Guo, S., et al. (2017). The human lncRNA LINC-PINT inhibits tumor cell invasion through a highly conserved sequence element. Genome Biol. *18*, 202.

Mercer, T.R., Clark, M.B., Crawford, J., Brunck, M.E., Gerhardt, D.J., Taft, R.J., Nielsen, L.K.,

Cell Systems
Previews

Dinger, M.E., and Mattick, J.S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. Nat. Protoc. *9*, 989–1009.

Nitsche, A., Rose, D., Fasold, M., Reiche, K., and Stadler, P.F. (2015). Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved. RNA *21*, 801–812.

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

Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Hubbard, T.J., Guigó, R., Harrow, J., and Bertone, P.; RGASP Consortium (2013). Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods *10*, 1177–1184.

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

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

IV

Relevant software written by the author

- **`LR-Seq`**
  - **Description**: Full Snakemake[344] CLS bioinformatics analysis workflow.
  - **URL**: https://github.com/julienlag/LR-Seq
- **`anchorTranscriptsEnds`**
  - **Description**: Prepare a GTF file of mapped transcriptome reads for anchored transcript merging
  - **URL**: https://github.com/julienlag/anchorTranscriptsEnds
- **`buildLoci`**
  - **Description**: Automatically build gene loci out of sets of overlapping transcripts
  - **URL**: https://github.com/julienlag/buildLoci
- **`matchDistribution`**
  - **Description**: Given distinct "subject" (S) and a "target" (T) distributions, this script attempts to mimic T's density by pseudo-randomly sampling from S's population.
  - **URL**: https://github.com/julienlag/matchDistribution
- **`samToPolyA`**
  - **Description**: Call polyA sites based on genome alignments in SAM format
  - **URL**: https://github.com/julienlag/samToPolyA
- **`tmerge`**
  - **Description**: Merge transcriptome read-to-genome alignments into non-redundant transcript models
  - **URL**: https://github.com/julienlag/tmerge

V

Image credits

- Cover design by the author. Photograph taken inside Gaudí's Basílica de la Sagrada Família, Barcelona, by the author[3].
- Figure 2 contains graphic elements from Wikimedia[4], licensed under the Creative Commons Attribution-Share Alike 4.0 International license[5].
- Figure 5 contains graphic elements from Wikimedia[6], licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license[7].

[3] https://flic.kr/p/5BZamS
[4] https://commons.wikimedia.org/wiki/File:Metabolic_Metro_Map.svg
[5] https://creativecommons.org/licenses/by-sa/4.0/deed.en
[6] https://commons.wikimedia.org/wiki/File:Top7.png
[7] https://creativecommons.org/licenses/by-sa/3.0/deed.en

VI

Miscellaneous

The present work was carried out using only Free and Open Source Software, under the Ubuntu Linux operating system[8]. The LaTeX source code of this Thesis, together with source images, relevant scripts and the PDF version of the document are available online[9] under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)[10]. All original figures were produced with Inkscape[11] and R[12] / ggplot2[13]. LaTeX code was edited using Texmaker[14], with Zotero[15] as the reference manager.

Contact: julienlag@gmail.com

[8] https://www.ubuntu.com/
[9] https://public-docs.crg.es/rguigo/jlagarde/PhD or QR code above
[10] https://creativecommons.org/licenses/by-nc-sa/4.0/
[11] https://inkscape.org/
[12] https://www.r-project.org/
[13] https://ggplot2.tidyverse.org/
[14] https://www.xm1math.net/texmaker/
[15] https://www.zotero.org/