BMB
reports

Invited Mini Review

Chimeric RNAs as potential biomarkers for tumor diagnosis

Jianhua Zhou[1], *Joshua Liao*[2], *Xuexiu Zheng*[3] *& Haihong Shen*[3,*]

[1]Nantong University, Nantong, JiangSu 226001, P. R. China, [2]The Hormel Institute, University of Minnesota 55912, Austin, MN, USA, [3]Gwangju Institute of Science and Technology, Gwangju 500-712, Korea

Cancers claim millions of lives each year. Early detection that can enable a higher chance of cure is of paramount importance to cancer patients. However, diagnostic tools for many forms of tumors have been lacking. Over the last few years, studies of chimeric RNAs as biomarkers have emerged. Numerous reports using bioinformatics and screening methodologies have described more than 30,000 expressed sequence tags (EST) or cDNA sequences as putative chimeric RNAs. While cancer cells have been well known to contain fusion genes derived from chromosomal translocations, rearrangements or deletions, recent studies suggest that trans-splicing in cells may be another source of chimeric RNA production. Unlike cis-splicing, trans-splicing takes place between two pre-mRNA molecules, which are in most cases derived from two different genes, generating a chimeric non-co-linear RNA. It is possible that trans-splicing occurs in normal cells at high frequencies but the resulting chimeric RNAs exist only at low levels. However the levels of certain RNA chimeras may be elevated in cancers, leading to the formation of fusion genes. In light of the fact that chimeric RNAs have been shown to be overrepresented in various tumors, studies of the mechanisms that produce chimeric RNAs and identification of signature RNA chimeras as biomarkers present an opportunity for the development of diagnoses for early tumor detection. (BMB reports 2012; 45(3): 133-140)

INTRODUCTION

In 1971, President Richard Nixon of the United States signed the National Cancer Act, which formally declared war on cancers. Billions of dollars have since been invested in research and development in order to better understand the mechanisms of cancer biology and to find more effective diagnostics and treatments. As a result, great strides have been made in every phase of cancer research, diagnosis and treatment, in particular for several specific types of cancers. Among many achievements, early detections, new drug therapeutics, surgical removal of tumors and other technologies have significantly increased the survival rates of cancer patients. However, despite this progress, the war has not yet been won and cancers are still the number one killer in the United States.

With the rapid expansion of research in genomics and proteomics during the last two decades, scientists now believe that cancer is close to being conquered. Genomics and proteomics are the studies of entire genomes of organisms at DNA, RNA and protein levels. Genomics at the DNA level is concerned with studies such as sequencing entire genomes and identifying mutations and single nucleotide polymorphisms (SNPs), which can be used as biomarkers in disease diagnoses, personalized medicine and treatments. On the other hand, proteomics involves investigation of proteins. Between DNA and proteins, the accumulation of genome knowledge has incited studies into functional genomics or gene expression at the RNA level. As a result of these studies, microarray, antibody array, protein array and next-generation sequencing have been developed in order to meet the requirements for high throughput approaches that can identify the bio-molecules that are specifically expressed in diseases such as cancers. Consequently, methods such as the classic and commonly used blood tests for detection of just a few proteins and biochemical molecules at a time have rapidly evolved into modem molecular and cellular biology tools that have spurred the discovery of biomarkers at several levels by a parallel high-throughput approach. Publications on genome sequences, mutations, SNPs, mRNAs, microRNA, non-coding RNAs and proteins have exploded in the last two decades. Even chimeric RNAs, one particular domain of biomarkers that had been until recently overlooked, seems to be gaining prominence.

A chimeric RNA is a RNA molecule that is made of two or more pieces of RNAs from different loci that should not be found on the same molecule. Chimeric RNAs are initially identified from transcripts of fusion genes due to chromosome translation, inversion or deletion (Table 1). Further studies have indicated that many other RNA chimeras are generated from mechanisms such as transcription read-through and splicing, transcription of short homologous sequence slippage or from so called trans-splicing (Table 1) (1). In the aid of high-throughput sequencing technology and bioinformatics analysis, a growing number of chimeric RNAs have been identified and validated during the last two years (2-8), raising the possibility that chimeric RNAs that are expressed specifically

*Corresponding author. Tel: +85-62-715-2507; Fax: +82-62-715-2484; E-mail: haihongshen@gist.ac.kr
http://dx.doi.org/10.5483/BMBRep.2012.45.3.133

Received 18 February 2012

Keyword: Biomarkers, Chimeric RNA, Fusion gene, Trans-splicing

Table 1. Sources of chimeric RNAs production

Source	Example	Fusion gene	Fusion RNA	Disease associated
Chromosome translocation	Philadelphia chromosome	Bcr-Abl	Bcr-Abl (9)	Chronic myelogenous leukemia (CML)
Interstitial deletion	Deletions at 11q23	MLL–FOXR1 and PAFAH1B2–FOXR1	MLL–FOXR1 and PAFAH1B2–FOXR1 (10)	Neuroblastomas
Chromosomal inversion	Inv(2)(p21;p23)	EML4-ALK	EML4-ALK (11)	Non-small cell lung cancer (NSCLC)
Trans-splicing	Anti-Apoptotic protein JAZF1-JJAZ1		JAZF1-JJAZ1 (22)	Endometrial stromal tumors
SHS slippage transcription	DMRT1-CD5R		DMRT1-CD5R (32)	
Read-through transcription	G-protein receptor $P2Y_{11}$		SSF1-$P2Y_{11}$ (18)	

SHS: short homologous sequence.

in diseases can be used as potentially valuable biomarkers in diagnostics and prognostics.

CHIMERIC RNAs PRODUCED FROM FUSION GENES

A fusion gene is defined as a gene that is produced from two previously separated genes. It has been demonstrated that many cancers such as prostate cancers and hematological cancers involve fusion genes as a result of chromosomal translocation, interstitial deletions or chromosomal inversions. One of the best known examples is the fusion Bcr-Abl gene that is defined in the Philadelphia chromosome, a specific reciprocal chromosomal translocation between chromosomes 9 and 22 that is responsible for chronic myelogenous leukemia (CML) (9). This translocation results in an elongated chromosome 9 and a truncated chromosome 22. Consequentially, the fusion gene is formed by the Abl gene on chromosome 9 (region q34) to a part of the Bcr gene on chromosome 22 (region q11). Depending on the precise fusion sites between the Bcr gene and the Abl gene and on alternative splicing events, the oncogenic fusion gene produces proteins with molecular weights ranging from 185 to 210 kDa. Since more than 95% of CML patients have the Philadelphia chromosome, the presence of this translocation and the fusion Bcr-Abl gene has been widely used as a test to detect the disease.

Fusion genes can be also formed from interstitial deletions. For instance, in a recent study (10), Santo *et al*. analyzed a series of neuroblastomas by comparative genomic hybridization and single-nucleotide polymorphism arrays and identified small interstitial deletions at 11q23, upstream of the forkhead-box R1 transcription factor (FOXR1) gene. These deletions lead to fusions between the genes at the proximal side of the deletions and the FOXR1 gene. More significantly, in contrast to the normal expression pattern of the FOXR1 gene that occurs in early embryogenesis, the resulting fusion transcripts of MLL–FOXR1 and PAFAH1B2-FOXR1 are exclusively expressed in neuroblastomas, indicating that the fusion products play a role in tumorigenesis of neuroblastomas.

Chromosome inversion is another source that can produce fusion genes. Chromosome inversion takes place when a segment of a chromosome is reversed end to end. Fusion genes that are created from chromosome inversion have been described in numerous tumors. For example, the chromosome inversion inv(2)(p21;p23) that leads to the echinoderm microtubule-associated protein-like 4-anaplastic lymphoma kinase (EML4-ALK) fusion gene was identified in non-small cell lung cancer (NSCLC) (11). Other examples of fusion genes that result from chromosome inversion include the *MLL* (Mixed-Lineage Leukemia) gene (12) with an inversion of 11q (inv (11)(q14q23)), producing a fusion gene of the *MLL* gene to the *CALM* (Clathrin Assembly Lymphoid Myeloid Leukemia); the inv(16) fusion gene CBFB-MYH11 (13); the NUP98- DDX10 fusion gene from the inversion (11) (p15q22) (14) and many others.

It is unclear how many fusion genes there are in the human genome. But it appears that chromosome rearrangements that lead to chromosome translocations, inversions or deletions occur quite frequently and that these events may play an important role in evolution. Most fusion loci may not be transcribed. Therefore the direct effects of these translocations, inversions or deletions would be minimal if any. However, it has been shown that a fraction of fusion genes are transcribed and produce chimeric RNA molecules that are made of RNAs from two or more genes. In fact, many so-called "fusion genes" in cancer cells (15) are initially discovered not based on the fusion genes themselves at the genome DNA level but rather on identification of their fusion RNA products using the RT-PCR approach that detects fusion cDNAs in various cases, without necessarily confirming the existence of a corresponding gene in the cell genome (16, 17). With the expansion of expressed sequence tag (EST) and cDNA databases and improvement of next generation sequencing technologies, an increased number of fusion RNAs and fusion genes have been identified and validated in recent years (2-4), especially in cancers. However, although it is commonly understood that fusion genes in cancers play a role in tumorigenesis, the mechanisms of this remain unclear. One explanation as to why fusion genes can cause tumors is that a fusion product from two or more genes

can have oncogenic functions that its parent genes do not possess. It is also possible that a pre-oncogene is fused to a strong promoter so that its expression is up-regulated, inducing tumors. Similarly, an anti-oncogene can be fused to a weaker promoter, leading to its down-regulation, which results in cancers. Nonetheless, the fact that many tumors have fusion genes leads to the development of diagnostic tests to detect cancers at their earliest stages.

CHIMERIC RNAs PRODUCED THROUGH RNA PROCESSING EVENTS (Fig. 1)

Chimeric RNAs can be generated from fusion genes as described above. It is also possible that some fusion RNAs may simply be RNA chimeras derived from other events such as transcription read-through, transcription of short homologous sequence slippage, and/or trans-splicing (Fig. 1) (1, 18-22). Interestingly, a growing number of studies have linked the trans-splicing process in normal cells to the formation of fusion genes in cancer cells, i.e., trans-splicing products may play a role in forming fusion genes.

Although there is not currently a survey of the frequencies of fusion RNAs occurring in cells, it is our impression that malignancies such as prostate cancers, breast cancers and ovarian cancers are reported at high frequency to involve abnormal RNA processing. For instance, a recent deep-sequencing analysis supported this assumption by validating the presence of more than 60 fusion RNAs in ovarian cancer tissues (23). Based on these studies, it is hypothesized that a profile of abnormal RNA chimera species could become a diagnostic biomarker tool, which would have an impact on tumor detection by making early diagnosis possible. This could lead to a significant increase in cancer survival.

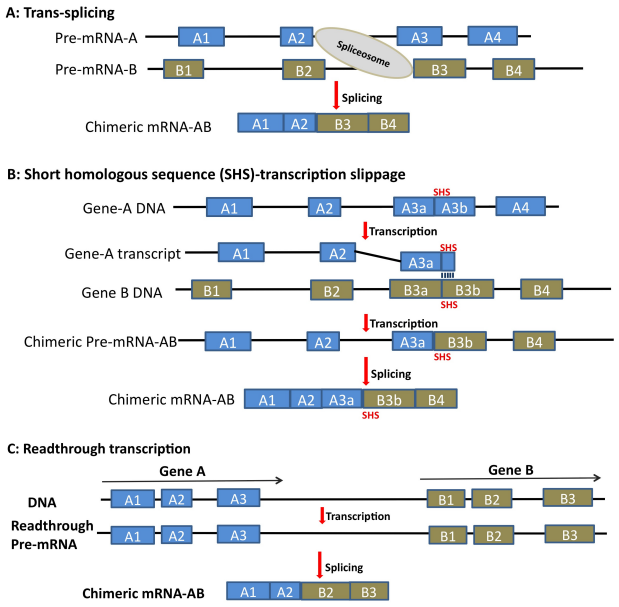

Fig. 1. RNA processing events that generate chimeric RNAs. (A) Trans-splicing: spliceosome links Exon 2 in Gene A to Exon 3 in Gene (B) producing chimeric mRNA-AB; (B) Short homologous sequence (SHS) slippage model: a transcript from Gene A with a SHS sequence in Exon A3b pairing with a SHS sequence in Exon B3b in Gene B continues transcription, generating chimeric pre-mRNA-AB. (C) Transcription reads through from Gene A to Gene B. Splicing then produces chimeric mRNA of Gene A and Gene B. Box: exon; line: intron.

Mechanisms of Cis-RNA splicing (see a recent reviewer by McManus & Graveley) (24)

In eukaryotes, an RNA segment with exons and introns is produced after transcription. The intron size is usually much larger than the exon size. The introns must be removed for the translation process to produce protein. Splicing varies in different organisms. In humans, at least ~90% of genes are spliced, whereas only ~4% of genes are spliced in yeast. The intron size of higher metazoans is much larger than the size in lower metazoans. The intron size is ten times larger than the exon size in humans, whereas it is much smaller in yeast. Although the regulation of splicing is varied in different organisms, the core splicing machinery is conserved.

Pre-mRNA splicing occurs in the large RNA-protein complex called spliceosome. Spliceosome is assembled on the pre-mRNA through a step-wise process. The components of spliceosome are divided into two parts. The first part is U1, U2, U4, U5 and U6 snRNPs (small nuclear RNA protein complexes); the second part comprises proteins including U2AF65 and SR proteins. The base pairing between U snRNA and splicing signals play key roles in the catalysis step of splicing.

When the pre-mRNA splicing occurs within a single pre-mRNA, it is considered to be cis-splicing. Cis pre-mRNA splicing happens in two steps pathways through phosphadiester bond formation. In the first step, a 2′-5′ phosphodiester bond is formed when the phosphate group at 5′ splice site attacks the hydroxyl group of the adenine at branch. In the second step, a 3′-5′ phosphodiester bond is formed when the phosphate group at 3′ splice site attacks the hydroxyl group of 5′ splice site. After cleavage, two exons are ligated to form a mature RNA. The key sequences that are required for pre-mRNA splicing include the 5′ splice site, 3′ splice site, branch point and polypyrimidine tract. The mechanisms by which this occurs still need to be identified.

RNA trans-splicing to produce chimeric RNA

Cis-splicing is a normal biological process that generates a mature mRNA from one single pre-mRNA. However, it has been recently found that splicing can also occur between two pre-mRNAs, which is known as trans-splicing (25). Trans-splicing can take place between two copies of the same pre-mRNA, resulting in an RNA containing duplicated exons. One example of this is the 77 kD estrogen receptor alpha (ERα) (26-28). Trans-splicing can also take place from two opposite strands of

the DNA double helix of the same gene, resulting in an RNA that contains oppositely oriented sequences (25). However, for most trans-splicing events, the two pre-mRNAs are transcribed from different genes, leading to a chimeric RNA (29). Trans-splicing is common in unicellular organisms and Caenorbabditis elegans. Trans-splicing has been also found in Drosophila and mammals. Astonishingly, the recent ENCODE pilot project demonstrated that transcripts from about 65% of the genes in human may be involved in the formation of chimeric RNAs (29, 30). In addition, paired-end transcriptome sequencing (16, 17) and/or the bioinformatics approach (31) have identified thousands more putative chimeric RNAs in human cells. For example, one of the databases established by a recent study contains more than 30,000 expressed sequence tags (EST) as putative chimeric RNAs (32), raising a serious question as to whether the concept of "gene" needs to be redefined (33). While most of these chimeric RNAs are not validated, there are many authentic examples that show the existence of trans-spliced chimeric RNAs which include examples such as the chimeric RNA between fatty-acid synthase (FAS) and ERα from prostate cancer cells. The chimeric RNA between fatty-acid synthase (FAS) and ERα has also been found to be expressed in cell lines of breast cancer and other cancer types (34). However, since chromosomal rearrangement of either FAS or the ERα gene is rare if it exists at all, its wide expression in many cell lines suggests that it is derived from RNA processing, likely trans-splicing. For the same reason, many other ERα-containing RNA chimeras that have been confirmed or reported (35, 36) may be also derived from trans-splicing events rather from fusion genes. Finally, as a better example of trans-splicing, the mouse Msh4 gene on chromosome 3 produces several chimeras mRNAs with other transcripts from either the same (chromosome 3) or different chromosomes including chromosomes 2,10 or 16 (37).

How do trans-splicing and other RNA processing events produce chimeric RNAs?

While cis-splicing occurs within the same pre-mRNA molecule through the formation of spliceosome, which pulls two adjacent exons together, it is not clear how trans-splicing could take place between two or more different pre-mRNA molecules. There have been numerous publications using both *in vitro* and in vivo assays that support the notion that trans-splicing may be responsible for the production of tens of thousands of chimeric RNAs. However, the exact mechanisms of this event for the majority of chimeric RNAs remain largely unconfirmed. In fact, in contrast to the popular and logical hypothesis that trans-splicing generates many chimeric RNAs, a recent paper by Li *et al.* proposed a short transcriptional slippage model (32). The authors conducted a large-scale search for chimeric RNAs in yeast, fruit fly, mouse and human samples and identified 5 chimeric RNAs in yeast, 4,084 in fruit fly, 10,586 in mouse, and more than 30,000 putative chimeric RNAs from humans. They estimated that approximately 1/3 of these chimeric RNAs are authentic based on small scale RT-PCR validation assay. Surprisingly, they found by further bioinformatics analysis that 50% of these putative chimerical RNAs did not possess the canonical GU-AU splice sites but rather had short homologous sequences (SHSs) at the chimerical junction sites of the source sequences. They hypothesized that this type of chimeric RNAs is probably produced through transcriptional slippage mediated by the pairing of short homologous sequences (SHSs) between two genes. As a result, they suggested that "chimeric transcripts" rather than "trans-splicing" should be used to define most chimeric RNAs. However, the authors also confirmed the existence of chimeric RNAs that are probably produced by the classic trans-splicing model, since about 20% of putative chimeric RNAs have a typical GU-AG splicing consensus sequence at the chimeric junctions. While the percentage (50%) of putative RNA chimeras that contain SHSs at junctions is striking, there is a possibility that this number is skewed because two thirds of the chimeric RNAs that the authors used to draw this conclusion were false positives. If false positives that contain SHSs sequences are disproportionally distributed, the percentage of chimeric RNAs that are produced from trans-splicing may increase or decrease. This notion is supported by a more recent paper by Kannan *et al.* (7), which demonstrated that the majority of confirmed chimeric RNAs in their studies contained the typical 5′ and 3′ sequences. However this does not adequately explain how two different pre-mRNAs come to join together. There is a possibility that two pre-mRNAs share common RNA binding proteins in a spliceosome and therefore may be pulled together for trans-splicing (Fig. 1). It is also possible that homologous sequences play a role in trans-splicing. A case in point is the fact that artificial trans-genes that have homologous sequences with endogenous genes can undergo trans-splicing in cells to replace endogenous exons. This strategy has been successfully used in cells to target several genes including tau and SMN with splicing defects for potential treatment of diseases such as tauopathies and spinal muscular atrophy (SMA) (20, 38-41). Nonetheless, it seems that there are at least three mechanisms to explain how chimeric RNAs can be produced from RNA processing events: trans-splicing; read-through transcription, and SHSs mediation (Fig. 1). However, to elucidate the exact mechanisms, a sufficient number of chimeric RNAs that are generated from RNA processing events need to be validated and analyzed. This is currently one of the major challenges in the studies of chimeric RNA.

THE CHALLENGE OF IDENTIFYING AND VALIDATING CHIMERIC RNAs

Sequencing transcriptomes and whole genomes has generated an enormous amount of data with more results being rapidly accumulated on a daily basis. As a result, mining these databases has become a major challenge as well as an opportunity for many scientists. Multiple software and platforms including

defuse, fusionmap, fusionseq, fusionhunter, tophat-fusion, chimerascan and snowshoes have been designed specifically to identify chimeric RNAs, leading to the discovery of tens of thousands of putative chimeric RNAs in humans as well as in other species from existing databases and newly acquired deep-sequencing results (2, 4, 6, 23, 42-44). However, despite this seeming progress, the number of chimeric RNAs that have been independently confirmed or validated is surprisingly small. Li *et al*. (32) predicted based on sequence analysis that more than 30,000 ESTs are putative chimeric RNAs in humans, but they only validated about one third of a small number of candidates from yeast and Drosophila in their publication. In another recent paper, only 60 chimeric RNAs were confirmed as positive while 61 putative chimeric RNAs were proved to be false chimeric RNAs from large amount of RNA-seq results of 40 ovarian cancer tissues (23). In another study of prostate cancers, only 27 out of 42 putative chimeric RNAs were confirmed among more than 2000 candidates (7). Furthermore, the situation is complicated because relatively few chimeric RNAs have been cross-confirmed by different studies. For instance, most chimeric RNAs identified by Edgren *et al*. (3) were shown to be derived from fusion genes while Kannan *et al* demonstrated that different sets of chimeric RNAs were more likely generated from trans-splicing (7). Even in the same study, different chimeric RNAs were most likely found in different samples (23), indicating that although we have the technologies to acquire chimeric RNA data, validation remains a challenge.

There are several factors that may have hindered efforts to identify and validate RNA chimeras. For one, the expression levels of many chimeric RNAs, especially those from events such as trans-splicing, SHS slippage (32) and transcription read-through that occurs at RNA levels, are low. Conventional Northern blot analysis has a low sensitivity which may not be sufficient to detect these low level chimeric RNAs. On the other hand, fusion RNAs that are generated from fusion genes would have an expression level that could be detectable by Northern blot. Fusion RNAs that are generated from fusion genes can be also validated on genome levels by methods such as in situ hybridization. Alternatively, high sensitivity of RT-PCR has been proven useful for validation of many known chimeric RNAs (23, 32). However RT-PCR has its own limitations and can sometimes introduce false results. Interestingly, RNase protection assay, the method that we believe could provide a reliable approach to confirm the existence of chimeric RNAs, has not yet been used in previous publications.

Another challenge for studies of chimeric RNA is what databases and which technologies should be chosen for reliably identifying these RNAs. With the rapid evolution of genomic research, next generation sequencing has become widely available. However, deep sequencing and existing databases also have limitations. For example, the length of sequencing is somewhat inadequate for complete analysis of exon junctions. Alignment of sequences to specific regions is also occasionally problematic, especially when repeated sequences are present or there are homologous fragments between chromosomal sequences and mitochondrial sequences. As a result, sequences in databases or sequences directly from RNA-seq often contain information that leads to false results when they are analyzed with bioinformatics approaches. This problem would be amplified with sequences such as chimeric RNAs that span multiple segments and/or that have low expression levels. Nevertheless, in order to realize the potential for chimeric RNAs to be utilized as biomarkers, it is essential to improve our technologies for confirmation of these molecules.

CAN CHIMERIC RNAs BE USED FOR DIAGNOSIS AND PROGNOSIS?

Chimeric RNA was discovered quite some years ago, but a majority of studies have focused on RNA chimeras that are produced from fusion genes. Next generation sequencing on the whole genome has facilitated identification of fusion genes. However it is estimated that only a small fraction of identified fusion genes from genome sequencing are transcribed into functional fusion RNAs, making this method inefficient for examining the roles of fusion genes in cancers. On the other hand, more recently developed RNA-seq technology has demonstrated that sequencing transcriptome can be efficiently used to detect fusion RNAs, leading to the characterization of fusion genes in cancers (3). The continued improvements of high-throughput sequencing technologies and bioinformatics analytic tools have also spurred studies into the roles of chimeric RNAs that are generated from RNA processing such as trans-splicing in cancer biology and diagnostics. In a recently published paper (7), Kannan *et al*. performed high-throughput sequencing analysis of 10 prostate cancer samples. They identified 2,349 chimeric RNA candidates and found that a subset of these chimeric RNAs are not detectable in primary human prostate epithelium control cells, indicating that these chimeras are only associated with prostate cancers. Specific chimeric RNAs that are produced through fusion genes or trans-splicing have been also described in other recent publications on tumors, including those relating to ovarian cancers and breast cancers (3, 5, 7, 23). While we have discussed the challenges involved in using chimeric RNAs as biomarkers for diagnostics and prognostics, these recent studies provide encouraging evidence that with improvements in sequencing technology and bioinformatics analytic tools, and more effective validation approaches, signature RNA chimeras could be identified for the early detection of tumors. In fact Skotheim *et al*. have recently proved that we can monitor expression of RNA chimeras in cancers with technologies such as microarray (45) if confirmed chimeric RNAs are available. This will bring hope to cancer patients.

Acknowledgements
This work was supported by the Priority Academic Program

Development of the Jiangsu Higher Education Institution (PAPD) (JZ), P. R. China; the Mid-career Researcher Program through a National Research Foundation (NRF) grant (2011-0000188 and 2011-0016757) funded by the Ministry of Education, Science, and Technology (MEST), Korea; the Korea Healthcare Technology R&D Project, Ministry for Health, Welfare and Family Affairs (A100733-1102-0000100) and a Systems Biology Infrastructure Establishment Grant provided by GIST in 2011.

REFERENCES

1. Akiva, P., Toporik, A., Edelheit, S., Peretz, Y., Diber, A., Shemesh, R., Novik, A. and Sorek, R. (2006) Transcription-mediated gene fusion in the human genome. *Genome Res.* **16**, 30-36.
2. Asmann, Y. W., Hossain, A., Necela, B. M., Middha, S., Kalari, K. R., Sun, Z., Chai, H. S., Williamson, D. W., Radisky, D., Schroth, G. P., Kocher, J. P., Perez, E. A. and Thompson, E. A. (2011) A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. *Nucleic Acids Res.* **39**, e100.
3. Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A. andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W. K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C. L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaoz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Loytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrimsdottir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B. and de Jong, P. J. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816.
4. Coady, T. H. and Lorson, C. L. (2010) Trans-splicing-mediated improvement in a severe mouse model of spinal muscular atrophy. *J. Neurosci.* **30**, 126-130.
5. Communi, D., Suarez-Huerta, N., Dussossoy, D., Savi, P. and Boeynaems, J. M. (2001) Cotranscription and intergenic splicing of human P2Y11 and SSF1 genes. *J. Biol. Chem.* **276**, 16561-16566.
6. Dotzlaw, H., Alkhalaf, M. and Murphy, L. C. (1992) Characterization of estrogen receptor variant mRNAs from human breast cancers. *Mol. Endocrinol.* **6**, 773-785.
7. Edgren, H., Murumagi, A., Kangaspeska, S., Nicorici, D., Hongisto, V., Kleivi, K., Rye, I. H., Nyberg, S., Wolf, M., Borresen-Dale, A. L. and Kallioniemi, O. (2011) Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome. Biol.* **12**, R6.
8. Flouriot, G., Brand, H., Seraphin, B. and Gannon, F. (2002) Natural trans-spliced mRNAs are generated from the human estrogen receptor-alpha (hER alpha) gene. *J. Biol. Chem.* **277**, 26244-26251.
9. Ge, H., Liu, K., Juan, T., Fang, F., Newman, M. and Hoeck,

W. (2011) FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics* **27**, 1922-1928.
10. Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S. and Snyder, M. (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res.* **17**, 669-681.
11. Gingeras, T. R. (2009) Implications of chimaeric non-co-linear transcripts. *Nature* **461**, 206-211.
12. Ha, K. C., Lalonde, E., Li, L., Cavallone, L., Natrajan, R., Lambros, M. B., Mitsopoulos, C., Hakas, J., Kozarewa, I., Fenwick, K., Lord, C. J., Ashworth, A., Vincent-Salomon, A., Basik, M., Reis-Filho, J. S., Majewski, J. and Foulkes, W. D. (2011) Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines. *BMC Med. Genomics* **4**, 75.
13. Hirano, M. and Noda, T. (2004) Genomic organization of the mouse Msh4 gene producing bicistronic, chimeric and antisense mRNA. *Gene* **342**, 165-177.
14. Horiuchi, T. and Aigaki, T. (2006) Alternative trans-splicing: a novel mode of pre-mRNA processing. *Biol. Cell* **98**, 135-140.
15. Iyer, M. K., Chinnaiyan, A. M. and Maher, C. A. (2011) ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* **27**, 2903-2904.
16. Kannan, K., Wang, L., Wang, J., Ittmann, M. M., Li, W. and Yen, L. (2011) Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 9172-9177.
17. Kim, D. and Salzberg, S. L. (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12**, R72.
18. Kundu, M. and Liu, P. P. (2001) Function of the inv(16) fusion gene CBFB-MYH11. *Curr. Opin. Hematol.* **8**, 201-205.
19. Li, H., Wang, J., Ma, X. and Sklar, J. (2009) Gene fusions and RNA trans-splicing in normal and neoplastic human cells. *Cell. Cycle.* **8**, 218-222.
20. Li, H., Wang, J., Mor, G. and Sklar, J. (2008) A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science* **321**, 1357-1361.
21. Li, X., Zhao, L., Jiang, H. and Wang, W. (2009) Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes. *J. Mol. Evol.* **68**, 56-65.
22. Li, Y., Chien, J., Smith, D. I. and Ma, J. (2011) FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics* **27**, 1708-1710.
23. Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N. and Chinnaiyan, A. M. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97-101.
24. Maher, C. A., Palanisamy, N., Brenner, J. C., Cao, X., Kalyana-Sundaram, S., Luo, S., Khrebtukova, I., Barrette, T. R., Grasso, C., Yu, J., Lonigro, R. J., Schroth, G., Kumar-Sinha, C. and Chinnaiyan, A. M. (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12353-12358.
25. McManus, C. J. and Graveley, B. R. (2011) RNA structure and the mechanisms of alternative splicing. *Curr. Opin. Genet. Dev.* **21**, 373-379.
26. McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M. G., Griffith, M., Heravi Moussavi, A., Senz, J., Melnyk, N., Pacheco, M., Marra, M. A., Hirst, M., Nielsen, T. O., Sahinalp, S. C., Huntsman, D. and Shah, S. P. (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.* **7**, e1001138.
27. Morerio, C., Acquila, M., Rapella, A., Tassano, E., Rosanda, C. and Panarello, C. (2006) Inversion (11)(p15q22) with NUP98-DDX10 fusion gene in pediatric acute myeloid leukemia. *Cancer Genet. Cytogenet.* **171**, 122-125.
28. Murphy, L. C., Dotzlaw, H., Hamerton, J. and Schwarz, J. (1993) Investigation of the origin of variant, truncated estrogen receptor-like mRNAs identified in some human breast cancer biopsy samples. *Breast. Cancer Res. Treat.* **26**, 149-161.
29. Nacu, S., Yuan, W., Kan, Z., Bhatt, D., Rivers, C. S., Stinson, J., Peters, B. A., Modrusan, Z., Jung, K., Seshagiri, S. and Wu, T. D. (2011) Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med. Genomics* **4**, 11.
30. Nowell, P. C. (1962) The minute chromosome (Phl) in chronic granulocytic leukemia. *Blut.* **8**, 65-66.
31. Pflueger, D., Terry, S., Sboner, A., Habegger, L., Esgueva, R., Lin, P. C., Svensson, M. A., Kitabayashi, N., Moss, B. J., MacDonald, T. Y., Cao, X., Barrette, T., Tewari, A. K., Chee, M. S., Chinnaiyan, A. M., Rickman, D. S., Demichelis, F., Gerstein, M. B. and Rubin, M. A. (2011) Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome. Res.* **21**, 56-67.
32. Pink, J. J., Fritsch, M., Bilimoria, M. M., Assikis, V. J. and Jordan, V. C. (1997) Cloning and characterization of a 77-kDa oestrogen receptor isolated from a human breast cancer cell line. *Br. J. Cancer* **75**, 17-27.
33. Pink, J. J., Wu, S. Q., Wolf, D. M., Bilimoria, M. M. and Jordan, V. C. (1996) A novel 80 kDa human estrogen receptor containing a duplication of exons 6 and 7. *Nucleic. Acids. Res.* **24**, 962-969.
34. Anthony, K., Garcia-Blanco, M. A., Mansfield, S. G., Anderton, B. H. and Gallo, J. M. (2009) Correction of tau mis-splicing caused by FTDP-17 MAPT mutations by spliceosome-mediated RNA trans-splicing. *Hum. Mol. Genet.* **18**, 3266-3273.
35. Santo, E. E., Ebus, M. E., Koster, J., Schulte, J. H., Lakeman, A., van Sluis, P., Vermeulen, J., Gisselsson, D., Ora, I., Lindner, S., Buckley, P. G., Stallings, R. L., Vandesompele, J., Eggert, A., Caron, H. N., Versteeg, R., and Molenaar, J. J. (2011) Oncogenic activation of FOXR1 by 11q23 intrachromosomal deletion-fusions in neuroblastoma. *Oncogene.* doi: 10.1038/onc.2011.344.
36. Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D. Z., Rozowsky, J. S., Tewari, A. K., Kitabayashi, N., Moss, B. J., Chee, M. S., Demichelis, F., Rubin, M. A. and Gerstein, M. B. (2010) FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol.* **11**, R104.
37. Shababi, M., Glascock, J. and Lorson, C. L. (2011)

Combination of SMN trans-splicing and a neurotrophic factor increases the life span and body mass in a severe model of spinal muscular atrophy. *Hum. Gene. Ther.* **22**, 135-144.
38. Shababi, M. and Lorson, C. L. (2012) Optimization of SMN Trans-Splicing Through the Analysis of SMN Introns. *J. Mol. Neurosci.* **46**, 459-469.
39. Shiga, Y., Sagawa, K., Takai, R., Sakaguchi, H., Yamagata, H. and Hayashi, S. (2006) Transcriptional readthrough of Hox genes Ubx and Antp and their divergent post-transcriptional control during crustacean evolution. *Evol. Dev.* **8**, 407-414.
40. Skotheim, R. I., Thomassen, G. O., Eken, M., Lind, G. E., Micci, F., Ribeiro, F. R., Cerveira, N., Teixeira, M. R., Heim, S., Rognes, T. and Lothe, R. A. (2009) A universal assay for detection of oncogenic fusion transcripts by oligo microarray analysis. *Mol. Cancer* **8**, 5.
41. Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., Bando, M., Ohno, S., Ishikawa, Y., Aburatani, H., Niki, T., Sohara, Y., Sugiyama, Y. and Mano, H. (2007) Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448**, 561-566.
42. Takahashi, T., Sonobe, M., Kobayashi, M., Yoshizawa, A., Menju, T., Nakayama, E., Mino, N., Iwakiri, S., Sato, K., Miyahara, R., Okubo, K., Manabe, T. and Date, H. (2010) Clinicopathologic features of non-small-cell lung cancer with EML4-ALK fusion gene. *Ann. Surg. Oncol.* **17**, 889-897.
43. Wechsler, D. S., Engstrom, L. D., Alexander, B. M., Motto, D. G. and Roulston, D. (2003) A novel chromosomal inversion at 11q23 in infant acute myeloid leukemia fuses MLL to CALM, a gene that encodes a clathrin assembly protein. *Genes Chromosomes Cancer* **36**, 26-36.
44. Yang, Y. and Walsh, C. E. (2005) Spliceosome-mediated RNA trans-splicing. *Mol. Ther.* **12**, 1006-1012.
45. Ye, Q., Chung, L. W., Li, S. and Zhau, H. E. (2000) Identification of a novel FAS/ER-alpha fusion transcript expressed in human cancer cells. *Biochim. Biophys. Acta.* **1493**, 373-377.