SUPPLEMENTAL TEXT SECTIONS

Supplemental Experimental Procedures

Cell culture. 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 (Camden, NJ, USA). The B-cells were grown to a density of $5x10^5$ cells/ml in RPMI 1640 supplemented with 15% fetal bovine serum, 100 units/ml penicillin and 100 µg/ml streptomycin, and 2 mmol/L L-glutamine. For downstream experiments, cells were harvested 24 hours after addition of fresh medium. Fibroblasts were cultured from a skin biopsy obtained at the National Institutes of Health Clinical Center from an adult female patient with autosomal dominant juvenile Amyotrophic Lateral Sclerosis due to a Senataxin mutation (L389S). Primary fibroblasts were prepared and cultured in MEM medium supplemented with 10% fetal bovine serum, 2 mM L-glutamine, and 100 U/mL penicillin-streptomycin. Fibroblasts were seeded to 70% confluency the day before experiments.

DNA sequencing. Genomic DNA was extracted from cultured B-cells of GM12004 and GM12750 using DNeasy blood and tissue kit (Qiagen, Valencia, CA, USA). DNA-seq libraries were prepared and sequenced on HiSeq 2000 instrument (Illumina, San Diego, CA, USA). Paired-end 100-nt reads were generated in order to achieve 60X and 30X coverage for GM12004 and GM12750, respectively. Low-quality bases as designated by Illumina were trimmed from the 3' end of reads, and reads shorter than 35bp were removed. Reads were aligned to an index comprising the human reference genome (hg18) and the Epstein-Barr virus genome (NC_009334.1) 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. Read pairs that aligned in the correct orientation (forward-reverse) were retained for further analyses. To select sites for further consideration: we identified those that 1) were covered by 10 or more reads, and number of reads was no greater than 3 times of the mean; 2) had only one type of nucleotide in all reads (homozygous sites).

mRNA-seq and chromatin-bound nascent RNA-seq. For mRNA sequencing, total RNA was extracted from cultured B-cells using the RNeasy Mini kit with DNase treatment (Qiagen). RNA-seq libraries were prepared following Illumina TruSeq RNA sample preparation protocol. The samples were sequenced using HiSeq 2000 instrument and 100-200 million 100-nt reads per sample were generated. For chromatin-associated nascent RNA-seq, cultured B cells were treated with cell fractionation buffer from the PARIS kit (Ambion) following manufacturer's protocol to obtain a nuclei pellet. Chromatin fraction was extracted from this pellet as previously described (Pandya-Jones and Black, 2009; Wuarin and Schibler, 1994). Briefly, the nuclei pellet was washed with ice-cold 1× PBS containng 1 mM EDTA, resuspended in an ice-cold glycerol buffer (20 mM Tris-HCl, pH 7.9, 75 mM NaCl, 0.5 mM EDTA, 0.85 mM DTT, 0.125 mM PMSF, 50% glycerol). An equal volume of ice-cold nuclei lysis buffer (10 mM HEPES, pH 7.6, 1 mM DTT, 7.5 mM MgCl2, 0.2 mM EDTA, 0.3 M NaCl, 1 M UREA, 1% NP-40) was added and the tube was gently vortexed, incubated for 2 min on ice, and centrifuged for 2 min, 4°C at 14,000 rpm. The supernatant was

removed and the chromatin pellet was gently washed with ice-cold 1×PBS containing 1 mM EDTA. Chromatin RNA was prepared from this pellet using TRI Reagent RT (Molecular Research Center). This RNA was further purified using the RNeasy Mini protocol (Qiagen) with on-column DNase digestion. After being shown to be free of genomic DNA, the chromatin RNA was sequenced on HiSeq 2000 as described above. The chromatin-bound RNA showed a 10-fold enrichment of U6 expression and a 4-fold depletion of ribosomal protein S14 expression compared to cytoplasmic RNA, confirming the quality of the nascent transcripts.

GRO-seq and PRO-seq

Nuclei isolation. Nuclei were isolated from B cells in a similar manner as Core et al, 2008 (Core et al., 2008). Briefly, 4 X 10^7 B cells were collected by centrifugation at 400 X g for 2 min at 4^{o}C. The cells were washed with 20 ml of ice-cold PBS and resuspended in 10 ml of ice-cold lysis buffer [20 mM Tris-HCl pH 7.4, 150 mM KCl, 1.5 mM $MgCl_2$, 1 mM DTT, 0.5% Igepal CA-630, 1X Complete Protease Inhibitor Cocktail (Roche) and 4 units/ml RNAseOUT (Invitrogen)]. Cell suspension was incubated on ice for 10 minutes before nuclei were centrifuged by 500 X g for 1 min at 4^{o}C. Pellets containing nuclei were washed carefully with 10 ml ice-cold lysis buffer, collected by centrifugation (500Xg, 1 min, 4^{o}C), and resuspended in ice-cold storage buffer [50 mM Tris-HCl pH 8.3, 5 mM $MgCl_2$, 0.1 mM EDTA, 40% glycerol] to 5X10^6 nuclei/100 µl. Nuclei were then snap frozen in liquid nitrogen until GRO-seq experiments.

GRO-seq library preparation. Libraries were prepared with 5X10^6 nuclei as in Core et al, 2008, with the following modifications. Trizol (Invitrogen) was used to stop the

reaction instead of DNase I and proteinase K treatment. The RNA was further extracted once with acid phenol:chloroform, and once with chloroform before precipitating with 2.5 volumes of -20^{o}C ethanol. Bead binding buffers all contained 4 units/ml of SUPERaseIn (Ambion) and the following buffers were slightly modified. Bead blocking buffer: 0.25 X SSPE, 1 mM EDTA, 0.05% Tween, 0.1% PVP, and 1 mg/ml ultrapure BSA (Ambion); Binding buffer: 0.25 X SSPE, 37.5 mM NaCl, 1 mM EDTA, 0.05% Tween; Low salt wash buffer: 0.2 X SSPE, 1 mM EDTA, 0.05% Tween. High salt wash buffer: 0.25 X SSPE, 137.5 mM NaCl, 1 mM EDTA, 0.05% Tween. The end repair steps were modified as following. Pelleted RNA from the first bead binding was resuspended in 20 µl DEPC-treated water and heated to 70^{o}C for 5 min, followed by incubation on ice for 2 min. 1.5 µl tobacco acid pyrophosphatase (TAP) buffer, 4.5 µl H_2O, 1 ul SUPERaseIn, and 1.5 µl TAP (Epicentre) were then added and the reaction incubated at 37^{o}C for 1.5 hours. 1 µl 300 mM $MgCl_2$ and 1 µl T4 polynucleotide kinase (PNK) were added to the reaction for an additional 30 min. To phosphorylate the 5'-ends, 20 µl T4 PNK buffer, 2 µl 100mM ATP, 143 µl water, 1 µl SUPERaseIn, and an additional 2 µl of PNK were added for 30 min at 37^{o}C. The reaction was then stopped by addition of 20 mM EDTA followed by acid phenol extraction and precipitation.

For the 5'-ligation (RNA oligo: 5'-GUUCAGAGUUCUACAGUCCGACGAUC-3') and 3'-ligation (RNA oligo: 5'-UCGUAUGCCGUCUUCUGCUUGUidT-3') reaction, RNA was resuspended in 10 µl ddH_2O and mixed with 1.5 µl of the adapter (25 uM) and 2 µl 50% PEG 8000. The components were heated to 70°C for 2 min and then placed on ice for 2 min. 2 µl 10 X RNA ligase 1 buffer (NEB), 2 µl 10mM ATP, 1 µl SUPERaseIn (Ambion), and 1.5 µl T4 RNA ligase 1 were then added and the mixture incubated at

22°C for 5-6 hours. 28 µl ddH_2O and 2 µl 0.5 M EDTA were added to stop the reaction, and the RNA was subjected to bead enrichment.

RNA was reverse transcribed into cDNA using 25 pmol Illumina reverse transcription primer (DNA oligo: 5'-CAAGCAGAAGACGGCATACGA-3') and SuperScript III Reverse Transcriptase (Invitrogen), according to the manufacturer's instruction except that the reaction was performed at 48°C for 15 min followed by 54°C for 45 minutes. Following a trial amplification to determine the optimal number of cycles, libraries were amplified for 15 cycles using Phusion polymerase (NEB), under standard PCR conditions with 500 nM final concentration of oligos GX1 (DNA oligo: 5'-CAAGCAGAAGACGGCATAC-GA-3') and GX2 (DNA oligo: 5'-AATGATACGGCGACCACCGACAGGTTCAGAGTTC-TACAGTCCGA-3'). Libraries were PAGE purified and all inserts greater than 20 bp were selected for high-throughput sequencing with Illumina HiSeq.

PRO-seq library preparation. 5×10^6 nuclei were added to the same volume of 2 X 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 using Trizol and precipitated in 75% ethanol. Extracted nascent RNA was fragmented by base hydrolysis in 0.2 N NaOH on ice for 10~12 min, and neutralized by adding 1X volume of 1 M Tris-HCl pH 6.8. Excessive salt was removed by using nuclease-free P-30 column (Bio-Rad). Fragmented nascent RNA was bound to 30 µl of Streptavidin M-280 magnetic beads (Invitrogen) following the manufacturer's instructions. The beads were washed once in high salt (2 M NaCl, 50 mM Tris-HCl pH 7.4, 0.5% Triton X-100), once in medium salt (300 mM NaCl, 10 mM Tris-HCl pH 7.4,

0.1% Triton X-100) and once in low salt (5 mM Tris-HCl pH 7.4, 0.1% Triton X-100). Bound RNA was extracted from the beads using Trizol (Invitrogen) in two consecutive extractions, and the RNA fractions were pooled, followed by ethanol precipitation.

For the first ligation reaction, fragmented nascent RNA was dissolved in ddH_2O and incubated with 10 pmol of reverse 3' RNA adaptor (5'p-GAUCGUCGGACUG-UAGAACUCUGAAC-/3'InvdT/) and T4 RNA ligase I (NEB) under manufacturer's condition for 6 hr at 20˚C. Ligated RNA was enriched with biotin-labeled products by another round of Streptavidin bead binding and extraction. For the 5' end repair, the RNA products were successively treated with tobacco acid pyrophosphatase (TAP, Epicentre) and polynucleotide kinase (PNK, NEB). Each reaction was followed by an ethanol precipitation step. 5' repaired RNA was ligated to reverse 5' RNA adaptor (5'-CUGAACAAGCAGAAGACGGCAUACGA-3'). Ligated RNA products were further enriched for biotin-labels by the third round of streptavidin bead binding and extraction. Adaptor ligated nascent RNA was reverse transcribed using 25 pmol RT primer (5'-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3') (GX2 primer, Illumina).

A portion of the RT product was removed and used for trial amplifications to determine the optimal number of PCR cycles. For the final amplification, 12.5 pmol of GX1 primer (Illumina) was added to the RT product with Phusion polymerase (NEB) under standard PCR conditions. Excess RT primer served as one of the primer pair for the PCR. 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 instrument. First PRO-seq experiment was carried out at Cornell University (Figure 2, PRO-seq), and the

replicated PRO-seq experiment was carried out at University of Pennsylvania (Figure S3A, PRO-seq-Replicated). Two PRO-seq datasets differ in number of sites that are covered by ≥10 RNA-seq reads and ≥10 DNA-seq reads; Number of sites covered in PRO-seq-replicated is 70% of that in PRO-seq. However, the position of RDDs relative to active Pol II is highly reproducible.

Sequence analysis. The GRO-seq and PRO-seq samples were sequenced using HiSeq 2000 instrument and 100-200 million 100-nt reads per sample were generated. Low-quality bases as designated by Illumina were trimmed from the 3' end of reads, 3' adapter was trimmed using FASTQ/A Clipper with default setting (Hannon lab) and reads shorter than 35 bp were removed. For analyses of PRO-seq data, the sequences were converted to the reverse-complements, and the terminal 10 bases (closest to the polymerase active site location) were removed to avoid complications that arise from mis-incorporation of biotin nucleotides. The resulting reads were aligned to an index comprising the human reference genome (hg18) and the Epstein-Barr virus genome (NC_009334.1) 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 to allow for SNP-tolerant alignments. The following parameters were used: Mismatches ≤[(read length+2)/12-2]; Mapping score ≥20; Soft-clipping on (-trim-mismatch-score=-3); Known exon-exon junctions (defined by RefSeq (downloaded March 7, 2011) and Gencode (version 3c)) and novel junctions (defined by GSNAP) were accepted. Although reads mapped to splicing junctions are exceedingly rare in GRO- and PRO-seq data, we included exon-exon junctions into the index and used BLAT to eliminate possible misalignment of spliced reads as described below. SNP

sites in the CEU population from Hapmap (release #28) and 1000 Genomes (pilot project) were included for SNP-tolerant alignments. Only reads that aligned to one genomic location (uniquely mapped reads) were used in further analyses. Read coverage was analyzed using using RSeQC (Wang et al., 2012). RPKM (read per kilobase per million reads) for each gene were calculated. The annotations of human gene are based on RefSeq (human NCBI36/hg18). 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. As described previously, the “first kilobase of each gene was omitted to better gauge the density of polymerase that actively elongates through the gene and to avoid over-counting from the increased density of paused polymerase in the 5’ end of the gene."(Core et al., 2008). In PRO-seq library, cDNA inserts > 70 bases were selected to maximize base coverage per read, thus many reads that come from promoter-proximally paused polymerases (20-60 nucleotides in length), were not sequenced in this study. In contrast, in GRO-seq library, cDNA inserts > 20 bases were selected, consistent with previous studies (Core et al., 2008). This explains why, in Figure 1B, the PRO-seq profile at the promoter appears to be less pronounced than the corresponding peak in the GRO-seq data set. These discrepancies do not affect our interpretation of results or conclusions for this current work regarding the timing of RDD incorporation into nascent RNAs. Average Phred quality scores of each base along RNA-seq reads from mRNA-seq, GRO-seq and PRO-seq are shown in Figure S3B. The quality scores of each library confirm that the increase in RDD around 55 nt is not a result of a loss of sequencing fidelity.

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 from consideration. To be included as RDD sites in the final lists, the following criteria had to be met: 1) a minimum of 10 total RNA-seq reads covering that site; 2) a minimum of 10 total DNA-seq reads covering that site; 3) DNA sequence at this site is 100% concordant, without any DNA-seq reads containing alternative alleles; 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. First, we removed all the sites that reside in repetitive genome regions annotated by RepeatMasker (version 3.2.7). Second, local sequences around each RDD site were aligned to the human reference genome to rule out misalignments to paralogous sequences or remaining pseudogenes. Specifically, for each RDD event, genomic sequences comprising sequences of length 25 nt, 50 nt, and 75 nt upstream and downstream of each site along with either the DNA variant or RNA variant were aligned to an index containing human reference genome (hg18) and sequences in hg19 but not present in hg18 using BLAT (Kent, 2002) (Stand-alone, v. 34x11). The settings '-stepSize=5' and 'repMatch=2253' were used to increase sensitivity. RDD events were removed if any of the 6 corresponding sequences aligned to another genomic location with ≤n mismatches (n=(read length + 2)/12–2) and with genomic sequences that explain the RDD call (that is if the genomic sequences match the RNA sequence). Lastly, to avoid potential misalignment of spliced reads in GSNAP

due to its high gap penalty algorithm, we re-aligned all the RNA-seq reads that contain putative RDD alleles using BLAT (Stand-alone, v. 34x11). Human genome sequences in hg19 that are not present in hg18 were included in our index in addition to sequences in hg18. Here, a low gap penalty was applied during BLAT alignment in order to compensate for high gap penalty of GSNAP alignment of spliced reads. Only RDD sites that were supported by both GSNAP and BLAT were retained for downstream analysis.

Prediction and analysis of R-loop forming sequences. Using the R-loop model developed by Kuznetsov and colleagues (Wongsurawat et al., 2012), we calculated R-loop scores for 2 kb of regions up and downstream of each RDD site. The average score at each position was smoothed using LOWESS method with 500 bp bandwidth (Cleveland, 1979). We compared these scores to those from random exonic sites in the genome. Exonic sites were picked because they have higher GC contents; we want to be sure that we did not randomly picked sites that happened to have lower GC contents which would lower R-loop scores. Despite this conservative comparison, RDD sites have significantly higher R-loop scores ($P<0.001$, t-test) than random sites.

Genome Walking. A GenomeWalker Human kit (Clontech, USA) was used to obtain upstream and downstream regions of human genomic DNA that flank an RDD site. These regions were the products of long range PCR employing RDD region-specific primers (designed to be within 100 nt or less of the RDD site) and the Clontech adaptor that is on each end of the DNA fragments in the GenomeWalker libraries. A primary PCR was followed by a secondary (nested) PCR using conditions recommended by the manufacturer. Ten to sixteen clones from each PCR were Sanger DNA sequenced and

results compared to the human genome sequence. Primers used in this assay are listed in Table S4.

ERCC RNA control library preparation. 49 plasmids containing 41 kb of ERCC control sequences (Table S5) flanked by a T7 promoter were obtained from Marc Salit. 1µg of plasmid was digested with either BamH1 or HindIII (Fermentas). The reactions were cleaned up using 96-well Minelute plates (Qiagen). RNAs were transcribed under standard conditions with T7 RNA polymerase, except that many reactions contained Br-UTP. DNA was degraded with DNAse I (Invitrogen), and reactions were cleaned up using a MEGA-Clear plate (Ambion). RNAs were quantified with a Qubit fluorimeter (Invitrogen), and several were analyzed on a bioanalyzer to verify full-length production. RNAs were grouped such that each group had 5-7 RNAs and each group had similar size and GC content distributions. Groups containing BrU-RNAs were serial diluted and pooled to give final amounts $1x10^8$, $2x10^7$, $1x10^7$, $4x10^6$, $2x10^6$, $8x10^5$, $1.6x10^5$, and $3.2x10^4$ copies per aliquot. An aliquot of each BrU-containing RNAs were used to generate libraries with protocol that is identical to the GRO-seq procedure from the point of the base hydrolysis step onwards. We obtained 172,497,660 sequenced bases (4,568X) from the ERCC samples (Table S5). At each of 37,765 unique sites, we compared the sequences to the template sequence. At each site, we calculated the "error rate" as the ratio of # reads with alternate base to the # total reads at the site; the result was 0.28%.

Droplet digital PCR. DNA probes specific for the DNA and RNA variant at RDD sites were synthesized and labeled by VIC and FAM, respectively (Applied Biosystems, USA). The PCR mixture was prepared using genomic DNA or cDNA (from GRO-seq

libraries or nuclear and cytoplasmic RNA fractions), gene-specific primers, VIC- and FAM- probes and Taqman reagents, and emulsion PCR was carried out following manufacturer's protocol (Bio-Rad Laboratories, USA). Fluorescent signal representing each variant was quantified utilizing QuantaLife Droplet Reader (Bio-Rad Laboratories, USA). Primers and probes used in this assay are listed in Table S6.

Supplemental Results and Discussion

Effect of duplicate reads.

It remains uncertain whether duplicate reads should be removed from analysis of RNA-seq data. Instead of making the decision one way or the other, we compared the gene expression, pausing index and RDD sites identified before and after removing the duplicate reads. The results were highly similar. The correlation coefficients for gene expression (RPKM) and pausing index before and after removal of duplicate reads are 0.91 and 0.87, respectively. Removal of identical reads has very little impact on RDD identifications; for instance, after removing identical reads, 23,024 RDDs of the 23,057 sites were still identified. The correlation coefficient of RDD levels before and after removing identical reads is 0.93. Given the effect of removing identical reads is so small and it is not clear how best to remove the identical reads, for current analyses, we kept them.

Probability-based error rate estimation. At a given single-nucleotide position, the base error probability was computed using quality scores for the reads mapped to the position. Given the set of reads supporting DNA-form allele S_0, and the set of reads supporting alternate allele S_1 in the same position, the base error probability

$P = \prod_{m \in S_1} p_m \prod_{n \in S_0} (1 - p_n)$ (Chepelev, 2012), where the *p* is the base-calling quality score generated from sequencing. Similarly, the mapping error probability was computed using mapping scores (from GSNAP) for the reads mapped to the position. Maximum P-value in a list of RDD sites was used to represent the P-value of the whole set. Bonferroni correction was performed to estimate the FDR (False Discovery Rate) of our RDD lists, which is essentially the FWER (Familywise error rate).

Experimental Validations of RDD Sites

Identification of RDDs relies on comparison of RNA and corresponding DNA sequences. Next-Gen sequencing (NGS) produces millions of short sequence reads that have to be assembled prior to sequence comparisons. Quality of the sequences, accuracy of the mapping and precision of sequence comparisons are critical in determining RDDs. Since NGS produces very short sequences (~100 nt/ read) relative to the sizes of the genomes and transcriptomes, for each sample, millions of reads representing DNA and RNA sequences have to be processed computationally. In this project, we inspect visually some of the RDDs but it is not possible to manually check all the sites. We use other computational tools such as alternate sequence alignment algorithms to check the data but we feel that the validation must include additional experimental steps. The following five sets of experiments and analyses examine the quality of the sequences, mapping accuracy and precision of the reverse transcriptase used for cDNA synthesis.

First, to confirm experimentally that the RDD sites are not false positive findings due to mis-mapping to highly similar sequences in the genome, we carried out “genome

walking" to isolate and map the flanking regions around the RDD sites (Siebert et al., 1995). PCR amplifications were performed with single-end sequence-specific primers that match the plus strand of genomic DNA near selected RDD sites and the resulting fragments were processed, cloned and sequenced (we then repeated the procedure with primers matching the minus strand near the RDD sites). We analyzed 10 sites; we found only the sequences corresponding to the regions of the RDD sites, no other regions were found. These data also confirmed the DNA sequences for those sites (Table 1, Table S4). Thus, the genome walking results showed that the RDDs cannot be explained by similar sequences in the genome.

Second, we compared the RDD sites to a comprehensive list of pseudogenes compiled by Chinnaiyan and colleagues (Kalyana-Sundaram et al., 2012). Among the 2,806 and 2,881 RDD sites in our two samples, only 11 and 5 of them overlap with pseudogenes, respectively. Most of these 16 RDD sites cannot be explained by the pseudogenes, because the pseudogenes and their "parent" genes have the same DNA sequences at the sites corresponding to the RDDs and they differ from the RNA sequences at those sites. Regardless, the small number of RDDs that overlap with pseudogenes and results of our genome walking confirm that the RDDs are located in unique regions of the human genome.

Third, we estimated the error rate for the Illumina-based sequencing used in this study to ensure that it cannot contribute considerably to the RDD findings. Sequencing errors are influenced by sequence features and therefore are not entirely random. The best match for sequence characteristics of our RDD sites is the corresponding DNA sequences. Our criteria for selecting sites to compare DNA and RNA sequences

require that the reads for DNA sequences to contain only a single nucleotide type (all A, C, G or T), so we already selected for sites with “no” errors in the corresponding DNA. To get an assessment of errors in nearby regions, we analyzed the DNA sequences within 50 nucleotides upstream and downstream of the RDD sites. We removed the SNP sites for the two individuals; then for the remaining sequences which should be monomorphic, at each site, we counted the number of reads that contain an alternate allele and calculated error rate as the ratio of the number of reads containing an alternate allele to the total number of reads at that site (the average coverage for the DNA samples are 30X and 60X). Most of the sites have a single nucleotide type (therefore no errors); for those sites (<8%) that have an alternate allele, the majority has only one read with an alternate allele (to be identified as a RDD, the alternate allele has to be represented by ≥ 2 unique reads). The error rate is very low (median <<0.01%). There are significantly ($P<0.0001$) fewer sequencing errors than RDDs. In addition, we assessed sequencing errors using the method developed by Chepelev; it also showed that the likelihoods that the RDDs are results of sequencing error are very low (Bonferroni corrected $P<0.001$, see Table S7) (Chepelev, 2012). Together, the data show that given their random patterns and low frequencies, sequencing errors cannot contribute considerably to RDDs.

Fourth, to assess errors from reverse transcription and Illumina-based sequencing, we carried out in vitro transcription with BrUTP and T7 polymerase using the reference samples from the External RNA Control Consortium (ERCC) (Baker et al., 2005) followed by reverse transcription and deep sequencing. The samples were sequenced very deeply (>4,500 fold coverage) to ensure a good sampling. The results contained

relatively few discrepancies (0.28%) from the expected sequences. We also compared the RNA sequences from these ERCC controls with the underlying DNA sequences for RDDs using the same criteria as those we used on our samples. Even with a much higher coverage in the ERCC than our experimental samples (>4,500X vs. 30X), we found only 19 sites that met our inclusion thresholds (at least 10 read coverage and 10% level). The RNA synthesis of the ERCC sample was carried out using T7 RNA polymerase which is more error prone than the RNA polymerase II in human cells (Huang et al., 2000). Despite this difference, the error rate is low. Thus, misincorporation by reverse transcriptase can be eliminated as the source of RDDs, since the combined errors of T7 polymerase and reverse transcriptase are not sufficient or consistent enough to produce appreciable levels of RDDs above the 10% threshold.

Fifth, the most direct way to assess errors from Illumina-based sequencing and mapping of the resulting reads is to use an alternate technology to analyze the DNA and RNA sequences. Other groups have used Sanger sequencing and droplet digital PCR to assess the false discovery rate of next generation sequencing-based identification of RNA editing and RDD events (Chen et al., 2012; Peng et al., 2012); here we adopted the same method to assess the false discovery rate. We carried out droplet digital PCR to compare the genomic DNA and RNA sequences at 15 randomly chosen RDDs identified in our GRO-seq samples. The results validated 14 of the 15 RDD sites in two samples (Table 2 and Figure S2), which correspond to a false discovery rate of 7%. In addition, we used RNA from nuclear fractions to assess RDDs. Nuclear RNA is a mixture of transcripts in different stages of processing and our PCR primers used here capture only unspliced immature transcripts thus only a few of the sites were expected

to be assayed. Nevertheless when we assayed for RDDs in the nuclear RNA samples, we found five of the 14 RDD sites (Table 2). The remaining 9 did not generate amplicons for sequence analysis. Four of the RDD sites (G-to-A in *DENND4B*; G-to-T in *MAP3K4*; T-to-C in *ARAP1* and *RAB12*) were further found in cytoplasmic RNA (Figure S2); this demonstrates that the RDD bearing transcripts can be processed into mature mRNA. To ensure that these RDDs are found in other cell types besides cultured B-cells, we analyzed DNA and RNA samples from foreskins of three newborns; and brain cortex, liver and muscles from autopsies of three accident victims. In all of these samples, the four RDD sites were found. The data also showed differences in RDD levels across cell types (Figure S2). While Sanger sequencing and droplet digital PCR allow us to check the RDD sites using alternate sequencing platforms, they are not the most efficient way to examine a large number of RDD sites. As another way to validate the RDDs, we isolated nascent RNA by preparing them from chromatin fractions (Wuarin and Schibler, 1994). The samples from both individuals were prepared; they showed enrichment of U6 expression (~10 fold) and depletion of transcript expression for the ribosomal protein S14 (~4 fold) compared to cytoplasmic RNA; thus we were confident that we had reliably isolated nascent RNAs. Then we looked for the RDDs that we had identified in PRO-seq and GRO-seq, and found 1,007 of the same RDD sites in the chromatin-bound nascent transcripts. These results show that RDD sites are found in nascent transcripts regardless of how they were isolated.

Together, these five analyses show that the RDDs are distinct from sequencing errors and they are found in unique regions of the genome and in different types of human cells.

SUPPLEMENTAL REFERENCES

Baker, S.C., Bauer, S.R., Beyer, R.P., Brenton, J.D., Bromley, B., Burrill, J., Causton, H., Conley, M.P., Elespuru, R., Fero, M., et al. (2005). The External RNA Controls Consortium: a progress report. Nat. Methods *2*, 731–734.

Chen, R., Mias, G.I., Li-Pook-Than, J., Jiang, L., Lam, H.Y.K., Chen, R., Miriami, E., Karczewski, K.J., Hariharan, M., Dewey, F.E., et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell *148*, 1293–1307.

Chepelev, I. (2012). Detection of RNA editing events in human cells using high-throughput sequencing. Methods Mol. Biol. Clifton NJ *815*, 91–102.

Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. J. Am. Stat. Assoc. *74*, 829–836.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science *322*, 1845 – 1848.

Huang, J., Brieba, L.G., and Sousa, R. (2000). Misincorporation by wild-type and mutant T7 RNA polymerases: identification of interactions that reduce misincorporation rates by stabilizing the catalytically incompetent open conformation. Biochemistry (Mosc.) *39*, 11571–11580.

Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D.R., Wu, Y.-M., Cao, X., Asangani, I.A., Kothari, V., Prensner, J.R., Lonigro, R.J., et al. (2012). Expressed Pseudogenes in the Transcriptional Landscape of Human Cancers. Cell *149*, 1622–1634.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res. *12*, 656–664.

Pandya-Jones, A., and Black, D.L. (2009). Co-transcriptional splicing of constitutive and alternative exons. RNA N. Y. N *15*, 1896–1908.

Peng, Z., Cheng, Y., Tan, B.C.-M., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X., et al. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat. Biotechnol. *30*, 253–260.

Siebert, P.D., Chenchik, A., Kellogg, D.E., Lukyanov, K.A., and Lukyanov, S.A. (1995). An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Res. *23*, 1087–1088.

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinforma. Oxf. Engl. *28*, 2184–2185.

Wongsurawat, T., Jenjaroenpun, P., Kwoh, C.K., and Kuznetsov, V. (2012). Quantitative model of R-loop forming structures reveals a novel level of RNA–DNA interactome complexity. Nucleic Acids Res. *40*, e16–e16.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinforma. Oxf. Engl. *26*, 873–881.

Wuarin, J., and Schibler, U. (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. Mol. Cell. Biol. *14*, 7219–7225.