# 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).

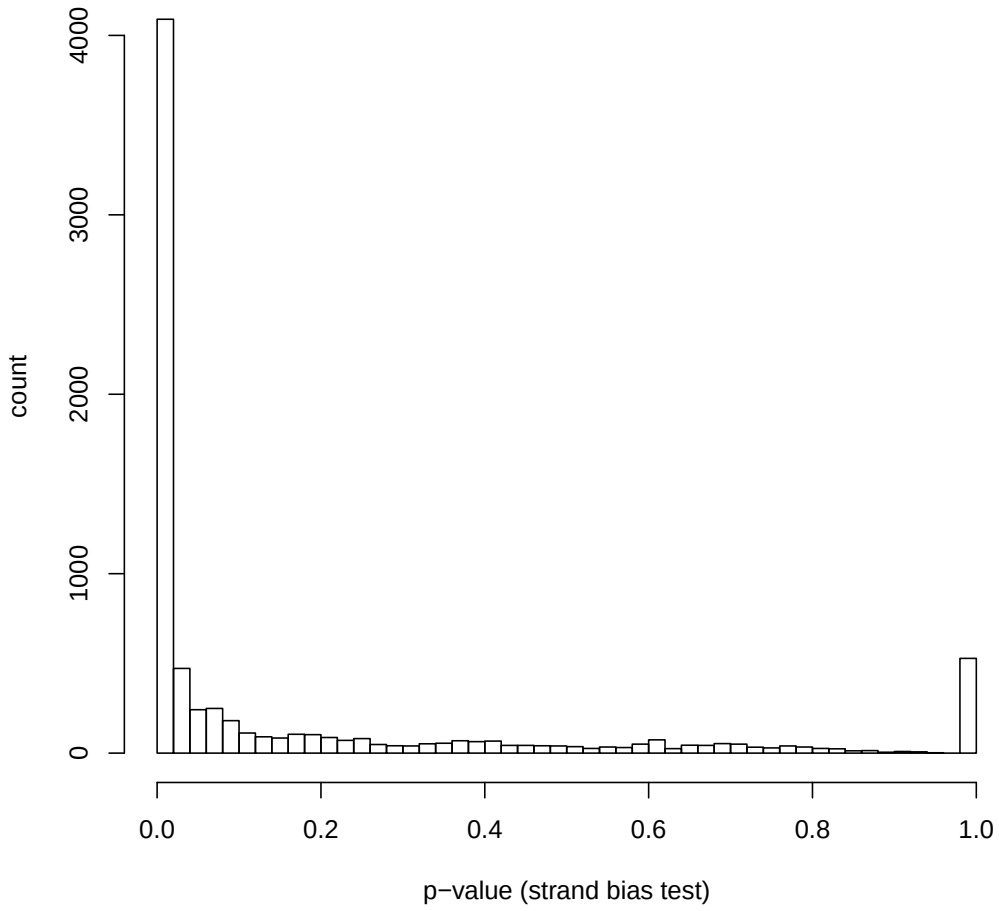


Figure 1: Histogram of p-values from test for strand bias. For each RDD site reported by Li et al. [1] and covered by at least five reads of both the "RNA form" and "DNA form", we calculated a test for strand bias. Plotted is the histogram of these p-values.

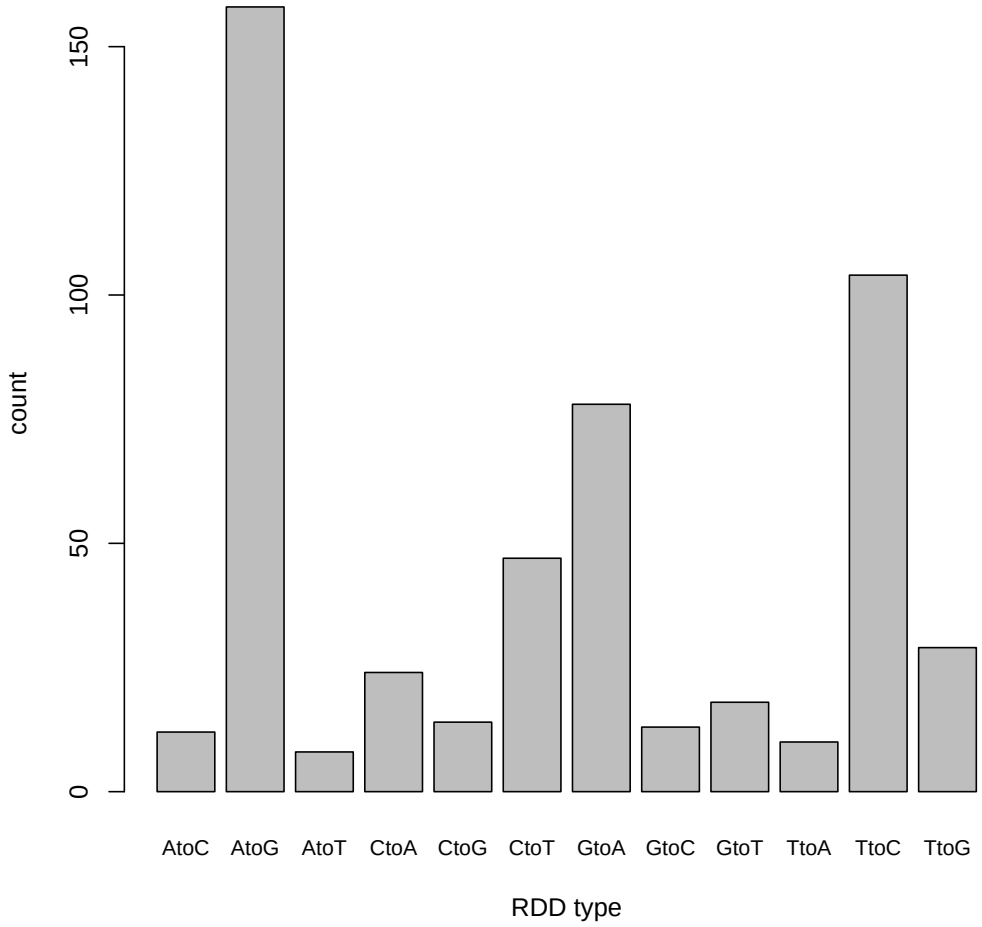


Figure 2: Histogram of RDD types remaining after filtering.