J Mol Evol (2009) 68:56–65
DOI 10.1007/s00239-008-9187-0

Short Homologous Sequences Are Strongly Associated with the Generation of Chimeric RNAs in Eukaryotes

Xin Li · Li Zhao · Huifeng Jiang · Wen Wang

Received: 3 September 2008 / Accepted: 17 November 2008 / Published online: 17 December 2008

Abstract Chimeric RNAs have been reported in varieties of organisms and are conventionally thought to be produced by *trans*-splicing of two or more distinct transcripts. Here, we conducted a large-scale search for chimeric RNAs in the budding yeast, fruit fly, mouse, and human. Thousands of chimeric transcripts were identified in these organisms except in yeast, in which five chimeric RNAs were observed. RT-PCR experiments for a sample of yeast and fly chimeric transcripts using specific primers show that about one-third of these chimeric RNAs can be reproduced. The results suggest that at least a considerable amount of chimeric RNAs is unlikely from aberrant transcription or splicing, and thus formation of chimeric RNAs is probably a widespread process and can greatly contribute to the complexity of the transcriptome and proteome of organisms. However, only a small fraction (<20%) of these chimeric RNAs has GU-AG at the junction sequences which fits the classical *trans*-splicing model. In contrast, we observed that about half of the chimeric RNAs have short homologous sequences (SHSs) at the junction sites of the source sequences. Our sequence mutation experiments in yeast showed that disruption of SHSs resulted in the disappearance of the corresponding chimeric RNAs, suggesting that SHSs are essential for generating this kind of chimeric RNA. In addition to the classical *trans*-splicing model, we propose a new model, the transcriptional slippage model, to explain the generation of those chimeric RNAs synthesized from templates with SHSs.

Keywords Chimeric RNAs · Short homologous sequences · *trans*-Splicing · Transcriptional slippage · Complexity of transcriptome

Introduction

Chimeric RNAs refer to three kinds of transcripts, i.e., transcripts encoded by two or more loci, transcripts encoded by two different stands of the same locus, and transcripts with shuffled exon order compared to genomic DNA sequences. Conceivably chimeric RNAs can increase the complexity of transcriptomes and proteomes and, thus, contribute to evolution of organisms. Chimeric RNAs have sporadically been reported in varieties of eukaryotes (Horiuchi et al. 2003; Mayer and Floeter-Winter 2005), but whether chimeric RNAs occur widely is still unknown. By taking advantage of enormous expression data for those model organisms, we extensively searched for chimeric RNAs in the budding yeast, fruit fly, mouse, and human in this study and identified many chimeric RNAs.

Trans-splicing is conventionally thought to be the process that produces chimeric RNAs (Horiuchi and Aigaki 2006; Mayer and Floeter-Winter 2005). This process can currently be classified into two types based on the molecular process of generation: SL (spliced leader)-addition

X. Li and L. Zhao contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9187-0) contains supplementary material, which is available to authorized users.

X. Li · L. Zhao · H. Jiang · W. Wang (✉)
CAS-Max Planck Junior Research Group on Evolutionary Genomics, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences (CAS), Kunming 650223, China
e-mail: wwang@mail.kiz.ac.cn

X. Li · L. Zhao · H. Jiang
Graduate School of Chinese Academy of Sciences, Beijing 100049, China

trans-splicing and non-SL-addition *trans*-splicing. SL-addition *trans*-splicing does not contribute to proteomic diversity since the SL sequence at the 5′ termini of their mRNAs is a noncoding sequence. So far, SL-addition *trans*-splicing has been observed in some protozoa, nematodes, and chordates (Horiuchi and Aigaki 2006; Maniatis and Tasic 2002; Mayer and Floeter-Winter 2005; Nilsen 2001), but no evidence has been found for this type of *trans*-splicing in yeast, fruit fly, and vertebrates. For non-SL-addition *trans*-splicing, several early in vitro experiments using cell-free system have shown that mammalian cells have the ability to join RNA segments from two separate precursor molecules by *trans*-splicing, suggesting that this kind of *trans*-splicing reaction could take place in vivo in eukaryotic cells (Konarska et al. 1985; Solnick 1985). Later, in vivo *trans*-splicing was reported in plant organelles (Chapdelaine and Bonen 1991; Kück et al. 1987; Koller et al. 1987). In the early 1990s, chimeric transcripts were discovered in mammalian cells, and *trans*-splicing was proposed to account for them (Joseph et al. 1991; Shimizu et al. 1989, 1991; Sullivan et al. 1991; Vellard et al. 1992). Subsequently such so-called "*trans*-splicing" events have been reported in many plant and animal species including rice, fruit fly, mosquito, chicken, mouse, rat, and human (Dorn et al. 2001; Fitzgerald et al. 2006; Hirano and Noda 2004; Horiuchi et al. 2003; Kawasaki et al. 1999; Robertson et al. 2007; Zhao et al. 2006). Some of these "*trans*-splicing" events have been shown to be essential and biologically significant (Horiuchi et al. 2003; Mongelard et al. 2002).

The process of non-SL-addition *trans*-splicing is thought to be similar to that of canonical *cis*-splicing, and both of them are proposed to use the same splicing machinery to generate mature mRNA molecules (Horiuchi and Aigaki 2006; Maniatis and Tasic 2002; Mayer and Floeter-Winter 2005). However, it is still far from clear how the spliceosome can join exons from different pre-mRNA molecules during *trans*-splicing process. It has been proposed that the individual *trans*-splicing precursors may interact through specific base pairing or through interactions among proteins binding to each of the precursors (Maniatis and Tasic 2002). The former proposal has been supported by some studies through analysis of the sequences involved (Dixon et al. 2007). This classical *trans*-splicing model can explain the reported intragenic *trans*-splicing cases by which exons from independently transcribed pre-mRNA molecules from one or different alleles of one gene are joined together, because all these cases of intragenic *trans*-splicing have the canonical splice sites (GU-AG) at the junction of the hybrid mRNA (Caudevilla et al. 1998; Frantz et al. 1999; Horiuchi et al. 2003; Takahara et al. 2000). Similarly, a few intergenic *trans*-splicing events occurring between closely linked genes are also consistent with this model (Finta and Zaphiropoulos 2000, 2002; Tasic et al. 2002; Zaphiropoulos 1999). However, a lot of other reported "*trans*-splicing," especially interchromosomal "*trans*-splicing" cases, does not conform to the classical splicing model that needs the canonical GU-AU splice sites. This absence of canonical splice sites at the junction positions of these chimeric RNAs suggests that other splicing-unrelated molecular mechanisms may be involved in these so-called "*trans*-splicing" events (Mayer and Floeter-Winter 2005; Unneberg and Claverie 2007). So it is prudent not to use the term "*trans*-splicing" for those chimeric RNAs without clear canonical splice sites at their junctions. Hereafter in this paper, we use "chimeric transcripts" instead of "*trans*-splicing" unless the splice site is clear.

In this study, we first performed a genome-wide screen for chimeric transcripts in budding yeast, fruit fly, mouse, and human. Surprisingly, we found that a high proportion (up to 25–49%) of total genes was involved in the formation of chimeric transcripts, except in the yeast, which has very low expressed sequence tag (EST) coverage. The numbers of detected chimeric RNAs are 5 in yeast, 4084 in fruit fly, 10,586 in mouse, and 31,005 in human. Our experimental data show that at least a considerable proportion of these chimeric RNAs is real rather than from artifacts. We also discovered that the classical *trans*-splicing model can only explain <20% of the formation of chimeric RNAs. However, we observed that about half of the inspected chimeric RNAs have short homologous sequences (SHSs) at the junction sites between the two source DNA sequences that encode them. These SHSs exist in a direct repeat manner, and sequence mutation experiments in yeast show that disruption of SHSs results in the disappearance of the corresponding chimeric transcripts, suggesting that SHSs are essential for generating chimeric RNAs. Therefore, to explain this result, we proposed that this kind of chimeric RNAs is probably produced by transcriptional slippage mediated by these SHSs rather than by "*trans*-splicing."

Materials and Methods

Identification of Chimeric Transcripts

We downloaded genome sequences and mRNAs/Refseqs/ESTs of yeast (sacCer1), fruit fly (dm2), mouse (mm7), and human (hg18) from UCSC (http://hgdownload.cse.ucsc.edu/downloads.html) and mapped these transcriptional sequences (mRNAs/Refseqs/ESTs) to their corresponding genomic sequences using BLAT with the default parameters (Kent 2002). EST library information and EST annotation were downloaded from NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/repository/UniLib/library.report

and ftp://ftp.ncbi.nih.gov/repository/UniGene/). We then parsed the raw BLAT results with a series of Perl scripts. First, to remove parologous and random spurious hits, we only retained those hits which have alignable lengths longer than 50 bp and sequence similarities in the aligned regions >95%. Second, if >80% of aligned regions of one hit overlap with that of another hit, we defined them as overlapped hits. We put all overlapped hits of one query sequence into one group, and thereby hits from a chimeric RNA were divided into two or more groups if they resulted from two or more source DNAs. To further remove paralogous hits, we only retained the best hit with the highest score within each group. If multiple hits within a group have similar scores exceeding our threshold, this transcript was excluded because of its ambiguous origin. Third, we further scrutinized these sequences and only retained those sequences which match one of the following three criteria: (a) sequences with at least two hits located within one chromosome but on different strands, (b) sequences with at least two hits located on different chromosomes, or (c) sequences which have exons with a different order from the corresponding genomic DNA. Fourth, we removed those chimeric sequences with poly(A) or poly(T) sequences longer than 10 bp or recognition sites of restriction enzymes used for their corresponding cDNA library construction at their junction positions, because these chimeric sequences may be generated by artificial processes during cDNA library construction. We also removed those chimeric sequences generated from mitochondrial genes. Finally, only sequences passing the above filtration process were considered as candidate chimeric RNAs. Then we clustered chimeric RNAs into different groups based on their mapping information, and each group contains chimeric RNAs from same source loci. We identify those groups whose RNAs were from two or more different ESTs library according to available EST annotation.

Confirmation of Chimeric mRNAs by RT-PCR and Sequencing

We extracted total RNA from embryo, larva, pupa, and adult of *Drosophila melanogaster* using the RNeasy Mini RNA extraction kit (Qiagen). Yeast cells were harvested from 20 ml of culture at $OD_{600} = 1.0$ and then resuspended in RNAlater solution (Ambion). We extracted yeast total RNA using Trizol (Tiangen, China) and then subjected it to DNase I (MBI) digestion. First-strand cDNA was synthesized using Oligo-dT and SuperScript II RNase H- reverse transcriptase (Invitrogen). Primers are designed based on the sequences of the chimeric transcripts to make the RT-PCR products cover the junction position. For fruit fly, we performed RT-PCR using cDNA pooled from developmental stages of egg, larva, pupa, and adult as templates. All RT-PCR products were purified and sequenced for verification.

Exclusion of the Possibility that Chimeric mRNAs Resulted from RT-PCR Artifacts

To exclude the possibility that the chimeric mRNAs resulted from RT-PCR artifacts, we amplified two DNA fragments which contain sequences corresponding to the 5′ and 3′ parts of one chimeric mRNA. DNA fragments were cut out from the gel and purified using TIANgel Purification Kit (Tiangen, China). The two DNA fragments were pooled and used as templates for RT-PCR.

We also conducted an in vitro transcription assay. The two DNA fragments were each cloned into the pBluescript II KS + plasmid. RNAs representing the 5′ and 3′ parts of one chimeric mRNA were then transcribed from the plasmids using the AmpliScribe T7 and T3-Flash Transcription kits (Epicentre), and their cDNAs were synthesized using gene-specific primers and SuperScript II RNase H- reverse transcriptase (Invitrogen). cDNAs were then purified using the QIAquick PCR Purification Kit (Qiagen). The two fragments of cDNA were also pooled and used as templates for RT-PCR.

If the chimeric product were a PCR artifact, we would see successful amplification of the observed chimeric cDNAs using the above two fragments of templates. If we did not see this, production of the chimeric cDNA must have been dependent on some cellular processes such as *trans*-splicing or other mechanisms.

Allele Replacement in Yeast

To test whether the observed SHSs at the junction positions of the two source DNAs of a chimeric RNA are essential to the formation of the chimeric RNA, we mutated SHS sites of the two yeast genes *SPT7* and *LYS12*. *Saccharomyces cerevisiae* strain BY4742 (*MATα*; *his3*; *leu2*; *ura3*; *lys2*) and two deletion strains, *SPT7* (*spt7Δ*::KanMX4) and *LYS12* (*lys12Δ*::KanMX4), were used for this study. The two deletion strains were from the complete set of yeast deletion strains in the BY4742 background purchased from EUROSCARF (http://web.uni-frankfurt.de/fb15/mikro/euroscarf/). Yeast strains were grown at 30°C on yeast-peptone-dextrose (YPD) medium.

Site-directed mutagenesis was performed according to the method of Ho et al. (1989). Specific alterations in nucleotide sequence were introduced by incorporating nucleotide changes into the overlapping oligonucleotide primers. The oligonucleotide primers used were as follows: spt7_F_1, TTGAACTTCGGATCAGTGAAAATTG; spt7_R_1, gctcc TGTtgaacgcatagccACTAATATCATATTCCTGTAGGAA TCTGGTAT; spt7_F_2, TggctatgcgttcaACAggagcTACGA

GGGAGTAAATACTAAAACATTAG; spt7_R_2, gtgcactctcagtacaatctTTATTGATTAAGGCGAGGAAGGC; lys12_F_1, AGAAACTGAACTAATGGCAGCAAGG; lys12_R_1, CTTtgcaagtgGtaaggtTCAATGTATGTTTTTTCAATTTTAATtgcaAttgaCTCAGTATTTTCTCTGACGATAACCA; lys12_F_2, ACATTGAaccttaCcacttgcaAAGAGTTGCTGATGCCACAAAG; and lys12_R_2, gtgcactctcagtacaatctCTATAATCTCGACAAAACGTCGTCA (lowercase letters indicate mutated sites). The mutations do not change the coding ability of the two genes. The pRS306 plasmid was used as PCR template to amplify the cassette of the *URA3* marker. The mutant allele and *URA3* marker were fused by adaptamer-mediated PCR. The fusion DNA fragment was transformed into the deletion strains to replace the KanMax4 marker using a lithium acetate procedure. Synthetic complete (SC) medium plates lacking uracil were used for *URA3*$^{+}$ transformant selection.

Results

Identification of Chimeric Transcripts

We downloaded total transcripts (ESTs/mRNAs/Refseqs) of budding yeast (a total of 35,140), fruit fly (a total of 572,146), mouse (a total of 4,987,242), and human (a total of 8,205,899) from UCSC for our initial screen. We finally identified 5 chimeric RNAs in yeast, 4084 in fruit fly, 10,586 in mouse, and 31,005 in human (chimeric RNAs listed in Supplementary Table S1). The numbers of genes involved in the formation of chimeric transcripts are 9 in yeast, 3558 in fruit fly, 7922 in mouse, and 11,643 in human, accounting for 0.13% (9/6697), 25% (3558/14,039), 33% (7922/23,786), and 49% (11,643/23,686) of the total genes based on the current annotation from the Ensembl database (Release 47, October 2007) in yeast, fruit fly, mouse, and human, respectively. It is interesting that multiple chimeric transcripts were identified in *S. cerevisiae* although very few genes in *S. cerevisiae* have introns, implying that a splicing-unrelated mechanism might underlie the generation of the chimeric transcripts in *S. cerevisiae*.

To characterize the chimeric RNAs identified in this study, we used those nonredundant chimeric RNAs which consist of only two distinct transcripts for all analyses conducted thereafter. We did not include chimeric RNAs containing more than two distinct transcripts due to the difficulty of classifying them into a particular group when we conducted statistical analysis. In fact, chimeric RNAs consisting of two transcripts are the majority of our data sets (all in yeast; fruit fly, 2700/2793 = 97%; mouse, 8685/9131 = 95%; human, 26,493/27,067 = 98%).

First, we examined the chromosomal distribution of these chimeric transcripts. We found that the number of source genes for chimeric transcripts on one chromosome is positively correlated with the gene number in that chromosome (Fig. 1a–c). This result indicates that there is no biased creation of chimeric mRNAs among chromosomes.

Second, we found that 16–30% of the chimeric transcripts (fruit fly, 430/2700 = 16%; mouse, 2645/8685 = 30%; human, 6848/26,493 = 26%) come from the same locus. The two parts of these chimeric RNAs either come from different strands of the same locus or have an exon order different from that of genomic DNA. This amount of intragenic chimeric RNAs suggests that the process of generation of chimeric RNAs may be space-related and most easily occurs between genomic regions that are in close proximity to each other. This result that chimeric RNAs prefer to be intragenic also suggests that at least parts of identified chimeric RNAs are genuine transcripts in the cell rather than artificial products during cDNA library construction. Because if chimeric RNAs are artificial products from recombination events during cDNA library construction, we would expect that two segments of these artificial chimeric sequences would come from different loci randomly rather than one locus preferably.

Third, we examined how many of these chimeric RNAs can encode chimeric proteins. If a putative protein encoded

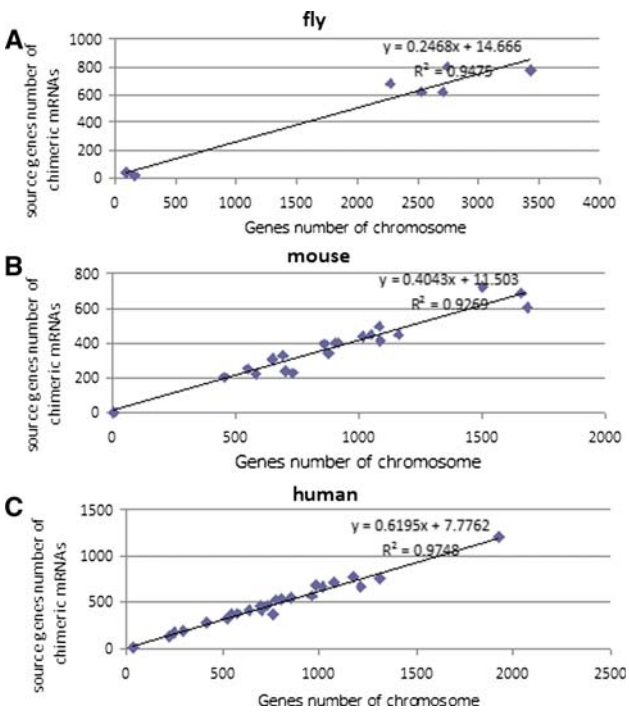

Fig. 1 Correlation between total gene number and source gene number of chimeric RNAs on a chromosome. **a** Fly; **b** mouse; **c** human

by a chimeric RNA covers through the junction position and each of the two distinct transcripts can encode more than 10 amino acids, we considered it a chimeric protein. Our results show that many of these chimeric RNAs (fruit fly, 362/2700 = 13%; mouse, 874/8685 = 10%; human, 2069/26,493 = 8%) can encode chimeric proteins, suggesting that these chmeric RNAs could contribute remarkably to the diversity of the proteome.

To confirm whether or not the identified chimeric RNAs are genuine transcripts in cells, we randomly picked 25 candidates from those in fruit fly and all candidates in yeast for RT-PCR confirmation. For fly, we used the pooled cDNA from embryo, larva, pupa, and adult of *Drosophila melanogaster* as the PCR template. In all, 8 of the 25 candidates in fruit fly and two of five candidates in yeast can be confirmed by our experiments (Fig. 2a, c). Because we identified a large number of chimeric RNAs in fly, mouse, and human, even if only one-third of the chimeric RNAs are real, the amount of chimeric RNAs in these organisms is still more than what we had appreciated. Furthermore, using the available EST and cDNA library annotation, we identified a considerable proportion of chimeric RNAs which occurred in at least two different EST libraries in mouse and human, suggesting the authenticity of these chimeric RNAs (see Supplementary Table S2).

To exclude the possibility that these chimeric transcripts were artifacts of random mismatching during the PCR processes, we picked five RT-PCR-confirmed cases for further verification. First, we amplified two DNA fragments which contained the source sequences corresponding to the 5′ and 3′ parts of one chimeric mRNA. Then the two DNA fragments were pooled and used as templates for RT-PCR. Second, we transcribed two RNAs which contained the corresponding 5′ and 3′ part of a chimeric RNA and synthesized their corresponding cDNAs by reverse transcription *in vitro*. We also pooled these two cDNAs and used them as templates for RT-PCR. If these chimeric mRNAs were RT-PCR artifacts, we would expect that we could also amplify them in these RT-PCR experiments using above two RT-PCR approaches. But our results showed that no such products were detected in any of the five cases tested (data not shown), suggesting the authenticity of these chimeric RNAs in cells.

Short Homologous Sequences are found at the Junction Sites of Many Chimeric RNAs

The classical *trans*-splicing model for generating chimeric RNAs requires splice sites (GU-AG) at the junctions of the source sequences (Fig. 3a). However, only a few reported interchromosomal chimeric RNAs are consistent with this prediction. We collected all the reported

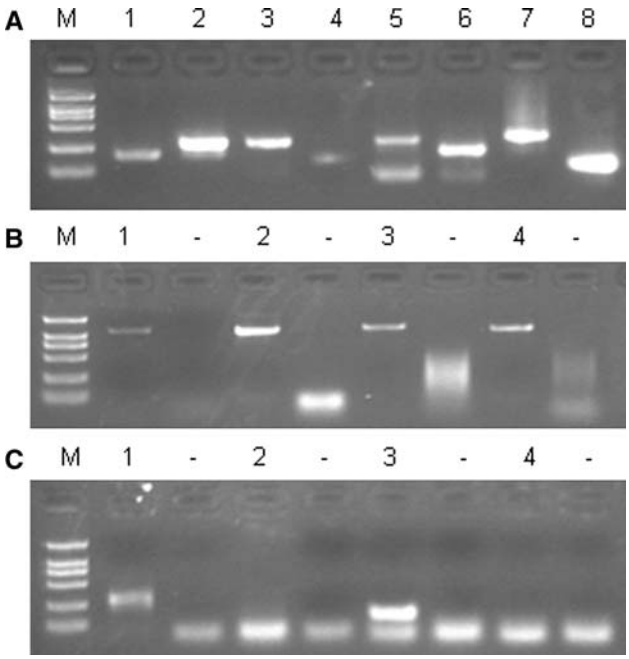

Fig. 2 RT-PCR results for verification of identified chimeric transcripts and experimental test for the transcriptional slippage model in yeast. M, DNA size marker. Minus signs represent negative controls of RT-PCRs. **a** RT-PCR results for CO327830 (lane 1), AI135867 (lane 2), EC067642 (lane 3), CO269652 (lane 4), CO281322 (lane 5), AI064259 (lane 6), EC243729 (lane 7), and CO187473 (lane 8), respectively. Amplified fragments were confirmed by sequencing. **b** RT-PCR results of *SPT7* (lane 1) and *LYS12* (lane 2) in the wild-type strain BY4742. RT-PCR results of *SPT7* (lane 3) and *LYS12* (lane 4) in each mutant strain. **C** RT-PCR results for chimeric transcripts of *LYS12* (lane 1) and *SPT7* (lane 3) in the wild-type strain BY4742. RT-PCR results for the chimeric transcripts of *LYS12* (lane 2) and *SPT7* (lane 4) in each mutant strain, showing disappearance of the chimeric RNAs

interchromosomal chimeric transcripts and analyzed the junction sequences of these cases in detail (Table 1). Only 3 of 13 (23%) reported chimeric RNAs (*ABP-HDC*, M38759; *Burs*, AY735442; and *DMRT1-CENP C1*, AY448020) have canonical splice sites (GU-AG) at the junctions (Fig. 4a–c). Interestingly, SHSs, which exist in a direct repeat manner, are present at the junction sites between the two source sequences of a chimeric RNA in four reported cases (31%) (*Oaz3*, DQ431007; *Msh4-Hspa*, AY351588; *DMRT1-CD5R*, AY448021; and *DMRT1-37LRP/p40*, AY448022) (Fig. 4d–f).

To assess the generality of the above observations, we manually scrutinized the junction sites for all the chimeric RNAs in yeast and the 200 junction sequences of the chimeric RNAs with the highest supported transcript numbers in each species of fruit fly, mouse, and human. We also considered CU-AC as canonical splice sites because some ESTs were deposited in the public databases as reverse complementary sequences. Only 9.5% (19/200), 18% (36/200), and 17.5% (35/200) of examined chimeric sequences

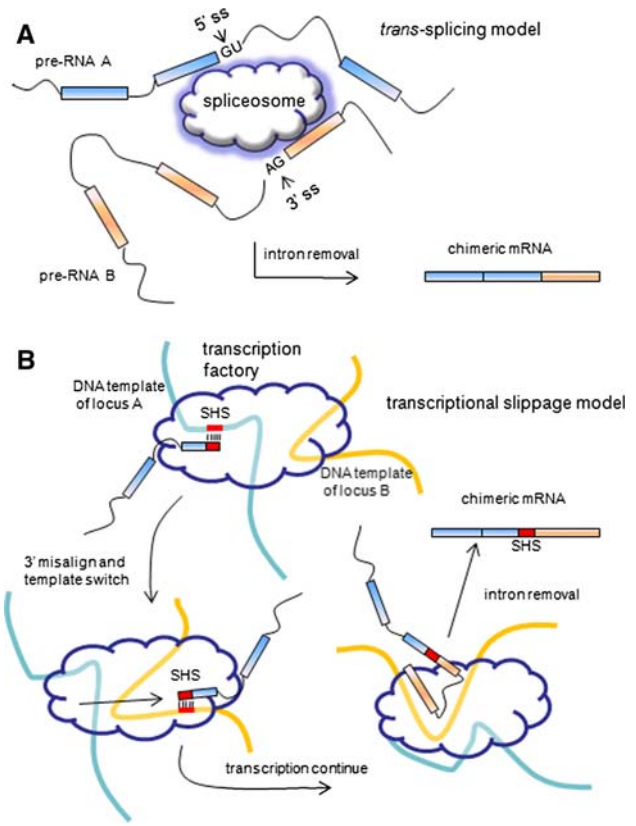

Fig. 3 Two models to explain the generation of chimeric transcripts. Exons and introns are shown as boxes and lines, respectively. Blue and orange represent different genes. **a** *Trans*-splicing model. Pre-RNA A and pre-RNA B provide the 5′ and 3′ splice sites, respectively. The two precursor pre-RNAs generate a chimeric transcript through the splicing process. **b** Transcriptional slippage model. Red boxes represent short homologous sequences (SHSs). Locus A and locus B are active and share one transcription factory. Locus A first transcribes a pre-RNA and then misaligns to the DNA template of locus B through the SHSs. Transcription continues at the locus B and the chimeric transcript is finally generated after intron removal

have canonical splice sites in fruit fly, mouse, and human, respectively (Fig. 5). To our surprise, about half of these sequences (fruit fly, 65.5%, 131/200; mouse, 52%, 104/200; human, 51.5%, 103/200) have SHSs ($\geq$4 bp) between two source sequences at their junctions (Fig. 5 and Supplementary Table S3). In yeast, none of the identified chimeric sequences has canonical splice sites, while all of them have SHSs between the two source sequences at their junctions (Supplementary Table S3). All these SHSs are directly repeated, and the majority of them are short (<10 bp). Some of these chimeric RNAs with SHSs are intragenic (fruit fly, 28%, 37/131; mouse, 57%, 59/104; human, 46%, 47/103) (from the same genomic locus but with different orientation or have a different exon order from that of genomic DNA) while many others are intergenic or interchromosomal (fruit fly, 72%, 94/131; mouse, 43%, 45/104; human, 54%, 56/103).

Short Homologous Sequences are Essential for Generating SHS-Containing Chimeric RNAs

Our analysis shows that SHSs exist widely at the very junction sites of chimeric RNAs. Because most of these kinds of chimeric RNAs do not have canonical splice sites, we speculated that SHSs may play a role in the process of generating chimeric RNAs. To test our hypothesis that SHSs may be necessary for the generation of the himeric transcripts, we used budding yeast for further experimental investigation due to convenient genetic tools in this organism. The above RT-PCR experiments confirmed the existence of chimeric transcripts at both the *SPT7* and the *LYS12* loci in yeast. The 5′ and 3′ parts of each of these two chimeric RNAs are located at the same genomic locus, respectively, but with a different orientation, and they have SHSs at the junctions of the two source sequences. We replaced the wild-type *SPT7* and *LYS12* loci with alleles mutated at the SHS junctions in the haploid yeast strain, BY4742, but without disrupting the two genes' coding ability. RT-PCRs were conducted for the wild-type and mutant strains to investigate whether the mutant strains still produce chimeric transcripts. Experimental results show that the regular *SPT7* and *LYS12* were both expressed in the wild-type and mutant strains (Fig. 2b), but the chimeric transcripts were detected only in the wild-type strain (Fig. 2c). These results provide direct evidence that mutation of SHSs can destroy the formation of the observed chimeric transcripts, strongly supporting our hypothesis that SHSs are essential for the formation of these kinds of chimeric RNAs.

Discussion

Our results here show that SHSs are essential for generating a novel kind of chimeric RNAs and about half of the chimeric RNAs may be generated through a SHS-dependent mechanism. This prompts us to propose a new molecular model, the "transcriptional slippage model" (Fig. 3b), to explain the generation of these chimeric transcripts. This model is different from the previous *trans*-splicing model (Fig. 3a). The new model proposes that when a pre-mRNA molecule is being transcribed, in some cases it dissociates from the template strand, and then the SHS at its 3′ end of pre-mRNA "misaligns" with the SHSs at another position of the same locus or another locus. A chimeric RNA can then be generated if the transcription process continues on the new template. This model has the following features. (1) It does not rely on canonical splice sites (GU-AG) at the junctions of source transcripts. The spliceosome is not required for generation of chimeric RNAs and thus it is much simpler than the *trans*-splicing

Table 1 Reported interchromosomal chimeric mRNAs

Species	5′ gene	3′ gene	5′ location	3′ location	Accession No.	Have GU-AG?	Fit our slippage model?	Reference
Rat	*Shbg*	*Hdc*	Chr10	Chr3	M38759	Yes		Sullivan et al. (1991)
Human	*CAMK2G*	*SRP72*	Chr10	Chr4	U81554			Breen and Ashcroft (1997)
Human	*KCND2*	*SOTA1*	Chr7	Chr1	L21934			Li et al. (1999)
Rat	*Ptprf*	noncoding	Chr5	Chr1	X83546			Zhang et al. (2003)
Mouse	noncoding	*Msh4*	Chr16	Chr3	AY351586			Hirano and Noda (2004)
Mouse	*Bcbp3*	*Msh4*	Chr10	Chr3	AY351589			Hirano and Noda (2004)
Mouse	*Hspa5*	*Msh4*	Chr2	Chr3	AY351588		Yes	Hirano and Noda (2004)
Rat	*Oaz3*	noncoding	Chr2	Chr4	DQ431007		Yes	Fitzgerald et al. (2006)
Rice	*Os06g0726600*	*Os03g0128700*	Chr6	Chr10	D13436			Kawasaki et al. (1999)
Mosquito	*burs124*	*burs3*	Chr2L	Chr2R	AY735442	Yes		Robertson et al. (2007)
Chicken	*DMRT1*	*Cenp C1*	ChrZ	Chr4	AY448020	Yes		Zhao et al. (2006)
Chicken	*DMRT1*	*CD5R*	chrZ	Chr5	AY448021		Yes	Zhao et al. (2006)
Chicken	*DMRT1*	*37LRP/p40*	chrZ	Chr2	AY448022		Yes	Zhao et al. (2006)

Fig. 4 Junction sequence analysis of reported interchromosomal chimeric mRNAs. Uppercases and lowercases represent exon and intron sequences, respectively. **a–c** Cases with canonical splice sites (underlined) at the junction. **b** AY735442 chimeric RNA is from two rounds of *trans*-splicing and thus has two junctions. Underlined letters represent splice sites. **d–g** Cases with homologous sequences at the junctions. Stars represent short homologous sequences (SHSs)

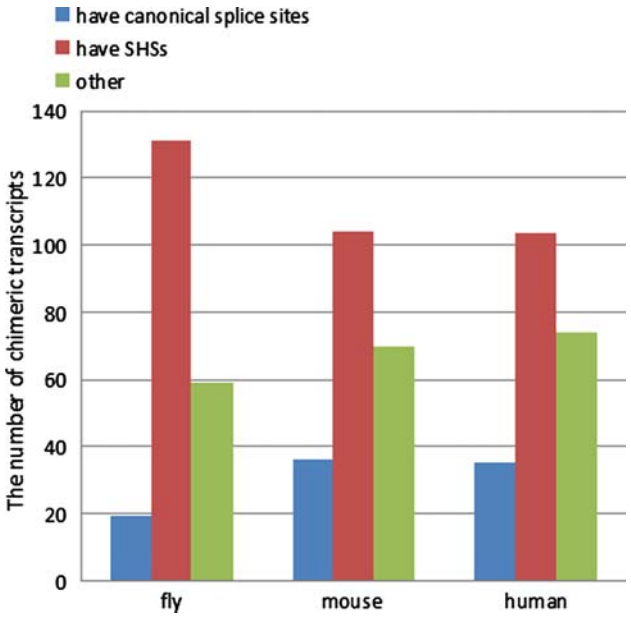

Fig. 5 Contribution of different molecular mechanisms to chimeric RNA formation in fly, mouse, and human

model. (2) For transcriptional slippage to occur, two loci involved in the generation of chimeric RNAs should be simultaneously active and close enough to occupy the same transcription factory. (3) SHSs between two loci are necessary for efficient template-switching to generate the mature chimeric RNAs.

Indeed many previous results suggest that this transcriptional slippage is possible. A study has shown that multiple genes can occupy the same transcription factory (Jackson et al. 1998). Another study showed that active

genes are dynamically organized into shared nuclear subcompartments and even distal genes can colocalize at the same transcription factory at high frequencies (Osborne et al. 2004). Later studies showed that genes from different chromosomes can also occupy the same transcription factory (Chuang and Belmont 2006; Ling et al. 2006). We speculate that colocalization of active genes to the same transcription factory is a prerequisite for the occurrence of generation of a SHS-dependent chimeric RNA. Most importantly transcriptional slippage has been reported in viruses, bacteria, and yeasts and some examples have been shown to have great functional significance (Baranov et al. 2005; Fabre et al. 2002; Wagner et al. 1990). Some human diseases are also proposed to be caused by transcriptional slippage (van Leeuwen et al. 1998; van den Hurk et al. 2001).

Alternatively it is also possible that these SHS-dependent chimeric RNAs were produced by *trans*-splicing reactions through sequence pairing or binding proteins, but the following two observations do not support this speculation. First, these SHSs exist in a direct repeat manner and thus cannot directly link the two precursor RNAs through complementary base pairing and mediate the *trans*-splicing reaction. Second, the SHSs are different in different chimeric RNAs (Supplementary Table S3), thus it is unlikely that cells produce enough specific binding proteins to bind to different SHSs to link two precursor RNAs together. Therefore, based on our data and previous studies, we speculate that transcriptional slippage model may be the most likely mechanism to generate SHS-dependent chimeric RNAs in vivo. Although the molecular mechanism and the key components of the "transcription factory" involved in this model still require further investigation, it is the most parsimonious explanation for the formation of related chimeric RNAs based on the current evidence.

It is noteworthy that about one-third of the chimeric RNAs in fruit fly, mouse, and human still can be explained neither by the *trans*-splicing model nor by our transcriptional slippage model (Fig. 5), indicating that there may be other, unknown molecular mechanisms leading to the generation of some chimeric RNAs. Alternatively, those chimeric RNAs may simply have resulted from aberrant transcription or splicing.

It is also noteworthy that our RT-PCR could confirm only about one-third of the cases tested. Four explanations may exist for this result. First, some chimeric RNAs may be under strict regulation and exist only at some specific stages or under conditions which were not collected in our RNA samples. Second, some chimeric RNAs may be present at a low abundance and our RT-PCR approach may be unable to detect them, but a large-scale cDNA sequencing approach such as emulsion PCR in 454 sequencing (Leamon et al. 2006) could detect them. Third, it is also possible that some of the chimeric RNAs may be "noise" from aberrant splicing or transcription, which has been reported in some previous studies (Tasic et al. 2002). Fourth, some of these chimeric RNAs could be from artifacts of cDNA library construction. Further experiments, for example, using northern blot to detect chimeric RNAs using junction sequences as the probes or, for some chimeric RNAs which may have the ability to encode chimeric proteins, using western blot to detect the existence of putative proteins in the cells, could provide the solid piece of evidence to verify the authenticity of the chimeric RNAs.

Despite these possibilities, some of the chimeric RNAs may be real and have important functional and evolutionary significances as revealed by the *mod (mdg4)* and *lola* genes in *Drosophila* (Horiuchi et al. 2003; Mongelard et al. 2002). It has also been reported that some chimeric RNAs are conserved among species (Gabler et al. 2005; Robertson et al. 2007). These cases may reflect functional constraint during long-term evolution.

The widespread existence of such chimeric RNAs suggests that the formation of chimeric RNAs might have important functional significance during the evolution of organisms. The discrepancy among the complexity of organisms and the relatively conserved gene numbers have been puzzling biologists for many years (Venter et al. 2001). In addition to alternative splicing (Boue et al. 2003; Graveley 2001), we have demonstrated that the origin of new exons in orthologous genes is also an important mechanism contributing to the complexity of the proteome in higher animals (Li et al. 2007; Wang et al. 2005). In this study, we identified a large number of chimeric transcripts in yeast, fly, mouse, and human. Considering that some chimeric RNAs have been shown to have important functions (Horiuchi et al. 2003; Mongelard et al. 2002), the large number of the chimeric transcripts identified in this study may also have important functional significance in cells and organisms, although such chimeric RNAs have not been sufficiently appreciated so far. These kinds of chimeric RNAs not only increase the complexity and diversity of the transcriptome and proteome of higher eukaryotes, but also enrich the concept of the gene, supporting the recently updated definition for genes that disjointed sets of genomic sequence also belong to one gene (Gerstein et al. 2007). In a sense, it also expands the concept of exon shuffling (Gilbert 1978) from the genomic DNA level to the transcriptional level.

Acknowledgments We thank Chris Tyler-Smith, Zhenglong Gu, Yali Xue, and Paul Lemetti's comments on and English editing of the manuscript. This work was supported by a CAS-Max Planck Society Fellowship, a NSFC key grant (No. 30430400), and a 973 Program grant (No. 2007CB815703-5) to W.W.

References

Baranov P, Hammer A, Zhou J, Gesteland R, Atkins J (2005) Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in IS element gene expression. Genome Biol 6:R25

Boue S, Letunic I, Bork P (2003) Alternative splicing and evolution. BioEssays 25:1031–1034

Breen MA, Ashcroft SJH (1997) A truncated isoform of Ca2+/calmodulin-dependent protein kinase II expressed in human islets of Langerhans may result from trans-splicing. FEBS Lett 409:375–379

Caudevilla C, Serra D, Miliar A, Codony C, Asins G, Bach M, Hegardt FG (1998) Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. Proc Natl Acad Sci USA 95:12185–12190

Chapdelaine Y, Bonen L (1991) The wheat mitochondrial gene for subunit I of the NADH dehydrogenase complex: a trans-splicing model for this gene-in-pieces. Cell 65:465–472

Chuang CH, Belmont AS (2006) Close encounters between active genes in the nucleus. Genome Biol 6:237

Dixon RJ, Eperon IC, Samani NJ (2007) Complementary intron sequence motifs associated with human exon repetition: a role for intragenic, inter-transcript interactions in gene expression. Bioinformatics 23:150–155

Dorn R, Reuter G, Loewendorf A (2001) Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in Drosophila. Proc Natl Acad Sci USA 98:9724–9729

Fabre E, Dujon B, Richard G-F (2002) Transcription and nuclear transport of CAG/CTG trinucleotide repeats in yeast. Nucleic Acids Res 30:3540–3547

Finta C, Zaphiropoulos PG (2000) The human CYP2C locus: a prototype for intergenic and exon repetition splicing events. Genomics 63:433–438

Finta C, Zaphiropoulos PG (2002) Intergenic mRNA molecules resulting from trans-splicing. J Biol Chem 277:5882–5890

Fitzgerald C, Sikora C, Lawson V, Dong K, Cheng M, Oko R, van der Hoorn FA (2006) Mammalian transcription in support of hybrid mRNA and protein synthesis in testis and lung. J Biol Chem 281:38172–38180

Frantz SA, Thiara AS, Lodwick D, Ng LL, Eperon IC, Samani NJ (1999) Exon repetition in mRNA. Proc Natl Acad Sci USA 96:5400–5405

Gabler M, Volkmar M, Weinlich S, Herbst A, Dobberthien P, Sklarss S, Fanti L, Pimpinelli S, Kress H, Reuter G, Dorn R (2005) Trans-splicing of the mod(mdg4) complex locus is conserved between the distantly related species Drosophila melanogaster and D. virilis. Genetics 169:723–736

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) What is a gene, post-ENCODE? History and updated definition. Genome Res 17:669–681

Gilbert W (1978) Why genes in pieces? Nature 271:501–501

Graveley BR (2001) Alternative splicing: increasing diversity in the proteomic world. Trends Genet 17:100–107

Hirano M, Noda T (2004) Genomic organization of the mouse Msh4 gene producing bicistronic, chimeric and antisense mRNA. Gene 342:165–177

Ho SN, Hunt HD, Horton RM, Pullen JK, Pease LR (1989) Site-directed mutagenesis by overlap extension using the polymerase chain reaction. Gene 77:51–59

Horiuchi T, Aigaki T (2006) Alternative trans-splicing: a novel mode of pre-mRNA processing. Biol Cell 98:135–140

Horiuchi T, Giniger E, Aigaki T (2003) Alternative trans-splicing of constant and variable exons of a Drosophila axon guidance gene, lola. Genes Dev 17:2496–2501

Jackson DA, Iborra FJ, Manders EMM, Cook PR (1998) Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei. Mol Biol Cell 9:1523–1536

Joseph DR, Sullivan PM, Wang Y-M, Millhorn DE, Bayliss DM (1991) Complex structure and regulation of the ABP/SHBG gene. J Steroid Biochem Mol Biol 40:771–775

Kück U, Choquet Y, Schneider M, Dron M, Bennoun P (1987) Structural and transcription analysis of two homologous genes for the P700 chlorophyll a-apoproteins in Chlamydomonas reinhardtii: evidence for in vivo trans-splicing. EMBO J 6:2185–2195

Kawasaki T, Okumura S, Kishimoto N, Shimada H, Higo K, Ichikawa N (1999) RNA maturation of the rice SPK gene may involve trans-splicing. Plant J 18:625–632

Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12:656–664

Koller B, Fromm H, Galun E, Edelman M (1987) Evidence for in vivo trans splicing of pre-mRNAs in tobacco chloroplasts. Cell 48:111–119

Konarska MM, Padgett RA, Sharp PA (1985) Trans-splicing of mRNA precursors in vitro. Cell 42:165–171

Leamon JH, Link DR, Egholm M, Rothberg JM (2006) Overview: methods and applications for droplet compartmentalization of biology. Nat Methods 3:541–543

Li B-L, Li X-L, Duan Z-J, Lee O, Lin S, Ma Z-M, Chang CCY, Yang X-Y, Park JP, Mohandas TK, Noll W, Chan L, Chang T-Y (1999) Human acyl-CoA:cholesterol acyltransferase-1 (ACAT-1) gene organization and evidence that the 4.3-kilobase ACAT–1 mRNA is produced from two different chromosomes. J Biol Chem 274:11060–11071

Li X, Liang J, Yu H, Su B, Xiao C, Shang Y, Wang W (2007) Functional consequences of new exon acquisition in mammalian chromodomain Y-like (CDYL) genes. Trends Genet 23:427–431

Ling JQ, Li T, Hu JF, Vu TH, Chen HL, Qiu XW, Cherry AM, Hoffman AR (2006) CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. Science 312:269–272

Maniatis T, Tasic B (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature 418:236–243

Mayer MG, Floeter-Winter LM (2005) Pre-mRNA trans-splicing: from kinetoplastids to mammals, an easy language for life diversity. Mem Inst Oswaldo Cruz 100:501–513

Mongelard F, Labrador M, Baxter EM, Gerasimova TI, Corces VG (2002) Trans-splicing as a novel mechanism to explain interallelic complementation in Drosophila. Genetics 160:1481–1487

Nilsen TW (2001) Evolutionary origin of SL-addition trans-splicing: still an enigma. Trends Genet 17:678–680

Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell JA, Lopes S, Reik W, Fraser P (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. Nature Genet 36:1065–1071

Robertson HM, Navik JA, Walden KKO, Honegger H-W (2007) The bursicon gene in mosquitoes: an unusual example of mRNA trans-splicing. Genetics 176:1351–1353

Shimizu A, Nussenzweig MC, Mizuta T-R, Leder P, Honjo T (1989) Immunoglobulin double-isotype expression by trans-mRNA in a human immunoglobulin transgenic mouse. Proc Natl Acad Sci USA 86:8020–8023

Shimizu A, Nussenzweig MC, Han H, Sanchez M, Honjo T (1991) Trans-splicing as a possible molecular mechanism for the

multiple isotype expression of the immunoglobulin gene. J Exp Med 173:1385–1393

Solnick D (1985) Trans-splicing of mRNA precursors. Cell 42:157–164

Sullivan PM, Petrusz P, Szpirer C, Joseph DR (1991) Alternative processing of androgen-binding protein RNA transcripts in fetal rat liver. Identification of a transcript formed by trans splicing. J Biol Chem 266:143–154

Takahara T, S-i Kanazu, Yanagisawa S, Akanuma H (2000) Heterogeneous sp1 mRNAs in human HepG2 cells Include a product of homotypic trans-splicing. J Biol Chem 275:38067–38072

Tasic B, Nabholz CE, Baldwin KK, Kim Y, Rueckert EH, Ribich SA, Cramer P, Wu Q, Axel R, Maniatis T (2002) Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. Mol Cell 10:21–33

Unneberg P, Claverie JM (2007) Tentative mapping of transcription-induced interchromosomal interaction using chimeric EST and mRNA data. PLoS ONE 2:e254

van den Hurk WH, Willems HJJ, Bloemen M, Martens GJM (2001) Novel frameshift mutations near short simple repeats. J Biol Chem 276:11496–11498

van Leeuwen FW, de Kleijn DP et al (1998) Frameshift mutants of beta amyloid precursor protein and ubiquitin-B in Alzheimer's and Down patients. Science 279:242–247

Vellard M, Sureau A, Soret J, Martinerie C, Perbal B (1992) A potential splicing factor is encoded by the opposite strand of the trans-spliced c-myb exon. Proc Natl Acad Sci USA 89:2511–2515

Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. Science 291:1304–1351

Wagner LA, Weiss RB, Driscoll R, Dunn DS, Gesteland RF (1990) Transcriptional slippage occurs during elongation at runs of adenine or thymine in Escherichia coli. Nucleic Acids Res 18:3529–3535

Wang W, Zheng H, Yang S, Yu H, Li J, Jiang H, Su J, Yang L, Zhang J, McDermott J, Samudrala R, Wang J, Yang H, Yu J, Kristiansen K, Wong GK-S, Wang J (2005) Origin and evolution of new exons in rodents. Genome Res 15:1258–1264

Zaphiropoulos PG (1999) RNA molecules containing exons originating from different members of the cytochrome P450 2C gene subfamily (CYP2C) in human epidermis and liver. Nucleic Acids Res 27:2585–2590

Zhang C, Xie Y, Martignetti JA, Yeo TT, Massa SM, Longo FM (2003) A candidate chimeric mammalian mRNA transcript Is derived from distinct chromosomes and Is associated with nonconsensus splice junction motifs. DNA Cell Biol 22:303–315

Zhao Y, H-s Yu, Lu H, Yao K, H-h Cheng, R-j Zhou (2006) Interchromosomal *trans-splicing* of DMRT1 gene on chicken chromosome Z. Zool Res 27:175–180