Published in final edited form as:
Cell Rep. 2014 March 13; 6(5): 906–915. doi:10.1016/j.celrep.2014.01.037.

RNA-DNA Differences Are Generated in Human Cells within Seconds After RNA Exits Pol II

Isabel X. Wang[1,*], **Leighton J. Core**[2,*], **Hojoong Kwak**[2,4], **Lauren Brady**[3], **Alan Bruzel**[1,4], **Lee McDaniel**[5], **Allison L. Richards**[6], **Ming Wu**[4], **Christopher Grunseich**[7], **John T. Lis**[2,†], and **Vivian G. Cheung**[1,4,8,†]

[1]Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, USA

[2]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA

[3]Cell and Molecular Biology Graduate Program, University of Pennsylvania, Philadelphia, PA 19104, USA

[4]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

[5]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA

[6]Human Genetics Graduate Program, University of Michigan, Ann Arbor, MI 48109, USA

[7]Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA

[8]Departments of Pediatrics and Genetics, University of Michigan, Ann Arbor, MI 48109, USA

Summary

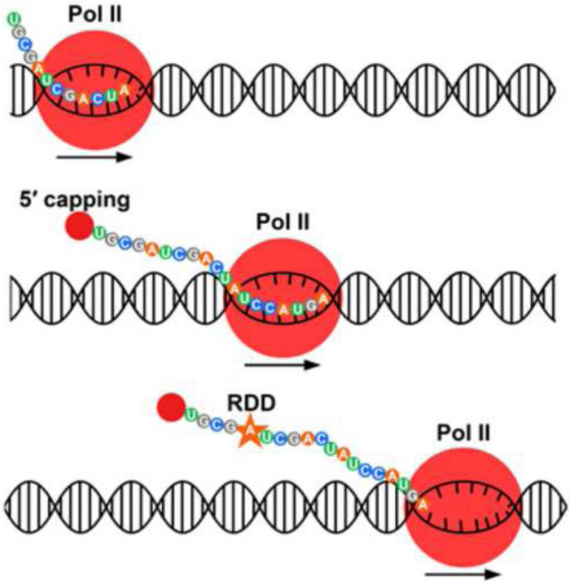
RNA sequences are expected to be identical to their corresponding DNA sequences. Here, we found all 12 types of RNA-DNA sequence differences (RDDs) in nascent RNA. Our results show that RDDs begin to occur in RNA chains about 55 nucleotides from the RNA polymerase II (Pol II) active site. These RDDs occur so soon after transcription that they are incompatible with known deaminase-mediated RNA editing mechanisms. Moreover, the 55-nucleotide delay in appearance indicates they do not arise during RNA synthesis by Pol II or as a direct consequence of modified base incorporation. Preliminary data suggest that RDD and R-loop formations may be coupled. These findings identify sequence substitution as an early step in co-transcriptional RNA processing.

Graphical Abstract

[†]Correspondence to: Vivian G. Cheung (vgcheung@umich.edu), John T. Lis (JohnLis@cornell.edu).
[*]These authors contributed equally to this work.
We dedicate this paper to the memory of Dr. Tom Kadesch who suggested the collaboration between the Lis and Cheung labs to study mechanisms that underlie RDDs.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Data: The sequence data have been deposited to the National Center for Biotechnology Information under the accession number GSE38233 (mRNA-seq), GSE39878 (chromatin-bound RNA-seq, GRO-seq and PRO-seq) and ERP001478 (DNA-seq).



DNA carries instructions for cellular proteins by providing the code that is transcribed into mRNA that in turn is translated into proteins. It is generally assumed that DNA sequences are copied faithfully into RNA. However, there are exceptions to this one-to-one relationship between RNA and its corresponding DNA sequences. The first example of a transcript sequence not encoded by the DNA was reported in 1986 by Benne and colleagues who showed that the *coxII* mRNA in trypanosome has 4 nucleotides not encoded in the DNA. They then coined the term RNA editing for this “novel mechanism of gene expression” (Benne et al., 1986). Other examples of RNA editing were soon discovered in organisms from plants to metazoans (Cattaneo et al., 1989; Driscoll et al., 1989; Gott et al., 1993; Gualberto et al., 1989). In humans, RNA editing occurs in processes mediated by ADAR (adenosine deaminases that act on RNA) (Bass and Weintraub, 1988) and APOBEC (apolipoprotein B mRNA editing enzymes) (Chen et al., 1987; Powell et al., 1987) families of proteins which lead to A-to-G (adenosine to inosine which is then recognized as guanosine) and C-to-U (cytidine to uridine) changes. Recently, advances in sequencing technologies have enabled deep sequencing of DNA and RNA which allowed us (Li et al., 2011) and others (Alon et al., 2012; Bar-Yaacov et al., 2013; Chen, 2013; Chen et al., 2012; Ju et al., 2011; Lagarrigue et al., 2013; Peng et al., 2012; Silberberg et al., 2012; Vesely et al., 2012) to uncover more RNA-DNA sequence differences (RDDs) than canonical RNA editing events. In different human cells and by using various sequencing and analytical methods, we and others have found all 12 types of RDDs.

While the mechanisms that mediate A-to-G and C-to-U editing in humans are known, we do not know how the other types of RDDs arise. For instance, A-to-C transversions are not likely to be mediated by ADAR and APOBEC families of deaminases. In this project, we ask when RDDs arise in order to distinguish the different types of underlying mechanisms. To address this, we compared nascent RNA sequences with their corresponding DNA sequences. The results show that all 12 types of RDDs occurred early during transcription. We found RDDs in transcripts beginning at approximately 55 bases from the active site or approximately 35 bases beyond the exit channel of RNA polymerase II (Pol II). This demonstrates that the RDD events occur by a mechanism distinct from altered base selectivity during catalysis of chain elongation by Pol II; nonetheless, the RNA processing events that mediate RDDs are closely coupled temporally and spatially to transcription in human cells. Given that RDDs emerge so soon after transcription, we studied cells from a patient with autosomal dominant form of juvenile ALS due to mutation in senataxin gene and found suggestive evidence that RDD formation may be coupled to R-loops.

Results

Nascent RNA from GRO-seq and PRO-seq

To determine whether RDDs occur during or after transcription, we sequenced nascent RNA using two global run-on sequencing methods, GRO-seq (Core et al., 2008) and precision run-on sequencing, PRO-seq (Figure 1A) (Kwak et al., 2013). We obtained ~ 100 million 100-nucleotide uniquely mapped GRO-seq reads from B-cells of two individuals. For one subject, we carried out two independent PRO-seq experiments and obtained ~60 million uniquely mapped reads in each. Additionally, we isolated and sequenced nascent RNA with an alternate method (Wuarin and Schibler, 1994) for comparison (chromatin-bound RNA-seq) (~190 million uniquely mapped reads). Finally, we carried out mRNA sequencing (mRNA-seq) and obtained ~135 million uniquely mapped RNA-seq reads, and sequenced the corresponding genomic DNA of the two individuals to 30× and 60× coverage.

We began by assessing the distributions of mapped reads from the libraries obtained by these four independent methods. As expected (Core et al., 2008; Kwak et al., 2013), GRO-seq and PRO-seq enriched for sequences near transcription start sites (Figure 1B, TSS). This enrichment in mammalian cells is due to promoter proximal pausing (sense strand) and upstream divergent transcription (antisense strand) (Core et al., 2008; Seila et al., 2008). Additionally, GRO-seq and PRO-seq data provide sensitive detection of active transcription units and identify over 9,000 transcriptionally-active genes. To ensure that we are looking at very nascent RNAs, we assessed the extent of splicing in GRO-seq and PRO-seq relative to chromatin-bound nascent RNA and mRNA. While about 20% of the mRNA-seq reads and 5% of the chromatin-bound nascent transcripts cover exon-exon junctions, less than 1% of the GRO-seq and PRO-seq reads span junctions. These nascent transcripts map throughout transcription units including introns (Core and Lis, 2008; Core et al., 2008; Core et al., 2012), while mRNA-seq libraries are dramatically depleted of introns but enriched in the 3' untranslated regions due to sample preparation for polyadenylated transcripts. These findings support that GRO-seq and PRO-seq correspond to greatly enriched short nascent RNA that is newly synthesized (also referred to as "very nascent RNAs" below), while chromatin-bound RNA represents longer transcripts on average from a later stage (referred to as "nascent RNA"). Figure 1C shows representative results for *UVRAG* and *CAPZB* from sequencing nascent and mature RNAs.

We also compared the expression levels of genes in the very nascent and mature mRNAs. The very nascent RNA differs from mature RNA in that the very nascent RNA levels depend on density of transcribing Pol II, while the mRNA levels depend on the rate of both transcription and mRNA decay. However, levels of transcripts in the two are significantly correlated (r=0.45, P<<0.0001) (Figure 1D) with outliers representing very stable or unstable mRNAs.

RDDs in very nascent RNA

Next, we turned to study RDDs in nascent RNA. Defining when RDDs arise during nascent transcription should help rule out or support particular mechanisms by which they are generated. Therefore, we analyzed the RNA sequences and their corresponding DNA sequences to assess how early during transcription the RNA-DNA differences arise. The steps to identify RDDs are shown in Figure 2. At sites that are covered by at least 10 uniquely mapped GRO-seq or PRO-seq reads and 10 monomorphic DNA reads (that contain only one nucleotide type: A, C, G or T), we compared the nascent RNA and corresponding DNA sequences and identified sites where RNA and DNA sequences are discordant. For a site to be identified as a candidate RDD, at least 10% of the GRO-seq or PRO-seq reads at that site (and a minimum of 2 unique reads) has to contain a sequence that differs from the underlying DNA sequences. All the resulting potential RDD sites were further processed in multiple steps to confirm their unique genomic locations.

The results uncovered 2,806 RDDs in one subject (GM12004), and 2,881 RDDs in the other individual (GM12750) (Tables S1 & S2). The orientation-specific sequencing allows us to distinguish all 12 possible types of mismatches between DNA and corresponding RNA sequences. In this analysis, we excluded C-to-T RDDs because the use of 5-bromouridine 5’-triphosphate (BrUTP) in GRO-seq may favor this type of misincorporation (Yu et al., 1993). All the 11 remaining types of RDDs were found; C-to-G was the most common in both samples (Figure 3A). We analyzed the PRO-seq data in the same way. Except for the 3’ most nucleotide, the sequenced RNAs from the PRO-seq sample are made in the cell (as opposed to about half in GRO-seq) thus it gives us longer segments of in vivo synthesized RNAs for analysis. We found 23,093 RDD sites out of about 115 million nucleotides screened, corresponding to one to two RDD per 10,000 bases screened and a frequency of ~ 2×10^{-4} RDD in the PRO-seq sample (Table S3) which is comparable to the frequency of RDD in mRNAs (also 10^{-4}) (Li et al., 2011). All 12 types of RDDs were identified (Figure 3B). Even though both GRO-seq and PRO-seq are global run-on assays coupled with deep sequencing, they are not identical, therefore different numbers of RDDs were detected in the two assays. Unlike GRO-seq, PRO-seq does not use BrUTP; thus miscorporation that favors C-to-T discordance is not a concern; therefore, we included all 12 types of RDDs in our analysis. This added more than 1,700 RDD sites (1,793 C-to-T). In addition, nearly the entire (except one or at most a few bases) PRO-seq transcripts as compared to ~15 to 20% of the GRO-seq transcripts are made in vivo. Together the addition of the C-to-T sites and the longer in vivo synthesized transcripts allowed us to identify about 8× more RDD sites in PRO-seq than GRO-seq. Despite the differences in the number, the distributions of RDD types are similar between GRO-seq and PRO-seq samples and across different thresholds of coverage and RDD levels (Figure S1). This reflects the robustness of our analysis. To be certain of our results, we confirmed the mapping and the sequences of the RDD sites with five different experiments and analyses, including genome walking, Sanger sequencing and droplet digital PCR using DNA and RNA from multiple tissues (Table 1&2, Figure S2) (see supplemental results and discussion for details).

Next, we examined RDDs from different experiments for overlaps. As one expects, the overlaps of RDD sites between the run-on experiments are low, since the ability to resample an RDD site in independent run-on assays depends on several parameters, including the density of transcribing Pol II, sequence depth, and RDD levels. GRO-seq and PRO-seq identify RDD sites in nascent RNA sequences that are closely associated (<100 nucleotides) with actively transcribing polymerases. Finding the same RDD event in two independent samples relies on sampling an RDD-bearing transcript bound to actively transcribing polymerases in both experiments; the chance of such occurrence is very low. The RDD identification also depends on sequence depths and the RDD levels (= number of RDD-containing reads/total number of reads at the site). The median RDD level among the sites detected in GRO-seq and PRO-seq is 0.24, therefore high coverage (~40×) is needed to obtain 80% of them in replicate samples (Chen, 2013). Nonetheless, 108 RDD sites were found in more than one sample (among the two GRO-seq and one PRO-seq datasets). The RDD sites we found in nascent RNAs were also present at a later stage of transcription. In chromatin-bound transcripts where we have longer transcripts and deeper coverage, we found over 1,000 RDD sites from one of the GRO-seq and/ or PRO-seq libraries. The distributions of these RDD sites are similar to those in GRO-seq and PRO-seq: T-to-G is one of the more abundant types and A-to-T is less frequent. These results show that the RDDs in nascent RNAs can be identified by different assays.

RDD formation occurs within seconds after transcription

To address how early during transcription do RDD events emerge, we first examined the GRO-seq results. As shown in Figure 1A, the GRO-seq reads comprise very nascent RNAs transcribed in vivo before nuclei isolation and a portion transcribed in vitro during the run-on. Since our very nascent RNAs are triple selected for BrU incorporation and we selectively analyzed reads with an identifiable 3'-end of the nascent RNA, the 3'-portion must contain the in vitro transcribed RNA and the 5'-portion contains some in vivo synthesized RNA. For both B cell samples, the majority of the RDDs are found in the 5' portion of the GRO-seq samples, which is enriched for the in vivo made nascent RNA (Figure 4A). These represent newly synthesized transcripts that have just exited the actively transcribing polymerase. These findings suggest that RDDs result from transcription-coupled RNA processing steps.

To further refine the time frame for these RDD events, we used PRO-seq to localize more precisely the RDD sites relative to actively transcribing RNA Pol II. In PRO-seq, the in vitro run-on assay was allowed only to proceed for one or at most a few nucleotides, thus the 3' ends of the PRO-seq reads mark precisely the locations of the transcriptionally active RNA polymerases in our B-cells. This offers an opportunity to examine nascent RNAs that have just exited the active site of Pol II. We examined where the RDDs were found relative to the actively transcribing Pol II, and as seen in the GRO-seq data, the RDD events occur after the RNA has exited the polymerase (Figure 4B). Moreover, the increased precision and accuracy afforded by PRO-seq allowed us to observe the abrupt increase at ~55 nucleotides from the active site of Pol II, corresponding to the sharp increase in RDD events around position 40 of the PRO-seq reads. As depicted in Figure 4B, the first ~20 bases from the 3' ends of the reads are nascent RNAs covered by RNA polymerase II, thus, RDD sites begin to appear about 35 bases after the RNA exits the polymerase. To confirm this observation, we repeated a PRO-seq experiment. The results confirmed our finding of an increase in RDD at ~55 nucleotides from the active site of Pol II (Figure S3A). In contrast, the RDDs found in mature mRNAs are more uniformly distributed as expected (Figure 4C). Moreover, analysis of sequencing quality score shows that the increase in RDD around 55 nt is not a result of a loss of fidelity (Figure S3B). These results are consistent with those from GRO-seq and demonstrate that RDD events appear to occur very rapidly (within seconds) after the nascent RNA is exposed, and are not occurring in the Pol II active site during the catalytic step of synthesizing RNA.

RDD frequency is lower in cells from a patient with Senataxin mutation

Our findings that RDDs emerge soon after nascent transcripts exit from transcription bubbles suggest the coupling of RDDs with R-loops (White and Hogness, 1977) which also initiate behind RNA polymerase. We therefore examined and found that RDDs are enriched significantly ($P<0.001$) in regions with R-loop forming sequences (Figure 4D) (Ginno et al., 2012; Wongsurawat et al., 2012). To study the co-occurrence of RDDs and R-loops, we carried out PRO-seq using cells from a patient with autosomal dominant form of juvenile Amyotrophic Lateral Sclerosis (ALS4) due to a mutation (L389S) in the Senataxin (*SETX*) gene that encodes a DNA/RNA helicase (Chen et al., 2004). The senataxin protein, SETX, interacts with RNA polymerase II (Chen et al., 2006; Ursic et al., 2004; Yuce and West, 2013) and plays a role in resolving R-loops particularly in transcription pause sites (Mischo et al., 2011; Skourti-Stathaki et al., 2011; Suraweera et al., 2009; Yuce and West, 2013). The mutation at position 389 corresponds to the N-terminus of SETX that interacts with other nuclear proteins including RNA polymerase II (Yuce and West, 2013). We found that there are 50% fewer RDDs in the very nascent RNA of the ALS4 sample; a frequency of 9×10^{-5} compared to 2×10^{-4}. Compared to controls, the RDD sites in the ALS4 sample skewed away from G-bearing transcripts; there are significantly more ($P=0.03$ (t-test)) RDD events that convert G in the DNA to other bases in the RNA (32% vs. 12% G-to-X, where X = A, C, or T (U)). Since R-loops preferentially form around nascent RNA that is G-rich (Roy and Lieber, 2009), this observation suggests the fewer RDDs in the ALS4 sample may be due to less efficient resolution of R-loops. These results encourage further studies to uncover the mechanistic connection of R-loops and RDDs.

A-to-G RDDs in very nascent RNA are not mediated by ADAR

In our B-cells, the only known editing mechanism is ADAR-mediated A-to-G editing (APOBEC1 is not expressed), so we asked if the A-to-G discrepancies in the nascent RNAs can be explained by ADAR proteins. Previously, ADAR-mediated editing was found in nascent RNA of Drosophila (Rodriguez et al., 2012) where nascent RNA was defined as chromatin-bound transcripts. We examined our chromatin-bound transcripts and mature poly-adenylated RNA, and found A-to-G editing events in both fractions, consistent with results in Drosophila. However, we did not find these A-to-G sites in GRO-seq or PRO-seq. For example, from mRNA-seq, we identified 65 A-to-G sites in *POLH*, and 48 of the adenosines were also edited in chromatin-bound RNA; but, none of these A-to-G sites were detected in the nascent RNA from GRO-seq or PRO-seq despite good sequence coverage (Figure S3C). For a more comprehensive analysis, we turned to results from several recent studies that have identified over 10,000 A>G editing sites (Bahn et al., 2011; Carmi et al., 2011; Kiran and Baranov, 2010; Li et al., 2009; Peng et al., 2012). None of the RDD sites in GRO-seq overlap with the editing sites reported in those studies. However, there are some A-to-G events in nascent RNA from GRO-seq and PRO-seq, so we compared the features of these A-to-G sites in nascent RNA with those known to be edited by ADAR-mediated deamination. We found that the sequence characteristics of the A-to-G sites in nascent and mature RNAs appear to be different. Most (>95%) of the ADAR-mediated A-to-G sites in polyadenylated mRNAs are found in Alu repeats (Athanasiadis et al., 2004; Chen, 2013), but in contrast, the A-to-G sites in very nascent (GRO- or PRO-seq) RNAs are not in Alu containing regions. In addition, the A-to-G sites in very nascent RNAs do not have the sequence motif (5’ depletion of G (Lehmann and Bass, 2000)) that flanks ADAR-edited adenosines (Fig S4A) (Wang et al., 2013). The data suggest that there are two distinct classes of A-to-G mismatches; those that are mediated by ADAR, and others that use a separate mechanism occurring on very nascent RNA during transcription.

Other characteristics of RDDs in very nascent RNA

Previous studies of RDDs focused on polyadenylated mRNAs (Bahn et al., 2011; Ju et al., 2011; Li et al., 2011; Peng et al., 2012); the very nascent RNAs in the present study allowed us to assess RDDs in regions such as introns that were spliced out in mature transcripts. Many of the RDDs in very nascent RNAs are found in intronic regions (28%), which could potentially affect downstream RNA processing steps. In addition, nearly half (44%) of the RDDs are intergenic (many of these correspond to gene isoforms with longer 5’ and 3’ UTRs relative to the REFseq forms). The remaining (28%) are found in exonic regions and evenly divided among coding exons and UTRs (48% and 52%, respectively). As we found previously (Li et al, 2011), unlike SNPs, there is no bias against nonsynonymous changes as ~70% of the coding RDD sites lead to alternate amino acids as predicted by the codon table. We studied the genes that contain RDD sites in nascent RNA and found that they are significantly ($P<10^{-30}$) enriched for roles in regulation and metabolism of nucleic acids and other macromolecules (see Table 3).

We also examined the sequences (10 bases) surrounding the RDD sites and showed that sequence context may be important. RDDs with the same DNA base share similar sequence characteristics. In particular, C-to-A and C-to-G, and the G-to-A, G-to-C and G-to-T RDDs share similar surrounding sequences. The RDDs whose DNA base is C reside in regions that are significantly more C-rich, while RDDs whose DNA base is G reside in regions that are significantly more G-rich than negative controls (Figure S4B & C) (t-test, $P< 0.05$). The enrichments of these nucleotides extend in both the 5’ and 3’ directions. These regions are more C-rich and G-rich, but they are not homopolymer tracts of Cs or Gs (Figure S4D). Thus, these are different from the co-transcriptional editing of homopolymer tracts in Ebola (Volchkov et al., 1995) and paramyxoviruses (Cattaneo et al., 1989; Paterson and Lamb, 1990). Additionally, RDDs whose DNA base is C show depletion of G at the base 3’ of the RDD, and those whose reference base is G show depletion of C at the base 5’ of the RDD. These features may affect the DNA and/or RNA structures, or possibly an RNA/DNA hybrid, which in turn signals for an RDD event as mentioned above.

Conclusions

We presented data from studying where RDDs occur and put them in context of known RNA editing mechanisms. We showed all 12 types of RDDs are found in RNAs that have recently extruded from the RNA Pol II exit channel. The RDD events occurred in vivo on transcripts about 35 nucleotides from the exit channel of Pol II. Pol II elongates in mammalian cells at 20 to 60 bases per second (Ardehali and Lis, 2009). Therefore, the RDD events found ~35 bases from the exit channel must occur very shortly after nascent RNA synthesis. Thus, our results indicate that RDDs are likely to occur within a few seconds of RNA synthesis and before classic RNA editing events. RNAs synthesized by RNA polymerase II are quickly modified: 5' caps are added as the RNA end exits the Pol II RNA channel (Rasmussen and Lis, 1993), introns are often spliced co-transcriptionally (Carrillo Oesterreich et al., 2010; Vargas et al., 2011) and 3'-ends are cleaved and polyadenylated before Pol II terminates transcription (Osheim et al., 2002). Based on knowledge of co-transcriptional processing events and results from the present study, we suggest that RDD occurs soon after the capping of the transcripts and before splicing.

The reason that we looked for timing of RDD is to help us to narrow the search for the underlying mechanisms that mediate its formation. A co-transcriptional event that coincides temporally with RDD formation is the emergence of R-loop (Broccoli et al., 2000; Drolet et al., 1995; Masse and Drolet, 1999). As a preliminary examination of whether there is association between RDD and R-loop, we studied RDDs in very nascent RNA of cells from a juvenile ALS patient with a mutation in the senataxin gene (Chen et al., 2004). The RNA/DNA helicase, senataxin, interacts with RNA polymerase and mediates the resolution of R-loops. We found that the patient has about 50% fewer RDDs in her nascent RNAs. The RDDs seem to be associated with R-loop since there is enrichment in R-loop forming sequences (Ginno et al., 2012) around RDD sites and depletion of G-bearing RDD transcripts in patient. These findings points to possible coupling of RDD and R-loop formations, and encourage further studies to uncover the molecular basis.

GRO-seq and PRO-seq assays allowed us to study very nascent RNA for RDD formation. But these methods also limit us to study sequences that are covered by or immediately adjacent (<100 bases) to actively transcribing polymerases. It is possible that there are other mechanisms, like ADAR-mediated editing, that modify RNA transcripts at a later stage of RNA processing. While our results show that RDD formation occurs very soon after RNA synthesis, they do not imply that all RDD formations have to occur as early co-transcriptional steps. Additional methods may be needed to identify or rule out existence of other processing steps that modify RNA sequences. Comparison of RNA sequences at different stages of maturity alone will not provide a comprehensive view because the levels of many RDD sites are low (below 30%), therefore the depth of sequencing necessary to conclude that a RDD site is absent in one stage of transcript synthesis but present in subsequent stages is difficult to achieve with current sequencing technologies given the constraints of error rate and cost. However, technologies to isolate RNA from different subcellular compartments and advances in sequence analysis are improving quickly; they soon will allow the tracking of individual transcripts through various processing steps and thus facilitate the determination of whether there are additional events that modify RNA sequences. In summary, we have identified sequence modification as an early RNA-processing step thus adding to the already complex set of events that add diversity to transcriptomes.

Experimental Procedures

DNA sequencing

Cultured B-cells from two normal individuals in the Centre d'Étude du Polymorphisme Humain database, GM12004 and GM12750, were obtained from Coriell Cell Repositories (NJ, USA). DNA-seq libraries were prepared and sequenced on HiSeq instrument to obtain 60× and 30× coverage, respectively (Illumina).

mRNA-seq and chromatin-bound nascent RNA-seq

For mRNA sequencing, RNA-seq libraries were prepared following Illumina TruSeq RNA sample preparation protocol. Chromatin-bound nascent RNA was extracted as previously described (Wuarin and Schibler, 1994). The mRNA and chromatin RNA were sequenced on HiSeq instrument.

GRO-seq and PRO-seq

Nuclei were isolated from cultured B cells and GRO-seq libraries were prepared with 5×10^6 nuclei as described previously (Core et al., 2008, 2012). PRO-seq libraries were prepared as described previously (Kwak et al., 2013). Briefly, 5×10^6 nuclei were added to 2 × Nuclear Run-On (NRO) reaction mixture (10 mM Tris-HCl pH 8.0, 300 mM KCl, 1% Sarkosyl, 5 mM $MgCl_2$, 1 mM DTT, 0.375 mM each of biotin-11-A/C/G/UTP (Perkin-Elmer), 0.8 u/µl RNase inhibitor) and incubated for 3 min at 30°C. Nascent RNA was extracted and fragmented by base hydrolysis in 0.2 N NaOH on ice for 10~12 min, and neutralized by adding 1 × volume of 1 M Tris-HCl pH 6.8. Fragmented nascent RNA was purified using streptavidin beads, ligated with reverse 3' RNA adapter (5'p-GAUCGUCGGACUGUAGAACUCUGAAC-/3'InvdT/), and biotin-labeled products were enriched by another round of streptavidin bead binding and extraction. For 5' end repair, the RNA products were successively treated with tobacco acid pyrophosphatase (TAP, Epicentre) and polynucleotide kinase (PNK, NEB). 5' repaired RNA was ligated to reverse 5' RNA adaptor (5'-CUGAACAAGCAGAAGACGGCAUACGA-3') before being further purified by the third round of streptavidin bead binding and extraction. RNA was reverse transcribed using 25 pmol RT primer (5'AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3'). The product was amplified 15±3 cycles and products greater than 150 bp (insert > 70 bp) were PAGE purified before being analyzed by Illumina HiSeq 2500 instrument. Two PRO-seq experiment was carried out, one at the Lis lab at Cornell University (Figure 3&4), and one at the Cheung lab at University of Pennsylvania (Figure S3A).

Sequence analysis

DNA-seq and RNA-seq reads were aligned to human reference genome (hg18) using GSNAP (Wu and Nacu, 2010) (version 2012-04-10). A list of SNP sites in the CEU population from Hapmap (release #28) and 1000 Genomes (pilot project) was used for SNP-tolerant alignments. Alignments with (read length + 2)/12 – 2 or fewer mismatches were obtained for each read. PRO-seq sequences were converted to the reverse-complements before alignment. For RNA sequence analysis, known exon-exon junctions (defined by RefSeq (downloaded March 7, 2011) and Gencode (version 3c)) and novel junctions (defined by GSNAP) were accepted. Read coverage was analyzed using RSeQC and RPKM (read per kilobase per million reads) for each gene were calculated (Wang et al., 2012). For GRO-seq and PRO-seq, we include all the reads covering exon or intron region in computing RPKM, while excluding 1kb-region downstream of TSS which is overrepresented by short transcripts associated with proximally paused Pol II.

RNA-DNA differences

To identify RDDs, we compared RNA sequence to its corresponding DNA sequence. Low-quality bases (Phred quality score < 20) in both the RNA and DNA were removed. To be included as RDD sites in the final lists, the following criteria had to be met: 1) a minimum of 10 total DNA-seq reads covering that site; 2) DNA sequence at this site is 100% concordant, without any DNA-seq reads containing alternative alleles; 3) a minimum of 10 total RNA-seq reads covering that site; 4) level of RDD (# of RNA-seq reads containing non-DNA allele/# all RNA-seq reads covering a given site) is ≥10% (a minimum of two RNA-seq reads containing RDD). To ensure the accuracy of the RDD sites, additional filtering steps were performed using two additional mapping algorithms. See supplemental experimental procedures for further details.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Nancy Zhang for discussions of estimation of sequencing errors, Jonathan Toung for analysis of the sequencing data, and Zhengwei Zhu for data analysis. We thank Dr. Kenneth Fischbeck for discussion of juvenile ALS and patient samples. Part of this work was carried out at the University of Pennsylvania prior to the Cheung lab's move to the University of Michigan. This work is supported by grants from the National Institutes of Health (to VGC MH087384, ES015733, JTL GM25232) and funds from the Howard Hughes Medical Institute (to VGC).

References

Alon S, Mor E, Vigneault F, Church G, Locatelli F, Galeano F, Gallo A, Shomron N, Eisenberg E. Systematic identification of edited microRNAs in the human brain. Genome Res. 2012; 22:1533–1540. [PubMed: 22499667]

Ardehali MB, Lis JT. Tracking rates of transcription and splicing in vivo. Nature structural & molecular biology. 2009; 16:1123–1124.

Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol. 2004; 2:e391. [PubMed: 15534692]

Bahn J, Lee J, Li G, Greer C, Peng G, Xiao X. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res. 2011; 29:29.

Bar-Yaacov D, Avital G, Levin L, Richards A, Hachen N, Rebolledo Jaramillo B, Nekrutenko A, Zarivach R, Mishmar D. RNA-DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA. Genome Res. 2013

Bass BL, Weintraub H. An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell. 1988; 55:1089–1098. [PubMed: 3203381]

Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell. 1986; 46:819–826. [PubMed: 3019552]

Broccoli S, Phoenix P, Drolet M. Isolation of the topB gene encoding DNA topoisomerase III as a multicopy suppressor of topA null mutations in Escherichia coli. Mol Microbiol. 2000; 35:58–68. [PubMed: 10632877]

Carmi S, Borukhov I, Levanon EY. Identification of widespread ultra-edited human RNAs. PLoS Genet. 2011; 7:e1002317. [PubMed: 22028664]

Carrillo Oesterreich F, Preibisch S, Neugebauer KM. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. Mol Cell. 2010; 40:571–581. [PubMed: 21095587]

Cattaneo R, Kaelin K, Baczko K, Billeter MA. Measles virus editing provides an additional cysteine-rich protein. Cell. 1989; 56:759–764. [PubMed: 2924348]

Chen L. Characterization and comparison of human nuclear and cytosolic editomes. Proc Natl Acad Sci U S A. 2013; 110:E2741–E2747. [PubMed: 23818636]

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012; 148:1293–1307. [PubMed: 22424236]

Chen SH, Habib G, Yang CY, Gu ZW, Lee BR, Weng SA, Silberman SR, Cai SJ, Deslypere JP, Rosseneu M, et al. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific inframe stop codon. Science. 1987; 238:363–366. [PubMed: 3659919]

Chen YZ, Bennett CL, Huynh HM, Blair IP, Puls I, Irobi J, Dierick I, Abel A, Kennerson ML, Rabin BA, et al. DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). Am J Hum Genet. 2004; 74:1128–1135. [PubMed: 15106121]

Chen YZ, Hashemi SH, Anderson SK, Huang Y, Moreira MC, Lynch DR, Glass IA, Chance PF, Bennett CL. Senataxin, the yeast Sen1p orthologue: characterization of a unique protein in which recessive mutations cause ataxia and dominant mutations cause motor neuron disease. Neurobiology of disease. 2006; 23:97–108. [PubMed: 16644229]

Core L, Lis J. Transcription regulation through promoter-proximal pausing of RNA polymerase II. Science. 2008; 319:1791–1792. [PubMed: 18369138]

Core L, Waterfall J, Lis J. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008; 322:1845–1848. Epub 2008 Dec 1844. [PubMed: 19056941]

Core LJ, Waterfall JJ, Gilchrist DA, Fargo DC, Kwak H, Adelman K, Lis JT. Defining the status of RNA polymerase at promoters. Cell reports. 2012; 2:1025–1035. [PubMed: 23062713]

Driscoll DM, Wynne JK, Wallis SC, Scott J. An in vitro system for the editing of apolipoprotein B mRNA. Cell. 1989; 58:519–525. [PubMed: 2758465]

Drolet M, Phoenix P, Menzel R, Masse E, Liu LF, Crouch RJ. Overexpression of RNase H partially complements the growth defect of an Escherichia coli delta topA mutant: R-loop formation is a major problem in the absence of DNA topoisomerase I. Proc Natl Acad Sci U S A. 1995; 92:3526–3530. [PubMed: 7536935]

Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell. 2012; 45:814–825. [PubMed: 22387027]

Gott JM, Visomirski LM, Hunter JL. Substitutional and insertional RNA editing of the cytochrome c oxidase subunit 1 mRNA of Physarum polycephalum. J Biol Chem. 1993; 268:25483–25486. [PubMed: 8244983]

Gualberto JM, Lamattina L, Bonnard G, Weil JH, Grienenberger JM. RNA editing in wheat mitochondria results in the conservation of protein sequences. Nature. 1989; 341:660–662. [PubMed: 2552325]

Ju YS, Kim JI, Kim S, Hong D, Park H, Shin JY, Lee S, Lee WC, Yu SB, Park SS, et al. Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. Nat Genet. 2011; 43:745–752. [PubMed: 21725310]

Kiran A, Baranov PV. DARNED: a DAtabase of RNa EDiting in humans. Bioinformatics. 2010; 26:1772–1776. [PubMed: 20547637]

Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science. 2013; 339:950–953. [PubMed: 23430654]

Lagarrigue S, Hormozdiari F, Martin LJ, Lecerf F, Hasin Y, Rau C, Hagopian R, Xiao Y, Yan J, Drake TA, et al. Limited RNA editing in exons of mouse liver and adipose. Genetics. 2013; 193:1107–1115. [PubMed: 23410828]

Lehmann KA, Bass BL. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry. 2000; 39:12875–12884. [PubMed: 11041852]

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science. 2009; 324:1210–1213. [PubMed: 19478186]

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 2011; 333:53–58. [PubMed: 21596952]

Masse E, Drolet M. Escherichia coli DNA topoisomerase I inhibits R-loop formation by relaxing transcription-induced negative supercoiling. J Biol Chem. 1999; 274:16659–16664. [PubMed: 10347234]

Mischo HE, Gomez-Gonzalez B, Grzechnik P, Rondon AG, Wei W, Steinmetz L, Aguilera A, Proudfoot NJ. Yeast Sen1 helicase protects the genome from transcription-associated instability. Mol Cell. 2011; 41:21–32. [PubMed: 21211720]

Osheim YN, Sikes ML, Beyer AL. EM visualization of Pol II genes in Drosophila: most genes terminate without prior 3' end cleavage of nascent transcripts. Chromosoma. 2002; 111:1–12. [PubMed: 12068918]

Paterson RG, Lamb RA. RNA editing by G-nucleotide insertion in mumps virus P-gene mRNA transcripts. Journal of virology. 1990; 64:4137–4145. [PubMed: 2166809]

Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat Biotechnol. 2012; 30:253–260. [PubMed: 22327324]

Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell. 1987; 50:831–840. [PubMed: 3621347]

Rasmussen EB, Lis JT. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. Proc Natl Acad Sci U S A. 1993; 90:7923–7927. [PubMed: 8367444]

Rodriguez J, Menet JS, Rosbash M. Nascent-seq indicates widespread cotranscriptional RNA editing in Drosophila. Mol Cell. 2012; 47:27–37. [PubMed: 22658416]

Roy D, Lieber MR. G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter. Mol Cell Biol. 2009; 29:3124–3133. [PubMed: 19307304]

Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. Divergent transcription from active promoters. Science. 2008; 322:1849–1851. [PubMed: 19056940]

Silberberg G, Lundin D, Navon R, Ohman M. Deregulation of the A-to-I RNA editing mechanism in psychiatric disorders. Hum Mol Genet. 2012; 21:311–321. [PubMed: 21984433]

Skourti-Stathaki K, Proudfoot NJ, Gromak N. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. Mol Cell. 2011; 42:794–805. [PubMed: 21700224]

Suraweera A, Lim Y, Woods R, Birrell GW, Nasim T, Becherel OJ, Lavin MF. Functional role for senataxin, defective in ataxia oculomotor apraxia type 2, in transcriptional regulation. Hum Mol Genet. 2009; 18:3384–3396. [PubMed: 19515850]

Ursic D, Chinchilla K, Finkel JS, Culbertson MR. Multiple protein/protein and protein/RNA interactions suggest roles for yeast DNA/RNA helicase Sen1p in transcription, transcription-coupled DNA repair and RNA processing. Nucleic Acids Res. 2004; 32:2441–2452. [PubMed: 15121901]

Vargas DY, Shah K, Batish M, Levandoski M, Sinha S, Marras SA, Schedl P, Tyagi S. Single-molecule imaging of transcriptionally coupled and uncoupled splicing. Cell. 2011; 147:1054–1065. [PubMed: 22118462]

Vesely C, Tauber S, Sedlazeck FJ, von Haeseler A, Jantsch MF. Adenosine deaminases that act on RNA induce reproducible changes in abundance and sequence of embryonic miRNAs. Genome Res. 2012; 22:1468–1476. [PubMed: 22310477]

Volchkov VE, Becker S, Volchkova VA, Ternovoj VA, Kotov AN, Netesov SV, Klenk HD. GP mRNA of Ebola virus is edited by the Ebola virus polymerase and by T7 and vaccinia virus polymerases. Virology. 1995; 214:421–430. [PubMed: 8553543]

White RL, Hogness DS. R loop mapping of the 18S and 28S sequences in the long and short repeating units of Drosophila melanogaster rDNA. Cell. 1977; 10:177–192. [PubMed: 402221]

Wongsurawat T, Jenjaroenpun P, Kwoh CK, Kuznetsov V. Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity. Nucleic Acids Res. 2012; 40:e16. [PubMed: 22121227]

Wuarin J, Schibler U. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. Mol Cell Biol. 1994; 14:7219–7225. [PubMed: 7523861]

Yu H, Eritja R, Bloom LB, Goodman MF. Ionization of bromouracil and fluorouracil stimulates base mispairing frequencies with guanine. J Biol Chem. 1993; 268:15935–15943. [PubMed: 7688001]

Yuce O, West SC. Senataxin, defective in the neurodegenerative disorder ataxia with oculomotor apraxia 2, lies at the interface of transcription and the DNA damage response. Mol Cell Biol. 2013; 33:406–417. [PubMed: 23149945]

Highlights

- RNA sequences differ from corresponding DNA sequences beyond canonical RNA editing;
- RNA-DNA sequence differences (RDDs) are found in nascent RNAs
- RDD formation occurs soon after transcript synthesis and before splicing.

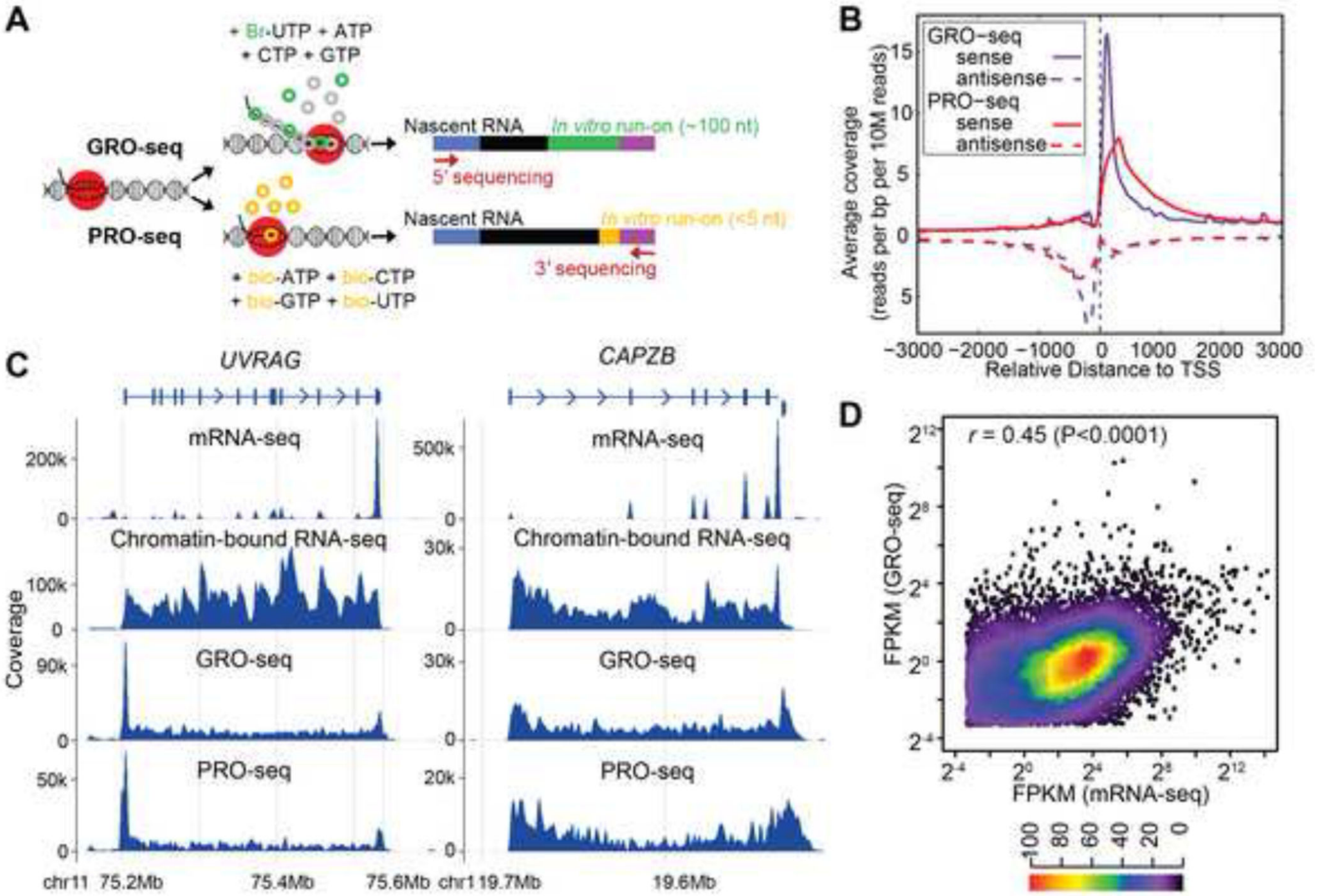


Figure 1.

GRO-seq and PRO-seq analysis. (**A**) Schematic of GRO-seq and PROseq. (**B**) Comparison between GRO-seq and PRO-seq. Sense and antisense transcripts associated with transcription start sites (TSS) are shown for GRO-seq and PRO-seq samples. The slight shift of the PRO-seq promoter-proximal peak downstream relative to the GRO-seq peak is because the PRO-seq reads that were less than 35 nucleotides were not mapped in the analysis, and because GRO-seq maps 5’ ends and PRO-seq maps 3’ ends of nascent RNAs. (**C**) mRNA-seq, chromatin-bound nascent RNA-seq, GRO-seq and PRO-seq results for two representative genes, *UVRAG* and *CAPZB*. For genes with proximal Pol II pausing such as *UVRAG*, there are more reads mapping to the 5’ ends of genes in both GRO-seq and PRO-seq samples. Schematic gene structure is aligned to mRNA-seq results, with boxes representing exons, lines representing introns and arrowheads showing direction of transcription. Coverage is calculated using bin size of ~ 1500 bp and 600 bp, respectively. (**D**) Scatter plot of gene expression levels from GRO-seq and mRNA-seq (FPKM>0.1). Results from GM12750 (shown) and GM12004 are similar (r=0.45 for both samples). Heatmap indicates frequency of different expression levels.

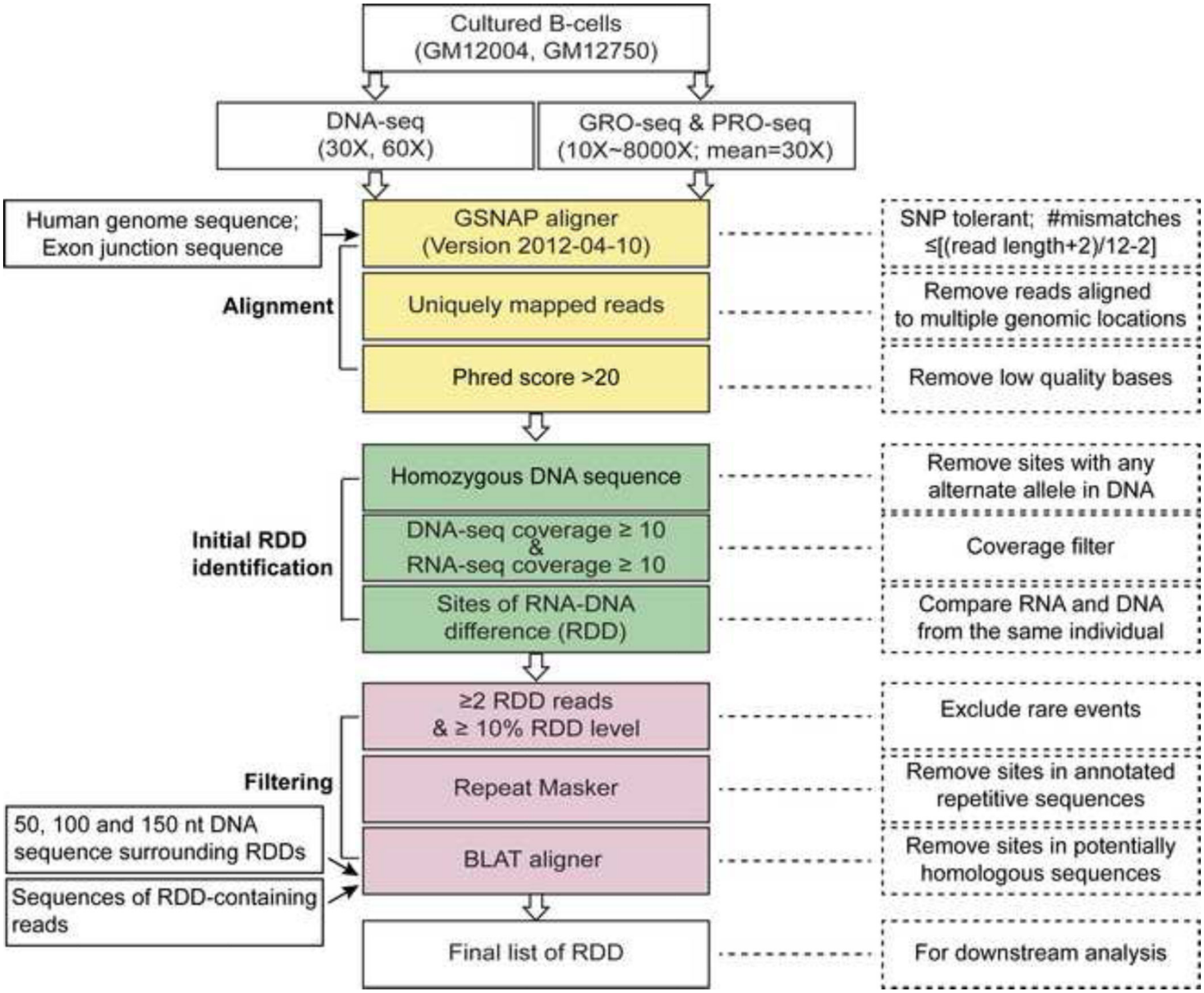

Cultured B-cells
(GM12004, GM12750)
DNA-seq
(30X, 60X)
GRO-seq & PRO-seq
(10X~8000X; mean=30X)
Human genome sequence;
Exon junction sequence
GSNAP aligner
(Version 2012-04-10)
SNP tolerant; #mismatches ≤[(read length+2)/12-2]
Alignment
Uniquely mapped reads
Remove reads aligned to multiple genomic locations
Phred score >20
Remove low quality bases
Homozygous DNA sequence
Remove sites with any alternate allele in DNA
Initial RDD identification
DNA-seq coverage ≥ 10 & RNA-seq coverage ≥ 10
Coverage filter
Sites of RNA-DNA difference (RDD)
Compare RNA and DNA from the same individual
≥2 RDD reads & ≥ 10% RDD level
Exclude rare events
Filtering
Repeat Masker
Remove sites in annotated repetitive sequences
50, 100 and 150 nt DNA sequence surrounding RDDs
Sequences of RDD-containing reads
BLAT aligner
Remove sites in potentially homologous sequences
Final list of RDD
For downstream analysis

Figure 2.

Analysis steps to identify RNA-DNA sequence differences. See also Table S1–S3.

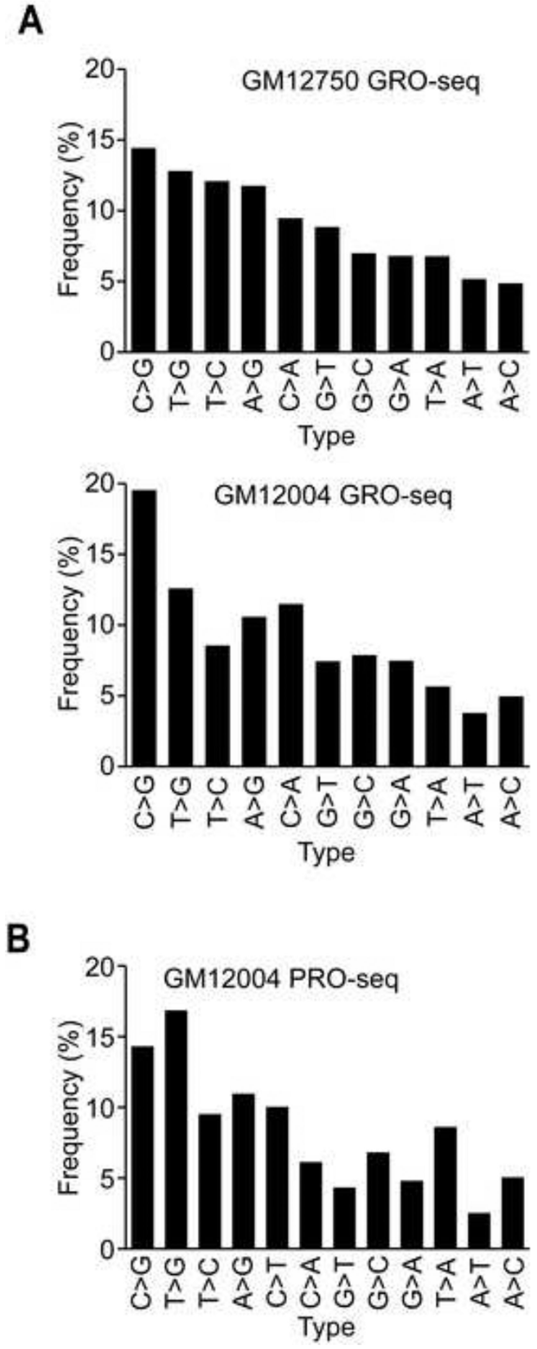


Figure 3.

RNA-DNA differences in very nascent transcripts. Distributions of RDD types (**A**) in GRO-seq samples of two individuals, (**B**) in PRO-seq. RDD types were ordered as in (**A**) and C-to-T RDDs for the PRO-seq sample.

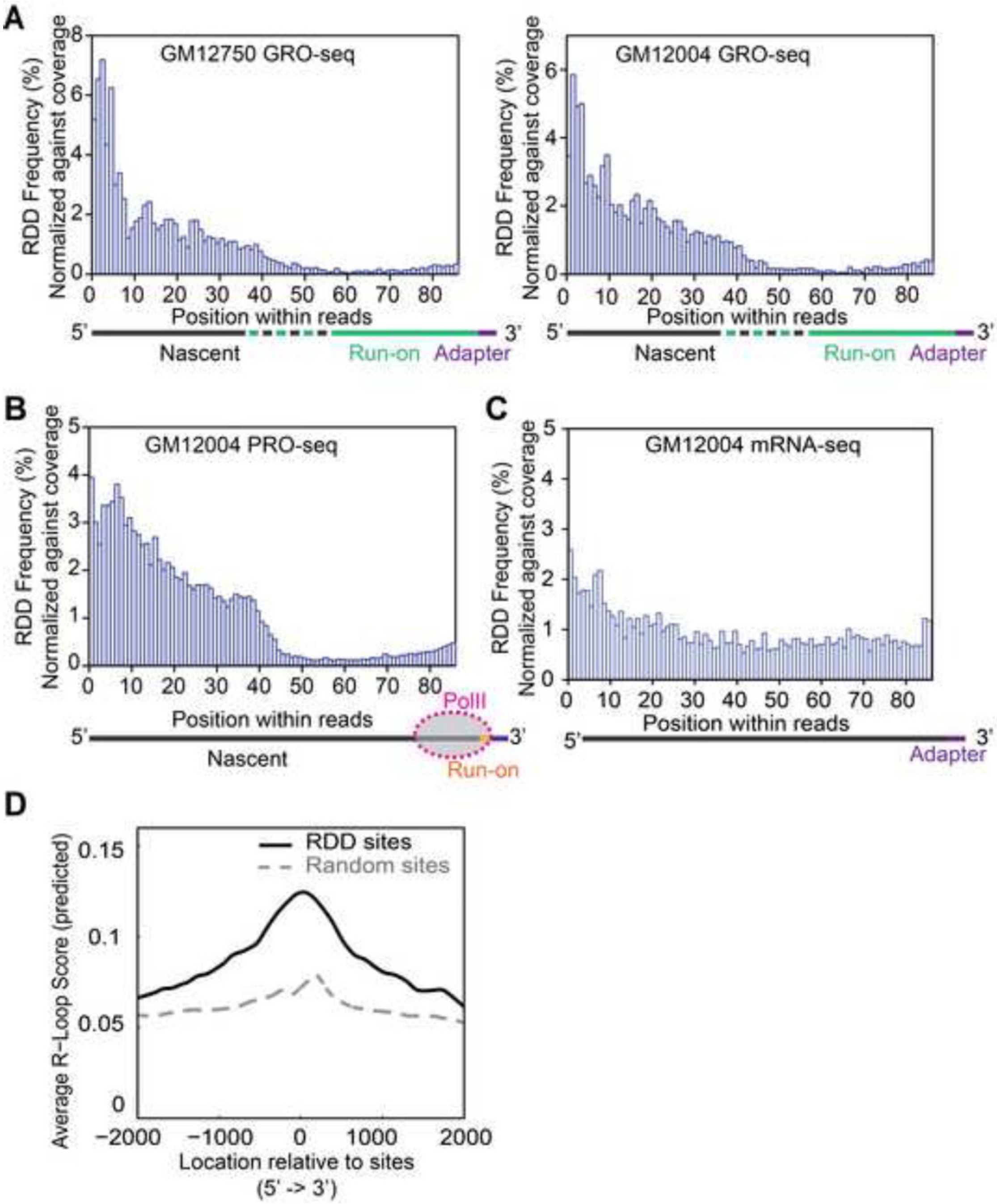


Figure 4.

Locations of RDD sites within sequencing reads. (**A**) Locations of RDD sites along GRO-seq reads. Only reads that have defined 3’ ends (reads that contain 3’ adapter sequences) were included in our analysis. (**B**) Locations of RDD sites along PRO-seq reads. Schematic diagrams indicate the locations of the different segments of GRO-seq (**A**) and PRO-seq (**B**) transcripts along the sequence reads. (**C**) Locations of RDD sites along mRNA-seq reads. (**D**) R-loop forming sequences are enriched in regions immediately adjacent to RDD sites.

Average R-loop scores for 2 kb of regions up and downstream of RDD sites are shown. RDD sites have significantly higher R-loop scores (P<0.001, t-test) than random control sites.

Table 1

Results of genome-walking confirm that RDDs are in unique regions of the genome.

Genomic location	RDD type	Plus strand		Minus strand	
		Sequence	# clones	Sequence	# clones
chr1:152175284	G-to-A	G	31	G	9
chr6: 107088915	T-to-A	T	19	T	21
chr9:34336911	G-to-A	G	14	G	10
chr11:72079055	T-to-C	T	10	T	10
chr12:100980077	A-to-C	A	11	A	18
chr14:20221257	G-to-T	G	14	G	16
chr16: 2140620	A-to-G	A	2	A	14
chr19: 2427122	C-to-A	C	15	C	1
chr22:42867269	T-to-C	T	13	T	14
chrX:7004437	G-to-T	G	11	G	10

Table 2

ddPCR validation of GRO-seq RDDs.

Genomic Location	Gene Name	RDD Type	Feature	Individual	Level in nascent RNA	
					GRO-seq (%)	ddPCR (%)
1:152175284	*DENND4B*[*]	G>A	Coding exon	GM12004	20	15
				GM12750	0	0
3:197100758	*TNK2*	G>T	Intron	GM12004	0	0
				GM12750	9	9
6:161450890	*MAP3K4*[*]	G>T	Coding exon	GM12004	17	1
				GM12750	0	2
6:37987903	*ZFAND3*	G>C	Intron	GM12004	0	0
				GM12750	7	3
9:34336911	----	G>A	Intergenic	GM12004	8	19
				GM12750	14	27
11:58103493	*ZFP91*	G>C	Coding exon	GM12004	0	0
				GM12750	18	10
11:72079055	*ARAP1*[*]	T>C	Intron	GM12004	0	0
				GM12750	11	7
12:100980077	----	A>C	Intergenic	GM12004	18	17
				GM12750	0	0
16:69880869	*FTSJD1*	C>G	5' UTR	GM12004	9	3
				GM12750	0	0
17:30949447	*AP2B1*	G>C	Coding exon	GM12004	0	0
				GM12750	10	16
18:8628755	*RAB12*[*]	T>C	3' UTR	GM12004	0	0
				GM12750	11	9
17:34815068	*MED1*	G>T	3' UTR	GM12004	0	0
				GM12750	13	10
19:2197783	*SF3A2*	G>C	Coding exon	GM12004	0	0

Genomic Location	Gene Name	RDD Type	Feature	Individual	Level in nascent RNA	
					GRO-seq (%)	ddPCR (%)
				GM12750	33	10
X:7004437	*HDHD1A*[*]	G>T	Intron	GM12004	43	50
				GM12750	0	0

[*] Also found in nuclear RNA fractions of both individuals (Figure S4) except the site in HDHD1A was found in GM12004 but not GM12750.

[†] We included a few RDD sites with levels <10% in the validations even though the analyses focused on sites whose levels are >10%. As shown, even the sites with lower levels were validated by ddPCR analysis of these same libraries.

Table 3

Genes with RDDs in their nascent RNAs are enriched for roles in regulation and metabolism of macromolecules.

GO Term	Examples	P-value
gene expression	*RNF10, ZNF791, KDM2B; DHX9; ELF4*	1.8×10^{-60}
nucleic acid metabolic process	*SP3, MAX, RPS6KA4; PSMD11; UTP23*	6.2×10^{-60}
RNA metabolic process	*RPS24; ELF1; CPEB2; DHX9; NFX1*	2.6×10^{-58}
cellular macromolecule biosynthetic process	*DPF1; SEC14L2; RPL18A; UPF1; HARS*	4.5×10^{-53}
macromolecule biosynthetic process	*ARFRP1;CTBP2; TSG101; GTF3C2; PARP10*	4.3×10^{-51}
regulation of macromolecule metabolic process	*AXIN1; FYN; VCP; SMARCA5; ZNF7*	3.9×10^{-50}
regulation of cellular metabolic process	*BCOR; ELL; MTF1; STAT5A; VPS36*	2.4×10^{-49}
cellular protein metabolic process	*CCT8; TCF3; RNF115; UBE4B; LNX1*	4.9×10^{-49}
regulation of primary metabolic process	*ATG7; CLIP3; YLPM1; CD44; POGK*	8.6×10^{-47}
regulation of nitrogen compound metabolic process	*AGRN; SMARCC1; MOV10; SUMO1; HSPA8*	4.6×10^{-36}