OPEN ACCESS Freely available online

PLOS ONE

Tandem RNA Chimeras Contribute to Transcriptome Diversity in Human Population and Are Associated with Intronic Genetic Variants

Liliana Greger[1]*, Jing Su[1], Johan Rung[1¤], Pedro G. Ferreira[2,3,4], Geuvadis consortium[¶], Tuuli Lappalainen[5], Emmanouil T. Dermitzakis[2,3,4], Alvis Brazma[1]

1 European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom, **2** Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland, **3** Institute for Genetics and Genomics in Geneva (iG3), University of Geneva, Geneva, Switzerland, **4** Swiss Institute of Bioinformatics, Geneva, Switzerland, **5** New York Genome Center, New York, New York, United States of America

Abstract

Chimeric RNAs originating from two or more different genes are known to exist not only in cancer, but also in normal tissues, where they can play a role in human evolution. However, the exact mechanism of their formation is unknown. Here, we use RNA sequencing data from 462 healthy individuals representing 5 human populations to systematically identify and in depth characterize 81 RNA tandem chimeric transcripts, 13 of which are novel. We observe that 6 out of these 81 chimeras have been regarded as cancer-specific. Moreover, we show that a prevalence of long introns at the fusion breakpoint is associated with the chimeric transcripts formation. We also find that tandem RNA chimeras have lower abundances as compared to their partner genes. Finally, by combining our results with genomic data from the same individuals we uncover intronic genetic variants associated with the chimeric RNA formation. Taken together our findings provide an important insight into the chimeric transcripts formation and open new avenues of research into the role of intronic genetic variants in post-transcriptional processing events.

Citation: Greger L, Su J, Rung J, Ferreira PG, Geuvadis consortium, et al. (2014) Tandem RNA Chimeras Contribute to Transcriptome Diversity in Human Population and Are Associated with Intronic Genetic Variants. PLoS ONE 9(8): e104567. doi:10.1371/journal.pone.0104567

Editor: Thomas Preiss, The John Curtin School of Medical Research, Australia

Received June 3, 2014; **Accepted** July 14, 2014; **Published** August 18, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by grants: the European Commission 7th Framework Programs (FP7) (GEUVADIS 261123) and EurocanPlatform (260791). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: lgreger@ebi.ac.uk

¤ Current address: Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden

¶ Membership of the Geuvadis consortium is provided in the Acknowledgments.

Introduction

Transcripts formed from exons of two or more different genes are known to play a major role in cancer development [1], where they often arise mostly from gene fusions caused by chromosomal rearrangements. However, chimeric transcripts are also present in healthy human tissues, where they may increase protein complexity and contribute to human evolution [2]–[7]. It is believed that chimeric transcripts in normal population arise from trans-splicing [8] or represent by-passing the transcription termination events [5], [6], [9].

The most prevalent type of chimeric transcripts found in the human genome are tandem RNA chimeras originating from genes residing on the same strand and chromosome [2]–[7]. These RNA chimeras, also known as read-throughs and conjoined genes have the potential for increasing the protein diversity as apparent from reports showing evidence of protein translation in organisms such as avian species [10]. These transcripts are potentially playing an important role in mammalian development, for example in regulating cell development in human [11] and in other organisms such as chick, quail and mouse [10], [12]. Furthermore, many tandem RNA chimeras were found to be evolutionary conserved [13]. Despite the numerous reports showing the existence of these chimeric transcripts in human, the exact mechanism of their formation is still unknown. Kim at all proposed that these events might be caused by removal of the poly (A) signal sequence from the upstream gene region by a deletion or a truncation mutation [14]. Trans-regulator of a poly(A) choice was reported to regulate the chimeric RNA formation in Arabidopsis thaliana [15]. There are also reports of intergenic splicing between tandem genes excluding the possibility of a read-through long transcript [11]. More detailed studies focused on chimeric transcript formation using large datasets of healthy individuals are needed to improve our knowledge of this phenomenon. Most of the reports that studied these events were previously based on ESTs data, which typically included also samples from non-healthy individuals, while at the same time were incomplete representations of the human genome.

Here, for a first time we identified and characterized chimeric transcripts found in a large cohort of population sample of individuals lacking severe phenotypes by using RNA sequencing data isolated from lymphoblastoid cell lines of 462 individuals

representing 5 populations from the 1000 Genomes project. By taking advantage of the genomic data available from the 1000 Genomes project, we could link for a first time the mechanism of tandem RNA chimera formation to specific genetic variants.

Materials and Methods

Chimeric RNAs discovery

In this study we used RNA sequencing data from lymphoblastoid cell lines of 462 individuals from five populations 91 CEPH (CEU, Utah residents with ancestry from northern and western Europe), 95 Finns (FIN, Finish in Finland), 94 British (GBR, British in England and Scotland), 93 Toscani (TIS, Toscani in Italia) and 89 Yoruba (YRI, Yoruba in Ibadan, Nigeria), who participated in the 1000 Genomes project (study design described in [16]. Chimeric transcripts were identified with a pipeline based on FusionMap v 2013-02-01 [17]. In essence, the raw sequence reads were aligned with bwa aligner [18]. Subsequently, the unmapped reads were extracted from the bam files with SAMtools using the flags for unmapped reads, [19] and then converted to fastq files with the Hydra package [20], which were subsequently subjected to fusion genes analysis with FusionMap [17]. The list of fusion genes was further filtered by the number of split reads supporting the fusion (>2). Finally, to reduce the number of false positive results chimeras, we used RepeatMasker (http://www.repeatmasker.org) to filter out breakpoints overlapping with repetitive elements.

To ensure that the observed population specific fusions were not due to the lack of expression of the partner genes in the other populations, only genes with positive RPKM values observed in all populations were considered. The expression values (RPKM) were estimated with Flux Capacitor v 1.2.3 [21].

The calculation of the open reading frame of the fusions was based on the CDS information in the gene model, provided by the reference GTF file and it is implemented in the FusionMap software.

Finally, the novelty of the fusions was assessed via queries against known annotation databases for read-through events: ConjoinG [7], AceView [22] and literature search.

The distances between the fusion breakpoints was estimated as $|bp2 - bp1|$, where bp1 is the breakpoint of upstream parent gene and bp2 is the breakpoint of the downstream parent gene involved in the RNA chimera. The frequency counts for each chimeric transcript were estimated as the number of individuals from each population harboring the chimeric transcript.

The sizes of the exons and introns residing at the fusion breakpoint were estimated by taking into account all possible transcript combinations between the chimeric genes.

Distances, frequency plots and intron lengths distribution plot were generated with ggplot2 package within R environment [23].

The exon exclusion pattern was assessed by estimating the frequencies of the excluded exons from the two parent genes during the tandem RNA chimera formation. We took into account all possible transcript combinations.

As a reference annotation for all analyses above we used Gencode v12.

Quantification of RNA chimeras

We created custom modified Gencode v12 reference gtf file to accommodate all possible transcript isoforms of tandem RNA chimeras by combining exons of the upstream partner gene up to the fusion breakpoint with exons of the downstream partner gene after the fusion breakpoint and assigning a new transcript ID. Then, for each fusion we assigned a new gene ID. In four cases the breakpoint was predicted to be away from the exon-intron boundaries and therefore, we modified the exon boundary to match the predicted breakpoint. Subsequently, all bam files mapped with GEM [24] were subjected to Flux Capacitor, which assigns the reads to the annotated in the custom gtf exon-intron structures [25] to quantify the transcript abundances for each gene. The gene level RPKMs were obtained by summing up the RPKMs for all transcripts sharing the same gene identity. The distribution of the chimeras RPKMs, estimated in individuals harboring the chimeric transcripts was further compared to the RPKMs of the partner genes and all other genes in the same individuals.

All custom scripts and the modified gtf file are available on request.

Association between tandem RNA chimera formation and genetic variants

The association between tandem RNA chimera formation and genetic variants has been performed on custom annotated 1000 Genomes Phase I, release v3 vcf files, based on Gencode v12 [16] with tools within SnpSift v3.2 toolset [26]. We selected only fusions with frequency higher than 5% in all combined populations and SNPs with MAF greater than 5% with SnpSift filter [26]. Initially, for each fusion we extracted all variants located on both fusion partner genes (+200 nucleotides outside the gene boundaries) with SnpSift Intervals [26]. Then, we tested for association of each fusion with genetic variants in the cases (samples, which have the fusion) compared to the controls (samples, which do not harbor the fusion) by applying Cochran–Armitage test with SnpSift Case-Control [26]. We further selected all variants with adjusted by bonferroni p value<0.05. Subsequently, for each variant, we estimated the OR (odd ratio i.e. proportions of individuals, with the fusion having the alternative allele relative to proportion of individuals without the fusion having the alternative allele).

Further, the identified variants were functionally annotated by utilizing data available from Ensembl regulatory build. Subsequently, the associated with RNA chimeras variants were overlapping with genomic locations for transcripts associated with RNA binding proteins by using data in bed format produced by RIP-chip GeneST, Ribonomic profiling on Affymetrix GeneChip® Human Gene 1.0 ST Arrays [27] as part of the ENCODE project [28]. Overlapping of genomic regions was assessed with bedtools [29]. The identified variants were further inspected with regional association plots performed with LocusZoom v 1.1 [32].

To explore variants in poly(A) sites and signals, we used annotated by Havana, Poly(A) sites and signals Gencode v12 [30] and predicted putative cleavage sites [31]. The prediction of putative cleavage sites was performed by selecting reads containing a poly(A) tail or a poly T head, filtering out low complexity reads, trimming and mapping the remaining reads uniquely to the genome. All 3′ UTR variants overlapping with these sites were inspected further. For each 5′ partner gene we filtered out the observed variants in poly(A) sites or signals by MAF (equal or higher than the gene frequency in the population).

Results

Detection and characterization of RNA chimeras

We used a pipeline based on FusionMap software tool [17] to detect fusion genes in RNA sequence data from 462 normal individuals from five different populations. This tool has been proved to be a good compromise in terms of specificity and sensitivity in normal tissues as compared to five other tools for fusion genes discovery [33]. The discovered fusions were classified

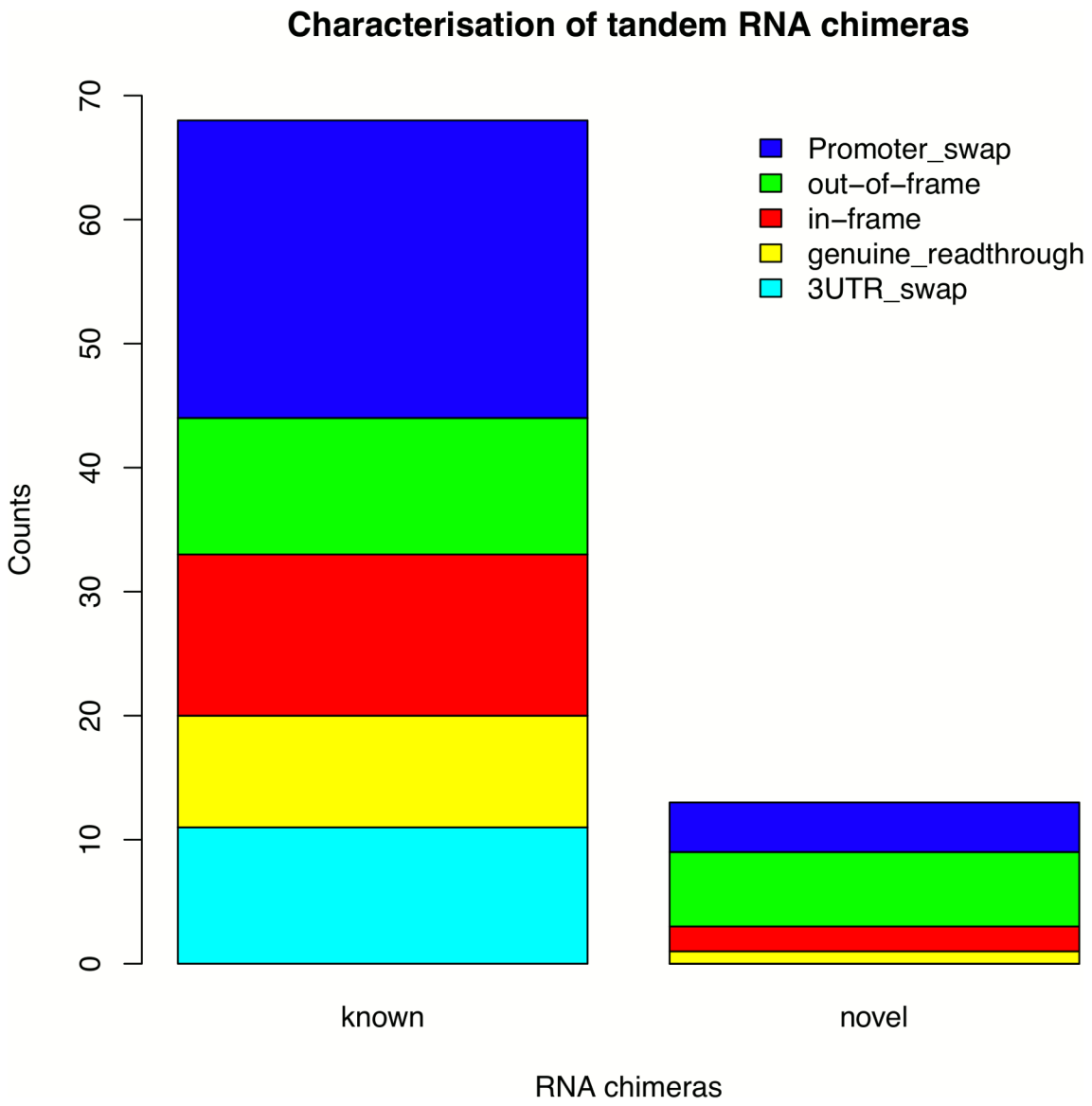

Figure 1. Characterization of tandem RNA chimeras. Barplot representing the frequency of the predicted structural classes for all tandem RNA chimeras. The most prevalent observed class is a promoter swap event.
doi:10.1371/journal.pone.0104567.g001

as intra-chromosomal fusions on the same strand, intra-chromosomal fusions on different strands and inter-chromosomal fusions. For chimeric transcripts arising from intra-chromosomal genes residing on the same strand we used the term "tandem RNA chimeras".

We observed 89 intrachromosomal genes residing on the same strand, 8 of which implied inversion since the direction of the fusion transcript was inverted relative to the annotation (Table S1 in File S1). In order to study in more detail fusions potentially formed by intergenic splicing than genomic rearrangements, we excluded these 8 fusions from further analysis. (Table S2 in File S1). Except in 10 cases, the breakpoint coincided with the exon boundaries. In 6 partner genes, the fusion breakpoint was located within 1–3 nucleotides from the exon-intron boundary and in 4 cases the breakpoint was predicted to be located inside the exon. Sixty-seven out of the 81 fusions were known, while 13 were novel.

We further classified the identified tandem RNA chimeras as "promoter-swap" events, "3′ UTR-swap" events and "genuine read-through" events. We considered a fusion as a promoter swap event if the coding sequence of the 3′ partner was placed under the promoter of the 5′ partner gene. Similarly, in 3′ UTR-swaps the coding sequence of the 5′ partner was placed under the control of the UTR of the 3′ partner. Fusions, where the whole coding sequences of both genes were conserved, were classified as "genuine read-through" events. Taking this approach, we identified 10 genuine read-through events, 28 promoter swap events, 11 3′ UTR swap events, 15 in-frame chimeras and 17 out-of-frame chimeras (Table S2 in File S1, Figure 1).

In addition to the tandem RNA chimeras we report 6 intrachromosomal fusions located on different strands, only one of which (*NAIP->OCLN*) was known [34] (Table S3 in File S1), and 15 interchromosomal fusions, 5 of which were previously described (Table S4 in File S1). While three of the chimeras

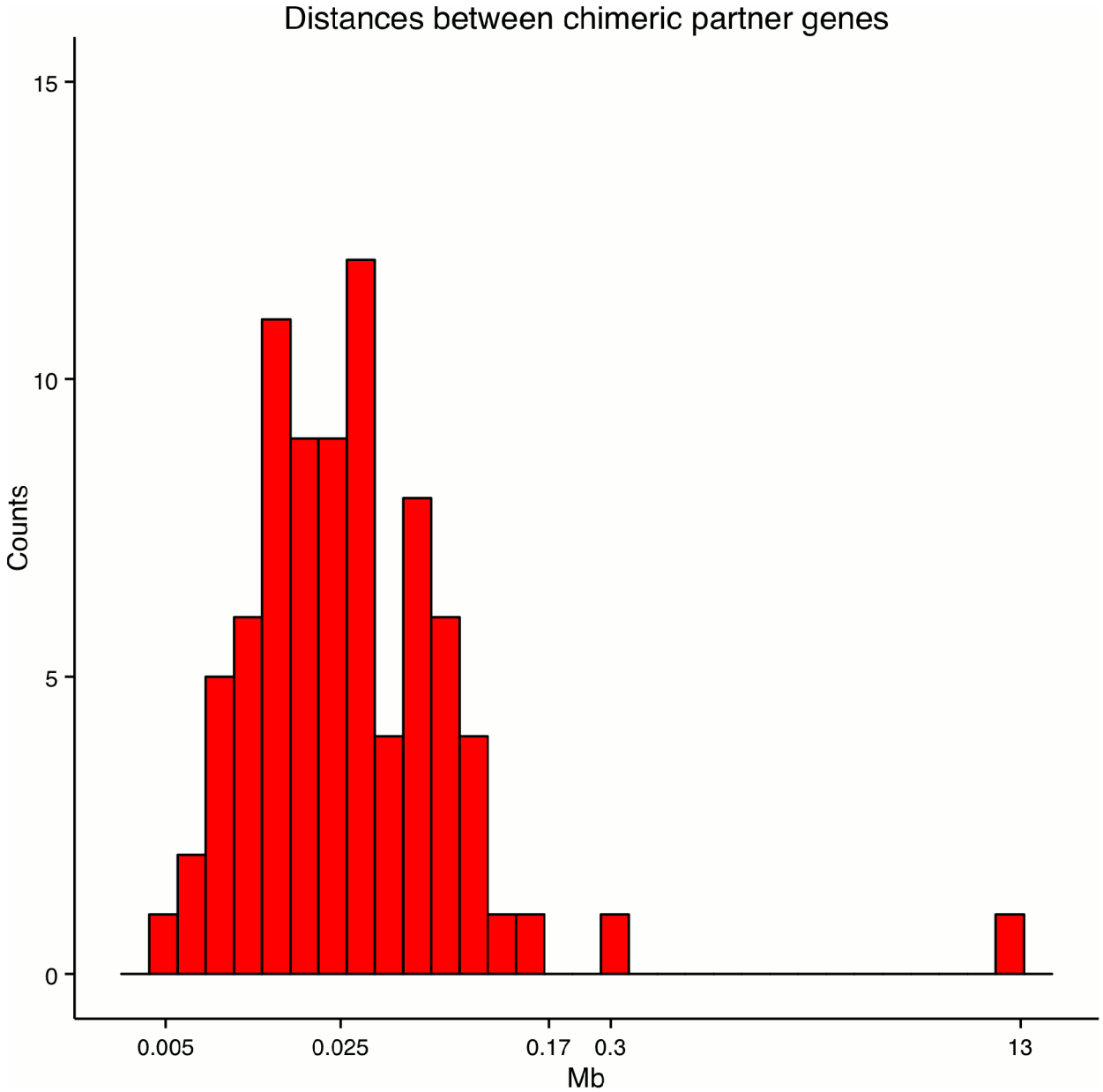

Figure 2. Distances between chimeric partner genes. Shown are distances between the tandem RNA chimeric partner genes in Mb. The majority of the distances are in the range between 5 kb and 170 kb.
doi:10.1371/journal.pone.0104567.g002

(*ACTB->POTEM*, *ACTB->POTEE* and *SP100->HMGB1*) have been found in normal population [35], [36], three of the chimeras (*C2orf27A->NBEA*, *HILPDA->EFCAB3*, *FARSB->TRIM61*) were previously identified only in cancer samples [37]. This indicates that some chimeric RNAs present in normal samples can be considered as cancer specific due to the lack of appropriate controls.

We further calculated the distances between the breakpoints of the tandem RNA chimeras (Figure 2). The range varied between 5kb and 170 kb with a median distance of 25 kb except for some outliers: *COX5A->EDC3* with a distance of 0.3 Mb and *NRXN1->EIF2AK2* with a distance of 13 MB. As previously reported, tandem RNA chimeras tend to originate from genes residing closer on the genome [6]. In comparison to the tandem RNA chimeras for chimeric partner genes residing on different strands, we observed larger distances, which implies a different formation mechanism. The range of the distances between these genes was in the interval from 25 kb to 227 Mb with a median value of 2 MB.

Frequency of chimeric transcripts in populations

To assess the distribution of the RNA chimeras across populations, we calculated the frequency of the tandem RNA chimeras in the studied five populations. Frequency plots and venn diagram of the tandem RNA chimeras showed that most of the chimeric transcripts are found only in a few individuals (Figure 3 and Figure 4A). We further observed that while the high-frequency chimeric transcripts were present in all five studied populations, some rare chimeric RNAs were detected in single individuals as for example "*COX5A->EDC3*" and "*COPE->*

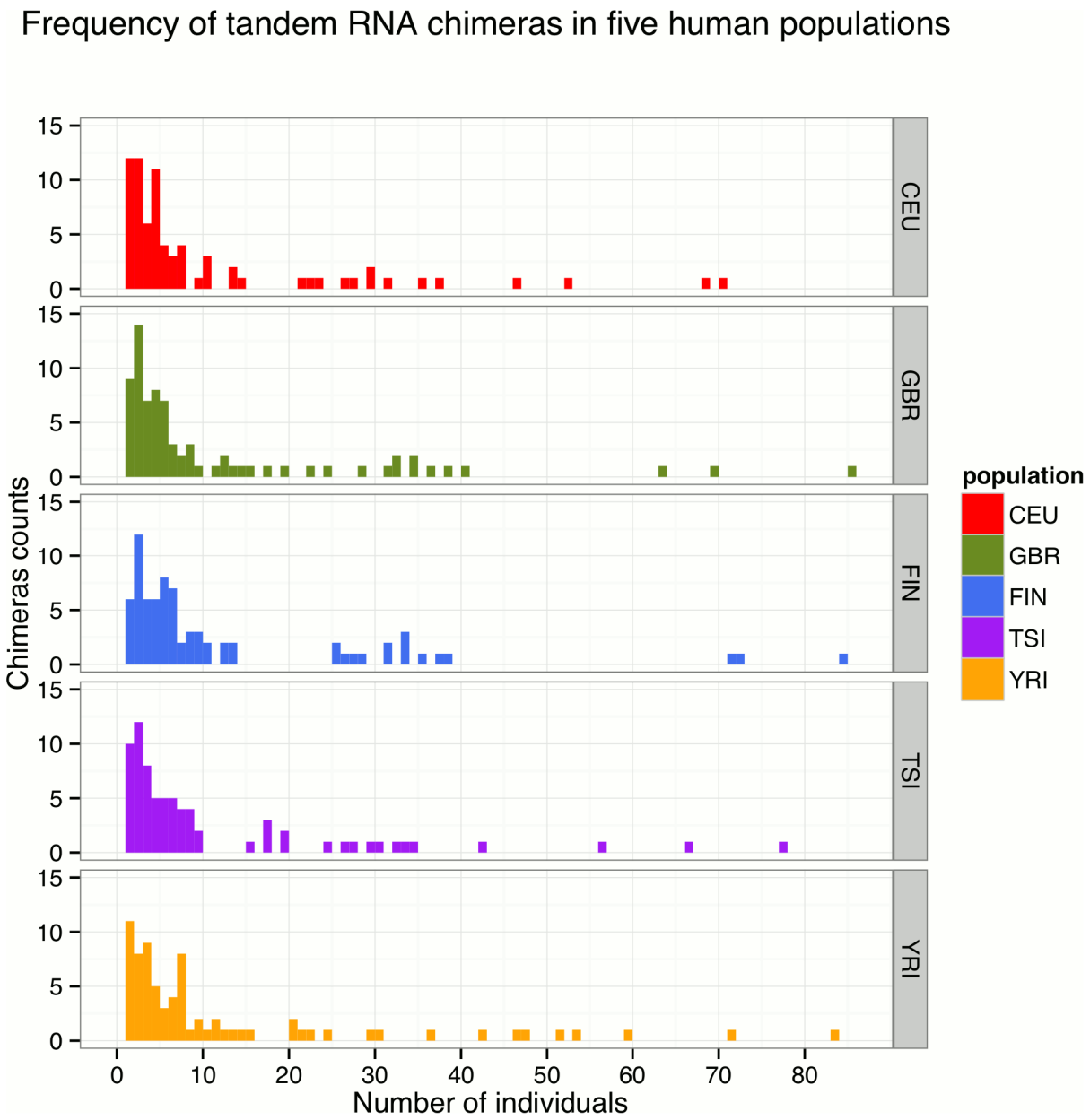

Figure 3. Tandem RNA chimeras frequency plots in five human populations. All populations show similar patterns with most of the chimeric transcripts found only in a few individuals.
doi:10.1371/journal.pone.0104567.g003

CERS1" were discovered only in one individual of the YRI population and the CEU population respectively (Table S2 in File S1). Moreover, we detected similar expression values for the partner genes in the 5 human populations suggesting that the varying frequency of the chimeric transcripts in the populations is not due to expression differences (Table S5 in File S1). However, providing the small number of samples for each population we could not determine if these rare RNA chimeras are rather low frequent or population specific.

Similar trend in the population frequencies was revealed for the other two identified chimeric transcripts sets (intrachromosomal RNA chimeras located on different strands and interchromosmal RNA chimeras) as shown in the venn diagrams (Figure 4B–C).

Exon exclusion pattern of tandem RNA chimeras

We further focused to study in more detail the tandem RNA chimeras as the most prevalent type amongst the identified RNA chimeras. In addition, the majority of the tandem RNA chimeras have been already confirmed in the literature. Moreover, tandem RNA chimeras are most likely to represent genuine read-through or intergenic splicing events rather than being a result from genomic structural variations. Furthermore, the data on structural variations available from the 1000 genome project [38] does not include inversions and translocations, therefore not allowing filtering of known structural variants in the intrachromosomal located on different strands and interchromosomal fusions data sets.

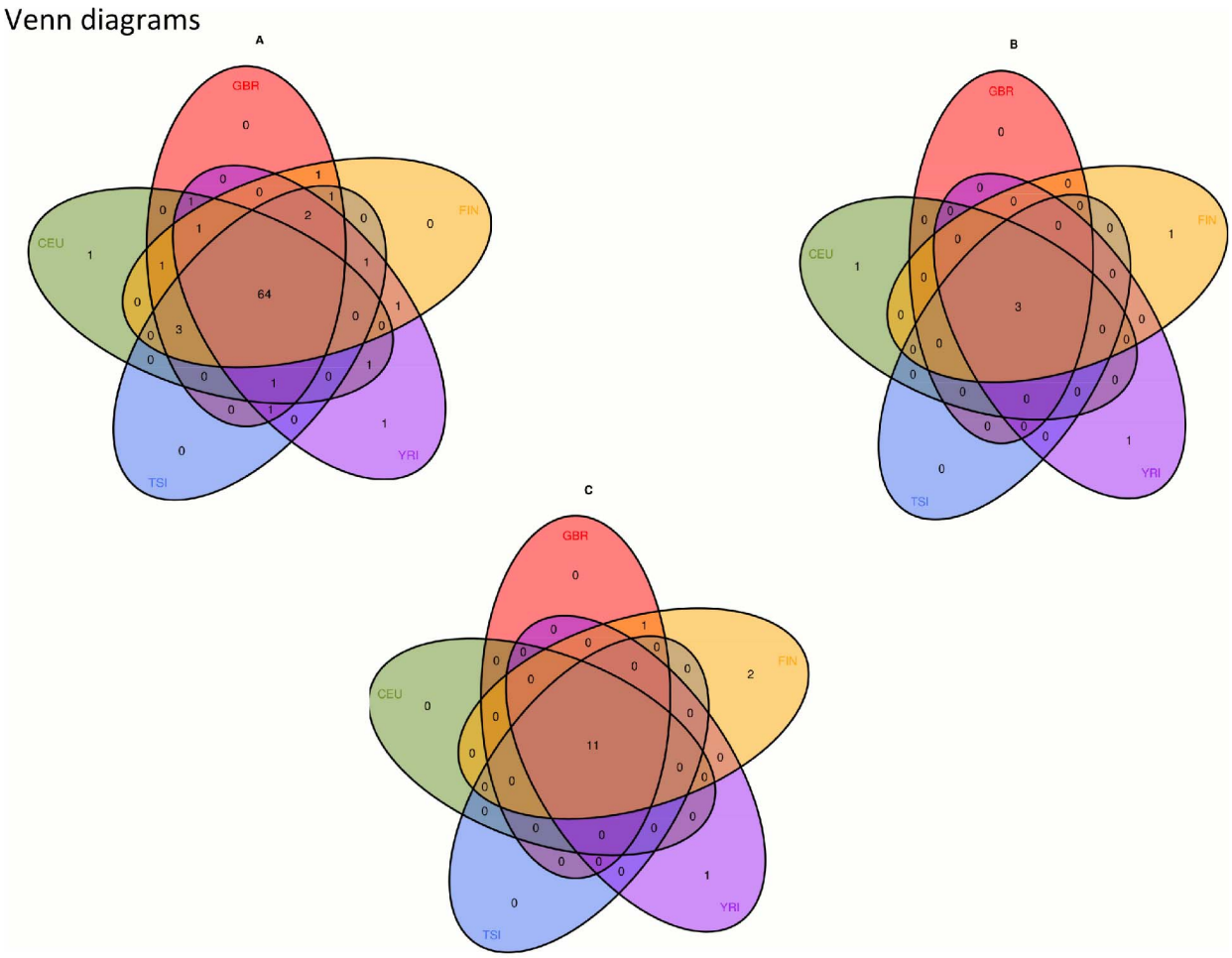

Figure 4. Venn diagrams. Venn diagrams displaying the relations between tandem RNA chimeras in five human populations (A); intrachromosomal chimeras located on different strands (B); and interchromosomal chimeras (C).
doi:10.1371/journal.pone.0104567.g004

We further sought to study the exon exclusion pattern in the tandem RNA chimera formation. We took into account all possible fusion transcripts (shown in Table S6 in File S1). Then, we estimated the proportions of the excluded exons. Figure 5 shows that the most abundant exon exclusion pattern (45%) involves the exclusion of the terminal exon of the 5′ partner gene and the first exon of the 3′ partner gene, a pattern already observed previously [6], [7]. The next most common exon exclusion pattern (16%) represented an omission of the last two exons of the 5′ partner and the first exon of 3′ partner, followed by the exclusion of the last terminal exon of the 5′ partner gene and the first two exons of the 3′ partner gene in 10% of the cases. Therefore, the last two exons of the 5′ partner gene and the first two exons of the 3′ partner genes are most likely to be spliced out during the tandem RNA chimera formation.

Tandem RNA chimeras are lowly expressed genes

Next, we explored in more detail the gene expression level of the tandem RNA chimeras. We created a custom gtf file with added all possible transcription combinations of the identified tandem RNA chimeras. Then, we compared the distribution of the summarized per gene expression values of the tandem RNA chimeras to the distribution of the expression values of the partner genes. Furthermore, given the fact that the chimeric transcripts were identified only in subset of individuals from each population, we compared only the expression values estimated in these individuals. As it shown in Figure 6, we observed that the tandem RNA chimeras are at lower abundances compared to the transcripts of the partner genes (Kolmogorov-Smirnov test, $D = 0.3262$, p-value$<$2.2e-16) with median values of 1.33 and 5.45 for the tandem RNA chimeras and for the partner genes respectively. This result is in agreement with a previous report, where the expression values of chimeric genes were estimated by a different approach using the reads located at the fusion junction [39]. However, it is worth mentioning that Frenkel-Morgenstern estimations were performed on datasets of chimeric transcripts, which included cancer specific fusions.

Long introns are overrepresented in tandem RNA chimeras

To check the possibility if the size of the exons and introns in the proximity of the fusion breakpoint could influence the RNA chimera formation, we compared the distributions of the exons and introns lengths located at the fusion boundaries to the lengths of the rest of the introns and exons comprising the fusion partner genes. While we did not observe a notable difference between the exon lengths (data not shown), we found that the chimeric genes tend to have larger introns at the fusion breakpoint as compared to

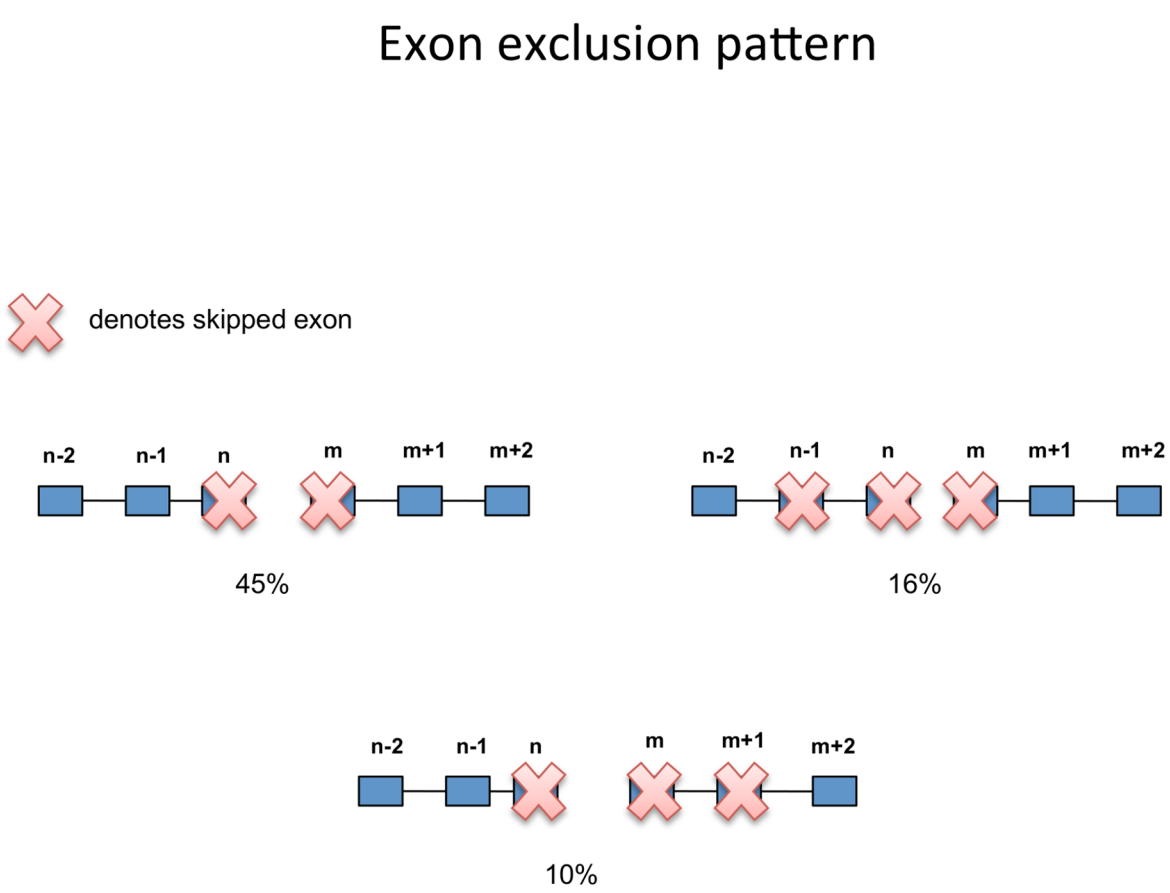

Figure 5. Exon exclusion pattern. Plot showing the proportions of the exon exclusion patterns in the tandem RNA chimeras. The numbers of the exons for the 5′ partner gene are counted from last to first, while the numbers of the exons for the 3′ partner gene are counted from first to last. The most abundant exon exclusion pattern involves the exclusion of the terminal exon of the 5′ partner gene and the first exon of the 3′ partner gene. doi:10.1371/journal.pone.0104567.g005

the rest of the introns comprising the partner genes, with estimated Kolmogorov-Smirnov tests, $D = 0.278$, p-value$<$2.2e-16 and $D = 0.389$, p-value$<$2.2e-16 for the 5′ intron and 3′ intron respectively (Figure 7). We observed median lengths of 2516 nt, 3734 nt for the intron following the fusion exon of the 5′ partner gene and for the intron proceeding the fusion exon of the 3′ partner gene respectively, as compared to median lengths of 1400 and 1164 estimated for the rest of the introns of the 5′ and 3′ partner genes.

Association of tandem RNA chimeras formation with genetic variants

To explore further the mechanism of tandem RNA chimera formation, we took advantage of the available genomic data from the same individuals provided by the 1000 genome project, with functional re-annotation with Gencode v12 as described in [16].

To check the hypothesis that deletions or mutation in the poly(A) signal of the 5′ partner gene drive the tandem RNA chimera formation, we inspected the poly(A) signals and sites as described in Methods. We did not observed in any 5′ partner genes deletions in the poly(A) signals. Two genes (LRRC33 and SDHD) out of 81 5′ partner genes were detected to have single nucleotide polymorphisms (A->G) in their poly(A) signal (rs77865386, rs17113461 respectively). Similarly, we could not identify any genes bearing deletions in their poly(A) sites. Only seven genes were identified to have single nucleotide polymorphisms in these sites (Table S7 in File S1).

Since the identified genetic variants in poly A signal or sites could not explain the mechanism of chimeras formation for the majority of the identified tandem RNA chimeras, we took an approach similar to genome wide association studies, where for each chimeric RNA, the cases consisted of individuals with chimera and controls were represented by individuals without chimera. For our analysis, we selected only fusions with frequency higher than 5% in all populations (37 chimeras) and SNPs with MAF $>5\%$.

By using this approach we identified 54 variants associated with the formation of 3 chimeras (Table S8 in File S1). All of the variants were located in intronic regions and more specifically, in introns surrounding the exon involved in the fusion. For all variants located at PABPC1 (Poly(A)denylate-binding protein 1) binding sites, which potentially could represent poly(A) sites or signals, we compared these to annotated from Gencode v12 or predicted poly(A) sites and signals found in intronic genomic locations. We did not observe any overlap. We further inspected the identified variants by regional association plots (Figure S1 in File S1, Figure S2 in File S1 and Figure S3 in File S1) and subsequently selected for each chimera the most significant variants having the lowest p-value (Table 1). Two out of the three variants rs9945924 and rs4148876, have been already reported to play an important role in controlling alternative splicing and disease progression [40].

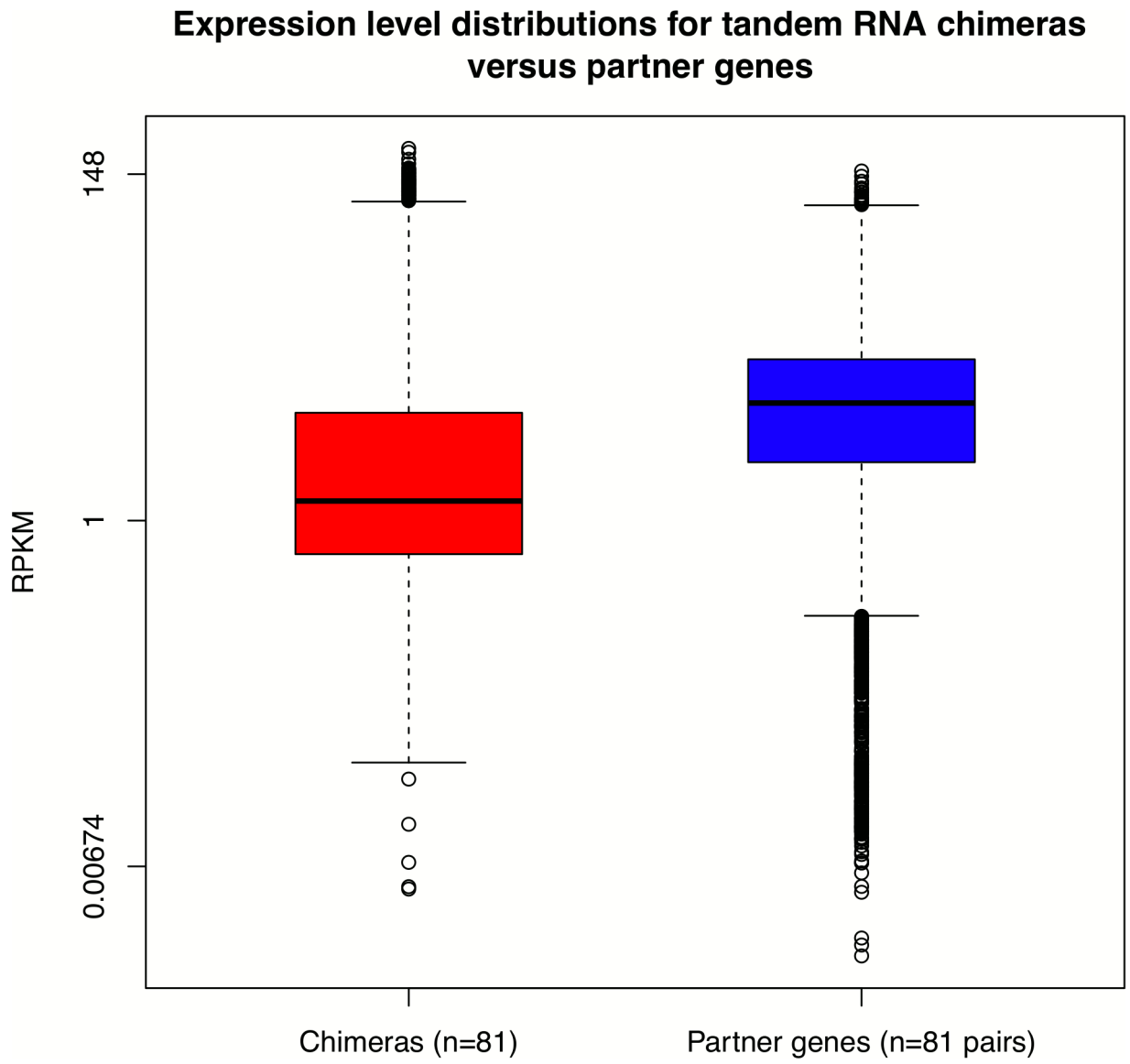

Figure 6. Expression level distributions of tandem RNA chimeras and their partner genes. RNA chimeras are lower expressed than their partner genes.
doi:10.1371/journal.pone.0104567.g006

Discussion

Chimeric transcripts, which contribute to the complexity of the human transcriptome, have already drowned the attention of the scientific community. However, the efforts to study this phenomenon are hampered by the lack of large datasets of normal populations.

Here, by combining expression and genetic variation data from the same individuals in a large cohort of samples, we identified and characterized chimeric transcripts in normal population. We focused on the most prevalent class of chimeric transcripts - tandem RNA chimeras. As a proof of concept most of the tandem RNA chimeras we identified were already present in public databases and in the literature, however we found 13 novel tandem RNA chimeras, demonstrating that there are more still to be discovered in normal populations. Our results showed prevalence of promoter swap events. An opposite trend was reported in cancer samples, where the most prevalent class of fusion transcripts occurs within the 3′ UTR sequence [41].

We observed varying frequency of RNA chimeras between individuals, which confirms that chimeric transcripts represent another layer of transcriptional variation existing in the human population.

Our results are in agreement with other groups [6], [7] showing evidence that the tandem RNA chimeras originate from genes that tend to reside closer on the genome. Perturbations in co-transcriptional coupling of terminal intron splicing and 3′ end processing [42]–[44] could explain the observations found in this and previous studies [6], [7], with the most frequent pattern featuring splicing out the terminal exon of the 5′ partner and the first exon of the 3′ partner. Furthermore, our results demonstrated prevalence of long introns in the vicinity of the 5′ and 3′ fusion exons. A long intron at the fusion breakpoint could allow enough time for the intergenic splicing to compete with the cis-splicing machinery for the 5′ unsaturated splice donor site. It has been

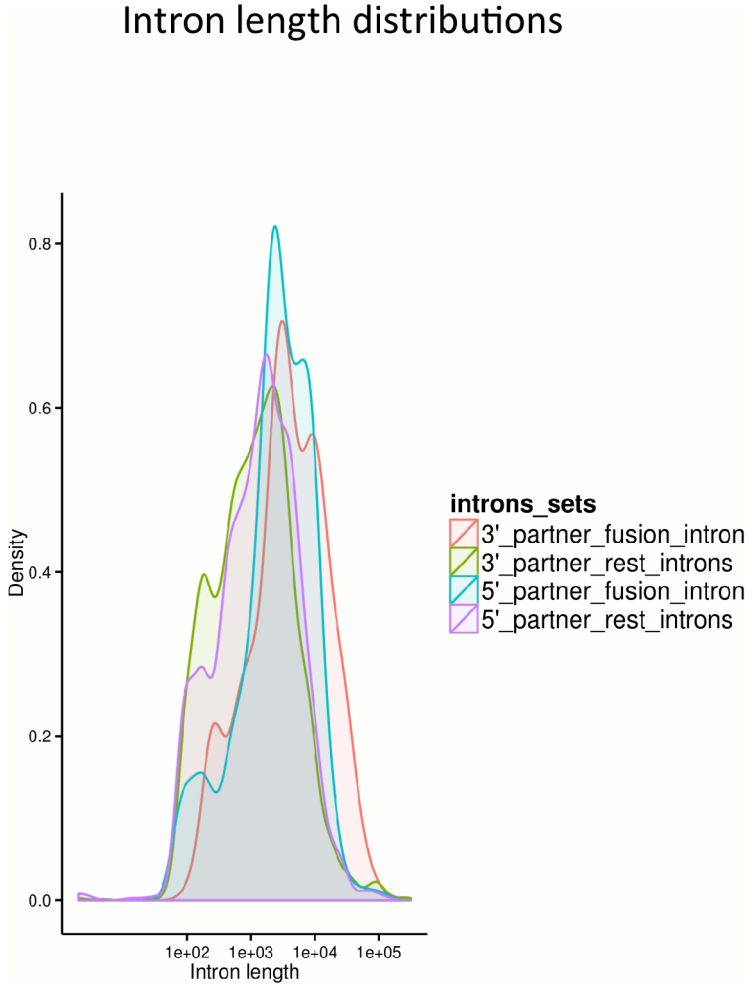

Figure 7. Intron length density plots. Density plots showing the distributions of the fusion 5′ and 3′ introns length versus the length of the rest of the introns comprising the fusion partner genes. Observed is a notable higher proportion of larger introns at the breakpoint.
doi:10.1371/journal.pone.0104567.g007

known that long introns can promote alternative splicing [45]. Moreover, a study on human cells expressing modified *Sp1* transgenes showed evidence that pausing of RNA polymerase II and presence of long intron promotes trans-splicing of genes residing on different chromosomes [46]. Herein, we confirmed this observation for tandem RNA chimera formation by using large scale high-throughput data.

Contrary to some reports [14], which suggests that deletions at poly(A) signals or poly(A) sites are responsible for the tandem RNA chimera formation, we could not identify such events. We observed genetic variants in poly A sites and signals in 9 of the 81 5′ partner genes, however these changes are unlikely to account for the mechanism behind the tandem RNA chimera origin. We do not exclude the possibility that other 3′ termination signal or trans-factors could play a role in the chimera formation.

By performing genetic association studies we were able to identify three variants associated with the presence of tandem RNA chimeras. All variants were located in intronic regions. Introns have increasing attention from the scientific community for their role in pre-mRNA splicing [47]. Reassuringly, two of the variants were already reported to play an important role in alternative splicing [48]. Thus, our results show that variants affecting splicing contribute to the RNA chimera origin.

By utilizing RNA binding sites and regulatory features genomic regions data we observed that two of the three variants are located in PABPC1 binding sites. PABPC1 has been mostly known to bind poly(A) tail and is involved in NMD pathway [49]. Several groups report involvement of PABPC1 in protein complexes playing a role in mRNA stability [50], [51]. Thus, PABPC1 could be also involved in RNA metabolism including splicing. More studies are needed to identify cis- and trans-factors involved in the formation of tandem RNA chimeras.

Our observations also demonstrate that many chimeras naturally occurring in normal populations have been mistaken for cancer-specific. For example *NRXN1->EIF2AK2*, *PRIM1->NACA* [52], *PRKAA1->TTC33*, *CTBS->GNG5* [53], *SLC35A3->HIAT1* [37] and *UBE2J2->FAM132A* [54]. In addition, the chimeras implying inversions were also considered as cancer-specific or being related to other diseases. For example *PPIP5K1->CATSPER2* and *YARS2->NAP1L1* [54] have been predicted in melanoma samples. The chimeras *TRIP12->SLC16A14* and *TFG->GPR128* have been found in cancer patients with clear cell renal carcinoma [55]. Nevertheless, another group reported identifying *TFG->GPR128* in healthy individuals [56]. Two of the chimeras implying inversion *MAPKAPK5->ACAD10* and *POLR1A->REEP1* have been detected in autistic individuals [57]. These findings suggest that lack of rigorous

Table 1. The most significant SNPs associated with RNA chimeras formation.

Gene	Chr	Position	SNP	p_value	adjusted p-value	Ref/Alt	Odd ratio	95% CI
GNG5	1	84965623	rs56212819	1.37E-04	2.87E-03	C/T	2.01	1.33–3.04
HMSD	18	61620766	rs9945924	5.55E-17	8.66E-15	G/A	4.79	3.22–7.14
TAP2	6	32796793	rs4148876	2.20E-16	4.77E-14	G/A	18.75	8.21–42.82

doi:10.1371/journal.pone.0104567.t001

normal controls could hamper the identification of genuine cancer-specific fusions. In addition, we noticed that 4% of the fusion genes identified in a clear cell renal cell carcinoma dataset represented RNA chimeras found in this study (data not shown). On the other hand we have to take into account that perturbations in the RNA chimera formation naturally occurring in normal population may lead to cancer progression. Such notion is supported by recent reports on differential expression in tandem RNA chimeras in tumor versus normal samples [58] or on identified tandem chimeric RNAs in cancer cells without an evidence for genomic rearrangements [59]. This complexity adds another challenge to the cancer biomarker discovery. In conclusion, more studies are needed to understand the mechanisms behind the chimeric transcripts regulation in healthy individuals and the perturbations of these naturally occurring events, which could lead to cancer.

Supporting Information

File S1 File includes Figures S1–S3 and Tables S1–S8. Figure S1: Regional association gene plots for CTBS->GNG5 chimera. Results are shown in the region flanking 100 kb both sides of the index SNPs. The marker SNPs are shown in purple and the color of the dots represent the degree of linkage disequilibrium (based on r2) in relation to the index SNP based on the March 2012 release of the 1000 Genomes data in European population. Figure S2: Regional association gene plots for HMSD->SERPINB8 chimera. Results are shown in the region flanking 100 kb both sides of the index SNPs. The marker SNPs are shown in purple and the color of the dots represent the degree of linkage disequilibrium (based on r2) in relation to the index SNP based on the March 2012 release of the 1000 Genomes data in European population. Figure S3: Regional association gene plots for TAP2->HLA-DOB chimera. Results are shown in the region flanking 100 kb both sides of the index SNPs. The marker SNPs are shown in purple and the color of the dots represent the degree of linkage disequilibrium (based on r2) in relation to the index SNP based on the March 2012 release of the 1000 Genomes data in European population. Table S1: Intrachromosomal RNA chimeras residing on the same strand, which orientation implies inversion. Table S2: Identified tandem RNA chimeras located on the same strand. Table S3: Identified intrachromosomal RNA chimeras located on different strands. Table S4: Identified interchromosomal RNA chimeras. Table S5: Gene expression values estimated for tandem RNA chimeras partner genes in each human population. Table S6: All annotated transcripts and exons residing at the fusion junction for the identified tandem RNA chimeras. Table S7: Genetic variants found in poly(A) sites of tandem RNA chimeras upstream genes. Table S8: All identified genetic variants associated with tandem RNA chimera formation. (PDF)

Acknowledgments

We thank Dr. Nuno Fonseca for the helpful discussion.

Geuvadis consortium

Tuuli Lappalainen[18], Michael Sammeth[19], Marc R. Friedländer[5], Peter A. C. Hoen[6], Jean Monlong[5], Manuel A. Rivas[7], Mar Gonzàlez-Porta[8], Natalja Kurbatova[8], Thasso Griebel[4], Pedro G. Ferreira[1,2,3], Matthias Barann[9], Thomas Wieland[10], Liliana Greger[8], Maarten van Iterson[6], Jonas Almlöf[11], Paolo Ribeca[4], Irina Pulyakhina[6], Daniela Esser[9], Thomas Giger[1], Andrew Tikhonov[8], Marc Sultan[12], Gabrielle Bertier[5], Daniel G. MacArthur[13], Monkol Lek[13], Esther Lizano[5], Henk P. I. Buermans[6,14], Ismael Padioleau[1,2,3], Thomas Schwarzmayr[10], Olof Karlberg[11], Halit Ongen[1,2,3], Sergi Beltran[4], Marta Gut[4], Katja Kahlem[4], Vyacheslav Amstislavskiy[12], Matti Pirinen[7], Stephen B. Montgomery[1], Peter Don-

nelly[7], Mark I. McCarthy[7,15], Paul Flicek[8], Tim M. Strom[10,16], The Geuvadis Consortium, Hans Lehrach[12], Stefan Schreiber[9], Ralf Sudbrak[12], Ángel Carracedo[17], Stylianos E. Antonarakis[1,2], Robert Häsler[9], Ann-Christine Syvänen[11], Gert-Jan van Ommen[6], Alvis Brazma[8], Thomas Meitinger[10,16], Philip Rosenstiel[9], Roderic Guigó[5], Ivo G. Gut[4], Xavier Estivill[5], Emmanouil T. Dermitzakis[1,2,3]

1 Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland,

2 Institute for Genetics and Genomics in Geneva (iG3), University of Geneva, Geneva, Switzerland,

3 Swiss Institute of Bioinformatics, Geneva, Switzerland,

4 Centro Nacional de Analisis Genomico, Barcelona, Spain,

5 Center for Genomic Regulation, Barcelona, Spain,

6 Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands,

7 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom,

8 European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom,

9 Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany,

10 Helmholtz Zentrum Munchen, Neuherberg, Germany,

11 Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Uppsala, Sweden,

12 Max Planck Institute for Molecular Genetics, Berlin, Germany,

13 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, United States of America,

14 Leiden Genome Technology Center, Leiden, the Netherlands,

15 Oxford Centre for DIabetes Endocrinology and Metabolism, University of Oxford, Oxford, United Kingdom,

16 Technische Universität München, Munich, Germany,

17 Fundación Publica Galega de Medicina Xenómica SERGAS, Genomic Medicine Group CIBERER, Universidade de Santiago de Compostela, Santiago de Compostela, Spain,

18 New York Genome Center, New York, New York, United States of America,

19 Bioinformatics Laboratory, National Laboratory of Scientific Computing, Petropolis, Rio de Janeiro, Brazil

Author Contributions

Conceived and designed the experiments: LG AB JR. Analyzed the data: LG JS. Contributed reagents/materials/analysis tools: TL ED Geuvadis consortium. Contributed to the writing of the manuscript: LG AB. Provided predicted putative cleavage sites: PGF.

References

1. Mitelman F, Johansson B, Mertens F (2007) The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 7: 233–245. doi:10.1038/nrc2091.
2. Magrangeas F, Pitiot G, Dubois S, Bragado-Nilsson E, Chérel M, et al. (1998) Cotranscription and intergenic splicing of human galactose-1-phosphate uridylyltransferase and interleukin-11 receptor alpha-chain genes generate a fusion mRNA in normal cells. Implication for the production of multidomain proteins during evolution. J Biol Chem 273: 16005–16010.
3. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, et al. (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res 15: 987–997. doi:10.1101/gr.3455305.
4. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, et al. (2007) Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. Genome Res 17: 746–759. doi:10.1101/gr.5660607.
5. Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, et al. (2006) Tandem chimerism as a means to increase protein complexity in the human genome. Genome Res 16: 37–44. doi:10.1101/gr.4145906.
6. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, et al. (2006) Transcription-mediated gene fusion in the human genome. Genome Res 16: 30–36. doi:10.1101/gr.4137606.
7. Prakash T, Sharma VK, Adati N, Ozawa R, Kumar N, et al. (2010) Expression of conjoined genes: another mechanism for gene regulation in eukaryotes. PLoS ONE 5: e13284. doi:10.1371/journal.pone.0013284.
8. Gingeras TR (2009) Implications of chimaeric non-co-linear transcripts. Nature 461: 206–211. doi:10.1038/nature08452.
9. Nacu S, Yuan W, Kan Z, Bhatt D, Rivers CS, et al. (2011) Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. BMC Med Genomics 4: 11. doi:10.1186/1755-8794-4-11.
10. Hernández-Sánchez C, Bártulos O, Valenciano AI, Mansilla A, de Pablo F (2006) The regulated expression of chimeric tyrosine hydroxylase-insulin transcripts during early development. Nucleic Acids Res 34: 3455–3464. doi:10.1093/nar/gkl436.
11. Brooks YS, Wang G, Yang Z, Smith KK, Bieberich E, et al. (2009) Functional pre- mRNA trans-splicing of coactivator CoAA and corepressor RBM4 during stem/progenitor cell differentiation. J Biol Chem 284: 18033–18046. doi:10.1074/jbc.M109.006999.
12. Kunarso G, Wong K-Y, Stanton LW, Lipovich L (2008) Detailed characterization of the mouse embryonic stem cell transcriptome reveals novel genes and intergenic splicing associated with pluripotency. BMC Genomics 9: 155. doi:10.1186/1471-2164-9-155.
13. Kim D-S, Kim D-W, Kim M-Y, Nam S-H, Choi S-H, et al. (2012) CACG: a database for comparative analysis of conjoined genes. Genomics 100: 14–17. doi:10.1016/j.ygeno.2012.05.005.
14. Kim RN, Kim A, Choi S-H, Kim D-S, Nam S-H, et al. (2012) Novel mechanism of conjoined gene formation in the human genome. Funct Integr Genomics 12: 45–61. doi:10.1007/s10142-011-0260-1.
15. Duc C, Sherstnev A, Cole C, Barton GJ, Simpson GG (2013) Transcription termination and chimeric RNA formation controlled by Arabidopsis thaliana FPA. PLoS Genet 9: e1003867. doi:10.1371/journal.pgen.1003867.
16. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506–511. doi:10.1038/nature12531.
17. Ge H, Liu K, Juan T, Fang F, Newman M, et al. (2011) FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics 27: 1922–1928. doi:10.1093/bioinformatics/btr310.
18. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589–595. doi:10.1093/bioinformatics/btp698.
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. doi:10.1093/bioinformatics/btp352.
20. Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, et al. (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res 20: 623–635. doi:10.1101/gr.102970.109.
21. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773–777. doi:10.1038/nature08903.
22. Thierry-Mieg D, Thierry-Mieg J (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol 7 Suppl 1: S12.1–14. doi:10.1186/gb-2006-7-s1-s12.
23. Wickham H (2009) Ggplot2: elegant graphics for data analysis. New York: Springer. 212 p.
24. Marco-Sola S, Sammeth M, Guigó R, Ribeca P (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods 9: 1185–1188. doi:10.1038/nmeth.2221.
25. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773–777. doi:10.1038/nature08903.
26. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, et al. (2012) Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet 3: 35. doi:10.3389/fgene.2012.00035.
27. Jain R, Devine T, George AD, Chittur SV, Baroni TE, et al. (2011) RIP-Chip analysis: RNA-Binding Protein Immunoprecipitation-Microarray (Chip) Profiling. Methods Mol Biol 703: 247–263. doi:10.1007/978-1-59745-248-9_17.
28. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, et al. (2013) ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res 41: D56–63. doi:10.1093/nar/gks1172.
29. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842. doi:10.1093/bioinformatics/btq033.
30. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22: 1760–1774. doi:10.1101/gr.135350.111.
31. Ferreira PG, Monlong A, Gonzàlez-Porta M, Barann M, Lappalainen T, et al. (1914) Small changes in the big picture: genetic fine-tuning in transcriptome. variation across human populations. PLoS Genet submitted.
32. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, et al. (2010) LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26: 2336–2337. doi:10.1093/bioinformatics/btq419.
33. Carrara M, Beccuti M, Cavallo F, Donatelli S, Lazzarato F, et al. (2013) State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? BMC Bioinformatics 14 Suppl 7: S2. doi:10.1186/1471-2105-14-S7-S2.
34. Courseaux A, Richard F, Grosgeorge J, Ortola C, Viale A, et al. (2003) Segmental duplications in euchromatic regions of human chromosome 5: a

source of evolutionary instability and transcriptional innovation. Genome Res 13: 369–381. doi:10.1101/gr.490303.
35. Lee Y, Ise T, Ha D, Saint Fleur A, Hahn Y, et al. (2006) Evolution and expression of chimeric POTE-actin genes in the human genome. Proc Natl Acad Sci USA 103: 17885–17890. doi:10.1073/pnas.0608344103.
36. Seeler JS, Marchio A, Sitterlin D, Transy C, Dejean A (1998) Interaction of SP100 with HP1 proteins: a link between the promyelocytic leukemia-associated nuclear bodies and the chromatin compartment. Proc Natl Acad Sci USA 95: 7316–7321.
37. Salzman J, Marinelli RJ, Wang PL, Green AE, Nielsen JS, et al. (2011) ESRRA-C11orf20 is a recurrent gene fusion in serous ovarian carcinoma. PLoS Biol 9: e1001156. doi:10.1371/journal.pbio.1001156.
38. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. doi:10.1038/nature11632.
39. Frenkel-Morgenstern M, Lacroix V, Ezkurdia I, Levin Y, Gabashvili A, et al. (2012) Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. Genome Res 22: 1231–1242. doi:10.1101/gr.130062.111.
40. Kawase T, Akatsuka Y, Torikai H, Morishima S, Oka A, et al. (2007) Alternative splicing due to an intronic SNP in HMSD generates a novel minor histocompatibility antigen. Blood 110: 1055–1063. doi:10.1182/blood-2007-02-075911.
41. Asmann YW, Necela BM, Kalari KR, Hossain A, Baker TR, et al. (2012) Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. Cancer Res 72: 1921–1928. doi:10.1158/0008-5472.CAN-11-3142.
42. Niwa M, Rose SD, Berget SM (1990) In vitro polyadenylation is stimulated by the presence of an upstream intron. Genes Dev 4: 1552–1559.
43. Niwa M, Berget SM (1991) Mutation of the AAUAAA polyadenylation signal depresses in vitro splicing of proximal but not distal introns. Genes Dev 5: 2086–2095.
44. Rigo F, Martinson HG (2008) Functional coupling of last-intron splicing and 3′-end processing to transcription in vitro: the poly(A) signal couples to splicing before committing to cleavage. Mol Cell Biol 28: 849–862. doi:10.1128/MCB.01410-07.
45. Kandul NP, Noor MAF (2009) Large introns in relation to alternative splicing and gene evolution: a case study of Drosophila bruno-3. BMC Genet 10: 67. doi:10.1186/1471-2156-10-67.
46. Takahara T, Tasic B, Maniatis T, Akanuma H, Yanagisawa S (2005) Delay in synthesis of the 3′ splice site promotes trans-splicing of the preceding 5′ splice site. Mol Cell 18: 245–251. doi:10.1016/j.molcel.2005.03.018.
47. Wang Y, Wang Z (2013) Systematical identification of splicing regulatory cis-elements and cognate trans-factors. Methods. doi:10.1016/j.ymeth.2013.08.019.
48. Qu H-Q, Lu Y, Marchand L, Bacot F, Fréchette R, et al. (2007) Genetic control of alternative splicing in the TAP2 gene: possible implication in the genetics of type 1 diabetes. Diabetes 56: 270–275. doi:10.2337/db06-0865.
49. Behm-Ansmant I, Gatfield D, Rehwinkel J, Hilgers V, Izaurralde E (2007) A conserved role for cytoplasmic poly(A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay. EMBO J 26: 1591–1601. doi:10.1038/sj.emboj.7601588.
50. Mangus DA, Evans MC, Jacobson A (2003) Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. Genome Biol 4: 223. doi:10.1186/gb-2003-4-7-223.
51. Bag J, Bhattacharjee RB (2010) Multiple levels of post-transcriptional control of expression of the poly (A)-binding protein. RNA Biol 7: 5–12.
52. Francis RW, Thompson-Wicking K, Carter KW, Anderson D, Kees UR, et al. (2012) FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data. PLoS ONE 7: e39987. doi:10.1371/journal.pone.0039987.
53. Plebani R, Oliver GR, Trerotola M, Guerra E, Cantanelli P, et al. (2012) Long-range transcriptome sequencing reveals cancer cell growth regulatory chimeric mRNA. Neoplasia 14: 1087–1096.
54. Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V (2011) Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. Bioinformatics 27: 1068–1075. doi:10.1093/bioinformatics/btr085.
55. Cancer Genome Atlas Research Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499: 43–49. doi:10.1038/nature12222.
56. Chase A, Ernst T, Fiebig A, Collins A, Grand F, et al. (2010) TFG, a target of chromosome translocations in lymphoma and soft tissue tumors, fuses to GPR128 in healthy individuals. Haematologica 95: 20–26. doi:10.3324/haematol.2009.011536.
57. Holt R, Sykes NH, Conceição IC, Cazier J-B, Anney RJL, et al. (2012) CNVs leading to fusion transcripts in individuals with autism spectrum disorder. Eur J Hum Genet 20: 1141–1147. doi:10.1038/ejhg.2012.73.
58. Wang K, Ubriaco G, Sutherland LC (2007) RBM6-RBM5 transcription-induced chimeras are differentially expressed in tumours. BMC Genomics 8: 348. doi:10.1186/1471-2164-8-348.
59. Zhang Y, Gong M, Yuan H, Park HG, Frierson HF, et al. (2012) Chimeric transcript generated by cis-splicing of adjacent genes regulates prostate cancer cell proliferation. Cancer Discov 2: 598–607. doi:10.1158/2159-8290.CD-12-0042.