BMC Genomics

RESEARCH ARTICLE Open Access

# Identification and analysis of pig chimeric mRNAs using RNA sequencing data

Lei Ma[1,2†], Shulin Yang[1†], Weiming Zhao[1†], Zhonglin Tang[1], Tingting Zhang[2] and Kui Li[1*]

## Abstract

**Background:** Gene fusion is ubiquitous over the course of evolution. It is expected to increase the diversity and complexity of transcriptomes and proteomes through chimeric sequence segments or altered regulation. However, chimeric mRNAs in pigs remain unclear. Here we identified some chimeric mRNAs in pigs and analyzed the expression of them across individuals and breeds using RNA-sequencing data.

**Results:** The present study identified 669 putative chimeric mRNAs in pigs, of which 251 chimeric candidates were detected in a set of RNA-sequencing data. The 618 candidates had clear trans-splicing sites, 537 of which obeyed the canonical GU-AG splice rule. Only two putative pig chimera variants whose fusion junction was overlapped with that of a known human chimeric mRNA were found. A set of unique chimeric events were considered middle variances in the expression across individuals and breeds, and revealed non-significant variance between sexes. Furthermore, the genomic region of the 5′ partner gene shares a similar DNA sequence with that of the 3′ partner gene for 458 putative chimeric mRNAs. The 81 of those shared DNA sequences significantly matched the known DNA-binding motifs in the JASPAR CORE database. Four DNA motifs shared in parental genomic regions had significant similarity with known human CTCF binding sites.

**Conclusions:** The present study provided detailed information on some pig chimeric mRNAs. We proposed a model that trans-acting factors, such as CTCF, induced the spatial organisation of parental genes to the same transcriptional factory so that parental genes were coordinatively transcribed to give birth to chimeric mRNAs.

**Keywords:** Chimeric mRNA, Trans-splicing, RNA-sequencing, CTCF, Pig

## Background

Chimeric mRNAs fused by two previously separate genes located on different genomic loci may allow a limited number of genes to encode a substantially large number of mRNAs and proteins. They are expected to increase proteomic diversity through chimeric proteins or altered regulation. As a consequence, gene fusion can change the properties of precursor proteins and can even perturb normal regulatory pathways and initiate or stimulate neoplastic cell growth. A well-known example is the *BCR-ABL1* fusion gene, which is the result of the chromosomal translocation t(9; 22)(q34; q11) and is responsible for 90% of chronic myelogenous leukemia cases [1]. In this sense, chimeric genes can be used as desirable therapeutic targets for cancers. For instance, matinib mesylate (Gleevec, Novartis) can target the oncogenic kinase activity of *BCR-ABL1* in chronic myeloid leukemia [2-4]. Therefore, the identification and analysis of novel chimeric genes will pave the way for a greater understanding of the role of gene fusion.

Chromosomal translocation is generally responsible for the generation of some chimeric mRNAs in cancer cells. Therefore, chimeric mRNAs are often viewed as potential diagnostic biomarkers for tumours caused by chromosomal translocation. However, a low amount of a chimeric RNA (*JAZF1-JJAZ1*) was detected in normal endometrial tissues, joining the *JAZF1* gene on chromosome band 7p15 to the *JJAZ1/SUZ12* gene on chromosome band 17q21 [5]. Chimeric RNAs and proteins are identical to those produced from a chromosomal rearrangement found in human endometrial stromal tumours [5]. The explanation generally offered for this

* Correspondence: kuili@iascaas.net.cn
†Equal contributors
[1]The Key Laboratory for Domestic Animal Genetic Resources and Breeding of Ministry of Agriculture of China, Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing 100193, P. R. China
Full list of author information is available at the end of the article

finding is that specific chromosomal rearrangements occur within small numbers of cells in healthy tissues. However, no rearranged bands t(7;17)(p15;q21) were detected in normal cells [5]. Given the absence of any detectable rearranged DNA in cells producing chimeric RNAs, the obvious explanation is the rearrangement at the RNA level. After incubation of mixed extracts from a human endometrial stromal cell line and from a rhesus monkey fibroblast cell line, rhesus *JAZF1* exons were joined to human *JJAZ1* exons, implying that the *JAZF1-JJAZ1* RNA is a result of trans-splicing [5].

In eukaryotes, trans-splicing is a special event in RNA processing where exons from two different primary RNA transcripts are joined from end to end and then ligated. In simulating the RNA cis-splicing mechanism, a cDNA is thought to be generated by trans-splicing when it is aligned to multiple non-contiguous genomic loci and the fusion junction obeys canonical GU-AG splice site. However, how precursor genes find each other before splicing remains to be elucidated, and where the trans-splicing event takes place is still poorly understood.

Some chimeras are derived from a non-spliceosome mechanism [6]. Short homologous sequences are proposed to be associated with the generation of chimeric mRNAs in eukaryotes, suggesting that the 'misaligns' of short homologous sequences could guide the chromosomal interaction for the proximity of distal genes [7]. In addition, read-through/splicing is another way of generating chimeric mRNAs [8-11]. In this process, an mRNA starts from the upstream gene, reads through intergenic regions, and ends at a termination point of the adjacent downstream gene, with the region in between removed by splicing. However, read-through/splicing cannot explain the chimeras derived from different chromosomes or opposite strands. Some chimeric mRNAs may have originated from the strand-switching feature of the reverse transcripatase [12]. In some cases, chimeric mRNAs are considered as artefacts from the reverse transcription polymerase chain reaction (RT-PCR) [12].

The presence of chimeric mRNAs in normal cells is a critical issue because the important pathways in normal cells would be disrupted by the potential therapy targeting chimeric mRNAs and proteins. The identification of chimeric mRNAs in normal cells will provide a wealth of biological information for this issue. The pig (*Sus scrofa*) is an economically important species and a potential medical model for some human health issues [13]. Therefore, research on chimeric mRNAs in normal cells can benefit from pigs. Results from the present study provide the first broad overview of chimeric mRNAs in pigs, and their analysis in normal tissue will aid in the further understanding of the molecular mechanisms of gene fusions.

## Results

### Identification of putative chimeric RNAs

After inspecting the chromosomal loci of mRNAs from the pigs, many mRNAs were located on non-contiguous positions. An issue whether any of these mRNAs are real chimeras fused from two previously separate transcripts was raised. Highly qualitative alignments of mRNAs to the *S. scrofa* chromosomes (SGSC Sscrofa9.2/susScr2, Nov. 2009) in the Genome Browser database of the University of California Santa Cruz (UCSC) [14,15] may shed light on this issue. Alignments having at least 96% sequence identity and a minimum length of 100 nt were used in this study. We only used mRNAs that were matched on two non-contiguous loci to ensure that inferred chimeras were results of actual fused transcripts rather than alignment artefacts. Alignments from two non-contiguous loci were required not to possess long similar sequences at the putative junction sites to discard false positive results from homologous, paralogous, or random spurious hits. In this step, we only allowed overlaps or gaps of up to 10 nt within the fusion junction to accommodate small errors in alignment that occur at the edges of the alignment. Consequently, 669 mRNAs were inferred as putative chimeras (Additional file 1), including 27 inter-chromosomal and 642 intra-chromosomal junction events. In the intra-chromosomal events, 494 and 148 mRNAs were inter-strand and intra-strand junction events, respectively. Only three candidates involved mRNAs from the mitochondrial genome. Figure 1 displays the distribution of putative chimeric mRNAs in chromosomes, showing that inter-strand events are over-represented in the set of predicted chimeric mRNAs.

For the confirmation of a hybrid transcript candidate, we inspected whether the fusion point corresponded to a pair of known splice sites. We separately extracted the chromosomal DNA sequences of the 5′ and 3′ partners of an inferred chimera and then connected the two non-contiguous genomic sequences to an artificially fused genomic sequence. Each inferred chimera was aligned to the corresponding artificially fused genomic sequence using the SIM4 program [16] to take into account consensus splice signals. The alignment around the fusion point was checked. Only the fusion points that were aligned precisely, without a gap or overlap, were retained. In addition, the reading frame must have structural integrity. Finally, 618 candidates had clear trans-splicing sites, 537 of which obeyed the GT/AG rule (Additional file 1).

To confirm further the trans-splicing events, 48 chimeric candidates were randomly selected for the RT-PCR assay using RNA from a number of tissues (see Methods). An RT-PCR product was required to span the fusion point. Through this assay, 36 out of randomly selected candidates showed identity with the expected

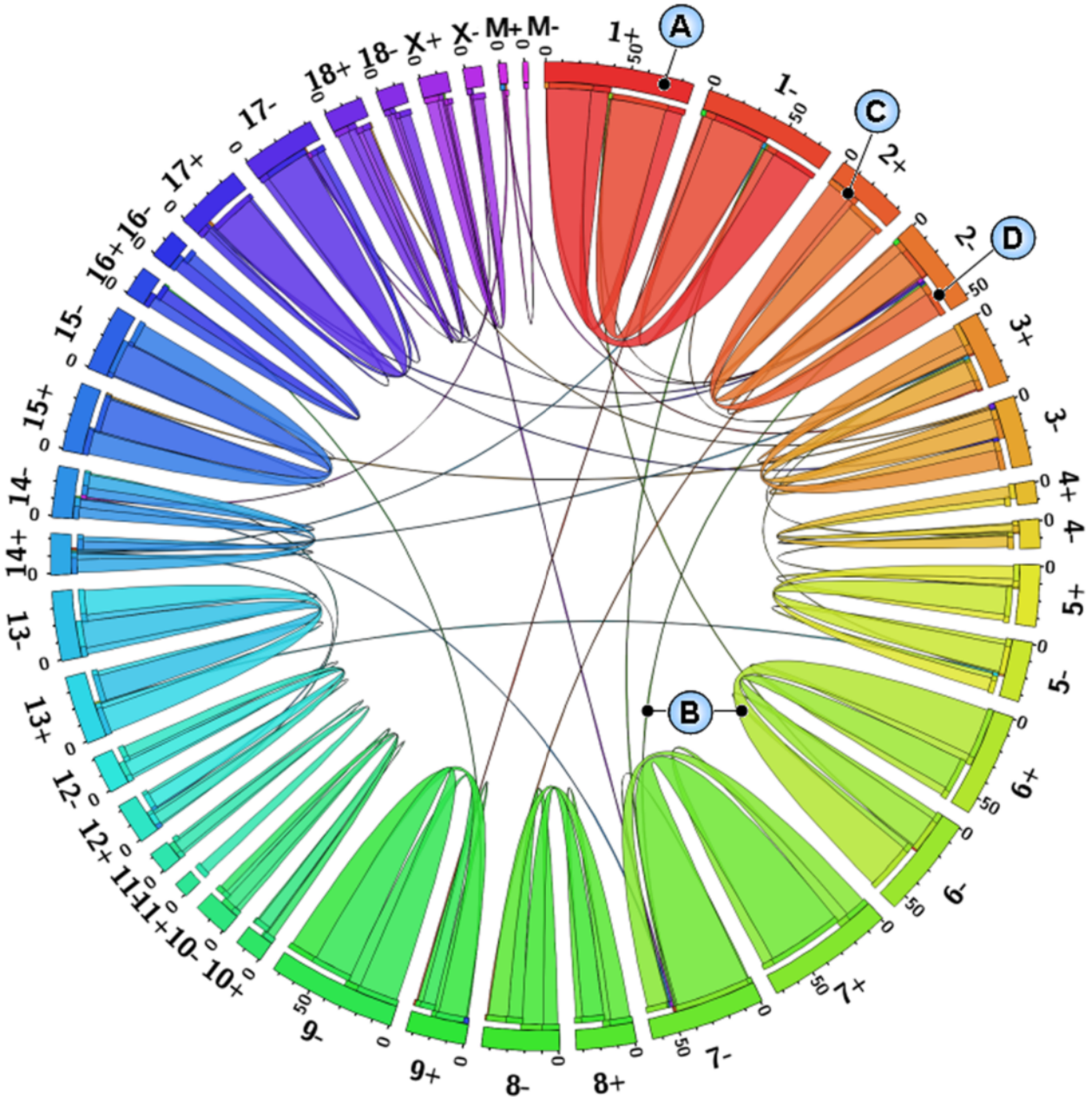


**Figure 1 Circular representation of the genome-wide distribution of putative chimeric mRNAs.** The outermost labels indicate the chromosome name with strand orientation. Each coloured segment (**A**) in the outermost circle encodes a chromosomal strand. Each bin in a segment represents ten events. Inner Ribbons (**B**) indicate the associated fusions from the 5′ partner to the 3′ partner. Ribbons are coloured according to the chromosomal strands in which the 5′ partners locate. Ribbons start from the 5′ partners with ribbon ends (**C**) coloured regarding the destination and stop at the 3′ partners with gaps (**D**) between the ribbons and associated segments. Ribbon size encodes the relative abundance of the associated fusion, that is, the count of putative chimeric mRNAs.

fusion sequences (Additional file 2). Given that the transcription of mRNAs may vary in different tissues or stages of life, the selected samples for the RT-PCR assay may not be suitable for their expression. In addition, all mRNAs used in the present study have prior biological studies annotated in databases of the UCSC and the NCBI (National Centre for Biotechnology Information). Thus, the rate detected by the RT-PCR assay might underestimate the positive rate of chimeric mRNA identification. The use of expressed sequence tag (EST) and RNA-sequencing data from more tissues or stages would supply the gaps of the RT-PCR assay.

Putative chimeras were aligned to ESTs downloaded from the UCSC database to seek support from external

experimental evidences and verify the putative fusion junctions. If at least 20 nt of the sequence on either side of a putative fusion point overlap with the ESTs, this candidate was retained for further analysis. The 431 candidates were supported by at least three ESTs (Additional file 3).

### Mapping putative pig chimeras to known human chimeric transcripts

Putative pig chimeric mRNAs were aligned to known human chimeric transcripts annotated in the chimera database (ChimeraDB 2.0) to estimate the relationship between two kinds of transcripts [17]. The fusion junctions of 21 putative pig chimeric mRNAs were matched to known human chimeric mRNAs (Additional file 4). However, only two putative pig chimera variants (AK239284 and AK349030) whose fusion junction was overlapped with that of a known human chimeric mRNA (AML1/AMP19 fusion gene) were found.

### Validation by transcriptome sequencing

We collected 396.2 million sequence reads from the transcriptome sequencing of liver tissue samples from 11 adult Bama miniature pigs (five males and six females, Additional file 5). This procedure was done to verify that the putative chimeric mRNAs were real expressed genes rather than involved in exonic coding sequences shared among multiple genes or homologous pseudogenes. The Illumina Genome Anlayzer II was employed to sequence these samples. Two length types of single-end reads, 76 and 101 nt, were generated (Additional file 5). For the uniformity of the read length, 101 nt reads were trimmed to 76 nt from a low-quality (right) end, which would increase the quality of 101 nt reads.

Reads from different samples were mapped on the pig genome (SGSC Sscrofa9.2/susScr2, Nov. 2009) using the Bowtie software (version 0.12.8) [18]. This Bowtie version does not report gapped alignments. Hence, a read mapped on the genome was derived from a contiguous genomic locus. In this sense, some unmapped reads may have originated from non-contiguous genomic loci and may therefore be suitable in inspecting splice junctions. Based on this fact, these unmapped reads were aligned to putative chimeric mRNA. In this step, we required that the junction reads should overlap with at least 5 nt of the sequence on either side of the chimeric junction. Furthermore, a fusion junction mapped by junction reads derived from at least three different start positions or at least three samples was considered a validated chimeric event. Consequently, up to 443 fusion junctions were validated by this strategy (Additional file 6). The 440 and 184 events were expressed in at least three and all samples, respectively (Figure 2A).

Estimation was further performed on the validity of junction reads that overlapped fusion points with a minimum of 5 nt. In the present study, reads were trimmed to 76 nt. Therefore, the length of fusion junctions was 142 nt (71 nt on either side of the fusion junction) by requiring a 5 nt overhang for read mapping fusion points. If the start position of a read located in the region from the 1st to the 67th nt of the fusion junction, the read was termed as a junction read. In this estimation, reads from 11 liver samples were pooled together. The 496 fusion junctions were matched by at least one read. Among these junctions, 89.3% (443/496) were overlapped by at least three reads and 89.7% (440/496) were validated by reads starting from at least three different positions (Additional file 7).

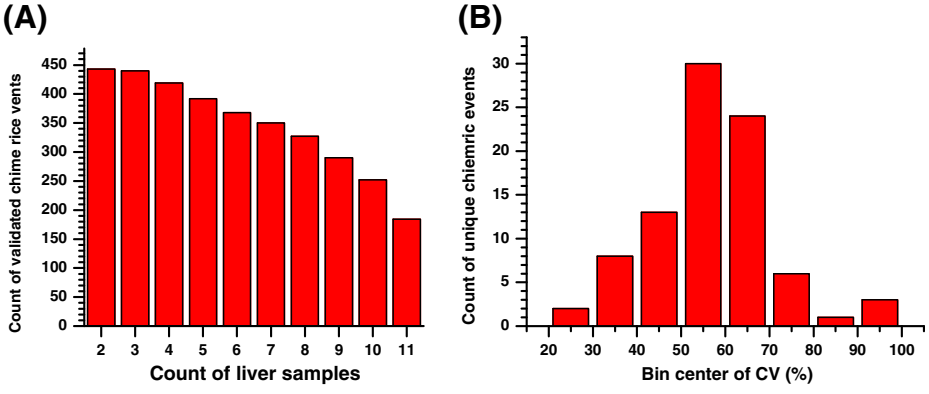


**Figure 2 Transcription of chimeric mRNAs in liver samples.** (**A**) Count of validated chimeric events as grouped according to the count of liver samples. A fusion junction mapped by junction reads derived from at least three different start positions or at least three samples was considered a validated chimeric event. (**B**) Distribution of unique chimeric events along the CV. The CV is the percentage ratio of the sample standard deviation to the sample mean of the junction reads for each event. The 87 unique chimeric events were put in eight bins according to the CV.

Interestingly, we observed a non-uniform distribution of reads along some mRNA sequences. For example, the read coverage showed multi-peaks along the mRNA sequence of AK346347 (Figure 3). Given that chimeric mRNAs share sequences with their precursor genes, determining which reads come from chimeras is necessary. Reads mapped on the fusion junctions were derived only from chimeric mRNAs. However, reads mapped on positions away from the fusion junctions would be derived from either chimeric mRNAs or their participating parental transcripts. An apparent trough was detected in the region from 400 nt to 420 nt, corresponding to the trans-splicing site at 403 nt (Figure 3). The lower read coverage along this chimeric junction indicated the lower expression of this chimeric gene relative to its precursors. In addition, except for the trans-splicing site, other troughs were closed to the cis-splicing sites of AK346347, indicating the existence of spliced variants of AK346347 among the samples.

### Variation of expression among individuals

We used a cut-off that required junction events to be present in all samples and unique without an overlap with other chimeras to access further the differential expression of unique chimeric mRNAs without the confounding issues of tissues. This cut-off resulted in 87 unique chimeric events. The dispersion of the expression of each unique chimeric event across the samples was measured using the coefficient of variation (CV), the percentage ratio of the sample standard deviation to the sample mean of the junction reads for each event. Figure 2B represents the distribution of junction events along the CVs. The mean of the CVs was 57%, with a standard deviation of 14%, following a normal distribution ($P$>0.57, *Kolmogorov-Smirinov test*). This result implies that most of these unique chimeric events were considered middle variances in the expression.

We compared the expression of these events between male and female samples to gain further insight. The mean CV of the females was lower than that of the males (52% versus 57%). Unique events were ordered using a nonparametric two-sided rank sum (RS) test, a statistical test that considers the difference in expression levels between male and female samples (Figure 4). The $P$-value of all events, except for three, was greater than 0.05, indicating non-significant variance in the expression of these events between sexes.

### Variation of expression among pig breeds

More attention was given to the variation in the expression of chimeric events among the pig breeds. A set of 49 nt single-end reads from three RNA-pooling samples of skeletal muscle was analysed in the same way as those from liver samples (Additional file 5, Additional file 6). These samples were obtained during embryo collection at slaughter. The first, second, and last samples were pooled using equivalent amounts of RNA from three

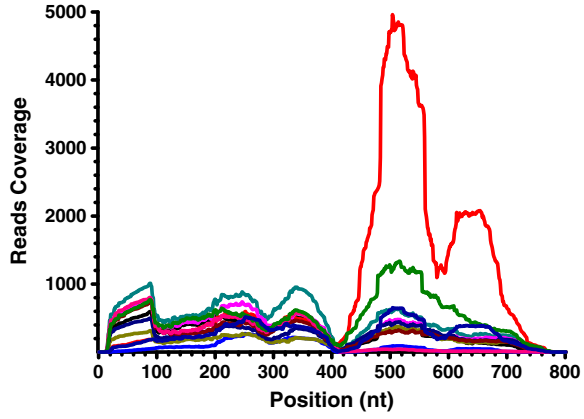


**Figure 3 Reads not uniformly distributed along mRNA.** The graph is an example showing that the read coverage is not uniformly distributed along the AK346347 transcript. Read coverage at a nucleotide position was determined by enumerating the total reads mapped on that position. The 11 liver samples were all represented.

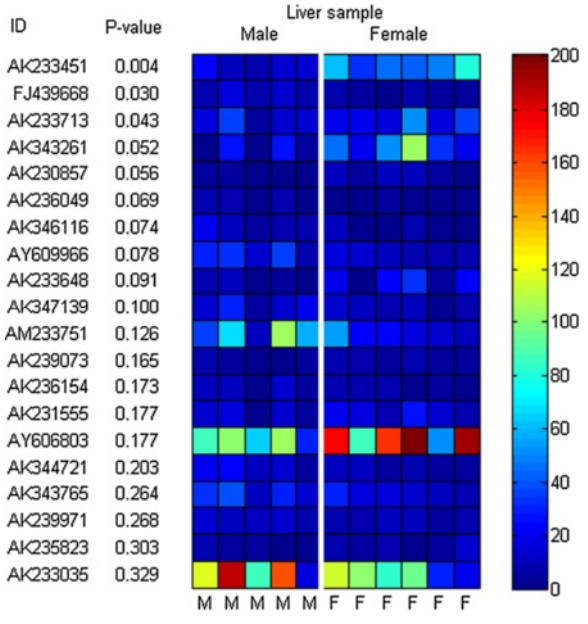


**Figure 4 Difference in the transcription between sexes.** The difference in the transcription of unique chimeric events between five males and six females was evaluated using a two-sided rank sum test. Top 20 events based on the $P$-value from the test were shown. The leftmost column shows the GenBank ID and the second displays the $P$-value. Each cell in the heat-map encodes the count of junction reads for each unique chimeric mRNAs in each sample. Each vertical column represents data from one sample and each horizontal row represents the relative abundance of one unique chimeric mRNAs across samples. Columns were grouped by sexes.

adult female Wuzhishan, Tongcheng, and Landrace pigs, respectively. These samples may remove the difference among female individuals to some extent. The mean of the CVs was 35%, with a standard deviation of 19%, spanning 0% to 89% and following a normal distribution ($P$>0.61, *Kolmogorov-Smirinov test*).

### DNA motif in the genomic region of chimeras

To exploit the putative mechanism responsible for the generation of chimeric mRNAs, we attempted to retrieve DNA motif sequences in two non-continuous genomic loci of chimeric mRNAs using the MEME software (Motif-based sequence analysis tools, version 4.6.1) [19]. In this step, for 445 putative chimeric mRNAs, similar DNA sequences were found between the 5′ and 3′ partners (Additional file 8). Similar sequences were prevalent in the upstream, intronic, and downstream sequences, but deficient in exons (Table 1). None of the similar sequences was found in the exonic pair of the 5′ and 3′ partners. The lack of similarity in the exonic pair may result from the elimination of chimeras with long overlapping sequences in the fusion junction, implying that although we cannot entirely exclude false positive results from homologous or paralogous genes, we minimised the effect of these events on the identification as much as possible. This result agrees with the suggestion that some regulatory elements, such as transcription factor binding sites or enhancers, are highly pronounced in non-coding regions.

Subsequently, these shared sequences were submitted to the TOMTOM [20] software in the MEME suite (4.6.1) [19] for comparison against the database of known motifs. This database is the JASPAR CORE (version 12-Oct-2009) that contains a curated, non-redundant set of profiles derived from published collections of experimentally defined transcription factor binding sites for multi-cellular eukaryotes [21]. The 81 shared sequences significantly matched known DNA motifs in the JASPAR CORE database ($P$<0.00065 and *false motif discovery rate* < 0.05, Additional file 9). Among these matched sequences, 6 were shared in the upstream regions of both partners ($P$<0.00009 and *false motif discovery rate* < 0.042). This finding suggests that the same or similar transcription factors would bind these potential shared DNA motifs to coordinate the transcription of parental genes, which may be necessary in generating chimeric mRNAs.

**Table 1 Distribution of potential shared DNA motifs in genomic regions**

| Type | 3′ Up | 3′ Exon | 3′ Intron | 3′ Down | Total |
|---|---|---|---|---|---|
| 5′ Up | 72 (6) | 1 (0) | 46 (9) | 55 (7) | 174 (22) |
| 5′ Exon | 0 (0) | 0 (0) | 2 (0) | 0 (0) | 2 (0) |
| 5′ Intron | 38 (2) | 0 (0) | 38 (10) | 42 (18) | 118 (30) |
| 5′ Down | 66 (11) | 1 (0) | 43 (4) | 41 (14) | 151 (29) |
| Total | 176 (19) | 2 (0) | 129 (23) | 138 (39) | 445 (81) |

The table indicates the count of potential DNA motifs shared within the genomic regions of the 5′ and 3′ partners. The rows represent the genomic regions for the upstream, exon, intron, and downstream regions of the 5′ partner and the columns of the 3′ partner. Numbers in parentheses indicate the count of shared sequences that significantly match the known DNA motifs in the JASPAR CORE ($P$<0.00065 and *false motif discovery rate* < 0.05). Up: upstream; Down: downstream.

The CCCTC-binding factor (CTCF) is a versatile trans-acting factor that binds distal regulatory elements such as enhancers, and CTCF binding sites are commonly distributed along the vertebrate genomes [22-26]. Thus, we placed efforts on computationally identifying potential CTCT binding sites shared in two non-continuous genomic regions of chimeric mRNAs. Four DNA motifs shared in parental genomic regions were significantly similar with known human CTCF binding sites ($P$<0.014 and *false motif discovery rate* < 0.029, Additional file 10). This result suggests that some trans-acting factors, such as the CTCT-binding factor, might bind these shared motifs to facilitate the approximation of the distal genomic parts and make up the subcellular environment for the generation of chimeric mRNAs. Communication between distal chromosomal elements would be an origin for the nuclear processes of gene fusions.

## Discussion

Following the hypothesis that a fusion transcription is derived by two non-continuously genomic loci, the present study revealed a list of pig chimeric mRNAs validated by the RNA-Seq and EST data (Figure 5). A set of unique chimeric mRNAs showed a middle variance among both individuals and breeds. The results provided detailed information regarding pig chimeric mRNAs and important implications for gene fusions.

Several factors including strand-switching, deep sequencing errors, or reference genome errors would result in false positive results. Therefore, we rigorously inspected each chimera using several criteria. First, all the mRNAs used in the present study have prior biological information annotated in the UCSC and NCBI databases to avoid reference genome errors as much as possible. To remove false results from homologous, paralogous, or random spurious hits, strict filtering was performed on the highly qualitative alignments of mRNAs to the *S. scrofa* chromosomes. Trans-splicing sites were then inspected for each candidate to exclude strand-switching or the random connection of two cDNAs. In addition, 14 independent samples were used to evaluate the expression of the fusion transcripts. We could not completely rule out the possibility of the creation of a false fusion in the process of cDNA library construction. However, random breakage and rejoining of two cDNAs are unlikely to happen at the exact exon boundaries of

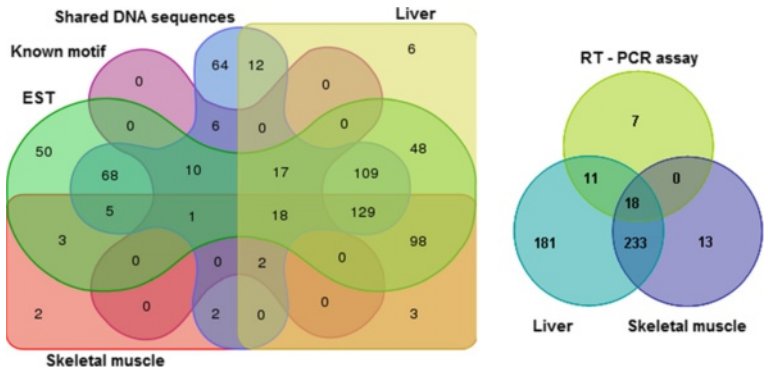


**Figure 5 Venn diagram showing the intersection of different groups.** The EST group represents chimeras that overlap ESTs ($n = 557$). The liver and skeletal muscle groups reveal chimeras that were present in two kind of samples ($n = 443$ and 264). The shared DNA sequences group represents chimeras with similar DNA motifs in the genomic regions of both partners ($n = 445$). The known motif group indicates the shared sequences that significant match the known DNA motifs ($n = 81$). The RT-PCR group shows chimeras validated by the RT-PCR assay ($n = 36$).

two genes and simultaneously in multiple samples. Thus, although the present identification of chimeric RNAs filters out some genuine fusion gene transcripts by stringent cut-offs, it is conservative and reliable.

Interestingly, the transcriptional reading-through was infrequently involved in the intra-strand chimeric candidates identified in the present study. RNA-polymerase generally ends at transcriptional terminators, preventing it from reading through the downstream gene. However, in unusual cases, long transcriptions span terminators and produce new, hybrid, multi-locus transcripts [8-11]. We checked the coordinates of putative chimeras on the pig genome and found that the exons of most intra-strand chimeras were out of order compared with those found in the genome. For example, the 5′ partner of AK238425 is located at the downstream of the 3′ partner on the plus strand in chromosome 16 (Figure 6A). AK351564 is another example that a 3′ partner lies in the body of the 5′ partner (Figure 6B). Among the 150 intra-strand chimeras, 142 follow the AK238425 way, five take the AK351564 way, and two occur in the third way that the 5′ partner lies in the body of the 3′ partner. The skewed representation of chimeras in the three ways was due to the elimination of chimeras with long overlapping sequences in the fusion junction. Only one chimera follows the order that the 5′ partner is located in the upstream of the 3′ partner. However, the distance between partners is 57,234,234 nt.

During transcription in vivo, different genes frequently share the same transcription factory where nascent RNA production and RNA polymerase II seem to be localised [27,28]. For example, the *Igh* on chromosome 12 is preferentially recruited to the same transcription factory where the *Myc* gene on chromosome 15 is highly transcribed [29]. Many active genes can dynamically co-localise to shared sites of ongoing transcription, which may be induced by the classical effectors of gene expression including trans-acting factors, enhancers, chromatin modifications, and chromosomal interaction [27]. For example, CTCF can create the dynamic nature of nuclear spatial organisation of different genes by binding to the elements on distal genomic regions or different chromosomes [25,30,31]. The recruitment of different genes into shared factories is expected to have a

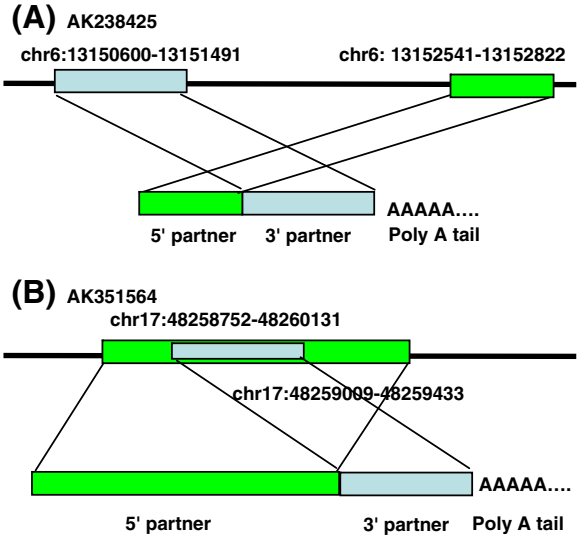


**Figure 6 Transcriptional read-through was infrequently involved in chimeric candidates.** The genomic and mRNA organization of AK238425 and AK351564 are depicted in the figure. (**A**) The 5′ partner of AK238425 is located downstream of the 3′ partner. (**B**) The 3′ partner of AK351564 lies in the body of the 5′ partner.

fundamental role in gene expression, which may efficiently share limited resources or perhaps coordinate the transcription of different genes.

The co-localisation of different genes into the same transcription factories provides insights into the origin of chimeric mRNAs. We found similar sequences shared in the 5′ and 3′ partners of some chimeric mRNAs. Some shared DNA motifs significantly matched the known DNA-binding motifs. For example, four shared DNA motifs have significant similarity with known human CTCF binding sites. The CTCF can recognise and bind to different DNA motifs by its zinc-finger domains [32]. Induced by trans-acting factors, such as the CTCF, parental genes may dynamically co-localise to the shared transcription factory, and then the same or similar transcription factors coordinate the transcription of them to give birth to chimeric mRNAs (Figure 7).To some extent, this result agrees with the suggestion that short homologous sequences at the junction sites may induce the formation of chimeric mRNAs [7].

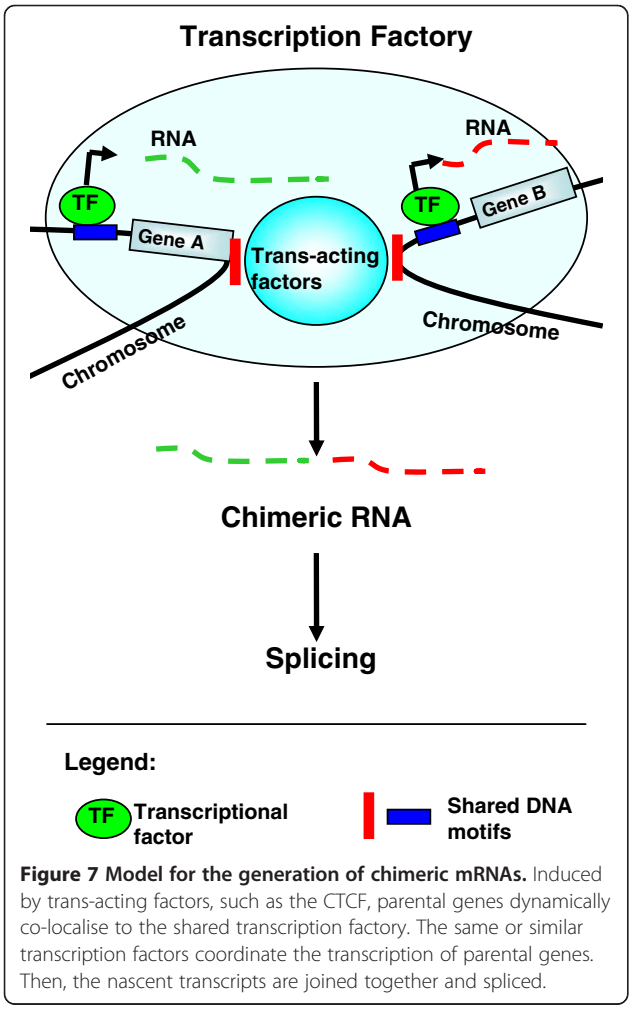


**Figure 7 Model for the generation of chimeric mRNAs.** Induced by trans-acting factors, such as the CTCF, parental genes dynamically co-localise to the shared transcription factory. The same or similar transcription factors coordinate the transcription of parental genes. Then, the nascent transcripts are joined together and spliced.

As earlier common computational methods for identifying precursor genes, a gene with the best alignments to a chimeric mRNA was considered as the precursor gene [7]. However, exons often overlap exons for some cases. For example, the 5′ partner of the chimeric mRNA AK343294 was precisely mapped on the exons of mRNAs AK233826, AK231250, and AK346646 in chromosome 5. Therefore, the precursor mRNA would be discretionary if multiple transcriptional start sites were present. Furthermore, the partners of chimeric mRNAs may be transcribed independently at their own transcriptional start sites that are not associated with other genes. Thus, the selection of which variant would serve as the precursor gene would need more molecular experimental identifications.

## Conclusions

The present study provided detailed information on pig chimeric mRNAs and further analysed the expression of unique chimeras among samples. Interestingly, similar DNA sequences widely shared in the two non-continuously genomic regions of chimeric mRNAs. Similar DNA sequences that share in the upstream regions of both partners significantly matched the known transcription factor binding sites in the JASPAR CORE database, suggesting the potential coordinated transcription of the parental genes. In addition, possible CTCF binding sites were also observed in the parental genomic regions. We supposed that trans-acting factors, such as CTCF, would induce the spatial organisation of parental genes to the same transcriptional factory so that parental genes would be coordinatively transcribed to give birth to chimeric mRNAs. Although this hypothesis needs further experimental evidence, it will provide useful information for the investigation of the mechanism for the generation of chimeric mRNAs. Overall, our results will aid in the further understanding of chimeric mRNAs.

## Methods

### Chimeric mRNA identification

The BED format table of all pig mRNAs were analysed for further study using the Galaxy [33-35] in the UCSC Table Browser (February 2011). According to the annotation of that table, GenBank pig mRNAs were aligned against the pig genome (SGSC Sscrofa9.2/susScr2, Nov. 2009) using the Blat program [36]. The alignment with the highest base identity was found when a single mRNA was aligned in multiple places. Only alignments with a base identity level within 0.5% of the best and at least 96% base identity with the genomic sequence were kept (http://genome.ucsc.edu/). An entry in that BED table annotates a chromosomal locus of an mRNA. We extracted mRNAs aligned to two non-contiguous loci. We required alignments from non-contiguous loci without long similar sequences at the

putative junction sites to remove homologous, paralogous, or random spurious hits. In accommodating small errors in alignment that occur at the edges of the alignment, we only allowed overlaps or gaps of up to 10 nt within the fusion junction. Using the Circos software [37], we represented the genome-wide distribution of putative chimeric mRNAs in Figure 1.

To validate putative chimeras by external experimental evidence, we aligned predicted chimeras to the EST sequences downloaded from the UCSC (May 2012) using the BLAST program (Basic Local Alignment Search Tool, version 2.2.26+) [38-40] with default parameters except at least 96% base identity. The candidate was retained for analysis when at least 20 nt of the sequence on either side of a putative fusion point overlapped ESTs. To compare with known human chimeras, we aligned pig putative chimeras to human chimeric mRNAs downloaded from the ChimeraDB 2.0 [17] using the BLAST with default parameters (May 2012).

### Inspection of splice sites

As previously described [41], we prepared an artificially fused genomic DNA sequence for putative chimeras by joining the genomic sequences of the 5′ and 3′ partners. The fusion transcript candidate was then aligned to the corresponding artificially fused genomic sequence using the SIM4 program (version 2002-03-03) [16] with default parameters. The alignment around the fusion point was inspected to take into account consensus splice signals.

### Validation by RT-PCR

We obtained total RNAs from Tongcheng pig tissues (liver, kidney, spleen, heart, lung, testis, ovary, embryo, skeletal muscle, small and large intestine) using the RNA Extraction Kit (BioTeke). The cDNA was prepared by reverse transcription using the Strand cDNA Synthesis Kit (BioTeke) with random hexamer priming and oligo dT's. PCR products covering the junction position were amplified using primers designed according to the hybrid transcripts (Additional file 2). PCR amplification was performed using the following thermocycling protocol: initial denaturation at 95°C for 4 min, followed by 30 cycles of denaturation at 95°C for 30 s, annealing at 60°C for 30 s, and elongation at 72°C for 30 s. The PCR products were then analyzed, cloned, and sequenced.

### Validation by RNA-seq data

Up to 400 million sequence reads from deep sequencing the transcriptome of pigs were recently acquired in our lab. In brief, the following steps were used for transcriptome sequencing using the Illumina Genome Analyser II at Shanghai Biotechnology Co., Ltd. We isolated mRNA from 10 μg of total RNA with an RNA integrity number (RIN) ≥ 8. The isolated mRNA was then fragmented and converted into double-stranded cDNA. The ends of cDNA were ligated to adapters. The fragments with 200 to 300 base pairs in length were amplified by PCR to make a library. Finally, the library was sequenced to yield single-end reads.

A set of reads was derived from the transcriptome of the liver tissue samples obtained from 11 adult Bama miniature pigs (five males and six females, Additional file 5). Reads with a Phred quality score lower than 20 were filtered out. The length of the reads from eight pigs was 76 nt, whereas that from the other three pigs was 101 nt. To obtain uniform lengths of reads, the 101 nt were trimmed from the low-quality (right) end of each read to only 76 nt before mapping. The remaining reads were aligned to the pig genome (SGSC Sscrofa9.2/susScr2, Nov. 2009) using the Bowtie software (version 0.12.8) [18] with default parameters except maximum two mismatches, unique mapping, and trimming from 101 to 76 nt for the three samples.

The present version of the Bowtie program (version 0.12.8) does not report gapped alignments. Thus, a read mapped on the genome was derived from a contiguous locus in the genome. However, some unmapped reads may arise from non-contiguous genomic loci, making them suitable for inspecting splice junctions. The unmapped reads were further aligned to the putative chimeric mRNAs by the Bowtie program with default parameters except maximum two mismatches and trimming from 101 to 76 nt for the three samples. The previously unmapped reads that were matched on the putative junctions with an overlap of at least 5 nt on either side of the RNA junction were remained for further analysis.

Another set of 49 nt single-end reads from three equivalently pooled RNA samples of skeletal muscle was analyzed as described above (Additional file 5). These samples were extracted during embryo collection at slaughter. The first, second, and last samples were pooled using equivalent amounts of RNA from three adult female Wuzhishan, Tongcheng, and Landrace pigs, respectively.

CV was calculated to represent the variance in the expression. The reads uniquely mapped on the pig genome and the junction reads were pooled together to reveal the read coverage along the transcript. The RS test was used to evaluate the difference in the expression levels between the male and female samples.

### DNA motif identification

The MEME software (version 4.6.1) [19] with default parameters (except DNA alphabet, zero or one occurrence of each motif per sequence, motif width between 10 and 30 nt, and maximum one motif to find) were used to search similar DNA sequences within two non-

continuous genomic sequences of chimeric mRNAs. Then, using the TOMTOM [20] tool, similar DNA sequences were compared with the database of 476 known motifs, the JASPAR CORE (version 12-Oct-2009).

## Additional files

**Additional file 1: Putative chimeric mRNAs.** The file lists the information on the putative chimeric mRNAs, including GenBank identifier, fusion type, overlap or gap between partners, junction site in the transcript, trans-splicing signals, and annotation in GenBank, as well as the chromosome location of the partners.

**Additional file 2: Putative chimeric mRNAs validated by RT-PCR.** The file shows the results of the RT-PCR assay.

**Additional file 3: Putative chimeric mRNAs validated by ESTs.** The table lists the chimeric mRNAs validated by ESTs. The columns of the table consist of the GenBank ID of putative chimeric mRNAs, the number of ESTs that overlap at least 20 nt of the sequence on either side of a putative fusion point, as well as the GenBank ID of each EST.

**Additional file 4: Pig putative chimeric mRNAs similar to known human chimeric mRNAs.** Pig putative chimeric mRNAs were aligned to known human chimeric transcripts. The table represents the BLAST results and the annotations of human chimeric mRNAs in ChimeraDB version 2.0.

**Additional file 5: Information on RNA-seq reads.** The raw reads, the cleaned reads, the uniquely mapped reads on the genome, the multi-mapped reads on the genome, the un-mapped reads on the genome, and the junction reads are shown in the table.

**Additional file 6: Pig chimeric mRNAs validated by RNA-seq reads.** The table shows the pig chimeric mRNAs validated by the RNA-seq reads derived from liver tissue and skeletal muscle. A fusion junction mapped by junction reads derived from at least three different start positions or at least three samples was considered a validated chimeric event. The junction reads should overlap with at least 5 nt of the sequence on either side of the chimeric junction.

**Additional file 7: Evaluation on the junction reads.** Figure (A) shows the count of the fusion junctions based on the count of junction reads. Figure (B) represents the count of the fusion junctions based on the count of positions that junction reads start.

**Additional file 8: Putative DNA motifs shared in the genomic regions of the partners.** The table gives some information on the putative DNA motifs shared in the genomic regions of the partners.

**Additional file 9: Similar DNA sequences that match known DNA motifs in the JASPAR CORE database.** The table shows similar DNA sequences matched known DNA motifs in the JASPAR CORE database.

**Additional file 10: Potential CTCF binding motifs.** The table shows potential CTCF binding motifs.

### Competing interests
The authors declare that they have no competing interests.

### Authors' contributions
LM designed, analysed, interpreted, and drafted the manuscript. SY and ZT handled the RNA sequencing. WZ and TZ performed the molecular genetic studies. KL conceived the study and participated in its design and coordination. All authors have read and approved the final manuscript.

### Acknowledgements
We thank Joshua Liao for his advice and two anonymous reviewers for their helpful suggestions on the manuscript. We also thank the China Postdoctoral Science Foundation (20110490045), the National Natural Science Foundation of China (30830080 and 31172189), and the Science Foundation of the Shihezi University (RCZX201137) for their support.

### Author details
[1]The Key Laboratory for Domestic Animal Genetic Resources and Breeding of Ministry of Agriculture of China, Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing 100193, P. R. China. [2]College of Life Science, Shihezi University, Xinjiang 832000, P. R. China.

Received: 11 December 2011 Accepted: 17 August 2012
Published: 28 August 2012

### References
1. Pane F, Intrieri M, Quintarelli C, Izzo B, Muccioli GC, Salvatore F: **BCR/ABL genes and leukemic phenotype: from molecular mechanisms to clinical correlations.** *Oncogene* 2002, **21:**8652–8667.
2. Druker BJ, Guilhot F, O'Brien SG, Gathmann I, Kantarjian H, Gattermann N, Deininger MW, Silver RT, Goldman JM, Stone RM, *et al*: **Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia.** *N Engl J Med* 2006, **355:**2408–2417.
3. Druker BJ, Tamura S, Buchdunger E, Ohno S, Segal GM, Fanning S, Zimmermann J, Lydon NB: **Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells.** *Nat Med* 1996, **2:**561–566.
4. Kantarjian H, Sawyers C, Hochhaus A, Guilhot F, Schiffer C, Gambacorti-Passerini C, Niederwieser D, Resta D, Capdeville R, Zoellner U, *et al*: **Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia.** *N Engl J Med* 2002, **346:**645–652.
5. Li H, Wang J, Mor G, Sklar J: **A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells.** *Science* 2008, **321**:1357–1361.
6. Di Segni G, Gastaldi S, Tocchini-Valentini GP: **Cis-and trans-splicing of mRNAs mediated by tRNA sequences in eukaryotic cells.** *Proc Natl Acad Sci* 2008, **105:**6864.
7. Li X, Zhao L, Jiang H, Wang W: **Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes.** *J Mol Evol* 2009, **68:**56–65.
8. Magrangeas F, Pitiot G, Dubois S, Bragado-Nilsson E, Cherel M, Jobert S, Lebeau B, Boisteau O, Lethe B, Mallet J, *et al*: **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* 1998, **273:**16005–16010.
9. Communi D, Suarez-Huerta N, Dussossoy D, Savi P, Boeynaems JM: **Cotranscription and intergenic splicing of human P2Y11 and SSF1 genes.** *J Biol Chem* 2001, **276:**16561–16566.
10. Moore RC, Lee IY, Silverman GL, Harrison PM, Strome R, Heinrich C, Karunaratne A, Pasternak SH, Chishti MA, Liang Y, *et al*: **Ataxia in prion protein (PrP)-deficient mice is associated with upregulation of the novel PrP-like protein doppel.** *J Mol Biol* 1999, **292:**797–817.
11. Finta C, Zaphiropoulos PG: **The human cytochrome P450 3A locus. Gene evolution by capture of downstream exons.** *Gene* 2000, **260:**13–23.
12. McManus CJ, Duff MO, Eipper-Mains J, Graveley BR: **Global analysis of trans-splicing in Drosophila.** *Proc Natl Acad Sci U S A* 2010, **107:**12975–12979.
13. Wernersson R, Schierup MH, Jorgensen FG, Gorodkin J, Panitz F, Staerfeldt HH, Christensen OF, Mailund T, Hornshoj H, Klein A, *et al*: **Pigs in sequence space: a 0.66X coverage pig genome survey based on shotgun sequencing.** *BMC Genomics* 2005, **6:**70.
14. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996–1006.
15. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A: **The UCSC genome browser database: update 2011.** *Nucleic Acids Res* 2011, **39:**D876–D882.
16. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: **A computer program for aligning a cDNA sequence with a genomic DNA sequence.** *Genome Res* 1998, **8**:967–974.
17. Kim N, Kim P, Nam S, Shin S, Lee S: **ChimerDB—a knowledgebase for fusion sequences.** *Nucleic Acids Res* 2006, **34:**D21–D24.
18. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10:**R25.
19. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching.** *Nucleic Acids Res* 2009, **37:**W202–W208.

20. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS: **Quantifying similarity between motifs.** *Genome Biol* 2007, **8:**R24.
21. Bryne JC, Valen E, Tang MHE, Marstrand T, Winther O, Da Piedade I, Krogh A, Lenhard B, Sandelin A: **JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update.** *Nucleic Acids Res* 2008, **36:**D102–D106.
22. Wallace JA, Felsenfeld G: **We gather together: insulators and genome organization.** *Curr Opin Genet Dev* 2007, **17:**400–407.
23. Ohlsson R, Renkawitz R, Lobanenkov V: **CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease.** *Trends Genet* 2001, **17:**520–527.
24. Klenova EM, Morse HC, Ohlsson R, Lobanenkov VV: **The novel BORIS+ CTCF gene family is uniquely involved in the epigenetics of normal biology and cancer.** *Seminars in Cancer Biology* 2002, **12:**399–414.
25. Dunn KL, Davie JR: **The many roles of the transcriptional regulator CTCF.** *Biochem Cell Biol* 2003, **81:**161–167.
26. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B: **Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.** *Cell* 2007, **128:**1231–1245.
27. Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell JA, Lopes S, Reik W, Fraser P: **Active genes dynamically colocalize to shared sites of ongoing transcription.** *Nat Genet* 2004, **36:**1065–1071.
28. Gingeras TR: **Implications of chimaeric non-co-linear transcripts.** *Nature* 2009, **461:**206–211.
29. Osborne CS, Chakalova L, Mitchell JA, Horton A, Wood AL, Bolland DJ, Corcoran AE, Fraser P: **Myc dynamically and preferentially relocates to a transcription factory occupied by Igh.** *PLoS Biol* 2007, **5:**e192.
30. Ren L, Shi M, Wang Y, Yang Z, Wang X, Zhao Z: **CTCF and cohesin cooperatively mediate the cell-type specific interchromatin interaction between Bcl11b and Arhgap6 loci.** *Mol Cell Biochem* 2012, **360:**243–251.
31. Ren L, Wang Y, Shi M, Wang X, Yang Z, Zhao Z: **CTCF Mediates the Cell-Type Specific Spatial Organization of the Kcnq5 Locus and the Local Gene Regulation.** *PLoS One* 2012, **7:**e31416.
32. Filippova GN, Qi CF, Ulmer JE, Moore JM, Ward MD, Hu YJ, Loukinov DI, Pugacheva EM, Klenova EM, Grundy P: **Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter its DNA-binding specificity.** *Cancer research* 2002, **62:**48.
33. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11:**R86.
34. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J: **Galaxy: a web-based genome analysis tool for experimentalists.** *Curr Protoc Mol Biol* 2010, **10:**11–21. Chapter 19:Unit 19.
35. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, *et al*: **Galaxy: a platform for interactive large-scale genome analysis.** *Genome Res* 2005, **15:**1451–1455.
36. Kent WJ: **BLAT—the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656–664.
37. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res* 2009, **19:**1639–1645.
38. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403–410.
39. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences.** *J Comput Biol* 2000, **7:**203–214.
40. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC Bioinforma* 2009, **10:**421.
41. Hahn Y, Bera TK, Gehlhaus K, Kirsch IR, Pastan IH, Lee B: **Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases.** *Proc Natl Acad Sci U S A* 2004, **101:**13257–13261.

doi:10.1186/1471-2164-13-429
**Cite this article as:** Ma *et al.*: Identification and analysis of pig chimeric mRNAs using RNA sequencing data. *BMC Genomics* 2012 **13**:429.