Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”

Joseph K. Pickrell[1], Yoav Gilad[1], Jonathan K. Pritchard[1,2]

[1]Department of Human Genetics and

[2]Howard Hughes Medical Institute

University of Chicago, 920 E. 58th St., CLSC 507, Chicago, IL 60637, USA

September 27, 2011

Abstract

Li et al. [1] reported over ten thousand mismatches between mRNA and DNA sequences from the same individuals, which they attributed to previously unrecognized mechanisms of gene regulation. We found that at least 88% of these sequence mismatches can likely be explained by technical artifacts such as errors in mapping sequencing reads to a reference genome, sequencing errors, and genetic variation.

Li et al.[1] sequenced cDNA from lymphoblastoid cell lines derived from 27 individuals whose genomes have been sequenced at low coverage [2], and identified 10,210 sites of mismatches between an individual's mRNA and DNA sequences (RDD sites, for RNA-DNA difference). RDD sites included all possible combinations of sequence mismatches, and the authors validated a subset of these mismatches by additional assays. These observations were interpreted as evidence for novel mechanisms of gene regulation, analogous perhaps to A→I RNA editing [3].

An alternative explanation is that some RDD sites are technical artifacts due to errors in mapping sequencing reads to a reference genome or systematic sequencing errors. To evaluate this possibility, we examined the sequence alignments used to call RDD sites (Supplementary Material). Visualizing these alignments revealed a number of anomalies. For example, at the RDD site presented in Figure 1A, all mismatches to the genome occur at the last base of reads aligned to the negative DNA strand. No such anomalies are seen in alignments around a positive control site (Figure 1B). The biases in the first example are consistent with several known issues that cause spurious differences between Illumina sequencing reads and a reference genome; these include read-mapping errors between paralogous genomic regions and around insertions and deletions [2; 4], as well as position and strand biases in the error rate of Illumina sequencing [5–7].

We asked whether the patterns seen in Figure 1A are typical among RDD sites. Indeed, mismatches to the genome at RDD sites are dramatically enriched at the ends of RNA sequencing reads; this contrasts with reads that match the genome at these sites (Figure 1C). This pattern is evidence that many of the RDD sites are false positives due to mapping or sequencing errors.

To quantify what fraction of RDD sites may be false positives, we used metrics developed in for calling single nucleotide polymorphisms (SNPs) from Illumina sequencing data. In this context, it is known that a search for mismatches between aligned reads and a genome will result in large numbers of false positive SNPs, many of which can be filtered out based on various criteria [2; 4; 8; 9]. We used two criteria based on comparing, at each RDD site, the alignments of RNA sequencing reads that match the genome with the alignments of reads that mismatch the genome–a test for *position bias* and a test for *strand bias* (Supplementary Information). These tests provide quantitative measures for the intuition that there should be no systematic differences in strand or start position between alignments of reads covering the two alternative genotypes at a site, and are similar to tests implemented in SNP-calling packages [4; 9].

In Figure 1D, we show the histogram of p-values for the position bias test for the 7,812 RDD sites with at least five reads supporting both bases. There is a clear skew towards low p-values, indicating pervasive technical artifacts. At a p-value threshold of 0.01, 87% of these RDD sites fail either the strand bias test or the position bias test (at a p-value threshold of 0.05, the corresponding number is 93%). To test the specificity of these filters, we compared the reported RDD sites to a database of known A→I RNA editing sites [10]. There are 23 sites in common between the two data sets; of these, 21 (91%) pass both of the filters. This indicates that we are largely only removing false positives.

Genetic variation is another source of false positives; an additional 1% of the putative RDD sites appear instead to be known genetic variants in these individuals (Supplementary Material). In total, we estimate that at least 88% (at a p-value threshold of 0.01) to 94% (at a p-value threshold of 0.05) of the RDD sites are likely false positives. This is probably an underestimate of the true false positive rate, since some false positive sites will pass the bias tests by chance and there are additional, unannotated SNPs in the genome.

Given the above results, we re-examined the validation experiments done by Li et al.[1]. These experiments are of two types. First, at 11 sites, the authors confirmed that the RDD event was absent from genomic DNA but present in cDNA by Sanger sequencing. At six of these 11 sites, the event is of the type A→G, and four of these six are present in a database of known A→I RNA editing sites [10]; these are likely true positives. Of the remaining five sites, three fall in a single gene–*HLA-DQB2*–that is copy number variable in these individuals [11], and one–in the gene *DPP7*–overlaps a known SNP (at which the reported RDD type matches the known alleles) [2]. We suggest that the authors have detected genetic variation rather than RNA-DNA differences at these sites. In sum, these experiments identify two previously unknown sites of A→I RNA editing, and provide evidence for a single G→A event.

The second validation experiment involved identifying peptides corresponding to RDD events. In their Table 3, Li et al.[1] provide 17 examples where both the "DNA form" (the unaltered version) and the "RNA form" (the modified version) of peptides were detected via mass spectrometry. All but one of these sites fail the bias tests described above. We propose that the "RNA forms" of these peptides are in most cases normal forms produced by paralogous genes. Indeed, examination of the "RNA forms" revealed that seven match both the reported protein and additional proteins equally well, and four of the remaining 10 match other proteins (in addition to the reported protein) with a single additional mismatch (Table 1; Supplementary Material). It cannot be ruled out that the "RNA forms" of these proteins are instead normal forms caused by genetic variation in their paralogs. An additional possibility is that some "RNA forms" result from sequencing errors in the peptides.

In summary, we estimate that a minimum of 88-94% of the RDD sites identified by Li et al.[1] are false positives due to mapping errors, sequencing errors, and genetic variation. It is possible that the remainder of RDD sites contain examples of novel mechanisms of gene regulation.

References

[1] M. Li, *et al.*, *Science* (2011).

[2] 1000 Genomes Project Consortium, *et al.*, *Nature* **467**, 1061 (2010).

[3] B. L. Bass, *Annu Rev Biochem* **71**, 817 (2002).

[4] M. A. Depristo, *et al.*, *Nat Genet* **43**, 491 (2011).

[5] K. Nakamura, *et al.*, *Nucleic Acids Res* (2011).

[6] Y. Erlich, P. P. Mitra, M. de la Bastide, W. R. McCombie, G. J. Hannon, *Nat Methods* **5**, 679 (2008).

[7] F. Meacham, *et al.*, *Nature Precedings* (2011).

[8] H. Li, J. Ruan, R. Durbin, *Genome Res* **18**, 1851 (2008).

[9] H. Li, *et al.*, *Bioinformatics* **25**, 2078 (2009).

[10] A. Kiran, P. V. Baranov, *Bioinformatics* **26**, 1772 (2010).

[11] D. F. Conrad, *et al.*, *Nature* **464**, 704 (2010).

[12] J. B. Li, *et al.*, *Science* **324**, 1210 (2009).

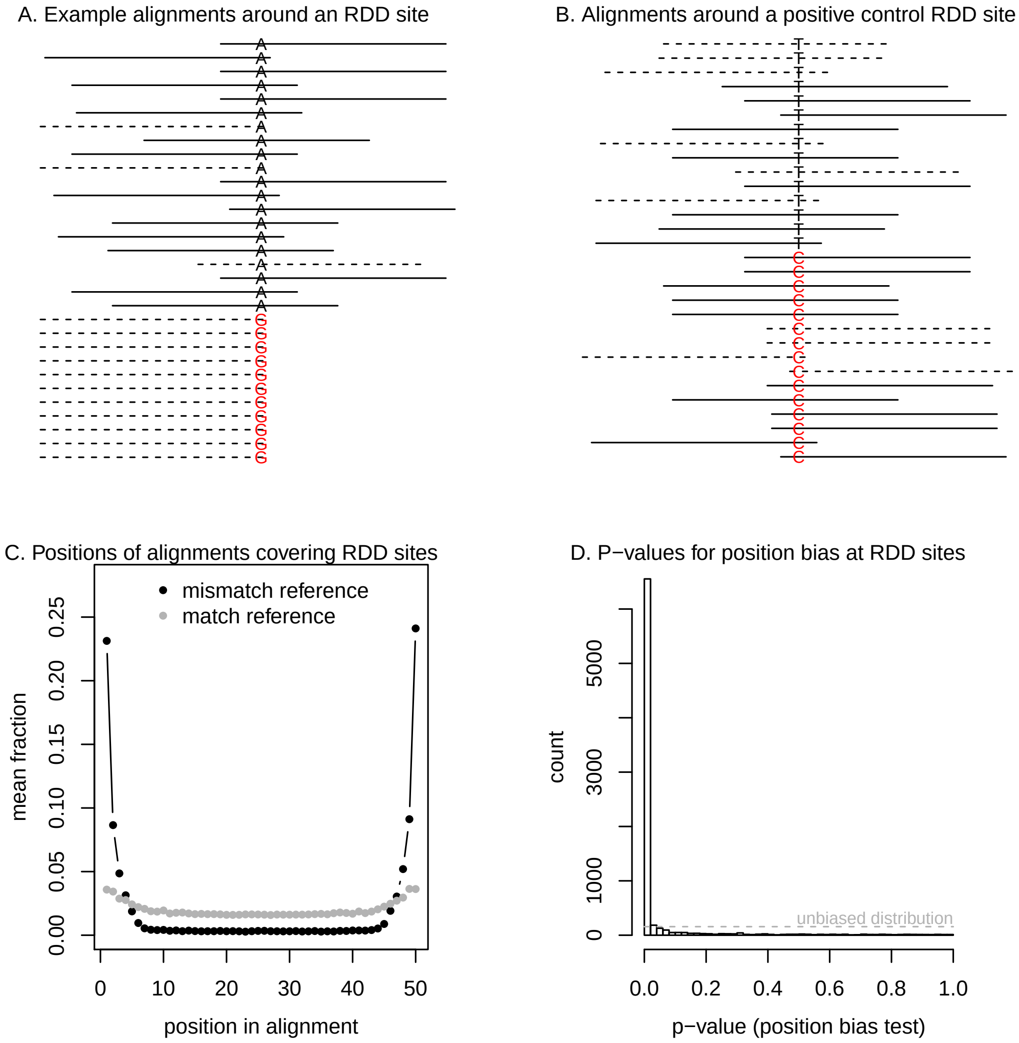

Figure 1:**Identifying false positive RDD calls. A. RNA-seq read alignments around an RDD call from Li et al. (2011).** Plotted are the positions of read alignments to the genome surrounding the RDD site at chromosome 11, position 105,473,792. The solid lines show sequencing reads aligning to the (+) strand of the genome, and dotted lines are alignments to the (-) strand of the genome. At the center of the plot is the base corresponding to the RDD site; the reference base is in black, and the non-reference base is in red, and both are labeled with respect to the (+) DNA strand. Alignments have been organized such that the mismatches to the genome are at the bottom of the figure. For plotting, we randomly sampled 20 alignments that match the genome at the RDD site; all 11 alignments that mismatch the genome are shown. **B. Read alignments around a positive control RDD site.** Plotted are the positions of read alignments to the genome surrounding the known A→I editing site in *AZIN1* [12] (on the forward strand this site appears as T→C). The format is the same as in **A**. For plotting, we randomly sampled 15 alignments that match the genome at the RDD site, and 15 alignments that do not match the genome at the site. **C. Position biases in alignments around RDD sites.** For each RDD site with at least five reads mismatching the genome, we calculated the fraction of reads with the mismatch (or the match) at each position in the alignment of the RNA-seq read to the genome (on the + DNA strand). Plotted is the average of this fraction across all sites, separately for the alignments which match and mismatch the genome. **D. Histogram of p-values for the position bias test.** For each RDD site with at least five reads mismatching the genome, we calculated a p-value for the position bias test (Supplementary Information). Plotted is the histogram of these p-values. If these sites were not consistently biased, the distribution of p-values would be uniform; this is indicated with the dashed grey line.

Protein	Position (hg18)	RDD type	# RDD reads	"RNA form" peptide sequence	P-values (dist., strand)	Equally good matches	# additional close matches
AP2A2	chr11:976858	T→G	3	DLALESMCTLASSEFSHEAVK	$0.01, 0.59$	AP2A1	0
DFNA5	chr7:24705225	T→A	23	VFPQLLCITLNGLCALGR	$8 \times 10^{-21}, 2 \times 10^{-7}$	-	0
ENO1	chr1:8848125	T→C	336	EGPELLK	$9 \times 10^{-65}, 8 \times 10^{-13}$	C7orf25, ABCF1	>20
ENO3	chr17:4800624	T→G	8	LAQSNGWGGMVSHR	$0.76, 0.0005$	-	2
FABP3	chr1:31618424	T→A	3	MVDAFLGTR	$0.007, 0.07$	-	1
FH	chr1:239747217	T→A	37	KEYDTFGELK	$1 \times 10^{-43}, 2 \times 10^{-20}$	-	0
HMGB1	chr13:29935772	T→A	10	MSSNAFFVQTCR	$1 \times 10^{-9}, 1 \times 10^{-8}$	HMGB2	2
NACA	chr12:55392932	G→A	16	DIELVMSQANVSR	$3 \times 10^{-8}, 0.80$	-	1
NSF	chr17:42161411	T→C	13	LLDYVPIGPR	$2 \times 10^{-9}, 0.07$	-	0
POL2RB	chr4:57567852	T→A	17	IISDGQK	$4 \times 10^{-10}, 0.0007$	MLKN1, CUL4B	>20
RAD50	chr5:131979610	T→G	9	WRQDNLTLR	$1 \times 10^{-6}, 0.01$	-	0
RPL12	chr9:129250509	A→G	518	HSGDITFDEIVNIAR	$1 \times 10^{-187}, 7 \times 10^{-12}$	-	0
RPL32	chr3:12852658	G→T	356	SAQLAIR	$6 \times 10^{-95}, 8 \times 10^{-12}$	RBM46	>20
RPS3AP47*	chr4:152243651	C→A	81	EVQKNDLK	$1 \times 10^{-62}, 1 \times 10^{-12}$	-	3
SLC25A17	chr22:39520485	A→G	3	TTHMVLLGIIK	$0.002, 0.06$	-	0
TUBA1*	chr2:219823379	A→G	33	EDMAALGK	$4 \times 10^{-6}, 6 \times 10^{-13}$	CCDC85B, TUBA1B, TUBA1C	9
TUBB2C	chr9:139257297	G→A	9	LHFFMPDFAPLTSR	$0.007, 0.31$	TUBB8, TUBB4Q, TUBB6, TUBB2B, TUBB2A, TUBB, TUBB4	1

Table 1: **Characteristics of RDD sites reported in peptides**. We re-evaluated the peptides presented in Table 3 of Li et al. (2011). Repeated from that table are the gene names, positions and types of RDD sites, and "RNA forms" of protein sequences. We additionally show the numbers of aligned reads that mismatch the genome at each site, and the p-values from the tests for position bias and strand bias at each site. P-values in red are less than 0.01. We used BLAST to search the human genome for matches to the peptides; given are the names of additional genes (apart from the one reported by Li et al. [1]) that match the peptide equally well (since these are the "RNA forms" of the peptides, the best matches have a single mismatching amino acid), and the number of genes with one additional mismatch (for a total of two mismatches) to the peptide. Mismatches are defined as either a substitution or an insertion/deletion of a single amino acid. * The RefSeq name for *TUBA1* is *TUBA4A*, and the RefSeq name for *RPS3AP47* is *RPS3A*.

Supplementary Material for "Comment on 'Widespread RNA and DNA sequence differences in the human transcriptome'"

Joseph K. Pickrell[1], Yoav Gilad[1], Jonathan K. Pritchard[1,2]
[1]Department of Human Genetics and
[2]Howard Hughes Medical Institute
University of Chicago, 920 E. 58th St., CLSC 507, Chicago, IL 60637, USA

September 27, 2011

Processing of data from Li et al.[1]. We downloaded the files containing the alignments of RNA-Seq reads used by Li et al. [1] from the Gene Expression Omnibus (accession GSE25840). We then sorted and indexed these files using SAMtools v.0.1.13 [2]. For each RDD site, we extracted the alignments covering the site, and combined reads across all individuals. These alignments were generated using bowtie v.0.12.7 [3]; this read mapping program outputs a flag denoting reads which mapped uniquely to the genome. We removed all reads which mapped non-unquely to the genome (i.e., those which have a bowtie "mapping quality" score less than 255), and all reads where the base covering the RDD site had a sequencing quality score less than 25. Both of these filters are identical to those reported by Li et al., though we found a few differences between the sites reported by Li et al. [1] and our analysis. For example, Li et al.[1] report an RDD site at chromosome 4, position 3,9141,595; as far as we can tell, all of the mismatching bases at that site have a low sequencing quality score. However, this is the only reported RDD site where we fail to find any evidence for mismatching reads in the data, indicating that we are able to analyze the RNA-seq alignments in the same way as Li et al. for nearly all RDD sites.

Tests for "position bias" and "strand bias". To test whether the alignments of RNA sequencing reads around RDD sites indicate the presence of a false positive RDD call, we used tests from the SNP-calling literature [2; 4; 5]. The test for position bias is as follows:

1. For each read alignment covering an RDD site, find which position in the alignment covers the RDD site.

2. Find the distance from that position to either end of the read. That is, if the position in the alignment is i, take $min\{i, 50 - i\}$, since the read length in this experiment is 50.

3. Split the reads into two classes: those which carry the "DNA form" of the base, and those which carry the "RNA form" of the base. The null hypothesis is that the distribution of the above distances is the same in both classes. This is tested by a t-test. In Figure 1D in the main text, we have plotted the distribution of p-values from this test.

The test for strand bias is as follows:

1. Split the alignments of reads covering an RDD site into four classes: those carrying the "DNA form" of the base and mapping to the (+) DNA strand, those carrying the "DNA form" of the base and mapping to the (-) DNA strand, those carrying the "RNA form" of the base and mapping to the (+) DNA strand, and those carrying the "RNA form" of the base and mapping to the (-) DNA strand.

2. Count the number of reads in each class. The null hypothesis is that the alignment strand is independent of the base at the RDD site; this is tested with a Fisher's exact test. The histogram of p-values for this test is presented in Supplementary Figure 1.

It is worth mentioning the types of artifacts which could cause a site to fail these tests. These artifacts are of two types: systematic errors in Illumina sequencing, and errors in identifying the correct genomic location of a sequencing read. We have not attempted to distinguish between these two types of artifact in this analysis, as both are non-biological. For the test of position bias, it is known that the error rate of Illumina sequencing depends on the position in the read [6]. Additionally, mapping errors around insertions/deletions relative to a reference genome can lead to mismatches occurring with positional biases, particularly towards the beginning and ends of alignments. For example, imagine the following sequence from a reference genome: ATGCGATG, and imagine an individual with the sequence ATGCTGCGGATG, where the red represents an insertion relative to the reference. Now, if we had a read with the sequence ATGCT, it would map to the reference sequence with a single mismatch at the end of the alignment, while other possible sequences having greater overlap with the insertion simply wouldn't align to the genome (if we assume a maximum of a single mismatch). This would lead to a spurious call of a G→T mismatch, but this error would be detected as a site showing a position bias.

For the test for strand bias, some types of Illumina sequencing error show a tendency to appear on one strand as opposed to the other; this is presumably because the error rates when sequencing a given sequence and its reverse complement can be different [7; 8]. A strand bias can also be caused by mapping errors, depending on the algorithm used. For example, imagine a sequencing read with mismatches in the first two bases (caused, for example, by an insertion relative to the reference, as in the above example). If the equivalent read were read on the opposite strand, these two mismatches would occur in the last two bases. Many alignment algorithms (including bowtie) use a seed-and-extend approach to mapping reads; the strand of the read influences whether the mismatches are part of the seed alignment used, and can thus influence whether a match is found.

It is likely difficult to differentiate between a sequencing error and a mapping error in many cases; however, both types of artifact can be detected using the above bias tests. In a sense, then, if a site fails one or more of these test, this is a symptom of a problematic site rather than a diagnosis of the exact problem.

Overlap of RDD sites with known SNPs. We downloaded the positions of single nucleotide polymorphisms (SNPs) identified from low-coverage sequencing of the same individuals used by Li et al. [1] from the 1000 Genomes Project (May 2011 release, www.1000genomes.org). Of the 1,033 RDD sites that have at least five mismatching reads and have p-values over 0.01 for both the bias tests, 113 (11%) overlap these SNPs. In nearly all cases (108/113 sites), the alleles of the SNP match the type of RDD event (e.g., if Li et al. [1] report a C→A event, the SNP at that position has the alleles C and A), indicating that these sites are positions of genetic variation rather than differences between RNA and DNA. If we remove the sites that fail the bias tests described above (at a p-value threshold of 0.05) as well as these SNPs, we are left with 515 RDD sites; the distribution of types is shown in Supplementary Figure 2. The proportion of A→G sites has increased to 31% (from 22%), but other types of sequence mismatch remain.

Analysis of peptides. For each peptide presented in Table 1 (from Table 3 of Li et al. [1]), we used BLAST to find matches in human RefSeq proteins. In particular, we used blastp to the human refseq_protein database. We counted the number of proteins with single mismatches to each peptide, as well as the number of proteins with two mismatches or insertion/deletions relative to the peptide. To be conservative, we counted an insertion or deletion of a single amino acid as a mismatch (such that, for example, the insertion of two amino acids would count as two mismatches). These counts are presented in Table 1 in the main text.

We note that the results of this BLAST analysis are different than those presented by Li et al. [1], who report that the RNA forms of the peptides are unique matches to single genes. It is unclear where this discrepancy comes from. One possibility is that the database used to determine whether a peptide is a unique match differs between our analysis and that of Li et al. [1]. In the section "Protein Database with RDD sites" from the Supplementary Information of Li et al. [1], the authors write: "We made a protein database using Gencode mRNA sequences. For genes that display non-synonymous RDDs, protein forms predicted from both DNA sequences and RNA sequences were included." This would seem to suggest that the authors included the predicted protein sequences of inferred RDD sites in the BLAST database. This would explain why the authors report that the RNA forms of peptides are unique (since they've added perfect matches to those peptides to the database); however, this approach assumes that the RDD sites are true positives, and thus would not be a true validation experiment. On the other hand, in the section "B-cells" later in the Supplementary Material, Li et al. [1] write "We carried out BLAST search and ensured that all 28 peptides that correspond to the RNA forms of the RDD-containing peptides are unique to the proteins of interests. For these alignments, we used nr to search all nonredundant sequences (which includes CDS translations+PDB+SwissProt+PIR+PRF)." This latter approach seems to be very similar to ours; it is thus unclear whether a difference in databases can explain the difference between our results and those reported by Li et al. [1].

References

[1] M. Li, *et al.*, *Science* (2011).

[2] H. Li, *et al.*, *Bioinformatics* **25**, 2078 (2009).

[3] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, *Genome Biol* **10**, R25 (2009).

[4] 1000 Genomes Project Consortium, *et al.*, *Nature* **467**, 1061 (2010).

[5] M. A. Depristo, *et al.*, *Nat Genet* **43**, 491 (2011).

[6] Y. Erlich, P. P. Mitra, M. delaBastide, W. R. McCombie, G. J. Hannon, *Nat Methods* **5**, 679 (2008).

[7] K. Nakamura, *et al.*, *Nucleic Acids Res* (2011).

[8] F. Meacham, *et al.*, *Nature Precedings* (2011).

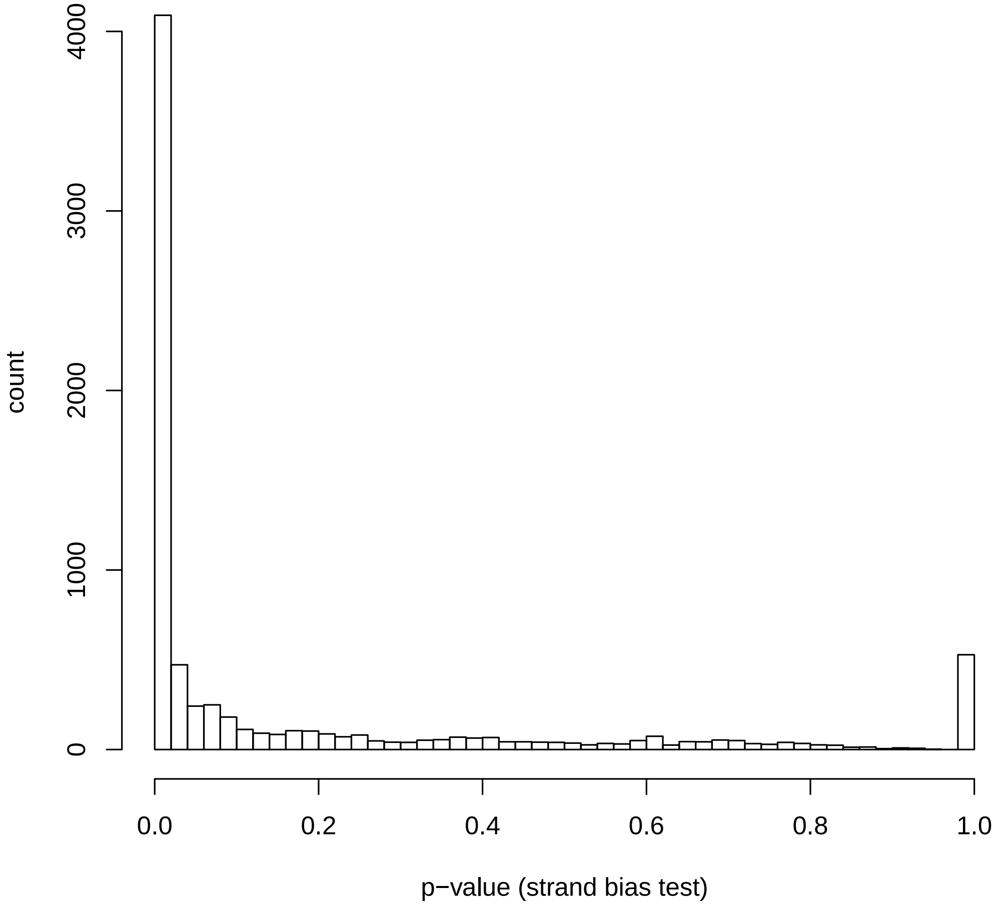

Figure 1: Histogram of p-values from test for strand bias. For each RDD site reported by Li et al. [1] and covered by at least five reads of both the "RNA form" and "DNA form", we calculated a test for strand bias. Plotted is the histogram of these p-values.

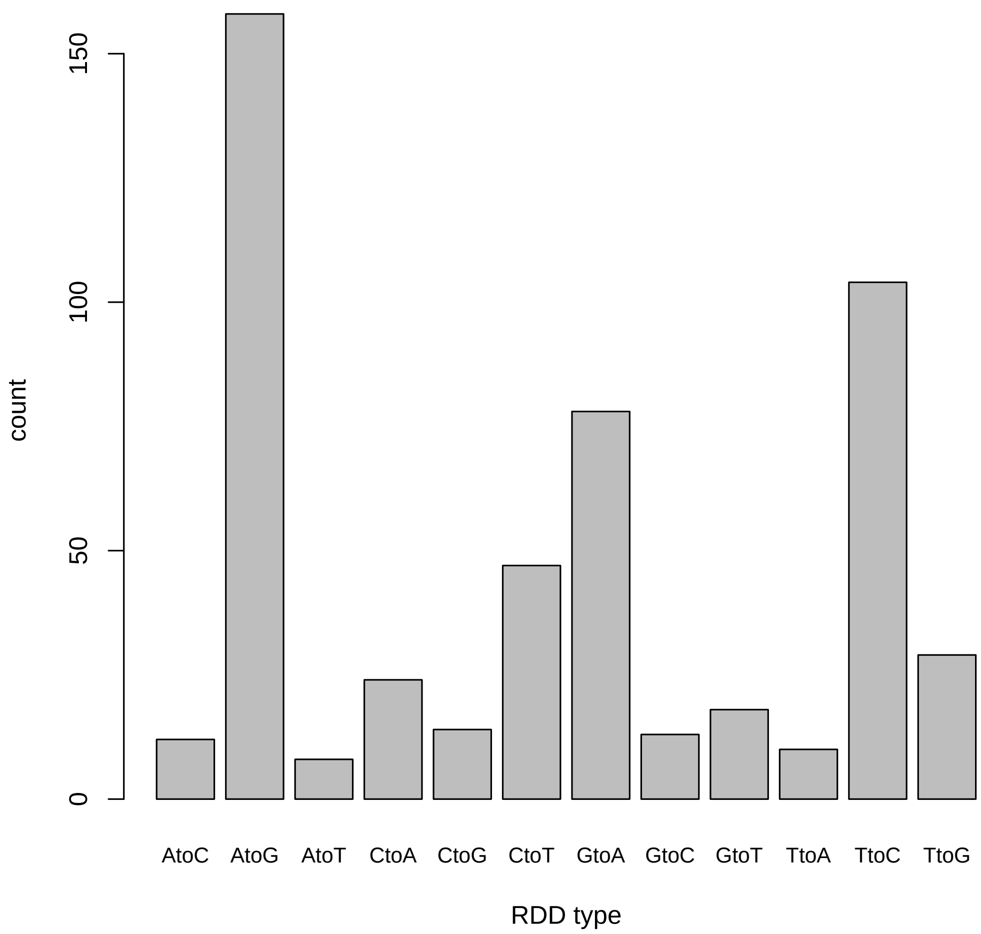

Figure 2: Histogram of RDD types remaining after filtering.