nature structural & molecular biology

Integrative classification of human coding and noncoding genes through RNA metabolism profiles

Neelanjan Mukherjee[1], Lorenzo Calviello[1,2], Antje Hirsekorn[1], Stefano de Pretis[3], Mattia Pelizzola[3] & Uwe Ohler[1,2,4]

Pervasive transcription of the human genome results in a heterogeneous mix of coding RNAs and long noncoding RNAs (lncRNAs). Only a small fraction of lncRNAs have demonstrated regulatory functions, thus making functional lncRNAs difficult to distinguish from nonfunctional transcriptional byproducts. This difficulty has resulted in numerous competing human lncRNA classifications that are complicated by a steady increase in the number of annotated lncRNAs. To address these challenges, we quantitatively examined transcription, splicing, degradation, localization and translation for coding and noncoding human genes. We observed that annotated lncRNAs had lower synthesis and higher degradation rates than mRNAs and discovered mechanistic differences explaining slower lncRNA splicing. We grouped genes into classes with similar RNA metabolism profiles, containing both mRNAs and lncRNAs to varying extents. These classes exhibited distinct RNA metabolism, different evolutionary patterns and differential sensitivity to cellular RNA-regulatory pathways. Our classification provides an alternative to genomic context-driven annotations of lncRNAs.

Pervasive transcription of the human genome has spurred efforts to identify and functionally characterize lncRNAs[1]. lncRNAs are noncoding transcripts longer than 200 nt; this length cutoff is an *ad hoc* convention to distinguish lncRNAs from well-characterized small noncoding RNAs, such as microRNAs (miRNAs), small nuclear and nucleolar RNAs (snRNAs and snoRNAs, respectively) and tRNAs[2]. Contemporary annotations include many tens of thousands of this expanding heterogeneous group of RNAs[3].

Like mRNAs, lncRNAs are transcribed by RNA polymerase II (pol II), 5′-capped and frequently spliced and polyadenylated[4]. The defining characteristic of lncRNA, the absence of a translated open reading frame (ORF), has received scrutiny, given lncRNAs' extensive polyribosomal association and detection in ribosome profiling experiments[5–9]. Many lncRNAs exhibit tissue-specific expression, thus suggesting that they are subject to regulation[10,11]. However, transcription is a low-fidelity process[12] constrained by cell-type-specific chromatin architecture and transcription-factor expression, and it is therefore difficult to discriminate 'transcriptional noise' from functionally important spatiotemporally restricted expression. Noncoding transcripts originating from regulatory regions have been shown to indicate activation status; many such transcripts have been annotated as lncRNAs but are actively degraded by nuclear surveillance mechanisms and are unlikely to have *trans*-regulatory functions[13].

Only a small proportion of human lncRNAs have known regulatory functions[14]. It remains unclear which lncRNAs are likely to be functional as distinct RNA species interacting with proteins and nucleic acids[15]; which lncRNAs may indicate or influence transcriptional activity via the process of transcription itself but not by acting as functional RNAs[16]; and which lncRNAs are nonfunctional byproducts of a stochastic cellular environment. Attempts to classify lncRNAs rely on sequence conservation, chromatin modifications, or the genomic position and orientation relative to coding genes[17]. However, the heterogeneity in form and function of lncRNAs remains a major obstacle that makes it difficult to prioritize lncRNAs for functional characterization and to generalize knowledge derived from individual lncRNAs to other lncRNAs.

RNAs that share common steps of RNA biogenesis and maturation often have similar functions, as in the case of miRNAs, snRNAs and tRNAs. This idea is generalizable to mRNAs, if common RNA metabolism behavior is assumed to reflect common regulation by *trans*-acting factors, which have been shown to coordinate the expression of mRNAs encoding functionally related proteins[18]. Because lncRNAs (and small noncoding RNAs) do not encode a separate molecule and are themselves the final actors, their biogenesis and maturation necessarily constrain their functional capacity, which we define on the basis of whether they are functional and the types of regulation in which they may participate. (Thus, lncRNAs exhibiting similar RNA metabolism behaviors may have similar functional capacities.) Therefore, we collected and generated transcriptome-wide profiles of six hallmarks of RNA metabolism. Quantitative examination of transcription, splicing, degradation, localization and translation revealed differences between annotated lncRNAs and mRNAs. These observations motivated us to perform an annotation-agnostic unsupervised classification of RNAs on the basis of their full RNA metabolism profiles. We identified seven classes, each of which contained both mRNAs and lncRNAs to different extents and exhibited distinct evolutionary patterns and fitness constraints, differential sensitivity to cellular RNA-regulatory pathways and different relationships

[1]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany. [2]Department of Biology, Humboldt University, Berlin, Germany. [3]Center for Genomic Science of IIT@SEMM, Fondazione Instituto Italiano di Tecnologia, Milan, Italy. [4]Department of Computer Science, Humboldt University, Berlin, Germany. Correspondence should be addressed to N.M. (neelanjanmukherjee@gmail.com) or U.O. (uwe.ohler@mdc-berlin.de).

Received 17 May; accepted 18 October; published online 21 November 2016; doi:10.1038/nsmb.3325

among the steps of RNA metabolism. For lncRNAs, these classes provide a roadmap that indicates which transcripts may be functional and broadly suggests the types of regulation that they may perform.

RESULTS

Differences in expression and maturation between coding and noncoding genes

We performed strand-specific paired-end RNA sequencing in triplicate by using synthetic spike-in RNAs in human embryonic kidney cells (HEK293 cells, average depth of ~31 million uniquely aligned reads) to determine average RNA copy number per cell (cpc) for 27,803 genes (**Supplementary Fig. 1a–c**). We detected a low-expression regime (n = 13,685 genes) and a high-expression regime (n = 14,118 genes) with average RNA cpc values of 0.05 and 11.45, respectively, results typical of RNA-seq experiments. Protein-coding genes exhibited higher expression than lncRNAs and pseudogenes (**Fig. 1a**), an expected observation that we confirmed by examination of 101 different tissues and cell lines from The Encyclopedia of DNA Elements (ENCODE) (**Supplementary Fig. 1d**).

Unlike mRNAs, which can be translated numerous times, lncRNAs must be expressed at sufficient levels for the RNA products to function in a given cellular context. Only 20% of lncRNAs were in the abundant population, compared with 68% of coding genes. In agreement with the concept that introns can enhance expression[19], only 2.6% of protein-coding genes in the high-expression regime were intronless, compared with 32% and 51.9% of lncRNAs and pseudogenes, respectively (**Supplementary Fig. 1e**,**f**). Furthermore, multiexonic lncRNAs were less robustly processed (**Fig. 1b**; estimation of primary and mature expression in Online Methods). These results confirmed known differences in human steady-state mature-RNA levels[10] and RNA maturation between coding and noncoding RNAs, and most lncRNAs were not expressed at a cpc consistent with a *trans*-acting function for the RNA product. Because our cpc estimates were derived from a population of cells, we cannot exclude that low-cpc RNAs represent high expression in a minor subset of cells. However, recent single-molecule imaging of numerous lncRNAs has indicated that low expression is not explained by such 'jackpot' cells[20].

Progressive metabolic labeling of RNA

To determine the mechanistic basis of the differences in behavior between mRNAs and lncRNAs, we generated progressive snapshots of RNA production and maturation by metabolic labeling of RNA with 4-thiouridine (4SU)[21]. After treating cells with 4SU for 7.5, 15, 30, 45 or 60 min, we purified total RNA and subsequently biochemically separated and strand-specifically paired-end sequenced newly transcribed 4SU-labeled RNAs; we performed the experiment three different times on three different frozen HEK293 stocks (**Fig. 1c**, average depth of 18 million uniquely mapping read pairs). The fraction of primary transcripts in 4SU samples was higher than that in other samples representing different stages of RNA maturation, including genomic run-on sequencing (GRO-seq)[22] and RNA-seq of cellular fractions (nuclear, cytoplasm, cytosol and polyribosomal[23,24]) (**Supplementary Fig. 1g**). We detected coverage for 79.6 million nucleotides in the 4SU samples; in comparison, 171.7 million nucleotides of the human genome are annotated in GENCODEv19 (ref. 25) (**Supplementary Fig. 1h**). Unstable regions, such as introns of coding genes and lncRNAs (**Supplementary Fig. 1i–m**), exhibited the most pronounced coverage differences between 4SU labeling and GRO-seq. Overall, the two methods were comparable to analysis of RNA synthesis (R = 0.83), but 4SU labeling but did not require *in vitro* perturbation of nuclei.

We inferred synthesis, processing and degradation rates for genes throughout the transcriptome by comparing primary and mature RNA concentrations of 4SU-labeled RNA and total RNA by using INSPEcT[26]. The inferred rates from different labeling times were precise and consistent with the observed behavior of individual genes (**Fig. 1d–f**). Scrutiny of these loci revealed similar time-dependent increases in the production of mature mRNA for the well-transcribed genes *ACTB* and *MYC* (**Fig. 1g**,**h**). Because *ACTB* had higher steady-state levels than those of *MYC*, we were able to correctly infer that *ACTB* mRNA was more stable than *MYC* mRNA. We detected progressive increases in the production of the lncRNAs XIST and GAS5, though they exhibited less complete intron excision (**Fig. 1i**,**j**). The two lncRNAs exhibited slower splicing of individual introns and lower gene-level processing rates than those of the two coding genes, thus suggesting differences in splicing mechanisms.

Splicing differences between coding and noncoding genes

Together with previous studies indicating that lncRNAs are less efficiently spliced[27], the differences that we observed for intron removal prompted us to analyze splicing differences in the excision of individual introns from lncRNAs and from coding genes. For each junction with sufficient coverage across all labeling time points (n = 21,782), we calculated an intron-centric splicing value, θ (**Supplementary Fig. 2a**), which ranged from 0 (unspliced) to 1 (spliced). We identified three clusters of introns representing fast, medium and slow intron-excision dynamics (**Fig. 2a**). Introns of lncRNAs were 17.3 times more likely than introns of coding genes to belong to the slow class (**Fig. 2b**). Additionally, the slow class was 8.3 times more likely to contain exons exhibiting higher skipping (low spliced-in values, Ψ; Online Methods) (**Fig. 2c**). Mirtrons were spliced out more quickly than introns not containing small RNAs, whereas snoRNA-containing introns were spliced more slowly (**Fig. 2d**).

To investigate the mechanisms underlying the different splicing behavior, we used features corresponding to nucleotide composition and length, canonical splicing signals and exonic-splicing-regulatory elements as well as RNA metabolism (synthesis and decay on the gene level), in a random forest classifier trained to discriminate between the fast and slow intron classes (**Supplementary Fig. 2b**). The model was trained by using all features as well as separately excluding rates derived from the metabolic labeling. Both models had similar performance, with a modest improvement when metabolic features were included (area under receiver operating characteristic curve from 0.79 to 0.82; **Fig. 2e**). We found that similar features were important in an orthogonal approach using individual splicing models to predict θ for coding introns, lncRNA introns, snoRNA host introns and mirtrons by using random forest regression (**Supplementary Fig. 2c–f**).

The distance of introns from the transcription start sites (TSSs) and from the transcription end sites (TESs) were important physical features for prediction and positively correlated with splicing speed (**Fig. 2f**). The GC content of introns and flanking exons was more important for the prediction of fast introns and was significantly lower for fast introns and flanking exons. Regarding canonical splicing signals, fast introns exhibited stronger splice sites and weaker branchpoints but similar polypyrimidine-tract scores. Upstream and downstream exons flanking fast introns, compared with slow introns, exhibited significantly higher levels of ESEs and lower levels of ESSs ($P < 0.05$), a result consistent with recent evidence of the use of purifying selection to preserve exonic splicing signals in lncRNAs[28,29].

Higher synthesis rates for genes containing slow introns (**Fig. 2f**, 'syn') prompted us to examine the phosphorylation status of serine

residues in the C-terminal domain (CTD) of RNA pol II[30] at splice sites by using native elongating transcript sequencing data (NET-seq)[31]. In NET-seq, we observed similar total signal and signal for phosphorylation at the second peptide in the CTD heptapeptide repeats (denoted Ser2p) at the 5′ splice site (**Fig. 2g,i**), but substantially more unphosphorylated RNA poll II signal at the flanking

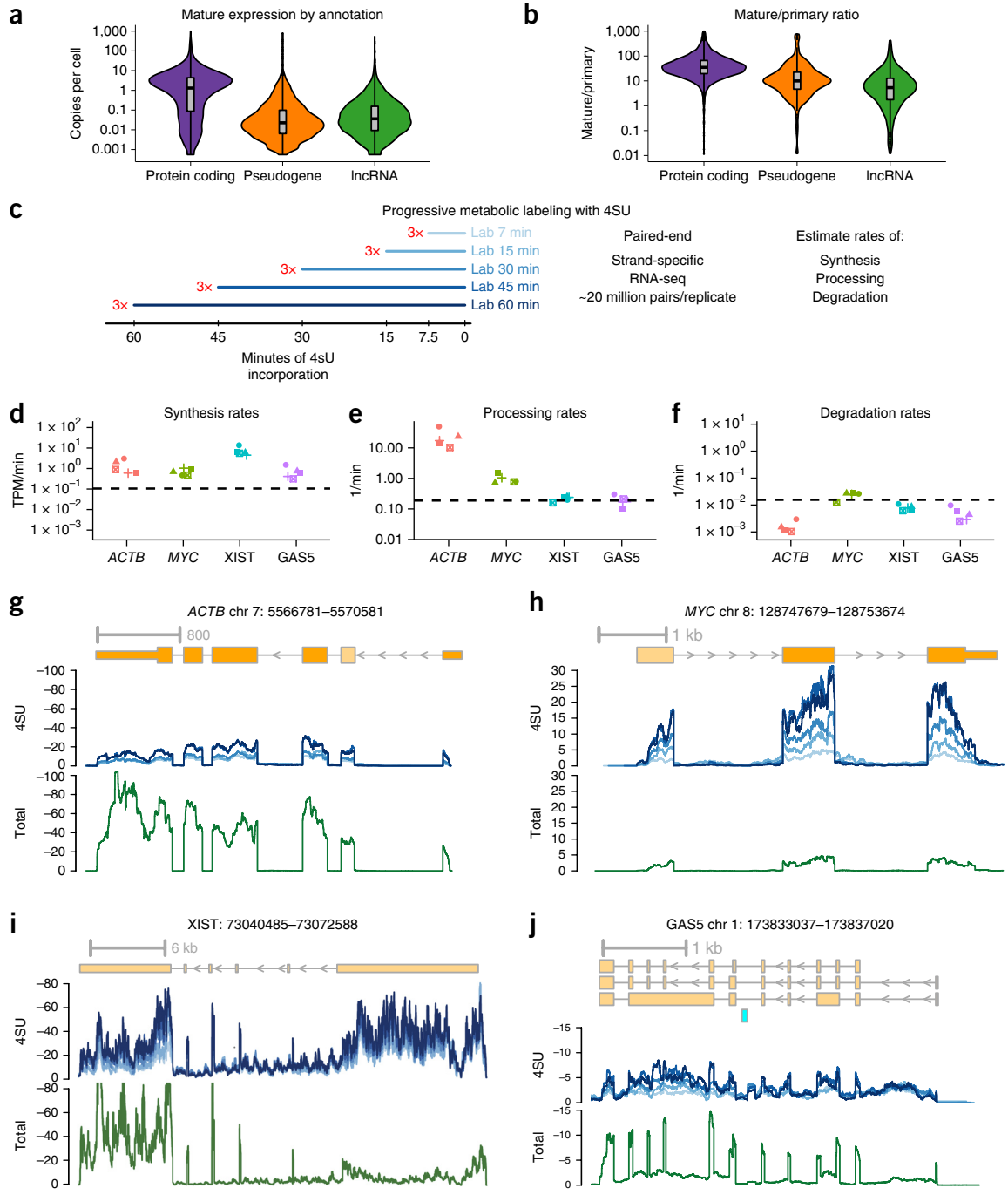

Figure 1 Progressive metabolic labeling of RNA. The average RNA cpc for endogenous genes was estimated from the fit determined for External RNA Controls Consortium (ERCC) spike-in RNAs. (**a**,**b**) Violin plots representing the density of the distribution with embedded box-and-whisker plots for the cpc of mature RNA for coding genes (n = 18,404), pseudogenes (n = 4,124) and lncRNAs (n = 5,275) (**a**) and the ratio of mature cpc values versus primary cpc values for the high-expression population of coding genes (n = 12,158), pseudogenes (n = 282) and lncRNAs (n = 713) (**b**). Center line, median; upper and lower hinges, first and third quartiles; whiskers, 1.5× interquartile range. (**c**) Time points of progressive 4SU labeling. The longer the labeling time (lab), the darker the blue line depicted (i.e., 7.5 min is the lightest blue, and 60 min is the darkest blue). All experiments were performed in triplicate (on three independent cell cultures). (**d**–**f**) Rates of synthesis (**d**), processing (**e**) and degradation (**f**) of *ACTB*, *MYC*, XIST and GAS5 for each of the five labeling times calculated with INSPEcT. The *y* axis represents the full range of values for each rate, and the dashed line is the average. For each of the five labeling time points (represented by different symbols), triplicates were used to estimate rates for all genes depicted. Chr, chromosome. (**g**–**j**) Coverage (library size in normalized fragments per million) of 4SU data (light to dark blue represents short to long labeling times, as depicted in **c**, and total RNA (green) profiles for *ACTB* (**g**), *MYC* (**h**), XIST (**i**) and GAS5 (**j**). Source data for **a**,**b**,**d**–**f** are available online.

5′ splice sites of slow introns (**Fig. 2h**). These data indicated the relative absence of proximal RNA pol II phosphorylation as a potential mechanism for the decreased splicing efficiency of these lncRNA introns, results consistent with previously reported *in vitro* results[32]. The GC content, splicing-regulatory elements and RNA pol II phosphorylation were very different between fast and slow introns

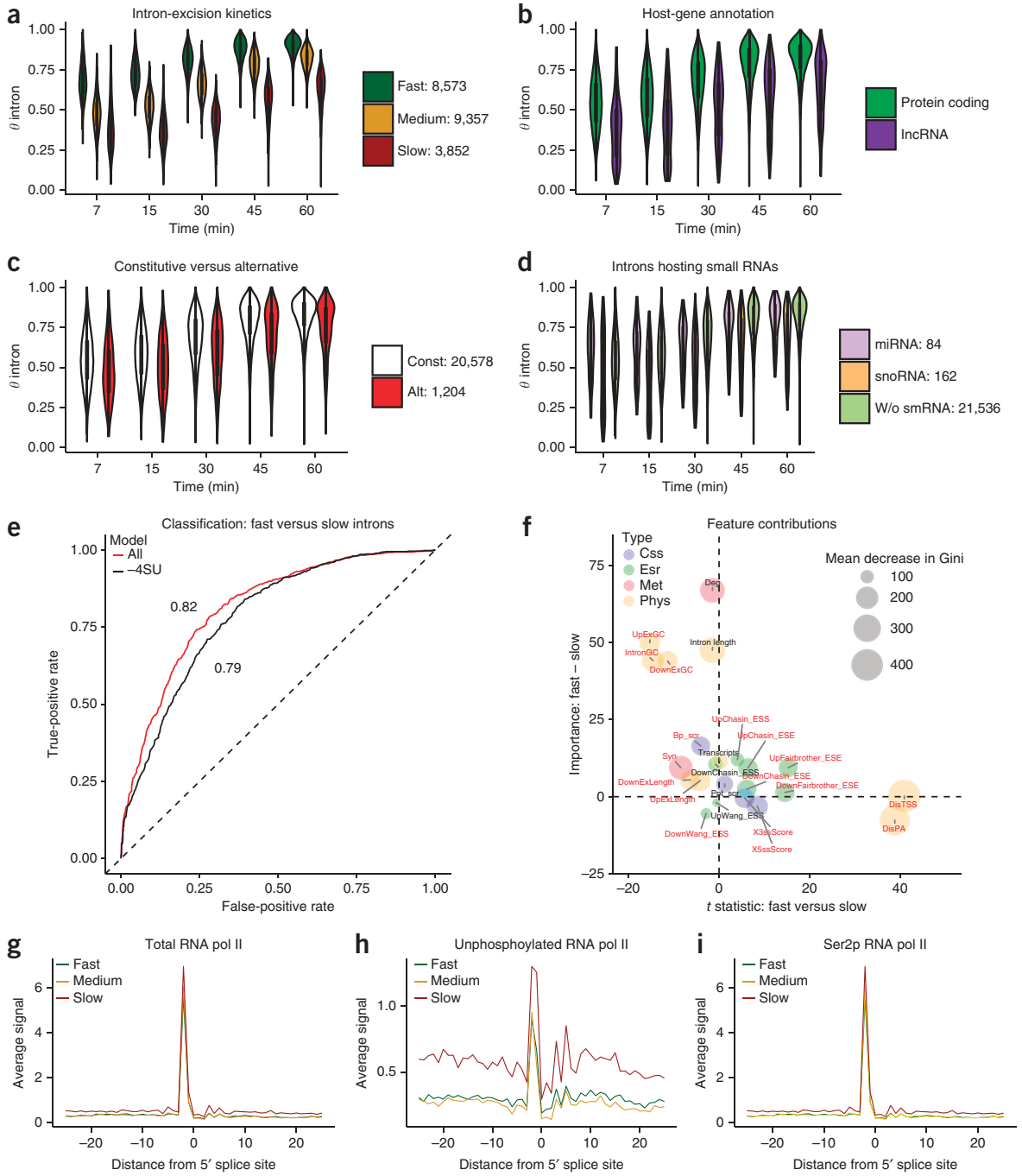

Figure 2 Dynamics of intron excision. Introns with fast, medium and slow intron-excision speeds were identified by clustering values of introns with sufficient data ($n = 21,782$) at all labeling time points. (**a**–**d**) Violin plots representing the density of the distribution with embedded box-and-whisker plots as in **Figure 1**, depicting the distribution of θ values for introns grouped on the basis of clustering excision dynamics (**a**), host-gene annotation (107,410 coding introns and 775 lncRNA introns) (**b**), constitutive or alternatively spliced adjacent exons (mean value = 0.37) (**c**) and the type of small RNA hosted by the intron (**d**). Const, consitutive; alt, alternative; w/o smRNA, without small RNA; TPM, transcripts per kilobase per million. (**e**) Average receiver operating characteristic (ROC) curve for predictions on the basis of five-fold cross-validation using all features (black) or all features excluding those derived from 4SU data, such as synthesis and degradation rates (red). (**f**) Bubble plot depicting the differential contribution and importance of features in classification. The differences (fast – slow) in the mean decrease in model accuracy for each class (y axis) plotted against the t statistic of the difference of means between intron classes (fast – slow). The circle size is the mean decrease in Gini coefficient representing the importance of that feature in the classification. Features with statistically significant differences (determined by two-sided t tests between fast introns, $n = 8,573$ and slow introns, $n = 3,852$; $P < 0.05$) are labeled in red. (**g**,**h**) The average NET-seq signal ±25 nt from the 5′ splice sites of total RNA poly II (**g**), unphosphorylated RNA pol II (**h**) and Ser2p RNA pol II (**i**) for fast (green), medium (yellow) and slow (red) introns. Source Data are available online.

and were important for classification as well as for discriminating coding and lncRNA introns.

Differences in RNA metabolism between coding and noncoding genes

Differences in steady-state transcript levels observed between lncRNAs and mRNAs must be due to differences in the metabolism of lncRNAs. To monitor the entirety of RNA life from synthesis to translation, we complemented the 4SU-derived features with the following quantitative estimates of subcellular localization and translation status all in HEK293 cells (Online Methods): the enrichment of cytosolic over nuclear expression (CytN)[23]; the enrichment of polyribosomal over cytosolic expression (PlyC)[24]; and translational potential (TrP), which represents the amount of translating ribosomes on a transcript on the basis of codon-by-codon movement[9]. We focused the rest of our analysis on the 15,120 genes for which we had complete data.

We observed a high correlation among rates calculated from different labeling times for all genes (**Supplementary Fig. 3a–d**), and the degradation rates were similar to previous estimates ($R = 0.53$)[33]. In agreement with earlier results from mouse cells[34], transcripts encoding ribosomal proteins exhibited high synthesis and processing as well as low degradation rates, thus confirming the quantitative behavior of the estimated rates (**Fig. 3a**). Very different RNA-metabolism patterns generated similar steady-state behavior: for example, although they exhibited similarly low steady-state RNA levels, mRNAs encoding KRAB-domain transcription factors exhibited average synthesis rates and high degradation rates, whereas mRNAs encoding extracellular proteins exhibited low synthesis rates and average degradation rates (**Fig. 3a**). These results highlight the limitations of standard RNA-seq protocols that assay only abundance.

The average synthesis and processing rates for lncRNAs were 3.4 and 2.9 times lower, respectively, than those for protein-coding genes (**Fig. 3b,c**). lncRNAs had average degradation rates 9.6 times higher than those of mRNAs (**Fig. 3d**), a result markedly different from those from previous less comprehensive studies reporting lncRNAs to be on average only 1.6 times[35] or 0.97 times[36] less stable than mRNAs. lncRNAs were modestly more nuclear than mRNAs (**Fig. 3e**), and their polyribosomal localization was similar (**Supplementary Fig. 3e**). The majority of lncRNAs did not contain actively translated ORFs[9] (**Fig. 3f**), thus confirming their status as bona fide noncoding RNAs. Among polyribosomal lncRNAs, those exhibiting evidence of translation were more transcribed (**Supplementary Fig. 3f**) and more stable (**Fig. 3g**).

Annotation-agnostic gene classification via RNA metabolism profiles

Although there were substantial differences in numerous aspects of RNA metabolism between annotated coding genes and lncRNAs, we also noticed that both exhibited a wide and overlapping range of behavior (**Figs. 2** and **3**). This result highlighted the heterogeneity in metabolism of both lncRNAs and mRNAs. We hypothesized that genes with similar RNA metabolism features would reflect the

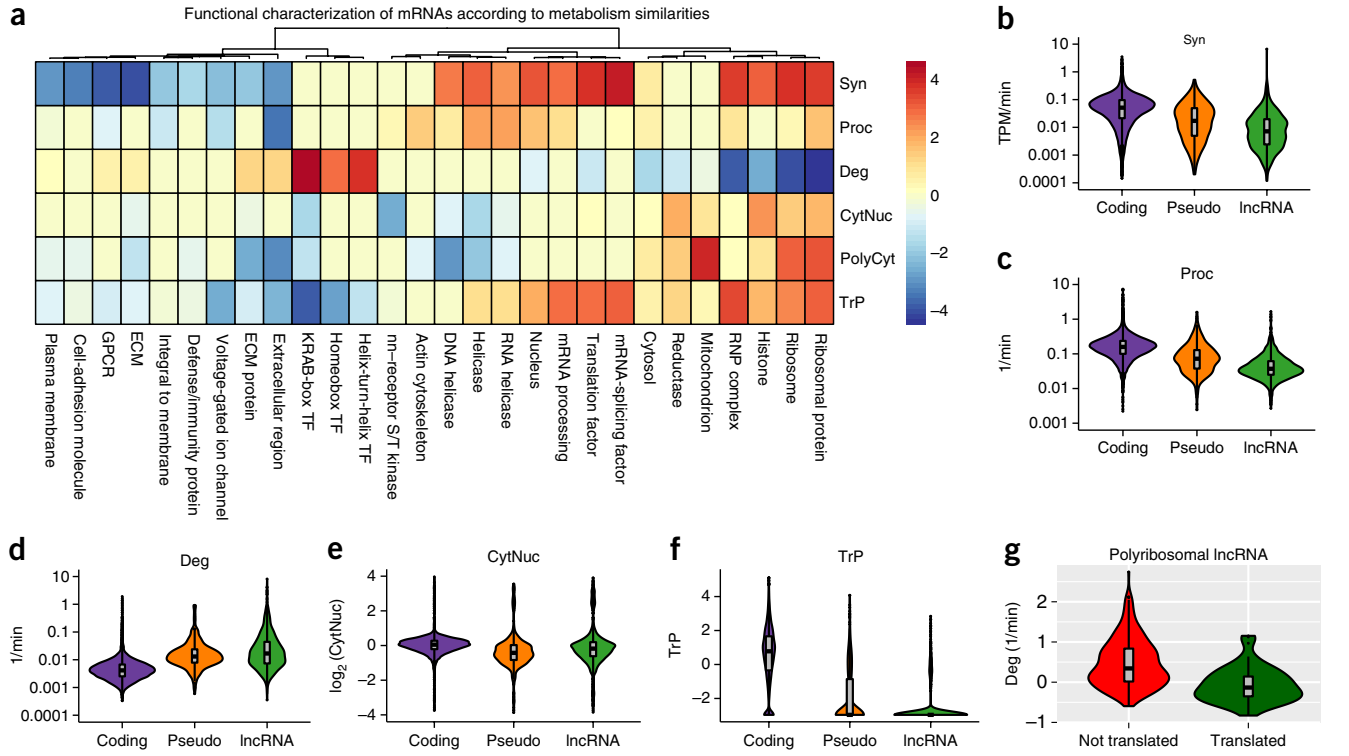

Figure 3 RNA metabolism of mRNA and lncRNA. (**a**) Clustering of gene-ontology enrichment for protein-coding genes on the basis of six hallmarks of RNA processing: synthesis rates (Syn), processing rates (Proc), degradation rates (Deg), cytoplasmic versus nuclear localization (CytNuc), polyribosomal versus cytoplasmic localization (PolyCyt) and translational status from ribosome profiling data (TrP). For each feature (for example, Syn), genes ($n = 15{,}120$) were rank-ordered, and the enrichment score (as in gene set enrichment analysis) was calculated for each PantherDB gene. Gene sets enriched at the top of the list (i.e., those with higher values for a given feature) have positive enrichment scores (red), and those enriched at the bottom of the list have negative enrichment scores (blue). (**b–f**) Box-and-whisker plots and violin plots as in **Figure 1**, for all detected genes ($n = 15{,}120$), depicting the distribution of synthesis rates (**b**), processing rates (**c**), degradation rates (**d**), cytoplasmic versus nuclear localization (**e**), translation for coding genes, lncRNAs and pseudogenes (pseudo) (**f**). (**g**) Distribution of degradation rates for polyribosomal lncRNA (PolyCyt > 0.1) divided into groups on the basis of the presence ($n = 58$) or absence ($n = 310$) of a RiboTaper-detected translated ORF from ribosome profiling data in HEK293 cells. Source data are available online.

coordinated activity of specific RNA-processing factors. Importantly, for noncoding RNAs this classification would help distinguish which, if any, of the types of regulatory mechanisms the RNAs might participate in. Therefore, we clustered all 15,120 genes *de novo* on the basis of all six features of the metabolism profiles. The number of clusters was determined by using the gap statistic (**Supplementary Fig. 4a**).

We identified seven RNA classes containing from 921 to 3,293 genes (**Fig. 4a**). Classes c1–c4 were enriched in coding genes (n = 10,793), although they contained 220 lncRNAs (**Fig. 4b**). Classes c5–c7 were enriched in lncRNA (n = 1,752), although they contained a similar number of coding genes (n = 1,838). Importantly, most GENCODE-defined lncRNA subcategories (for example, long intergenic noncoding RNAs or antisense) were primarily nonspecifically distributed across clusters (**Fig. 4c**), thus demonstrating that classification on the basis of positional genomic features does not coincide with specific lncRNA behavior. In agreement with results from earlier studies indicating that many processed-transcript RNAs are translated[9], this biotype was more likely than other lncRNA biotypes to be in c1–c4.

Even though coding genes were enriched in classes c1–c4, they occurred in all seven classes enriched in distinct characteristics and encoding functionally related proteins (**Supplementary Fig. 4b**). Genes in c1 and c2 exhibited the most similar RNA metabolism profiles. The c1 genes showed the highest expression and lowest tissue specificity, in agreement with 'housekeeping' genes such as mRNAs encoding ribosomal proteins. Genes in c3 were synthesized and processed well but had higher degradation rates and were less cytoplasmic; this class was enriched in transcription factors, particularly the aforementioned KRAB-domain family members. These genes may be subject to nuclear retention as a mechanism reducing cytoplasmic gene-expression noise created by transcriptional bursts[37,38]; indeed, c3 genes, relative to other classes, have been found to exhibit less cytoplasmic localization in a recent study in mouse liver tissue (**Supplementary Fig. 4d**). Genes in c4, compared with c1–c3, encoded receptors and had relatively lower synthesis, processing and translation rates along with higher degradation rates. These genes exhibited the most tissue specificity within c1–c4 and had steady-state expression levels similar to those of genes in c6. Within c5–c7, clusters with transcripts showing little to no evidence of translation, the c6 mRNAs had the highest synthesis rates and the most nuclear localization, and were enriched in pseudogenes along with calcium-channel- and ion-channel-encoding mRNAs. Genes in c5, typified by G-protein-coupled receptors and ion channels, and c7, typified by peptide hormones, exhibited similar steady-state levels of expression but very different degradation rates and localization. The c7 genes had the highest proportion of protein-coding genes that had no functional classification and were part of the 'missing' proteome[39] (**Supplementary Fig. 4c**), thus raising the possibility that some may not be (or may no longer be) protein coding.

Classes exhibited overlapping gene-expression distributions (**Supplementary Fig. 4e**) and thus were not recapitulated by steady-state expression or any individual RNA metabolism features. For all classes, the maximum expression levels of genes across 101 ENCODE tissues and cell lines were correlated with their expression in HEK293 cells (**Fig. 4d** and Online Methods). Classes c4–c7 exhibited the most tissue-specific expression (**Supplementary Fig. 4f**). Genes not expressed in HEK293 cells had the lowest expression levels across tissues (**Fig. 4e**). These results supported the generality of RNA-metabolism-derived classes and not simply HEK293 cell-specific behavior. Importantly, they highlighted the distinct utility of our new classes compared with existing classifications or biotypes (**Supplementary Fig. 4g**).

Classes exhibit specific evolution and fitness signatures

The seven classes exhibited differences in gene age (origination in vertebrate phylogeny derived from ref. 40). Classes c1, c2 and c3 were enriched in ancestral protein-coding genes predating the divergence of vertebrates (**Fig. 5a**). Many lncRNAs originated throughout mammalian and primate evolution, whereas pseudogenes were gained primarily in primates. Genes in c4 and c5 were gained before the divergence of eutherian mammals. Classes c3, c6 and c7 showed significant enrichment for gains in the primate lineage, and c6 and c7 were enriched in human-specific genes. These two classes had the highest degradation rates, which, compared with synthesis rates, were more strongly correlated with gene age (**Fig. 5b**,**c**). This result was consistent with observations of high turnover of lineage-specific lncRNAs in tetrapods[41] as well as rodents[42]. We found that young genes have managed to be synthesized but have not avoided being degraded, as has been suggested for the balance between splicing and polyadenylation of RNAs produced from divergent transcription[43].

Distinct regulatory pathways shape RNA classes

The classes responded in specific ways to perturbations in cellular RNA-quality-control and RNA-regulatory pathways (**Fig. 5d**). Genes in c3 were the most downregulated after depletion of ELAVL1 (HuR), an RNA-binding protein (RBP) antagonizing AU-rich element (ARE)-mediated decay. The genes' nuclear preference is thus aligned with the pre-mRNA-stabilizing function of HuR[44]. These genes also exhibited the highest 3′ untranslated region (UTR) ARE content and degradation rates except for c7, thus suggesting a particular sensitivity of these genes to cytoplasmic ARE-mediated-decay mechanisms. The nuclear poly(A)-binding protein (*PABPN1*) and poly(A) polymerase (*PAP*; official symbol *PAPOLA*)-mediated RNA decay (PPD) pathway limits the accumulation of inefficiently processed nuclear RNAs[45]. Both c5 and c7 genes exhibited the lowest processing rates and increased the most after overexpression of a dominant-negative *PABPN1* mutant (LALA) that binds RNA but cannot stimulate *PAP*[45]. Interestingly, only c5 and c7 genes exhibited strong negative correlation between processing rate and export to the cytoplasm (**Fig. 5f**,**h**), thus explaining the lower sensitivity of c6 genes to PPD. Furthermore, the classes exhibited unique response patterns to depletion of different RBPs in K562 cells from ENCODE (**Supplementary Fig. 5a**). Altogether, these results indicated that RNA classes are controlled by distinct regulatory pathways in multiple human cell types.

We examined the relationship between different steps of RNA metabolism by quantifying each pairwise association between two features while accounting for all other features (partial correlation analysis; Online Methods). Class c1 genes exhibited the most interdependence among different steps of RNA metabolism (**Fig. 5e**), whereas c6 and c7 (**Fig. 5g**,**h**) exhibited the least. The coupling among various steps of RNA metabolism is expected to be dependent on *cis*-regulatory elements allowing these genes to interact with the responsible RNA-processing machinery. Thus, less coupling of RNA metabolism is indicative of classes enriched in younger genes (**Fig. 5a**), which have a lower fraction of their sequences under evolutionary constraint.

lncRNAs in different classes exhibit distinct behavior

Finally, we focused on lncRNAs, first asking which classes were enriched in known functional lncRNAs from lncRNADB v2.0 (ref. 14). Classes c1, c2 and c3 exhibited the strongest overlap with lncRNADB (**Fig. 6a**), whereas c4 and c6 were only ~1.5 times more likely to be found in lncRNADB than expected. In contrast, lncRNADB genes were depleted in c5 and particularly in c7. GENCODE lncRNA biotypes showed less difference in overlap

with lncRNADB, and processed transcripts showed the strongest enrichment (**Supplementary Fig. 6a**).

To complement the gene origination analysis, we compared genic regions of different classes by using fitCons, an approach that integrates sequence divergence in primates and sequence polymorphisms among humans with functional genomic data to estimate the probability that a point mutation will have a fitness consequence[46]. The exonic sequences from lncRNA genes in c1–c4 and c6 had higher fitCons scores than

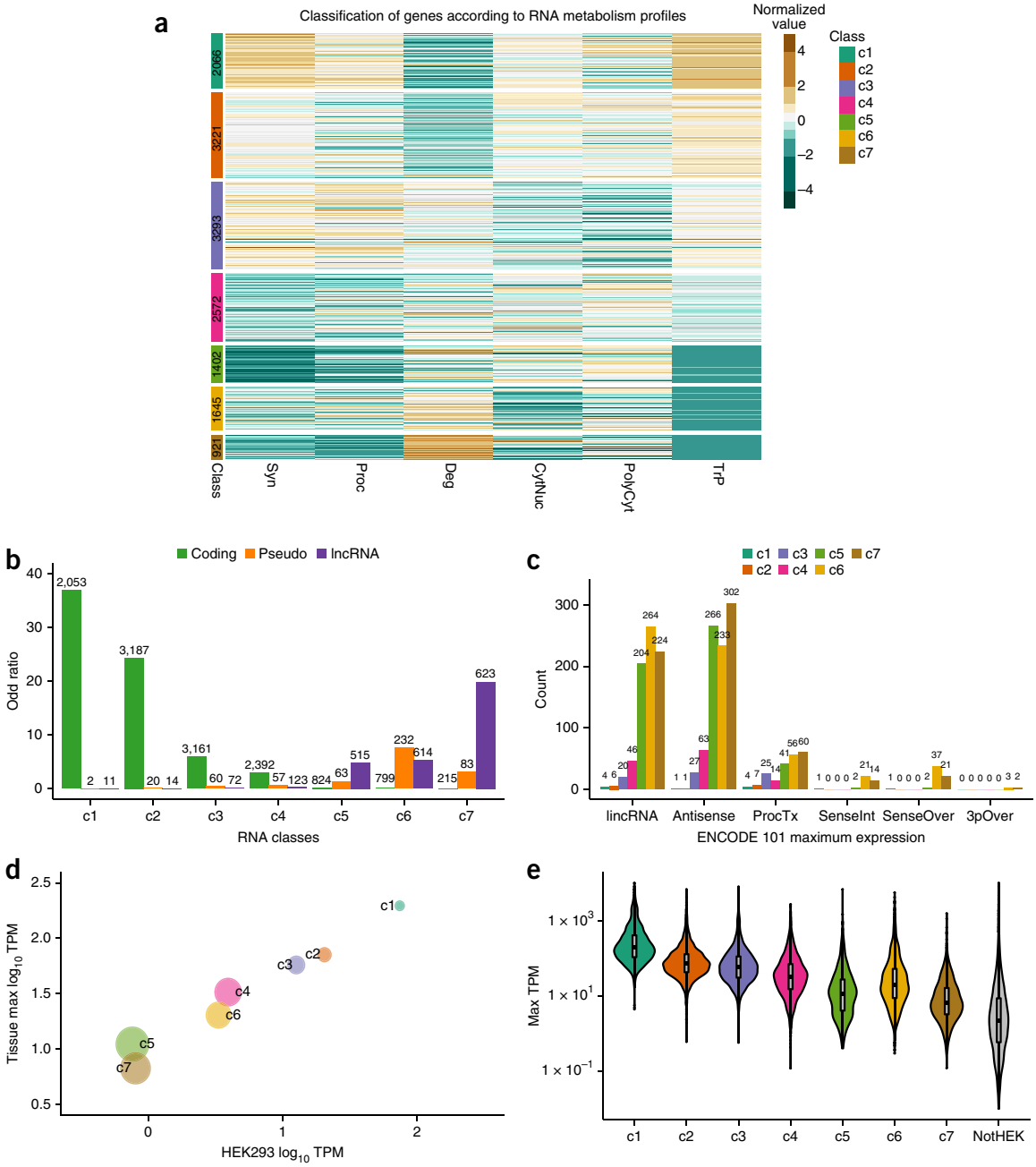

Figure 4 Classification of genes according to RNA metabolism profiles. (**a**) Centered and scaled values for each metabolic feature for each of seven classes determined by *k*-means clustering. The number on the class labels represents the number of genes in each class. (**b**) Odds ratio (Fishers exact test) for overlap between class membership and GENCODE V19 annotation classes. Numbers of genes in each class and annotation categories are shown above bars. (**c**) The number of lncRNA genes in a given GENCODE V19 biotype for each RNA class. Numbers of genes in each class and biotype category are shown above the bars. lincRNA, long intervening noncoding RNA; ProcTx, processed transcript; SenseInt, sense intronic; SenseOver, sense overlapping; 3pOver, 3′-UTR overlapping. (**d**) Bubble plot of median value of HEK293 expression (*x* axis) and of maximum (max) TPM across 101 tissues and cell lines from ENCODE (*y* axis) for all genes in a class (number of genes in each class shown in **a**. The bubble size represents the average tissue specificity score (a higher score indicates more tissue restricted). (**e**) Violin plots representing the density of the distribution with embedded box-and-whisker plots as in **Figure 1**, showing maximum TPM across 101 ENCODE cell lines and tissues for genes in each RNA class and for genes excluded from the clustering of RNA features because of insufficient data in HEK293 cells. The numbers of gene in each class are the same as in **a**; additionally, there were 37,408 genes in the 'NotHEK' category comprising genes not expressed or with insufficient expression in HEK293 cells. Source data are available online.

those in c5 and c7 (**Fig. 6b**). Examination of coding and UTR sequences of coding genes revealed a similar pattern (**Supplementary Fig. 6b**,**c**). Among the classes enriched in lncRNAs (c5–c7), c6 had higher fitCons scores and behaved more like c1–c4 genes. c5 lncRNA exons had lower fitCons scores than those from c7, which had substantially higher degradation rates. The lower constraint on younger c7 exonic

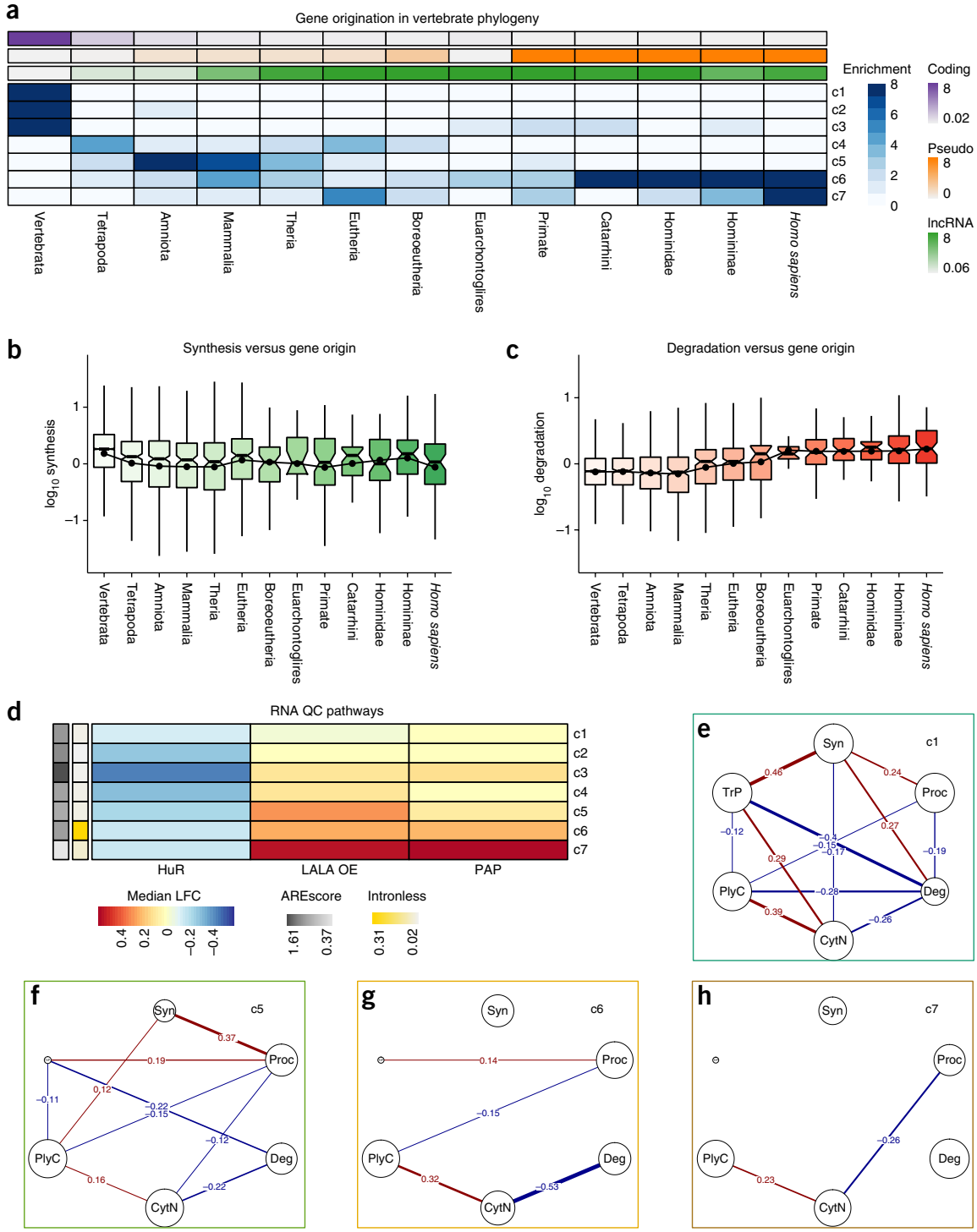

Figure 5 Evolutionary and regulatory differences among RNA classes. (**a**) Heat map of $-\log_{10} P$ value of overlap enrichment between the phylogenetic point of gene origination and RNA classes. Odds ratio for overlap between gene origination and annotation category, specifically for coding genes (purple), pseudogenes (orange) or lncRNAs (green). (**b**,**c**) Box-and-whisker plot of synthesis (**b**) and degradation (**c**) rates for genes grouped by origination class. The 'hinges' are the first and third quartile, the notches extend to ±1.58 the interquartile range divided by the square root of the number of genes in that group, and the whiskers extend to the first and 99th percentiles, respectively. A line connects the means (black dot) for each group (from left to right, the number of genes in each group is 9,004, 495, 556, 266, 203, 323, 58, 24, 66, 52, 58, 83 and 69, respectively). (**d**) $\log_2$ fold change of RBP perturbation minus control in HEK293 cells. Median AREscores of all 3′-UTR sequences from class (gray) and percentage intronless (yellow). (**e**–**h**) Partial correlation coefficient between different RNA metabolism features c1, c5, c6 and c7. Source data are available online.

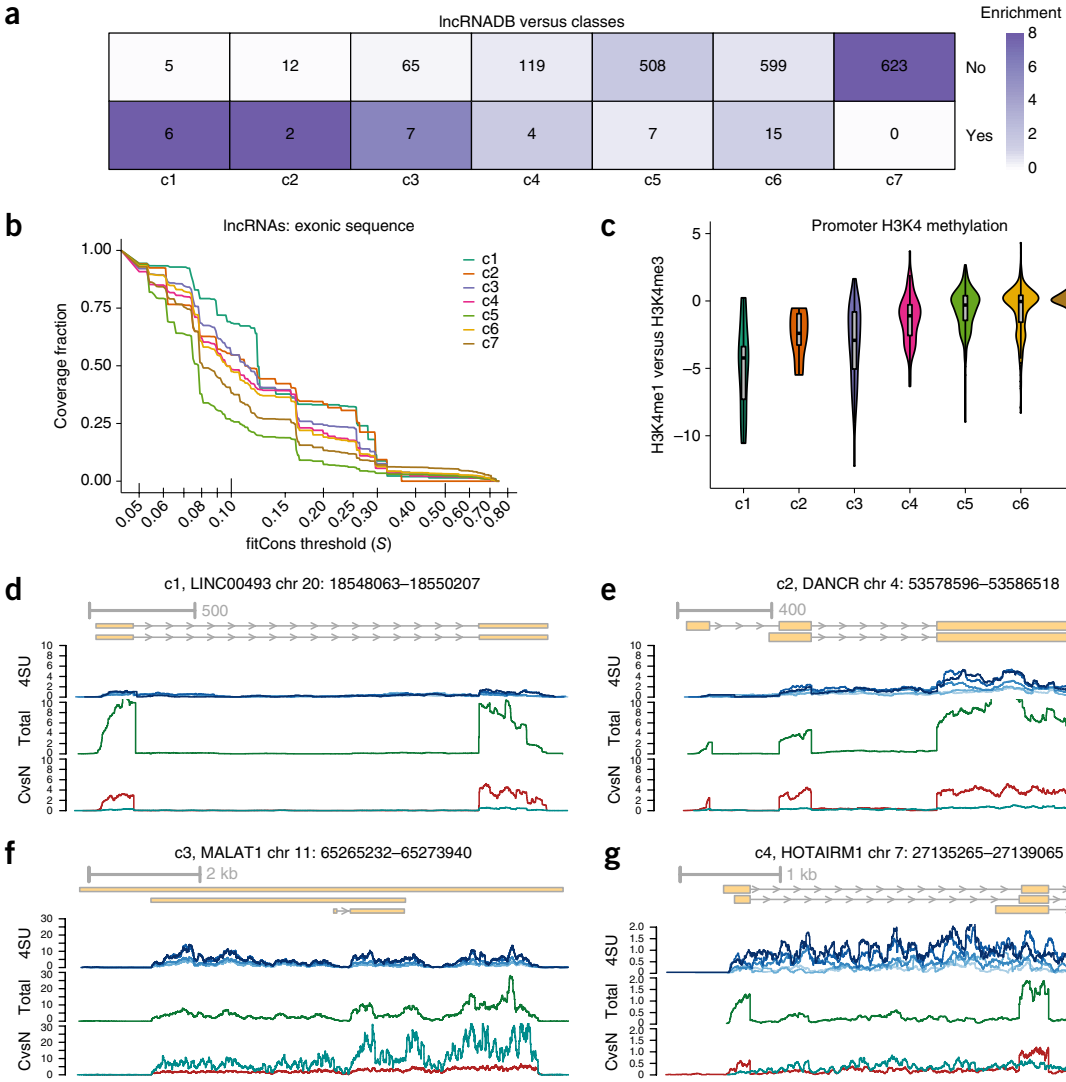

Figure 6 Distinct behavior of lncRNA classes. (**a**) The odds ratio of the overlap between lncRNAs that were either found ('yes') or not found ('no') in lncRNADB for each RNA class. The numbers represent the gene count in each category. (**b**) The fraction of nucleotides of lncRNA exons in a particular class with a fitCons score > *S*. (**c**) Violin plots representing the density of the distribution with embedded box-and-whisker plots of H3K4me1 versus H3K4me3 input normalized chromatin immunoprecipitation (ChIP)-seq signals in the promoter (−250 to +750 nt from the annotated TSS) of genes in each class (the number of genes are 10, 10, 55, 82, 363, 387 and 379, from left to right). (**d**–**g**) Coverage of 4SU (blue), total (green), cytoplasmic (red) and nuclear (cyan) RNA profiles for lncRNAs belonging to c1–c4. Source data for **a**–**c** are available online.

lncRNA sequences was consistent with results from earlier evolutionary studies[41,42]. These differences in 'fitness consequences' in the exonic sequence encoded by both coding genes and lncRNAs strongly support the utility of our classes (**Supplementary Fig. 6d**).

Histone modifications at TSSs have been used to distinguish between promoter loci (with trimethylated histone H3 K4 (H3K4me3)) and enhancer-like loci (with monomethylated H3 K4 (H3K4me1)). This approach has been used to define different classes of intergenic lncRNAs[47]. We compared the ratio of H3K4me1 versus H3K4me3 at annotated lncRNA TSSs of different classes (**Fig. 6c**). We found that c1–c4 promoters had low H3K4me1 and high H3K4me3 signals, a result consistent with 'promoter-like' behavior. The c5–c7 promoters had similarly low signals for both H3K4me1 and H3K4me3, a result more consistent with background transcription. In these classes, a minority of promoters had higher H3K4me1/H3K4me3 ratios, which may indicate enhancer-like properties that may be *cis* regulatory but are unlikely to function as RNAs.

The classification of lncRNAs on the basis of their RNA metabolism provided insights regarding whether they are bona fide lncRNAs and what their potential modes of action might be. For example, LINC00493 (**Fig. 6d**) belonging to c1, contained a highly translated ORF with peptide evidence, thus providing additional evidence to annul its status as a lncRNA[48]. DANCR (**Fig. 6e**), a member of c2, is involved in the antagonism of miRNAs[49,50], in agreement with its high stability and cytoplasmic localization. MALAT1 belonged to the category of stable and nuclear c3 genes (**Fig. 6f**), in agreement with its known role in splicing[51]. Genes in c4 and c6 had similar synthesis, processing and degradation rates, but c6 RNAs were strongly localized to the nucleus. Accordingly, TUG1, a lncRNA capable of recruiting Polycomb repressive complex 2 and silencing a subset of genes[52], was a member of c6 (**Supplementary Fig. 6e**). The lncRNAs in c4 may have cytoplasmic and/or nuclear functions. HOTAIRM1A, a c4 gene, has a nuclear function, given its ability to modulate the expression of neighboring HOX-encoding genes[53], although we also detected

a spliced cytoplasmic isoform (**Fig. 6g**). In this way, well-studied lncRNAs provide hypotheses regarding other lncRNAs belonging to the same class. The lncRNAs in c5 and especially c7 were probably transcriptional byproducts and thus were unlikely to be functional as RNAs (**Supplementary Fig. 6f**,**g**).

DISCUSSION

The comprehensive analysis presented here hinges on a multivariate integrated view of RNA processing, but it is instructive to place this analysis within the context of previous work examining only individual aspects. We found that lncRNA genes exhibited lower synthesis, processing and stability than coding genes. In agreement with the results from other studies in humans, we also observed that lncRNAs were only modestly more enriched than mRNAs in the nucleus[2,11,20], in spite of the popular notion of their strong nuclear localization[16]. Poorer processing and higher degradation of lncRNAs would partially explain this observation independently of the regulatory capacity of the lncRNA.

Lower steady-state lncRNA expression was explained by differences in both synthesis and degradation. Few studies have explicitly compared mRNA and lncRNA stability. A report on human cells has observed minimal differences in RNA stability[36]. Experiments in mouse neuroblastoma cells have also indicated similar stability[35]; in agreement with our findings, experiments in mouse dendritic cells have indicated high degradation rates for intergenic lncRNAs but similar synthesis and processing rates[54]. Beyond experimental, annotation, species- and context-specific differences, the heterogeneity of lncRNAs and the lack of a common classification make it inherently difficult to reconcile conclusions from different studies. Finally, a low-abundance lncRNA might have a *trans*-acting function if it were sufficiently stable, and it might even participate in sequential interactions, thereby 'amplifying' its regulatory capacity, similarly to mRNAs being translated multiple times. Thus our integrative classification of RNAs, which covered 15,120 genes, including 1,972 lncRNAs, should provide an important step toward gaining functional and relational coherence in the field.

The lncRNAs and coding genes were much more similar within a given class than across classes (**Supplementary Fig. 6e**), thus indicating that we did not randomly assign a smaller number of lncRNAs driven by the larger number of mRNAs. Attempts to classify lncRNAs independently of gene orientation (such as GENCODE biotypes) have largely relied on evolutionary sequence conservation and chromatin modifications, and in some cases have been restricted to intergenic lncRNAs. Overlapping and complex gene models limit the comprehensiveness and resolution of these classifications, but did not affect our classification, because our data are strand specific. Although this study focused on a single human cell line, we expect to see transcript- and context-specific differences in behavior. Furthermore, methods to address the evolutionary conservation and selection of RNA structural elements may be important for inferring lncRNA-mediated regulation in spite of the observation that lncRNAs fold less stably than mRNAs[55].

Our approach uncovered a considerable fraction of annotated protein-coding genes transcribed in a given cell type that are not processed and translated into proteins. This finding demonstrates that generating transcriptomic data beyond steady-state RNA-seq is both beneficial and necessary. Some of these transcripts are under active regulation, for instance through their retention in the nucleus, and are translated under specific conditions. Furthermore, any given gene exists on a continuum within a specific cellular and evolutionary context and may be in the process of gaining or losing specific characteristics. As such, some of the protein-coding genes that cluster with pseudogenes or unstable lncRNAs may never generate proteins, but this apparent pseudogeneization may have occurred too recently to allow for accumulating sequence changes typically used to identify pseudogenes. Likewise, some currently annotated lncRNAs encode translated ORFs and produce detectable peptides, and moreover their entire processing resembles that of protein-coding genes to such an extent that they can safely be reannotated. However, such RNAs represent a relatively small fraction; furthermore, most known functional lncRNAs are within RNA classes that show evidence of translation but lack known detectable peptides. The extent to which the translational signature relates to the noncoding function of these RNAs, or whether they may represent bifunctional genes, is unclear[56].

Our classification provides information on whether a given individual lncRNA is functional as well as the types of function (i.e., regulation of transcription, splicing, stability or translation) it may perform; however, additional data and analyses are necessary to accurately predict a precise molecular mechanism or targets of an individual lncRNA. Nevertheless, the processing patterns indicate that the majority of detectable lncRNAs do not encode RNAs that are likely to regulate gene expression *in trans*; comparatively few lncRNAs show evidence of 'enhancer-like' marks that may indicate *cis*-regulatory function. Altogether, these classes provide a rationale for prioritizing lncRNAs for which genetic evidence should be painstakingly collected[57,58] and suggest the potential types of regulation that they may impart.

METHODS

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

Accession codes. Total RNA-seq and metabolic labeling data have been deposited in the Sequence Read Archive under accession code GSE84722.

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

ACKNOWLEDGMENTS

U.O. acknowledges support from an award from the US National Institutes of Health (R01-GM104962) and the Simons Institute for the Theory of Computing at UC Berkeley, where he was a long-term visitor in the Algorithmic Challenges in Genomics Program in the spring of 2015. N.M. acknowledges support from EU Marie Curie IIF.

AUTHOR CONTRIBUTIONS

N.M. and U.O. conceived the project; N.M. and U.O. developed the methodology; N.M., L.C. and S.d.P. developed software and performed formal analysis; N.M. and A.H. conducted the investigation; N.M. conducted the visualization; N.M. and U.O. wrote the original draft; L.C., S.d.P. and M.P. reviewed and edited the paper; N.M. and U.O. acquired funding; N.M. and U.O. provided resources; N.M. and U.O. supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
2. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
3. Iyer, M.K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
4. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
5. van Heesch, S. *et al.* Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol.* **15**, R6 (2014).

6. Ingolia, N.T. *et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* **8**, 1365–1379 (2014).
7. 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).
8. Bánfai, B. *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22**, 1646–1657 (2012).
9. Calviello, L. *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13**, 165–170 (2016).
10. 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).
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. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* **14**, 103–105 (2007).
13. Andersson, R. *et al.* Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat. Commun.* **5**, 5336 (2014).
14. Quek, X.C. *et al.* lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.* **43**, D168–D173 (2015).
15. Rinn, J.L. & Chang, H.Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166 (2012).
16. Ulitsky, I. & Bartel, D.P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26–46 (2013).
17. St Laurent, G., Wahlestedt, C. & Kapranov, P. The landscape of long noncoding RNA classification. *Trends Genet.* **31**, 239–251 (2015).
18. Keene, J.D. RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.* **8**, 533–543 (2007).
19. Le Hir, H., Nott, A. & Moore, M.J. How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci.* **28**, 215–220 (2003).
20. Cabili, M.N. *et al.* Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* **16**, 20 (2015).
21. Windhager, L. *et al.* Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Res.* **22**, 2031–2042 (2012).
22. Fong, N. *et al.* Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* **28**, 2663–2676 (2014).
23. Sultan, M. *et al.* Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* **15**, 675 (2014).
24. Sterne-Weiler, T. *et al.* Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res.* **23**, 1615–1623 (2013).
25. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
26. de Pretis, S. *et al.* INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments. *Bioinformatics* **31**, 2829–2835 (2015).
27. 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).
28. Haerty, W. & Ponting, C.P. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* **21**, 333–346 (2015).
29. Schüler, A., Ghanbarian, A.T. & Hurst, L.D. Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol. Biol. Evol.* **31**, 3164–3183 (2014).
30. Hsin, J.-P. & Manley, J.L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev.* **26**, 2119–2137 (2012).
31. Nojima, T. *et al.* Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**, 526–540 (2015).
32. Hirose, Y., Tacke, R. & Manley, J.L. Phosphorylated RNA polymerase II stimulates pre-mRNA splicing. *Genes Dev.* **13**, 1234–1239 (1999).
33. Gregersen, L.H. *et al.* MOV10 Is a 5′ to 3′ RNA helicase contributing to UPF1 mRNA target degradation by translocation along 3′ UTRs. *Mol. Cell* **54**, 573–585 (2014).
34. Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29**, 436–442 (2011).
35. Clark, M.B. *et al.* Genome-wide analysis of long noncoding RNA stability. *Genome Res.* **22**, 885–898 (2012).
36. Tani, H. *et al.* Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.* **22**, 947–956 (2012).
37. Bahar Halpern, K. *et al.* Nuclear retention of mRNA in mammalian tissues. *Cell Rep.* **13**, 2653–2662 (2015).
38. Battich, N., Stoeger, T. & Pelkmans, L. Control of transcript variability in single mammalian cells. *Cell* **163**, 1596–1610 (2015).
39. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
40. Zhang, Y.E., Vibranovski, M.D., Landback, P., Marais, G.A.B. & Long, M. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol.* **8**, e1000494 (2010).
41. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).
42. Kutter, C. *et al.* Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* **8**, e1002841 (2012).
43. Wu, X. & Sharp, P.A. Divergent transcription: a driving force for new gene origination? *Cell* **155**, 990–996 (2013).
44. Mukherjee, N. *et al.* Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell* **43**, 327–339 (2011).
45. Bresson, S.M., Hunter, O.V., Hunter, A.C. & Conrad, N.K. Canonical poly(A) polymerase activity promotes the decay of a wide variety of mammalian nuclear RNAs. *PLoS Genet.* **11**, e1005610 (2015).
46. Gulko, B., Hubisz, M.J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).
47. 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).
48. Michalik, K.M. *et al.* Long noncoding RNA MALAT1 regulates endothelial cell function and vessel growth. *Circ. Res.* **114**, 1389–1397 (2014).
49. Kretz, M. *et al.* Suppression of progenitor differentiation requires the long noncoding RNA ANCR. *Genes Dev.* **26**, 338–343 (2012).
50. Yuan, S.X. *et al.* Long noncoding RNA DANCR increases stemness features of hepatocellular carcinoma by derepression of CTNNB1. *Hepatology* **63**, 499–511 (2016).
51. Tripathi, V. *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* **39**, 925–938 (2010).
52. 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).
53. Zhang, X. *et al.* A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood* **113**, 2526–2534 (2009).
54. Rabani, M. *et al.* High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell* **159**, 1698–1710 (2014).
55. Yang, J.-R. & Zhang, J. Human long noncoding RNAs are substantially less folded than messenger RNAs. *Mol. Biol. Evol.* **32**, 970–977 (2015).
56. Ulveling, D., Francastel, C. & Hubé, F. When one is better than two: RNA with dual functions. *Biochimie* **93**, 633–644 (2011).
57. Sauvageau, M. *et al.* Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife* **2**, e01749 (2013).
58. Bassett, A.R. *et al.* Considerations when investigating lncRNA function *in vivo*. *eLife* **3**, e03058 (2014).

ONLINE METHODS

Cell culture. HEK293 cells were cultured in high-glucose DMEM (Thermo Fisher, 41965039) supplemented with 10% FBS (Thermo Fisher, 16000044) and 1% pen/strep (Thermo Fisher, #15070063). Cells were authenticated and were free of mycoplasma contamination.

Total RNA. RNA from human embryonic kidney cells (HEK293) was extracted and purified with TRIzol (Life Technologies) and Direct-zol (Zymo), respectively. ERCC RNA standards were spiked into 2-µg aliquots. Next, three different approaches were used to exclude abundant rRNAs and enrich for RNAs of interest. Ribosomal RNA was depleted with (i) Ribozero Gold (Illumina) and (ii) hybridization with DNA probes complementary to rRNA, then treated with thermostable RNase H (Epicentre) as described in ref. 59. The final sample was subjected to one round of poly(A) selection with Dynabeads oligo$(dT)_{25}$ (Life Technologies). Expression levels of cellular transcripts and ERCCs were quantified with RSEM[60]. The fit of the known number of ERCC molecules and the amount of RNA per cell were used to estimate the transcripts per million units of abundance and to calculate the absolute average RNA copy number per cell. The strong positive association ($R^2 = 0.94$, **Supplementary Fig. 1a**) between the RNA-seq ERCC estimates and known ERCC concentrations permitted accurate RNA copy-number estimates. High- and low-expression regimes were determined by fitting a Gaussian mixture model with two components with *mclust*[61,62].

Metabolic labeling. Each replicate was performed on different days from different HEK293 stocks. Cells were thawed and passaged twice before each experiment. For each experiment 2.5×10^6 HEK293 cells were seeded in a 10 cm^2 tissue culture plate and incubated at 37 °C overnight. Cells were exposed to 500 µM 4SU for 7.5, 15, 30, 45 or 60 min. For the 7.5- and 15-min samples, 2×10 cm^2 tissue-culture plates were used. At the time of collection, medium was removed, and cells were washed in PBS. Cells were directly collected in 3 mL of TRIzol. Separation of labeled RNA was performed as previously described[63].

RNA-seq data processing. Reads were first mapped to rRNA sequences with bowtie v1.0.1 (https://sourceforge.net/projects/bowtie-bio/files/bowtie/1.0.1/bowtie-1.0.1-linux-x86_64.zip/download/)[64]. Reads that did not map to rRNA were subjected to our analysis pipeline performed with a combination of in-house scripts and published software.

The pipeline performs the following tasks:

1. Quantification of primary- and mature-transcript expression

Mature- and primary-transcript expression was quantified with RSEM v 1.2.11 (http://deweylab.biostat.wisc.edu/rsem/src/rsem-1.2.11.tar.gz)[60]. To calculate primary-transcript expression, we included an additional isoform corresponding to the sequence of the full gene locus. Specifically, we modified the GENCODEv19 gtf and used this as the input for the 'rsem-prepare-reference' function to generate a modified index used for quantification. For each gene, we calculated the expression of 'mature' RNA as the sum of all isoforms for that gene excluding the 'primary' transcript. For intronless genes, primary and mature expression values were summed.

2. Alignment and coverage

For alignment, reads were mapped to hg19 with STAR v2.4.0j (https://github.com/alexdobin/STAR/releases/tag/STAR_2.4.0j/) and an index built on gencode v19 (ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/gencode.v19.chr_patch_hapl_scaff.annotation.gtf.gz/)[65].

For coverage tracks, the unique read-depth normalized bedgraph output from STAR was converted into bigwig format with bedGraphToBigWig (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/bedGraphToBigWig). GRO-seq coverage comparison analysis and all coverage plots for specific genes were performed with these normalized bigwig files. We included comparisons to a deeper GRO-seq library generated in HeLa cells[66].

3. Calculation of splicing information

We used bam2ssj (https://github.com/pervouchine/bam2ssj)[67] to quantify reads overlapping exon-exon splice junctions and exon-intron or intron-exon splice sites. The software was slightly modified to output the geneIDs associated with each splice junction. A given junction/splice site was included in downstream analysis if it was supported by more than five reads. The output was used to create bed and fasta files necessary to calculate splicing features (described below).

4. Splicing features

5′- and 3′-splice-site strength was calculated with MaxENT (http://genes.mit.edu/burgelab/maxent/download/)[68]. Polypyrimidine and branchpoint scores were calculated with SVM-BPfinder (https://github.com/RegulatoryGenomicsUPF/svm-bpfinder/)[69] on the first 75 nt upstream of each 3′ splice site. Exonic enhancer and silencer density was calculated with SROOGLE (http://sroogle.tau.ac.il/SROOGLE.rar)[70]. Other features for exons and introns (length, GC content, distance to TSS/pA site) were calculated with bedtools and custom scripts.

Annotation. Gene annotation was retrieved from the GENCODE V19 according to the 'gene_type' tag in column 9. The 'Annotation' section of the **Supplementary Note** includes definitions of the annotation categories.

Inferring metabolism rates with INSPEcT. Rates of synthesis, processing and degradation were inferred for triplicates from independent cell cultures with INSPEcT[26]. We considered using DRiLL; however such an analysis would have required software that was not freely available. Earlier comparisons between the two methodologies have indicated that INSPEcT (freely available in Bioconductor) performs as well or better than DRiLL[26]. Instead of intronic and exonic RPKMs, which are typically used in INSPEcT, we provided primary and mature TPMs (described above) as input. Doing so prevented the potential overestimation of mature sequences due to consideration of all exonic sequences as being mature. We used rate estimates that accounted for the degradation of transcripts during the labeling pulse. Rates estimated from 7 and 15 min were systematically higher and exhibited a larger coefficient of variation (data not shown) that could potentially have been resolved by more efficient biotin conjugation methods[71].

Modeling intron excision. Only introns/junctions with more than five reads in all three replicates were used for downstream analysis. For each intron, we calculated the average θ and Ψ (**Supplementary Fig. 4a**) along with features described above (**Supplementary Fig. 4b**). We clustered Ψ values into either a 'constitutive' class (median $\Psi = 1$) or an 'alternative' class (median $\Psi = 0.37$) with *k*-means clustering ($k = 2$)[72]. We performed two tasks with the randomForest package in R[73]:

1. Classification of introns by excision kinetics

Individual introns were clustered on the basis of the θ value throughout the time course with *k*-means ($k = 3$) clustering.

Classification of fast versus slow introns was performed with (k) all features and (ii) all features except those derived from labeling data (for example, synthesis and degradation rates).

Similar results were achieved for both classifications independently of the fraction of data withheld for testing ($x = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$ and 0.9). Data were randomized, and then *x* fractions were set aside for testing, and the rest were used for training.

Feature importance was extracted with the *importance()* function from the randomForest package. The 'Importance' reported in **Figure 4c** is the node impurity, as measured by the Gini index.

2. Prediction of θ with random forest regression for different intron types.

We performed predictive modeling on four non–mutually exclusive classes of introns: (i) introns within protein-coding genes; (ii) introns within lncRNA genes; (iii) introns considered to be mirtrons (i.e., introns reported as mirtrons in ref. 74 and supported by HEK293 AGO2 PAR-CLIP data from ref. 75); and (iv) introns containing snoRNAs (i.e., introns annotated as snoRNAs and supported by HEK293 snoRNA-RBP PAR-CLIP data[76]).

We used all features as in the classification example, except that only Chasin ESE and ESS density were used, although similar results were achieved with other ESE/ESS sets.

The reported average R^2 for all regression models corresponded to the 'pseudo R-squared' value (1 − mse / Var(y)) calculated in the R randomForest package.

The reported importance represented the increase in the residual sum of squares.

We calculated two-sided *t* tests to identify features that were significantly different ($P < 0.05$) between fast and slow introns.

More details about importance are provided in the randomForest documentation: "The first measure is computed from permuting OOB data: For each tree, the prediction error on the out-of-bag portion of the data is recorded (error rate for classification, MSE for regression)[73]. Then the same is done after permuting

 doi:10.1038/nsmb.3325

each predictor variable. The difference between the two are then averaged over all trees, and normalized by the standard deviation of the differences. The second measure is the total decrease in node impurities from splitting on the variable, averaged over all trees. For classification, the node impurity is measured by the Gini index. For regression, it is measured by residual sum of squares."

Spatial meta-analysis. We used bigwig tracks of mammalian NET-seq data for RNA pol II with different phosphorylation status[31]. Meta-analysis was performed on specific sets of introns described at 50 nt upstream and downstream of the 5′ splice site. We used genomation to calculate the average signal of winsorized data (exclusion of outliers outside the 99th percentile that can skew the average profile)[77].

Overlap analysis. The odds ratios depicted in heat maps representing overlap between two lists were calculated with the GeneOverlap R package[78]. Unless otherwise indicated, $-\log_{10}(P \text{ values})$ and odds ratios >8 were set to 8.

Annotation-agnostic classification. We performed *k*-means clustering with the gap statistic to determine the number of clusters[79] on the following six hallmarks of RNA metabolism:

1. Synthesis rates

We used rates from all labeling times (7.5, 15, 30, 45, 60 min) as pseudo-replicates for all rates.

2. Processing rates

Processing rates for intronless genes were imputed with the MICE R package[80].

3. Degradation rates

Genes that had unassigned ('NA') degradation rates were excluded.

4. Cytoplasmic versus nuclear ratio

We calculated the $\log_2$ (mature cytoplasmic TPM + 1×10^{-4}) − $\log_2$ (mature nuclear TPM + 1×10^{-4}) from the above-described RSEM-based quantification of rRNA-depleted strand-specific paired-end RNA-seq libraries generated in ref. 23.

5. Polyribosomal versus cytosolic ratio

We calculated the $\log_2$ (mature polyribosomal TPM + 1×10^{-4}) − $\log_2$ (mature cytosolic TPM + 1×10^{-4}) from the above-described RSEM-based quantification of rRNA-depleted strand-specific paired-end RNA-seq libraries generated in ref. 24.

6. Translation

We used the Multitaper spectral coefficient at a frequency of 3 nt along the translated ORFs, with the R package 'multitaper'. This estimate, termed the 'translational potential' (TrP), quantitatively represents the amount of translating ribosomes weighted on the basis of their codon-by-codon (3-nt periodicity) movement, which has previously been shown to represent bona fide translation elongation[9]. The TrP was calculated with Ribo-seq data from HEK293 cells previously generated in our laboratory[9]. After the addition of a pseudocount (specifically the minimum TrP value), these values were $\log_{10}$ transformed and centered.

ENCODE data. Paired-end strand-specific total RNA-seq data were quantified with kallisto (https://github.com/pachterlab/kallisto/). We used the calculated TPM for only ~15,000 genes that were in the HEK293 clustering analysis. To facilitate comparisons with HEK293 data, we also quantified transcripts for the HEK293 total RNA with kallisto. These TPMs were quantile normalized and used for calculating the median or maximum expression, correlations, fold changes and tissue specificity. Samples exhibiting low correlations with other samples (Spearman's $\rho < 0.42$) were excluded. The mean was calulated for replicate samples. We used ENCODE data representing paired-end and strand-specifically sequenced total RNA (not poly(A)-enriched RNA) from 101 cell lines and tissues (described in the **Supplementary Note**). Tissue specificity scores were calculated with the Shannon entropy of the expression of a gene across all samples as described previously[81].

ChIP-seq data for H3K4me1 and H3K4me3 and respective inputs from HEK293 was downloaded from ENCODE. Reads were aligned with bowtie2 and then converted into a depth-normalized bigwig file. Bigwig files for each replicate were used to calculate promoter coverage defined as ±250 from the annotated start site for each gene analyzed and then averaged. To calculate the input-normalized signal for H3K4me3 and H3K4me1, the $\log_2$ of the input signal was subtracted from the $\log_2$ of matched H3K4me signal. Next, the normalized H3K4me3 was subtracted from the normalized H3K4me1. Promoter regions with no signal for either modification were excluded.

Miscellaneous. Heat maps were made with the *aheatmap* function from refs. 82–84. Coverage tracks for specific gene loci were made with *Gviz*[85]. Partial correlation analysis was performed with *ppcor*[86] and plotted with *qgraph*[87]. Edges with correlation coefficients <0.1 were not plotted. AU-rich-element content was calculated with AREscore[88].

Statistics. All statistical tests were performed in R. The *t* tests and nonparametric tests (KS and Wilcoxon tests) comparing distributions were performed with the base statistical functions. All specific tests are described in their respective sections of the Online Methods.

Code availability. Code and interactive-data visualization is available at https://ohlerlab.mdc-berlin.de/software/Classification_of_human_genes_by_RNA_metabolism_profiles_130/.

Data availability. Paired-end-read data for total RNA-seq and metabolic labeling data have been deposited in the Sequence Read Archive under accession code GSE84722. These data also include processed data that may serve as a starting point for many analyses. Source data for **Figures 1–6** are available with the paper online. Any other data are available from the corresponding authors upon request.

59. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–629 (2013).
60. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
61. Fraley, C., Raftery, A.E., Murphy, T.B. & Scrucca, L. *Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation* (Department of Statistics, University of Washington, 2012).
62. Fraley, C. & Raftery, A.E. Model-based clustering, discriminant analysis and density estimation. *J. Am. Stat. Assoc.* **97**, 611–631 (2002).
63. Doelken, P., Huggins, J.T., Goldblatt, M., Nietert, P. & Sahn, S.A. Effects of coexisting pneumonia and end-stage renal disease on pleural fluid analysis in patients with hydrostatic pleural effusion. *Chest* **143**, 1709–1716 (2013).
64. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
65. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
66. Duttke, S.H. *et al.* Human promoters are intrinsically directional. *Mol. Cell* **57**, 674–684 (2015).
67. Pervouchine, D.D., Knowles, D.G. & Guigó, R. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29**, 273–274 (2013).
68. Yeo, G. & Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
69. Corvelo, A., Hallegger, M., Smith, C.W.J. & Eyras, E. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput. Biol.* **6**, e1001016 (2010).
70. Schwartz, S., Hall, E. & Ast, G. SROOGLE: webserver for integrative, user-friendly visualization of splicing signals. *Nucleic Acids Res.* **37**, W189–W192 (2009).
71. Duffy, E.E. *et al.* Tracking distinct RNA populations using efficient and reversible covalent chemistry. *Mol. Cell* **59**, 858–866 (2015).
72. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2015).
73. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
74. Ladewig, E., Okamura, K., Flynt, A.S., Westholm, J.O. & Lai, E.C. Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome Res.* **22**, 1634–1645 (2012).
75. Wiwie, C., Baumbach, J. & Röttger, R. Comparing the performance of biomedical clustering methods. *Nat. Methods* **12**, 1033–1038 (2015).
76. Kishore, S. *et al.* Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome Biol.* **14**, R45 (2013).
77. Akalin, A., Franke, V., Vlahoviček, K., Mason, C.E. & Schübeler, D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* **31**, 1127–1129 (2015).
78. Shen, L. *GeneOverlap: Test and Visualize Gene Overlaps* (Mount Sinai, 2013).

79. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Series B Stat. Methodol.* **63**, 411–423 (2001).
80. van Buuren, S. & Groothuis-Oudshoorn, K. Mice: multivariate imputation by chained equations in r. *J. Stat. Softw.* **45**, 1–67 (2011).
81. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).
82. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).
83. Gaujoux, R. & Seoighe, C. *Using the Package nMF* (CRAN, 2015).
84. Gaujoux, R. & Seoighe, C. *The Package nMF: Manual Pages* (CRAN, 2015).
85. Hahne, F. & Ivanek, R. in *Statistical Genomics: Methods and Protocols* (eds. Mathé, E. & Davis, S.) 335–351 (Springer, 2016).
86. Kim, S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Commun. Stat. Appl. Methods* **22**, 665–674 (2015).
87. Epskamp, S., Cramer, A.O.J., Waldorp, L.J., Schmittmann, V.D. & Borsboom, D. Qgraph: network visualizations of relationships in psychometric data. *J. Stat. Softw.* **48**, 1–18 (2012).
88. Spasic, M. *et al.* Genome-wide assessment of AU-rich elements by the AREScore algorithm. *PLoS Genet.* **8**, e1002433 (2012).

 doi:10.1038/nsmb.3325