Artificial DNA: PNA & XNA 2:3, 69-70; July/August/September 2011; © 2011 Landes Bioscience

RNA-DNA sequence differences spell genetic code ambiguities

Thomas Bentin[1,*] and Michael L. Nielsen[2]

[1]Department of Cellular and Molecular Medicine; [2]Department of Proteomics; The Novo Nordisk Center for Protein Research; University of Copenhagen; Copenhagen, Denmark

A recent paper in Science by Li et al. 2011[1] reports widespread sequence differences in the human transcriptome between RNAs and their encoding genes termed RNA-DNA differences (RDDs). The findings could add a new layer of complexity to gene expression but the study has been criticized.

RDDs are reminiscent of adenosine to inosine enzymatic mRNA editing (and less frequently cytosine to uracil editing) but are generated via entirely unknown mechanism(s). A distinguishing feature of RDDs is that they encompass all 12 possible nucleotide exchanges. The implications are potentially profound because the sheer scale of RDDs reported including 28.766 events at 10.000 exonic sites, means that there are a considerable number of exceptions to our digital one-to-one understanding of how DNA encodes RNA. Correspondingly, non-synonymous RDDs located in open reading frames lead to sequence heterogeneous proteins whose identity cannot be predicted as based on DNA sequencing alone. RDDs, also prevalent in the 3' untranslated region, could potentially be of significance to mRNA- localization, translation and turnover. From an applied perspective, design of antisense and siRNA should be contemplated within this new framework.

The report by Li and co-workers has already been the subject of extensive scrutiny. The challenge is to discriminate whether two RDD variants of RNA in fact originate from one gene or from different but homologous genes. The main objections concern challenges in assigning short RNA sequence reads (50 nt) to the corresponding parent DNA locus due to the occurrence of paralogs (multiple copies of the same gene), difficulties in read assignments located across intron-exon boundaries, and potential errors generated during cDNA preparation. Even the validation by Sanger sequencing capable of much longer reads (typically around 800 bases) carried out for a subset of RDDs has been contested at www.genomesunzipped.org.

We should like to point out that the present study also faces challenges in relation to peptide identification by mass spectrometry as only briefly touched upon previously. The authors have used two proteomic datasets for the protein validation of RDD events. One publicly available dataset was acquired with low-resolution instrumentation, which holds the potential for large false discovery identifications. This dataset was searched in the protein database with a tolerance window of 4 dalton for peptide parent masses, and 0.5 dalton for fragment ions. Particularly, a peptide tolerance window of 4 dalton creates a possibility for larger false-positive identification, especially for identification of peptide amino acid substitutions considering the number of amino acids, which differ in mass by less than 4 dalton (e.g. proline/valine/threonine, leucine/isoleucine/asparagines/aspartic acid and lysine/glutamine/glutamic acid/methionine).

The authors correctly used high-mass accuracy instrumentation (LTQ Orbitrap) for proteome analysis of B-cells, but utilized low-resolution settings when searching their data in the protein sequence dataset. In this search the authors allowed for a peptide mass tolerance of 0.3 dalton, whereas it would

Key words: Transcription, RNA editing, gene expression, RNA-DNA differences, transcriptome

Submitted: 06/29/11

Accepted: 06/30/11

DOI: 10.4161/ADNA.2.3.17086

Correspondence to: Thomas Bentin;
Email: bentin@sund.ku.dk

Commentary to: Li et al. Widespread RNA and DNA Sequence differences in the human transcriptome. Scienceexpress 19 May 2011

be expected that the appropriate search tolerance should have been 5-10 ppm instead. Considering that the average tryptic peptide contains 9-12 amino acids, a peptide mass tolerance window of 0.3 dalton corresponds to a mass accuracy of >150ppm – i.e. significantly larger than what would be necessary for high-resolution datasets.

Overall, the authors identify 38,572 peptides belonging to 3,217 proteins at a false-discovery rate of <1%. In such a dataset it would therefore be assumed that up to 386 peptides are wrongly identified (1% of 38,572). Considering this false discovery rate and comparing its size to the total number of identified peptide level RDD events (in total 327), it can be questioned whether these identified RDD events might be overrepresented among the 1% falsely identified peptides.

From the above deliberations, and in particular from those of Joe Pickrell and others at genomesunzipped.org, it should be painfully evident just how complex a topic global RDD analysis constitute demanding specialized expertise within several disciplines and high-end instrumentation. The present work of Li et al.[1] formulated as a "RDD hypothesis" provides a fundamental deviation from Crick's "Central Dogma" and "Sequence Hypothesis",[2] which is biologicaly interesting, experimentally testable, and undoubtedly will be the focus of much contemporary investigation.

References

1. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM et al. Widespread RNA and DNA sequence differences in the human transcriptome. Science 2011; 33:53-8.
2. Crick FH. On protein synthesis. Symp Soc Exp Biol 1958; 12:138-163.