



## Looking ultra deep: Short identical sequences and transcriptional slippage

Katja Ritz <sup>a,d</sup>, Barbera D.C. van Schaik <sup>b</sup>, Marja E. Jakobs <sup>a</sup>, Eleonora Aronica <sup>c</sup>,  
Marina A. Tijssen <sup>d</sup>, Antoine H.C. van Kampen <sup>b,e</sup>, Frank Baas <sup>a,\*</sup>

<sup>a</sup> Department of Genome Analysis, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105AZ Amsterdam, The Netherlands

<sup>b</sup> Bioinformatics Laboratory, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105AZ Amsterdam, The Netherlands

<sup>c</sup> Department of (Neuro)Pathology, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105AZ Amsterdam, The Netherlands

<sup>d</sup> Department of Neurology, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105AZ Amsterdam, The Netherlands

<sup>e</sup> Biosystems Data Analysis, Swammerdam Institute for Life Science, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

### ARTICLE INFO

#### Article history:

Received 15 November 2010

Accepted 16 May 2011

Available online 23 May 2011

#### Keywords:

Deep sequencing

Template switching

Reverse transcriptase

Transcriptional slippage

Short homologous sequences

Chimeric RNA

### ABSTRACT

Studying transcriptomes by ultra deep sequencing provides an in-depth picture of transcriptional regulation and it facilitates the detection of rare transcriptional events. Using ultra deep sequencing of amplicons we identified known isoforms and also various new low frequency variants. Most of these variants likely involve the splicing machinery except for two events that we named variations affecting multiple exons, which are mainly deletions affecting parts of adjacent exons and intra-exonic deletions. Both events involve short identical sequences of 1 to 8 nucleotides at the junction and canonical splice sites are missing. They were identified in different genes and species at very low frequencies. We excluded that they are an artifact of PCR, sequencing, or reverse transcription. We propose that these variants represent intramolecular slippage events that require short identical sequences for reannealing of dissociated transcripts.

© 2011 Elsevier Inc. All rights reserved.

### 1. Introduction

Next generation sequencing technologies provide unprecedented insight in posttranscriptional processes. Whole transcriptome analysis revealed a more complex posttranscriptional regulation than initially believed. A recent deep sequencing study showed a much higher complexity of the rice transcriptome, many novel alternative splicing events, and transcripts, and variation in untranslated regions were identified [1]. Other studies reported millions of novel splice sites by deep sequencing of different tissues suggesting that more than 95% of human multi-exon genes are alternatively spliced [2–4]. In depth sequencing of transcriptomes facilitates the detection of rare events such as chimeric RNAs. Chimeric RNAs are characterized by showing partial alignment to two genes which do not always originate by genomic rearrangements [5]. They can either be evolved by trans-splicing [6], by co-transcription of adjacent genes and intergenic splicing [7], by transcriptional slippage [8], or they could be an artifact of cDNA synthesis and generated *in vitro* by the reverse transcriptase [9–11]. Some studies argue that many reported chimeric RNAs may be an artifact [12–14].

We recently conducted an ultra deep sequencing study to investigate alternative splicing events in different human brain regions of *SGCE*, a gene associated with the neurological movement

disorder myoclonus–dystonia. The entire *SGCE* cDNA was analyzed by ultra deep amplicon sequencing. We characterized the expression pattern of the major isoforms and reported 19 novel low frequency alternative exons [15]. Most of those new exons lead to a frameshift and early stop codons. They are likely to represent “noise” of RNA processing and are not considered to be functional. In addition to the inclusion of novel alternative exons many other events have been identified at very low frequencies (low frequency variants) that have not been previously reported. Most of these low frequency variants can be explained by conventional splicing except for two events that did not exhibit canonical splice sites but have short identical sequences at the junction suggesting a distinct mechanism. In this report we present and characterize these low frequency variants. We made similar observations for two other genes (*POLR2G* and *SLC25A3*) and showed that our observations are not an artifact of PCR, sequencing, or reverse transcription. Our data support the existence of transcriptional slippage and provides new evidence for a slippage driven mechanism generating chimeric RNAs.

### 2. Material and methods

#### 2.1. Sample collection and processing

Human tissue (brain (motor cortex), heart) was obtained from two control cases without history of neurological diseases (Case 1: male, 77 years, 7 h post-mortem delay; Case 2: male, 7 years, <8 h post mortem delay) from the Department of Neuropathology of the

\* Corresponding author at: Department of Genome Analysis, Academic Medical Center, Meibergdreef 9, 1105AZ Amsterdam, The Netherlands. Fax: +31 205669312.  
E-mail address: [f.baas@amc.uva.nl](mailto:f.baas@amc.uva.nl) (F. Baas).

Academic Medical Center (University of Amsterdam, The Netherlands). A blood sample from a third control case was included. Informed consent was obtained for research purposes in all cases. Mouse (cerebral cortex), rat (cerebral cortex) and zebrafish (100 embryos, 24 hpf) tissue was obtained from wild type animals (approved by local committee; strains: mouse: NMRI nu/nu, rat: PVG, zebrafish: TL). For total RNA isolation, 20 to 40 sections of frozen brain tissue (depending on tissue size, 20  $\mu$ m) were cut per sample and dissolved in TRIzol Reagent (Invitrogen, Breda, The Netherlands) and subsequently processed by the QIAcube instrument (RNeasy protocol, Qiagen, Venlo, The Netherlands). The PAXgene system (Qiagen) was used for total RNA isolation from blood samples.

## 2.2. Ultra deep amplicon sequencing

To cover the entire cDNA of genes that were ultra deep sequenced (*SGCE*, *POLR2G*, *SLC25A3*) overlapping PCR products were generated. cDNA was synthesized from 1  $\mu$ g of total RNA with SuperScript II reverse transcriptase (Invitrogen), oligodT<sub>12</sub>-VN primers and 1  $\mu$ l cDNA was subjected to a 10  $\mu$ l PCR reaction. PCR reactions were performed using fusion primers consisting of a 19 bp fixed sequence (Roche/454 GS FLX, A or B sequence at the 5' end) and a target-specific sequence (3' end). Also, a 5 nucleotide MID (multiplex identifier)-tag was incorporated to allow for multiplexing of samples. All amplicons were further processed according to manufacturer's instructions and sequenced with a 454 GS FLX system (Roche Diagnostics, Almere, The Netherlands) aiming for 10,000 sequence reads per PCR product. For primers and conditions see Supplementary Table 1.

## 2.3. Exclusion of PCR and RT-PCR artifacts

*SGCE* full-length cDNA was cloned into a pGEM-T Easy vector (Promega, Leiden, The Netherlands). This *SGCE* plasmid was used as template for a PCR to exclude that variants may be an artifact of the PCR reaction. As second control to exclude RT-PCR artifacts, we synthesized RNA from the *SGCE* plasmid with the HiScribe T7 *In Vitro* Transcription Kit (NEB, Westburg BV, Leusden, The Netherlands). The *in vitro* synthesized RNA was used for cDNA synthesis with SuperScript II reverse transcriptase (Invitrogen) and a gene-specific primer followed by a PCR and 454 deep sequencing. cDNA synthesis and the PCR were performed with the same conditions used for non-control samples. The input of cDNA was diluted so that the same number of cycles was necessary to yield similar amount of product. Primers amplified *SGCE* exon 7 to 12 (Supplementary Table 1 for primers and conditions) and both control samples were ultra deep sequenced from the reverse strand.

## 2.4. Bioinformatics

All sequence reads obtained by ultra deep sequencing were grouped by MID. The BLAT algorithm (BLAST-like alignment tool) was used to do a pairwise comparison between all sequences with a common MID to identify similar sequences within each sample. Grouping criteria were: (1) sequence length  $\geq 210$  nt, (2) percent identity  $\geq 98\%$ , (3) score  $\geq 105$ , and (4) query coverage  $\geq 98\%$ . The resulting groups were mapped to the respective chromosome (BLAT algorithm, hg18, mm9, rn4, danRer6). All groups that contained at least two sequence reads were analyzed in the UCSC genome browser and using CodonCode Aligner software 3.0.1 (Dedham, MA, USA).

## 3. Results

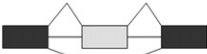

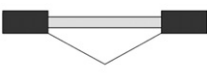

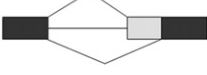

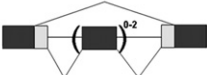
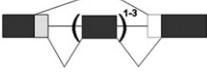
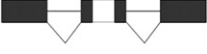
### 3.1. Characterization of low frequency variants

Analysis of the entire *SGCE* gene (four PCR products) in three tissues (heart, brain, blood; Case 1) showed that the majority of

sequence reads (average 98.7%, range 93.5%–100% for different amplicons and tissues) could be assigned to known isoforms (NM\_003919.2, NM\_001099400.1, NM\_001099401.1). In addition, 78 different low frequency variants were found (median frequency 0.05%, range 0.01%–5.4%,  $12,994 \pm 2459$  sequence reads analyzed on average per tissue and PCR product, see Supplementary Table 2). We only included variants that were detected at least twice. Eight different types of low frequency variants were identified: (1) inclusion of new alternative exons, (2) exon skipping (skipping of an exon previously reported as constitutive exon), (3) intron retention, (4) alternative 5' splice site, (5) alternative 3' splice site, (6) exon duplication, (7) variations affecting multiple exons and (8) intra-exonic deletions (Fig. 1A and B). Events 7 and 8 have not been reported before. Event 7 describes two observations: 7a) deletions affecting parts of adjacent exons and 7b) a deletion of part of the 3' end of an exon, followed by a 5' extended exon (both types involved skipping of 0 to 3 exons; Fig. 2A). Event 8 describes small deletions within one exon (Fig. 2B). In contrast to the first six events, the latter two (variations affecting multiple exons and intra-exonic deletions) did not exhibit canonical splice sites making involvement of the splicing machinery unlikely. Instead, 97.4% of variants classified to both types of events showed short identical sequences at the junctions from 1 to 8 nucleotides (Fig. 2). Short identical sequences were not a feature of either "new alternative exons" or "alternative 5' or 3' splice site". We identified 16 different variations affecting multiple exons and 23 different intra-exonic deletions in the *SGCE* gene (Fig. 1B). Both events were detected in brain and muscle tissue and not in blood.

To test whether the observed low frequency variants are gene- or sequence-specific two additional genes, *POLR2G* and *SLC25A3*, were analyzed by ultra deep sequencing in the same control case (brain, heart, blood; Case 1). *POLR2G* is a subunit of the RNA polymerase. This gene is highly conserved, essential and thus expressed in all tissues and only one isoform has been reported so far. A strict regulation and a low evolutionary turnover rate are expected for this gene. The mitochondrial phosphate carrier *SLC25A3* is a gene with tissue-specific expression like *SGCE*.

Analysis of ultra deep sequencing data for both genes (*POLR2G* and *SLC25A3*: two PCR products each) gave results similar to *SGCE*. For both genes, on average 99.1% (range: *POLR2G* 96.9%–100%; *SLC25A3* 96.8%–100% for different amplicons and tissues) of analyzed sequence reads could be assigned to known isoforms (*POLR2G*: NM\_002696.2; *SLC25A3*: NM\_002635.3, NM\_005888.3, NM\_213611.2), but also many low frequency variants were identified in all tissues (Fig. 1C and D). For *POLR2G*, 64 low frequency variants were identified (median frequency 0.03%, range 0.02%–2.0%,  $10,654 \pm 869$  sequence reads analyzed on average per tissue and PCR product, see Supplementary Table 2). For *SLC25A3*, 94 low frequency variants were found (median frequency 0.04%, range 0.02%–0.85%,  $8650 \pm 1167$  sequence reads in total, see Supplementary Table 2). All types of low frequency variants were found in both genes except for exon duplications which were only identified in *SGCE*. The two novel events were highly represented: 22 different variations affecting multiple exons and 22 different intra-exonic deletions for *POLR2G* and 60 variations affecting multiple exons and 26 different intra-exonic deletions for *SLC25A3*. The majority of variants classified as variations affecting multiple exons and intra-exonic deletions showed short identical sequences at the junction (*POLR2G*: 95.5%, *SLC25A3*: 93%; Fig. 3). The range of deletions was highly variable for all three genes and skipping of 14 to 158 nucleotides was observed for variations affecting multiple exons and 6 to 100 nucleotides for intra-exonic deletions. For variations affecting multiple exons three variants were identified that involved a partial deletion of one exon, followed by skipping of 1 to 3 exons, followed by a 5' extended exon (Fig. 1A, 7b). The range of the extensions varied from 3 to 186 nucleotides. In the three genes tested, variations affecting multiple exons and intra-exonic deletions were mainly identified in heart and brain tissue and to a lower extent or

| A                                      |  | B           | C             | D              |
|--|--|-------------|---------------|----------------|
|  |  | <i>SGCE</i> | <i>POLR2G</i> | <i>SLC25A3</i> |
| 1. new alternative exon                |     | 23 (27.7)   | 1 (1.5)       | 0              |
| 2. skipping of a constitutive exon     |     | 8 (9.6)     | 9 (13.0)      | 5 (5.2)        |
| 3. intron retention                    |     | 0           | 2 (2.9)       | 1 (1.0)        |
| 4. alternative 5' splice site          |     | 0           | 2 (2.9)       | 2 (2.1)        |
| 5. alternative 3' splice site          |     | 12 (14.5)   | 11 (15.9)     | 2 (2.1)        |
| 6. exon duplication                    |     | 1 (1.2)     | 0             | 0              |
| -----                                  |  |             |               |                |
| 7. variations affecting multiple exons | 7a  | 15 (18.1)   | 20 (29.0)     | 60 (62.5)      |
|  | 7b  | 1 (1.2)     | 2 (2.9)       | 0              |
| 8. intra-exonic deletion               |    | 23 (27.7)   | 22 (31.9)     | 26 (27.1)      |

**Fig. 1.** Events observed by ultra deep sequencing. Panel A) Eight different events were identified. Events 1 to 6 likely involve the splicing machinery. Event 7 and 8 did not exhibit canonical splice sites, but short identical sequences at the junctions. Event 7 covers two observations: (7a) deletions of parts of adjacent exons, including skipping of 0 to 2 exons and (7b) deletion of part of the 3' end of an exon, followed by skipping of 1 to 3 exons and a 5' extended exon. Panels B–D) Number of variants (percentage in brackets) identified per event for *SGCE* (panel B), for *POLR2G* (panel C), and *SLC25A3* (panel D). For *SGCE*, 78 low frequency variants were identified with 83 different events (some low frequency variants exhibited two different events), for *POLR2G* we found 64 low frequency variants and 69 different events, and 94 low frequency variants with 96 events for *SLC25A3*. Low frequency variants represent 1.3% of all analyzed sequence reads for *SGCE* (meaning that 98.7% sequence reads could be assigned to known isoforms), 0.9% for *POLR2G* and 0.9% for *SLC25A3*. Results for the three different tissues tested were combined per gene.

even absent in blood (Fig. 4). The majority of low frequency variants lead to frameshift and premature stop codon (*SGCE*: 75.6%; *POLR2G*: 64.1%; *SLC25A3*: 53.2%). All low frequency variants are listed in Supplementary Table 2, including their frequency, short identical sequences, splice site predictions, the effect on the reading frame as well as predictions for nonsense mediated decay.

### 3.2. Exclusion of artifacts

To exclude that low frequency variants represent artifacts we included two control samples: First, we performed a PCR using a *SGCE* plasmid as template (control 1) assuring that we start with only one sequence. If other variants than the template are detected after deep sequencing of control 1, they must have been introduced during the amplification step. Second, we synthesized RNA by *in vitro* transcription using the *SGCE* plasmid as template. This synthetic RNA was converted to cDNA synthesis followed by a PCR and 454 sequencing (control 2). This experiment should show whether the reverse transcriptase with its ability to switch templates is responsible for the generation of variants. Ultra deep sequencing yielded 12,289 sequence reads for control 1 and 12,852 sequence reads for control 2. Analysis of the sequence reads did not reveal any variant in either control; all sequence reads could be assigned to the template *SGCE* sequence. This supports our hypothesis that identified variants are not

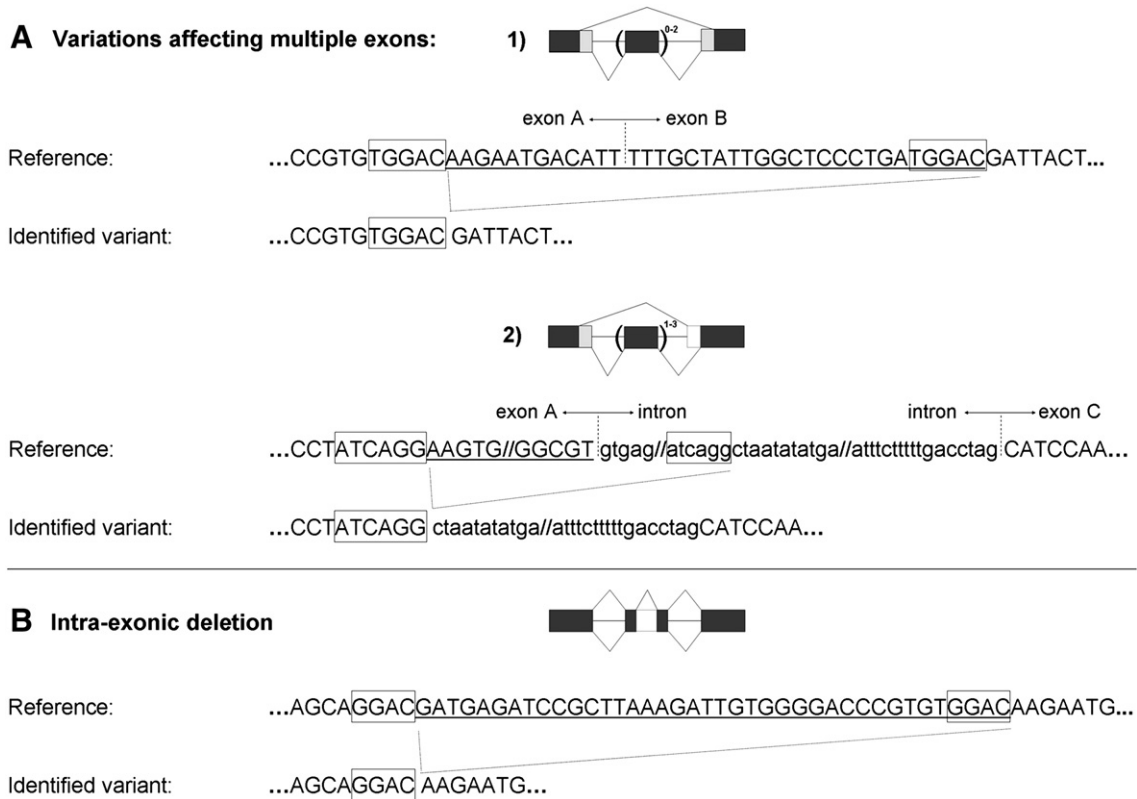
an artifact of the library preparation, but reflect mRNA molecules generated *in vivo*.

### 3.3. Low frequency variants are not age-related

To test whether the occurrence of low frequency variants is an age-dependent effect we analyzed two tissues (brain, heart) of two cases: a man (Case 1, 77 years old) and a boy (Case 2, 7 years old). We sequenced *SGCE* exon 7 to exon 12 and identified 14 low frequency variants in both Case 1 and in Case 2 (both tissues combined). Variations affecting multiple exons as well as intra-exonic deletions involving short identical sequences were identified (Tables 1 and 2) suggesting that our observations are not dependent on age.

### 3.4. Low frequency variants are not species-specific

To test whether the low frequency variants are also present in other species, *SGCE* (exon 7 to 12) was analyzed in mouse and rat brain and zebrafish embryos. In all cases we detected low frequency variants, including variations affecting multiple exons and intra-exonic deletions with the involvement of short identical sequences of 1 to 9 nucleotides (Tables 1 and 2) and the lack of canonical splice sites (Supplementary Table 2). Analysis of the zebrafish sample revealed a considerably higher number of low frequency variants and



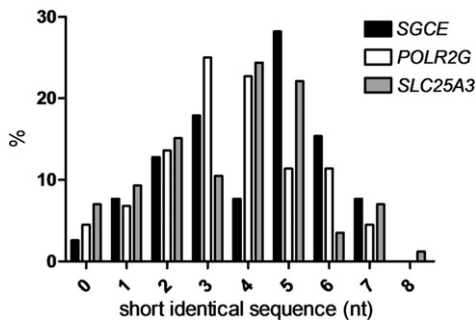
**Fig. 2.** Variants classified as variation affecting multiple exons and intra-exonic deletions and the role of short identical sequences. Panel A shows examples of both types of variations affecting multiple exons. (A1) represents a deletion (underlined) that affects the 3' end of exon A and the 5' end of exon B. A short identical sequence of 5 nucleotides (box) is present at the junction. This variant has been identified three times in brain sample of Case 1 in *POLR2G*. (A2) represents a deletion that affects the 3' end of exon A and a 5' extended exon B. Intronic sequences are shown in lower case, the deletion is underlined. We identified a short identical sequence of 6 nucleotides (box). This variant has been identified seven times in heart sample of Case 1 in the *SGCE* gene. Panel B illustrates an example of an intra-exonic deletion identified in a *POLR2G* transcript in heart tissue of Case 1. A short identical sequence of 4 nucleotides is present (box). The deleted sequence is underlined.

especially a higher number of intra-exonic deletions compared to other species (total number of low frequency variants: zebrafish: 36 vs. human: 14 (77 years old case), mouse: 18, rat: 12; see Table 1).

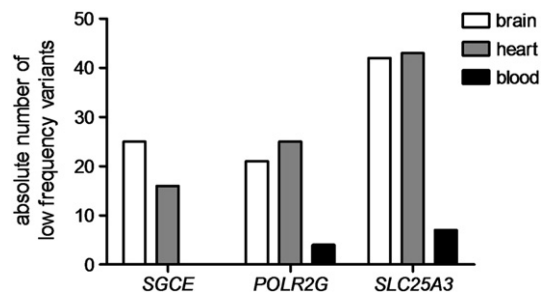
**4. Discussion**

We investigated alternative splicing events of single genes. Next to known isoforms several low frequency variants were identified. The majority of events are most likely due to alternative splicing requiring splice sites and splice factors: inclusion of new alternative exons, skipping of constitutive exons, intron retention, alternative 3' splice site, and alternative 5' splice site. One exon duplication event was

identified, which is likely a product of intragenic trans-splicing [6,16]. The low frequency of identified variants and the fact that most of them lead to a frameshift and a premature stop which targets them for nonsense mediated decay [17] suggests that they are not functional. Two novel events unlikely refer to the splicing machinery: variations affecting multiple exons and small intra-exonic deletions. Both events lack consensus splice sites and the majority of transcripts revealed short identical sequences of 1 to 8 nucleotides at the junction. These events were not species-specific; they were identified in mouse, rat and zebrafish. In zebrafish we identified significantly more variations affecting multiple exons and especially intra-exonic deletions than in other species. A possible explanation could be that 100 entire zebrafish embryos were investigated in contrast to other species where single tissues and just one individual were analyzed. The larger



**Fig. 3.** Frequency of short identical sequences. Displayed is the frequency of short identical sequences (nt, nucleotides) that were observed in variants classified as variations affecting multiple exons and intra-exonic deletions. Y-axis shows the percentage of low frequency variants with 0–8 nucleotides short identical sequences (X-axis) at the junction.



**Fig. 4.** Tissue-specific trends of variations affecting multiple exons and intra-exonic deletions. Absolute number of different low frequency variants that were classified as variations affecting multiple exons and intra-exonic deletions is shown per gene and tissue.



**Table 1**

Absolute number of different variants identified per event for 7 years old vs. 77 years old human cases and different species in *SGCE* (exon 9–12).

| <i>SGCE</i>                         | HS            |                | MM | RN | DR |
|-------------------------------------|---------------|----------------|----|----|----|
|                                     | (7 years old) | (77 years old) |    |    |    |
| New alternative exon                | 1             | 1              | 4  | 4  | 0  |
| Skipping of a constitutive exon     | 1             | 1              | 1  | 0  | 1  |
| Intron retention                    | 0             | 0              | 0  | 0  | 0  |
| Alternative 5' ss                   | 0             | 0              | 0  | 0  | 0  |
| Alternative 3' ss                   | 5             | 7              | 4  | 4  | 2  |
| Exon duplication                    | 0             | 0              | 0  | 0  | 0  |
| Variations affecting multiple exons | 1             | 0              | 4  | 1  | 8  |
| Intra-exonic deletion               | 6             | 5              | 5  | 3  | 25 |

HS *Homo sapiens*, MM *Mus musculus*, RN *Rattus norvegicus*, DR *Danio rerio*, ss splice site.

number of both events could be a summation of interindividual differences in transcriptional slippage or the transcriptional regulation in zebrafish may be less strict resulting in more slippage events.

The occurrence of short identical sequences near junctions is a common observation in genomic rearrangements due to different mechanisms: microhomology-mediated end joining (MMEJ) and microhomology-mediated break-induced replication (MMBIR) or fork stalling and template switch (FoSTeS) are using sequence homologies to align and anneal broken single stranded DNA ends [18,19]. The majority of complex genomic rearrangements like non-recurrent copy number variations (CNV) reveal microhomologies of 1–10 bp at the breakpoints [20]. At RNA level short identical sequences have been described in chimeric transcripts [8]. They were thought to be generated by trans-splicing of two or more transcripts [16], but recently a different mechanism based on the presence of short identical sequences was suggested: a large-scale search for chimeric RNAs in different species revealed that about 50% have short identical sequences with  $\geq 4$  bp at the junction sites of the source sequence which is similar to our observations. Li and colleagues showed that disruption of the short identical sequence lead to loss of the respective chimeric RNA, suggesting that they are essential for their formation [8]. They proposed an intermolecular transcriptional slippage model in which under certain circumstances a transcription complex with its pre-mRNA molecule dissociates from the template strand during transcription. The presence of the short identical sequences may facilitate strand annealing to continue transcription on the same or a different DNA template. This requires in case of intermolecular slippage, that both loci involved in the generation of the chimeric RNA should be active and occupy the same transcription factory, which are discrete sites in the nucleus where multiple active RNA polymerases are concentrated [21]. For intramo-

**Table 2**

Frequency of short identical sequences identified in *SGCE* (exon 9–12).

| <i>SGCE</i> | HS            |                | MM       | RN       | DR        |
|-------------|---------------|----------------|----------|----------|-----------|
|             | (7 years old) | (77 years old) |          |          |           |
| SIS (nt)    |               |                |          |          |           |
| 0           | 0             | 0              | 1 (11.1) | 0        | 1 (3.0)   |
| 1           | 1 (14.3)      | 1 (20.0)       | 2 (22.2) | 0        | 2 (6.1)   |
| 2           | 1 (14.3)      | 1 (20.0)       | 0        | 0        | 4 (12.1)  |
| 3           | 1 (14.3)      | 0              | 0        | 0        | 6 (18.2)  |
| 4           | 1 (14.3)      | 0              | 0        | 0        | 13 (39.4) |
| 5           | 2 (28.6)      | 3 (60.0)       | 5 (55.6) | 2 (50.0) | 4 (12.1)  |
| 6           | 1 (14.3)      | 0              | 1 (11.1) | 2 (50.0) | 0         |
| 7           | 0             | 0              | 0        | 0        | 1 (3.0)   |
| 8           | 0             | 0              | 0        | 0        | 0         |
| 9           | 0             | 0              | 0        | 0        | 2 (6.1)   |

Frequency of short identical sequences (SIS) identified in variants classified as variations affecting multiple exons and intra-exonic deletions, for 7 years old vs. 77 years old human cases (HS), mouse (MM), rat (RN), and zebrafish (DR). The absolute number of different variants identified (percentage in brackets) per nucleotides (nt) of short identical sequences is given.

lecular slippage, in which the same template is used, events such as those we described in this study are generated. On average 95% of variations affecting multiple exons and intra-exonic deletions display short identical sequences (1–8 nucleotides; 56% with  $\geq 4$  nucleotides identical sequence). Our approach did not allow for the detection of chimeric RNAs since the sequencing template was generated by a gene-specific PCR, thus defining start and end position of the amplicon. The observed deletions together with the short identical sequences support and argue for involvement of transcriptional slippage and make trans-splicing unlikely. Another fact arguing against involvement of the spliceosome is that the minimum intron size is limited by the length of splicing signals and the shortest introns that have been identified in eukaryotes were between 20 and 30 bp [22]. In our study intra-exonic deletions of even less nucleotides were found, which favors the transcriptional slippage model upon splicing as a generation mechanism.

Recently, Houseley and colleagues suggested that technical artifact can lead to presumed trans-splicing or transcriptional slippage events. Events that lack canonical splice sites can arise by template switching of reverse transcriptase during cDNA preparation [12]. Template switching, an intrinsic property of reverse transcriptase enzyme, can produce cDNAs which can be misinterpreted as splicing events [9–11]. However, our control experiments show that the observations made in this study are not artifacts of reverse transcriptase, PCR reaction or sequencing method. In addition, we observed tissue-specific differences in the occurrence of both types of low frequency variants (Fig. 4); they were predominantly expressed in brain and muscle and their occurrence in blood was considerably lower and even absent in *SGCE*.

The question remains whether transcriptional slippage occurs at random or whether it is influenced by certain factors. We observed some recurrent events in different tissues of the same individual. The same deletion was identified whereas the flanking sequence contains a combination of different exons suggesting that they did not evolve by a completely random process (e.g. an intra-exonic deletion in *SGCE* exon 9 was observed in combination with or without the known alternatively spliced exon 8). A possible explanation is that at some regions the polymerase dissociates more often from the template than at others. Upon reannealing at short identical sequences these sites show a higher frequency of slippage. Dissociation from the template could be influenced by sequence variation, epigenetic effects or secondary structures. In stress situations, transcriptional regulation can be less strict resulting in enhanced slippage and accumulation of observed low frequency, non-functional variants or to variants with (toxic) gain-of-functions. We excluded age as a condition that may have an effect on transcriptional slippage. However, the hypothesis of a role of slippage in disease remains speculation.

To conclude, we present an in-depth analysis of ultra deep amplicon sequencing data and identified various low frequency variants. Next to known splicing events, we report two novel types of events, variations affecting multiple exons and intra-exonic deletions that lack canonical splice sites and involve short identical sequences at the junction. The presence of short identical sequences and the size of deletions favor the transcriptional slippage model upon (trans-) splicing as a generation mechanism. Our results suggest that intramolecular slippage occurs *in vivo* and support the hypothesis that intermolecular transcriptional slippage is a mechanism to create chimeric RNAs.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2011.05.005.

## References

- [1] G. Zhang, G. Guo, X. Hu, Y. Zhang, Q. Li, R. Li, R. Zhuang, Z. Lu, Z. He, X. Fang, L. Chen, W. Tian, Y. Tao, K. Kristiansen, X. Zhang, S. Li, H. Yang, J. Wang, J. Wang, Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome, *Genome Res.* 20 (2010) 646–654.

- [2] Q. Pan, O. Shai, L.J. Lee, B.J. Frey, B.J. Blencowe, Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, *Nat. Genet.* 40 (2008) 1413–1415.
- [3] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nat. Methods* 5 (2008) 621–628.
- [4] P. Carninci, Is sequencing enlightenment ending the dark age of the transcriptome? *Nat. Methods* 6 (2009) 711–713.
- [5] H. Li, J. Wang, G. Mor, J. Sklar, A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells, *Science* 321 (2008) 1357–1361.
- [6] T. Takahara, B. Tasic, T. Maniatis, H. Akanuma, S. Yanagisawa, Delay in synthesis of the 3' splice site promotes trans-splicing of the preceding 5' splice site, *Mol. Cell* 18 (2005) 245–251.
- [7] D. Communi, N. Suarez-Huerta, D. Dussosoy, P. Savi, J.M. Boeynaems, Cotranscription and intergenic splicing of human P2Y11 and SSF1 genes, *J. Biol. Chem.* 276 (2001) 16561–16566.
- [8] X. Li, L. Zhao, H. Jiang, W. Wang, Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes, *J. Mol. Evol.* 68 (2009) 56–65.
- [9] E. Gilboa, S.W. Mitra, S. Goff, D. Baltimore, A detailed model of reverse transcription and tests of crucial aspects, *Cell* 18 (1979) 93–100.
- [10] H.M. Temin, Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation, *Proc. Natl. Acad. Sci. U. S. A.* 90 (1993) 6900–6903.
- [11] J. Cocquet, A. Chong, G. Zhang, R.A. Veitia, Reverse transcriptase template switching and false alternative transcripts, *Genomics* 88 (2006) 127–131.
- [12] J. Houseley, D. Tollervey, Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro, *PLoS One* 5 (2010) e12271.
- [13] C.J. McManus, M.O. Duff, J. Eipper-Mains, B.R. Graveley, Global analysis of trans-splicing in *Drosophila*, *Proc. Natl. Acad. Sci. U. S. A.* 107 (2010) 12975–12979.
- [14] F. Ozsolak, P.M. Milos, RNA sequencing: advances, challenges and opportunities, *Nat. Rev. Genet.* 12 (2011) 87–98.
- [15] K. Ritz, B.D. van Schaik, M.E. Jakobs, A.H. van Kampen, E. Aronica, M.A. Tijssen, F. Baas, SGCE isoform characterization and expression in human brain: implications for myoclonus-dystonia pathogenesis? *Eur. J. Hum. Genet.* 19 (2011) 438–444.
- [16] T. Horiuchi, T. Aigaki, Alternative trans-splicing: a novel mode of pre-mRNA processing, *Biol. Cell* 98 (2006) 135–140.
- [17] L.E. Maquat, Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics, *Nat. Rev. Mol. Cell Biol.* 5 (2004) 89–99.
- [18] M. McVey, S.E. Lee, MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings, *Trends Genet.* 24 (2008) 529–538.
- [19] P.J. Hastings, J.R. Lupski, S.M. Rosenberg, G. Ira, Mechanisms of change in gene copy number, *Nat. Rev. Genet.* 10 (2009) 551–564.
- [20] F. Zhang, C.M. Carvalho, J.R. Lupski, Complex human chromosomal and genomic rearrangements, *Trends Genet.* 25 (2009) 298–307.
- [21] H. Sutherland, W.A. Bickmore, Transcription factories: gene expression in unions? *Nat. Rev. Genet.* 10 (2009) 457–466.
- [22] M. Deutsch, M. Long, Intron-exon structures of eukaryotic model organisms, *Nucleic Acids Res.* 27 (1999) 3219–3228.