The American Journal of Pathology, Vol. 185, No. 7, July 2015

The American Journal of
PATHOLOGY
ajp.amjpathol.org

REVIEW

Discovery and Classification of Fusion Transcripts in Prostate Cancer and Normal Prostate Tissue

Jian-Hua Luo,* Silvia Liu,† Ze-Hua Zuo,* Rui Chen,† George C. Tseng,† and Yan P. Yu*

From the Departments of Pathology and Biostatistics,† University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania*

Accepted for publication March 12, 2015.

Address correspondence to Jian-Hua Luo, M.D., Ph.D., Department of Pathology, Scaife S-728, 3550 Terrace St., University of Pittsburgh School of Medicine, Pittsburgh, PA 15261. E-mail: luoj@msx.upmc.edu.

Fusion transcript formation is one of the fundamental mechanisms that drives the development of prostate cancer. Because of the advance of high-throughput parallel sequencing, many fusion transcripts have been discovered. However, the discovery rate of fusion transcripts specific for prostate cancer is lagging behind the discoveries made on chromosome abnormalities of prostate cancer. Recent analyses suggest that many fusion transcripts are present in both benign and cancerous tissues. Some of these fusion transcripts likely represent important components of normal gene expression in cells. It is necessary to identify the criteria and features of fusion transcripts that are specific for cancer. In this review, we discuss optimization of RNA sequencing depth for fusion transcript discovery and the characteristics of fusion transcripts in normal prostate tissues and prostate cancer. We also propose a new classification of cancer-specific fusion transcripts on the basis of their tail gene fusion protein product and the roles that these fusions may play in cancer development. *(Am J Pathol 2015, 185: 1834–1845; http://dx.doi.org/10.1016/j.ajpath.2015.03.008)*

Genomic alteration is the hallmark of human malignancies. These alterations are manifested as changes in copy number in a region of chromosome such as amplification, deletion, duplication, or chromosome rearrangement. Single nucleotide mutations that alter the protein structure may also have relevant impact on the development of cancers. The invariable association between human malignancies and these genomic alterations argues strongly that genome alterations, rather than epigenomic changes, are the most critical and common underlying mechanisms for the development of human cancers. The genomic and epigenomic alterations and their corresponding gene expression alteration in prostate cancer were well documented and characterized in the past 15 years.[1–11]

Studies that used high-throughput genome array and whole-genome sequencing reveal a large number of copy number changes in human prostate cancer genomes, affecting almost all chromosomes.[6,12,13] Because of the high level of genomic rearrangement identified in the prostate cancer genome, the term *chromoplexy* was coined to describe the complexity and abundance of chromosome rearrangements.[14] Some of the grave consequences of chromosome rearrangement are the aberrant joining of unrelated genes and the production of a fusion transcript.

The events of joining together two unrelated genes often lead to the creation of a new oncogene that might be analogous to BCR-Abl seen in leukemias. Even though genomic rearrangement in prostate cancer is extensive, only a handful of fusion transcripts related to prostate cancer are properly validated, suggesting that the prevailing approach of fusion transcript discovery is suboptimal. RNA sequencing could be modified to improve the discovery rate of fusion transcripts.

High Coverage of RNA Sequencing Is Necessary to Discover Most Low-Abundance Fusion Transcripts

Abundantly expressed housekeeping genes usually have the dominant presence in whole-transcriptome sequencing. Indeed, we found that the top quartile of expressed genes generally account for 90% of the mapped reads in our

Supported by NIH grant RO1 CA098249 (J.-H.L.), American Cancer Society grant RSG-08-137-01-CNE (Y.P.Y.), and a grant from the University of Pittsburgh Cancer Institute.

Disclosures: None declared.

http://dx.doi.org/10.1016/j.ajpath.2015.03.008

transcriptome analyses.[15] This leaves few detected reads representing low abundantly expressed genes.

To detect fusion transcripts, accurate detection relies on the finding of split reads that contain fusion-joining sequences of two unrelated genes in a 100-bp sequence when using Illumina HiSeq2500 sequencer. Reads mapped to other regions of a fusion transcript are irrelevant because they overlap a fusion transcript and its wild-type counterpart and thus will be classified as wild-type gene transcripts.

Even though spanning reads that contain pair-end mapped to the head and tail parts of a fusion transcript may provide important support for the presence of a fusion transcript, the rate of validation is low because of uncertainty of the location of the fusion joining point. Thus, the focus of fusion transcript detection is identification of sequences that contain the fusion split juncture. Because of the limit of sequence mapping, the split juncture has to be at least 10 bp away from the end of the read to avoid ambiguous mapping. Sequences not >90 bp from either side of the fusion split juncture will be used when using Illumina 200 cycle paired-end sequencing scheme. As a result, the rate of detecting a split read may be low if the expression of the fusion transcript is not abundant.

In our recent transcriptome sequencing that used 1200× to 1400× coverage per gene,[15] of the eight validated fusion transcripts and TMPRSS2-ERG, the numbers of split reads ranged from 2 to 13 (Table 1).[15] Most of the fusion transcripts have five or less split reads by using FusionCatcher (*https://code.google.com/p/fusioncatcher*, last accessed March 12, 2015). This translates to the bottom half of expression levels for all genes. Two of nine fusion transcripts are in the bottom 30% of expressed genes.

To determine the threshold of coverage to detect these transcripts, we randomized and subsampled the mapped reads to equivalence of 1000×, 600×, 400×, and 200× coverage. Although TMPRSS2-ERG and MTOR-TP53BP1 were consistently detected even when coverage was reduced to 300×, all other fusion transcripts disappeared when sequencing coverage was reduced to 600×. However, when sequencing depth was raised to 1000× coverage, only LRRC59-FLJ60017 was not detectable. Thus, sequencing depth is crucial to detecting novel fusion transcripts.

To detect most fusion transcripts, a minimum of 1000× coverage is required. Because whole-genome sequencing from these prostate cancer samples indicates far more chromosome rearrangements and translocations than the number of fusion transcripts found,[14,15] a substantial number of fusion transcripts may have escaped our detection even with 1300× coverage. Quite likely, the undetected fusion transcripts are expressed in low abundance.

Extensive Presence of Fusion Transcripts in Normal OD Prostate Tissues

Similar to nucleotide and copy number variants, most fusion transcripts can be physiologic. Recent transcriptome sequencing of normal prostate tissues from healthy organ donors (OD) revealed numerous fusion transcripts in normal prostate tissues.[15] These fusion transcripts can be classified into two categories: The fusion transcripts connect two adjacent transcripts. This type of fusion transcripts appears to result from the splicing of the 5′ end of a head gene and the 3′ end of a tail gene. A chimera mRNA is produced with the arrangement of head—tail gene in the same direction of transcription from the chromosome. Such fusion transcript formation requires no chromosome rearrangement. Some of these transcripts may have a chimera protein translated. The second type of fusion transcript involves the combination of head and tail genes from a far distance, or from different chromosomes, or in different chromosome transcription directions. The formation of such fusion transcripts may require chromosome recombination or may be RNA recombination if the evidence of chromosome rearrangement is absent.

Among the 20 samples of OD prostate samples, >80% of validated fusion transcripts are the results of splicing of a continuous transcript that spans two genes (category 1). The most common iteration of type 1 category 1 fusion is *TTTY15-USP9Y* from Y chromosome (Table 2). *TTTY15* is a testis-specific non-coding RNA with no known function. USP9Y is an ubiquitin-specific peptidase that may be involved in spermatogenesis.[16] Two different fusion points in *TTTY15* were identified (Figure 1A): one fusion isoform produced chimera RNA with the first 3 exons of *TTTY15*

Table 1 Coverage Dilution Simulation to Detect Fusion Transcripts

				Whole		1000×		600×		300×	
Sample	Head gene	Tail gene	Expression percentile	Span	Split	Span	Split	Span	Split	Span	Split
1T	*TMPRSS2*	*ERG*	30^{th}—40^{th}	7	6	6	2	3	1	3	1
5T	*TMPRSS2*	*ERG*	70^{th}—80^{th}	6	13	2	4	0	2	0	0
3T	*TRMT11*	*GRIK2*	50^{th}—60^{th}	3	12	5	4	0	0	0	0
1T	*LRRC59*	*FLJ60017*	30^{th}—40^{th}	3	7	0	0	0	0	0	0
4T	*TMEM135*	*CCDC67*	20^{th}—30^{th}	3	3	3	1	0	0	0	0
2T	*MTOR*	*TP53BP1*	50^{th}—60^{th}	12	5	10	5	5	5	5	3
3T	*MAN2A1*	*FER*	60^{th}—70^{th}	3	13	3	8	0	0	0	0
3T	*KDM4B*	*AC011523.2*	20^{th}—30^{th}	7	2	7	1	0	0	0	0
3T	*CCNH*	*C5orf30*	30^{th}—40^{th}	3	5	2	4	0	0	0	0