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

Isabel X. Wang, Leighton J. Core, Hojoong Kwak, Lauren Brady, Alan Bruzel, Lee McDaniel, Allison L. Richards, Ming Wu, Christopher Grunseich, John T. Lis, Vivian G. Cheung

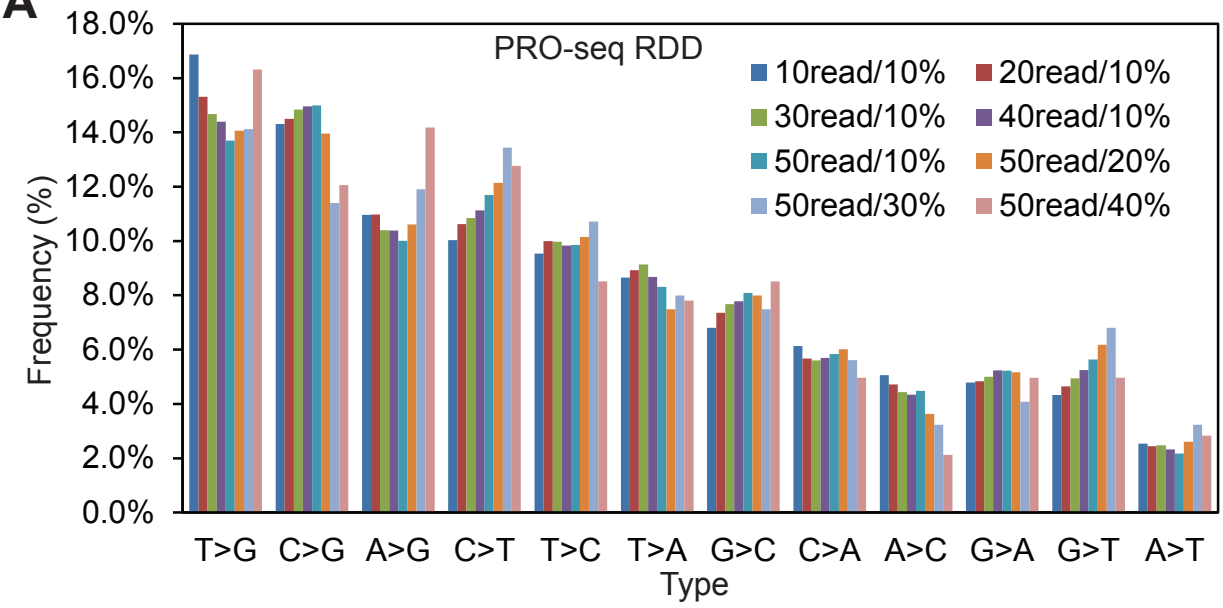
Figure S1. Distribution of RDD types remains similar when thresholds of RNA-seq coverage and RDD level are raised. Related to Figure 2. (**A**) Frequency of each RDD type in PRO-seq (upper panel) and GRO-seq (lower panel) at different thresholds. (**B**) Number of RDDs detected at different thresholds shown in (**A**). Distribution of RDD types in GRO-seq is also similar at different thresholds (data not shown).

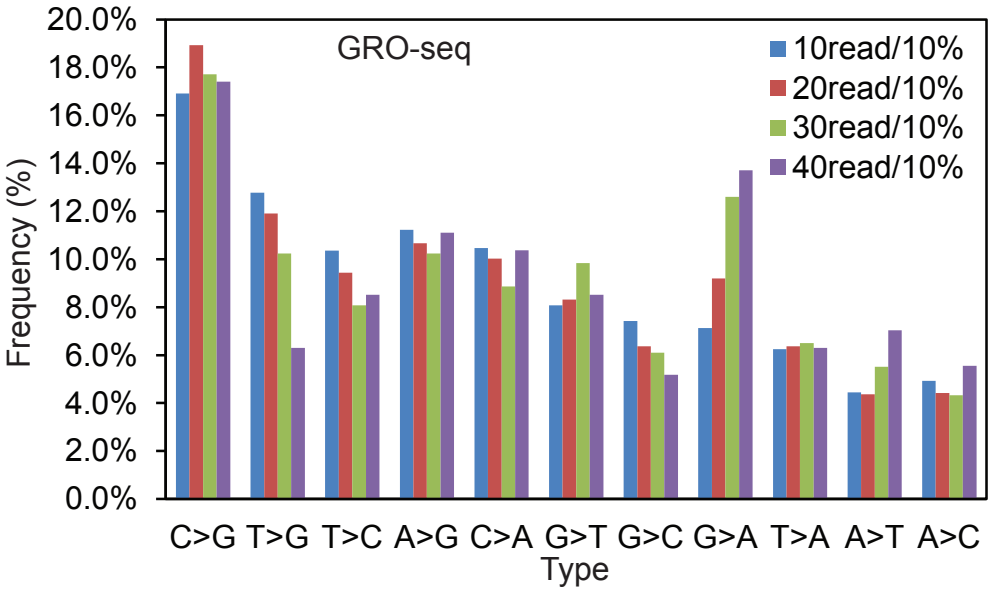
Figure S2. Validation of RNA-DNA differences. Related to Figure 3 and 4. (**A**) RDD levels differ among different tissue types and different individuals. Heatmap shows RDD levels measured by droplet digital PCR. Each row represents RDD levels measured in one tissue from a given individual. (**B**) RDD in *DENND4B* is validated by digital droplet PCR in both GRO-seq, nuclear and cytoplasmic RNAs. Genomic DNA, global run-on cDNA, nuclear cDNA and cytoplasmic cDNA were assayed for G-to-A RDD in *DENND4B* (chr1:152175284). An assay specific for unspliced transcripts was used for genomic DNA and global run-on cDNA while an assay specific for spliced transcripts was used for nuclear and cytoplasmic cDNA. Droplets with only the A allele appear in the top left quadrant; those with only the G allele are in the bottom right quadrant. The top right quadrant contains droplets with transcripts containing both the A and G allele while the bottom left quadrant shows droplets with no PCR amplification.

Figure S3. Replication of RDDs in very nascent transcripts. Related to Figure 4. (**A**) A second PRO-seq experiment replicated distribution of RDDs. RDDs occur after transcripts exit Pol II active site. (**B**) The average Phred quality score of each base along reads in mRNA-, GRO- and PRO-seq. (**C**) ADAR mediated A-to-G editing sites are present in mature mRNA and chromatin-bound RNA, but not in nascent RNA identified in GRO-seq or PRO-seq. Shown here is IGViewer screenshot of part of *POLH* gene.

Figure S4. Features of sequences surrounding RDDs in very nascent transcripts. Related to Figure 3. (**A**) Sequences near A-to-G sites detected in very nascent transcripts differ from those detected in mRNA. A-to-G sites in mRNA were obtained from a recent study (Wang et al., 2013). For each type of RDDs, 10 nucleotides upstream and downstream of RDD sites were analyzed. At each position, the average percentage of each nucleotide is reported. (**B**) Features of sequences surrounding C-to-X and G-to-Y (X= A or G; Y = A, C, or T) RDD sites in very nascent RNAs. (**C**) To compare the sequence context around RDD sites with controls, we analyzed the nucleotide composition 10 nucleotides upstream and downstream of the RDD sites and those in three sets of “far” sites which are 50, 100 and 1,000 bases away from the RDDs. Asterisks denote p-value <0.05, T-test. Results for GM12004 and GM12750 are similar; data for GM12004 are shown. (**D**) Examples of sequences surrounding C-to-X and G-to-Y RDDs.

A

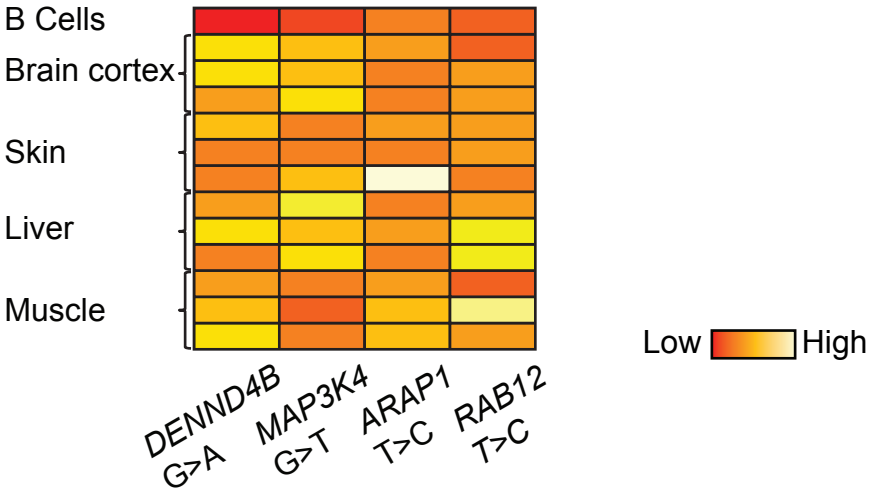




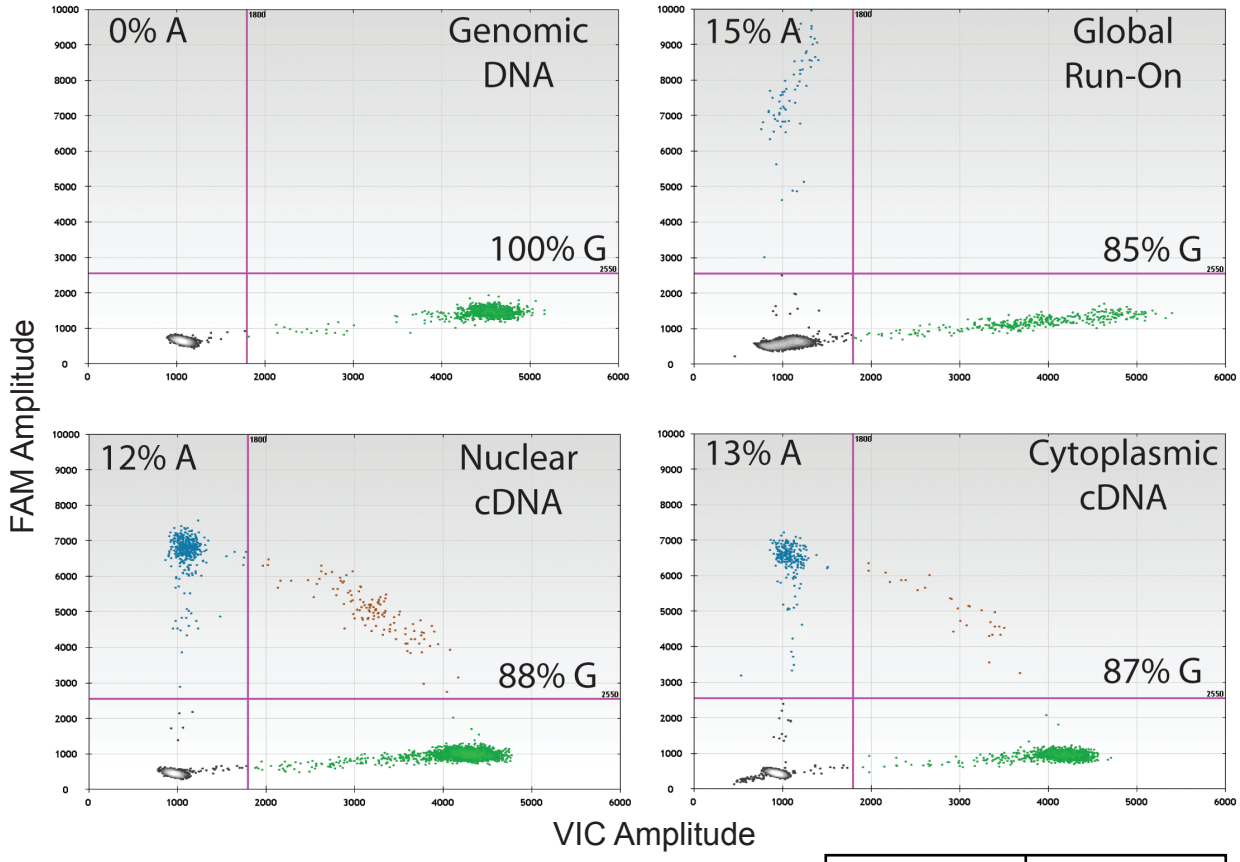
B

Threshold		# RDD in PRO-seq
RNA-seq coverage	**RDD level**	
10 reads	≥ 10%	23,093
20 reads	≥ 10%	14,642
30 reads	≥ 10%	9,560
40 reads	≥ 10%	6,778
50 reads	≥ 10%	4,774
50 reads	≥ 20%	1,738
50 reads	≥ 30%	571
50 reads	≥ 40%	136

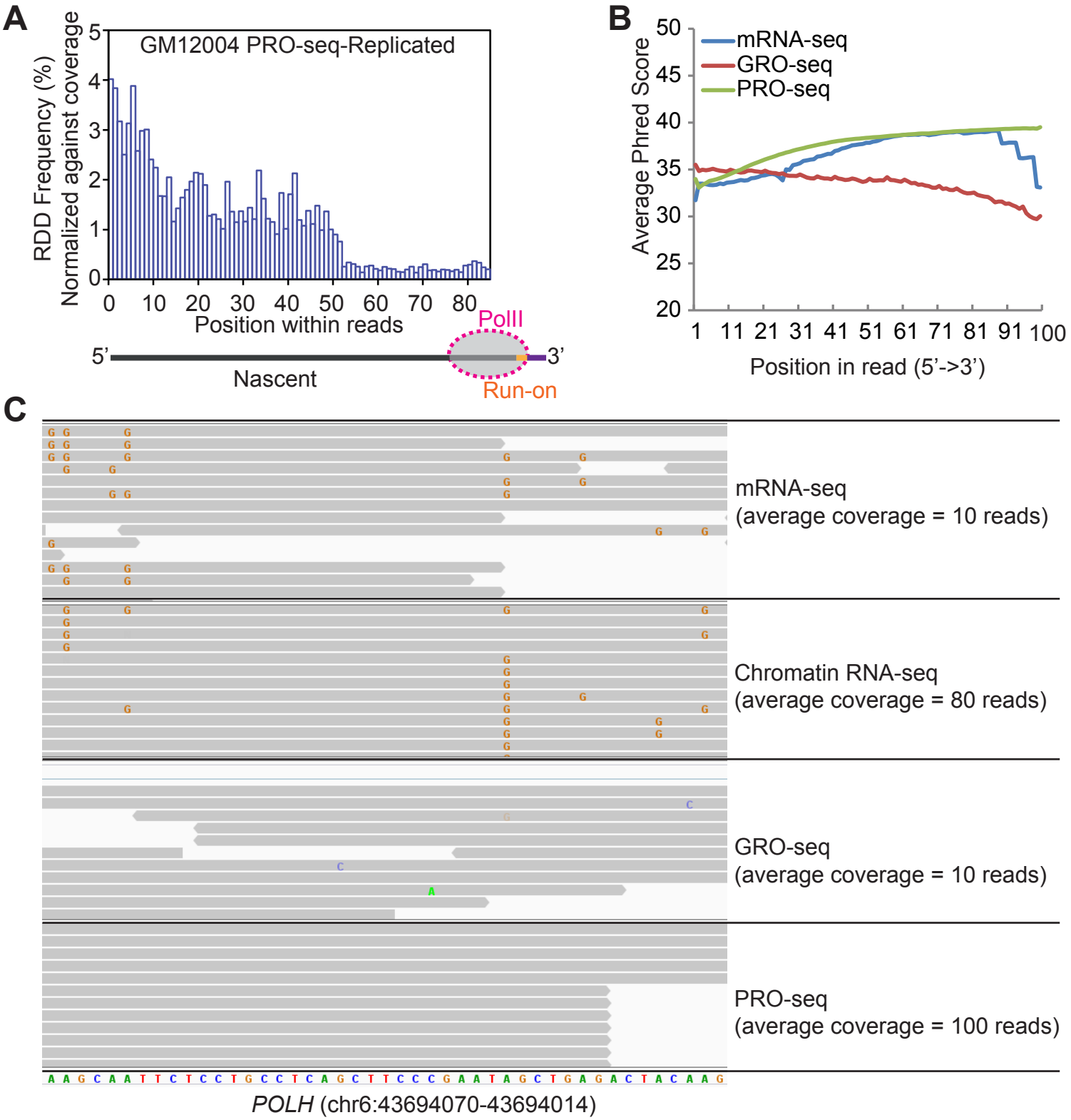
A



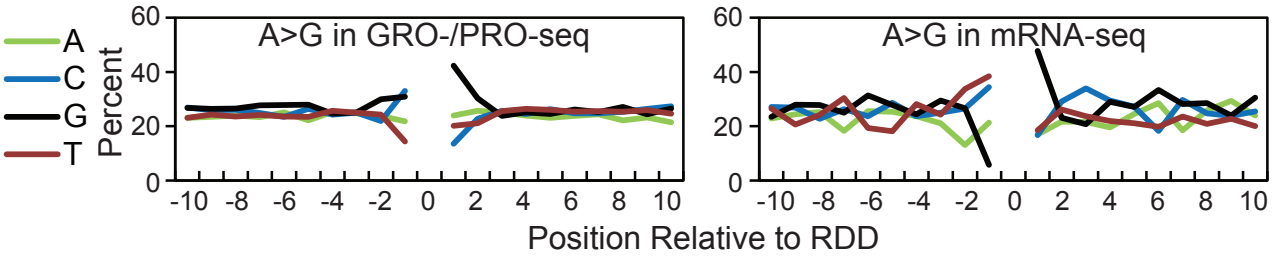
B



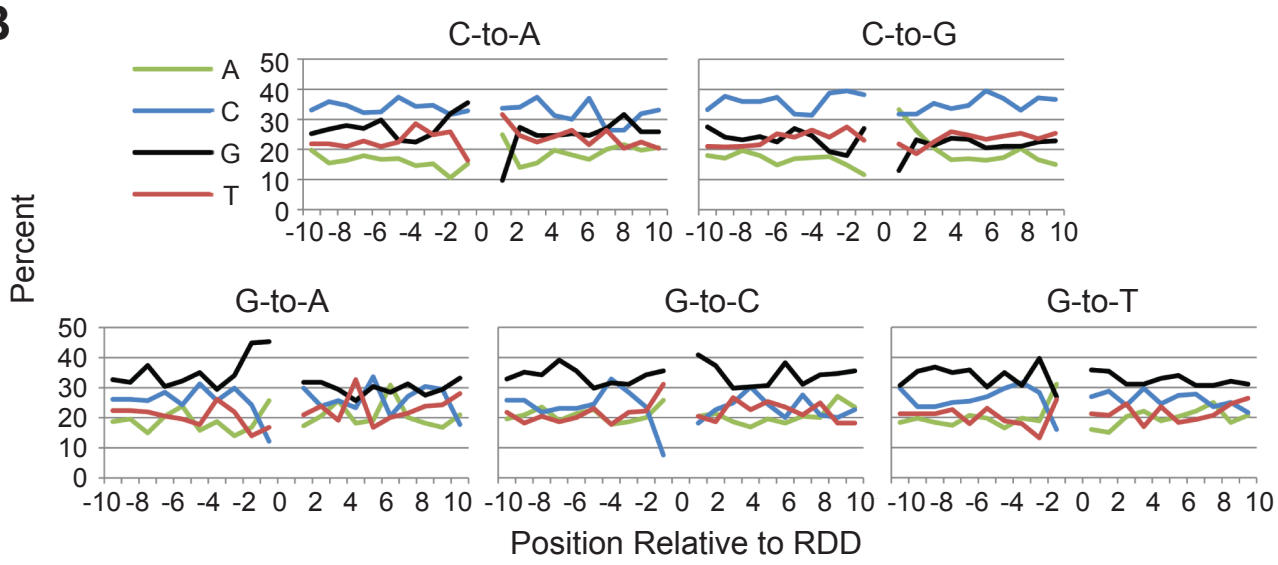
A (RDD allele)	Both
Negative control	G (Ref. allele)


A
GM12004 PRO-seq-Replicated
RDD Frequency (%)
Normalized against coverage
0
1
2
3
4
5
0 10 20 30 40 50 60 70 80
Position within reads
PolII
5'
3'
Nascent
Run-on
B
mRNA-seq
GRO-seq
PRO-seq
Average Phred Score
20
25
30
35
40
45
50
1 11 21 31 41 51 61 71 81 91 100
Position in read (5'->3')
C
mRNA-seq
(average coverage = 10 reads)
Chromatin RNA-seq
(average coverage = 80 reads)
GRO-seq
(average coverage = 10 reads)
PRO-seq
(average coverage = 100 reads)
POLH (chr6:43694070-43694014)

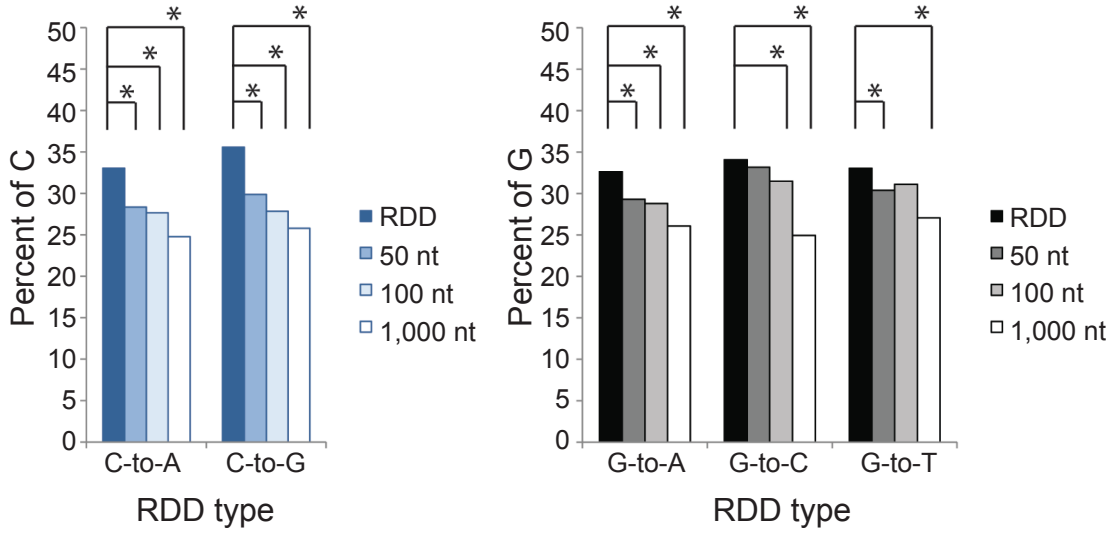
A



B



C



D

Chr	Position	Strand	Type	Gene Name	Sequence
1	152175284	-	G>A	*DENND4B*	GCCCAGGTGTGTTACCGGGTA
8	97384747	+	C>G	*PTDSS1*	TCTCCCTGCTCCCACCCACTA
10	97616031	+	C>G	*ENTPD1*	CCACCTATGTCTTCCTCATGG
17	39757073	-	G>T	*SLC25A39*	GTGGACACGTGCGAGGTGGAC

Table S1 RNA-DNA differences identified in GRO-seq data from GM12004. Related to Figure 2.

Table S2 RNA-DNA differences identified in GRO-seq data from GM12750. Related to Figure 2.

Table S3 RNA-DNA differences identified in PRO-seq data from GM12004. Related to Figure 2.

Table S4. Primers used in genome walking. Related to experimental procedures.

RDD		primers that amplify minus strand		primers that amplify plus strand	
location	type	primary	secondary (nested)	primary	secondary (nested)
chr1:152175284	G-to-A	5'-ATCTCCAGCATGACCCGCACAGACAGCACA-3'	5'-CCCATAGTGTGAGCAGAGCTGCATCAGTAC-3'	5'-GGTGCTTGCTGAGACTTTGGTTTTGGTTGA-3'	5'-GTGAGTCTCCCATTGTCCTGTGCCTCTCCT-3'
chr6:107088915	T-to-A	5'-GTCACAGCTTACAATTTATGACAGGTTTGT-3'	5'-GGTGAGCATCTTCTTTTCTGGCTGTTCATT-3'	5'-TGTCTAAGCTCACCAACCAAGTTCAACGAG-3'	5'-ACTTTGAATTAACATCTGAGGCAGTGTTCA-3'
chr9:34336911	G-to-A	5'-ATTCTTTGCAGGAATAACTGACAGTAGGGT-3'	5'-GACTAAAGCATGTTATGGGAGGTGGATGAG-3'	5'-GCAAAGTGGTGATGTTGTCAAAGCCTACTC-3'	5'-CTGCCACATCCTAATCTCAGCTCCTCTCTC-3'
chr11:72079055	T-to-C	5'-CCTCCGATTCCTTGGGGTCTGATGTGCATG-3'	5'-ATGTGTGTGATGTAAGGGTGGCTGAGTCAG-3'	5'-TGGGGGATTCATTCTCAAAAGGAGATTGCA-3'	5'-CATCCTTCCCTCCCAAGACCCACTTCTGCA-3'
chr12:100980077	A-to-C	5'-GAAGAAAGCTTCTCCGCTCACTAATCACTC-3'	5'-CCTCCTCGCTCGGCATTCTGGGATTGGTAG-3'	5'-GATTTCAGTGCTCTAACTGCATCCTTGCAC-3'	5'-CGCCACCCCGAAATTGACGTCACAATCTGG-3'
chr14:20221257	G-to-T	5'-AAAGAGATTTCTTTCTTTCCCGTCTCTGGT-3'	5'-TAAGGATGTCTCGAGTAGGGATTTTTAGGA-3'	5'-CACATTAAGTACTACAGTCCTCCGCTCTAC-3'	5'-CAGCTGAGCTATCGAAGGAAGCACGGAAAC-3'
chr16:2140620	A-to-G	5'-TTAATTAGAGCCAGTTAGAGGGAGGATGGA-3'	5'-GGAAGGGCTGTTCCAGGGACAGGAAAAGGG-3'	5'-CTCGGGCACTTCAGTCGTCACTCCATCCCA-3'	5'-GGACTTGAGGGGACTTGGAACCCCCAGACT-3'
chr19:2427122	C-to-A	5'-CCAATAGGAGGGGCGAATGACTCCACTGAG-3'	5'-CACGCCCAATGTTCAAGTCTATAAAAGTCG-3'	5'-GTCCTTGGCGAGGATAATCCAGGAAGTTGC-3'	5'-ACCAACGGTAGTCCGACGCTTCGGCGATCT-3'
chr22:42867269	T-to-C	5'-CTTGCGTTCTTATCAGTAGGCCTTTACCTG-3'	5'-GCTTTCTGGCTGGCTCTGTCCCTCATCCTC-3'	5'-CAAACAAGAACACAGAGAGAGCCATGAGAT-3'	5'-GGGTCACTCAGAACACAGTTTCACAAGCAG-3'
chrX:7004437	G-to-T	5'-ACAAACCACAAAAGTCCTCACATGCAGGAA-3'	5'-GCAGGCATTTGCAGAGACTCCCCGACGGAT-3'	5'-AGTCCTTGGCCTGAACGATCTGCAGGTACA-3'	5'-TTCCAGAATCACAGGAAGTCACCCTTGAAA-3'

Table S5 Reference sequence of ERCC libraries used in this study. Related to experimental procedures.

Table S6. Sequences of primers and reporters for droplet digital PCR. Related to experimental procedures.

Genomic Location	Gene Name	RDD type	Forward Primer	Reverse Primer	VIC Probe Sequence	FAM Probe Sequence
1:152175284	*DENND4B*	G>A	5'-ATTGTCCTGTGCCTCTCCTG-3'	5'-GTGAGCAGAGCTGCATCAGTA-3'	5'-CCAGGTGTGTTACCG-3'	5'-CCAGGTGTATTACCG-3'
1:224242710	*C1orf55*	T>C	5'-CCTTCTGCAGTCTCTAGTCCCT-3'	5'-GCTGTTAAACCTTTCTGTTGCTTCT-3'	5'-CATGCCCGACCTGTC-3'	5'-CATGCCCAACCTGTC-3'
3:197100758	*TNK2*	G>T	5'-CGTGTAGGGACAACAGAGT-3'	5'-CCAGGAGTTGCCAGTGATGT-3'	5'-TGAAGCGTCCGCAGAAG-3'	5'-CTGAAGCGTCCTCAGAAG-3'
6:161450890	*MAP3K4*	G>T	5'-TAGGCTGTATTCAAAGCAGATCAC-3'	5'-GGTGGACTATGCCATGCTCATG-3'	5'-AGGACGTTGATAGCAATG-3'	5'-AGGACGTTGATCGCAATG-3'
6:37987903	*ZFAND3*	G>C	5'-GTGGCTAAGAAGATAGACTGACC-3'	5'-GATGAGCCACAGGTAATTGTTCTC-3'	5'-TAGGACTATCAAGATCCTC-3'	5'-TAGGACTATCAACATCCTC-3'
9:34336911	*(intergenic)*	G>A	5'-GGTGGATGTTATGGGAGGTGGAT-3'	5'-CACCTGCCACATCCTAATCTCA-3'	5'-CCTCTCTCTCTCACCCTC-3'	5'-CTCTCTCTCCCACCCTC-3'
11:58103493	*ZFP91*	G>C	5'-AGCCGCGTGCTGAGG-3'	5'-CTCCGGCGGGACACA-3'	5'-TCGGTGCCGACCTC-3'	5'-TCGGTCCCGACCTC-3'
11:72079055	*ARAP1*	T>C	5'-GTGTGATGTAAGGGTGGCTGA-3'	5'-GACATCCTTCCCTCCCAAGAC-3'	5'-CTTCTGCAGACTTTCTGAC-3'	5'-CTGCAGACCTTCTGAC-3'
12:100980077	(intergenic)	A>C	5'-CTCGGCATTCTGGGATTGGT-3'	5'-CCACCCCGAAATTGACGTC-3'	5'-CAATCTGGACTAAAACT-3'	5'-TCTGGACGAAAACT-3'
16:69880869	*FTSJD1*	C>G	5'-GCCTAGGCGAGGCGAGA-3'	5'-CACGTGCGCCAACGT-3'	5'-TACGCACGCCGGCGAC-3'	5'-TACGCACCCCGGCGAC-3'
17:30949447	*AP2B1*	G>C	5'-GGAGGCTGTGAAGAAAGTGATTG-3'	5'-TTACTCTTACCTAACATCCTTCC-3'	5'-CCACGGTGATAGCAG-3'	5'-CCACGGTCATAGCAG-3'
18:8628755	*RAB12*	T>C	5'-CCAGTGGTAGCATCTGAAGT-3'	5'-CATGAATGGAGCTTTGCTTCGT-3'	5'-ATTGGAGTGTTTCTGCC-3'	5'-TTGGAGTGTCTCTGCC-3'
19:2197783	*SF3A2*	G>C	5'-ACCAAGCAGAGAGACTCGGA-3'	5'-TCCCTGATCTCACCTGGAAGAG-3'	5'-ATGGGCCAGCAGAGC-3'	5'-ATGGGCCACCAGAGC-3'
X:7004437	*HDHD1A*	G>T	5'-AAGTCACCCTTGAAAGGG-3'	5'-GACTCCCCGACGGATGC-3'	5'-TGGCGATGTAAGTGTAAG-3'	5'-TGGCGATGTAATTGTAAG-3'

Table S7. Analysis of mapping error and sequencing error. Related to Figure 2.

Sample	Individual	Mapping Error (p-value)	Mapping Error (FDR)	Sequencing Error (p-value)	Sequencing Error (FDR)
GRO-seq	GM12004	0.0002	0.0007	0.0005	0.0029
GRO-seq	GM12750	0.0001	0.0005	0.0016	0.004
PRO-seq	GM12004	0.0002	0.0012	0.0063	0.0261