Comment on “Widespread RNA and DNA Sequence Differences in the Human Transcriptome”

Wei Lin, Robert Piskol, Meng How Tan, Jin Billy Li

Li et al. (Research Articles, 1 July 2011, p. 53; published online 19 May 2011) reported widespread differences between the RNA and DNA sequences of the same human cells, including all 12 possible mismatch types. Before accepting such a fundamental claim, a deeper analysis of the sequencing data is required to discern true differences between RNA and DNA from potential artifacts.

By comparing RNA and DNA sequences generated by high-throughput sequencing of human B cells, Li et al. (1) reported finding 10,210 RNA-DNA differences (RDDs) of all 12 possible mismatch categories. Before this study, only two types of RDDs, known as RNA editing, were known to exist in humans, both of which are catalyzed by deaminases (2, 3). One converts adenosine to inosine (recognized as guanine); the other converts cytosine to uracil. Li et al.’s finding, if confirmed, would add tremendous informational complexity to the transcriptome. However, after examining their data, we discovered several issues that may lead to an exceedingly high rate of false-positive calls of RDDs.

Accurately mapping the short high-throughput sequencing reads is critical for calling RDDs to ensure that RNA- and DNA-derived sequences originate from the same genomic locus. A read derived from one genomic locus can be incorrectly mapped to another paralogous region that arose through genomic duplication. Slight differences between duplicated sequences can be mistaken for RDDs. To estimate the fraction of RDDs that may be misclassified due to duplications, we extracted the flanking sequences (25 bases on each side) of the 10,210 RDDs identified in (1) and aligned them against the human genome with BLAT (4). A total of 1284 sites (~13%) mapped to two or more regions in the genome with >90% identity (Fig. 1). Although Li et al. removed all of the segmental duplications (~1 kb) and repetitive regions (1), a large number of known shorter duplications were not removed as they are not considered to be repetitive. In addition, Li et al. mapped reads to the transcriptome rather than the genome. Their reference transcriptome (Gencode; www.gencodegenes.org) is potentially missing some paralogous genes, which could lead to erroneous RDD calls.

Even when a read is mapped to the correct genomic locus, that read’s alignment with the genome can contain differences primarily for two reasons. First, when a read spans across the splicing junction of an alternatively spliced gene, a misalignment can be easily introduced due to ambiguous mapping of the read end to one of two (or more) possible exons. This is especially true when reads are being mapped to the transcriptome that contains a possibly incomplete annotation of isoforms. Second, the 5’ read end, derived from either end of a double-stranded cDNA fragment, may have a higher error rate due to mispriming events introduced by the random hexamers used in the RNA sequencing protocol. Because the alignment tool used in (1), Bowtie, performs semiglobal alignment that forces the read ends to be aligned even when there is a mismatch (5), the two issues above tend to concentrate mismatches to the ends of the reads.

To investigate this further, we took all reads that carry the variant bases at the RDDs and asked where the variants occur along the reads. Consistent with our hypotheses above, there is a clear enrichment of variants at the read ends, particularly at the 5’ end (Fig. 2A). However, one could argue that the observed 5’ read end bias is precisely due to the presence of RDDs, whereby the first-strand cDNA synthesis may be prematurely terminated immediately after the RDD site is transcribed due to reverse transcriptase incompatibilities with a modified base. To test this possibility, we annotated the originating RNA strand of each read to separate the sequencing reads into two groups: one derived from the 5’ end of the first-strand cDNA, and the other from the 5’ end of the second-strand cDNA. If the premature termination hypothesis were true, we would expect RDDs to be enriched in the 5’ end of the second-strand cDNA. In fact, we found that the opposite is true, where the variants are strongly favored at the 5’ ends of the first-strand cDNA (Fig. 2, B and C). We speculate that this phenomenon is primarily rooted in the RNA sequencing protocol. The 5’-end mismatch may be introduced by the random hexamer during the first- and second-strand syntheses. However, the DNA polymerase I used in the second-strand synthesis has 5’->3’ exonuclease activity that can remove 5’-end mismatches from the second-strand cDNA. Based on these observations, we trimmed the ends (positions 1, 2, 48, 49, and 50) of all RNA sequencing reads.

**Fig. 1.** Percentage of RDDs in duplicated regions. We extracted regions of 25 bp to each side of all RDDs. For each region, the number of hits to the reference genome and the percentage of matched nucleotides was determined. An RDD was considered to be in a duplicated region if more than one hit to the genome was found.

---

1Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA. 2Department of Genetics, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: jin.billy.li@stanford.edu
reads and recalled the RDDs. As a result, a vast majority (8364, 82%) of the 10,210 RDDs were eliminated.

In addition to read mapping–related issues, many RDD sites identified in (1) may actually be heterozygous single-nucleotide polymorphisms (SNPs) that were incorrectly called as homozygous due to the low coverage of genome sequences (as few as four reads were used to determine homozygosity in the DNA). In fact, we identified 556 RDDs that are annotated in the most recent SNP database (dbSNP132), and 476 (~86%) of them match the SNP type. Because the SNP database is far from being comprehensive, it is likely that many other RDDs are also derived from heterozygous SNPs that are not called due to low genome sequence coverage.

Taken together, we identified 9037 (~89%) of 10,210 RDD sites that may be false positives derived from gene duplications, read-end misalignment, and/or SNPs. Although it is possible that some of the remaining 1173 sites are real RDDs, only 14.8% of them are A-to-G type, which is believed to be the most common type of editing. This low fraction probably indicates a high false-positive rate even in the remaining set of sites. We manually examined many of these positions, and a large fraction of the variant calls appeared to be derived from misalignments, particularly at the splicing junctions. In the future, it will be of paramount importance to confirm their presence with alternative methods to rule out additional artifacts. High-throughput sequencing will continue to revolutionize the field of RNA editing by delivering enormous amounts of data at single-base resolution (6, 7). However, such data have their intrinsic limitations and pitfalls (8) and should therefore be used with caution. Rigorous sequence data analyses are crucial to reliably identify RDDs that should then be subject to further experimental validations.

**References and Notes**

4. W. J. Kent, Genome Res. 12, 656 (2002).

**Acknowledgments:** We thank U. Laserson for critical reading of the manuscript. This work was supported by Stanford startup funds to J.B.L.

4 July 2011; accepted 19 December 2011
10.1126/science.1210624

---

**Fig. 2.** Distribution of RDD variants across read positions. (A) We took all reads used in (1) that carry the variant type of RDDs and examined the position of RDD variants in the reads. Each of the 27 individuals is shown in a different color. The same analysis was repeated separately for (B) reads derived from the 5′ end of the first-strand cDNA (opposite to the RNA strand) and (C) reads derived from the 3′ end of the second-strand cDNA (same as the RNA strand). In the inset diagrams, red and blue thick arrows highlight the orientation of reads used in analysis.
Comment on "Widespread RNA and DNA Sequence Differences in the Human Transcriptome"
Wei Lin, Robert Piskol, Meng How Tan and Jin Billy Li

Science 335 (6074), 1302.
DOI: 10.1126/science.1210624