[CANCER RESEARCH 63, 6996–6997, October 15, 2003]

Letters to the Editor

Expressed Sequence Tags: Clean before Using. Correspondence re: Z. Wang *et al.*, Computational Analysis and Experimental Validation of Tumor-associated Alternative RNA Splicing in Human Cancer. Cancer Res., *63:* 655–657, 2003.

Letter

Wang *et al.* (1) report on finding 845 alternative splicing isoforms that were highly associated with human tumors. They further found that canonical (GT-AG) splice junctions were noticeably less common in the tumor-associated isoforms. Their study was based on inferring alternative splicing using ESTs[1] aligned to full-length mRNAs.

We have recently identified, as part of an investigation of contaminated EST libraries (2), 18 libraries that contain abnormally high fractions (as high as 39%) of ESTs that span noncanonical introns. The noncanonical introns in these 18 libraries were much shorter than normal introns and had a very different length distribution than introns in the genome as a whole (3). In particular, 64% of noncanonical introns from these libraries were 51–59 bases long, and 46% were exactly 54 bases, as opposed to an average intron length of >3000 bases over the entire genome (3). Furthermore, all 18 identified EST libraries were generated in the same institute. We therefore concluded that an error of some sort occurred in the creation or reporting of the ESTs from those libraries, so that these EST sequences are artifacts rather than real biological phenomena. Unfortunately, Wang *et al.* could not have been aware of our work, as our studies were published contemporaneously.

Examining the data reported by Wang *et al.*, we found that 151 of the 573 reported deletions were found using only ESTs from the 18 libraries that we identified as problematic. Indeed, 141 (93%) of these deletions were noncanonical, and 129 (85%) were 51–59 bases long. We note that although ESTs from the 18 contaminated libraries comprise only ~1% of all of the ESTs in dbEST, they are responsible for 26% of the deletions reported by Wang *et al.* We therefore conclude that these 151 isoforms are artifacts and should not be considered real tumor-associated splice variants.

We believe that the conclusions of Wang *et al.* about the potential for using alternative splicing isoforms as diagnostic markers still hold and are interesting, although we think that they have found fewer useful markers than they initially thought. However, in light of our findings, their claim that "canonical GT-AG splice junctions were used significantly less frequently in the alternative splicing isoforms in tumors" must be reexamined.

In conclusion, ESTs are extremely useful for many applications, as the article by Wang *et al.* demonstrates. However, since EST databases contain many kinds of artifacts, they must be thoroughly cleaned before being used for large scale analyses, in order for such studies to yield meaningful results.

Rotem Sorek
Compugen Ltd.
Tel Aviv 69512, Israel
Department of Human Genetics and Molecular Medicine
Sackler School of Medicine
Tel Aviv University, Ramat Aviv 69978, Israel

Ortal Basechess
Hershel M. Safer
Compugen LtD.
Tel Aviv 69512, Israel

References

1. Wang, Z., Lo, H. S., Yang, H., Gere, S., Hu, Y., Buetow, K. H., and Lee, M. P. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. Cancer Res., *63:* 655–657, 2003.
2. Sorek, R., and Safer, H. M. A novel algorithm for computational identification of contaminated EST libraries. Nucleic Acids Res., *31:* 1067–1074, 2003.
3. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature, *409:* 860–921, 2001.

Reply

We reported that 845 alternative splicing isoforms were associated with human cancers, and we also found that canonical GT-AG splicing junctions were used significantly less frequently in the alternative splicing isoforms in tumors (1). Our analysis was based on using the basic local alignment search tool (BLAST) algorithm. Alignments generated from BLAST required high-quality expressed sequence tag (EST) sequences. More importantly, 54 of 55 reverse transcription-PCR assays (98%) detected the predicted alternative splicing products, and the sequence of these products was confirmed by direct sequencing (1). We also performed experimental validation for 27 of the 151 deletion splicing isoforms that Sorek and Safer identified as problematic. We found that 13 of 27 reverse transcription-PCR assays (48%) detected the predicted alternative splicing products. This experimental validation clearly supported the validity of our analysis for alternative splicing products.

Sorek and Safer (2) identified 18 EST libraries as being contaminated with noncanonical introns through a computational analysis. These 18 libraries contained noncanonical introns exceeding three SDs above the mean. There was no other convincing evidence why these libraries were problematic other than the statistical deviation. In fact, the low EST counts in these libraries might cause such a deviation (8 of 18 contained <1000 ESTs). Although it may be suspicious that all 18 libraries identified originate from the same institute, laboratory-based experiments are required to prove that ESTs from these 18 libraries are indeed artifacts.

Regarding the frequency of splicing junctions, the relevant issue is the splicing acceptor site of the type II deletion and the splicing donor of the type III deletion described in Fig. 1 in our article (1). We have reexamined 151 of the 573 deletion splicing products that came from the 18 libraries that Sorek and Safer (2) identified as problematic.

Received 2/23/03; accepted 5/21/03.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] The abbreviation used is: EST, expressed sequence tag.

Received 6/12/03; revised 8/4/03; accepted 8/6/03.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

None of the ESTs from the 151 deletion splicing isoforms from the 18 libraries in question were in the data set used to calculate the frequency of the type II deletion shown in Fig. 1 (1). Only 1 EST from the 151 deletion splicing isoforms in the type III deletion was included in our data in Fig. 1, and this did not affect our finding.

Zhining Wang
Maxwell P. Lee
Laboratory of Population Genetics
National Cancer Institute
Bethesda, Maryland 20892

References

1. Wang, Z., Lo, H. S., Yang, H., Gere, S., Hu, Y., Buetow, K. H., and Lee, M. P. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. Cancer Res., *63:* 655–657, 2003.
2. Sorek, R., and Safer, H. M. A novel algorithm for computational identification of contaminated EST libraries. Nucleic Acids Res., *31:* 1067–1074, 2003.