Intergenic RNA mainly derives from nascent transcripts of known genes

Author	Federico Agostini, Julian Zagalak, Jan Attig, Jernej Ule, Nicholas M. Luscombe
journal or publication title	Genome Biology
volume	22
number	1
page range	136
year	2021-05-05
Publisher	BMC
Rights	(C) 2021 The Author(s).
Author's flag	publisher
URL	http://id.nii.ac.jp/1394/00001955/

doi: info:doi/10.1186/s13059-021-02350-x

Agostini et al. Genome Biology (2021) 22:136
https://doi.org/10.1186/s13059-021-02350-x

Genome Biology

RESEARCH

Open Access

Intergenic RNA mainly derives from nascent transcripts of known genes

Federico Agostini[1*], Julian Zagalak[1,2], Jan Attig[1], Jernej Ule[1,2†] and Nicholas M. Luscombe[1,3,4†]

* Correspondence: federico.agostini@scilifelab.se
†Jernej Ule and Nicholas M. Luscombe are senior authors.
[1]The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK
Full list of author information is available at the end of the article

Abstract

Background: Eukaryotic genomes undergo pervasive transcription, leading to the production of many types of stable and unstable RNAs. Transcription is not restricted to regions with annotated gene features but includes almost any genomic context. Currently, the source and function of most RNAs originating from intergenic regions in the human genome remain unclear.

Results: We hypothesize that many intergenic RNAs can be ascribed to the presence of as-yet unannotated genes or the "fuzzy" transcription of known genes that extends beyond the annotated boundaries. To elucidate the contributions of these two sources, we assemble a dataset of more than 2.5 billion publicly available RNA-seq reads across 5 human cell lines and multiple cellular compartments to annotate transcriptional units in the human genome. About 80% of transcripts from unannotated intergenic regions can be attributed to the fuzzy transcription of existing genes; the remaining transcripts originate mainly from putative long non-coding RNA loci that are rarely spliced. We validate the transcriptional activity of these intergenic RNAs using independent measurements, including transcriptional start sites, chromatin signatures, and genomic occupancies of RNA polymerase II in various phosphorylation states. We also analyze the nuclear localization and sensitivities of intergenic transcripts to nucleases to illustrate that they tend to be rapidly degraded either on-chromatin by XRN2 or off-chromatin by the exosome.

Conclusions: We provide a curated atlas of intergenic RNAs that distinguishes between alternative processing of well-annotated genes from independent transcriptional units based on the combined analysis of chromatin signatures, nuclear RNA localization, and degradation pathways.

Keywords: RNA, RNA-seq, Transcription, Gene annotation

Background

Studies estimate that up to 85% of the human genome is pervasively transcribed by RNA polymerase II (Pol II), resulting in a plethora of RNA products [1–4]. Many of these transcripts belong to well-established categories, such as messenger RNAs (mRNAs) which are characterized by the presence of 5′ cap, coding sequence (CDS), and poly(A) tail. Other transcripts are categorized as long non-coding RNAs (lncRNAs), generally defined as RNA molecules longer than 200 nt with little coding potential. Currently, lncRNAs are divided into three major groups depending on their genomic location relative to protein-coding genes: promoter upstream transcripts (PROMPTs), produced up to 2.5 kb upstream of active transcription start sites (TSSs) [5]; enhancer RNAs (eRNAs), bi-directionally transcribed from enhancer DNA elements [6, 7]; and large intervening non-coding RNAs (lincRNAs), located in intergenic regions, distal from protein-coding genes, and regulated as independent transcriptional units [8]. Gene and transcript annotations for the human genome are continuously updated, and their assignment to specific biotype categories can change across reference databases [9]. In particular, in the past decade, efforts towards the identification and characterization of novel lncRNA genes have been made, either through computational predictions or functional assays [10, 11]. Despite such endeavors, a marked proportion of RNA-seq reads from human cells still maps to unannotated, ostensibly intergenic portions of the human genome [12]. It is therefore often challenging to understand whether such reads originate from independent transcription units or are associated with annotated genes.

Many well-characterized lncRNAs, such as the X-inactive-specific transcript *Xist* [13], share-processing features (e.g., 5′ m^7G cap and poly(A) tail) with mRNAs [8] and have specific, experimentally validated functions. However, the majority of lncRNA gene loci might not function through their resulting products, but rather through the act of transcription itself, which for instance can affect the expression of neighboring genes [14–16]. In support of this view, studies have highlighted how ncRNA genes are associated with early transcriptional termination of Pol II and their products undergo rapid post-transcriptional degradation [3, 17–20], thus explaining their low nuclear abundance. Further, recent studies indicate a possible scenario in which nascent transcripts from protein-coding genes play a similar role by regulating chromatin remodeling [21]. For example, the binding of Polycomb repressive complex 2 (PRC2) to genomic targets was initially ascribed to a specific set of lncRNAs [22–25]. However, it was later shown that PRC2 also binds nascent, unspliced mRNAs, which sequester the complex, thus preventing gene silencing [26–29].

In addition to mRNAs and lncRNAs described above, downstream of gene transcripts (DoGs) arises when Pol II terminates far downstream of the ends of genes [30]. These readthrough transcripts appear to be linked to stress conditions, such as osmotic and oxidative stress [30, 31]. It remains unclear whether transcription of DoGs has any gene regulatory function, but possible roles range from antisense-mediated gene expression control [32] to maintenance of local open chromatin structure. Moreover, their regulation remains largely unknown. Nevertheless, the existence of DoGs increases the complexity of transcriptome annotation, posing additional challenges to the understanding of function and regulation of intergenic transcripts.

In a recent study [33], we performed RNA-seq of the nuclear and cytoplasmic compartments of untreated HeLa cells and found that an unexpectedly large fraction (7.63%) of nuclear RNA-seq reads derived from intergenic genomic regions. Since the majority of these reads (60.3%) could not be detected in the cytoplasmic samples, here, we seek to investigate their transcriptional origin. We developed a computational method to identify and classify the sources of intergenic transcription. We investigate their characteristics, expression patterns, and epigenetic environment. Specifically, we observe that the largest fraction of intergenic RNA corresponds to DoGs, upstream of gene transcripts (UoGs), which likely result from alternative TSSs upstream of annotated genes, and linker of genes (LoGs), which are DoGs that continue into the neighboring gene body. We find that most intergenic RNA is generated during transcription associated with annotated genes and is confined to chromatin due to efficient degradation of DoGs and LoGs by XRN2, and UoGs by the exosome. Most remaining intergenic RNA corresponds to poorly spliced lncRNAs that are degraded by the exosome. We conclude that most of the unannotated intergenic RNAs are the consequence of non-productive transcription associated with known genes, and exert their potential functions locally, before being rapidly removed through cellular quality control mechanisms.

Results

Identification of intergenic transcriptional units

To gain a comprehensive overview of the transcriptional landscape, we identified 38 publicly available datasets containing chromatin and nuclear fractionated RNA-seq samples. These cover 5 human cell lines (HeLa, HEK293, HepG2, K562, HCT116) and four subcellular fractions (cytosolic, nuclear, chromatin, and nucleoplasm). The initial processing and mapping to the human genome yielded > 2.5 billion uniquely mapped reads (Fig. 1 and Additional file 1: Table S1, S2, and S3). We employed StringTie [34] to generate preliminary annotations of the transcriptional units expressed within each dataset. We then merged the results into a comprehensive transcriptomic assembly across the entire dataset and also included all genes present in the GENCODE reference annotation [35]. Finally, we employed a custom pipeline (see the “Methods” section) to annotate transcripts expressed in intergenic regions and to define their relationship with annotated genes (Figs. 1 and 2a; see the “Methods” section). We defined transcriptional units (TU) as products of transcription from intergenic portions of the genome, which can either take place as an independent event or in association with features in the reference annotation.

We classified TUs into two broad groups based on their genomic location relative to existing gene annotations (Fig. 2a, b). (i) Gene-associated TUs are those showing continuity of transcription from the body of annotated genes. These were further divided into upstream of gene (UoG), downstream of gene (DoG), and linker of genes TUs (LoG). (ii) Independent TUs are > 10 kb away from existing gene annotations and so classed as purely intergenic. Although we use the term “independent TUs” for the sake of clarity, it is possible that some might in the future end up annotated as new genes that produce functional non-coding RNAs or perhaps even protein-coding mRNAs. Indeed, up to 80% (1453) of these independent TUs overlap with transcriptional products annotated in public databases (e.g., NONCODE [36], CHESS [37], and RefSeq [38]; Additional file 1: Table S4), while a number of the independent TUs we identified (373) have not been reported in any of the aforementioned databases, thus indicating that our results, on top of being supported by external sources, also expand the current catalog of transcribed RNAs. In total, we classified 7411 TUs covering ~ 5.6% of the human genome (Fig. 2c). Both gene-associated and independent TUs are of comparable lengths to previously annotated long non-coding RNA genes, ranging from 1 kb (the minimum length threshold for a TU) to hundreds of kb (Fig. 2d).

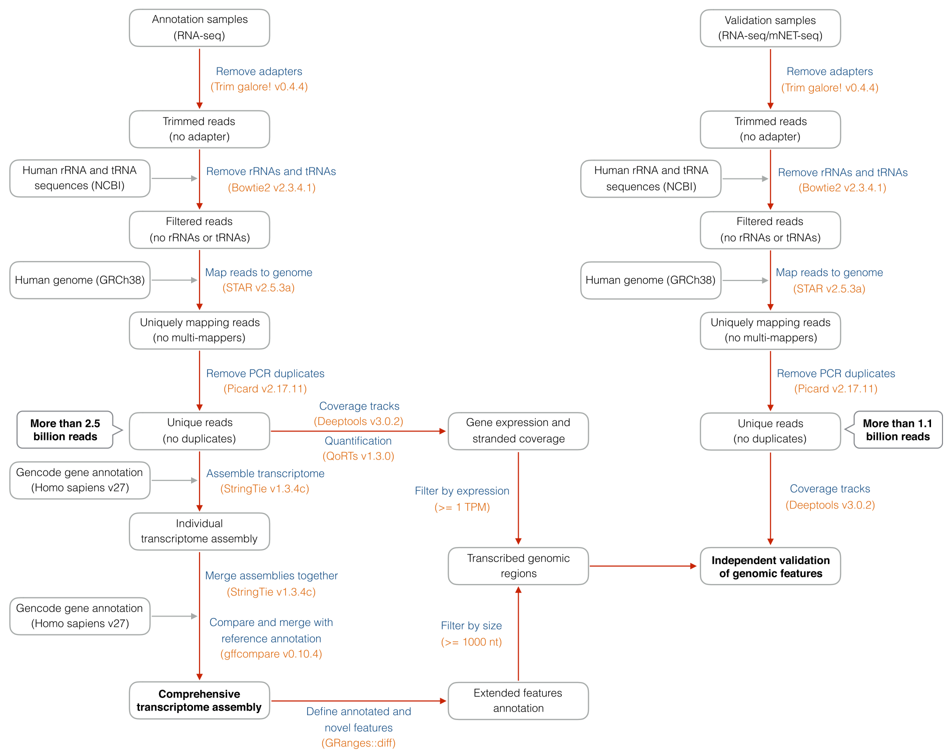

Fig. 1 Flow chart of the data analysis pipeline. Schematic describing the main data processing steps, intermediate and final outputs of the analysis pipeline, applied to RNA-seq (left side) and other sequencing (NET-seq, right side) data. Procedures (blue) and tools (orange) are indicated

Assessing the expression levels in HeLa cells, it is apparent that the relative abundance of TUs is higher in the nucleus (4.28% of mapped reads) compared with the cytoplasm (1.34% of mapped reads; Fig. 2e). Moreover, within the nucleus, TUs tend to be chromatin-associated (5.2% of mapped reads) rather than the nucleoplasm (1.21%) (Fig. 2e). Intriguingly, we noticed that about 80% of these reads mapped to features linked to transcription of previously annotated loci (i.e., gene-associated TUs), while the remainder belongs to independent TUs (Fig. 2e). We also compared the normalized expression levels of TUs with annotated genes (Fig. 2f). Protein-coding transcripts tend to be the most highly expressed within the chromatin-associated and nucleoplasmic compartments; however, in these subcellular fractions, both gene-associated and independent TUs are more highly expressed than annotated lncRNAs. Additionally, we found independent TUs tend to undergo less splicing than lncRNAs (Additional file 1: Figure S1B). In agreement with previous reports [30, 39], DoGs and LoGs show the highest expression among TUs, suggesting that levels of transcription outside annotated loci primarily depend on the activity of annotated upstream features (Fig. 2f).

To investigate the properties of TUs in greater detail, we focused further analysis to those with the strongest evidence:

- Both the gene-associated TU and neighboring genes must have TPM expression ≥1 (Additional file 1: Figure S1D) and length ≥ 5 kb (to avoid overlaps when assessing meta-profiles).

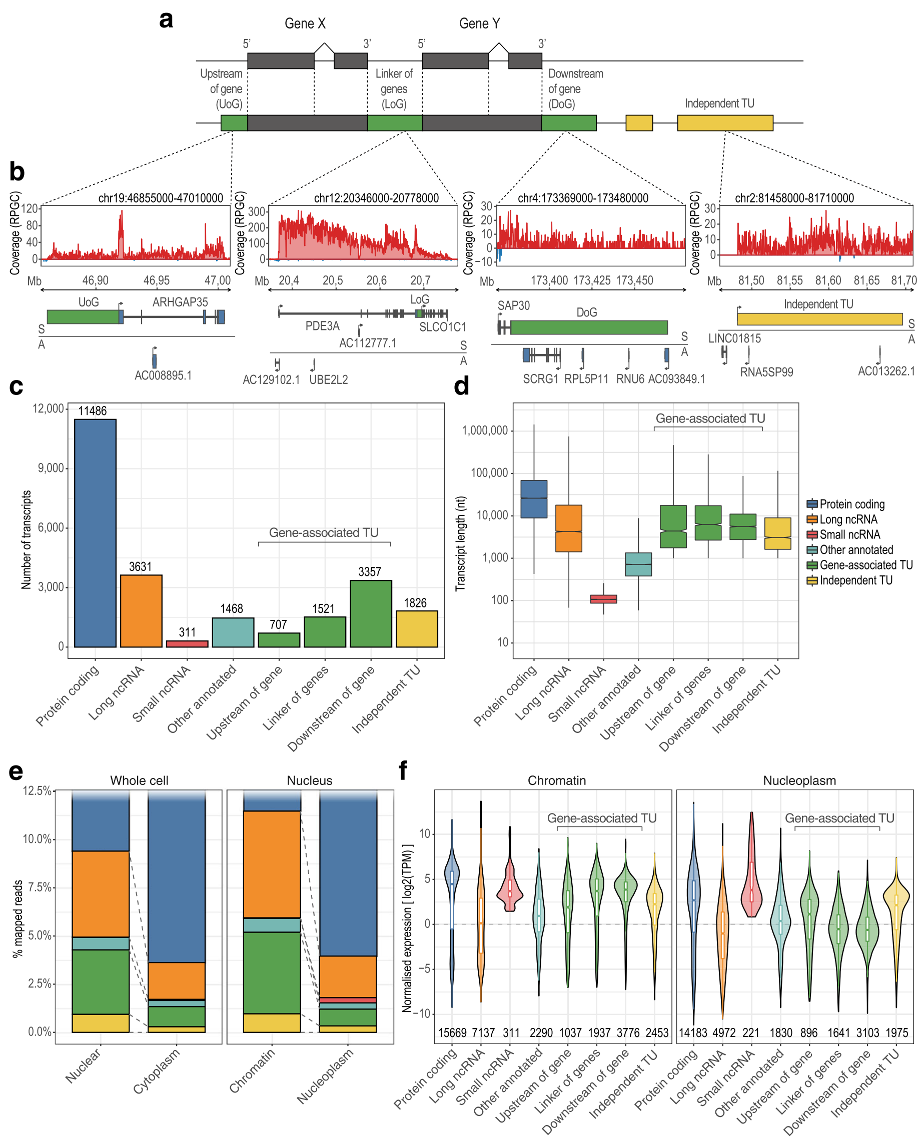

Fig. 2 General features of newly identified transcriptional units (TUs). **a** Schematic representation of the gene-associated (green) and independent (yellow) transcriptional units annotated in this study. **b** Upper panels: genome browser views of nuclear RNA-seq signals in HeLa cells, for example, TUs (red and blue indicate RNA-seq reads mapping to the sense and antisense strands, respectively). Lower panels: genomic annotations of pre-existing genes and newly identified TUs; horizontal line divides the features on the sense (S) and antisense (A) orientations. Coverage is reported at 1× depth (reads per genome coverage (RPGC)). **c** Comparison of the number of annotated and newly identified transcripts detected in the current RNA-seq dataset (TPM ≥ 1). **d** Comparison of transcript lengths. **e** Proportions of uniquely mapping RNA-seq reads originating from different transcript types for the whole cell (left) and nuclear (right) subcellular fractions of HeLa cells. **f** Distributions of expression levels of annotated and newly identified TUs for the chromatin-associated (left panel) and nucleoplasm (right panel) subcellular fractions of HeLa cells

- UoG, LoG, and DoG TUs must be associated with a protein-coding gene (Additional file 1: Figure S1C), thus reducing the chance of including poorly annotated genes with relatively unreliable start and end genomic coordinates (e.g., pseudogenes).

- Independent TUs must be ≥10 kb from any annotated feature on the same strand orientation to ensure that they are not transcribed as part of a known gene (Additional file 1: Figure S1A).

These filtering criteria left 1604 gene-associated TUs (88 UoG, 1329 DoG, 187 LoG) and 571 independent TUs. As controls, we paired gene-associated TUs with their corresponding protein-coding genes, and we identified 3462 lncRNA genes in a similar size range to independent TUs.

Gene-associated transcription breaks gene boundaries

Next, we sought to understand the transcriptional origins of gene-associated TUs. Here, we focus on data from HeLa cells unless stated otherwise, as it is the cell type with the largest variety of measurements.

Figure 3 displays meta-profiles of diverse transcriptional measurements aligned to the start and ends of TUs and protein-coding genes (Figure S2 for LoGs). All categories of TUs display clear RNA-seq coverage in the nuclear and chromatin-associated fractions, but in contrast to protein-coding genes, the signal is virtually lost in the cytoplasm (Figs. 3a and S2A). There is a clear jump in expression levels at the gene boundaries

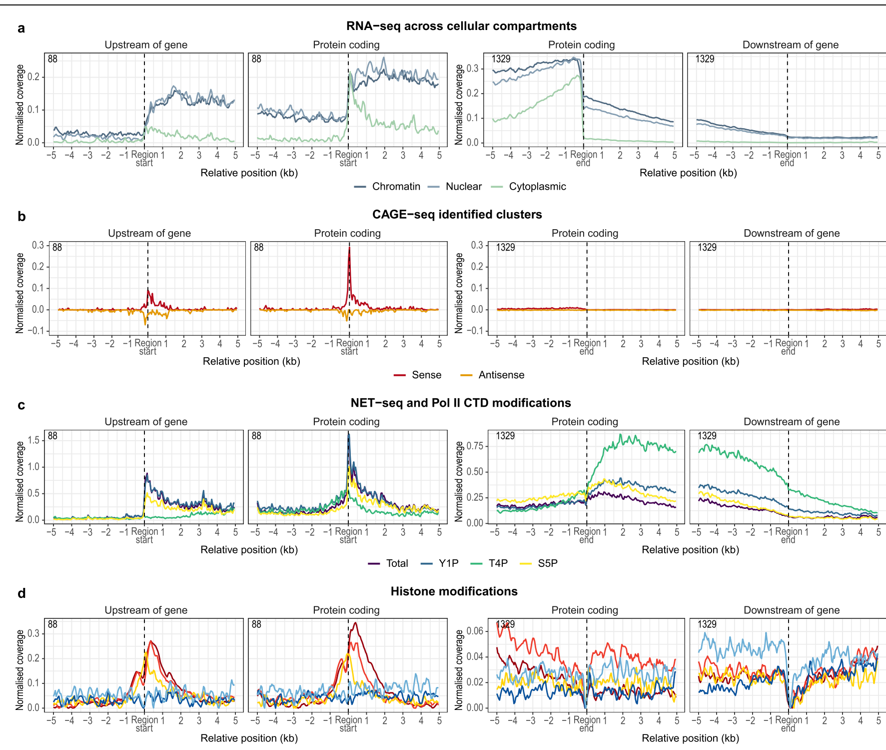

Fig. 3 Meta-profiles of transcriptional measurements around gene-associated TUs. Meta-profiles of transcriptional measurements plotted relative to the start positions of UoGs and their associated protein-coding genes (left-hand panels) and relative to the end positions of DoGs and their associated genes (right-hand panels). **a** RNA-seq measurements in different subcellular compartments. **b** CAGE-seq measurements in the sense and antisense strands. **c** NET-seq measurements for different Pol II CTD modifications. **d** ChIP-seq measurements for histone marks and EP300 occupancy-associated transcriptional activities

upon transition between the TU and associated gene, but TUs nonetheless display remarkably high relative expression levels in the nuclear and chromatin compartments.

Since the novel TUs are transcribed by RNA Pol II, we asked if the unannotated TSSs initiating UoG transcription have previously been detected through cap analysis of gene expression (CAGE). To this end, we used annotated CAGE peaks derived from a large collection of cell lines and tissues [40, 41]. UoGs display a slight enrichment of CAGE peaks at the start site, but we could hardly detect any signal for DoGs and LoGs (Figs. 3b and S2B); this suggests that whereas UoGs show evidence of independent transcriptional initiation, DoGs and LoGs are most likely generated from transcriptional readthrough of the upstream gene. The modest CAGE signal for UoGs (detected for 48 out of 98 UoGs) suggests that they are not efficiently capped, in contrast to mRNAs initiating at annotated start sites of protein-coding genes (Fig. 3b). This indicates that the majority of intergenic TUs might be designated as substrates for exonucleases and prone to degradation [42, 43].

Figure 3c (and Additional file 1: Figure S3A) shows prominent Pol II occupancies at the start sites of UoGs, albeit at lower levels than at the TSS of associated genes. Together with the CAGE data, this possibly indicates the formation of a pre-initiation complex (PIC) and therefore the existence of unannotated, upstream TSSs. Active transcription of TUs is supported by mammalian native elongating transcript sequencing (NET-seq) data, which identifies nascent RNA fragments attached to transcriptionally engaged RNA Pol II [39]. NET-seq is capable of differentiating between distinct transcriptional stages by mapping nascent RNAs associated with different patterns of RNA Pol II C-terminal heptad repeat domain (CTD) phosphorylation. The annotated and UoG TSSs display similar NET-seq profiles, thus suggesting that the TUs are not the result of stochastic Pol II binding but rather the outcome of coordinated transcriptional initiation events. Indeed, the profile for tyrosine-1 (Y1P) phosphorylated Pol II—a hallmark of TSS-paused protein-coding gene transcripts [17]—displays the highest signal at the start positions of both UoGs and protein-coding genes, with the former having a less pronounced peak and a broader distribution of the signal. Moreover, serine-5 (S5P) phosphorylated Pol II, which is mainly associated with TSS events such as co-transcriptional capping and early transcriptional elongation [44], follows a pattern similar to the total and Y1P profiles around these regions.

Threonine-4 (T4P) phosphorylation is a hallmark of terminating Pol II and causes a characteristic NET-seq signal near transcription end sites (TESs) of protein-coding genes [17]. Among protein-coding genes, the T4P profile peaks immediately after canonical TESs and remains high, while gradually decreasing towards the end of the associated DoG (Fig. 3c). This observation implies that although Pol II is poised to terminate after encountering the canonical TES, actual Pol II detachment might occur several kilobases downstream, as previously reported [45]. LoGs represent a special case, in which a high T4P signal after the TES of the upstream gene is maintained throughout the intergenic space only to peak again at the TSS of the downstream gene (Additional file 1: Figure S2C and S3B). This suggests either that transcription of LoGs joins two adjacent transcripts thereby generating pseudo-bicistronic nascent RNAs or, alternatively, that Pol II reaches the downstream gene and reinitiates transcription from a T4P state. In both cases, the downstream gene is potentially dependent on the transcription and by extension the promoter state of its upstream gene, thus implying the existence of co-regulation.

Finally, we examined the ChIP-seq profiles for four histone marks associated with transcriptional activity, as well as the histone acetyltransferase EP300 (Fig. 3d and Additional file 1: Figure S4A). Epigenetic modifications such as H3K4me3 and H3K27ac, which are associated with active promoters and enhancers respectively [46, 47], are enriched at both protein-coding and UoG start sites (Fig. 3d). Furthermore, the trimethylated forms of H3K27 and H3K9, commonly found at transcriptionally silenced regions [46, 47], are depleted. Interestingly, the histone acetyltransferase EP300, which regulates transcription of genes via chromatin remodeling, shows a comparable enrichment in binding at both UoG TSSs and annotated TSSs (Fig. 3d). EP300 is also known as a transcriptional coactivator, due to its ability to bind to transcription factors and the transcription machinery, and consequently activates transcription [46, 47]. Therefore, the presence of this protein more than 5 kb (size used for selecting the intergenic features) upstream of the canonical TSS is intriguing, as it suggests that transcription from the upstream intergenic regions is not merely the consequence of stochastic initiation events, but rather a concerted and precisely regulated process.

Transcription from deep intergenic regions

Next, we focused on independent TUs. We noticed a number of similarities between these elements and the 3426 control lncRNAs. Specifically, for both classes, we could detect RNA-seq signal upstream of the TSS and downstream of the TES (Fig. 4a). This is probably due to the sub-optimal annotation of these reference positions, a challenging task considering the intrinsically low level of expression of such transcripts [48, 49]. Interestingly, the CAGE signal displays equal enrichment in both orientations around the TSS of lncRNAs, possibly indicating that most of these RNAs originate from divergent transcription (Fig. 4b). The NET-seq profiles show similar enrichment patterns for total RNA Pol II and the CTD modifications TSSs and TESs of lncRNAs and independent TUs (Fig. 4c and Additional file 1: Figure S3C). Finally, the H3K9me3 and H3K27ac profiles around the TSSs of both lncRNAs and independent TUs resemble those of protein-coding genes and UoGs (Figs. 3d and 4d), highlighting equivalent chromatin statuses. Thus, based on the transcriptional and related measurements, independent TUs appear to be bona fide lincRNAs that eluded reference annotation.

Rapid degradation of chromatin-associated intergenic RNAs

We showed that both gene-associated and independent TUs are widespread across the genome, and their expression levels in the nucleus are comparable to those of annotated genes. Moreover, analysis of the transcribed loci did not highlight distinctive characteristics that explain why TUs are found only in the chromatin cellular compartment. Therefore, we hypothesized that there may be differences in the control of retention and stability of these transcripts.

First, we compared the expression levels of annotated RNAs and intergenic TUs between chromatin-associated and nucleoplasmic fractions. We found that unspliced protein-coding and long ncRNA transcripts tend to be equally distributed between the

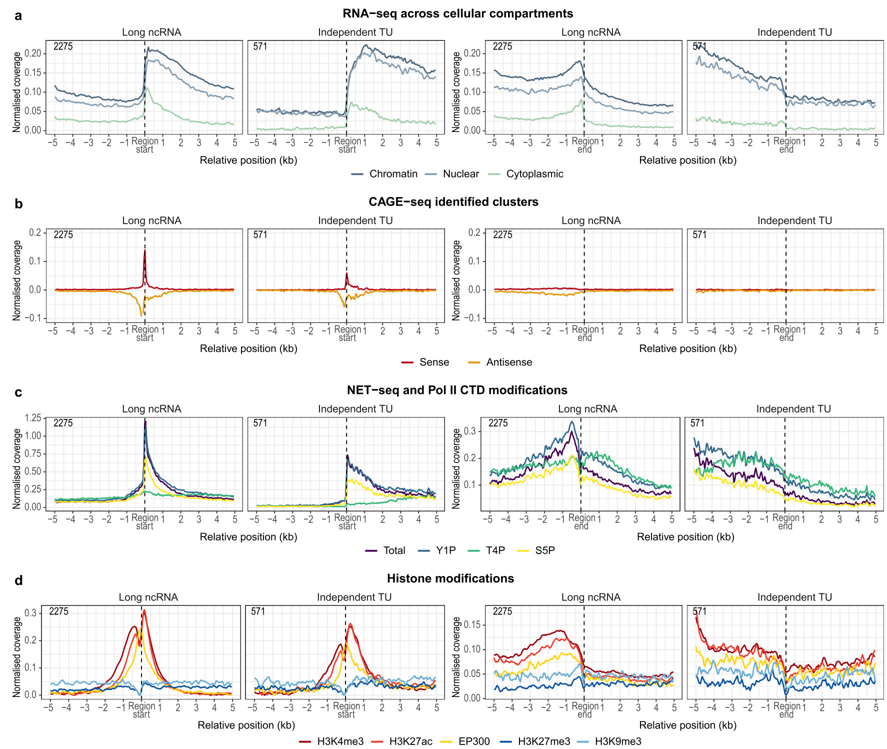

Fig. 4 Meta-profiles of transcriptional measurements around independent TUs. Meta-profiles of transcriptional measurements plotted relative to the start and end positions of independent TUs and control long non-coding RNA genes. **a**–**d** as in Fig. 3

two fractions, whereas TUs, in particular, DoGs and LoGs, are preferentially confined to the chromatin-associated fraction (Fig. 5a and Additional file 1: Figure S5A).

The scarcity of these transcripts in the nucleoplasm suggests that they exert their function, if any, bound to the chromatin fraction or that they are transcriptional by-products that are rapidly degraded. We examined recently published RNA-seq data following knockdown or depletion of proteins involved in the processing and degradation of transcriptional products: specifically, EXOSC3 [17], CSTF2 (and its paralog CSTF2T), CPSF3 (also known as CPSF73) knockdowns in HeLa cells [39], and XRN2 depletion in HCT116 cells [50].

EXOSC3 is part of the RNA exosome complex; it possesses 3′ to 5′ exoribonuclease activity, and it is involved in eliminating transcriptional by-products. Known substrates include non-coding transcripts, such as promoter-upstream transcripts (PROMPTs), mRNAs with processing defects [51, 52], and most prominently rRNA and snoRNAs, as part of their normal processing and maturation in the nucleolus [53]. The EXOSC3 knockdown had little or no effect on transcripts of protein-coding genes and their associated DoGs and LoGs; however, there is a marked effect on the stability of lncRNAs, UoGs, and independent TUs in the nucleoplasmic fraction (Fig. 5b and Additional file 1: Figure S5B). Moreover, the accumulation of these transcriptional products, caused by the loss of a functional nuclear RNA exosome, is more dramatic in the nucleoplasm than in the chromatin fraction, suggesting that they are generally targeted post-transcriptionally and cleared once they move away from the chromatin environment.

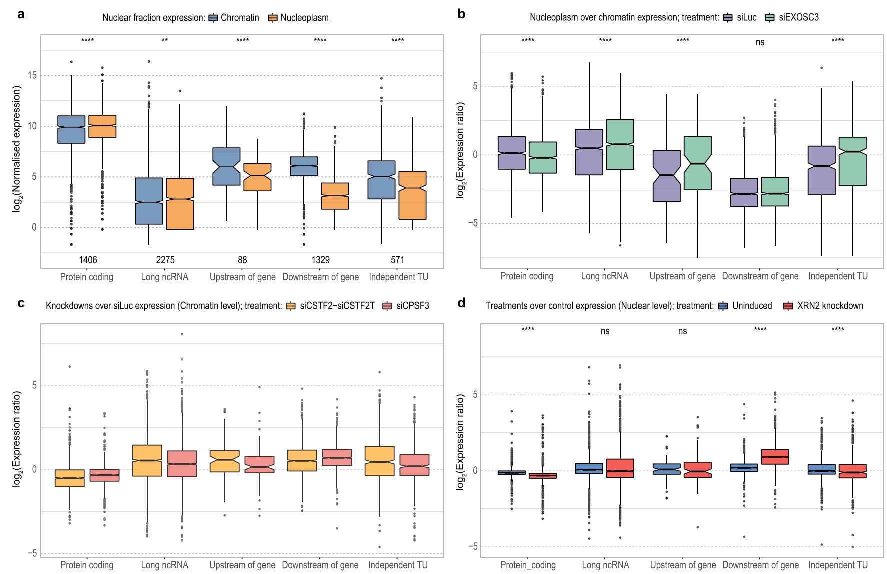

Fig. 5 Impact of nuclease depletion on TU expression. **a** Expression levels of protein-coding genes and TUs in the chromatin and nucleoplasm fractions. **b** Relative nucleoplasmic-to-chromatin expression levels in response to EXOSC3 knockdown and control siLuc treatments. **c** Expression levels in CSTF2+CSTF2T and CPSF3 knockdowns relative to control in the chromatin fraction. **d** Expression levels in XRN2 knockdown (via activation of the auxin-inducible degron system) and basal (uninduced; minus auxin) treatments relative to unmodified XRN2 control in the nuclear fraction. *p* values were calculated using the two-sided Wilcoxon rank sum test, with asterisks indicating statistical significance at the following thresholds: $^{ns}p > 0.05$; $^{*}p \leq 0.05$; $^{**}p \leq 0.01$; $^{***}p \leq 0.001$; $^{****}p \leq 0.0001$

Since we observed a predominant chromatin retention (Fig. 5a) and no effect of EXOSC3 knockdown on DoGs and LoGs (Fig. 5b and Additional file 1: Figure S5B), we hypothesized that other mechanisms must regulate these TUs. We examined the factors involved in processing the terminal regions of nascent transcripts: CSTF2 (and its paralog CSTF2T), implicated in 3′ end cleavage and polyadenylation of pre-mRNAs; CPSF3 (also known as CPSF73), a 3′ end-processing endonuclease; and XRN2, an exoribonuclease with 5′ to 3′ activity. Indeed, knockdowns of CPSF3 and of CSTF2+CSTF2T lead to increased levels of DoGs and LoGs (Fig. 5c and Additional file 1: Figure S5C), suggesting that degradation of these transcripts is strongly dependent on the correct processing of the 3′ end of nascent transcripts. Downstream of cleavage at the poly(A) signal by the CPSF/CSTF complex, the remaining 3′ by-product is depleted by the 3′→5′ exonuclease XRN2 [54]. Hence, we evaluated the expression of these transcripts in XRN2 depletion [50] to assess whether LoGs, like DoGs, are coupled with 3′ end processing of the upstream gene. XRN2 depletion greatly increased the expression of DoGs and LoGs, while leaving other transcript types unchanged (Fig. 5d and Additional file 1: Figure S5D), thus indicating that XRN2 activity indeed regulates DoG and LoG abundance.

Discussion

In 2010, van Bakel and colleagues explored transcribed unannotated transcripts by using tiling arrays and poly(A)$^{+}$ RNA-seq technologies. Despite their study proving valuable to better characterize intergenic fragments associated with transcription of annotated genes, the methods used limited their detection potential. Indeed, the authors were able to identify a large number of novel alternative exons, although the functional implications of the intergenic transcripts they defined remained elusive [55]. Here, we take advantage of the large amounts of nuclear and chromatin subcellular fraction RNA-seq and NET-seq data, which allow a deeper and more comprehensive detection of short-lived long RNAs, and investigate their stability and degradation mechanisms. As a result, we were able to expand the current catalog of intergenic RNAs that are either associated with the transcription of annotated genes or produced as independent transcriptional units.

Non-canonical transcription upstream of genes

To date, transcription upstream of canonical genes has been reported as a consequence of bidirectional transcription from neighboring promoters or enhancers, with the transcript being generated in the antisense direction. In contrast to these transcripts, the UoGs identified here originate from the same strand as the associated downstream genes, thus limiting the possibility that these are products of enhancer- or promoter-derived divergent transcription. The presence of CAGE peaks on opposite strands around the beginning of these transcripts suggests that a minor fraction could instead originate from convergent transcription [56]. Either way, transcription close and across the canonical promoter region of the respective gene is expected to result in the regulatory impact of UoG units, such as altering chromatin accessibility, or recruitment of Pol II and co-factors. Our study highlights examples in widespread used cell lines that can be studied in depth leveraging further genome-wide transcription data (such as Hi-C) or mechanistic analysis through genome editing.

Non-canonical transcription downstream of genes

Studies have recently highlighted the presence of widespread transcription of intergenic regions downstream of protein-coding genes in mice and humans in response to heat shock, osmotic stress, or oxidative stress [30, 31]. Although this form of transcriptional readthrough has been ascribed to the mammalian stress response, here, we found evidence for such behavior in unstimulated, normally proliferating cell lines. We observed two categories of readthrough, which are characterized by distinctive patterns of Pol II CTD phosphorylation. In the first group, DoGs arise from the transcription of canonical genes that then continues for a few to hundreds of kilobases across intergenic space (DoG), as previously reported [17, 30, 39]. These are marked by the sharp increase in threonine 4 phosphorylation of Pol II (T4P) after the annotated TES, and the gradual and eventual loss of Pol II binding with distance. In the second group, Pol II continues transcribing to the next gene (LoG), thus hinting at the possibility of polycistronic transcription in higher eukaryotes [57]. In this case, the T4P signal does not fade, suggesting that most Pol II continues transcribing until it reaches the downstream gene. It is not clear whether Pol II proceeds uninterrupted through the next gene or reinitiates a separate transcriptional event. Co-regulation of genes in close proximity on the same chromosome has previously been described [58], and the existence of LoGs could be one of the factors explaining such observations.

Functional consequences of non-canonical transcription on canonical genes

Although intergenic transcription has been commonly considered a consequence of pervasive transcription and, therefore, having no apparent functional role, accumulating evidence indicates that such processes can have major repercussions on the activities of neighboring genes [14]. Indeed, the effect of lncRNA transcription on gene activation or repression has been reported by a few studies [15, 59–61]. Interestingly, this phenomenon does not seem to be restricted to lncRNAs but also extends to protein-coding mRNAs and, potentially, to all transcriptional events [21]. Gene-associated RNAs can recruit chromatin remodelers that are able to maintain an open chromatin state or act as binding platforms for protein complexes at gene-proximal sites, such as the transcriptional factor Yin and Yang 1 (YY1) [62] and the MLL complex subunit WD repeat-containing 5 (WDR5) [63]. As a result, transcription upstream and downstream of annotated genes that we identified in this study might be functionally important for maintaining an open chromatin state and for the correct expression of neighboring genes. That these transcripts are tightly associated with chromatin and are rapidly degraded by nuclear surveillance processes, suggesting that their functions do not go beyond the course of transcription. For example, DoGs are highly sensitive to XRN2-mediated degradation; it has been previously reported that this protein promotes transcriptional termination at protein-coding genes via the torpedo mechanism model, in which the exonuclease degrades the gene-associated RNA until it reaches the elongation complex, consequently causing its termination [64–66]. Hence, transcription of very long DoGs might underlie a longer engagement of RNA Pol II and, consequently, its inability to readily detach from DNA and restart transcription elsewhere and its contribution to maintain chromatin in an open state.

These mechanisms of transcription-associated chromatin regulation are not necessarily confined to intergenic regions linked to previously annotated genes but might be broadened to independent TUs. However, since these regions are usually found in gene-poor portions of the genome, their products are more likely to exert their functional role in trans. This and their similarities in terms of transcriptional activity, chromatin state, and degradation patterns to lncRNAs support the hypothesis that independent TUs could be novel lncRNA loci.

Conclusions

In summary, we assembled publicly available RNA-seq data to identify and classify intergenic transcripts based on their expression and location relative to annotated genes. We showed that gene-associated and independent RNAs have characteristic patterns of transcription and that they are highly sensitive to nuclear degradation processes. Our data are consistent with recently reported chromatin remodeling and gene expression regulatory mechanisms associated with transcription. Collectively, the results expand the current categories in gene annotation and provide the tools to further investigate the underappreciated role of intergenic transcription as a function of gene expression and regulation.

Methods

Read alignment and post-processing

Sequencing quality checks were performed on all experiments using FastQC [67]. Adaptor sequences were removed using TrimGalore (v0.4.4_dev) [68] with default parameters. Reads were filtered against human rRNA and tRNA sequences obtained from the NCBI using Bowtie2 (v2.3.3.1) [69] with the option --sensitive-local. Reads that failed to align were mapped with STAR (v2.5.3a) [70] to UCSC hg38/GRCh38 genome assembly using GENCODE (v27) gene annotation [35] as a reference, with the following parameters: *--twopassMode Basic --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --sjdbScore 1 --outFilterMultimapNmax 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --outFilterType BySJout --outSAMattributes All --outSAMtype BAM SortedByCoordinate*, and specific options for gapped (*--alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000*) and ungapped (*--alignIntronMax 1 --alignMatesGapMax 300*) alignments. PCR duplicates were removed using Picard MarkDuplicates (v2.18.3) with default parameters. Quantification of expression was performed using QoRTs (v1.3.0) [71] and the GENCODE (v27) gene annotation [35].

Genomic coverage tracks

Deduplicated unique alignments were converted to stranded normalized coverage bigWig files using deeptools (v3.0.2) [72] with *--normalizeUsing CPM --binSize 20 --smoothLength 60* options, and *--filterRNAstrand* for the selection of forward and reverse strands.

De novo transcriptome assembly

Deduplicated uniquely mapped reads were assembled into a de novo annotation GTF using StringTie (v1.3.4c) [34] with the GENCODE (v27) gene annotation [35] as a reference, and the following parameters: *-f 0.2 -g 100 -j 3 -t*. The individual annotation GTFs from all wild-type RNA-seq datasets (no treatment condition was used to annotate the intergenic regions) were then used as input for StringTie with the *--merge* option to generate a non-redundant set of predicted transcripts. The output, which consists of a GTF file with merged gene models, was filtered using the gffcompare utility [73] with the *-C* option to discard predicted transcripts that were fully contained within larger annotated regions.

Identification of intergenic transcriptional units

A custom pipeline written in R was used to process the GTF file generated as described above. The pipeline performs several steps, the first of which is the discrimination of the purely intergenic regions (i.e., defined using the sequencing data) from the known features (i.e., already present in the GENCODE gene annotation). This operation is performed by the *setdiff()* function from the *GenomicFeatures* R package [74] on the gffcompare-generated GTF and GENCODE reference annotation files. Intergenic regions with length ≤ 1 kb are discarded. Although these can represent true signals, the low number of supporting reads observed in these small regions can be seen as a consequence of misannotation or misalignment. The remaining are further divided into "gene-associated" and "independent" transcriptional units ("gene-associated TUs" and "independent TUs," respectively) based on whether they originate from annotated genes, thus showing transcriptional continuity with the gene body, or from regions devoid of annotated features, and therefore, they are considered independent events of transcription. Gene-associated transcriptional units are assigned to different sub-groups depending on their position and connection to the neighboring gene(s):

- Upstream of gene (UoG): the unit is located upstream of the associated gene.
- Downstream of gene (DoG): the unit is located downstream of the associated gene.
- Linker of genes (LoG): the unit is located between two genes and transcriptionally associated with them.

To confirm the co-occurrence of the annotated gene(s) and gene-associated features, their expression is re-assessed across the RNA-seq datasets within R using a customized *countOverlapsUnion()* function. Features are considered co-transcribed if expressed (TPM ≥ 1) in the same cell line and in at least two datasets. At this level, the categorization is also re-evaluated and, if necessary, TUs can be re-assigned to the proper sub-group (e.g., a linker of gene TU whose downstream gene is not expressed will become a downstream of gene TU).

Independent TU comparison with public databases

Expressed independent TUs were used as input for BEDtools *intersect* to calculate the individual overlaps with annotated features retrieved from NONCODE [36], CHESS [37], and RefSeq [38] databases, using the options *-s* (stranded) and *-f* (to test the fractions of overlaps from 0.1 to 1). The number of independent TUs not annotated in any of the databases was obtained by performing sequential intersections (adding the options *-wa -v*).

Splicing analysis

Deduplicated unique alignments were parsed using samtools [75] *view*, and gapped alignments (i.e., reads encompassing known or putative splice junctions) were extracted based on their CIGAR information (i.e., whether or not it contained “N”). Reads were then assigned to “long ncRNA” or “independent TU” features using the *countOverlapsUnion()* function from the *GenomicFeatures* R package [74]. For each dataset, the fraction of junction reads was calculated over the total number of deduplicated unique reads.

Selection of HeLa TUs and metadata profiles

Since the large majority of data available for validation derived from HeLa cells, we decided to focus our analysis of intergenic features only to those expressed in this cell line. Therefore, we generated a set of annotated genes and gene-associated and independent TUs where each feature had an average expression of ≥1 TPM across the HeLa RNA-seq datasets. In addition, we required the gene-associated features to be connected to annotated protein-coding genes, thus reducing the chance to include poorly annotated genes for which start and end genomic coordinates are not reliable (e.g., pseudogenes). We retained only the independent TUs located ≥10 kb from any annotated feature on the same strand orientation, to ensure that their transcription is not directly linked to known genes. Finally, we discarded features with length < 5 kb to avoid signal overlaps between the start and end positions in metadata profiles.

The metadata profiles were generated using the CPM-normalized coverage bigWig files (see "Genomic coverage tracks" section) and a custom wrapper of the *ScoreMatrixBin()* function from the *genomation* R package [76]. The wrapper function is used to facilitate strand splitting, centering, and resizing (i.e., ±5 kb from region start or end position), binning (i.e., 200 bins over the 10-kb window), and normalization and averaging of the signal. When not specified in the figure legend, normalization was performed by dividing the bins of each feature (or group of features in case of paired annotated gene and its gene-associated TU) by the value of the bin with the higher count across the region.

Epigenetic modification profiles

We collected the "fold change over control" and merged replicates ChIP-seq bigWig files from ENCODE. The list of epigenetic modifications and associated accession numbers can be found in Additional file 1: Table S3. The ChIP-seq signals across the regions of interest were calculated using the wrapper function described in the previous section.

CAGE peak profiles

We retrieved the hg38 CAGE reprocessed data [40] from the FANTOM Consortium [77]. The density of the CAGE peaks (phases 1 and 2) was calculated using the wrapper function described in the "Selection of HeLa TUs and metadata profiles" section, without applying any normalization.

Quantification of expression and degradation

We collected data from cells in wild-type/untreated conditions and after the knockdown of several proteins from different sources (Additional file 1: Table S2). The datasets were processed as described in the "Read alignment and post-processing" section. Deduplicated uniquely mapped reads were loaded into R using the *GenomicAlignments* R package [74], and the expression of the features quantified with the *countOverlapsUnion()* function. The *estimateSizeFactorForMatrix* function from the DESeq2 R package [78] was used to normalize the counts for each group of experiments. The *ggpubr* R package was used to visualize the results and perform the statistical tests (i.e., two-sided Wilcoxon rank sum test).

Supplementary Information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-021-02350-x.

Additional file 1. Integrated supplementary Figures and Tables. Contains figures from S1 to S5 and tables from S1 to S4.

Additional file 2. Review history.

Acknowledgements

The authors are grateful to C. Pederiva and A. M. Chakrabarti for the valuable assistance and C. Pederiva for the comments on this manuscript and for the valuable advice.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 2.

Authors' contributions

F.A., J.U., and N.M.L. conceived the original idea and managed the project. F.A. developed the bioinformatic pipeline. F.A. and J.U. planned the computational analyses. F.A. performed the computational analyses. J.Z., J.A., J.U., and N.M.L. provided valuable feedback on computational analyses and provided interpretation of the results. All the authors provided critical feedback and helped to shape the research, analysis, and manuscript. F.A. and J.Z. wrote the manuscript with valuable feedback from all authors. The authors read and approved the final manuscript.

Authors' information

F.A. is currently employed as a researcher at the Science for Life Laboratory, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden.

Funding

This work was supported by the European Research Council (617837-Translate to J.U.) and the Wellcome Trust with a Joint Investigator Award (103760/Z/14/Z to J.U. and N.M.L.). N.M.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the Francis Crick Institute and is additionally funded by the MRC eMedLab Medical Bioinformatics Infrastructure Award (MR/L016311/1) and core funding from the Okinawa Institute of Science & Technology Graduate University. The Francis Crick Institute receives its core funding from Cancer Research UK (FC001110), the UK Medical Research Council (FC001110), and the Wellcome Trust (FC001110) (to N.M.L., J.A., and F.A.).

Availability of data and materials

The datasets analyzed in the current study are available on GEO and ArrayExpress with the following accession numbers: GSE66478 [79], GSE81662 [17], GSE90238 [80], GSE90256 [80], GSE90249 [80], GSE90230 [80], GSE90220 [80], E-MTAB-6204 [81], GSE90248 [80], GSE90250 [80], GSE90228 [80], GSE90236 [80], GSE109003 [50], GSE39878 [82], GSE60358 [39], and GSE29611 [80]. These accession numbers are also available in Additional file 1: Tables S1, S2 and S3. The scripts used to perform the identification of the transcriptional units are available under GPL3 open-source license at https://github.com/luslab/IntergenicTranscription. The version used in this article is available as a Zenodo archive with DOI https://doi.org/10.5281/zenodo.4644300 [83].

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK. [2]Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, Queen Square, London WC1N 3BG, UK. [3]UCL Genetics Institute, Department of Genetics, Environment and Evolution, University College London, Gower Street, London WC1E 6BT, UK. [4]Okinawa Institute of Science & Technology Graduate University, 1919-1 Tancha, Onna-son, Kunigami-gun, Okinawa 904-0495, Japan.

Received: 7 January 2020 Accepted: 12 April 2021
Published online: 05 May 2021

References

1. Jacquier A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet. 2009;10(12):833–44. https://doi.org/10.1038/nrg2683.
2. Hangauer MJ, Vaughn IW, McManus MT. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. 2013;9(6):e1003569. https://doi.org/10.1371/journal.pgen.1003569.
3. Jensen TH, Jacquier A, Libri D. Dealing with pervasive transcription. Mol Cell. 2013;52(4):473–84. https://doi.org/10.1016/j.molcel.2013.10.032.
4. Porrua O, Libri D. Transcription termination and the control of the transcriptome: why, where and how to stop. Nat Rev Mol Cell Biol. 2015;16(3):190–202. https://doi.org/10.1038/nrm3943.
5. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH. RNA exosome depletion reveals transcription upstream of active human promoters. Science. 2008;322(5909):1851–4. https://doi.org/10.1126/science.1164096.
6. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465(7295):182–7. https://doi.org/10.1038/nature09033.
7. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. 2010;8(5):e1000384. https://doi.org/10.1371/journal.pbio.1000384.
8. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES. Chromatin signature reveals

over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458(7235):223–7. https://doi.org/10.1038/nature07672.

9. Abascal F, Juan D, Jungreis I, Kellis M, Martinez L, Rigau M, Rodriguez JM, Vazquez J, Tress ML. Loose ends: almost one in five human genes still have unresolved coding status. Nucleic Acids Res. 2018;46(14):7070–84. https://doi.org/10.1093/nar/gky587.
10. Kopp F, Mendell JT. Functional classification and experimental dissection of long noncoding RNAs. Cell. 2018;172(3):393–407. https://doi.org/10.1016/j.cell.2018.01.011.
11. Ransohoff JD, Wei Y, Khavari PA. The functions and unique features of long intergenic non-coding RNA. Nat Rev Mol Cell Biol. 2018;19(3):143–57. https://doi.org/10.1038/nrm.2017.104.
12. 3 Characterization of intergenic regions and gene definition. Nature. 2019; Available from: https://doi.org/10.1038/nature28172.
13. Sahakyan A, Yang Y, Plath K. The role of Xist in X-chromosome dosage compensation. Trends Cell Biol. 2018;28(12):999–1013. https://doi.org/10.1016/j.tcb.2018.05.005.
14. Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, McDonel PE, Guttman M, Lander ES. Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature. 2016;539(7629):452–5. https://doi.org/10.1038/nature20149.
15. Joung J, Engreitz JM, Konermann S, Abudayyeh OO, Verdine VK, Aguet F, Gootenberg JS, Sanjana NE, Wright JB, Fulco CP, Tseng YY, Yoon CH, Boehm JS, Lander ES, Zhang F. Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. Nature. 2017;548(7667):343–6. https://doi.org/10.1038/nature23451.
16. Ard R, Allshire RC, Marquardt S. Emerging properties and functional consequences of noncoding transcription. Genetics. 2017;207(2):357–67. https://doi.org/10.1534/genetics.117.300095.
17. Schlackow M, Nojima T, Gomes T, Dhir A, Carmo-Fonseca M, Proudfoot NJ. Distinctive patterns of transcription and RNA processing for human lincRNAs. Mol Cell. 2017;65(1):25–38. https://doi.org/10.1016/j.molcel.2016.11.029.
18. Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. Cell. 2013;154(1):26–46. https://doi.org/10.1016/j.cell.2013.06.020.
19. Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature. 2013;499(7458):360–3. https://doi.org/10.1038/nature12349.
20. Ntini E, Järvelin AI, Bornholdt J, Chen Y, Boyd M, Jørgensen M, Andersson R, Hoof I, Schein A, Andersen PR, Andersen PK, Preker P, Valen E, Zhao X, Pelechano V, Steinmetz LM, Sandelin A, Jensen TH. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat Struct Mol Biol. 2013;20(8):923–8. https://doi.org/10.1038/nsmb.2640.
21. Skalska L, Beltran-Nebot M, Ule J, Jenner RG. Regulatory feedback from nascent RNA to chromatin and transcription. Nat Rev Mol Cell Biol. 2017;18(5):331–7. https://doi.org/10.1038/nrm.2017.12.
22. Tsai M-C, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. Long noncoding RNA as modular scaffold of histone modification complexes. Science. 2010;329(5992):689–93. https://doi.org/10.1126/science.1192002.
23. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 2007;129(7):1311–23. https://doi.org/10.1016/j.cell.2007.05.022.
24. Zhao J, Sun BK, Erwin JA, Song J-J, Lee JT. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science. 2008;322(5902):750–6. https://doi.org/10.1126/science.1163045.
25. Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, Nagano T, Mancini-DiNardo D, Kanduri C. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell. 2008;32(2):232–46. https://doi.org/10.1016/j.molcel.2008.08.022.
26. Davidovich C, Zheng L, Goodrich KJ, Cech TR. Promiscuous RNA binding by Polycomb repressive complex 2. Nat Struct Mol Biol. 2013;20(11):1250–7. https://doi.org/10.1038/nsmb.2679.
27. Kaneko S, Son J, Shen SS, Reinberg D, Bonasio R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. Nat Struct Mol Biol. 2013;20(11):1258–64. https://doi.org/10.1038/nsmb.2700.
28. Kaneko S, Son J, Bonasio R, Shen SS, Reinberg D. Nascent RNA interaction keeps PRC2 activity poised and in check. Genes Dev. 2014;28(18):1983–8. https://doi.org/10.1101/gad.247940.114.
29. Beltran M, Yates CM, Skalska L, Dawson M, Reis FP, Viiri K, Fisher CL, Sibley CR, Foster BM, Bartke T, Ule J, Jenner RG. The interaction of PRC2 with RNA or chromatin is mutually antagonistic. Genome Res. 2016;26(7):896–907. https://doi.org/10.1101/gr.197632.115.
30. Vilborg A, Passarelli MC, Yario TA, Tycowski KT, Steitz JA. Widespread inducible transcription downstream of human genes. Mol Cell. 2015;59(3):449–61. https://doi.org/10.1016/j.molcel.2015.06.016.
31. Vilborg A, Sabath N, Wiesel Y, Nathans J, Levy-Adam F, Yario TA, Steitz JA, Shalgi R. Comparative analysis reveals genomic features of stress-induced transcriptional readthrough. Proc Natl Acad Sci U S A. 2017;114(40):E8362–71. https://doi.org/10.1073/pnas.1711120114.
32. Muniz L, Deb MK, Aguirrebengoa M, Lazorthes S, Trouche D, Nicolas E. Control of gene expression in senescence through transcriptional read-through of convergent protein-coding genes. Cell Rep. 2017;21(9):2433–46. https://doi.org/10.1016/j.celrep.2017.11.006.
33. Attig J, Agostini F, Gooding C, Chakrabarti AM, Singh A, Haberman N, et al. Heteromeric RNP assembly at LINEs controls lineage-specific RNA processing. Cell. 2018;174:1067–81.e17.
34. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5. https://doi.org/10.1038/nbt.3122.
35. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ. GENCODE: the reference human genome annotation for the ENCODE Project. Genome Res. 2012;22(9):1760–74. https://doi.org/10.1101/gr.135350.111.

36. Fang S, Zhang L, Guo J, Niu Y, Wu Y, Li H, Zhao LH, Li XY, Teng XY, Sun XH, Sun L, Zhang MQ, Chen RS, Zhao Y. NONCODEV5: a comprehensive annotation database for long non-coding RNAs. Nucleic Acids Res. 2018;46(D1):D308–14. https://doi.org/10.1093/nar/gkx1107.
37. Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang Y-C, Madugundu AK, Pandey A, Salzberg SL. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. 2018;19(1):208. https://doi.org/10.1186/s13059-018-1590-2.
38. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45. https://doi.org/10.1093/nar/gkv1189.
39. Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ. Mammalian NET-Seq reveals genome-wide nascent transcription coupled to RNA processing. Cell. 2015;161(3):526–40. https://doi.org/10.1016/j.cell.2015.03.027.
40. Abugessaisa I, Noguchi S, Hasegawa A, Harshbarger J, Kondo A, Lizio M, Severin J, Carninci P, Kawaji H, Kasukawa T. FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies. Sci Data. 2017;4(1):170107. https://doi.org/10.1038/sdata.2017.107.
41. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455–61. https://doi.org/10.1038/nature12787.
42. Jurado AR, Tan D, Jiao X, Kiledjian M, Tong L. Structure and function of pre-mRNA 5′-end capping quality control and 3′-end processing. Biochemistry. 2014;53(12):1882–98. https://doi.org/10.1021/bi401715v.
43. Ramanathan A, Robb GB, Chan S-H. mRNA capping: biological functions and applications. Nucleic Acids Res. 2016;44(16):7511–26. https://doi.org/10.1093/nar/gkw551.
44. Hsin J-P, Manley JL. The RNA polymerase II CTD coordinates transcription and RNA processing. Genes Dev. 2012;26(19):2119–37. https://doi.org/10.1101/gad.200303.112.
45. Lian Z, Karpikov A, Lian J, Mahajan MC, Hartman S, Gerstein M, Snyder M, Weissman SM. A genomic analysis of RNA polymerase II modification and chromatin architecture related to 3′ end RNA polyadenylation. Genome Res. 2008;18(8):1224–37. https://doi.org/10.1101/gr.075804.107.
46. Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet. 2011;12(1):7–18. https://doi.org/10.1038/nrg2905.
47. Gates LA, Foulds CE, O'Malley BW. Histone marks in the "driver's seat": functional roles in steering the transcription cycle. Trends Biochem Sci. 2017;42(12):977–89. https://doi.org/10.1016/j.tibs.2017.10.004.
48. Salviano-Silva A, Lobo-Alves SC, de Almeida RC, Malheiros D, Petzl-Erler ML. Besides pathology: long non-coding RNA in cell and tissue homeostasis. Noncoding RNA. 2018;4 Available from: https://doi.org/10.3390/ncrna4010003
49. Tuck AC, Natarajan KN, Rice GM, Borawski J, Mohn F, Rankova A, Flemr M, Wenger A, Nutiu R, Teichmann S, Bühler M. Distinctive features of lincRNA gene expression suggest widespread RNA-independent functions. Life Sci Alliance. 2018;1(4):e201800124. https://doi.org/10.26508/lsa.201800124.
50. Eaton JD, Davidson L, Bauer DLV, Natsume T, Kanemaki MT, West S. Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity. Genes Dev. 2018;32(2):127–39. https://doi.org/10.1101/gad.308528.117.
51. Pefanis E, Wang J, Rothschild G, Lim J, Kazadi D, Sun J, Federation A, Chao J, Elliott O, Liu ZP, Economides AN, Bradner JE, Rabadan R, Basu U. RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. Cell. 2015;161(4):774–89. https://doi.org/10.1016/j.cell.2015.04.034.
52. Rialdi A, Hultquist J, Jimenez-Morales D, Peralta Z, Campisi L, Fenouil R, et al. The RNA exosome syncs IAV-RNAPII transcription to promote viral ribogenesis and infectivity. Cell. 2017;169:679–92.e14.
53. Iarovaia OV, Minina EP, Sheval EV, Onichtchouk D, Dokudovskaya S, Razin SV, Vassetzky YS. Nucleolus: a central hub for nuclear functions. Trends Cell Biol. 2019;29(8):647–59. https://doi.org/10.1016/j.tcb.2019.04.003.
54. Kaneko S, Rozenblatt-Rosen O, Meyerson M, Manley JL. The multifunctional protein p54nrb/PSF recruits the exonuclease XRN2 to facilitate pre-mRNA 3′ processing and transcription termination. Genes Dev. 2007;21(14):1779–89. https://doi.org/10.1101/gad.1565207.
55. van Bakel H, Nislow C, Blencowe BJ, Hughes TR. Most "dark matter" transcripts are associated with known genes. PLoS Biol. 2010;8(5):e1000371. https://doi.org/10.1371/journal.pbio.1000371.
56. Meng F-L, Du Z, Federation A, Hu J, Wang Q, Kieffer-Kwon K-R, et al. Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. Cell. 2014;159(7):1538–48. https://doi.org/10.1016/j.cell.2014.11.014.
57. Miki TS, Carl SH, Stadler MB, Großhans H. XRN2 autoregulation and control of polycistronic gene expression in Caenorhabditis elegans. PLoS Genet. 2016;12(9):e1006313. https://doi.org/10.1371/journal.pgen.1006313.
58. Kustatscher G, Grabowski P, Rappsilber J. Pervasive coexpression of spatially proximal genes is buffered at the protein level. Mol Syst Biol. 2017;13(8):937. https://doi.org/10.15252/msb.20177548.
59. Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, Wysocka J, Lei M, Dekker J, Helms JA, Chang HY. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature. 2011;472(7341):120–4. https://doi.org/10.1038/nature09819.
60. Anderson KM, Anderson DM, McAnally JR, Shelton JM, Bassel-Duby R, Olson EN. Transcription of the non-coding RNA upperhand controls Hand2 expression and heart development. Nature. 2016;539(7629):433–6. https://doi.org/10.1038/nature20128.
61. Paralkar VR, Taborda CC, Huang P, Yao Y, Kossenkov AV, Prasad R, Luan J, Davies JOJ, Hughes JR, Hardison RC, Blobel GA, Weiss MJ. Unlinking an lncRNA from its associated cis element. Mol Cell. 2016;62(1):104–10. https://doi.org/10.1016/j.molcel.2016.02.029.
62. Sigova AA, Abraham BJ, Ji X, Molinie B, Hannett NM, Guo YE, Jangi M, Giallourakis CC, Sharp PA, Young RA. Transcription factor trapping by RNA in gene regulatory elements. Science. 2015;350(6263):978–81. https://doi.org/10.1126/science.aad3346.
63. Hendrickson DG, Kelley DR, Tenen D, Bernstein B, Rinn JL, et al. Genome Biol. 2016;17:28.
64. Connelly S, Manley JL. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. Genes Dev. 1988;2(4):440–52. https://doi.org/10.1101/gad.2.4.440.

65. West S, Gromak N, Proudfoot NJ. Human 5′ --> 3′ exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature. 2004;432(7016):522–5. https://doi.org/10.1038/nature03035.
66. Fong N, Brannan K, Erickson B, Kim H, Cortazar MA, Sheridan RM, Nguyen T, Karp S, Bentley DL. Effects of transcription elongation rate and Xrn2 exonuclease activity on RNA polymerase II termination suggest widespread kinetic competition. Mol Cell. 2015;60(2):256–67. https://doi.org/10.1016/j.molcel.2015.09.026.
67. Andrews S. FastQC: a quality control tool for high throughput sequence data [online] [internet]. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
68. Krueger F. TrimGalore [Internet]. GitHub Repository. 2012. Available from: https://github.com/FelixKrueger/TrimGalore
69. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923.
70. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.
71. Hartley SW, Mullikin JC. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. BMC Bioinformatics. 2015;16(1):224. https://doi.org/10.1186/s12859-015-0670-5.
72. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42(W1):W187–91. https://doi.org/10.1093/nar/gku365.
73. Pertea G, Pertea M. GFF utilities: GffRead and GffCompare. F1000Res. 2020;9 Available from: https://doi.org/10.12688/f1000research.23297.2
74. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118. https://doi.org/10.1371/journal.pcbi.1003118.
75. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.
76. Akalin A, Franke V, Vlahoviček K, Mason CE, Schübeler D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. Bioinformatics. 2015;31(7):1127–9. https://doi.org/10.1093/bioinformatics/btu775.
77. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, et al. A promoter-level mammalian expression atlas. Nature. 2014;507:462–70.
78. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013;8(9):1765–86. https://doi.org/10.1038/nprot.2013.099.
79. Werner MS, Ruthenburg AJ. Nuclear fractionation reveals thousands of chromatin-tethered noncoding RNAs adjacent to active genes. Cell Rep. 2015;12(7):1089–98. https://doi.org/10.1016/j.celrep.2015.07.033.
80. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57–74. https://doi.org/10.1038/nature11247.
81. Coelho MB, Attig J, Bellora N, König J, Hallegger M, Kayikci M, Eyras E, Ule J, Smith CWJ. Nuclear matrix protein Matrin3 regulates alternative splicing and forms overlapping regulatory networks with PTB. EMBO J. 2015;34(5):653–68. https://doi.org/10.15252/embj.201489852.
82. Wang IX, Core LJ, Kwak H, Brady L, Bruzel A, McDaniel L, Richards AL, Wu M, Grunseich C, Lis JT, Cheung VG. RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase II. Cell Rep. 2014;6(5):906–15. https://doi.org/10.1016/j.celrep.2014.01.037.
83. Agostini F. IntergenicTranscription: code release [GitHub]. 2021. Available from: https://zenodo.org/record/4662579

Publisher's Note